单位代码 10006
学　　号 17373190
分 类 号 TP312

北京航空航天大學
BEIHANG UNIVERSITY

毕业设计 (论文)

面向特定人物的细粒度建模和深度伪造检测方法

学 院 名 称　计算机学院
专 业 名 称　计算机科学与技术专业
学 生 姓 名　朱正阳
指 导 教 师　阮利

2022 年 04 月

目 录

1 绪论

1.1 研究背景

1.1.1 深度伪造检测

近年来，随着深度学习技术的迅猛发展，深度学习为计算机视觉、自然语言处理和语音识别、无人驾驶等领域的创新提供了有力的技术支撑。“深度伪造”是指利用深度学习算法实现音频和视频的模拟和伪造，包括语音模拟、换脸、表情操纵等。

然而，深度伪造技术的不当使用也造成了诸多不良影响。例如，利用语音伪造进行电信诈骗；将一些知名歌星、影星等公众人物的脸拼接到色情明星身上，伪造色情片非法牟利，或者是伪造恶搞视频，严重侵犯公众人物的肖像权、名誉权；借助互联网炮制并传播虚假新闻与政治谣言，操纵舆论，对社会与国家安全造成严重威胁。

针对深度伪造带来的各种社会问题，我国出台了相关监管办法。2019 年 11 月，网信办发布《网络音视频信息服务管理规定》[1]，指出自 2020 年 1 月 1 日起，AI 造假视频不得随意发布。规定第十一条提及“网络音视频信息服务提供者和网络音视频信息服务使用者不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息”，第十二条提及“网络音视频信息服务提供者应当加强对网络音视频信息服务使用者发布的音视频信息的管理，部署应用违法违规音视频以及非真实音视频鉴别技术”，深度伪造检测技术的重要性可见一斑。

为防范深度伪造技术带来的诸多社会危害，发展深度伪造检测技术，开展主动技术防范与检测工作是必要的。

1.1.2 面向特定人物细粒度建模对深度伪造检测技术的意义

细粒度图像分类是在区分出基本类别的基础上，进行更精细的子类划分，如区分鸟的种类、车的款式、狗的品种等，目前在工业界和实际生活中有着广泛的业务需求和应用场景。

以特朗普这一公众人物为例，可以通过对特朗普相关的视频及音频进行特征提取，进行特朗普人物多模态细粒度建模。

对人物的多模态细粒度建模利用更多的信息来源、更准确的特征库，可以做出更优

[1]https://www.12371.cn/2020/09/14/ARTI1600059089203553.shtml

的决策，这样在针对疑似特朗普造假视频后能够快速识别视频中的可疑点并判断视频的真伪，及时制止虚假信息的传播。

1.2 相关研究进展及问题分析

近年出现的视觉深度伪造技术主要有换脸、表情迁移和动作迁移等方式，对“眼见为实”的传统观念产生巨大挑战，造成数字信任危机。

深度伪造及其检测技术的发展极度依赖数据集的构建，规模大、覆盖广、质量好的数据集对深度伪造检测技术的发展起到积极促进作用。除去 FaceForensics++q[1] 以及 DFDC[2] 等经典深度伪造数据集，He 等人[3] 构建了一个目前维度最大最丰富的深度伪造公开数据集和评测基准 ForgeryNet，数据规模方面拥有 290 万张图像及 221247 个视频，评测方法上，提出了伪造图片分类、伪造空间定位、伪造视频分类、伪造时序定位四种不同的图片级和视频级评测方法。

为应对深度伪造视频带来的挑战，当前深度伪造视频检测方法主要有基于跨视频帧组时序特征的检测方法，以及基于视频帧内视觉伪像的检测方法。

基于跨视频帧组时序特征的检测，由于深度伪造模型多以静态面部图像作为训练数据，难以实现对眨眼、呼吸、心跳等生理信息的伪造，Li 等人[4] 提出一种基于眨眼检测的深度伪造视频检测方法：首先在视频帧层面提取出面部区域和眼睛区域，其次通过人脸对齐、提取和缩放眼睛区域标点的边界框等操作创建新的帧序列，并分配至长期循环卷积网络 LRCN[5]，预测眨眼行为。该模型的不足在于仅仅对是否眨眼进行检测，而不考虑眨眼频率的合理性。Güera 等人[6] 证明深度伪造视频帧内和帧之间时序具有不一致的特性, 进而基于 CNN 和 LSTM 提出了一种时间感知管道方法来检测深度伪造视频。

基于视频帧内视觉伪像的检测，首先通过探索视频帧内视觉伪像并提取判别特征，其次将特征分配至分类器中进行训练，最终实现对视频真伪性的判断。由于伪造时经常需要经过旋转、缩放、剪切等人脸仿射变换方法，易致使合成视频伪造部位与原视频背景之间分辨率不一致，Li 等人[7] 基于 CNN 模型提出一种基于面部变形后帧内视觉伪像特征的检测模型。Nguyen 等人[7] 提出了一种基于胶囊网络的视觉伪造检测方法，Sabour 等人[8] 证明了胶囊网络能够准确描述对象部件之间的层次关系，胶囊网络通过动态路由算法, 以胶囊作为基本的训练单元，在多次迭代后将三个胶囊的输出路由到对应的输出胶囊, 进而分离伪造图像和真实图像。

