Layer-wise Supervised Neural Network for Face Alignment with Multi-task Regularization

Saijie Ni Botao Wang Hongkai Xiong
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
Email: {nsj161, botaowang, xionghongkai}@sjtu.edu.cn

***Abstract*—Convolutional neural networks (CNN) have achieved prominent performance in facial landmark detection in recent years. However, the training of such deep network is non-trivial due to the over-fitting problem caused by the insufficient training data and the diminishing gradients problem occurred in the back-propagation. To address these problems, we propose a multi-task learning framework with supervised neural networks to jointly detect facial landmarks with a set of related tasks. On the one hand, to handle the over-fitting problem, the proposed method takes the advantage of additional task labels to train the model in a multi-task learning fashion to generate a shared feature representation for high-level recognition tasks. On the other hand, in order to tackle the transparency and diminishing gradients problem, the proposed method enforces supervision to the intermediate layers of the network, augmenting the gradient signal propagated from the final layer. Experiments on public benchmarks validate the effectiveness of the proposed method.**

I. Introduction

Locating facial landmarks such as eyes, nose and mouth is essential for tasks like face recognition, face tracking and 3D face modeling. Although extensive studies have been made in facial landmark detection [1]–[3], it is still challenging for current approaches in unconstrained environments due to large head pose variations and partial occlusions. In general, face alignment research can be categorized into two groups: template-fitting methods [1] and regression based methods [2]. Regression based methods map the learned features to the facial landmark space after extracting features from the image. Cao et al. [2] employs cascaded fern regression with pixel-difference to predict landmark localization. It is improved in [3], where the regression problem is formulated with multiple deep models. To reduce the complexity, [4] proposes a multi-task learning framework with a single deep network. However, the relations between the main task and related tasks are not taken into consideration.

Multi-task learning is a way to train a universal model for several different but related tasks using a shared representation. It is generally acknowledged that the model learned in the multi-task learning fashion [5] has stronger generalization capability than the one learned in a single task fashion. Existing multi-task learning methods model the relationships among different tasks in two ways. One way is to assume the share of common parameters with other tasks such as a Bayesian model sharing a common prior [6]. The other way is to find latent feature representation among these tasks, for example, learning a sparse representation shared cross tasks [7]. Multi-task learning is an appealing approach which improves the generalization capability of a neural network with shared features, especially when the data is insufficient. Accordingly, the unique character of CNN model makes it possible to learn regression and classification tasks simultaneously.

In addition, most of the conventional feature extraction approaches for face alignment are handcrafted, which is tricky and lacking of flexibility. On the contrary, we derive a method to learn the features automatically in a data-driven manner, leveraging the deep learning architectures developed in recent years. Deep learning models are a class of multi-layer networks that can act on the raw input images to compute high-level representations automatically. One particular type of deep learning models that have achieved great practical success is the deep convolutional neural networks (CNNs) [8]. These models stack many layers of linear filters and underlying receptive fields followed by a nonlinear activation function, thereby computing abstract features. However, learning a deep CNN is usually associated with the estimation of millions of parameters, which often leads to: 1) uncertainty of transparency and robustness of the learned features; 2) over-fitting.

We address these two problems with the employment of supervision objective functions and multi-task feature learning. Specifically, this paper proposes a multi-task learning framework optimized with supervised neural network to jointly detect facial landmarks with a set of related tasks. The contribution of this paper is two-fold. First, we impose supervision on multiple layers, instead of only the final layer like conventional methods, of the neural network to alleviate the transparency deficiency of the learned features. The model is improved by adding auxiliary supervised layers connected to intermediate layers, which encourages discrimination of learned features in the lower stages. It strengthens the gradient signal passed in the back-propagation process, and provides extra guidance in the early layers to avoid the diminishing gradient problem. Second, we apply the multi-task learning strategy to improve the generalization capability of the model by utilizing auxiliary labels instead of samples. To be concrete, the features of the main task, i.e., facial landmark detection, is learned simultaneously with several other related high-level tasks, including glasses detection, smiling detection, gender

The work has been partially supported by the NSFC grants 61425011, U1201255, 61271218, 61529101, and Shu Guang project (13SG13).

