Semi-Supervised Learning for Facial Landmarks with Confidence and Augmentation Sifting Mechanisms

Hao-Wen Chia * and Jian-Jiun Ding†
*† National Taiwan University, Taipei, Taiwan
E-mail: * r08942119@ntu.edu.tw, † jjding@ntu.edu.tw, Tel: +886-2-33669652

Abstract— **Facial landmarks are important for various facial analysis tasks, including face recognition, age estimation, expression identification, medical image processing, and forensics. Influenced by the popularity of self-training in recent years, we propose a semi-supervised based human face landmark detection algorithm. First, we train a model with labeled data. Then, a huge amount of unlabeled data is fed into the model to generate pseudo labels. In order to filter out the pseudo labels with higher credibility, we propose a probabilistic model and determine how close the output feature distribution corresponding to the pseudo labels to the Gaussian distribution is. Then, the data with the pseudo labels are adopted to improve the performance. Moreover, different thresholds are applied for screening. Experiments show that, with the proposed semi-supervised based algorithm, the accuracy of landmark extraction can be improved.**

I. INTRODUCTION

Facial landmark detection is critical in many facial analysis tasks. There are many existing facial landmark detection algorithms, including conventional and learning-based ones. However, due to the limited amount of labeled data, it is still hard to take all the conditions into account. Thus, in this paper, a semi-supervised algorithm is proposed. We acquire unlabeled images from the CelebA Dataset [1] and train the network with the following four main steps: (1) train a teacher model on labeled images, (2) use the teacher model to generate pseudo labels on unlabeled images, (3) choose the pseudo-labeled images with higher credibility, and (4) train a student model on both labeled images and pseudo labeled images. Compared to other semi-supervised learning algorithms, the main contributions of the proposed method are that the confidence sifting mechanism is applied to sift the pseudo-labeled data with less confidence and that the augmentation sifting mechanism is applied to remove the data with less robustness.

Experiments show that, on the 300W-common test dataset [2], the proposed model can get a 5.5% lower error rate compared to the teacher model. The proposed model can improve the accuracy of facial landmark extraction.

II. RELATED WORKS

A. Summary of Facial Landmark Detection Methods

There are many existing facial landmark detection algorithms, including the active shape model using principal component analysis (PCA) [3] and the model of assembling regression trees [4]. The heatmap regression method, which regresses a heatmap generated from landmark coordinates, is widely used in facial landmark detection [5-8]. The 2D heatmap is generated by plotting a Gaussian distribution at the landmark location at each channel. Then, the predicted heatmap is used to infer the landmark coordinates. By heatmap regression, the accuracy of facial landmark extraction in the 2D space can be significantly improved.

Several CNN-based heatmap regression models were proposed in recent years. A joint bottom-up and top-down stacked hourglass network was proposed in [5]. Tang *et al.* proposed a quantized densely connected U-net [6] with fewer parameters than the stacked hourglass network. Valle *et al.* [7] combined the CNN models to assemble regression trees in a coarse-to-fine fashion. Li *et al.* [8] focused on the structure of the face by using the cascaded graph convolutional network. Wu *et al.* [9]. proposed a two-stage stacked hourglass model to predict the facial boundary map with boundary information.

The loss function assignment is also important for facial landmark detection. Based on different regression methods, it is suitable to regress with different loss functions. The adaptive wing loss [10], a two-stage loss function, was proposed in 2019. It is to emphasize the gradient of loss when the error is small. With the natural log function, the loss function can magnify the gradient of loss with small error. It is defined as follows

$$Awing(y,\hat{y}) = \begin{cases} \omega ln(1+\left|\frac{y-\hat{y}}{\epsilon}\right|^{\alpha-y}), & |y-\hat{y}|<\theta \\ A|y-\hat{y}|-C, & otherwise \end{cases} \quad (1)$$

where y and $\hat{y}$ are the values of the ground truth and the predicted value, respectively,

$$A = \omega\left(\frac{1}{1+\left(\theta/\epsilon\right)^{\alpha-y}}\right)(\alpha-y)\left(\left(\theta/\epsilon\right)^{\alpha-y-1}\right)\left(1/\epsilon\right),$$

$$C = (\theta A - \omega \ln\left(1+\left(\theta/\epsilon\right)^{\alpha-y}\right)),$$

and $\omega = 14, \theta = 0.5, \epsilon = 1, \alpha = 2.1$.

