文章编号：2095-4980(2021)02-0295-08

基于卷积神经网络的大姿态人脸对齐方法

蓝　敏

(长沙职业技术学院 经济贸易与信息技术学院，湖南 长沙 410217)

摘　要：大姿态人脸对齐是人脸识别和三维人脸重构等很多重要视觉任务的先决条件。现有的对齐方法大多使用二维界标位置来进行对齐，且使用的界标数量有限，影响大姿态人脸对齐的准确性。提出一种采用三维形变模型(3DMM)来表示二维人脸图像，将具有任意姿态的人脸对齐问题建模为基于 3DMM 的拟合问题。采用基于卷积神经网络(CNN)的级联回归方法学习二维人脸图像及其表示之间的映射关系。提出 2 种新的姿态不变局部特征作为卷积神经网络学习的输入层，通过训练得到 CNN 用于大姿态人脸对齐。在 2 个经典的人脸图像数据集上的仿真实验结果表明，与目前最新的人脸对齐方法相比，该方法的效果较优。

关键词：人脸对齐；界标；三维形变模型；卷积神经网络；姿态不变局部特征

中图分类号：TN957.52　　**文献标志码：**A　　**doi：**10.11805/TKYDA2019447

Large pose face alignment method based on convolutional neural network

LAN Min

(College of Economics，Trade and Information Technology，Changsha Vocational & Technical College，Changsha Hunan 410217，China)

Abstract：Large pose face alignment is a prerequisite for many important visual tasks such as face recognition and 3D face reconstruction. However，most of the existing alignment methods use two-dimensional boundary markers to align，and the number of boundary markers used is limited，which greatly affects the accuracy of large pose face alignment. Therefore，an improved large pose face alignment method is proposed. Firstly，3D deformable model is utilized to represent 2D face image. And the problem of face alignment with arbitrary pose is modeled as a fitting problem based on Three Dimensional Deformation Model(3DMM). And then a cascade regression method based on Convolutional Neural Network(CNN) is adopted to learn the mapping relationship between two-dimensional face image and its representation. Finally，two new pose invariant local features are proposed as the input layer of CNN learning，and CNN is applied for large pose face alignment through training. Simulation results on two classic face image data sets show that the proposed method is better than the latest face alignment method.

Keywords：face alignment；boundary markers；3D Deformable Model；Convolutional Neural Network；pose invariant local features

人脸对齐[1]指的是对齐人脸图像并识别特定基准点的过程，比如眼角、鼻尖等。尽管人脸对齐已经研究了多年，但仍然还是一个艰巨的任务，实际场景中，摄像机拍摄到的人脸会出现各种光照、姿态、视角，当人脸图片的差异很大时，会给人脸对齐算法带来难度。提高人脸对齐准确性有利于很多和面部分析相关的计算机视觉任务(如人脸识别[2]、三维人脸重构[3]等)。

文献[4–5]集成了人脸识别和姿态估计来实现人脸对齐，提出采用三维形状模型来对齐包含各种姿态的人脸。文献[6]提出一个基于级联回归量的人脸对齐方法，将该方法用于 FERET 数据库的人脸轮廓对齐时效果较好。然而该方法使用固定大小的特征提取框来提取特征，经常会得到无用的特征，不利于对人脸上的关键点进行准确定位。文献[7]使用遮挡不变的人脸对齐方法(Robust Cascaded Pose Regression，RCPR)来处理大姿态人脸对齐问题，通

收稿日期：2019-11-06；**修回日期：**2019-12-13

基金项目：国家自然科学基金资助项目(61402410/F020501)；湖南省教育厅科学研究资助项目(17C0195)

作者简介：蓝　敏(1982–)，男，硕士，主要研究方向为图像处理、数据挖掘。email:929138392@qq.com

过回归量进行三维界标估计。但是该方法当选取的初始关键点位置远离真实位置时，回归的结果会很差。

文献[8]提出一种自适应监督下降方法(Supervised Descent Method，SDM)的姿态鲁棒人脸对齐算法。该算法将所有样本放到一个统一的训练模型中训练，没有考虑样本之间存在的姿态差异，对于一些复杂姿态的样本定位效果不好。文献[9]在估计二维轮廓界标的位置和人脸形状的前提下进行人脸对齐，该方法缺乏对三维人脸姿态的考虑，在不同环境下算法对齐效果的稳定性较差。文献[10]基于正面人脸图像的相似性提出一个三维人脸模型拟合方法。SUN等[11]提出一个基于卷积神经网络(Convolutional Neural Network，CNN)的三步人脸对齐算法。ZHANG等[12]在文献[11]的基础上进一步开发了基于多步CNN的人脸对齐算法。此外，还有文献[13]在估计人脸上有限个界标位置的基础上，提出姿态不变的三维人脸对齐(Pose-Invariant 3D Face Alignment，PIFA)算法来进行大姿态人脸对齐。总的来说，这些前期研究的共同点在于它们大多使用二维界标位置，并且界标数量大多在6个以内。针对现有方法的不足，文中提出一种改进的大姿态人脸对齐方法。首先考虑到，在不同姿态下人脸具有不同数量的可见界标，而界标的空间分布在很大程度上取决于脸部姿态。这对现有人脸对齐方法构成巨大挑战，因为现有对齐方法大多基于二维形状模型，该模型在建模过程中经常会导致三维平面外变形，影响了人脸对齐方法的性能。为此，本文采用级联回归方法来学习二维人脸图像及其表示之间的映射关系。具体而言，提出采用卷积神经网络作为级联框架中的回归量来学习映射。先前研究采用CNN进行人脸对齐时，对于每张图片所估计的二维界标大多不超过6个，而本文的级联CNN方法所估计的二维和三维界标数量要高得多，可以达到34个。另外，通过采用界标移动技术，所提方法可以在拟合过程中自适应调整三维界标，有助于拟合脸颊界标。最后，传统二维人脸对齐方法通常基于每个估计的二维界标周围的局部特征补丁。然而对于地面实况界标(比如外眼角)，很难确保来自人脸的不同姿态下的局部特征补丁能涵盖解剖学意义上的完全相同部分的人脸皮肤，这对算法学习造成了额外困难，因为难以找到统一学习模式同地面实况界标联系起来实现人脸对齐[14]。为此，本文利用个体特异三维曲面法线来估计每个界标的可见性，并提出了2种新的姿态不变局部特征作为CNN学习的输入层，通过训练得到CNN用于大姿态人脸对齐。

