联合人脸检测和人脸标点--2014

Joint Cascade Face Detection and Alignment

Dong Chen[1], Shaoqing Ren[1], Yichen Wei[2], Xudong Cao[2], and Jian Sun[2]

第一节：

关键思想是将人脸检测和人脸标点结合起来。

一个应用比较广泛的人脸检测方法，Viola-Jones检测器是基于以下两个原则进行检测的：1，逐步提升的级联结构；2，简单的特征。这种方法在日常生活场景中效果不甚理想。

其他有许多工作是针对多视角的人脸检测【10，17，27，7】他们采用分治策略——在不同视角和头部姿态下，分别训练不同的检测器。这种做法往往更加麻烦，使得系统性能和准确度降低。

一些创新的方法不再使用Boosted-cascade方法，如【32】（2012年）文献之中，在多视角和多表情情况下，采用**部分可变形混合模型**来捕获大规模人脸变化。这种模型很复杂并且可以同时进行人脸检测，表情评估和面部标点。【24】使用基于模范样本的人脸检测器并利用图像检索技术，避免了费时的滑动窗口搜索。这些新方法在一些比较难的图库上的效果都比V-J要好得多【32，9】。然而这些方法，由于复杂性太高（【32】处理一张图需要40秒），不太容易实用。

在本文中，算法的性能和准确度都很高。仍然遵从boosted-cascade方法的原则——“提升级联结构+简单特征”。使用像素差作为特征使得高性能得到保证。我们的算法在**VGA图像（640*480pixel）上的检测性能是28.6ms/张。**

本文中的系统明显优于其他系统【32，9，22】和一些商业系统，图一所示说明我们的系统在多视角，遮挡和低光照度的情况下也能很好的完成检测。
系统运行结果——在比较困难的场景也能完成检测和标点

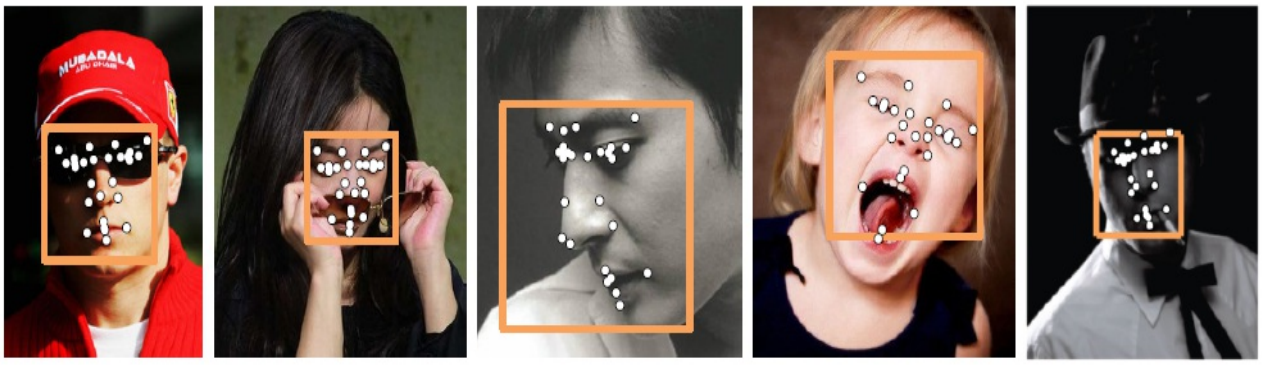

Fig. 1. Face detection and alignment results on challenging examples.

