保护世界领导人免受深度造假之害

Shruti Agarwal和Hany Farid 美
国加州大学伯克利分校
Berkeley CA, USA
{shruti.agarwal, hfarid}@berkeley.edu

顾玉明、何明明、长野国基和李浩 南加州大学/南加州大学创意技术研究所
美国加州洛杉矶
{ygu, he}@ict.usc.edu, koki.nagano0219@gmail.com, hao@hao-li.com

摘要

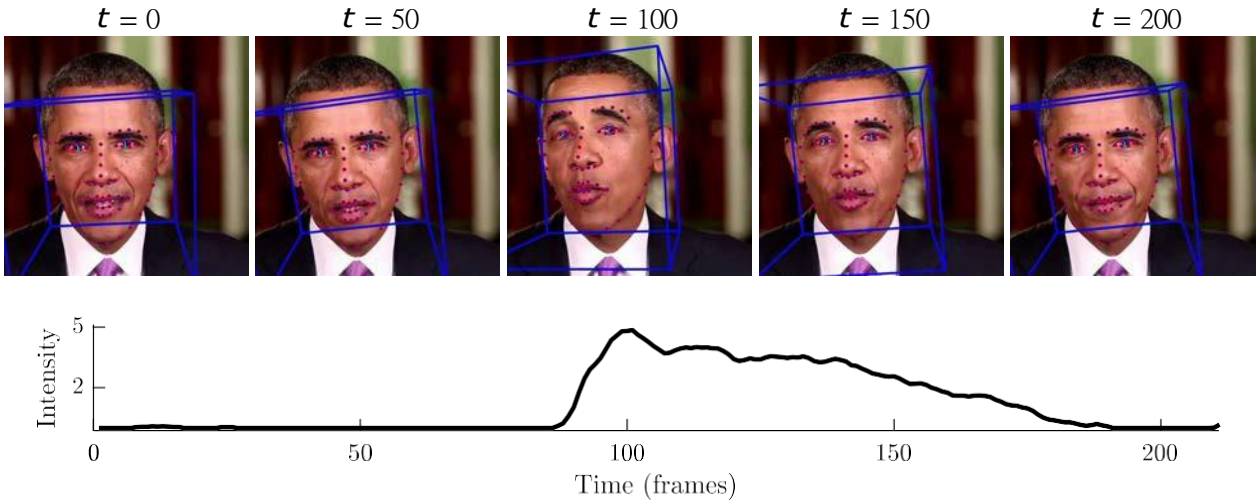
复杂的假视频的制作在很大程度上是由好莱坞电影公司或国家行为者负责的。然而，深度学习的再一次进步，使得创造复杂和引人注目的假视频变得非常容易。例如，通过相对较少的数据和计算机能力，普通人可以制作一个世界领导人承认非法活动并导致宪法危机的视频，一个军事领导人说一些对种族不敏感的话并导致军事活动地区的内乱，或者一个企业巨头声称他们的利润很低并导致全球股票被操纵。这些所谓的深度造假对我们的民主、国家安全和社会构成了重大威胁。为了应对这种日益增长的威胁，我们描述了一种法医技术，该技术可以模拟面部表情和动作，以确定一个人的说话模式。虽然在视觉上并不明显，但这些关联性经常被深度伪造视频的创建方式所违反，因此可以用来进行认证。

1. 简介

虽然通过使用视觉效果对数字图像和视频进行令人信服的操纵已经有几十年的历史了，但最近深度学习的进展导致了虚假内容的真实性和可创建性的大幅提高[27, 14, 29, 6, 19, 21]。这些所谓的人工智能合成的媒体（俗称深度造假）分为三类：（1）换脸，即视频中的脸被自动替换成另一个人的脸。这种类型的技术已经被用来将著名演员插入各种电影片段中，而他们在其中从未出现过。在这种情况下，一个人在原始视频中的相似度被重新置于另一个人的相似度中[13]；（2）唇语，在这种情况下，源视频被修改，使嘴部区域与一个任意的音频记录一致。例如，演员兼导演Jordan Peele制作了一个特别引人注目的例子，奥巴马总统的视频被修改为说 "特朗普总统是个彻头彻尾的笨蛋****"；以及（3）木偶大师，在这种情况下，目标人物被坐在摄像机前的表演者制作成动画（头部动作、眼部动作、面部表情），并表演他们希望木偶说什么和做什么。

虽然这些方法肯定有娱乐性和非恶意的应用，但人们对这些技术可能的武器化表示担忧[7]。例如，在过去的几年里，从针对我国公民的暴力行为到选举篡改，错误信息的严重后果令人不安地上升[22, 28, 26]。复杂和引人注目的假视频的加入可能会使错误信息运动变得更加危险。