1　人脸对齐

本文提出的人脸对齐方法的基本思路是：首先，采用三维形变模型拟合具有任意姿态的二维人脸图像。然后，基于CNN回归量的级联按顺序来估计未知拟合参数、三维形状参数和投影矩阵参数。最后，通过采用密集三维形状模型估计脸颊界标的位置，提取姿态不变的局部特征来进行有效的CNN学习，进而实现人脸对齐。图1展示了所提方法的全过程。

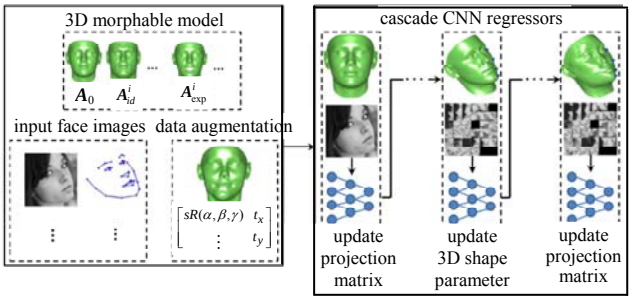


Fig.1 Whole process of the proposed method
图1 本文方法的全过程

1.1 问题描述

为了表示个体人脸的密集三维形状，文中采用三维形变模型(3DMM)来拟合具有任意姿态的二维人脸图像，其表示为：

$$\boldsymbol{A}=\boldsymbol{A}_0+\sum_{i=1}^{N_{id}}\boldsymbol{p}_{id}^i\boldsymbol{A}_{id}^i+\sum_{i=1}^{N_{exp}}\boldsymbol{p}_{exp}^i\boldsymbol{A}_{exp}^i \tag{1}$$

式中：$\boldsymbol{A}$ 为三维形状矩阵；$\boldsymbol{A}_0$ 为平均形状；$\boldsymbol{A}_{id}^i$ 为第 i 个身份基；$\boldsymbol{A}_{exp}^i$ 为第 i 个表达基；$\boldsymbol{p}_{id}^i$ 为第 i 个身份系数；$\boldsymbol{p}_{exp}^i$ 为第 i 个表达系数。2个系数共同表示为一个三维人脸的形状参数：$\boldsymbol{p}=(\boldsymbol{p}_{id}^{\mathrm{T}},\boldsymbol{p}_{exp}^{\mathrm{T}})^{\mathrm{T}}$。采用Basel三维人脸模型[15]作为身份基，采用人脸仓库[16]作为表达基。三维形状矩阵 $\boldsymbol{A}$ 以及 $\boldsymbol{A}_0$，$\boldsymbol{A}_{id}^i$ 和 $\boldsymbol{A}_{exp}^i$，是 $3\times Q$ 矩阵，它包含三维人脸表面 Q 个顶点的 x，y 和 z 坐标：

$$\boldsymbol{A}=\begin{pmatrix}x_1 & x_2 & \cdots & x_Q\\ y_1 & y_2 & \cdots & y_Q\\ z_1 & z_2 & \cdots & z_Q\end{pmatrix} \tag{2}$$

可将任意的三维人脸模型投影到一个二维图像上。在二维图像中，人脸形状可以表示为在面部基准点上包含 N 个二维界标的稀疏集，将这些二维界标的x和y坐标表示为矩阵 $\boldsymbol{U}$，有：

$$\boldsymbol{U}=\begin{pmatrix} u_1 & u_2 & \cdots & u_N \\ v_1 & v_2 & \cdots & v_N \end{pmatrix} \tag{3}$$