我们最初的动机是发现了精确的人脸标点对于判断是否有脸很有裨益。

在第二节，我们实验发现在后级处理阶段，一个简单的基于面部标定点特征的SVM分类器可以极大地提高V-J检测器的准确度。这一现象的原因是，面部标定可以找到人脸区域直接的对应关系，使得各区域可以直接对比，这简化了人脸/非人脸的分类问题。

我们的问题是，当需求得到高召回率的时候，级联的检测器会返回很多错误的正样本窗口，这使得SVM分类器判别减慢。

我们的方法灵感源于最近在级联人脸标点上的研究【19，4，28，29，21】，这些研究中，人脸形状通过提升的回归器逐渐更新。每一级上的回归器学习不仅取决于图像信息，也取决于从前一级回归器估计到的形状。这种方式的特征学习称之为形状索引特征。这种特征在人脸形状发生几何变化上呈现出更高的不变性。这对获得高的标点准确率和速度而言是很重要的。

由于级联结构在检测和标点上具有如此好的表现，我们提出的方法，将二者结合起来，相得益彰。

在第三节，我们提出一个同一的框架流程，将检测和标点结合起来。由于嵌入了标点信息，检测环节的训练学习变得更加高效。而且，人脸标点也是同时完成了。

在第四节，在我们新的框架下，我们拓展近期一个很前沿的标点方法【21】，我们展现了怎样同时使用简单的形状索引特征进行人脸检测，以使得系统更高效。我们首次提出了联合人脸检测和人脸标点的方法，并首次展现了简单形状索引特征对于人脸检测也很高效。

在第五节，我们证明了我们方法在准确度和速度方面的超高性能。

第二节：标点如何帮助检测人脸：一个后级的分类器。

1，使用OpenCv中的V-J检测器，在一个比较低的阈值下检测，以保证一个高的召回率。
2，将得到许多图像窗口，其中有许多是false positive(错误的正样本)，将这些窗口分为真假正样本两部分，用于训练一个线性SVM分类器。
3，在测试的时候，将V-J检测器输出的窗口，送入SVM分类器中进行筛选。
所有的窗口都被归一化到**96*96**pixel大小
我们实验了几种不同的特征，是否使用标定点，结果如下：
a. 将窗口分为6*6各不重叠的部分，在每一块区域上进行SIFT描述子的提取。
b. 选用27个面部的平均形状点，并以这些点为中心，提取SIFT描述子。
c. 按照【21】中算法得到面部的27标点，并且以此为中心提取SIFT描述子。
上述三种情况下，所有的SIFT描述子都被连接成一个向量，送入SVM训练。
画出测试实验中的分类情况分布图，（1）是原始的V-J检测器，（2）（3）（4）分别是结合了上述三种特征下训练的SVM分类器的分类结果。

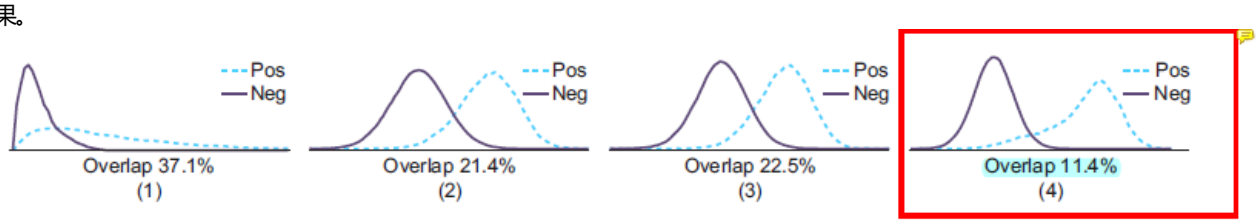

Fig. 2. The distribution of classification scores. (1): original cascade detector; (2)(3)(4) post SVM classifiers using three types of features, as described in Section 2.