尽管目前已经存在大量深度伪造视频检测研究，这些研究也存在以下共性的问题：

- 研究所使用的训练数据集一般为具有多样性的数据集，但在有一定泛化性的情况

下，在对于特定人物的检测上精度不够。

- 现有研究对于音频特征的利用不足，通过在音频及视频所表现出的特征一致性上的不足判断疑似伪造点的研究较少。

1.3 研究目标与研究内容

本文针对面向特定人物的深度伪造检测研究较少，且深度伪造视频难以同时伪造视频图像信息及其匹配的音频信息而导致音视频特征不一致的问题，设计并实现一个面向特定人物的细粒度建模及深度伪造检测方法。

1.3.1 特定人物细粒度建模

针对面向特定人物的伪造视频面临可用于训练的样本少等特点，难以使用传统的基于伪造样本进行模型训练的检测方法的挑战，本论文收集整理特定人物的图像数据集以及音频数据集，并在图像与音频两个模态上对人物进行细粒度建模。

针对人物细粒度建模的问题，首先利用预训练好的 CNN 模型对特定人物的图像训练数据集和音频数据集分别进行特征提取，其次对提取出的特征进行基于密度聚集[9] 的特征域划分，实现人物建模。

1.3.2 音视频情感一致性判别算法

针对深度伪造视频的音频特征未被合理利用的问题，本文从音画情感一致性的角度出发，设计一个基于音画情感一致性的深度伪造视频判定算法。

针对提取出的人脸图像特征以及声纹特征，采用基于梅尔频率倒谱系数 MFCC[10] 与 CNN 的音频情绪识别算法[11] 进行音频模态上的情感识别，并根据识别结果对视频进行一定的时间片分割，对每个时间片上的视频图像数据进行抽帧，并采用基于 CNN 的微表情情感识别算法[12] 进行图像模态上的情感识别。最后采用情感一致性判别算法，找出情绪不一致的可疑伪造点。

1.3.3 面向特定人物细粒度建模的深度伪造检测方法

利用面向特定人物的细粒度建模以及音画情感一致性判别算法设计并实现一个深度伪造

1.4 主要贡献和创新性

本文通过对特定人物人脸特征以及声纹特征的提取和处理，设计了一套面向特定人物细粒度建模和基于音画情绪一致性判别的深度伪造视频检测方法。本文的主要贡献如下：

- 构建特定人物视频及对应音频的数据集。
- 提取图像及音频两个模态上的特征并处理，实现面向特定人物的细粒度建模。
- 针对提取出的特定人物人脸特征和声纹特征，设计一种基于 MFCC 和 CNN 的音画情感一致性判别算法。
- 根据特定人物细粒度建模与音画情感一致性判别算法，设计并实现一个完整的深度伪造视频检测系统。

本文的主要创新如下：

- 提出了一种基于密度聚类的特征域划分算法，包含对密度函数的定义，样本空间中密度的计算以及临界密度的确定等操作，对特定人物图像及音频数据中提取出的特征进行特征划分，建立该特定人物的特征域。
- 提出了一种基于 MFCC 和 CNN 的音画情感一致性判别算法，并通过情感一致性的判断来寻找视频中的疑似伪造点，进行深度伪造视频检测。