借鉴文献[3]的工作，文中采用弱透视投影来描述三维形状 $\boldsymbol{A}$ 和二维界标 $\boldsymbol{U}$ 之间的关系，即：

$$\boldsymbol{U}=sR\boldsymbol{A}(:,\boldsymbol{d})+t \tag{4}$$

式中：s 为一个尺度参数；R 为一个通过 3 个旋转角 α,β 和 γ 控制的 3×3 旋转矩阵的前 2 行；t 为一个平移参数，由 t_x 和 t_y 组成；$\boldsymbol{d}$ 为一个 $N-\dim$ 指标向量，表示对应于二维界标的有语义意义的三维顶点的指标。通过收集和该投影相关的所有参数，可以得到一个投影向量 $\boldsymbol{m}=(s,\alpha,\beta,\gamma,t_x,t_y)^{\mathrm{T}}$。此时，可以将任意二维人脸形状表示为三维人脸形状的投影。换句话说，投影参数 $\boldsymbol{m}$ 和形状参数 $\boldsymbol{p}$ 可唯一地表示一个二维人脸形状。因此，任意给定一幅人脸图像，人脸对齐问题等价于估计给定人脸图像的参数 $\boldsymbol{m}$ 和参数 $\boldsymbol{p}$。

1.2 界标移动

对于给定的指标向量 $\boldsymbol{d}$，式(4)中的投影关系对于正面人脸图像是准确的。但是，随着人脸变成侧视图，在二维图像中，脸颊上的原三维界标变得不可见，如图2所示。在图2中，从左边到右边分别表示：初始界标、拟合的三维密集形状和具有可见性的估计界标。右边一列的绿色/红色/黄色分别表示：可见/不可见/脸颊界标。此时，为了保证式(4)中的映射关系的准确性，需要估计匹配这些脸颊界标的三维顶点。已有研究[11,17]提出了多种方案来解决这个问题，但都缺乏对于脸颊界标的有效估计。为此，文中提出如下的界标移动策略：定义一组路径，每个路径储存顶点的指标，这些顶点不仅距离原三维脸颊界标最近，也在三维人脸的轮廓上。给定一个非正面三维人脸 $\boldsymbol{A}$，使用 α 和 β (俯仰角和偏航角)旋转 $\boldsymbol{A}$，在每个路径上寻找一个具有最高(最低)x坐标的顶点，即，右边(左边)脸颊上的边界顶点。这些搜索到的顶点为对应于二维脸颊界标的新三维界标。然后，更新 $\boldsymbol{d}$ 的相关元素，以确保在式(4)的投影中选择这些顶点。界标移动的具体过程如算法1所示。在算法1，将界标移动过程总结为函数 $\boldsymbol{d}\leftarrow g(\boldsymbol{A},\boldsymbol{m})$。需指出的是，当人脸几乎是侧视图时 ($|\beta|>70^\circ$)，不使用界标移动，因为移动的界标会和现有的鼻子和嘴巴中间的二维界标重叠。

Fig.2 Fitting dense 3D shape to estimate the boundary of large pose face
图2 拟合密集三维形状来估计大姿态人脸的界标

算法1：界标移动 $g(\boldsymbol{A},\boldsymbol{m})$

输入：估计得到的三维人脸形状矩阵 $\boldsymbol{A}$ 和投影参数 $\boldsymbol{m}$

输出：指标向量 $\boldsymbol{d}$

/*通过估计的 α 和 β 来旋转 $\boldsymbol{A}$*/

1. $\hat{\boldsymbol{A}}=R(\alpha,\beta,0)\boldsymbol{A}$
2. if $0^\circ<\beta<70^\circ$ then
3. for each $i=1,2,\cdots,4$ do
4. $V_{\text{cheek}}(i)=\arg\max_{id}\left(\hat{\boldsymbol{A}}\left(1,path_{\text{cheek}}(i)\right)\right)$
5. if $-70^\circ<\beta<0^\circ$ then
6. for each $i=5,6,\cdots,8$ do
7. $V_{\text{cheek}}(i)=\arg\max_{id}\left(\hat{\boldsymbol{A}}\left(1,path_{\text{cheek}}(i)\right)\right)$
8. 利用 V_{cheek} 更新 $\boldsymbol{d}$ 的8个元素。

2 CNN模型的训练

2.1 数据增广

由于投影参数 $\boldsymbol{m}$ 和形状参数 $\boldsymbol{p}$ 可有效表示人脸图像，应考虑更多包含环境实况的人脸图像，以便学习算法能够利用它来提高人脸对齐的效果。对于大多数现有人脸对齐数据库，仅二维界标位置和界标的可见性被手动标注，没有关联的三维信息(如 $\boldsymbol{m}$ 和 $\boldsymbol{p}$)。为提高学习效果，针对二维人脸图像设计一个数据增广过程，其目标是有

效地估计包含了环境实况的 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 的表示。

具体来说，对于任意给定标记的可见二维界标 $\boldsymbol{U}$ 和界标可见性 v，使用式(5)目标函数来估计 $\boldsymbol{m}$ 和 $\boldsymbol{p}$：

$$J(\boldsymbol{m},\boldsymbol{p})=\| (sR\boldsymbol{A}(:,g(\boldsymbol{A},\boldsymbol{m}))+t-\boldsymbol{U})\odot V\|_F^2 \tag{5}$$

即要最小化三维界标和二维标记界标之间的投影的差异。在式(5)中，界标移动 $g(:,:)$ 可以使非侧视图的脸颊界标变得“可见”，可见性 v 可用来避免不可见的界标(如外眼角和侧视图中半张脸)成为优化的一部分。为了最小化式(5)所示的目标函数，在每次迭代中交替地实现 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 的最小化。初始化时，设置初始化三维形状参数 $\boldsymbol{p}=0$，并先估计 $\boldsymbol{m}$。然后，在每次迭代中使用当前估计得到的 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 来计算 $g(\boldsymbol{A},\boldsymbol{m})$，其值为一个常数。

