Headline

大家好：

这是**2018年度第1篇Arxiv Weekly**。

本文是 人脸 方向的文章。

Highlight

> 本文通过多轮廓估计层（**shape prediction layer, SPL**）更好地在遮挡、多相外观等复杂数据条件下解决了**face alignment**问题。

Information

Title

Deep Multi-Center Learning for Face Alignment

Link

https://arxiv.org/pdf/1808.01558.pdf

Codes

https://github.com/ZhiwenShao/MCNet-Extension

Source

- 上海交通大学（Shanghai Jiao Tong University）
- 华东师范大学（East China Normal University）

Introduction

人脸特征点（Facial Landmarks）之间有较强的相关性，显然一个特征点的位置可以从其相邻特征点的位置进行后验推断。然而传统的DL算法一般只使用一个FC层（也即所谓的轮廓估计层shape prediction layer, SPL）进行人脸特征点估计。

本文提出了face alignment的最新架构：包含多个轮廓估计层的多中心点架构（Multi-Center Learning with multiple shape prediction layers， MCL）。具体来说，每个SPL主要负责检测一簇语义相关的特征点，其中难以检测的特征点优先检测，而后对每一簇特征点分别针对性优化。

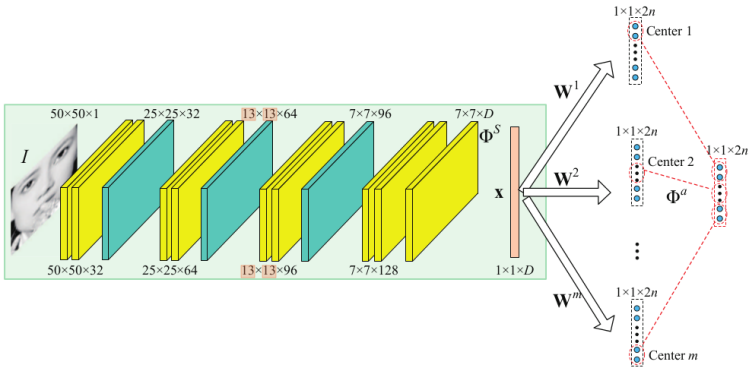
另外，为了降低模型的复杂度，本文利用模型组合的思路把所有的sub-SPL整合成一个大SPL。

实验表明本文的方法能够更好地解决面部遮挡、同一个人不同的appearance等复杂的face alignment问题[复杂例子和alignment结果如下图所示，下图为本篇的前序工作MCNet文章中的图例]，并且维持实时性。

Keys

本文有两个需要解析的关键点：网络结构和训练方式。

1.本文采用的网络结构如下：



可以看到，预处理之后的图片输入网络后，通过三组(Conv, Conv, MaxPooling)模块后，输入后续的三层卷积中，再通过global pooling得到最终的特征。这个特征被同时送入m个SPL，最终通过assemble获得n个final landmarks。[其中每个Conv后都附加了BN和ReLU]

2.本文的训练过程有如下要点：

2.1 训练总体流程

Algorithm 1 Multi-Center Learning Algorithm.

Input: A network MCL, Ω^t, Ω^v, initialized Θ.
Output: Θ.

1: Pre-train shared layers and one shape prediction layer until convergence;
2: Fix the parameters of the first six convolutional layers and fine-tune subsequent layers until convergence;
3: Fine-tune all the layers until convergence;
4: **for** $i = 1$ to m **do**
5: Fix Θ^S and fine-tune the i-th shape prediction layer until convergence;
6: **end for**
7: $\Theta = \Theta^S \cup \mathbf{W}^a$;
8: Return Θ.

分为pre-train---weighting finetune---multi-center finetune---model assembling几个阶段，下面分别解析。

2.2 Loss函数设计

$$E = \sum_{j=1}^{n} w_j[(y_{2j-1} - \hat{y}_{2j-1})^2 + (y_{2j} - \hat{y}_{2j})^2]/(2d^2), \quad (4)$$

是一个含有weight，也即w_j的$;\text{L}_2$ loss。

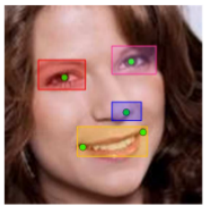
2.3 weighting finetune设计

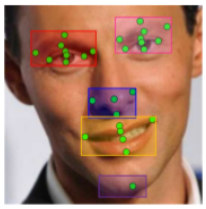
在训练的Step2中，前面六层卷积被固定，后三层卷积先进行finetune；而后在Step3中，整个网络进行最终的finetune。之所以称为weighting finetune，是因为在微调的时候，loss越大的路径上调整力度越大，由如下的weighting控制。这样一来，能够将有限的力量集中在challenging case上。

