具有粗到细卷积网络级联的广泛面部地标定位

摘要

我们提出了一种新的方法，用粗到细的卷积网络级联来定位广泛的面部标志。深度卷积神经网络（DCNN）已成功应用于面部标志定位中，具有双重优势：1）隐含地利用了面部点之间的几何约束; 2）可以利用大量的培训数据。然而，在广泛的面部标志定位任务中，需要将大量的面部标志（超过50个点）放置在统一的系统中，这给传统卷积网络的结构设计和训练过程带来了很大的困难。在本文中，我们设计了一个四级卷积网络级联，以粗略到精细的方式解决问题。在我们的系统中，训练每个网络级别以局部细化由先前网络级别生成的面部地标的子集。此外，每个级别预测显式几何约束（特定面部组件的位置和旋转角度）以纠正当前网络级别的输入。粗到细级联和几何细化的组合使我们的系统能够在300-W面部地标定位挑战中准确定位广泛的面部标志（68点）。

1.简介

面部地标定位在人脸识别和面部分析系统中起着关键作用。在Chen最近的一篇论文[4]中，可以看出，如果可以使用准确的面部标志，简单的特征可以在人脸识别方面取得领先的表现。因此，面部地标定位的问题在过去几年中引起了广泛的兴趣。通常，从面部图像中定位面部标志有三种主要方法：第一类基于局部补丁分类器执行滑动窗口搜索，其遇到局部特征中的模糊或损坏的问题。此外，很难将全球背景信息纳入本地搜索框架;第二类方法是活动形状模型（ASM）[2]和主动外观模型（AAM）[5]的众所周知的框架。这些方法适合全局面部外观的生成模型，因此对局部腐败具有鲁棒性。然而，为了估计生成模型中的参数，需要昂贵的迭代步骤。



图1.地标定位系统的比较。 第一行是原始面部图像。 第二行由OpenCV [3]中包含的局部补丁检测器产生。 第三行由Stasm [9]生成，这是一个开源的AAM实现。 我们的结果显示在第四行，显着优于其他行。

最近，提出了一种基于明确回归方法的新框架[10,11]。在该框架中，地标定位的问题被直接视为回归任务，并且整体回归器用于计算地标坐标。与上述方法相比，该框架更加健壮和稳定，因为全球背景信息最初被纳入其中;它也更有效，因为不需要迭代拟合步骤或滑动窗口搜索。 Sun [11]在回归框架中应用了更强大的深度卷积神经网络（DCNN），而不是[10]中使用的随机蕨类植物，而是实现了最先进的性能。