关于图像和视频取证有大量的文献[11]。但是，由于人工智能合成的内容是一个相对较新的现象，专门用于检测深度假货的取证技术非常少。其中一个例子是基于一个巧妙的观察，即第一代换脸深层假象中描述的个体要么不眨眼，要么不按预期的频率眨眼[15]。这种假象是由于用于合成人脸的数据通常没有描绘出人的闭眼。可以预见的是，在这项取证技术被公开后不久，下一代的合成技术就将眨眼纳入了他们的系统，因此这项技术现在已经不那么有效了。这个团队还开发了一种技术[31]，用于检测



图一。上面显示的是一个250帧的视频片段中的5个等距的帧，上面注有OpenFace跟踪的结果。下面显示的是在这个视频片段中测量的一个动作单元AU01（眉毛上扬）的强度。

通过利用从整个脸部周围的特征和仅在中央（可能被调换的）脸部区域的特征计算出来的三维头部姿势的差异，来检测换脸的深度假象。虽然在检测换脸方面很有效，但这种方法在检测唇语同步或傀儡师深度造假方面并不有效。

其他取证技术利用了在合成过程中引入的低级像素假象[16, 1, 20, 23, 32, 12, 24, 18]。尽管这些技术以相对较高的精度检测出各种假象，但它们和其他基于像素的技术一样，受到简单的洗钱反措施的影响，这些反措施很容易破坏所测量的假象（例如，添加噪音、重新压缩、调整大小）。我们描述了一种取证技术，旨在检测个人的深度造假。我们为特定的个人定制了我们的取证技术，由于对社会和民主选举的风险，我们把重点放在世界和国家领导人以及高级职位的候选人上。具体来说，我们首先表明，当个人说话时，他们会表现出相对明显的面部和头部运动特征（例如见[9]以及[30]，其中上半身运动被用于识别说话者）。我们还表明，所有这三种类型的深度伪造往往会破坏这些规律，因为表情是由模仿者控制的（换脸和木偶大师），或者嘴巴与脸部其他部位脱钩（唇语同步）。我们通过建立我们所说的高知名度人士的软生物识别模型来利用这些规律性，并利用这些模型来区分真实和虚假的视频。我们在大量美国政治家的深度假象上展示了这种方法的功效，这些政治家包括希拉里-克林顿、巴拉克-奥巴马、伯尼-桑德斯、唐纳德-特朗普和伊丽莎-贝斯-沃伦。与以前的方法不同，这种方法对洗钱有弹性，因为它依赖于相对粗略的不易被破坏的测量结果，并且能够检测出所有三种形式的深度造假。

2. 方法

我们假设，当一个人说话时，他们有不同的（但可能不是唯一的）面部表情和动作。给定一个单一的视频作为输入，我们首先跟踪面部和头部运动，然后提取特定动作单元的存在和强度[10]。然后我们建立一个新奇的检测模型（单类支持向量机（SVM）[25]），将一个人与其他个人以及喜剧性的模仿者和深度模仿者区分开。

2.1. 面部跟踪和测量

我们使用开源的面部行为分析工具包OpenFace2[3, 2, 4] 来提取视频中的面部和头部动作。这个库提供了给定视频中每一帧的二维和三维面部地标位置、头部姿势、眼睛注视和面部动作单元。图1显示了提取的测量结果的一个例子。

面部肌肉的运动可以用面部动作单元（AU）进行编码[10]。OpenFace2工具箱提供了17个动作单位的强度和发生率。眉毛内侧提升器（AU01）、眉毛外侧提升器（AU02）、眉毛下垂器（AU04）、上睑提升器（AU05）、脸颊提升器（AU06）、眼睑收紧器（AU07）、鼻子皱纹器（AU09）、上唇提升器（AU10）。唇角拉紧器(AU12)、调光器(AU14)、唇角压紧器(AU15)、下巴提升器(AU17)、唇部拉伸器(AU20)、唇部收紧器(AU23)、唇部(AU25)、下颌下降器(AU26)和眨眼器(AU45)。

我们的模型包含了16个AU--眼睛眨动的AU被取消了，因为它被发现没有足够的破坏力。

有关人士 (POI)	视频 (小时)	阶层 (小时)	段 (计数)	10秒 夹子(计数)
真正的				
希拉里-克林	5.56	2.37	150	22, 059
顿 巴拉克-奥	18.93	12.51	972	207, 590
巴马 伯尼-桑	8.18	4.14	405	63, 624
德斯 唐纳德-	11.21	6.08	881	72, 522
特朗普	4.44	2.22	260	31, 713
伊丽莎白-沃伦				
喜剧模仿者 希拉里-克				
林顿 巴拉克-奥巴	0.82	0.17	28	1, 529
马 伯尼-桑德斯 唐	0.70	0.17	21	2, 308
纳德-特朗普	0.39	0.11	12	1, 519
伊丽莎白-沃伦	0.53	0.19	24	2, 616
	0.11	0.04	10	264
换脸深假 希拉里-				
克林顿 巴拉克-奥	0.20	0.16	25	1, 576
巴马 伯尼-桑德斯	0.20	11	12	1, 691
唐纳德-特朗普	0.07	0.06	5	1, 084
伊丽莎白-沃伦	0.22	0.19	24	2, 460
	0.04	0.04	10	277
唇齿相依的深假				
巴拉克-奥巴马	0.99	0.99	111	13, 176
木偶大师深藏不露				
巴拉克-奥巴马	0.19	0.20	20	2, 516