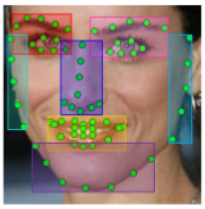
$$w_j = n\epsilon_j^b / \sum_{j=1}^{n} \epsilon_j^b. \quad (5)$$

2.4 multi-center finetune设计

进行到这一步后，前面的特征抽象网络全部训练完毕。开始训练SPL层。所谓的multi-center，是指把最终的landmarks分配到不同的几个簇，每个簇是一块面部特征对应的特征点集合，例如眼睛、鼻子、嘴、脸颊轮廓等[如下图所示]。然后每个SPL层针以某个簇为优化的中心，着力准确地刻画自己簇中所有landmark。


(a) 5 landmarks.


(b) 29 landmarks.


(c) 68 landmarks.

而进行multi-center focus的方案，也是设计启发式的weighting参数，最终公式如下。这个公式能够保障SPL对本簇内landmark的优化力度是簇外landmark的alpha倍。[alpha >> 1]

$$w_j = \begin{cases} w_{P^{i(c)}} |P^{i(c)}| \cdot \epsilon_j^w / \sum_{j \in P^{i(c)}} \epsilon_j^w, & j \in P^{i(c)}, \\ w_{P^{i(r)}} (n - |P^{i(c)}|) \cdot \epsilon_j^w / \sum_{j \in P^{i(r)}} \epsilon_j^w, & j \in P^{i(r)}. \end{cases} \quad (9)$$

2.5 model assembling方案

最终你会得到m个SPL，它们有不彼此重叠的center/簇。因此进行合并的最自然方案，就是使得最终生成的所有landmark都来自自己簇对应的SPL。而因为SPL之间簇没有重叠，这个融合过程可以通过直接融合weighting进行。[这里有些绕，可能需要一定时间理解]

$$\mathbf{W}^a = \Phi^a(\mathbf{W}^1, \cdots, \mathbf{W}^m), \quad (3)$$

where $\mathbf{W}^a = (\mathbf{w}_1^a, \mathbf{w}_2^a, \cdots, \mathbf{w}_{2n}^a)$ is the assembled weight matrix. Specifically, $\mathbf{w}_{2j-1}^a = \mathbf{w}_{2j-1}^i$, $\mathbf{w}_{2j}^a = \mathbf{w}_{2j}^i$, $j \in P^{i(c)}$, $i = 1, \cdots, m$, where $P^{i(c)}$ is the i-th cluster of indexes of landmarks. The final prediction of our MCL is $\hat{\mathbf{y}} = \mathbf{W}^{aT}\mathbf{x}$.

2.6 weight matrix与反传的结合方式

注意这个部分原文的表述比较有误导性。代表weighting的w和代表FC layer的$\textbf{W}$是完全没关系的，weighting会自然融合在反传里，起到的作用就是提高被focus的landmark的lr。

$$\mathbf{w}_k = \mathbf{w}_k - \eta w_j(\hat{y}_k - y_k)\mathbf{x}/d^2, \qquad (11)$$

Results

下图给出了本文算法和常见同类算法的Mean Error对比。

Method	AFLW 5 landmarks	COFW 29 landmarks	IBUG 68 landmarks
ESR [4]	12.4*	11.2*	17.00*
SDM [5]	8.5*	11.14*	15.40*
Cascaded CNN [8]	8.72	-	-
RCPR [6]	11.6*	8.5	17.26*
CFAN [10]	7.83[2]	-	16.78*
LBF [7]	-	-	11.98
cGPRT [19]	-	-	11.03
CFSS [42]	-	-	9.98
TCDCN [11], [12]	8.0	8.05	8.60
ALR [43]	7.42	-	-
CFT [26]	-	6.33	10.06
Wu et al. [24]	-	**5.93**	-
Honari et al. [25]	5.60	-	8.44
RAR [27]	7.23	6.03	**8.35**
Sim-Wu et al. [44]	-	6.40	-
MCL	**5.38**	6.00	8.51

下图给出了本文算法和前序工作MCNet的Mean Error对比：

Method	AFLW 5 landmarks	COFW 29 landmarks	IBUG 68 landmarks
pre-BM [13]	**5.61**	6.40	9.23
BM	5.67	**6.25**	**8.89**

Insights

文章中设计的结构，在作者分析中有三个主要的好处：

1. 和主流CNN网络相比，本文网络结构显然轻量级很多，因此无论是training还是inference都会更加高效。
2. 过深的网络结构会削弱spatial information的信息，抽象出来的更多是semantic information，因此本文的网络更加适用于facial landmark这样对spatial information要求很高的任务。
3. overfitting问题上，本文网络有天然的优势。

除了从这些points中收到启发，本身文中设计的weighting机制也是比较精巧和值得分析的。