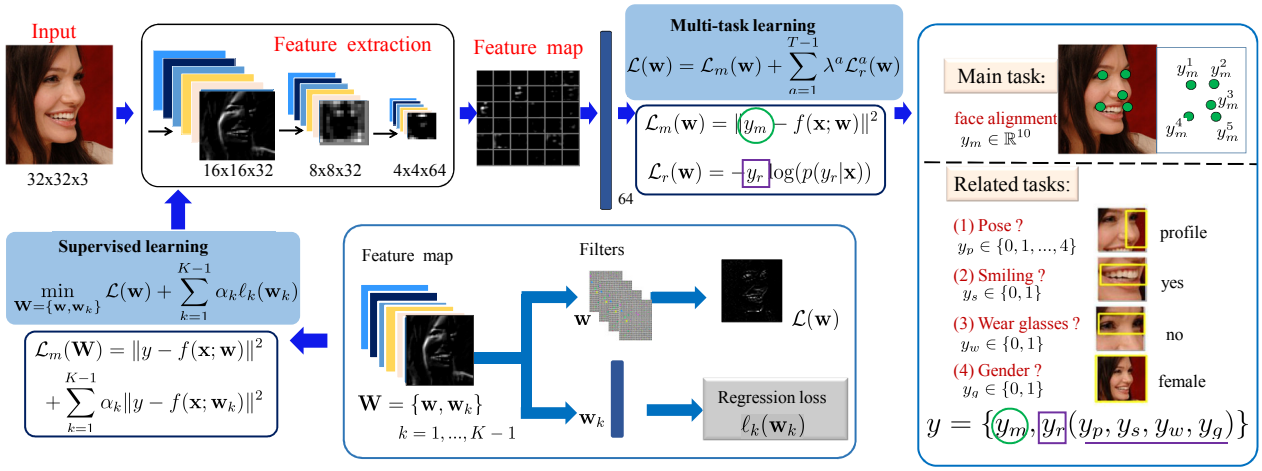

Fig. 1: The framework of the proposed model. Given an input face image, a CNN-based supervised network extracts shared features, which are the inputs to the facial landmark detection task and the related classification tasks.

prediction, and pose estimation, which provide additional regularization to the learning of network parameters.

II. Multi-task Learning of Supervised Networks

The framework of the proposed method is illustrated in Fig. 1. The main task is face alignment regression, which aims at detecting facial landmarks, including eyes, nose and mouth, in the face image. To obtain a robust feature representation, several related classification tasks, i.e., pose estimation, smiling detection, glasses detection, and gender prediction, are incorporated by associating additional labels to the training samples. Finally, a convolutional neural network is trained over these tasks in a multi-task learning fashion, which introduces direct supervision to the intermediate layers to alleviate the transparency and gradient diminishing problems.

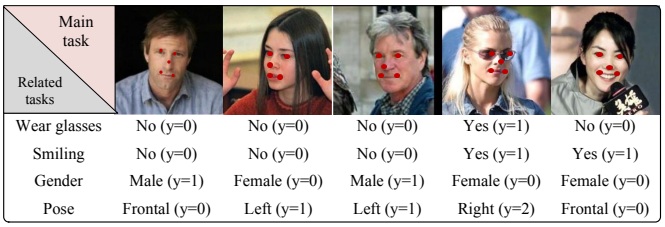

Fig. 2: Examples of the main task – face alignment, and the related appearance attribute tasks.