表1.下载的视频总时长和其中有POI发言的片段，以及片段的总数和从片段中提取的10秒片段。

图2.从上到下显示的是一个10秒钟的片段的五个例子帧，分别是原始的、唇语深假的、喜剧性的、换脸深假的和木偶大师深假的。

对于我们的目的来说，这16个单位具有鲜明的特点。这16个单位被增加了以下四个特征。(1）头部围绕X轴的旋转（俯仰）；（2）头部围绕Z轴的旋转（滚动）。
(3）嘴角之间的三维水平距离（嘴$_h$ ）；以及（4）下唇和上唇之间的三维垂直距离（嘴$_v$ ）。第一对特征捕捉一般的头部运动（我们不考虑围绕Y轴的旋转（偏航），因为在直接对个人说话时与在大群人中说话时有不同）。第二对特征捕捉嘴部伸展（AU27）和嘴唇吸吮（AU28），这些都是默认的16个单位所不能捕捉的。

我们使用皮尔逊相关度来衡量这些特征之间的线性关系，以描述个人的运动特征。在总共20个面部/头部特征中，我们计算了所有20个特征之间的皮尔逊相关度。
这些特征，产生$_{20}C_2 = (20 \times 19)/2 = 190$对
所有10秒重叠的视频片段的特征（见2.2节）。因此，每个10秒的视频片段都被重新划分为一个维度为190的特征向量，正如接下来所描述的那样，它被用来对视频进行真假分类。

2.2. 数据集

我们专注于感兴趣的人（POI）在正式场合的谈话视频，例如，每周的广告、新闻采访和公开演讲。所有的视频都是从YouTube上手动下载的，其中的POI主要是朝向摄像机的。对于每个下载的

我们手动提取了符合以下条件的视频*片段*
以下要求。(1)该段至少有10秒的时间。

(2) POI在整个片段中说话；(3) 片段中只有一张脸--POI--是可见的；(4) 在片段中摄像机是相对静止的（允许缓慢变焦）。所有的片段都以30帧的速度保存，使用mp4格式，质量相对较高，为20。然后将每个片段分割成10秒以上的片段（片段的提取是通过在片段上滑动一个窗口，每次5帧）。表1显示的是五个POI的视频和片段的持续时间以及提取的片段数量。

我们用以下数据集测试了我们的方法：1)
来自FaceForensics数据集[23]的1,004个独特人物的5.6小时视频片段，产生30,683个10秒的片段；2）每个POI的喜剧模仿者，（表1）。
3）换脸式深度伪造、唇语式深度伪造和木偶大师式深度伪造（表1）。图2显示的是来自原始视频的10秒片段的5个例子帧，一个唇语深度伪造，一个喜剧模仿者，一个换脸深度伪造，以及巴拉克-奥巴马的木偶大师深度伪造。

2.2.1 深度造假

使用他们的喜剧模仿者的视频作为基础，我们为每个POI生成了换脸的深度假象。为了在每个POI和他们的模仿者之间交换面孔，我们根据深度假象架构[1]，训练了一个生成式广告网络（GAN）。每个GAN都是用每个POI大约5000张图片来训练的。然后，GAN取代了

[1]github.com/shaoanlu/faceswap-GAN

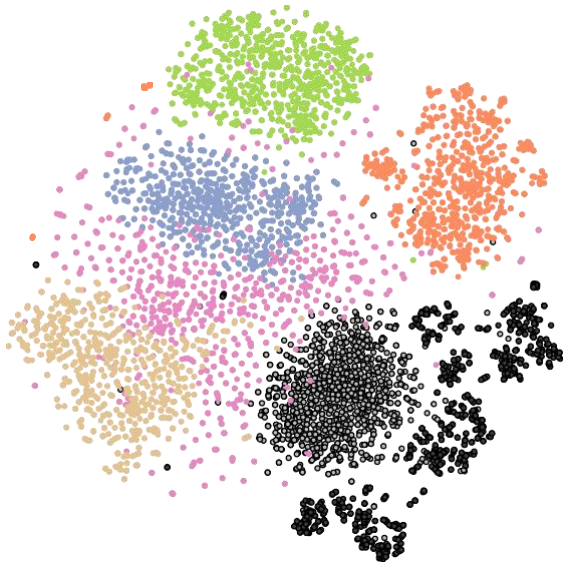


图3.显示的是希拉里-克林顿（棕色）、巴拉克-奥巴马（浅灰色带黑边）、伯尼-桑德斯（绿色）、唐纳德-特朗普（橙色）、伊丽莎-贝斯-沃伦（蓝色）、随机人物[23]（粉色）和巴拉克-奥巴马的唇语深假（深灰色带黑边）190-D特征的2维可视化。

