Facial Landmark Detection by Deep Multi-task Learning

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang

Dept. of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Abstract. Facial landmark detection has long been impeded by the problems of occlusion and pose variation. Instead of treating the detection task as a single and independent problem, we investigate the possibility of improving detection robustness through multi-task learning. Specifically, we wish to optimize facial landmark detection together with heterogeneous but subtly correlated tasks, e.g. head pose estimation and facial attribute inference. This is non-trivial since different tasks have different learning difficulties and convergence rates. To address this problem, we formulate a novel tasks-constrained deep model, with task-wise early stopping to facilitate learning convergence. Extensive evaluations show that the proposed task-constrained learning (i) outperforms existing methods, especially in dealing with faces with severe occlusion and pose variation, and (ii) reduces model complexity drastically compared to the state-of-the-art method based on cascaded deep model [21].

1 Introduction

F10, 16], robust facial landmark detection remains a formidable challenge in the presence of partial occlusion and large head pose variations (Figure 1).

Facial landmark detection is traditionally approached as a single and independent problem. Popular approaches include template fitting approaches [8, 32, 27] and regression-based methods [3, 4, 9, 26, 31]. For example, Sun et al. [21] propose to detect facial landmarks by coarse-to-fine regression using a cascade of deep convolutional neural networks (CNN). This method shows superior accuracy compared to previous methods [2, 4] and existing commercial systems. Nevertheless, the method requires a complex and unwieldy cascade architecture of deep model.

We believe that facial landmark detection is not a standalone problem, but its estimation can be influenced by a number of heterogeneous and subtly correlated factors. For instance, when a kid is smiling, his mouth is widely opened (second image in Figure 1). Effectively discovering and exploiting such an intrinsically correlated facial attribute would help in detecting the mouth corners more accurately. Also, the inter-ocular distance is smaller in faces with large yaw

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 94–108, 2014.

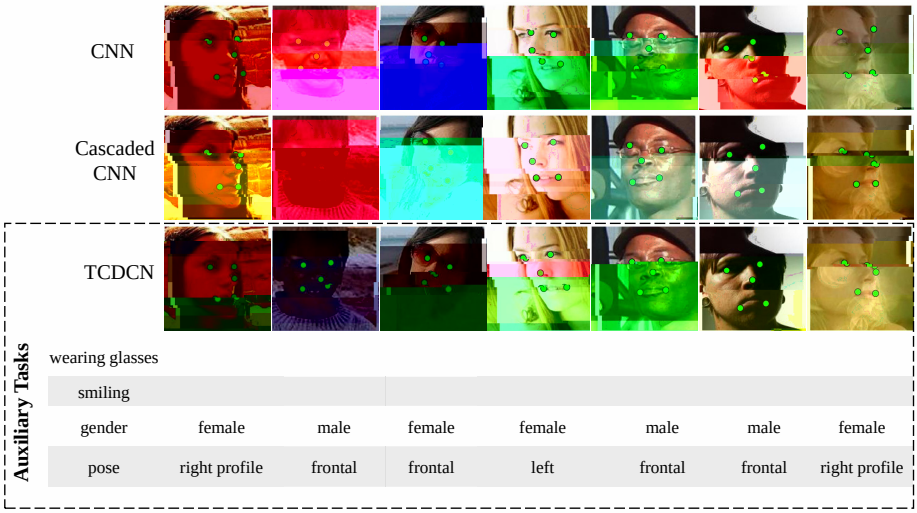

CNN
Cascaded CNN
TCDCN
Auxiliary Tasks
wearing glasses
smiling
gender female male female female male male female
pose right profile frontal frontal left frontal frontal right profile

Fig. 1.

limi(t)-343.3999939thes e-1.394999972rda

where λ^a denotes the importance coefficient of a-th task's error and the regularization terms are omitted for simplification. Beside the aforementioned difference, Eq.(1) and Eq.(2) are distinct in two aspects. First, different types of loss functions can be optimized together by Eq.(2), e.g. regression and classification can be combined, while existing methods [30] that employ Eq.(1) assume implicitly that the loss functions across all tasks are identical. Second, Eq.(1) allows data $\mathbf{x}_i^t$ in different tasks to have different input representations, while Eq.(2) focuses on a shared input representation $\mathbf{x}_i$. The latter is more suitable for our problem, since all tasks share similar facial representation.