实验得出，利用了标定点信息使得那些很难判断的窗口可以得到有效的判决。

考虑到高性能的需求，对于一个标准的级联检测器来说，若需求高召回率时，基于人脸标点的后级分类器会显得很慢。我们的实验中，设置了前级V-J检测器的召回率为99%，每个输入图片会得到3000个输出窗口，再应用后级分类器往往需要几秒的时间。——这很慢

第三节：一个统一的级联人脸检测和标点的框架

为了更好的利用标点信息，我们提出了一个统一的框架。
级联的检测：
不失一般性，分类分数为：

$$f^N = \sum_{i=1}^{N} \mathcal{C}^i(\mathbf{x}). \qquad (1)$$

测试一个待判决的窗口x时，每个弱分类器循序地评估，一旦出现f^n<cita^n，n=1,2,3...N。则立即将此窗口拒绝掉。级联的检测中，负窗口很快会被拒绝，使得其性能很高。

级联的标定：
我们使用了一种结合姿态索引特征和提升回归的姿态回归框架。这种框架在【4，28，29，21】中都证明了其人脸标点的高效性。
我们定义形状S为一个2L维的向量，L是标点总数。逐步回归的过程是：

$$\mathbf{S}^t = \mathbf{S}^{t-1} + \mathcal{R}^t(\mathbf{x}, \mathbf{S}^{t-1}), t = 1, ..., T. \qquad (2)$$

每一个 $\mathcal{R}^t$ 是一个回归函数，它会在前一级形状的基础上增加一个形状增量。学习时使当前形状St与真实形状 $\hat{\mathbf{S}}$ 之间的误差最小。——在所有的样本上考虑这个误差。如下式所示：

$$\mathcal{R}^t = \arg\min_{\mathcal{R}} \sum_i ||\hat{\mathbf{S}}_i - (\mathbf{S}_i^{t-1} + \mathcal{R}(\mathbf{x}_i, \mathbf{S}_i^{t-1}))||^2, \qquad (3)$$

where index i iterates over all the training samples.

一个关键的创新点在于，在级联标点框架中，每一个回归器Rt依赖于前一个形状S(t-1),在训练学习的过程中，特征定义为与S(t-1)相关，所以称之为姿态/形状索引特征【19，4】，这种特征对面部形状变化时，呈现出很好的几何不变性。

将这种姿态/形状索引特征同样的应用在检测环节，**方法是让学习弱分类器的式（1）也依赖于人脸形状信息**。注意到式（1）中的弱分类器的数目N通常是几百上千的规模，远比标点过程中层级T要大（T一般小于10）。为了统一这两个训练学习任务，我们将所有的N个弱分类器分为T个层级，每个层级有K=N/T个弱分类器。这样方程（1）的形式变为如下（4）所示：

$$f = \sum_{t=1}^{T} \sum_{k=1}^{K} \mathcal{C}_k^t(\mathbf{x}, \mathbf{S}^{t-1}). \qquad (4)$$

原则上，每一层级的回归函数和分类函数，不是一定需要同时训练学习和应用。但是为了让两个环节使用同样的特征，为了加速性能，我们同时进行

两个环节的训练学习。

演示了我们测试算法的过程，负图像窗口逐渐被拒绝，并且正图像窗中人脸标定点的位置也不断被校正。下一节，我们将呈现这种联合学习方法。

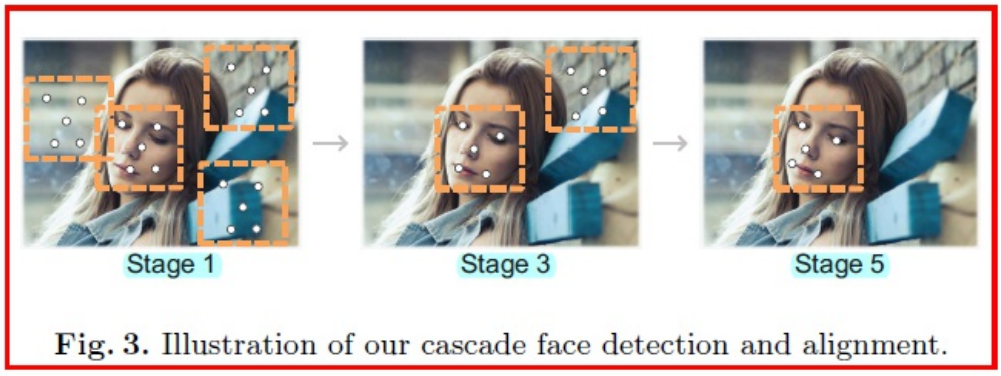

Fig. 3. Illustration of our cascade face detection and alignment.