B. Semi-supervision

In recent years, the technique of semi-supervised learning has been widely adopted in classification problems [11]. Typically, when applying semi-supervised learning, first a teacher model is trained on labeled data. Then, a huge amount of unlabeled data is acquired and the teacher model is used to generate pseudo labels on these data. Then, the student model is trained on the combination of labeled and pseudo labeled data. With the use of semi-supervised learning, a very accurate model can be achieved even if the number of labeled images is limited initially.

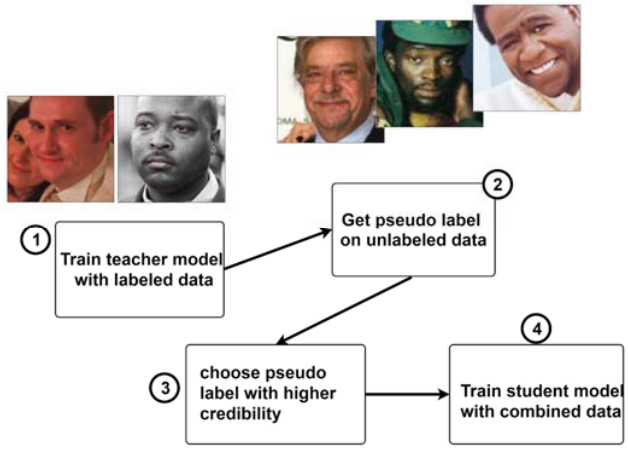

Fig. 1. Pipeline of the training phase for the proposed semi-supervised algorithm. In Step 3, we remove images with unconfident pseudo labels.

III. Proposed Landmark Detection System

The proposed facial landmark detection system has three stages. The first one is preprocessing. In this stage, the initial face image is cropped according to the ground truth bounding box. Then, data augmentation is performed. The second stage is to predict the heatmap. The third stage is post-processing. In this stage, the maximum value of the predicted heatmap at each landmark channel is applied to get the final landmark coordinates.

The proposed model is the improvement of the stacked hourglass model [10] by using semi-supervised learning with the confidence and the distance sifting mechanisms. The model is a fully convolutional network and can regress on the ground truth heatmap directly.

IV. Proposed Semi-supervised Based Approach

The proposed semi-supervised learning approach has the following four main steps: (1) Train a teacher model with labeled images. We use the images in the 300W [4] or WFLW [9] dataset as our labeled data. (2) Generate pseudo labels on unlabeled images with the teacher model. We choose unlabeled data from CelebA [3] about 3 to 4 times the labeled data. (3) Choose the data with the pseudo label that has higher credibility. (4) Train a student model on the combination of labeled images and pseudo labeled images.

In Step 3, choose the pseudo labeled image with higher credibility is a very important step, which can decide whether the prediction result is precise or not. Here we proposed two mechanisms to sift pseudo labeled data.

A. Confidence Sifting Mechanism

The teacher model is trained to predict the heatmap consisting of a 2D Gaussian distribution with the same variance along x and y axes centering on (x_k, y_k), where (x_k, y_k) is the ground truth of the k^{th} landmark. If the predicted heatmap patch around the prediction point is closer to the 2D Gaussian distribution, then the prediction result has higher confidence.