冒充者的脸和POI的脸，在每一帧视频上匹配冒充者的表情和头部姿态。我们首先使用dlib检测面部地标和面部包围盒。边界框的82%的中心部分被用来生成POI的脸。然后用面部地标将生成的脸与原始脸对齐。面部标志的轮廓被用来生成一个用于后期处理的面具，其中包括阿尔法混合和颜色匹配，以提高最终面部交换视频的空间-时间一致性。

使用巴拉克-奥巴马的喜剧模仿者，我们还为奥巴马生成了木偶大师的深度假象。照片真实化身GAN（paGAN）[19]从一张照片中合成照片真实的脸。这个基本过程生成的视频只有一个在静态黑色背景上的浮动头部。除了创造这些类型的假象外，我们还修改了这个合成过程，在训练过程中去掉了人脸面具，使我们能够生成具有完整背景的视频。这些视频的时间一致性是通过调节网络的多个帧来证明的，使网络能够及时看到[14]。这个修改后的模型只用巴拉克-奥巴马的图像进行训练。

虽然这两种类型的假货在视觉上都很吸引人，但它们偶尔会包含时空上的故障。然而，这些故障正在不断地被排除，我们期望未来的版本将重新产生几乎没有故障的视频。

2.3. 建模

图3显示的是希拉里-克林顿、巴拉克-奥巴马、伯尼-桑德斯、唐纳德-特朗普、伊丽莎白-沃伦、随机人物[23]和巴拉克-奥巴马的唇语深假的二维t-SNE[17]可视化的190维特征。请注意，在这个低维代表中，POI彼此之间有很好的分离。这表明，所提出的动作单元和头部运动的相关性可以用来区分不同的人。我们还注意到，这种可视化支持使用单类支持向量机（SVM）的决定。特别是，如果我们训练一个两类SVM来区分奥巴马（浅灰色）和随机的人（粉红色），那么这个分类器几乎会完全错误地分类出深度假象（深灰色，黑色边框）。

在理想的世界里，我们会建立一个大型的个人真实视频数据集和一个大型的同一个人的假视频数据集。然而，在实践中，这并不实际，因为它需要一个广泛的假视频集，而此时制造假货的技术正在迅速发展。因此，我们训练一个新奇的检测模型（一类SVM[25]），它只需要一个POI的真实视频。对于在YouTube等视频分享网站上有大量存在的世界和国家领导人以及高官候选人来说，获取这些数据相对容易。

控制高斯核宽度和离群点百分比的SVM超参数γ和ν是利用FaceForensics原始视频数据集[23]中10%的随机人的视频片段进行优化的。具体来说，我们对γ和ν进行了网格搜索，并选择了在POI和这些随机人之间产生最高区分度的超参数。这些超参数针对每个POI进行了调整。在从重叠的10秒片段中提取的190个特征上训练SVM。在测试过程中，SVM符号决策函数的输入被用作一个新的10秒片段的分类分数[25]（负分对应于假视频，正分对应于真视频，分数的大小对应于与决策边界的距离，可以作为一种信心的衡量标准）。我们接下来报告我们的分类器的测试准确性，所有10秒的视频片段被分成80:20的训练：测试数据集，其中没有重叠的数据。训练和测试视频片段。

3. 结果

每个特定的POI模型的性能是使用特定的喜剧模仿者和深层假货（第2.2节）来测试的。我们将测试准确性报告为接收者操作曲线（ROC）的曲线下面积（AUC）和在固定的假阳性率下正确识别原件的真阳性率（TPR）。

	随机人	喜剧性冒名顶替者	换脸	唇齿相依	傀儡-掌握
190-特征					
10秒短片					
TPR (1% FPR)	0.62	0.56	0.61	0.30	0.40
TPR (5% FPR)	0.79	0.75	0.81	0.49	0.85
TPR (10% FPR)	0.87	0.84	0.87	0.60	0.96
AUC	0.95	0.94	0.95	0.83	0.97
段					
TPR (1% FPR)	0.78	0.97	0.96	0.70	0.93
TPR (5% FPR)	0.85	0.98	0.96	0.76	0.93
TPR (10% FPR)	0.99	0.98	0.97	0.88	1.00
AUC	0.98	0.99	0.99	0.93	1.00
29-特征					
10秒短片					
AUC	0.98	0.94	0.93	0.95	0.98
段					
AUC	1.00	0.98	0.96	0.99	1.00