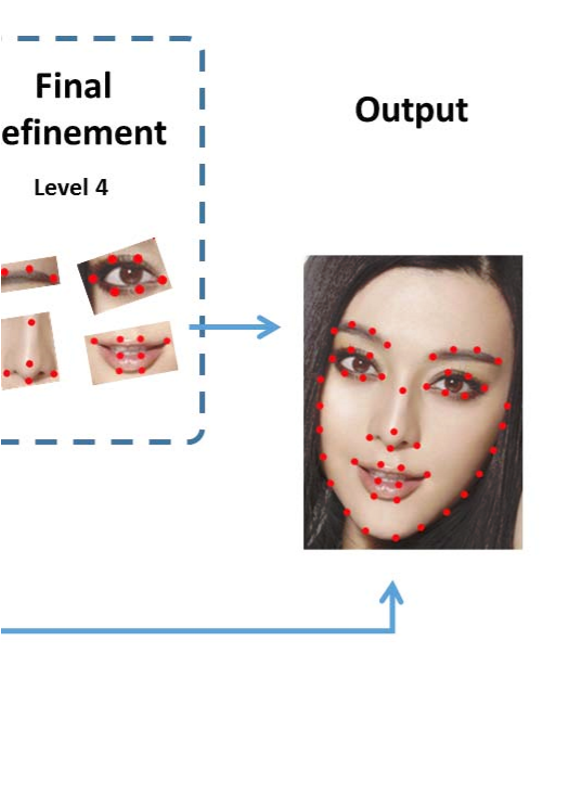
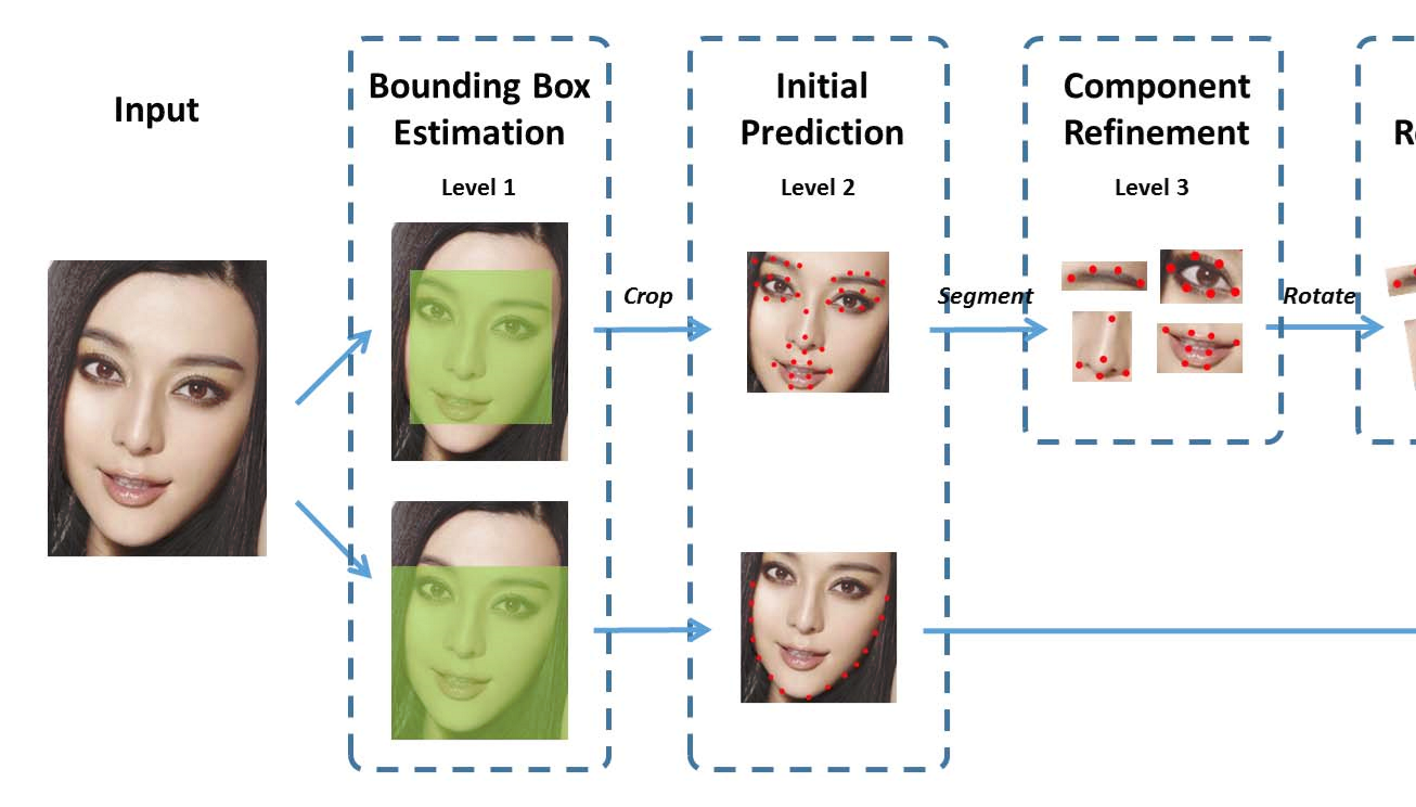


图2.系统概述。 第一级网络分别预测内点和轮廓点的边界框。 对于内点，第二级预测对于每个组件由第三级细化的位置的初始估计。 第四级用于通过将旋转的图像块作为新输入来改善口和眼的预测。 两个级别的独立网络用于轮廓点。 为清楚起见，并非所有68个点都在图中呈现。