A. Multi-task Learning Formulation

The conventional multi-task learning aims to improve the performance of multiple related tasks by exploiting the intrinsic relationships among them. Accordingly, although the main objective of our model is to locate facial landmarks, we incorporate several related facial appearance tasks to derive a common feature representation. Consequently, the overall loss is the linear combination of the losses of the main task $\mathcal{L}_m$ and the related tasks $\mathcal{L}_r$.

$$\mathcal{L}(\mathbf{w}) = \mathcal{L}_m(\mathbf{w}) + \sum_{a=1}^{T-1} \lambda^a \mathcal{L}_r^a(\mathbf{w}) \tag{1}$$

where λ^a weights the cost of the a-th task, which can be determined by the correlation of the related tasks.

We denote the training set by $S = (\mathbf{x}_i^t, y_i^t)$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, where sample $\mathbf{x}_i^t \in \mathbb{R}^d$ denotes the raw input of the t-th task and $y_i^t \in \mathbb{R}$ denotes the corresponding ground truth label. An example of the labels for the main task and the related tasks is demonstrated in Fig. 2. We drop the subscript i for notational simplicity.

The regression task is to predict five facial points, so the coordinates of the landmarks are the target value. Squared-error is used as the cost function for the regression task

$$\mathcal{L}_m(\mathbf{w}) = \|y - f(\mathbf{x};\mathbf{w})\|^2$$

where $f(\mathbf{x})$ is the estimate of the five facial points.

For each related task, we employ the cross-entropy function,

$$\mathcal{L}_r(\mathbf{w}) = -y \log(p(y|\mathbf{x}))$$

where $p(y|\mathbf{x})$ is a softmax function, which models the class posterior probability.

Overall, the following optimization problem will be solved:

$$\min_{\mathbf{w}} \|y - f(\mathbf{x};\mathbf{w})\|^2 + \sum_{a=1}^{T-1} \lambda^a (-y \log(p(y^a|\mathbf{x}))) \tag{2}$$

Different types of tasks result in different output spaces, which is difficult to optimize by traditional MTL algorithms. Thus, we proposed an optimization algorithm with supervised learning to solve Eq. (2), which is described in Section II-B.

B. Supervised Learning with Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class of deep learning models that were designed to automatically capture highly nonlinear mappings between inputs and outputs.

Unlike traditional optimization algorithm, the unique structure of CNN makes multi-task with different types of loss functions (regression and classification) and shared representation possible. CNNs are usually composed of alternate convolutional and max-pooling layers to extract hierarchical features, followed by several fully connected layers. We denote a recursive function for each layer $k = 1, \cdots, K$ as

$$\mathbf{Z}_k = pool(\mathbf{Z}_{k-1} * \mathbf{W}_k + \mathbf{b}_k), \tag{3}$$

where $\mathbf{Z}_k$ is the feature map of the k-th layer, $\mathbf{W}_k$ denotes the filters to be learned, and $\mathbf{b}_k$ is the bias term. Note that $\mathbf{Z}_k$ is the shared representation between the main task and related tasks. Eq. (3) and Eq. (2) can be trained jointly to solve the minimization problem.

Despite of the attractive qualities of CNNs, there nonetheless remain some fundamental questions: 1) the features learned at hidden layers are not always discriminative; 2) overfitting occurs when the dataset is small; 3) gradients vanish when signal propagates back.

We adopt a deeply-supervised method, which provides integrated direct supervision to hidden layers rather than just to output layers. Instead of associating a classifier with each hidden layer [9], we employ regression supervision with every convolved response. To solve Eq. (1) more accurately and

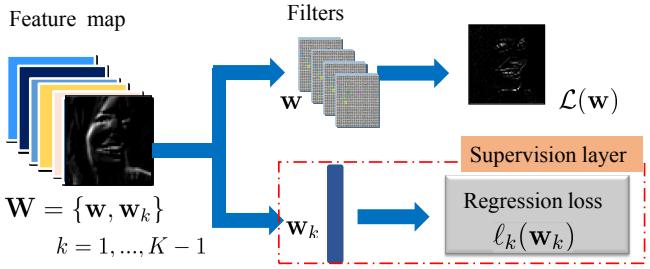

Fig. 3: Supervised learning-based feature extraction model.