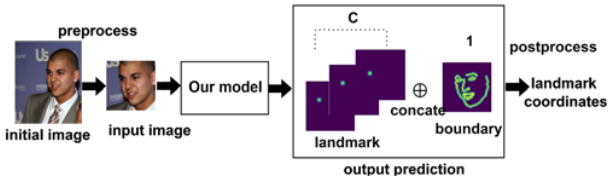

Fig. 2 Pipeline of the proposed facial landmark detection system. The model predicts the output heatmaps with size $(C+1)\times H\times W$, where C is the number of output landmark channels and H and W are the height and width of the output heatmap, respectively.

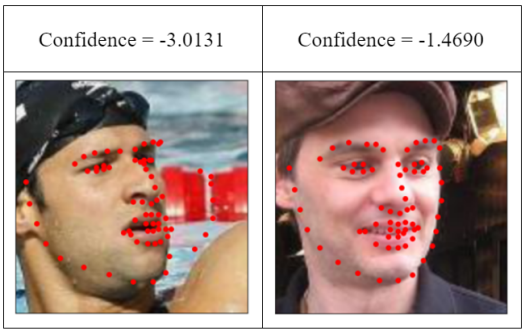

Fig. 3 Facial landmark detection results using the pseudo labeled data with different thresholds for confidence value. One can see that a higher threshold for confidence can achieve a better result.

We apply the Pearson chi-square test to evaluate the confidence:

$$Confidence = \frac{1}{K}\sum\nolimits_k -\chi_k^2\left(P \mid A;W\right), \tag{2}$$

where the chi-square function is:

$$\chi_k^2\left(P \mid A;W\right) = \sum\nolimits_i \frac{\left(E_i - \Phi_i\left(P \mid A;W\right)\right)^2}{E_i}, \tag{3}$$

E_i is a Gaussian heatmap, which is the template representing the ideal response, i is pixel index, A is the input image, W is the set of model parameters; P is the predicted output, and Φ_i is the cropped patch (with the same size of the Gaussian template) centered on P.

We use (2) to calculate the confidence of each prediction and remove the data with very low confidence from the pseudo labeled data.

In Fig. 3, we can find that the prediction result has higher accuracy when applying the pseudo labeled with higher confidence.

B. Augmentation Sifting Mechanism

According [12], if a predictor is valid, then the prediction result is robust to data augmentation. Hence, we propose a method based on this idea to verify the landmark coordinates: (1) Perform data augmentation on the input image. Data augmentation includes rotation ($\pm 5^{\circ}$), rescaling ($\pm 5\%$), and flipping (100%). (2) Predict landmark coordinates for both images. (3) Calculate the distance function of predicted landmarks of the two images.

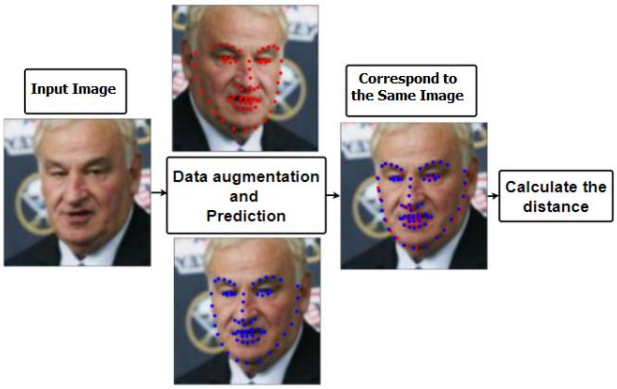

Fig. 4 The pipeline to verify with data augmentation.

The distance function is as follows:

$$Distance = \frac{1}{M}\sum_{i=1}^{M}\frac{\|p_i-\hat{p}_i\|_2^2}{d} \quad (4)$$

where p_i and $\hat{p}_i$ are the i^{th} predicted landmark coordinates for the original and augmented images, respectively, M is the number of landmarks, and d is the normalization factor. Here, we use the inter-pupil distance (the distance of eye centers) as the normalization factor.

