Optimizing Facial Landmark Detection by Facial Attribute Learning

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

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

Abstract. Instead of treating the facial landmark 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 multiple facial attributes learning. 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 [5]. [1]

1 Introduction

Facial landmark detection is a fundamental component in many face analysis tasks, such as face verification. Though great strides have been made in this field [1,2,5], robust facial landmark detection remains a formidable challenge in the presence of partial occlusion and large head pose variations. Facial landmark detection is traditionally approached as a single and independent problem [1,2]. However, we believe that it is not standalone, but its estimation can be influenced by a number of heterogeneous and subtly correlated factors such as facial attributes or expressions. For example, when a kid is smiling, his mouth is opened. Exploiting such an correlated facial attribute would help in detecting the mouth corners.

This study aims to investigate the possibility of *optimizing facial landmark detection with related/auxiliary tasks*, which include head pose estimation, gender classification, age estimation, facial expression recognition, or more generally, facial attribute inference. In particular, we propose a *Tasks-Constrained Deep Convolutional Network* (TCDCN) to jointly optimize facial landmark detection with a set of related tasks. Specifically, we formulate a task-constrained loss function to allow the errors of related tasks to be back-propagated jointly to improve the generalization of landmark detection. To accommodate related tasks with different convergence rates, we devise a task-wise early stopping criterion to facilitate learning convergence. The proposed approach outperforms the cascaded CNN model [5] and other existing methods [1,2,6,7,8] in our experiments.

[1] The full version of this paper has been accepted by the ECCV main conference.

2 Tasks-Constrained Facial Landmark Detection

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 $a \in A$, including inferences of 'pose', 'gender', 'wear glasses', and 'smiling'. More specifically, $\mathbf{y}_i^r$ is the coordinates of the landmarks, $y_i^p \in \{0, 1, .., 4\}$ 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

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

where $f(\mathbf{x}_i; \mathbf{W}^r) = (\mathbf{W}^r)^\mathsf{T}\mathbf{x}_i$ in the first term is a linear function. The second term is a softmax function,modeling the class posterior probability. 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 shared representation.

A straightforward way to learn the proposed network is by stochastic gradient descent. However, it is non-trivial to optimize multiple tasks simultaneously using this method. The reason is that different tasks have different loss functions and learning difficulties, and thus with different convergence rates. Here, 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. Now we introduce a criterion to automatically determine when to stop learning an auxiliary task. Let E_{val}^a and E_{tr}^a 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

$$\frac{k \cdot \text{med}_{j=t-k}^{t} E_{tr}^a(j)}{\sum_{j=t-k}^{t} E_{tr}^a(j) - k \cdot \text{med}_{j=t-k}^{t} E_{tr}^a(j)} \cdot \frac{E_{val}^a(t) - \min_{j=1..t} E_{tr}^a(j)}{\lambda^a \cdot \min_{j=1..t} E_{tr}^a(j)} > \epsilon, \tag{2}$$

where t denotes the current iteration and k controls a training strip of length k. The 'med' denotes the function for calculating median value.

3 Experiments

The training dataset we use is identical to [5], with 10,000 face images from the web. Each face is additionally annotated with attributes of the related tasks. This dataset, known as Multi-Task Facial Landmark (MTFL) dataset, and the landmark detector will be released for research usage. We report our results using mean error and failure rate. The mean error is measured by the localization error (normalized by the inter-ocular distance). Mean error larger than 10% is reported as a failure.

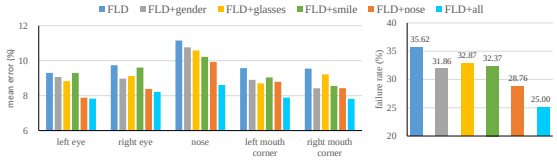


Fig. 1. Comparison of different model variants of TCDCN: the mean error over different landmarks, and the overall failure rate.

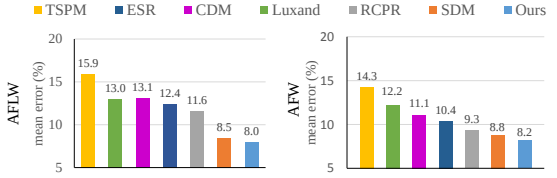


Fig. 2. Comparison with RCPR [1], TSPM [8], CDM [7], Luxand [4], and SDM [6] on AFLW [3] and AFW [8] datasets.