2.2 级联CNN耦合回归量

给定一组训练人脸图像集合(大小为 N_d)和参数 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 的增广表示(包含了环境实况)，希望学习一个映射函数，能够从人脸图像的外观预测 $\boldsymbol{m}$ 和 $\boldsymbol{p}$。很明显，这是一个复杂的非线性映射函数。考虑到CNN在视觉任务上(如姿态估计、人脸识别和人脸对齐等)取得的成功，本文通过学习一系列基于CNN的回归量将CNN和级联回归量结合起来以更改 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 的估计。

除了包含地面实况的 $\boldsymbol{m}$ 和 $\boldsymbol{p}$，文中假设每个训练图像已知 $\boldsymbol{m}$ 和 $\boldsymbol{p}$ 的初始值：$\boldsymbol{m}_0$ 和 $\boldsymbol{p}_0$。在级联CNN的第 k 步，学习一个CNN来估计期望的投影参数的更新：

$$\Theta_{\boldsymbol{m}}^k=\arg\min_{\Theta_{\boldsymbol{m}}^k}\sum_{i=1}^{N_d}\|\Delta\boldsymbol{m}_i^k-CNN_{\boldsymbol{m}}^k(\boldsymbol{I}_i,\boldsymbol{U}_i,v_i^{k-1};\Theta_{\boldsymbol{m}}^k)\|^2 \tag{6}$$

式中：$\Delta\boldsymbol{m}_i^k=\boldsymbol{m}_i-\boldsymbol{m}_i^{k-1}$，即真实的投影更新值为当前投影参数和地面实况值之差；$\boldsymbol{U}_i$ 为当前估计的二维界标，通过式(4)可知，它基于 $\boldsymbol{m}_i^{k-1}$ 和 d_i^{k-1} 计算得到；v_i^{k-1} 为在第 $k-1$ 步估计得到的界标可见性。类似地，可学习另外一个CNN回归量来估计形状参数的更新值：

$$\Theta_{\boldsymbol{p}}^k=\arg\min_{\Theta_{\boldsymbol{p}}^k}\sum_{i=1}^{N_d}\|\Delta\boldsymbol{p}_i^k-CNN(\boldsymbol{I}_i,\boldsymbol{U}_i,v_i^k;\Theta_{\boldsymbol{p}}^k)\|^2 \tag{7}$$

通过 $CNN_{\boldsymbol{m}}$ 更新得到的最新的 $\boldsymbol{m}_i^k$ 和 d_i^k 来重新计算 $\boldsymbol{U}_i$，如式(4)所示。使用一个6步级联CNN，包括 $CNN_{\boldsymbol{m}}^1$，$CNN_{\boldsymbol{m}}^2$, $CNN_{\boldsymbol{p}}^3$, $CNN_{\boldsymbol{m}}^4$, $CNN_{\boldsymbol{p}}^5$ 和 $CNN_{\boldsymbol{m}}^6$。在第1步，$CNN_{\boldsymbol{m}}^1$ 的输入层为整个脸部区域，包含初始边界框，其目的是粗略估计人脸姿态。第2步到第6步的输入是一个114×114的图像，包含从当前估计的二维界标 $\boldsymbol{U}_i$ 提取的19×19姿态不变特征补丁阵列。这些串接的特征补丁可编码关于当前二维界标周围局部外观的充分信息，促使CNN去优化参数 $\Theta_{\boldsymbol{m}}^k$ 或 $\Theta_{\boldsymbol{p}}^k$。可对该方法进行拓展，使用更多界标来估计更准确的密集三维模型。需要注意的是，由于采用了界标移动技术，估计的二维界标 $\boldsymbol{U}_i$ 包含移动的三维界标的投影，即二维脸颊界标。因此，这些脸颊界标周围的外观特征也是CNN输入的一部分。这些附加脸颊界标可有效地作为约束条件，来影响各姿态面部轮廓的外观，这基本上就是三维人脸表面的形状。使用修正线性单元(Rectified Linear Unit，RLU)[18]作为激活函数，在没有无监督预训练的情况下，可以使CNN取得最优性能。在全部6步中，使用相同的CNN架构如图3所示。

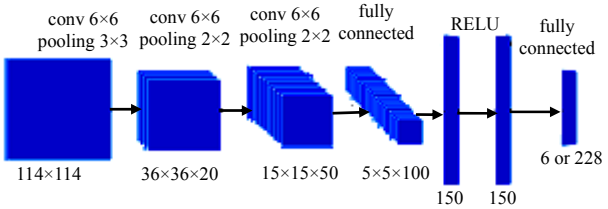


Fig.3 CNN architecture used in each step of the proposed method

图 3 本文方法每一步所使用的 CNN 架构

2.3 可见性和二维外观特征

在采用密集三维形状模型时，一个显著的优点是可提取更高级的二维特征，该优点有助于级联CNN学习。在本文中，这些二维特征是指二维界标可见性和每个二维界标周围的外观补丁。