然而，面部地标本地化仍然是一个非常具有挑战性挑战来自于由于姿势，闪电，表情等的变化导致的面部外观的大变化。当需要大量的地标点时，任务更具挑战性。挑战的性质在不同的面部点之间变化很大，因此单模型方法可能会失败。另一方面，对每个点采用单独的系统会大大增加计算时间。然而，大量的点是双刃剑：关于地标的相对位置的内部结构的有价值的信息变得存在。面部组件的全局排列的几何约束以及组件内的点的相互作用为系统充分利用它们提供了准确性和鲁棒性的改进的希望。

为了应对这一挑战，我们精心设计了一个多级卷积网络级联，通过粗粒度网络级联来解决广泛的面部地标定位问题。 我们的贡献是三方面的：1）不像[11]预测稀疏的面部标志（5分）与网络级联，我们验证卷积网络级联对广泛的面部地标定位问题的有效性; 2）我们卷积网络; 3）我们表明显式几何细化（估计面部组件的位置/旋转并纠正每个网络级别的输入）可以显着提高准确性和鲁棒性。 大量实验表明，我们的系统准确，稳健。

2.概述

图2简要介绍了我们的多层面部地标定位系统。我们使用术语内部点来表示眼睛，眉毛，嘴巴和鼻子的51个点以及轮廓上17个点的轮廓点。内点和轮廓点的子系统与第一级分开。在第一级中，训练两个神经网络以独立地估计内点和轮廓点的边界框（x-y坐标的最大值和最小值）。这些盒子分别送入系统的其余部分。

内在点。对于内部51点，另外还训练了三级卷积神经网络。在获得内点的边界框之后，最初通过第二级估计51个内部界标。基于初始估计，分别计算6个面部成分（即眉毛，眼睛，嘴和鼻子）的区域。训练第三级以独立地细化每个面部组件的界标。估计每个部件的旋转角度并将其校正为直立，并且将旋转的补片馈送到第四级网络以获得最终结果。

轮廓点。 使用更简单的网络级联

用于轮廓点的定位。 给定覆盖脸颊的边界框，第二级将裁剪后的图像作为输入，并计算来自原始像素的轮廓点的坐标。 由于时间有限，未使用三级和四级网络，我们将更深层次的网络级联进一步利用到未来的工作中。

3.粗到细的DCNN级联

我们框架的核心思想是粗到细级联的设计。 每个网络级别细化由先前级别计算的区域内的地标子集。 在第一级，面部分为两部分：内部和轮廓。 在第二级之后，内部的面部组件进一步分离。 我们不为每个面部地标训练单独的网络以降低计算成本。 粗到精框架有许多优点。

3.1。分离损失函数

不同地标的定位硬度不平衡。特别是，由于两个原因，轮廓比内点明显更困难。首先，与内部地标相比，面部图像为轮廓点提供较少的局部纹理信息，但是来自这些点附近的背景的无关信息明显更多。此外，这些点的基本事实本质上更嘈杂，因为每个点的确切位置的定义更加模糊。这些因素导致两部分的训练误差之间的严重不平衡，因此如果所有68个点被一起训练，则L2损失函数将由轮廓支配。因此，训练两个独立的子系统使整个系统有机会学习内点的详细结构，而不是将其大部分能力用于拟合“困难”轮廓。我们的实验支持这一论点。

在内部点中，面部组件的相对困难仍然不均匀。如第5节所示，眉毛明显更难，而系统对眼睛的预测更准确。

3.2。多级细化

将定位任务分解为多个阶段，在每个阶段中考虑点或组件之间的交互。在第一级中，计算与面部姿势密切相关的面部轮廓的相对位置。在更高级别，逐步显示更详细的信息。第二级网络学习面部组件的相对位置，并且识别组件内部形状的任务由后续级别处理。第三级网络可能会受到本地腐败的影响。但是，由于全球信息被考虑在第二级，因此最终输出仍然有意义。

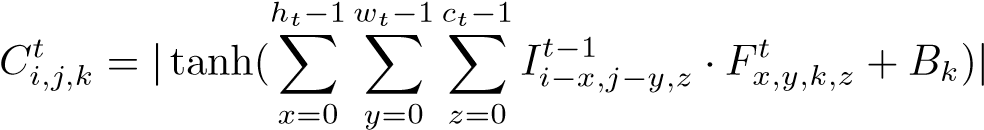
边界框将点组的位置和范围的信息携带到下一级。因此，框内的图像通常在平移和缩放方面很好地对准。相反，由面部检测器产生的矩形远非令人满意。在某些情况下，它包含太多无关的背景信息，这些信息会混淆神经网络。此外，面部并不总是以矩形为中心，这进一步使系统的定位任务复杂化。