表2.所显示的是巴拉克-奥巴马的总体准确率，即曲线下面积（AUC）和三种不同的假阳性率（FPR）的真阳性率（TPR）。上半部分对应的是使用全套190个特征的10秒视频片段和完整视频片段的准确性。下半部分对应的是只使用29个特征。

(FPR)为1%、5%和10%。这些准确率是针对10秒钟的片段和完整的视频片段报告的。一个视频片段的分类是基于所有重叠的10秒片段的SVM得分的中位数。我们首先介绍了 对原版和假奥巴马视频的详细分析，然后是对其他POI的分析。

3.1. 巴拉克-奥巴马

表2的上半部分显示的是根据190个特征对奥巴马的视频进行分类的准确率。前四行对应的是10秒片段的准确度，后四行对应的是完整视频片段的准确度。10秒片段和完整片段的平均AUC为0.93和0.98。唇部同步假象的AUC最低，分别为0.83和0.93，这可能是因为与其他假象相比，这些假象只操纵了嘴部区域。因此，许多面部表情和动作在这些假动作中得到了保留。然而，正如接下来所显示的，通过一个简单的特征剪裁技术，可以提高唇部同步假动作的准确性。

为了选择最佳的分类特征，用1到190个特征对190个模型进行了反复的训练。具体来说，在第一次迭代中，190个模型只使用一个特征进行训练。挑选出总体训练精度最好的特征。在第二次迭代中，189个模型使用两个特征进行训练，其中第一个特征是在第一次迭代中确定的。挑选出总体训练精度最好的第二个特征。整个过程重复了190次。图4显示的是这个过程的前29次迭代中测试精度与特征数量的关系（训练精度在29个特征时达到最大值）。这

	随机人	喜剧性冒名顶替者	换脸
希拉里-克林顿			
TPR (1% FPR)	0.31	0.22	0.48
TPR (5% FPR)	0.60	0.55	0.77
TPR (10% FPR)	0.75	0.76	0.89
AUC	0.91	0.93	0.95
伯尼-桑德斯			
TPR (1% FPR)	0.78	0.48	0.58
TPR (5% FPR)	0.92	0.70	0.84
TPR (10% FPR)	0.95	0.84	0.92
AUC	0.98	0.94	0.96
唐纳德-特朗普			
TPR (1% FPR)	0.30	0.39	0.31
TPR (5% FPR)	0.65	0.72	0.60
TPR (10% FPR)	0.77	0.83	0.74
AUC	0.92	0.94	0.90
伊丽莎白-沃伦			
TPR (1% FPR)	0.75	0.97	0.86
TPR (5% FPR)	0.91	0.98	0.91
TPR (10% FPR)	0.95	0.99	0.92
AUC	0.98	1.00	0.98

表3.显示的是希拉里-克林顿、伯尼-桑德斯、唐纳德-特朗普和伊丽莎白-沃伦的10秒视频剪辑的总体准确率。准确率以曲线下面积（AUC）和三种不同假阳性率（FPR）的真阳性率（TPR）报告。

迭代训练是在10%的随机人物、喜剧模仿者和所有三种类型的深度伪造的10秒视频片段上进行的。

在只有13个特征的情况下，AUC几乎达到了0.95的平均水平。该图中没有显示的是，在包括30个特征后，accu- racy开始慢慢减少。排名前五位的区分特征是以下几个方面的相关性。
(1) 上唇提升器(AU10)和嘴角之间的三维水平距离(嘴部$_h$)；(2) 唇角去压器(AU15)和嘴部$_h$ ；(3) 头部围绕X轴的旋转(间距)和嘴部$_v$ ；(4) 减速器(AU14)和间距；和
(5）唇角压舌板（AU15）和嘴唇部分（AU25）。值得注意的是，这前五名的关联中至少有一个与嘴部相对应的成分。我们假设，这些特征是最重要的，因为唇语假唱的性质只修改嘴部区域，而换脸、木偶大师和喜剧模仿者根本无法捕捉到微妙的嘴部动作。

表2的下半部分显示了全部190个特征和图4中列举的29个特征的准确性比较。表中黑体字的数值表示相对于全部190个特征集而言，准确率有所提高。接下来，我们测试了这29个特征对简单的洗钱攻击、对提取的视频片段的长度以及对说话背景的稳健性。