transparently, we solve Eq. (4) instead,

$$\min_{\mathbf{W}=\{\mathbf{w},\mathbf{w}_k\}} \mathcal{L}(\mathbf{w}) + \sum_{k=1}^{K-1} \alpha_k \ell_k(\mathbf{w}_k) \tag{4}$$

where $\mathcal{L}$ is the output objective of the last layer, ℓ_k is the companion objective of the k-th layer which provides an additional constraint to propagate the supervision to early layers. Hence,

$$\mathcal{L}_m(\mathbf{W}) = \|y - f(\mathbf{x}; \mathbf{w})\|^2 + \sum_{k=1}^{K-1} \alpha_k \|y - f(\mathbf{x}; \mathbf{w}_k)\|^2$$

$$\mathcal{L}_r(\mathbf{w}) = -y \log(p(y|\mathbf{x}))$$

where $\mathbf{w}$ and $\mathbf{w}_k$ denote the weight parameters of filters with the output layer and the hidden layers, respectively.

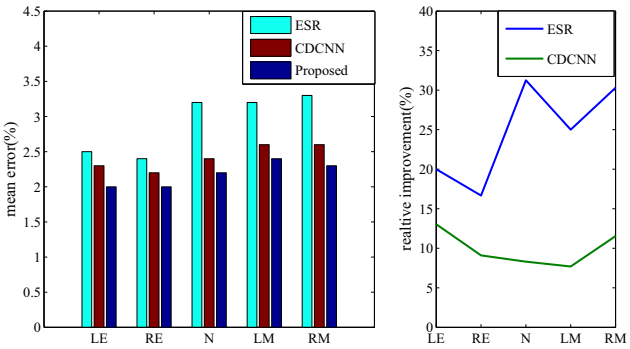

Fig. 4: The mean error and relative improvement comparison with ESR [2] and CDCNN [3] over five landmarks: left eye, right eye, nose, left mouth corner and right mouth corner on LFPW test dataset.

Note that the supervision is only employed based on the main task, which guarantees the priority of face alignment task. Moreover, rather than giving equal importance to the tasks, we explore the relationship of tasks to eventually maximize the performance of the main task.

Eq. (4) optimizes the tasks with stochastic gradient descent to firstly learn the shared representation and then propagate the errors to refine the representation. The main advantage of Eq. (4), the supervised learning method, is that it provides direct hidden layer supervision by introducing regression functions for each intermediate layer, which can be seen as a regularization term within the learning process. Meanwhile, the main task-only property of the supervision is of great importance for multi-task framework to optimize the desired problem.

III. Experiments

A. Datasets

The training set [4] is composed of 10000 face images from LFW dataset and the Internet. Each image is annotated with five landmarks, i.e., centers of the eyes, nose and corners of the mouth, as depicted in Fig. 2. All the coordinates are normalized by the size of bounding box so that their values range within $[0, 1]$. We augment the dataset by randomly selecting 16 bounding boxes, resizing them to the size of 32×32, and then applying a mirror transformation to double the training set. The testing set is composed of AFLW, AFW, and LFPW, all of which are widely used by previous facial landmark detection methods [1]–[4].

Performance is measured by the *average detection error* and the *failure rate* of the facial points. The mean error is measured by the distance between the predicted landmark position and the ground truth position normalized by inter-ocular distance or face width. Note that inter-ocular distance is not suitable when dataset has large variations because two eyes are not necessarily visible. So we use the width of the face bounding box for evaluation in Table I and switch to inter-ocular distance for fair comparison with other published results. Mean error larger than 5% is reported as a failure.

