**基于多任务级联的卷积神经网络**

在实际环境中，人的各种姿势、灯饰和遮挡是影响人脸检测和定位的重要影响因素。近期的研究表明：一种基于多任务深度级联卷积神经网络的算法可以完美的解决了这两个影响因素。本设计使用该算法利用检测和对齐之间的内在相关性的分析来提高性能，该算法使用了层叠式结构，通过精心设计的深度卷积神经网络的三个阶段，由浅入深的方式预测人脸和地标位置。此外，该算法提出了一种全新的在线硬件样本挖掘策略，在实践中进一步提高了性能。本节在详细介绍了基于多任务深度级联卷积神经网络算法的详细原理的同时，通过对比本算法与其他算法基于FDDB（人脸检测数据集）和WIDER FACE和LFW（人脸比对数据集）三种不同的训练集得到的人脸检测准确率进行对比，得出了在保持了实时性性能的同时，本算法比当今最先进的技术具有更高的精度的结果。

### 1 基于多任务深度级联卷积神经网络算法的提出

人脸检测与校正对于诸如面部识别和表情分析等许多面部应用是必不可少的，然而面部巨大的视觉变化诸如：遮挡、大幅度的姿势变化和极端的光照，都会造成识别过程的压力，无疑这是对实际面部应用的一个巨大的挑战。

Viola和JonesP. Viola and M. J. Jones, “Robust real-time face detection. International journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004提出的cascade人脸检测方法利用Haar-Like特征和AdaBoost自适应学习算法对cascade分类器进行训练，取得了很好的实时性和实时性。然而，大量研究表明：该检测器在人脸视觉变化较大的实际应用中，即使具有更先进的特征和分类器，其性能也会显著下降。即使经过改进，在层叠结构的基础上，引入可变形部件模型（DPM，Deformable Part Model）进行人脸检测，取得了显著的效果，但仍然有着计算成本昂贵、在培训阶段需要大批量注释的弊端。近年来，卷积神经网络(Convolutional Neural Network, CNN)在图像分类和人脸识别等计算机视觉任务中取得了显著的进展。受机器视觉中深度学习方法取得重大成功的启发，Yang S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684等人利用基于深度学习框架的卷积神经网络人脸识别算法的研究，训练用于人脸识别的深层卷积神经网络，在人脸区域获得高响应，进而得到人脸候选窗口。然而，由于卷积神经网络结构复杂，这种方法在实际应用中耗费时间。LiH. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334等人采用层叠卷积神经网络进行人脸检测，但需要对人脸检测的边界框（bounding box，一种求解离散点集最优包围空间的算法，是用体积稍大且特性简单的几何体来近似地代替复杂的几何对象）进行标定，需要额外的计算量，从而忽略了人脸地标定位与边界框回归之间的内在相关性。

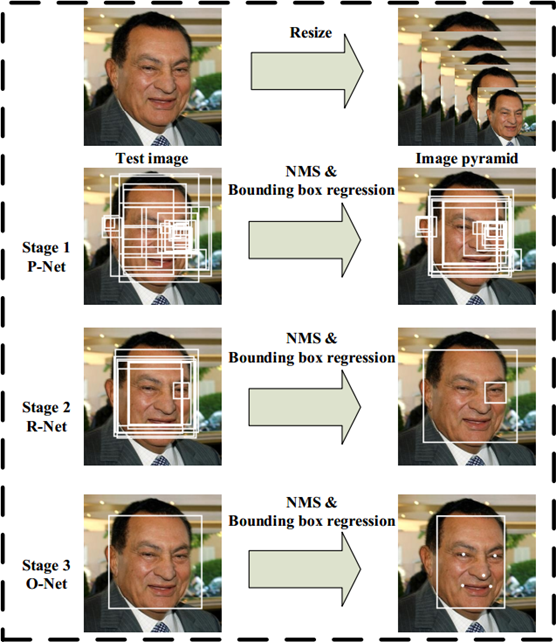
人脸校正也引起了广泛的研究兴趣。在这个领域的研究大致分为两类：基于回归的方法和基于模板匹配的方法。Zhang提出了一种使用脸部属性识别的辅助方法以增强使用深度卷积网络的人脸对齐性能。。