为了计算每个二维界标的可见性，采用了在当前相机或摄像头的投影矩阵下，检查相应三维界标的三维曲面法线是否指向相机或摄像头的基本思想。需要注意的是，不使用针对所有人脸的平均三维曲面法线，而是使用特定于个人的个体特异三维曲面法线来衡量界标的可见性。具体来说，对于当前估计的三维形状 $\boldsymbol{A}$，计算感兴趣的三维界标周围一组稀疏顶点的三维曲面法线，这些三维法线的平均值表示为 $\boldsymbol{N}$。图4显示了使用平均三维曲面法线的优点。个体特异三维曲面法线是一个三维界标(黑色箭头)周围法线的平均值。一般而言，包含噪声的三维“左眼角”界标(蓝色箭头)的噪音相对有较高的曲面法线。在给定 $\boldsymbol{N}$ 的情况下，计算 $v=\boldsymbol{N}^{\mathrm{T}}(R_1\times R_2)$，其中，$R_1$ 和 R_2 为 R 的前2行。如果 v 是正的，则认为二维界标是可见的，其二维外观特征将是CNN输入的一部分。否则，是不可

见的，那么对于CNN输入而言，相应特征为零。

除了可见性估计，三维形状模型也可辅助生成高级外观特征，作为CNN的输入层。具体来说，希望在每个估计的二维界标周围提取一个姿态不变的外观补丁，这些补丁的阵列将构成输入层。下面描述了2种方法来提取外观特征(一个19×19补丁，用于第 n 个二维界标)。

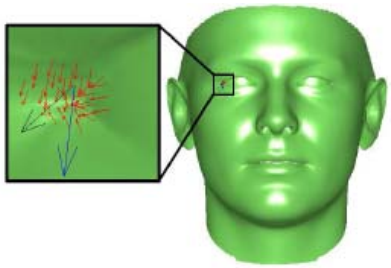


Fig.4 Example of average 3D surface normal
图 4 平均三维曲面法线的例子

1) 分段仿射翘曲特征(Piecewise Affine-Warped Feature，PAWF)：特征对应关系对于任何视觉学习均非常重要。但是，由于二维人脸是具有任意视角的三维表面的投影，很难确保从该二维图像提取的局部补丁对应于来自其他二维图像的补丁，即使2个补丁都以相同的第 n 个二维界标的地面实况位置为中心。这里，“对应于”指的是补丁覆盖解剖学意义上人脸的完全相同区域。但是，通过密集三维形状模型，可以提取不同对象和姿态的局部补丁，并获得解剖学意义上的对应关系。

在离线学习阶段，首先在最靠近第 n 个界标的平均三维形状 $\boldsymbol{A}_0$ 上搜索 T 个顶点。其次，旋转 T 个顶点，使第 n 个界标的三维曲面法线指向摄像头。最后，在 T 个顶点中，找到具有最小和最大的 x 坐标和 y 坐标的4个“邻域顶点”，将这4个顶点的ID表示为一个 4 dim 向量 $\boldsymbol{d}_p^{(n)}$。

在CNN学习阶段，针对第 i 个图像的第 n 个界标，基于当前估计的投影参数 m，将4个邻域顶点投影到第 i 个图像上，并获得4个邻域点，$\boldsymbol{U}_i^{(n)} = s\boldsymbol{R}\boldsymbol{A}(:,\boldsymbol{d}_p^{(n)}) + t$。在所有二维人脸图像，$\boldsymbol{U}_i^{(n)}$ 在解剖学意义上对应于相同的人脸顶点。因此，通过使用分段仿射变换，将这些邻域点范围内的图像内容翘曲到一个19×19补丁。

在大多数情况下，可以很好地提取这种新的特征表示，除了在一些特殊情况下，比如侧视图的鼻尖，在这种情况下，通常会发生2种情形，一种是第 n 个界标的投影在邻域点确定的区域之外；另外一种情形是其中一个邻域点是不可见的。在发生这种情况时，通过根据不可见点到投影界标位置的相对距离来改变不可见点的位置，如图5所示。当4个邻域点中的一个点(右下方的红点)变得不可见时，它会连接到二维界标进一步拓展相同的距离，并生成新邻域点。这样一来，可以包含鼻子周围的背景信息。

2) 直接三维投影特征(Direct 3D Projected Feature，D3PF)：D3PF和PAWF均始于第 n 个三维界标周围的 T 个顶点。不是像在PAWF那样找到4个邻域顶点，D3PF使用一个19×19网格来覆盖 T 个顶点，并将网格点的顶点存储在 $\boldsymbol{d}_d^{(n)}$。和PAWF类似，将一组3D顶点 $\boldsymbol{A}(:,\boldsymbol{d}_d^{(n)})$ 投影到二维图像，并通过双线性插值提取一个19×19补丁，如图6所示。还通过其曲面法线估计这些三维顶点的可见性，并将零值放在不可见顶点的补丁上。对于D3PF，补丁中的每个像素对应于其他图像补丁中的相同像素；而对于PAWF，这仅针对4个邻域点成立。