TABLE I: Mean Error Comparison of Model Variants(%)

Landmark	CNN base	Multi-task	Supervised	Proposed
left eye	4.82	2.58	2.46	**1.82**
right eye	4.79	2.47	2.32	**1.78**
nose	5.35	2.78	2.80	**2.15**
left mouth corner	5.50	3.01	2.98	**2.46**
right mouth corner	5.45	3.15	2.98	**2.50**
average error	5.18	2.80	2.71	**2.14**
failure rate	39.51	30.57	28.06	**23.24**

B. Network Structure

We use CNN as the basic building block of the system. The network takes the raw pixels as input and performs regression on the coordinates of the desired points and classification on the whole image. Three convolutional layers are stacked after the input 32×32 RGB image patch. Each convolutional layer applies 5×5 filters to the multichannel input image and each convolution layer is not directly connected to pooling layer. It splits to two ways, one of which propagates as normal to produce filter responses followed by 3×3 pooling and the other calculates the regression loss of this hidden layer with five facial points. The fourth layer is a fully-connected layer that has 64 neurons to represent the shared features of both main and related tasks. The final layer is composed of separate sub-networks for the face alignment problem and related classification tasks, as shown in Fig. 1.

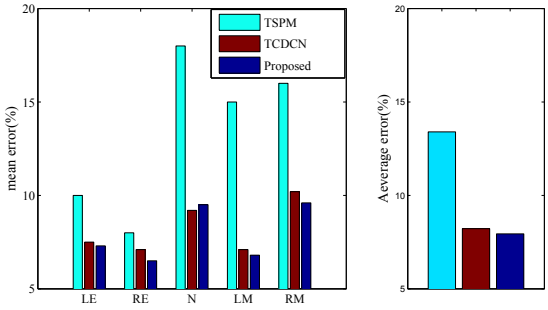

Fig. 5: The mean error comparison with TSPM [1] and TCDCN [4] over five landmarks: left eye, right eye, nose, left mouth corner and right mouth corner on AFW.

C. Implementation Details

We subtract the mean of each pixel from the image and then scale it to a standard deviation of 1. We adopt the training procedure used by Krizhevsky et al. [8]. The network is trained using stochastic gradient decent with a batch size of 128. The training process starts from the initial momentum 0.9 and learning rates 0.01 which is adapted during training. More specifically, if the overall loss is not reduced for 1000 iterations in a row, the learning rate is dropped by 50%. This procedure is repeated until convergence.

D. Experiment Results

To validate the effectiveness of the proposed multi-task framework with supervised learning, we evaluate the results visually and quantitatively. Table I shows the results with 2500 test images on AFLW. Single-task face alignment trained with traditional CNNs serves as the baseline. The first two variants are trained with either multi-task or supervised learning. The final model combines both multi-task and supervised learning, which is the performance of the proposed method. Obviously, the effectiveness of the multi-task regularization and layer-wise supervision is clearly validated in Table I. Fig. 6 presents some example results on AFLW. Furthermore, Fig. 4 compares with ESR [2] and CDCNN [3] methods on LFPW test dataset. Finally, Fig. 5 presents the similar result in comparison with TSPM [1] and TCDCN [4] methods on AFW.

Fig. 6: Example results on AFLW: faces with occlusion, pose variation, lighting condition variations and different expressions. The last three cases in red line are inaccurate examples.

IV. Conclusion

This paper presents a novel deep learning-based face alignment method with two primary contributions compared with conventional CNNs: 1) the multi-task learning framework with shared features of multiple related tasks to improve the generalization capability of the network; 2) the employment of supervision layers to strengthen the discriminativeness of the learned features. Experiments on face alignment validated the effectiveness of the proposed model in comparison with traditional convolutional neural networks of single task learning and other state-of-the-art methods.

References

[1] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int'l J. Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[3] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *CVPR*, 2013, pp. 3476–3483.

[4] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*, 2014, pp. 94–108.

[5] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[6] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *J. Machine Learning Research*, vol. 4, pp. 83–99, 2003.

[7] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[9] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *arXiv:1409.5185*, 2014.