In the following, we formulate our facial landmark detection model based on Eq.(2). Suppose we have a set of feature vectors in a shared feature space across tasks $\{\mathbf{x}_i\}_{i=1}^N$ and their corresponding labels $\{\mathbf{y}_i^r, y_i^p, y_i^g, y_i^w, y_i^s\}_{i=1}^N$, where $\mathbf{y}_i^r$ is the target of landmark detection and the remaining are the targets of auxiliary tasks, including inferences of 'pose', 'gender', 'wear glasses', and 'smiling'. More specifically, $\mathbf{y}_i^r \in \mathbb{R}^{10}$ is the 2D coordinates of the five landmarks (centers of the eyes, nose, corners of the mouth), $y_i^p \in \{0, 1, .., 4\}$ indicates five different poses $(0°, \pm 30°, \pm 60°)$, and $y_i^g, y_i^w, y_i^s \in \{0, 1\}$ are binary attributes. It is reasonable to employ the least square and cross-entropy as the loss functions for the main task (regression) and the auxiliary tasks (classification), respectively. Therefore, the objective function can be rewritten as

$$\underset{\mathbf{W}^r, \{\mathbf{W}^a\}}{\operatorname{argmin}} \ \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{y}_i^r - f(\mathbf{x}_i; \mathbf{W}^r)\|^2 - \sum_{i=1}^{N} \sum_{a \in A} \lambda^a y_i^a \log(p(y_i^a | \mathbf{x}_i; \mathbf{W}^a)) + \sum_{t=1}^{T} \|\mathbf{W}\|_2^2, \tag{3}$$

where $f(\mathbf{x}_i; \mathbf{W}^r) = (\mathbf{W}^r)^\top \mathbf{x}_i$ in the first term is a linear function. The second term is a softmax function $p(y_i = m|\mathbf{x}_i) = \frac{\exp\{(\mathbf{W}_m^a)^\top \mathbf{x}_i\}}{\sum_j \exp\{(\mathbf{W}_j^a)^\top \mathbf{x}_i\}}$, which models the class posterior probability ($\mathbf{W}_j^a$ denotes the jth column of the matrix), and the third term penalizes large weights ($W = \{\mathbf{W}^r, \{\mathbf{W}^a\}\}$). In this work, we adopt the deep convolutional network (DCN) to jointly learn the share feature space $\mathbf{x}$, since the unique structure of DCN allows for multitask and s6 0 0 9.9626 30

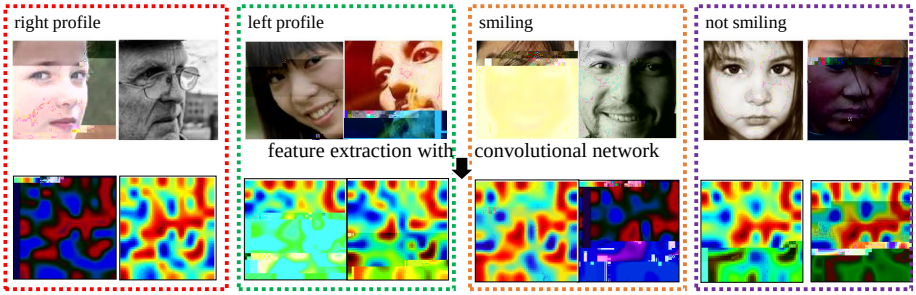


Fig. 2. The TCDCN extracts shared features for facial landmark detection and related tasks. The first row shows the face images and the second row shows the corresponding features in the shared feature space, where the face images with similar poses and attributes are close with each other. This reveals that the learned feature space is robust to pose, expression ('smiling'), and occlusion ('wearing glasses').