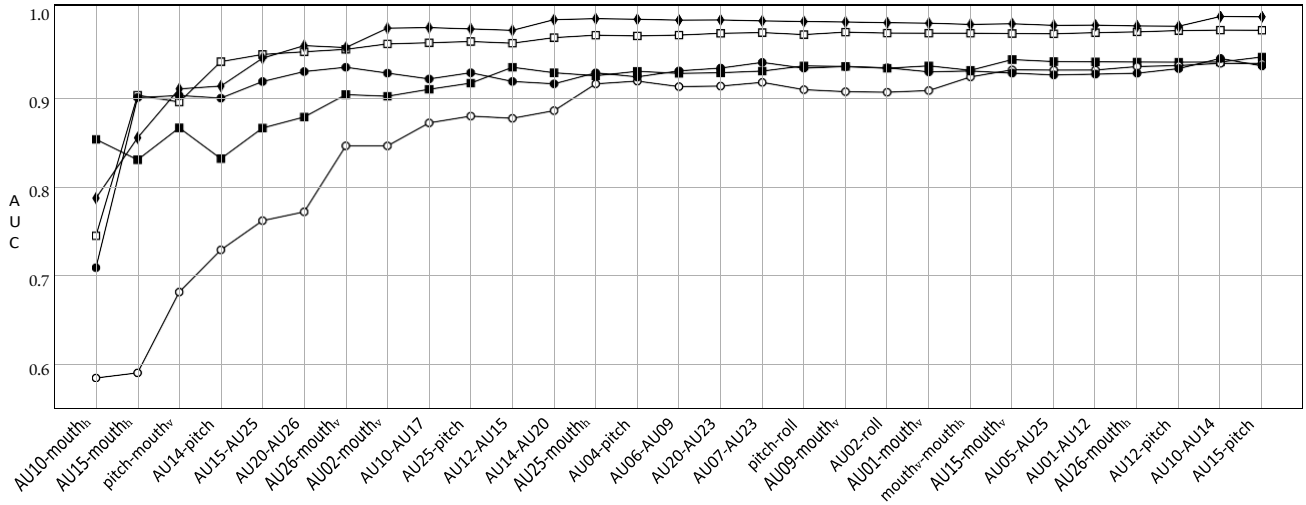


图4.喜剧模仿者（黑色方块）、随机人（白色方块）、唇语深度假唱（黑色圆圈）、换脸深度假唱（白色圆圈）和木偶大师（黑色钻石）的准确率（作为AUC），其分类器在横轴上列举了1到29个特征。特别是，AU10-口$_h$ ，*其*准确率对应于仅在此特征上训练的SVM。AU15嘴$_h$ *的*准确率对应于在两个特征上训练的SVM，AU10嘴$_h$ 和AU15嘴$_h$ 。总体准确率在大约13个特征时趋于平稳。

3.1.1 稳健性

如前所述，许多取证技术在面对像重新压缩这样的简单攻击时都会失败，因此我们测试了我们的方法对这种类型的洗钱行为的稳健性。每个原始和伪造的视频片段最初都以H.264量化质量为20保存。然后，每个片段以较低的40质量重新压缩。区分巴拉克-奥巴马与随机人物、喜剧模仿者、换脸、对口型和木偶大师的10秒片段的AUCs在这种清洗之后是：。0.97、0.93、0.93、0.92和0.96，与高质量的视频几乎没有变化（见表2）。正如预期的那样，由于我们的分析不依赖于像素级的假象，我们的技术对简单的清洗攻击是稳健的。

为了确定对片段长度的稳健性，我们使用长度为2、5、15和20秒的片段重新训练了四个新模型。所有视频的平均AUC为0.80、0.91、0.97和0.98，而剪辑长度为10秒的 AUC为0.96。正如预期的那样，较短的片段的准确率下降，但基本上不受10到20秒之间的片段长度的影响。

一个人的谈话风格和面部行为会随着他的谈话环境而变化。例如，在发表准备好的演讲时的面部行为，与在现场采访中回答一个充满压力的问题相比，会有很大的不同。在两个后续的体验中，我们测试了奥巴马模型在各种背景下的稳健性，这些背景不同于用于训练的每周讲话。

在第一个实验中，我们收集了一些视频，其中，像

在每周的讲话中，奥巴马都在对着摄像机说话。然而，这些视频跨越了各种背景，从宣布奥萨马-本-拉登的死亡到总统辩论视频，以及宣传视频。我们共收集了1.5小时的此类视频，产生了91段1.3小时的视频和21 ，152个超过10秒的片段。在AUC方面，将这些视频与喜剧演员、随机人物、唇语假唱、换脸假唱和木偶大师假唱区分开来的平均准确率，10秒片段为0.91，完整片段为0.98，而以前的准确率为0.96和0.99。尽管背景不同，我们的模型似乎对这些新的背景有合理的概括。