DCNN通常被认为足够强大以处理输入图像中的巨大变化，但单个网络的容量仍受其大小的限制。鉴于先前的知识不足，网络将投入相当大的力量来寻找面部的位置。为了解决这个问题，采用了“分而治之”的策略，将任务分为两个步骤：首先找到整体位置，然后计算区域内的相对位置。对于整个面部，第一步由第一级网络执行，其监督信号不包括边界框内的点的详细结构，并且任务的其余部分留给后续级别。以这种方式，负载在不同级别的网络之间共享，并且通过仅中等大小的网络实现良好性能。

该想法在第三和第四级进一步扩展，其中通过图像块的旋转将该方向规范化。仅在第三级之后才考虑旋转，因为在早期级别中未能预测稳健旋转角度所导致的后果是严重的。实验结果表明，第四级的性能增益不如之前的水平那么显着，但绝对不可忽略。

4.实施细节

深度卷积神经网络。 我们使用DCNN作为系统的基本构建块。 网络将原始像素作为输入，并对所需点的坐标执行回归。 图3是深层架构的图示。 在输入节点之后堆叠三个卷积层。 每个卷积层将多个滤波器应用于多通道输入图像并输出响应。 让第t个卷积层的输入为It，然后根据输出计算输出



其中I表示卷积层的输入，F和B是可调参数。 按照标准做法，将超切和绝对值函数应用于滤波器响应，从而为系统带来非线性。

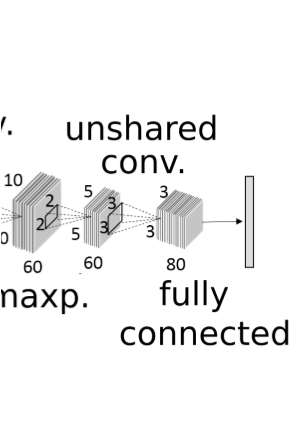
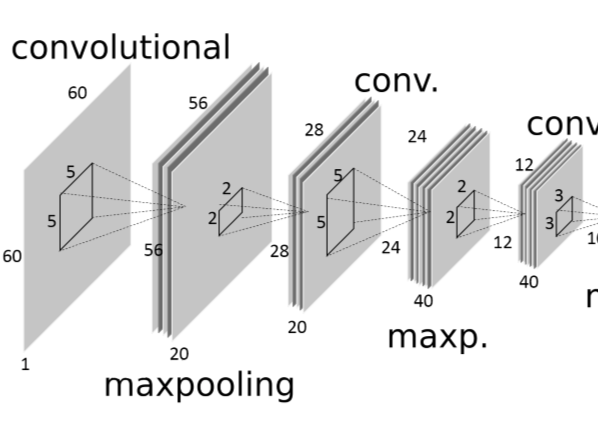
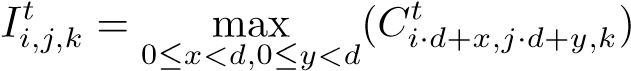
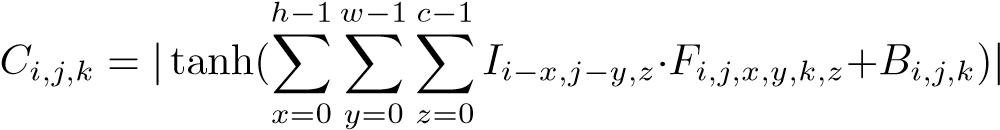


图3.我们系统中网络的典型结构。 该网络由卷积层，非共享卷积层和完全连接层组成。 在卷积层之后执行最大池化。 在非共享卷积层中，在不同位置使用的权重是不同的。 在层之间插入Tanh和绝对值非线性。 其他网络的架构与此类似。