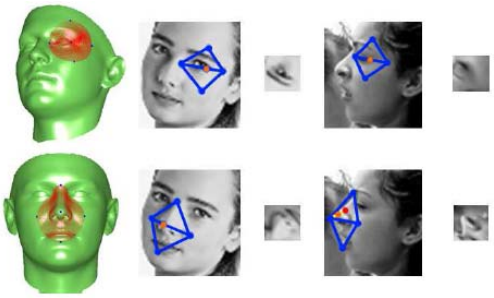


Fig.5 Example of extracting features of PAWF
图 5 提取 PAWF 特征的例子

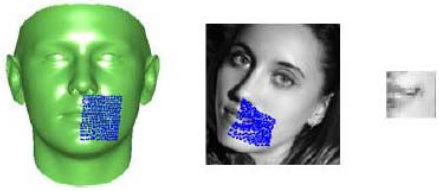


Fig.6 Example of extracting D3PF
图 6 提取 D3PF 的例子

3 仿真实验

3.1 数据集

针对本文研究的人脸对齐问题，选择了2个公开的人脸数据集进行仿真实验，测试所使用的软硬件环境如下：Inter(R) Core (M) i3-3240 CPU 3.40 GHz,500 GB 硬盘，4.0 GB内存，Microsoft Windows 10 Professional,Python3.7, Anaconda平台和Tensorflow开源库。数据集的具体情况如下：

1) AFLW数据集[19]：是一个包含25 000张人脸图像的大型人脸数据集。其中的每个图像均手动加上21个界标，每个界标包括可见性标签。借鉴文献[13]的工作，选择了AFLW的子集来获得旋角的平衡分布，包括3 901张图像用于训练，1 299张图像用于测试。使用相同的子集，并为所有5 200张图像手动加上另外13个界标。原始界标和

添加的界标的定义如图7(a)所示。根据每张图像的地面实况界标，找到最紧致的边界框，将其尺寸扩展10%，向边界框的左上角、宽度和高度添加10%的噪音。这些随机生成的边界框模拟不精确的人脸识别窗口，可用于训练和测试。

2) AFW数据集[4]：包含205张图像中的468个人脸。每个人脸手动加上6个界标，每个界标均有可见性标签。使用Basel人脸模型[15]的 $N_{id}=199$ 个基来表示身份变化，使用人脸仓库[16]的 $N_{\exp}=29$ 个基来表示表达变化。总共有228个基来表示具有53 215个顶点的三维人脸形状。

3.2 实验设置

1) 参与比较的基准方法：选择最新的人脸对齐方法和提出的方法进行比较。在AFLW数据集上，将提出的方法同PIFA[13]和RCPR[7]进行比较；在AFW数据集上，将提出的方法同PIFA[9]、CDM(Cascaded Deformable shape Model)[5]和混合树状模型(Tree Structured Part Model，TSPM)[4]进行比较。

2) 参数设置：就提出的方法而言，在训练期间，CNN的学习率取0.000 1。对于RCPR，使用了其论文中报告的参数，进行100次迭代，采用15个增强回归量。对于PIFA，进行200次迭代，采用5个增强回归量。针对PAWF和D3PF，在第2步，T为5 000；在其他步，T为3 000。根据实证估计，CNN的6步已经足以保证拟合过程的收敛。

3) 估计标准：使用2种常规标准来衡量34个界标的误差。在AFLW数据集上，使用由边界框大小确定的可见地标的归一化平均误差(Normalized Mean Error，NME)[20]。需要注意的是，在NME中不使用眼睛到眼睛的距离(如侧面图)，因为它在大姿态中不能准确地定义。在AFW数据集上，采用了平均像素误差(Mean Average Pixel Error，MAPE)[5]。

4) 特征提取：为了衡量文中提出特征的有效性，表1基于多种特征(即，CNN^2到CNN^6的输入层)比较了本文方法在AFLW数据集上进行人脸对齐的准确性。其中，“提取的补丁”指从使用边界框的人脸图像中提取的具有恒定尺寸(19×19)的补丁，这作为基准特征。针对特征“+脸颊界标”，轮廓界标的另外4个19×19 补丁(非正面人脸的此类补丁是不可见的)将被脸颊界标补丁替换，并用在CNN学习的输入层。从表1可知，PAWF可以获得比D3PF更高的人脸对齐精确度。通过比较表1的第1列和第3列发现，针对级联CNN回归量，从脸颊界标提取特征可作为有效的附加视觉线索。通过结合脸颊界标的使用和PAWE特征的提取，可获得最高的准确率，将用于剩余的实验。此外注意到，AFLW训练集的规模相对于训练一个CNN来说太小。但基于CNN的回归量仍然能够学习并很好地对齐未看见的图像。将其归因于文中所提出的有效外观特征，即良好的特征匹配能够降低CNN对大量训练数据的要求，这也从侧面表明了本文方法的有效性。

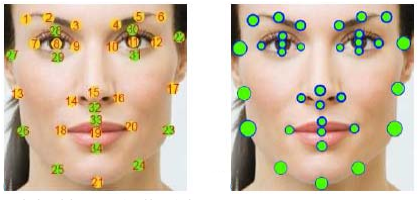
表1 基于不同特征的本文方法的NME(%)比较

Table1 Comparison of NME(%) of the proposed methods

PAWF+ cheek landmarks	D3PF+ cheek landmarks	PAWF	extracted patch
4.72	5.02	5.19	5.51