第四节：我们的方法

如第三节所示，【4，28，29，21】中的标点方法都可以用来进行人脸检测，我们选择其中最新并且效果最好的一篇【21】，它其中的算法最精准，最快，并且容易集成检测过程中的弱分类器学习。

4.1 回顾【21】中的标点方法

回归函数 $\mathcal{R}^t$ 是K个基于树的回归量的和

$$\mathcal{R}^t(\mathbf{x}, \mathbf{S}^{t-1}) = \sum_{k=1}^{K} \mathcal{R}_k^t(\mathbf{x}, \mathbf{S}^{t-1}). \qquad (5)$$

每一个回归量 $\mathcal{R}_k^t$ 是一个决策树（蕨）——它的每个叶子上存储了一个形状增量。图像窗口X落入哪一个叶子上，则决策树（回归量）即输出该叶子上的形状增量。

怎么学习所有的决策树（蕨） $\mathcal{R}_k^t$ 呢？？？

1，估计每一个标定点位置的增量——局部点的树结构学习：
对每一个标定点，使用形状索引像素差特征【4】，训练学习一个标准的回归森林【2】，来估计这个点的增量。

2，全局的树输出的训练学习：
叶子上存储单个标定点的增量的方法被弃置，相反的，每个叶子上存储一个形状增量（多个标定点的增量），并且所有的形状增量使用方程（3）进行优化。注意到这是一个简单的全局线性回归问题。

算法2是使用级联人脸检测和标点方法，对图像窗口x测试的过程。这个模型由所有的弱学习器 $\{\mathcal{CR}_k^t\}$ 和对应的分类阈值 $\{\theta_k^t\}$ 组成。

Algorithm 2 Our testing algorithm for cascade face detection and alignment for an image window $\mathbf{x}$. The model consists of all weak learners $\{\mathcal{CR}_k^t\}$ and classification thresholds $\{\theta_k^t\}$.

initialize the face shape $\mathbf{S}$ as the mean shape in window of $\mathbf{x}$
initialize the detection score $f = 0$
for $t = 1$ to T do
  $\Delta\mathbf{S} = \mathbf{0}$
  for $k = 1$ to K do
    $(f', \Delta\mathbf{S}') = \mathcal{CR}_k^t(\mathbf{x}, \mathbf{S})$
    $f = f + f'$
    if $f < \theta_k^t$ then
      return "not a face"
    end if
    $\Delta\mathbf{S} = \Delta\mathbf{S} + \Delta\mathbf{S}'$
  end for
  $\mathbf{S} = \mathbf{S} + \Delta\mathbf{S}$
end for
return "is a face with shape $\mathbf{S}$"

这个方法的优点在于这两个步骤：
1，局部学习即是在局部块上的单个点的回归，它在简单的像素特征上可以达到更高的抵抗噪声的能力（相比于【4】），同时，这种学习到的特征比人工的SIFT特征更加高效（【28】）。
2，全局学习可以降低局部标点的误差和依赖，一旦给出固定的树的结构，这个步骤能实现全局的最优解，因此，这两个训练学习步骤实现了很强的局部最优解（由方程（3）保证）

4.2 检测和标点的联合学习：

注意到，方程（4）中的弱分类器和方程（5）中的树回归量具有相同的增加形式，所以我们提出在单个决策树中同时学习分类器和回归量。也即，方程5中的每个回归树 $\mathcal{R}_k^t$ 升级为一个混合树 $\mathcal{CR}_k^t$ ——它既能输出分类分数又能输出形状增量。因此，在测试中，分类和回归的部分同时进行评估，如算法2所示。因为混合树 $\mathcal{CR}_k^t$ 使用了与分类和回归相同的特征，因此这种测试比普通的测试更加快速（算法1）