The TCDCN has four convolutional layers and a fully connected layer on the top. Each convolutional layer is followed by a max pooling layer. It is worth noting that in comparison to the cascaded CNN approach [21] that deploys 23 CNNs, our formulation constructs only one single CNN, of which complexity is similar to that of a CNN in the first-level cascade of [21]. We compare the complexity of these two approaches in Section 4.3. Further details of the network architecture is provided in Section 4 to facilitate re-implementation of the proposed model. Several pairs of face images and their features of the shared space of TCDCN are visualized in Figure 2, which shows that the learned features are robust to large poses and expressions. For example, the features of smiling faces or faces have similar poses exhibit similar patterns.

3.2 Learning Tasks-Constrained Deep Convolutional Network

A straightforward way to learn the proposed network is by stochastic gradient descent, whose effectiveness has been proven when a single task is present [12]. However, it is non-trivial to optimize multiple tasks simultaneously using the same method. The reason is that different tasks have different loss functions and learning difficulties, and thus with different convergence rates. Existing methods [30] solve this problem by exploring the relationship of tasks, e.g. through learning a covariance matrix of the weights of all tasks. Nevertheless, such methods can only be applied if the loss functions of all tasks are identical. This assumption is not valid when we wish to perform joint learning on heterogeneous tasks. Moreover, it is computationally impractical in dealing with weight vectors in high dimension.

Task-Wise Early Stopping: We propose an efficient yet effective approach to "early stop" the auxiliary tasks, before they begin to over-fit the training set and thus harm the main task. The intuition behind is that at the beginning

of the training process, the TCDCN is constrained by all tasks to avoid being trapped at a bad local minima. As training proceeds, certain auxiliary tasks are no longer beneficial to the main task after they reach their peak performance, their learning process thus should be halted. Note that the regularization offered by early stopping is different from weight regularization in Eq.(3). The latter globally helps to prevent over-fitting in each task through penalizing certain parameter configurations. In Section 4.2, we show that task-wise early stopping is critical for multi-task learning convergence even with weight regularization.

Now we introduce a criterion to automatically determine when to stop learning an auxiliary task. Let E^a_{val} and E^a_{tr} be the values of the loss function of task a on the validation set and training set, respectively. We stop the task if its measure exceeds a threshold ϵ as below

$$\frac{k \cdot \text{med}^t_{j=t-k} E^a_{tr}(j)}{\sum^t_{j=t-k} E^a_{tr}(j) - k \cdot \text{med}^t_{j=t-k} E^a_{tr}(j)} \cdot \frac{E^a_{val}(t) - \min_{j=1..t} E^a_{tr}(j)}{\lambda^a \cdot \min_{j=1..t} E^a_{tr}(j)} > \epsilon, \qquad (5)$$

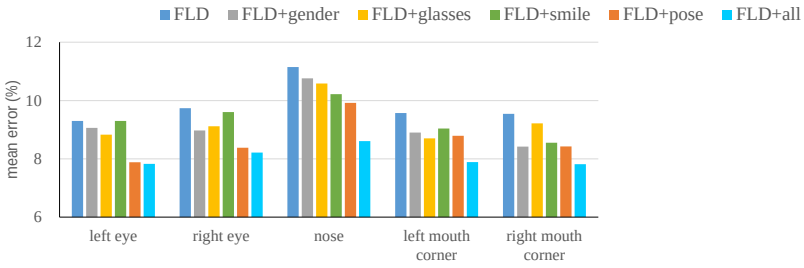
where t denotes the current iteration and k controls a training strip of length k. The 'med' denotes the function for calculating median value. The first term in Eq.(5) represents the tendency of the training error. If the training error drops rapidly within a period of length k, the value of the first term is small, indicating that training can be continued as the task is still valuable; otherwise, the first term is large, then the task is more likely to be stopped. The second term measures the generalization error compared to the training error. The λ^a is the importance coefficient of a-th task's error, which can be learned through gradient descent. Its magnitude reveals that more important task tends to have longer impact. This strategy achieves satisfactory results for learning deep convolution network given multiple tasks. Its superior performance is demonstrated in Section 4.2.