3.3 AFLW数据集上的实验与分析

通过对齐具有任意姿态的人脸，在AFLW数据集上将本文方法同2种最相关的方法(PIFA和RCPR)进行了比较。使用RCPR的源代码对其进行训练和试验。类似地，对于PIFA，使用它的源代码来训练界标数量高达13个的AFLW训练集。3种方法的准确性如表2所示。从表2可以看到，本文方法获得的结果要优于PIFA和RCPR这2种基准方法。针对每种界标的误差比较如图7所示。其中，图7(a)是AFLW中的原始图像(黄色)和添加界标后的图像(绿色)；图7(b)是RCPR(蓝色)和本文方法(绿色)对于每个界标进行估计的结果对比。根据平均NME和人脸方框尺寸的乘积来确定圆圈半径。从图7可以看到，RCPR方法对于轮廓界标的估计误差更高。而针对所有界标，本文方法的误差均低于RCPR。



(a) original image (yellow) in AFLW and the image after adding boundary markers (green)　(b) comparison of the results of RCPR(blue) and our method(green) for each boundary marker

Fig.7 Performance comparison of different methods

图7 不同方法的性能对比

表2 三种方法在AFLW上的NME(%)比较

Table2 Comparison of NME(%) of three methods on AFLW

proposed method	PIFA	RCPR
4.72	8.04	6.26

通过利用实验图像的地面实况界标位置，根据每个图像估计旋角，将AFLW数据集中的图像分成6个子集。图8比较了本文方法和RCPR的NME。可以看到，本文方法在不同的姿态下均可取得更优结果，更重要的是，本文方法的鲁棒性更好，在各个姿态上的变化也更小。为了更详细地比较不同方法的NME分布，图9给出了不同方法在进行人脸对齐时的累积误差分布(Cumulative Error Distribution，CED)

图。从图9可以明显地观察到，随着测试图像的比例增加，所有方法的NME都在增加，但无论测试图像的比例如何变化，本文方法的NME值始终最优。

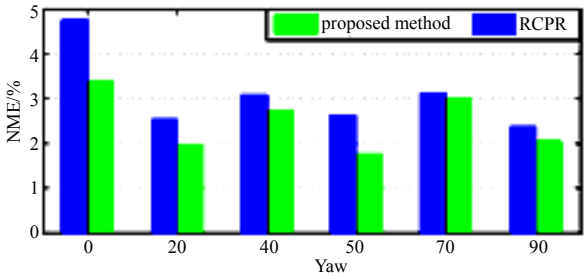


Fig.8 NME comparison for each pose
图8 针对每个姿态的NME比较

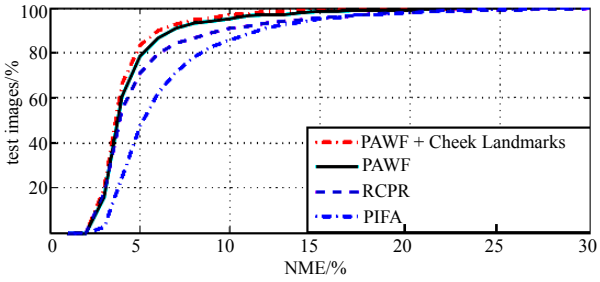


Fig.9 CED comparison of different methods
图9 不同方法的CED比较

3.4 AFW数据集上的实验与分析

AFW数据集包含所有姿态范围的人脸图像，带有6个界标。表3给出了5种方法在AFW上的MAPE比较结果。对于PIFA,CDM和TSPM，使用了其论文中所报告的误差。从表3的结果可以看到，和PIFA,CDM和TSPM方法相比，本文提出的方法仍能获得更优结果。

表3 五种方法在AFW上的MAPE比较

Table3 MAPE comparison of five methods on AFW

proposed method(PAWF)	proposed method(D3PF)	PIFA	CDM	TSPM
7.43	7.83	8.61	9.13	11.09

3.5 在2种数据集上的对齐结果展示

在AFLW和AFW数据集上展示本文方法所获得的一些对齐结果，如图10所示。其中，绿色/红色/黄色点分别表示可见/不可见/脸颊界标。第1行：针对AFLW的初始界标；第2行：估计的三维密集形状；第3行：估计的界标；第4行和第5行：针对AFLW估计的界标；第6行：针对AFW估计的界标。本文方法在每一步所获得的结果如图11所示。其中，第1行表示提取的特征；第2行表示对齐结果。

Fig.10 Results of the method in this paper on AFLW and AFW
图 10 本文方法在 AFLW 和 AFW 上获得的结果

Fig.11 Results obtained in each step of the proposed method
图11 本文方法在各步骤所获得的结果

3.6 算法效率分析

最后，以规模更大的AFLW数据集作为测试对象，在Anaconda平台上测试了PAWF方法和D3PF方法的效率。仿真运行结果表明，PAWF方法和D3PF方法的速度分别为0.6帧/秒和0.26帧/秒，可以满足大多数场景下图像对齐的要求。仔细分析其原因可知，这主要是因为对卷积神经网络的输入作了优化，在考虑可见性和二维外观特征的基础上，采用分段仿射翘曲特征和直接三维投影特征进行网络的训练，在保证图像对齐质量的同时，降低了时间复杂度，取得了更好的效率。

4 结论

针对现有人脸对齐方法的不足，提出一种改进的大姿态人脸对齐方法，通过结合级联CNN回归量和三维形变模型(3DMM)，将三维密集形状拟合到大姿态人脸图像，提出2种姿态不变的特征来提高人脸对齐的准确性。最后，在2个具有大姿态的极具挑战性的人脸数据集上，验证了所提出的方法的优越性。在下一步工作中，将针对现有基于深度学习的人脸对齐算法难以实现真正意义上“端对端”、浅层特征表征能力及鲁棒性差的问题，拟提

出一种基于图卷积神经网络的快速人脸对齐算法。

参考文献:

[1] 高军山,陈杭,林慧平,等. 高分辨力极化 SAR 图像城市区域车辆目标检测[J]. 太赫兹科学与电子信息学报, 2018, 16(4):603-608. (GAO Junshan,CHEN Hang,LIN Huiping,et al. Vehicle detection over urban areas in high resolution polarimetric SAR images[J]. Journal of Terahertz Science and Electronic Information Technology, 2018,16(4):603-608.)