然而，以往大多数的人脸检测和人脸对齐方法都忽略了这两个任务之间的内在联系。现有的几部作品都试图共同解决这些问题，但都存在一定的局限性。例如，Chen等人利用像素值差的特征，联合利用随机森林进行比对和检测。但是，这些手工特征极大地限制了它的性能。Zhang Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision,2014, pp. 94-108利用多任务卷积神经网络来提高多视图人脸检测的准确性，但是检测的召回受到弱人脸检测器产生的初始检测窗口的限制。

另一方面，在训练过程中困难样本挖掘对于增强检测器能力至关重要。然而，传统的挖掘较大损失样本通常以离线方式进行，这显著增加了人工操作。所以，设计一种自动适用于当前训练状态的在线困难样本挖掘的人脸检测方法是可取的。

### 2 基于多任务深度级联卷积神经网络算法的基本工作原理

这一节详细介绍联合人脸检测和人脸对齐的方法。算法包括三个阶段：首先，通过快速提案网络(P-Net)生成候选窗口。在此之后，将在下一阶段通过细化网络(R-Net)来细化这些候选项。在第三阶段，输出网络(O-Net)产生最终的人脸框与标志定位。算法流程如图所示。



A 总体框架

对于一个给定的图像，首先将其调整到不同的比例，以构建一个图像金字塔，这是以下三级级联框架的输入，其推理流程如图所示：



图 MTCNN的推理流程图

第1阶段：首先，利用一个名为提议网络(P-Net)的全连接卷积神经网络获得候选脸部窗口和边界坐标回归向量；然后，估计边界框回归向量进行候选窗口的校准；最后，使用了非最大抑制（NMS，Non-Maximum Suppression）进行合并重叠的候选窗口。

第二阶段:所有候选窗口被传到下一层卷积神经网络——提炼网络（R-Net）该结构进一步排除大量错误候选窗口，进行边界回归校准，并进行NMS抑制。

第三阶段:这一阶段与第二阶段相似，但这一阶段，用更多的监督去识别脸部区域。最终会在网络输出五个脸部标志位置——左/右眼、鼻尖、左/右嘴角。

B 卷积神经网络架构

然而考虑到基础的卷积神经网络可能受到以下因素的影响：

(1)一些卷积层的卷积核缺少多样性，可能限制了它们的识别能力；

(2)与其他多类目标检测和分类任务相比，人脸检测是一个非常具有挑战性的二值类别任务，所以它每一层需要少量的卷积核。为此，我们减少了核的数量并且将5x5的卷积核换成3x3的卷积核以减少计算量，同时增加网络深度以取得更好的效果。有了这些改进，和先前的结构相比，我们用更少的时间取得了更好的效果（训练阶段的结果如表1所示。为了公平比较，我们在每个组中使用了相同的训练集和验证集）。基于多任务深度级联卷积神经网络的网络结构如图2所示。其使用PReLUK. He, X. Zhang, S. Ren, J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in IEEE International Conference on Computer Vision, 2015, pp. 1026-1034作为非线性激活函数，并应用在卷积层和全连接层之后（除了输出层）。

表 CNN与MTCNN处理图像的速度与验证精度对比

|  |  |  |  |
| --- | --- | --- | --- |
| Group | CNN | 300 × Forward  Propagation | Validation Accuracy |
| Group1 | 12-Net  P-Net | 0.038s  0.031s | 94.4%  94.6% |
| Group2 | 24-Net  R-NET | 0.738s  0.458s | 95.1%  95.4% |
| Group3 | 48-Net  O-Net | 3.577s  1.347s | 93.2%  95.4% |