Learning Procedure: We have discussed when and how to switch off an auxiliary task during training before it over-fits. For each iteration, we perform stochastic gradient descent to update the weights of the tasks and filters of the network. For example, the weight matrix of the main task is updated by $\Delta \mathbf{W}^r = -\eta \frac{\partial E^r}{\partial \mathbf{W}^r}$ with η being the learning rate ($\eta = 0.003$ in our implementation), and $\frac{\partial E^r}{\partial \mathbf{W}^r} = (\mathbf{y}^r_i - (\mathbf{W}^r)^\mathsf{T} \mathbf{x}_i)\mathbf{x}^\mathsf{T}_i$. Also, the derivative of the auxiliary task's weights can be calculated in a similar manner as $\frac{\partial E^a}{\partial \mathbf{W}}$

Prediction: First, a test face image $\mathbf{x}^0$ is projected to the shared space to obtain $\mathbf{x}^l$. Second, we predict the landmark positions by $(\mathbf{W}^r)^\top \mathbf{x}^l$ and the results of the auxiliary tasks by $p(y^a|\mathbf{x}^l; \mathbf{W}^a)$. This process is efficient and its complexity is discussed in Section 4.3.

4 Implementation and Experiments

Network Structure: Figure 3 shows the network structure of TCDCN. The input of the network is 40×40 gray-scale face image. The feature extraction stage contains four convolutional layers, three pooling layers, and one fully connected layer. Each convolutional layer contains a filter bank producing multiple feature maps. The filter weights are not spatially shared, that means a different set of filters is applied at every location in the input map. The absolute tangent function is selected as the activation function. For the pooling layers, we conduct


FLD
FLD+gender
FLD+glasses
FLD+smile
FLD+pose
FLD+all
mean error (%)
12
10
8
6
left eye
right eye
nose
left mouth corner
right mouth corner

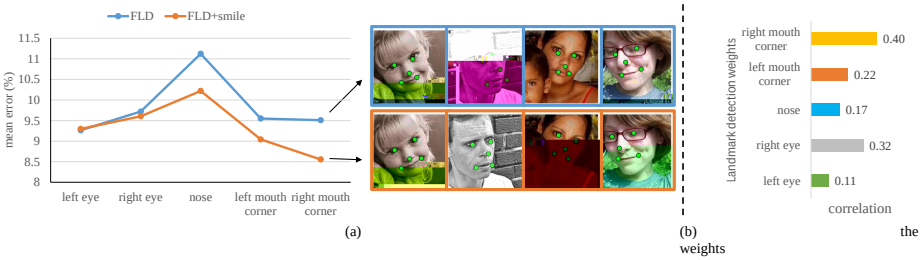


Fig. 5. FLD vs. FLD+smile. The smiling attribute helps detection more on the nose and corners of mouth, than the centers of eyes, since 'smiling' mainly a ects the lower part of a face.

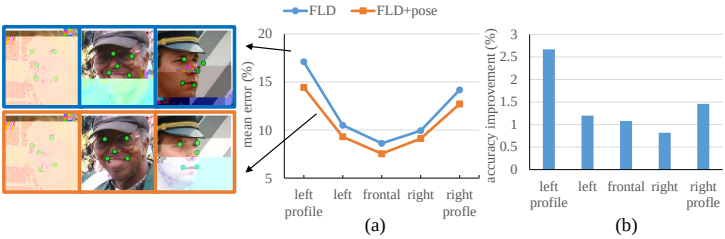


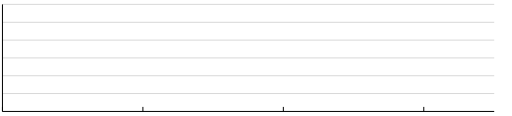
Fig. 6. FLD vs. FLD+pose. (a) Mean error in di erent poses, and (b) Accuracy improvement by the FLD+pose in di erent poses.