[2] 姚乃明,郭清沛,乔逢春,等. 基于生成式对抗网络的鲁棒人脸表情识别[J]. 自动化学报, 2018,44(5):865-877. (YAO Naiming,GUO Qingpei,QIAO Fengchun,et al. Robust facial expression recognition with generative adversarial networks[J]. Acta Automatica Sinica, 2018,44(5):865-877.)

[3] 洪金华,张荣,郭立君. 基于 L1/2 正则化的三维人体姿态重构[J]. 自动化学报, 2018,44(6):1086-1095. (HONG Jinhua, ZHANG Rong,GUO Lijun. 3D human body pose reconstruction via L1/2 regularization[J]. Acta Automatica Sinica, 2018, 44(6):1086-1095.)

[4] ZHU X,RAMANAN D. Face detection,pose estimation,and landmark localization in the wild[C]// IEEE conference on computer vision and pattern recognition. RI,USA:IEEE Computer Society, 2012:2879-2886.

[5] YU X,HUANG J,ZHANG S,et al. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model[C]// Proceedings of the IEEE International Conference on Computer Vision. Sydney,NSW,Australia:IEEE Computer Society, 2013:1944-1951.

[6] WU Y,JI Q. Robust facial landmark detection under significant head poses and occlusion[C]// Proceedings of the IEEE International Conference on Computer Vision. Chile:IEEE Computer Society, 2015:3658-3666.

[7] BURGOS-ARTIZZU X P,PERONA P,DOLLÁR P. Robust face landmark estimation under occlusion[C]// Proceedings of the IEEEInternational Conference on Computer Vision. Australia:IEEE Computer Society, 2013:1513-1520.

[8] 赵慧,景丽萍,于剑. 自适应监督下降方法的姿态鲁棒人脸对齐算法[J]. 计算机科学与探索, 2019,11(2):1-8. (ZHAO Hui, JING Liping,YU Jian. Pose-robust face alignment with adaptive supervised descent method[J]. Journal of Frontiers of Computer Science & Technology, 2019,11(2):1-8.)

[9] QU Chengchao,MONARI Eduardo,SCHUCHERT Tobias,et al. Adaptive contour fitting for pose-invariant 3D face shape reconstruction[C]// 26th British Machine Vision Conference(BMVC). Swansea,UK:BMVA Press, 2015:1-12.

[10] ZHU X,YAN J, YI D, et al. Discriminative 3D morphable model fitting[C]// IEEE International Conference and Workshops on Automatic Face and Gesture Recognition(FG). Ljubljana,Slovenia:IEEE Press, 2015:1-8.

[11] SUN Y,WANG X,TANG X. Deep convolutional network cascade for facial point detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Portland,OR,USA:IEEE Computer Society, 2013:3476-3483.

[12] ZHANG J,SHAN S,KAN M,et al. Coarse to fine auto-encoder networks(cfan) for real-time face alignment[C]// European Conference on Computer Vision. Zurich,Switzerland:Springer,Cham, 2014:1-16.

[13] JOURABLOO A,LIU X. Pose-invariant 3D face alignment[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago de Chile:IEEE Computer Society, 2015:3694-3702.

[14] JIN K H,MC CANN M T,FROUSTEY E,et al. Deep convolutional neural network for inverse problems in imaging[J]. IEEE Transactions on Image Processing, 2017,26(9):4509-4522.

[15] TRAN L,YIN X,LIU X. Disentangled representation learning GAN for pose-invariant face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,HI,USA:IEEE Computer Society, 2017:1415-1424.

[16] CAO C,WENG Y,ZHOU S,et al. Facewarehouse:a 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2013,20(3):413-425.

[17] CAO C,HOU Q,ZHOU K. Displaced dynamic expression regression for real-time facial tracking and animation[J]. ACM Transactions on Graphics(TOG), 2014,33(4):43.

[18] POLSON N G,ROČKOVÁ V. Posterior concentration for sparse deep learning[C]// Advances in Neural Information Processing Systems. Montréal,Canada:IEEE Press, 2018:930-941.

[19] KOESTINGER M,WOHLHART P,ROTH P M,et al. Annotated facial landmarks in the wild:a large-scale,real-world database for facial landmark localization[C]// IEEE international conference on computer vision workps(ICCV workshops). Barcelona, Spain:IEEE Computer Society, 2011:2144-2151.

[20] ZHANG Z,LUO P,LOY C C,et al. Facial landmark detection by deep multi-task learning[C]// European Conference on Computer Vision. Springer,Cham:IEEE Press, 2014:94-108.