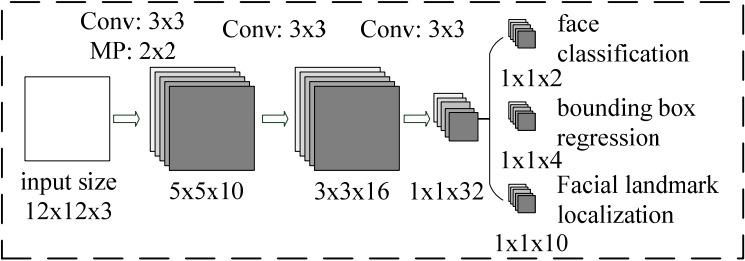


图 P-Net网络结构图

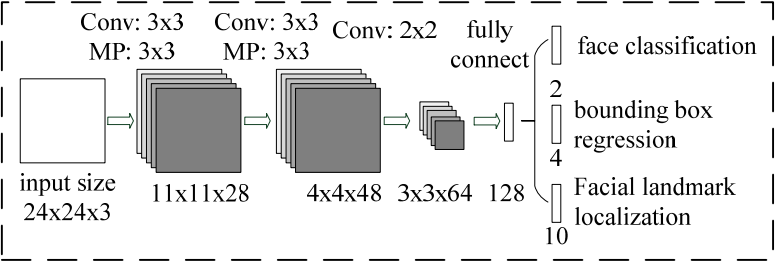


图 R-Net网络结构图

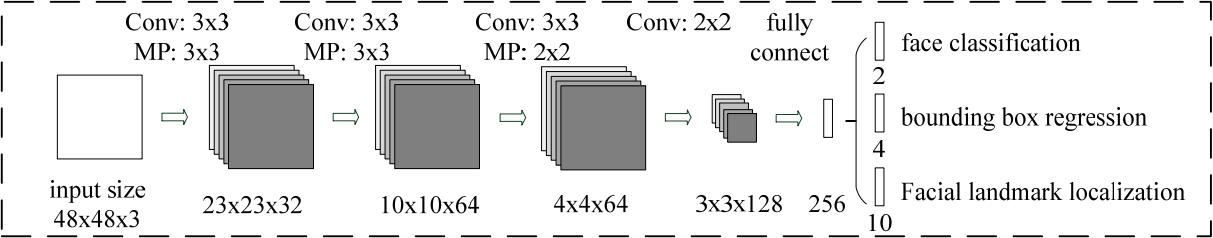


图 O-Net网络结构图

C 训练

在此，利用三个任务来训练MTCNN检测器:人脸/非人脸分类、边界框回归和人脸标志定位。

（1）人脸分类:将学习目标表示为二分类问题。对于每个样本，使用交叉熵损失:

（1）

其中，pi是网络输出的概率，xi表示标准样本为人脸概率。评分表示真实的标签。

（2）边界框回归：对于每个候选窗口，预测它与最近的真实位置的偏移（边界框的左边、顶部、高度和宽度）。学习目标可视为一个回归问题，对每个样本xi使用欧几里得损失：

（2）

其中，为网络输出的回归目标，是真实坐标、一共有四个坐标，包括左边、顶部、高度和宽度，因此。

（3）人脸标志定位：和边界框回归任务一样，人脸标志定位也是一个回归问题，其最小化欧式损失为：

（3）

其中，为网络得到的人脸标志定位坐标，是第i个样本的真实坐标值。该网络共有五个面部标志，包括左眼，右眼，鼻子，左嘴角，右嘴角，因此，。

（4）多源训练：由于在每个卷积神经网络中使用不同的任务，所以在学习过程中有不同类型的训练图像，比如人脸/非人脸、部分对齐脸，在这种情况下，上述三个任务中的将会失效。例如，对于背景区域的样本，通常只计算交叉熵的损失值，而边界框回归与人脸地标定位的损失值则为0。这时，可以通过示例类型指示器直接实现，该整体学习目标表达式为：

（4）

该式中，N为训练样本个数，αi为任务的重要性。为了获得更加准确的人脸标志定位，在P-Net和R-Net中，使用（αdet=1，αbox=0.5，αlandmark=0.5），在O-Net中，使用（αdet=1，αbox=0.5，αlandmark=1）。是样本类型指标。在这种情况下，使用随机梯度下降法来训练多任务深度级联卷积神经网络的网络。