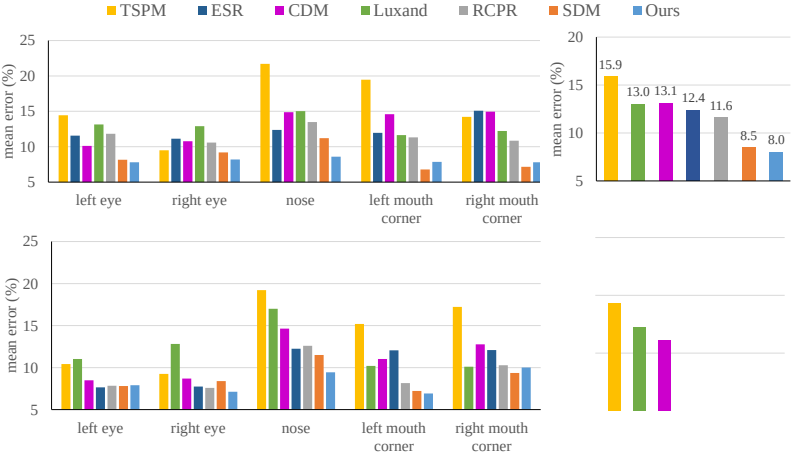
weight vectors, which are learned to predict the positions of the mouth's corners have high correlation with the weights of 'smiling' inference. This demonstrates that TCDCN implicitly learns relationship between tasks.

FLD vs. FLD+pose: As observed in Figure 6(a), detection errors of FLD increase along with the degree of head pose deviation from the frontal view to profiles, while these errors can be partially recovered by FLD+pose as depicted in Figure 6(b).

4.2 The Benefits of Task-Wise Early Stopping

To verify the effectiveness of the task-wise early stopping, we train the proposed TCDCN with and without this technique and compare the landmark detection rates in Figure 7(a), which shows that without task-wise early stopping, the accuracy is much lower. Figure 7(b) plots the main task's loss errors of the training set and the validation set within 2,600 iterations. Without early stopping, the training error converges slowly and exhibits substantial oscillations. However, convergence rates of both the training and validation sets are fast and stable




TSPM
ESR
CDM
Luxand
RCPR
SDM
Ours
mean error (%)
25
20
15
10
5
left eye
right eye
nose
left mouth corner
right mouth corner
mean error (%)
20
15
10
5
15.9
13.0
13.1
12.4
11.6
8.5
8.0
mean error (%)
25
20
15
10
5
left eye
right eye
nose
left mouth corner
right mouth corner

take RCPR [3] as an example. Instead of drawing training samples randomly as initialization as did in [3], we initialize RCPR by first applying TCDCN on the test image to estimate the five landmarks. We compare the results of RCPR with and without TCDCN as initialization on the COFW dataset [3], which includes 507 test faces that are annotated with 29 landmarks. Figure 11(a) shows the relative improvement for each landmark on the COFW dataset and Figure 11(b) visualizes several examples. It is demonstrated that with our robust initialization, the algorithm can obtain improvement in difficult cases with rotation and occlusion.

5 Conclusions

Instead of learning facial landmark detection in isolation, we have shown that more robust landmark detection can be achieved through joint learning with heterogeneous but subtly correlated tasks, such as appearance attribute, expression, demographic, and head pose. The proposed Tasks-Constrained DCN allows errors of related tasks to be back-propagated in deep hidden layers for constructing a shared representation to be relevant to the main task. We have shown that task-wise early stopping scheme is critical to ensure convergence of the model. Thanks to multi-task learning, the proposed model is more robust to faces with severe occlusions and large pose variations compared to existing methods. We have observed that a deep model needs not be cascaded [21] to achieve the better performance. The lighter-weight CNN allows real-time performance without the

10. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: CVPR, pp. 2578–2585 (2012)
11. Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCV Workshops, pp. 2144–2151 (2011)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
13. Li, H., Shen, C., Shi, Q.: Real-time visual tracking using compressive sensing. In: CVPR, pp. 1305–1312 (2011)
14. Liu, X.: Generic face alignment using boosted appearance model. In: CVPR (2007)
15. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with GaussianFace. Tech. rep., arXiv:1404.3840 (2014)
16. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: CVPR, pp. 2480–2487 (2012)
17. Luo, P., Wang, X., Tang, X.: A deep sum-product architecture for robust facial attributes analysis. In: CVPR, pp. 2864–2871 (2013)
18. Luxand Incorporated: Luxand face SDK. K. K. K. K.21