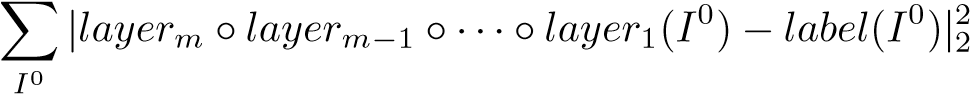
在卷积之后使用具有非重叠池区域的最大池



在寻求像素级精度的本地化任务中使用max-pooling层似乎不自然。 然而，仍然采用这些层，认为由这些层引起的整个系统的鲁棒性很好地补偿了池化操作中的信息损失，并且地标的整体形状和相对位置比像素级更重要。 输入图像中的细节。 在卷积层之后是非共享卷积层。 应用的滤波器在不同的位置上是不同的，因此该层是局部接受的而不是卷积的。



最终预测由一个或两个完全连接的层产生。 调整参数以最小化L2损失：



网络规模我们的DCNN的架构受到[11]的工作的推动。表1总结了网络架构。我们在系统的不同部分使用三种网络。在第二级中使用的网络N1具有更高的分辨率，因为其输入覆盖整个面的范围。

|  |  |  |  |
| --- | --- | --- | --- |
| network | N1 | N2 | N3 |
| input | 60x60 | 40x40 | 40x40 |
| conv. 1 | 5x5x20 | 5x5x20 | 5x5x20 |
| conv. 2 | 5x5x40 | 3x3x40 | 3x3x40 |
| conv. 3 | 3x3x60 | 3x3x60 | 3x3x60 |
| unshared | 3x3x80 | 2x2x80 | 2x2x80 |
| hidden |  |  | 120 |

表1.网络的分辨率，过滤器大小和通道数。 N1用于第二级中的内点。 N2用于轮廓点。 N3用于其他人。 在N3中使用两个完全连接的层，它们之间有120个隐藏单元。 在N1和N2中，一个完全连接的层直接连接输出单元和非共享卷积层。

训练。神经网络通过随机梯度下降和手动调整的超参数进行训练。为了避免严重的过度拟合，在进入网络之前，通过轻微的相似变换（旋转，平移和缩放）随机改变图像。此步骤创建几乎无限数量的训练样本，并使训练错误接近我们验证集上的错误。此外，我们翻转图像以重复使用右眼的左眼模型，并为右眼眉毛重新设置左眼眉。

图像处理。将图像块标准化为零和单位方差，然后应用超正切函数，使像素值落在[-1,1]范围内。在边界框内裁剪图像时，框会放大10％到20％。放大保留了更多的上下文信息，并且它允许系统容忍边界框估计步骤中的小故障。在第四级中，根据面部成分中的两个角点的位置计算旋转角度。

5.实验

我们在一个包含383个图像的数据集上进行了实验，这些图像由300-Faces in the Wild Challenge提供。 图像和注释来自AFW，LFPW，HELEN和IBUG [6,1,12,7,8]。 随机选择500个图像的子集作为我们的验证集。 在验证集上使用两个性能指标：第一个是预测的地标位置与通过眼间距离归一化的地面实况之间的平均距离