We use (4) to calculate the distance of each pseudo labeled image and remove the one with too large distance. The flowchart of the proposed augmentation sifting mechanism is plotted in Fig. 4.

After these two sifting mechanisms, we get pseudo labels with higher confidence. Then, we train the student model with the combination of labeled and pseudo labeled data.

V. EXPERIMENTS

A. Labeled and Unlabeled Datasets and Metrics

We train the proposed approach with 300W dataset [2], which consists of the images with large variations of identity, expression, pose, occlusion, and illumination. A method that can achieve high accuracy on this dataset will always perform better in realistic data. The 300W dataset consists of **a training subset** (3148 images), **a common test subset** (554 images), and **a challenging test subset** (135 images). The images in these three datasets are not overlapped.

The unlabeled images are from the CelebA dataset [1]. The CelebA dataset contains 202,599 face pictures with 10,177 celebrity identities. We take 20,260 images from it.

The normalized mean error (NME) is commonly used to evaluate the localization quality. It is defined as:

$$NME(P,\hat{P}) = \frac{1}{M}\sum_{i=1}^{M}\frac{\|p_i-\hat{p}_i\|_2}{d} \quad (5)$$

where P and $\hat{P}$ are the ground truth and the predicted landmark coordinates for each image, respectively. p_i and $\hat{p}_i$ are the i^{th} landmark coordinates in the ground truth and in the prediction result, respectively, M is the number of landmarks, d is the normalization factor. We use the inter-pupil distance (the distance of eye centers) as the normalization factor.

Fig. 5 Visualization of the landmark detection results of the TCDCN [17] and the proposed algorithm on the 300W test dataset.

The failure rate (RF) is also a metric to evaluation the landmark detection result. If the NME is larger than a threshold, then it is considered a failed prediction. Here, we use 10% as the threshold.

B. Implementation Details

We use the bounding boxes provided by the dataset to crop the input images. The input size of the proposed model is $256 \times 256 \times 3$ and the output size is 64×64. During the training phase, we use RMSProp with an initial learning rate of 10^{-4} and momentum = 0. We use the adaptive wing loss in (1). We train the teacher model for 220 epochs and the learning rate is reduced to 10^{-5} and 10^{-6} after 80 and 160 epochs, respectively. The student model is trained for 80 epochs and the learning rate is reduced to 10^{-5} and 10^{-6} after 30 and 50 epochs, respectively.

C. Evaluation

From Table I, one can see that the best performance can be achieved if 10,600 pseudo labeled images are adopted (i.e., 52.32% of the pseudo labeled images are chosen). In Tables II and III, the performance of the proposed algorithm and some other landmark detection algorithms are compared. The results show that, with the proposed semi-supervised learning algorithm with confidence and augmentation sifting mechanisms, a more accurate facial landmark detection result can be achieved. Some visual results of facial landmark detection are shown in Fig. 5.

D. Ablation Study

The ablation study of the two adopted sifting mechanisms is shown in Table IV. With confidence and augmentation sifting mechanisms in (2) and (4), the result is the best on the 300W Common and Full test dataset. However, for the 300W Challenge test dataset, the best result is achieved when the augmentation sifting algorithm is adopted only. It is worth mentioning that, without confidence and augmentation sifting mechanisms, the result is much worse. Hence, the confidence and augmentation sifting mechanisms are helpful for improving the performance of landmark detection.

TABLE I: Comparison of adding different amounts of pseudo labeled data. The number in brackets represents the number of adopted data. The evaluation metric is the NME (%) here.

Method	Common	Challenge	All
(i) Labeled data (3148)	4.35	8.72	5.21
(ii) Labeled data (3148) + pseudo labeled data (7200)	4.20	**8.54**	5.05
(iii) Labeled data (3148) + pseudo labeled data (10600)	**4.11**	8.60	**4.99**
(iv) Labeled data (3148) + pseudo labeled data (14000)	4.29	8.7	5.15

TABLE II: Comparison with other classic methods. The evaluation metric is the NME (%) here (with inter-pupil normalization).