在第二个实验中，我们收集了另一轮奥巴马的视频，其内容差异更大，从他看着采访者而不是摄像机的采访到他在回答问题时明显停顿并倾向于向下沉思的现场采访。我们总共收集了4.1小时的视频，产生了140个1.5小时的视频片段和19 855个10秒的重叠片段。平均AUC明显下降到
10秒钟的片段为0.61，片段为0.66。在这种情况下，视频的背景明显不同，所以我们的原始模型没有捕捉到必要的特征。然而，在原始数据集和这些采访风格的视频上重新训练奥巴马模型时，10秒钟的片段和片段的AUC增加到了0.82和0.87。尽管有了这一改进，我们看到准确率没有以前那么高了，这表明我们可能要训练POI和PPT。

特定背景下的模型和/或用更稳定和针对POI的特征来扩展当前的功能。

3.1.2 与FaceForensics++比较

我们将我们的技术与FaceForensics++[24]中使用的基于CNN的方法进行比较，其中多个模型被训练来检测三种类型的人脸操作，包括换脸的深度伪造。我们评估了使用XceptionNet[8]架构训练的性能较高的模型，并将裁剪过的脸作为输入。这些模型的性能在真实的、换脸深度伪造的、唇语同步深度伪造的和木偶大师深度伪造的奥巴马视频上进行了测试，这些视频以高低质量保存（喜剧模仿者和随机人物数据集没有被使用，因为它们不是合成的内容）。我们测试了作者提供的模型[2]，没有对我们的数据集进行任何微调。

真实类的每帧CNN输出被用来计算准确率（AUC）。在质量为20/40的情况下，检测换脸、木偶大师和唇语深层造假帧的总体准确率为0.84/0.71、0.53/0.76和0.50/0.50，而我们的平均AUC为0.96/0.94。尽管FaceForensics++在换脸深层假象上工作得相当好，但它却不能推广到它在训练过程中没有见过的唇部同步深层假象。

3.2. 其他领导人/候选人

在这一部分，我们分析了为希拉里-克林顿、伯尼-桑德斯、唐-特朗普和伊丽莎白-沃伦训练的SVM模型的性能。图5显示的是为这四位领导人收集的视频的样本帧（见表1）。对于每个POI，使用190个特征的完整集合训练一个模型。表3显示的是对希拉里-克林顿、伯尼-桑德斯、唐纳德-特朗普和伊丽莎白-沃伦的10秒短片进行分类的准确率。这些POI的平均AUC为0.93、0.96、0.92和0.98。

4. 讨论

我们描述了一种取证方法，利用独特和一致的面部表情来检测深度造假。我们表明，面部表情和头部动作之间的相关性可以用来将一个人与其他人以及他们的深度伪造视频区分开来。这项技术的可靠性针对压缩、视频片段长度和人说话的环境进行了测试。与现有的基于像素的检测方法相比，我们的技术对压缩是稳健的。然而，我们发现，我们的方法的适用性很容易受到人说话的不同环境的影响（例如，正式的准备好的讲话直接看向摄像机和现场采访时看向摄像机之外）。我们将通过两种方式之一来解决这一局限性。

[2] niessnerlab.org/projects/roessler2019faceforensicspp.html

图5.图中显示的是(a)真实；(b)喜剧人物；和(c)四个POI的换脸样本框架。

只需在广泛的背景下收集更大、更多样化的视频，或建立针对POI和背景的模型。除了这种语境效应外，我们发现当POI持续看向远处的摄像机时，动作单元的可靠性可能会被大大压缩。为了解决这些局限性，我们建议用语言学分析来增强我们的模型，以捕捉正在说什么和如何说之间的相关性。

鸣谢

这项研究由谷歌、微软和美国国防部高级研究计划局（FA8750-16-C-0166）资助。所表达的观点、意见和发现是作者的，不应该被解释为代表国防部或美国政府的官方观点或政策。李浩隶属于南加州大学、南加州大学/信息通信技术学院和Pinscreen。这个项目不是由Pinscreen公司资助的，也不是在Pinscreen公司进行的。Koki Nagano隶属于Pinscreen，但通过他在USC/ICT的关系参与了这个项目。我们感谢Ira Kemelmacher-Shlizerman和Supasorn Suwajanakorn提供的唇部同步深层假象。

参考文献

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen.Mesonet：一个紧凑的面部视频伪造检测网络。在*IEEE信息取证与安全国际研讨会上*，第1-7页。IEEE，2018年。2

[2] Tadas Baltrus ˇ aitis, Marwa Mahmoud, and Peter Robinson.用于自动动作单元检测的跨数据集学习和特定人的规范化。在*第11届IEEE自动脸部和手势识别国际会议和研讨会上*，第6卷，第1-6页。IEEE，2015年。2

[3] Tadas Baltrus ˇ aitis, Peter Robinson, and Louis-Philippe Morency.Openface：一个开源的面部行为分析工具包。在*IEEE计算机视觉应用冬季会议上*，第1-10页。IEEE，2016年。2