其中M是地标的数量，p是预测，g是地面实况，l和r是左眼角和右眼角的位置。 第二个是累积误差曲线，绘制点与标准化距离的百分比。

5.1。 验证我们的方法

68个地标的本地化难度差别很大。 图4显示了不同面部组件的验证错误。 轮廓点的性能明显变差。 这种观察是我们将轮廓与内部点分开的想法的动机。

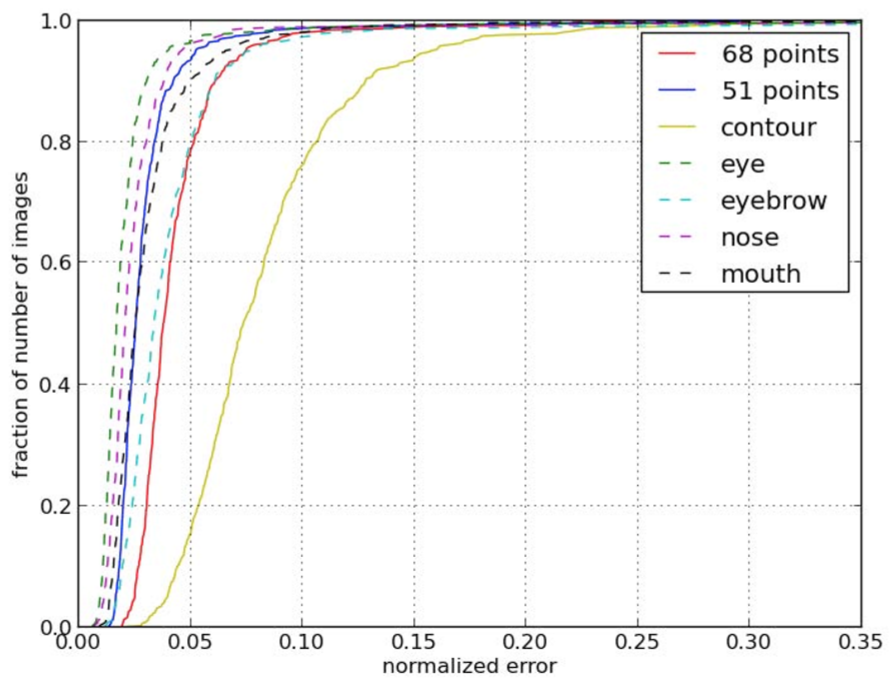


图4.验证集上的累积错误曲线。 比较整个面部，轮廓点，内部点和不同面部成分的误差。 结果表明，不同面部标志点的硬度严重不平衡。

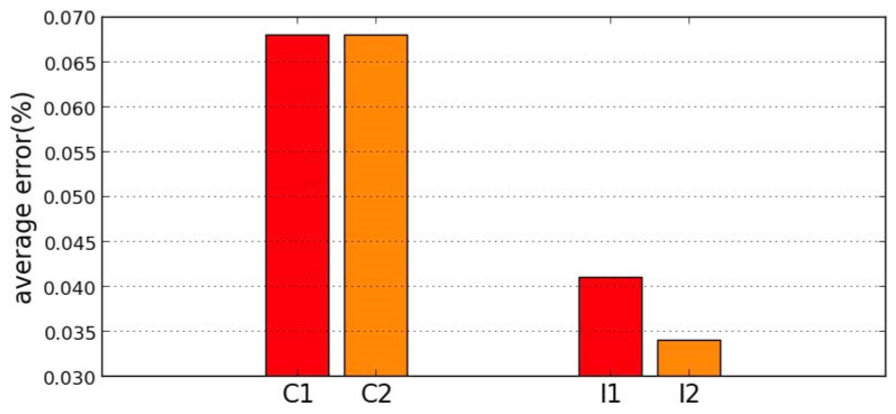


图5.轮廓和内点分离的影响。 一个网络一起预测68个点，其内点和轮廓点的误差分别为C1和I1。 C2和I2由两个独立预测这些点的网络实现

将轮廓与内点分开对我们的表现至关重要。 我们进行了一项实验，其中培训了三个网络。 其中一个，作为一个更大的尺寸，预测68点在一起。 另外两个分别学习了轮廓和内点，并且两个网络的训练和测试所涉及的计算量的总和大致匹配大网络。 为了消除其他影响因素，网络的输入区域根据地面实况值计算。 图5显示分离提高了内点的性能，而轮廓点上的性能并未恶化。

在我们的系统中，面部检测器给出的矩形不直接用于计算产生实际面部标志的网络输入区域。 相反，根据第一级别的预测裁剪第二级别的图像。 如果直接使用面部检测器的盒子，性能会下降。 表2列出了在各种条件下实现的验证错误。 它表示通过在第一级估计的边界框在内点上改善了性能。

|  |  |
| --- | --- |
| output | error value |
| (51 points) detector box | 0.0662 |
| (51 points) level 1 box | 0.0401 |
| level 2 | 0.0510 |
| level 3 | 0.0438 |
| level 4 | 0.0431 |