Method	Common	Challenge	All
CFAN	5.50	16.78	7.69
SDM	5.57	15.40	7.52
3DDFA [14]	6.15	10.59.	7.01
LBF [15]	4.95	11.98	6.32
CFSS [16]	4.73	9.98	5.76
TCDCN [17]	4.80	8.60	5.54
RCN [18]	4.67	8.44	5.41
Proposed	**4.11**	**8.6**	**4.99**

TABLE III: Evaluation on the test subsets of the 300W dataset. The fail rate (FR) is adopted for evaluation.

Method	FR (8%)
ESR [19]	17.00
cGPRT [20]	12.83
CFSS [16]	12.30
Proposed	11.9

TABLE IV: Ablation study of the proposed algorithm with and without the confidence and the augmentation sifting mechanisms.

Method	Common	Challenge	Full.
w/ Confidence, w/ Augment	**4.11**	8.60	**4.99**
w/o Confidence, w/ Augment	4.49	**8.38**	5.25
w/ Confidence, w/o Augment	4.19	8.88	5.11
w/o Confidence, w/o Augment	5.11	9.37	5.95

VI. CONCLUSIONS

In this paper, a semi-supervised algorithm with confidence and augmentation sifting mechanisms is proposed to detect the facial landmarks. With the two sifting mechanisms, only the pseudo labeled images with high confidence are adopted for semi-supervised learning. The proposed algorithm is helpful for improving the accuracy of facial landmark extraction and useful for facial image processing.

ACKNOWLEDGMENT

The authors thank for the support of Ministry of Science and Technology, Taiwan.

REFERENCES

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *IEEE Int. Conf. Computer Vision*, pp. 3730-3738, 2015.

[2] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," *IEEE Int. Conf. Computer Vision Workshops*, pp. 397-403, 2013.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, “Active shape models-their training and application,” *Computer Vision and Image Understanding*, vol. 61, issue 1, pp. 38-59, 1995.

[4] V. Kazemi and J. Sullivan. "One millisecond face alignment with an ensemble of regression trees," *IEEE Conf. Computer Vision and Pattern Recognition*. Pp. 1867-1874, 2014.

[5] A. Newell, K. Yang, and J. Deng. "Stacked hourglass networks for human pose estimation," *European Conf. Computer Vision*, Springer, Cham, pp. 483-499, 2016.

[6] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-nets for efficient landmark localization," *European Conf. Computer Vision*, Springer, Cham, pp. 339-354, 2018.

[7] R. Valle, J. M. Buenaposada, A. Valdes, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," *European Conf. Computer Vision*, Springer, Cham, pp. 585-606, 2018.

[8] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C. T. Cheng, J. Xiao, L. Lu, C. F. Kuo, and S. Miao, "Structured landmark detection via topology-adapting deep graph learning." *arXiv preprint arXiv:2004.08190*, 2020.

[9] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2129-2138, 2018.

[10] X. Wang, L. Bo, and L. Fuxin. "Adaptive wing loss for robust face alignment via heatmap regression," *IEEE/CVF Conf. Computer Vision*, pp. 6971-6981, 2019.

[11] Q. Xie, M. T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 10687-10698, 2020.

[12] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, “Improving landmark localization with semi-supervised learning,” *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1546-1555, 2018.

[13] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3467-3476, 2019.

[14] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 146-155, 2016.

[15] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1685-1692, 2014.

[16] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4998-5006, 2015.

[17] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, issue 5, pp. 918-930, 2015.

[18] S. Honari, J. Yosinski, P. Vincent, and C. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 5743-5752 2016.

[19] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” *Int. J. Computer Vision*, vol. 107, issue 2, pp. 177–190, 2014.

[20] D. Lee, H. Park, and C. D. Yoo, “Face alignment using cascade gaussian process regression trees,” *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4204–4212, 2015.