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis- Philippe Morency.Openface 2.0:面部行为分析工具包。在*第13届IEEE自动人脸和手势识别国际会议上*，第59-66页。IEEE，2018年。2

[5] Jennifer Finney Boylan.深度造假技术会摧毁民主吗，2018。1

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros.Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018.1

[7] Robert Chesney和Danielle Keats Citron.深度造假。对隐私、民主和国家安全的一个迫在眉睫的挑战。技术报告公法研究论文第692号，德克萨斯大学法学院，2018年。1

[8] Franc¸ois Chollet.Xception:带有深度可分离卷积的深度学习。在*IEEE计算机视觉和模式识别会议上*，第1251-1258页，2017。7

[9] Jeffrey F Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman.面部表情的个体差异。面部表情的个体差异：随时间变化的稳定性，与自我报告的情绪的关系，以及对人的识别的能力。在*第四届IEEE多模态界面国际会议上*，第491页。IEEE计算机协会，2002年。2

[10] Paul Ekman 和 Wallace V Friesen.测量面部动作。*环境心理学和非语言行为*，1（1）：56-75，1976。2

[11] H.法里德.*Photo Forensics*.麻省理工学院出版社，2016年。1

[12] David Gu¨era和Edward J Delp.使用递归神经网络的深度虚假视频检测。In *IEEE International Con- ference on Advanced Video and Signal Based Surveillance*, pages 1-6, 2018.2

[13] 德鲁-哈维尔。斯嘉丽-约翰逊谈人工智能生成的假性爱视频：没有什么能阻止有人剪切和粘贴我的形象，2018年。1

[14] H.Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P.Pe´rez, C. Richardt, M. Zollho¨fer, and C. Theobalt.深度视频肖像。*ACM Transactions on Graphics*, 2018.1, 4

[15] 李岳尊，张明清，和刘思伟.In ictu oculi: 通过检测眼睛的眨动来揭露AI创建的假视频。在*IEEE信息取证和安全研讨会*上。
香港，2018年。1

[16] Yuezun Li and Siwei Lyu. *ArXiv preprint arXiv:1811.00656*, 2018.2

[17] Laurens van der Maaten 和 Geoffrey Hinton.使用t-SNE对数据进行可视化。*机器学习研究杂志*，9(11):2579-2605, 2008.4

[18] Falko Matern, Christian Riess, and Marc Stamminger.利用视觉假象来揭露深层假象和人脸操纵。In *IEEE Winter Applications of Computer Vision Workshops*, pages 83-92.IEEE，2019年。2

[19] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. paGAN: real-real avatars using dynamic textures.在*SIGGRAPH亚洲技术文件*中，第258.ACM, 2018.1, 4

[20] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen.Capsule-forensics:Using capsule networks to detect forged images and videos. *arXiv preprint arXiv:1810.11215*, 2018.2

[21] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer.GANimation:从单一图像中获得解剖学意义上的面部动画。*CoRR*, abs/1807.09251, 2018.1

[22] Kevin Roose和Paul Mozur。扎克伯格因缅甸暴力事件被骂。这是他的道歉，2018年。1

[23] Andreas Ro¨ssler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner.Faceforensics:*arXiv preprint arXiv:1803.09179*, 2018.2, 3, 4

[24] Andreas Ro¨ssler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner.Faceforensics++:*arXiv preprint arXiv:1901.08971*, 2019.2, 7

[25] Bernhard Scho¨lkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson.估计高维分布的支持度。*Neural Compu- tation*，13（7）：1443-1471。2, 4

[26] Scott Shane和Mark Mazzetti。俄罗斯影响美国选民的3年活动内幕，2018年。1

[27] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman.合成奥巴马：从音频中学习唇部同步。*ACM Transactions on Graphics*，36（4）：95，2017。1

[28] Amanda Taub 和 Max Fisher。国家是火柴盒，而Facebook是匹配的地方，2018年。1

[29] J.Thies, M. Zollho¨fer, M. Stamminger, C. Theobalt, and M. Nießner.Headon。人类肖像视频的实时重演。*ACM Transactions on Graphics*, 2018.1

[30] George Williams, Graham Taylor, Kirill Smolskiy, and Chris Bregler.用于多模态身份验证的身体运动分析。*在第20届国际模式识别会议上*，第2198-2201页。IEEE, 2010.2

[31] 杨欣，李岳尊，和柳思伟。利用不一致的头部姿势暴露深度假象。在*IEEE国际声学、语音和信号处理会议上*，布里斯托尔。
联合王国，2019年。1

[32] Ning Yu, Larry Davis, and Mario Fritz.将假图像归于GANs。分析生成图像中的指纹。*CoRR*, abs/1811.08180, 2018.2