表2.在各种条件下实现的验证错误。 误差值是预测和地面实况之间的平均归一化距离。 前两行仅显示在内点上计算的误差，而其他行对应于所有68点上的平均误差。

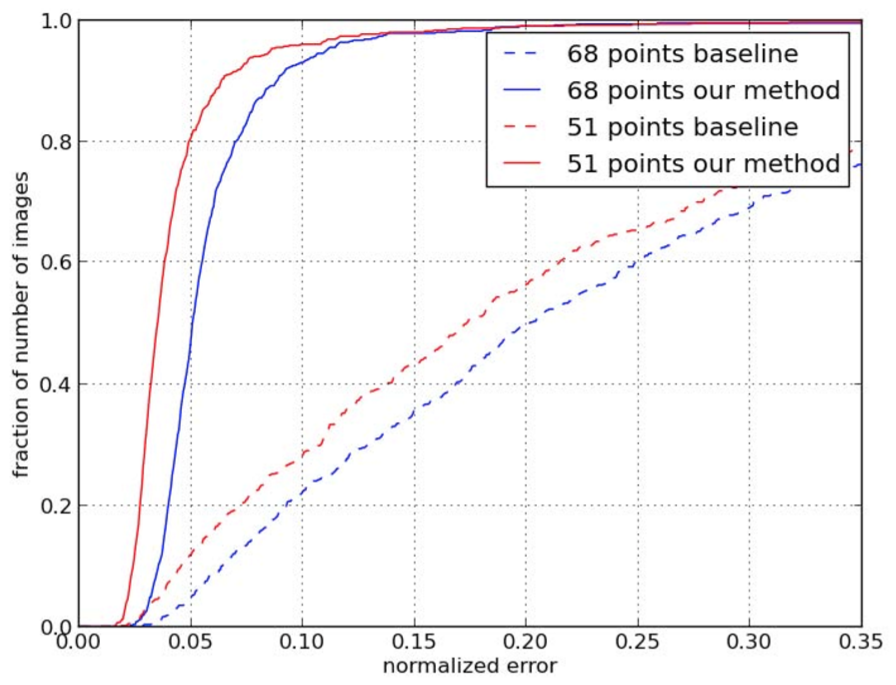


图6. 300-W挑战中的结果比较。 我们的方法显着优于基线。

为了定量研究第三和第四级的影响，我们计算了在每个网络级别实现的验证错误。 为每个组件培训单独的网络允许第三级网络将性能提高14％。 性能增益仍然是通过纠正旋转的第四级获得的。

5.2与其他方法比较

图6是我们的系统在300面临野外挑战中的结果。 我们的结果远远好于基于AAM的基线。 我们还将验证集上的性能与其他系统进行了比较。 由于这些探测器产生了不同的面部地标集，我们只在与它们进行比较时显示共同地标点的相对改进。 表3列出了结果。 我们的系统优于那些公共可用或商业地标检测系统。



图7.验证集中的一些示例。 数据集包含姿势和闪电条件的巨大变化，但我们的系统仍然能够给出良好的结果。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| system | # points | error | ours | improvement |
| Intraface[[1]](#footnote-1) | 37 | 0.046 | 0.029 | 37% |
| FACE++(1.0)[[2]](#footnote-2) | 3 | 0.075 | 0.029 | 61% |
| FACE++ | 11 | 0.034 | 0.026 | 25% |
| Lambda Lab3 | 3 | 0.097 | 0.026 | 73% |

表3.与验证集上的其他公共系统的比较。 误差值是由眼间距离归一化的平均欧几里德距离。 我们的系统优于其他系统。

图7给出了从验证集中获取的一些示例。 我们的系统能够处理包含姿势和闪电条件差异很大的图像。 即使在存在遮挡的情况下，它也可以预测脸部的形状。 尽管取得了成功，仍然存在进一步改进的机会，特别是对于眉毛或脸部轮廓上的点。

六，结论

我们提出了一种新的面部地标定位自动系统。 在我们的方法中，精心设计了四个DCNN级别，以形成粗到细的网络级联。 为了验证我们设计的有效性，我们表明我们的系统可以在300 W面部地标定位挑战中实现领先的性能。

1. http://www.humansensing.cs.cmu.edu/intraface/ [↑](#footnote-ref-1)
2. http://en.faceplusplus.com/ 3http://www.lambdal.com/ [↑](#footnote-ref-2)