Algorithm 1 The general ~~testing algorithm~~ for cascade face detection and alignment for an image window $\mathbf{x}$.

initialize the face shape $\mathbf{S}$ as the mean shape in window of $\mathbf{x}$
initialize the detection score $f = 0$
for $t = 1$ to T **do**
  for $k = 1$ to K **do**
    $f = f + \mathcal{C}_k^t(\mathbf{x}, \mathbf{S})$
    if $f < \theta_k^t$ **then**
      return "not a face"
    end if
  end for
  $\Delta\mathbf{S} = \mathcal{R}^t(\mathbf{x}, \mathbf{S})$
  $\mathbf{S} = \mathbf{S} + \Delta\mathbf{S}$
end for
return "is a face with shape $\mathbf{S}$"

为了学习一个混合的分类/回归决策树，我们在霍夫森林中使用了与文献【6】中相同的策略（http://note.youdao.com/share/?id=14adbab56d90125d305e90c7268aa848&type=note）：在进行每个节点的划分时，我们随机的选取要么是最小化分类器的**二进制熵**（概率为p），要么最小化面部标定点在**回归器中的增量的变化**（概率为1-p），很直观地，p在前级应该。尽量大，加快分类器拒绝负样本窗口，后级时p应该小，使得回归器标点精度高

根据经验，我们选取了一组随回归层级t线性递减的参数作为概率p

$$\rho(t) = 1 - 0.1t, t = 1, ..., T.$$

在进行树内每个节点的分割时，我们对【4】中的形状索引像素差特征进行了拓展——拓展为多尺度下的特征。具体的，我们通过下采样得到输入图像在三种尺度下的图像（源图像，二分之一下采样，四分之一下采样）。为了生成特征，我们随机的选择了一个图像尺度（三选一），在当前面部标点中选择**随机**的两个点，根据对应点生成两个随机的偏移量，将这两个偏移下的像素差作为特征。我们发现这种多尺度的像素差向量对噪声更加鲁棒，对人脸检测也更加必要。

在级联分类器训练过程中，我们使用ReaLBoost算法，在学习决策树之前，每个样本i都赋予一个权重 w_i

$$w_i = e^{-y_i f_i}, \qquad (6)$$

$y_i \in \{-1, 1\}$，y=1表示是正样本（脸），y=-1代表负样本（非脸）。 f_i 是当前分类器的分类分数。Wi将用于计算分割时的加权的二进制熵。

在每个树的叶子节点处，分类分数由下式给出：

$$\frac{1}{2}\ln\left(\frac{\sum_{\{i \in leaf \bigcap y_i = 1\}} w_i}{\sum_{\{i \in leaf \bigcap y_i = -1\}} w_i}\right), \qquad (7)$$

分子和分母分别是叶子节点上的所有正样本和所有负样本的权重之和

为了实现算法3中，我们的训练算法。所有的人脸形状都按照人脸框进行归一化了。

Algorithm 3 Training of cascade and joint face detection and alignment.

Input: all training samples $\{\mathbf{x}_i\}$, class labels $\{y_i\}$
Input: ground truth shapes $\hat{\mathbf{S}}_i$ for positive samples, $y_i = 1$
Output: all weak learners $\{\mathcal{CR}_k^t\}$, classification thresholds $\{\theta_k^t\}$
set the initial face shapes $\mathbf{S}_i^0$ as random perturbations of the mean shapes in windows of $\mathbf{x}_i$
set all initial classification scores $f_i = 0$
for $t = 1$ to T **do**
  for $k = 1$ to K **do**
    for each training sample i **do**
      compute its weight w_i according to Eq. (6)
    end for
    select a point $(k \bmod L)$ for regression /*local learning in Section 4.1*/
    learn the structure of classification/regression tree $\mathcal{CR}_k^t$ as in Section 4.2
    for each tree leaf **do**
      set its classification score according to Eq. (7)
    end for
    for each training sample i **do**
      update its classification score as $f_i = f_i + \mathcal{CR}_k^t(\mathbf{x}_i, \mathbf{S}_i^{t-1})$
    end for
    use all $\{f_i\}$ to set the bias θ_k^t, according to a preset precision-recall condition
    remove samples whose $f_i < \theta_k^t$ from training set
    perform hard negative sample mining if negative samples are insufficient
  end for
  learn the shape increments of all leaves /* global learning in Section 4.1 */
  compute $\mathbf{S}_i^t$ for all samples according to Eq. (2) and (5)
end for