（5）在线困难样本挖掘：与传统分类器训练后进行的传统困难样本挖掘不同，在人脸/非人脸分类任务中进行了在线困难样本挖掘，以适应训练过程。特别是在每一个小批量样本中，对所有样本前向传播计算的损失进行排序，并选择其中的前70%作为困难挖掘样本。然后在反向传播中只计算该梯度，这意味着忽略了那些在训练过程中对增强检测器帮助较小的简单样本。实验表明，该策略在没有人工选择样本的情况下收获了更好的性能，下一节将通过具体的实验数据分析证明其有效性。

### 3 基于多任务深度级联卷积神经网络算法的实验数据

本节先从评估困难样本挖掘的有效性，然后，通过使用MTCNN人脸检测与校准算法与当今最前沿的人脸检测与校准方法进行比较，其中使用的数据集有：FDDB（Face Detection Data Set and Benchmark，人脸检测数据集和基准测试）V. Jain, and E. G. Learned-Miller, “FDDB: A benchmark for face detection in unconstrained settings,” Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.、WIDER FACE （一种用于训练人脸检测的模型）S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A Face Detection Benchmark”. arXiv preprint arXiv:1511.06523.和AFLW（Annotated Facial Landmarks in the Wild，带注释的随机人脸标志）M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2011, pp. 2144-2151.其中，FDDB包含2,845张图像中的5,171张脸的注释；WIDER FACE由32,203张图像中的393,703个带标签的人脸边框样本图片组成，其中50%用于测试(根据图像难度分为三个子集)，40%用于训练，10%用于交叉验证；AFLW包含了24,386个人脸的面部地标注释，使用的Z.Zhang Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision,2014, pp. 94-108.提供的子集进行测试；最后，通过实验数据的分析，评估了基于MTCNN人脸检测器的计算效率。

（1）训练数据

由于实验共同执行人脸检测和对齐，因此在训练过程中使用四种不同类型的数据注释：

1）非人脸负样本：交并比（IoU）比率小于0.3的区域（在目标检测的评价体系中，IoU是模型产生的目标窗口和原来标记窗口的交叠率。可以简单的理解为检测结果与真实人脸的交集比上它们的并集，即为检测的准确率 IoU）;

2）人脸正样本：高于0.65的IoU；

3）部分脸：在0.4至0.65之间的IoU;

4）人脸关键点：标记为左眼、右眼、鼻子、左嘴角和右嘴角等五个关键点的面孔。部分脸和非人脸负样本之间的差距不明确，不同脸部标注之间存在差异；

其中，模糊人脸与清晰人脸样本用来完成人脸分类任务，模糊人脸样本与部分人脸的样本用来完成边界框回归任务，带有五个脸部标志位的人脸样本用于完成人脸关键点定位任务。总训练集中四种数据的占比为3:1:1:2（模糊人脸：清晰人脸：部分人脸：带有五个脸部标志位的人脸），每个网络的训练数据收集描述如下：

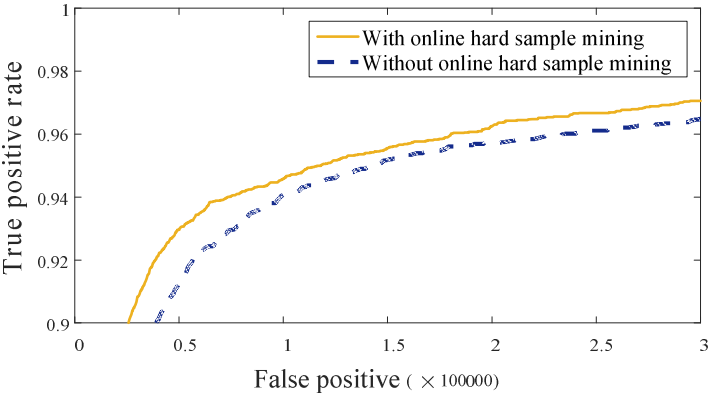
1）P-Net：随机从WIDER FACE中裁剪出几个图片来收集正面，负面和部分脸部。然后将CelebA（香港中文大学提供开放数据，包含10177个名人身份的202599张图片，并且都做好了特征标记，一种人脸相关的训练非常好用的数据集）的人脸作为面部标记。