3.1 Evaluating the Effectiveness of Learning with Related Task

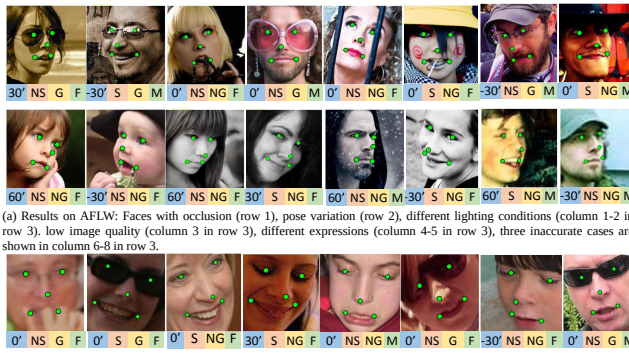
To examine the influence of related tasks, we evaluate five variants of the proposed model. The first variant is trained only on facial landmark detection, and another four model variants on facial landmark detection along with the auxiliary task of recognizing 'pose', 'gender', 'wearing glasses', and 'smiling', respectively. The full model is trained using all the related tasks. For simplicity, we name each variants by facial landmark detection (FLD) and the related task, such as "FLD", "FLD+pose", "FLD+all". We employ 3,000 faces form the popular AFLW [3] for evaluation. It is evident from Figure 1 that optimizing landmark detection with related tasks is beneficial. In particular, FLD+all outperforms FLD by a large margin, with a reduction over 10% in failure rate. When single related task is present, FLD+pose performs the best. This is not surprising since pose variation affects locations of all landmarks globally and directly. The other related tasks such as 'smiling' and 'wearing glasses' have comparatively smaller influence, since they mainly capture local information of the face.

3.2 Comparison with the Cascaded CNN [5]

Similar to Section 3.1, we employ AFLW images for evaluation. Note that we use the same 10,000 training faces as in the cascaded CNN method. Thus the only difference is that we exploit a multi-task learning approach. We achieve higher accuracy with mean error of **8.04**, while that of the cascaded CNN is **8.97**. On the other hand, the proposed method only has one CNN, whereas the cascaded CNN [5] deploys 23 CNNs in the cascade. Hence, TCDCN has much lower computational cost. The cascaded CNN requires 0.12s to process a face on an Intel Core i5 CPU, whilst TCDCN only takes 17ms, which is *7 times faster*.

3.3 Comparison with other State-of-the-art Methods

We compare against: (1) Robust Cascaded Pose Regression (RCPR) [1] (2) Tree Structured Part Model (TSPM) [8](3) A commercial software, Luxand face SDK [4]; (4) Explicit Shape Regression (ESR) [2]; (5) A Cascaded Deformable Model (CDM) [7]; (6) Supervised Descent Method (SDM) [6]. The evaluation is



(a) Results on AFLW: Faces with occlusion (row 1), pose variation (row 2), different lighting conditions (column 1-2 in row 3). low image quality (column 3 in row 3), different expressions (column 4-5 in row 3), three inaccurate cases are shown in column 6-8 in row 3.

(b) Results on AFW

Fig. 3. Example detections by the proposed model on AFLW [3] and AFW [8] images. The labels below each image denote the tagging results for the related tasks: $(0°, \pm 30°, \pm 60°)$ for pose; S/NS = smiling/not-smiling; G/NG = with-glasses/without-glasses; M/F = male/female.

based on AFLW [3] and AFW [8]. Figure 2 shows that TCDCN outperforms all the state-of-the-art methods. Figure 3(a) shows several examples of TCDCN's detection, with additional tags generated from related tasks. We observe that our method is robust to faces with large pose variation, lighting, and occlusion.

4 Conclusions

We have shown that more robust landmark detection can be achieved through joint learning with heterogeneous but subtly correlated tasks, such as facial attributes. 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. Thanks to multi-task learning, the proposed model is more robust to faces with severe occlusions and large pose variations compared to existing methods.

References

1. Burgos-Artizzu, X.P., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: ICCV. pp. 1513–1520
2. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR. pp. 2887–2894 (2012)
3. 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)
4. Luxand Incorporated: Luxand face SDK, http://www.luxand.com/
5. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. pp. 3476–3483 (2013)
6. Xiong, X., De La Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR. pp. 532–539 (2013)
7. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: ICCV. pp. 1944–1951 (2013)
8. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. pp. 2879–2886 (2012)