1，输入：所有训练图像窗口样本Xi，类别Yi(1/-1)

2，输入：正样本窗口中的标定点真实形状，人为标定获取

3，输出：输出所有得到的弱分类器$\{\mathcal{CR}_k^t\}$和分类器的阈值$\{\theta_k^t\}$

4，初始化所有样本的形状 $\mathbf{S}_i^0$，——通过在平均形状（针对所有 $\mathbf{x}_i$ ）上增加一个随机扰量

5 实验

我们从互联网上收集了20000张人脸图片和20000张其他图片（无脸）。所有的脸被手工标注了27个面部点的位置。同时也可使用他们的翻转图片做训练，全部使用灰度图像。

每个分类/回归树的深度为4，对树内每个节点的分割测试次数为2000.这个训练在16核的机器上花费了3天。

然后，我们如第二节所述，使用训练图片集的输出窗口（有很多，分正负）训练了基于标定点的SVM分类器，在检测环节，我们使用了标准的滑动窗口搜索。检测器的输出窗口再送入SVM分类器进行判决。所有通过SVM分类器的窗口，再通过一个裁剪过程：窗口矩形框会聚集起来，从矩形框集群中选择一个框作为最终的输出。

评估时采用了三个很具挑战的公开库：FDDB,AFW，AMU-MIT。这些库不包含我们的训练图片。FDDB和AFW是在外部环境下采集的。CMU-MIT是一个稍微过时但是人脸质量较低且与其他两个库区别较大的库，我们用来测试检测器的容量。。。

5.1 标点对检测的影响：
下图给出我们的方法与基准方法的对比，以证明标点对人脸检测的影响。基准检测器是采用相似的方式进行训练的，只是缺失了算法3的24行，基准的算法中人脸形状始终保持为最初的平均形状。它有点类似标准V-J风格的级联检测器。

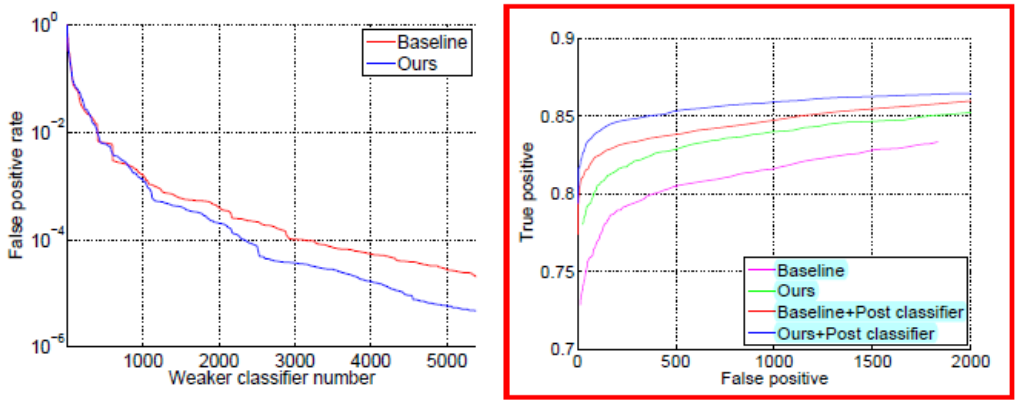

Fig. 4. Comparison of our detector and a baseline detector without alignment. Left: false positive rates all over the detection cascades. Right: recall versus number of false positives, with and without a post classifier.