2）R-Net：使用本算法框架的第一阶段检测WIDER FACE中的脸部来收集正面，负面和部分脸部，并检测CelebA中的面部标记。

3）O-Net：类似于R-Net收集数据，但使用框架的前两个阶段来检测人脸并收集数据。

（2）在线困难样本挖掘的有效性

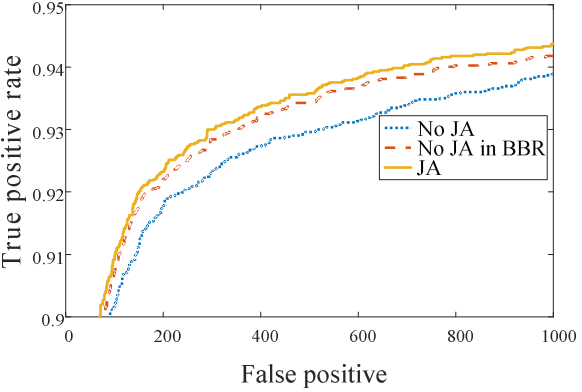
为了评估所提出的在线困难样本挖掘策略的贡献，训练两个P-Nets（有和没有在线困难样本挖掘）并比较它们在FDDB上的表现。 图3（a）显示了在FDDB上两个不同P-Nets检测的结果。很明显，在线困难样本挖掘有利于提高性能，在FDDB可以使整体性能提高1.5％。



有无在线困难样本挖掘的性能对比

（3）联合人脸检测和对齐联合检测的有效性

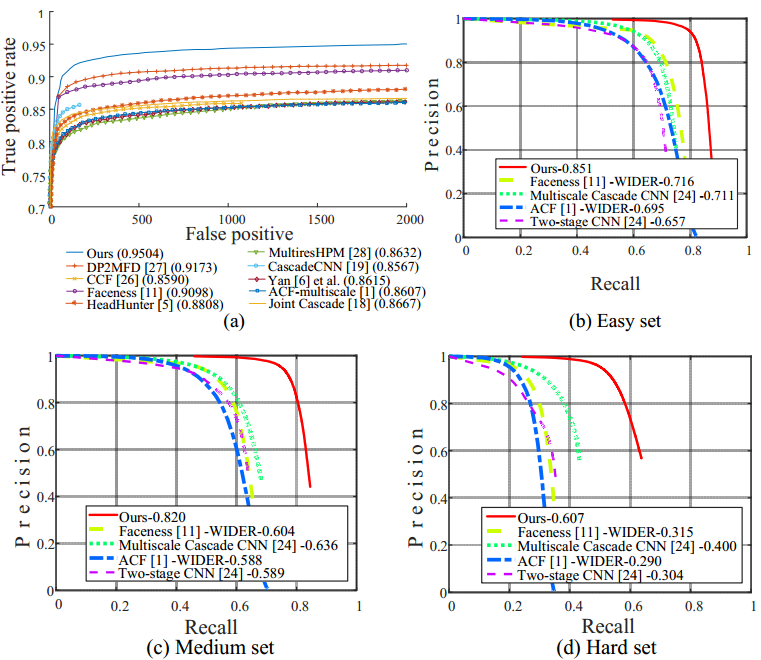
为了评估联合检测和对齐的贡献，评估了的两个不同的O-Nets（有和没有联合面部标志回归学习）在FDDB（具有相同的P-Net和R-Net）上的性能。 还比较了这两个O-Nets中边界框回归的性能。 图表明，联合地标本地化任务学习有助于增强人脸分类和边界框回归任务。



有无联合人脸检测和对齐联合检测的性能对比

（4）人脸检测评估

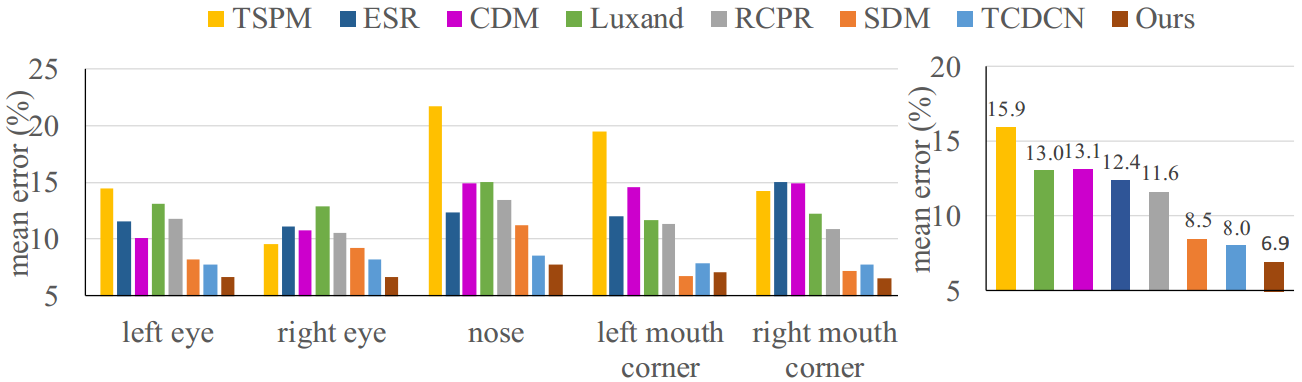
为了评估MTCNN人脸检测方法的性能，将MTCNN与FDDB中的最新的人脸检测方法以及WIDER FACE中的最新的人脸检测方法进行了比较。 图4（a） -（d）表明，MTCNN在两个基准测试中始终优于所有与其比较方法。



MTCNN与其他人脸检测算法的基准测试

（5）面部对齐评估

该部分进行MTCNN中人脸对齐性能与以下算法的人脸对齐性能进行比较：RCPRX. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in IEEE International Conference on Computer Vision, 2013, pp. 1513-1520.，TSPMX. Zhu, and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879-2886.，Luxand face SDKLuxand Incorporated: Luxand face SDK, http://www.luxand.com/，ESRX. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” International Journal of Computer Vision, vol 107, no. 2, pp. 177-190, 2012.，CDMX. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” International Journal of Computer Vision, vol 107, no. 2, pp. 177-190, 2012.，SDMX. Xiong, and F. Torre, “Supervised descent method and its applications to face alignment,” in IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532-539.，和TCDCNZ. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision, 2014, pp. 94-108.。平均误差通过估计值与真实值之间的距离来测量，并且相对于眼间距离进行归一化。图 显示MTCNN优于所有具有余量的最先进的算法。 这也表明MTCNN在嘴角定位方面显示出较小的优势，这可能是由于在训练数据中的过程中对嘴角位置有显著影响的表情变化很小。



脸部对齐在AFLW上的评估

（6）实时性能评估

鉴于级联结构，MTCNN算法可以在联合人脸检测和对齐方面实现高速。如表2所示，将MTCNN与GPU上的最新技术记性比较。值得注意的是MTCNN算法目前的实现是基于未经优化的MATLAB代码。

|  |  |  |
| --- | --- | --- |
| 算法 | GPU型号 | 处理速度 |
| MTCNN | Nvidia Titan Black | 99FPS |
| Cascade CNN | Nvidia Titan Black | 100FPS |
| Faceness | Nvidia Titan Black | 20FPS |
| DP2MFD | Nvidia Tesla K20 | 0.285FPS |

### 4 基于多任务深度级联卷积神经网络算法的实验数据分析

对于一种基于多任务级联卷积神经网络联合人脸检测和对齐的框架。实验结果表明，MTCNN算法始终在多个挑战性数据集（包括用于人脸检测的FDDB、WIDER FACE以及用于面部对齐的AFLW数据集）上均优于最先进的方法，同时对最小面部尺寸为20x20的640x480的VGA图像实现了实时性能 。性能改进的三个主要贡献是仔细设计级联卷积神经网络架构，在线困难样本挖掘的有效性策略和联合人脸对齐学习。

引用

[1] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for

multi-view face detection,” in IEEE International Joint Conference on

Biometrics, 2014, pp. 1-8.

[2] P. Viola and M. J. Jones, “Robust real-time face detection. International

journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004

[3] M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, “Fast polygonal

integration and its application in extending haar-like features to improve

object detection,” in IEEE Conference on Computer Vision and Pattern

Recognition, 2010, pp. 942-949.

[4] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, “Fast human detection

using a cascade of histograms of oriented gradients,” in IEEE Computer

Conference on Computer Vision and Pattern Recognition, 2006, pp.

1491-1498.

[5] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection

without bells and whistles,” in European Conference on Computer Vision,

2014, pp. 720-735.

[6] J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for

object detection,” in IEEE Conference on Computer Vision and Pattern

Recognition, 2014, pp. 2497-2504.

[7] X. Zhu, and D. Ramanan, “Face detection, pose estimation, and landmark

localization in the wild,” in IEEE Conference on Computer Vision and

Pattern Recognition, 2012, pp. 2879-2886.

[8] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial

landmarks in the wild: A large-scale, real-world database for facial land-

mark localization,” in IEEE Conference on Computer Vision and Pattern

Recognition Workshops, 2011, pp. 2144-2151.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification

with deep convolutional neural networks,” in Advances in neural infor-

mation processing systems, 2012, pp. 1097-1105.

[10] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representa-

tion by joint identification-verification,” in Advances in Neural Infor-

mation Processing Systems, 2014, pp. 1988-1996.

[11] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to

face detection: A deep learning approach,” in IEEE International Confer-

ence on Computer Vision, 2015, pp. 3676-3684.

[12] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark

estimation under occlusion,” in IEEE International Conference on Com-

puter Vision, 2013, pp. 1513-1520.

[13] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape

regression,” International Journal of Computer Vision, vol 107, no. 2, pp.

177-190, 2012.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,”

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23,

no. 6, pp. 681-685, 2001.

[15] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, “Pose-free facial

landmark fitting via optimized part mixtures and cascaded deformable

shape model,” in IEEE International Conference on Computer Vision,

2013, pp. 1944-1951.

[16] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder

networks (CFAN) for real-time face alignment,” in European Conference

on Computer Vision, 2014, pp. 1-16.

[17] Luxand Incorporated: Luxand face SDK, http://www.luxand.com/

[18] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection

and alignment,” in European Conference on Computer Vision, 2014, pp.

109-122.

[19] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural

network cascade for face detection,” in IEEE Conference on Computer

Vision and Pattern Recognition, 2015, pp. 5325-5334.

[20] C. Zhang, and Z. Zhang, “Improving multiview face detection with mul-

ti-task deep convolutional neural networks,” IEEE Winter Conference

on Applications of Computer Vision, 2014, pp. 1036-1041.

[21] X. Xiong, and F. Torre, “Supervised descent method and its applications to

face alignment,” in IEEE Conference on Computer Vision and Pattern

Recognition, 2013, pp. 532-539.

[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by

deep multi-task learning,” in European Conference on Computer Vision,

2014, pp. 94-108.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the

wild,” in IEEE International Conference on Computer Vision, 2015, pp.

3730-3738.

[24] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A Face Detec-

tion Benchmark”. arXiv preprint arXiv:1511.06523.

[25] V. Jain, and E. G. Learned-Miller, “FDDB: A benchmark for face detec-

tion in unconstrained settings,” Technical Report UMCS-2010-009, Uni-

versity of Massachusetts, Amherst, 2010.

[26] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional channel features,” in

IEEE International Conference on Computer Vision, 2015, pp. 82-90.

[27] R. Ranjan, V. M. Patel, and R. Chellappa, “A deep pyramid deformable

part model for face detection,” in IEEE International Conference on Bio-

metrics Theory, Applications and Systems, 2015, pp. 1-8.

[28] G. Ghiasi, and C. C. Fowlkes, “Occlusion Coherence: Detecting and

Localizing Occluded Faces,” arXiv preprint arXiv:1506.08347.

[29] S. S. Farfade, M. J. Saberian, and L. J. Li, “Multi-view face detection using

deep convolutional neural networks,” in ACM on International Conference

on Multimedia Retrieval, 2015, pp. 643-650.

[30] K. He, X. Zhang, S. Ren, J. Sun, “Delving deep into rectifiers: Surpassing

human-level performance on imagenet classification,” in IEEE Interna-

tional Conference on Computer Vision, 2015, pp. 1026-1034