2 研究成果

2.1 整体设计

本文提出的面向特定人物的细粒度建模和深度伪造检测方法整体设计如下图所示：

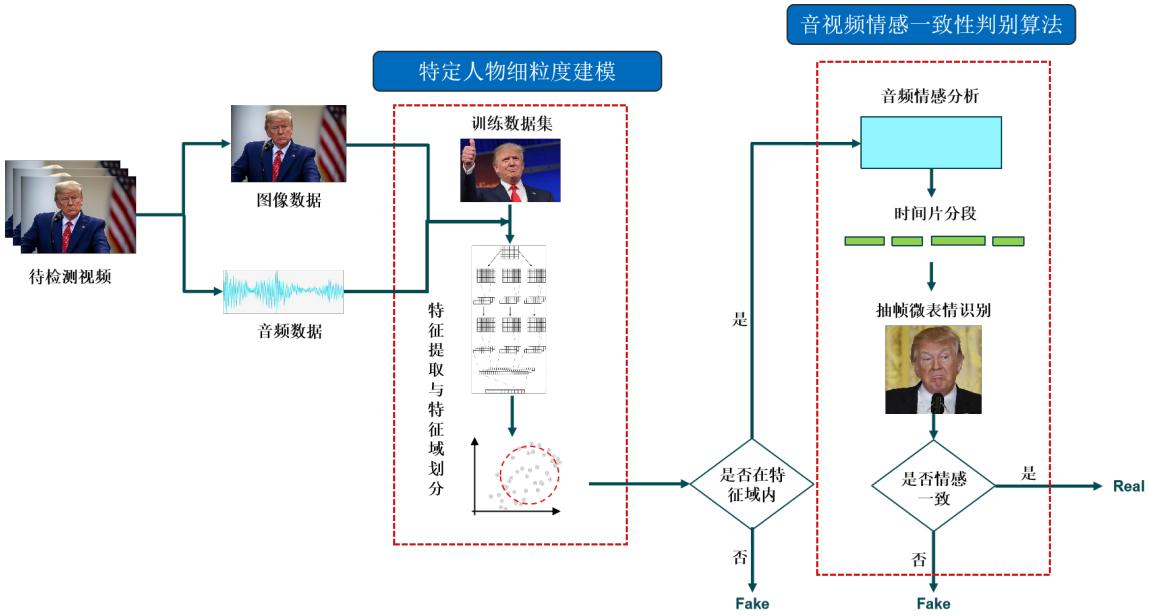

图 2.1　面向特定人物的细粒度建模和深度伪造检测方法整体设计

从左至右可以分为四个部分：

1. 第一部分为输入，输入为该特定人物的待检测视频数据。需要对输入进行视频和音频的分离，分别得到对应的视频数据和音频数据。

2. 第二部分为面向特定人物的细粒度建模。该部分首先需要用特定人物的真实准确的训练数据集进行细粒度建模，具体为利用预训练好的 CNN 模型分别对人脸图像数据集和音频数据集提取人脸特征与声纹特征，再对特征进行基于密度聚类的特征划分，找到该特定人物的特征所在区域，实现建模。

3. 第三部分为特征判断。首先对第一部分的输入进行特征提取，其次利用第二部分所提前得到的特定人物特征区域，判断输入数据提取出的特征是否在特定人物特征区域内。若不在特征区域内，则说明输入数据的特征不是该特定人物的特征，疑似存在伪造。若在特征区域内，则身份验证成功，进行下一步检测。

4. 第四部分为音画情感一致性判别。对上一部分提取出的特征进行情感分析，并对比由音频和图像特征分析出的情感，若情感基本一致则为 Real，情感不一致则为 Fake，判断之后输出结果。

2.2 特定人物细粒度建模

2.2.1 特征提取

采用预训练好的 CNN 模型对特定人物的人脸图像和音频数据进行特征提取，提取出的特征结构为向量，在样本空间中可以表示为一特征点。

2.2.2 基于密度聚类的特征划分算法

针对提取出的特定人物的特征与其他人的特征以及噪声的边界不确定的问题，本节提出了一种基于密度聚类的特征划分算法。

算法 1 特征点特征划分算法

输入: 特征点数组 $data$，特征点数量 n，影响值函数参数 σ
输出: 密度边界 ρ_0

ρ_list = [ ]
for i *in range*$(0, n-1)$ **do**
  $\rho = 0$
  for j *inrange*$(0, n-1)$ **do**
    $d = dist(data[i], data[j])$
    $\rho = \rho + exp\{-\frac{d^2}{2\sigma^2}\}$
  end for
  $\rho\_list[i] = \rho$
end for
sort(ρ_list)
$\rho_0 = \rho\_list[\lfloor 0.05n \rfloor]$
return ρ_0

算法首先设置了影响函数

$$f = e^{-\frac{x^2}{2\sigma^2}}$$

该影响函数所表示的是特征点对空间中某一点的影响，或者看作在这一点的权重，其中 x 表示特征点到空间中该点的距离，σ 为参数，σ 的大小影响 f 随 x 的增大的衰减程度，σ 的值越小，衰减速度越快。

定义空间中某点的密度 ρ 为样本空间中所有特征点在此处的影响的叠加。ρ 衡量的是样本空间中特征点的分布紧密性。该算法的目的为找到一个临界密度 ρ_0，若空间中某

点密度大 ρ_0，则表示该点附近的特征点分布较为密集，视为符合该特定人物的特征；否则该点即视为在其他人特征点的分布区域或噪声区域。

对于 σ 的值的确定，若 σ 过小，f 衰减过快会导致特征点之间的空间的密度仍旧很小，无法反映附近的特征点的密度；若 σ 过大，f 衰减过慢会导致特征点稀疏的空间的密度与密集的空间的密度的差距过小，进而导致特征边界划分误差过大。设置合理的 σ 值需要引入一定数量的非特征点或噪声来对影响函数的划分进行测试。

该算法的优点在于能够进行进行任意形状的划分而非仅限于凸函数，且对噪声数据不敏感。

2.3 音画情感一致性算法

输入包含音频数据的视频数据，首先对音频数据进行情感分析，得到分析结果后，将时间轴按照情感的不同进行划分，对每个音频情绪相同的时间段内进行一定数量的抽帧操作，并通过检测抽取图像的人物微表情检测并分析情感，与音频分析出的情感逐一对比，判断情感一致性。

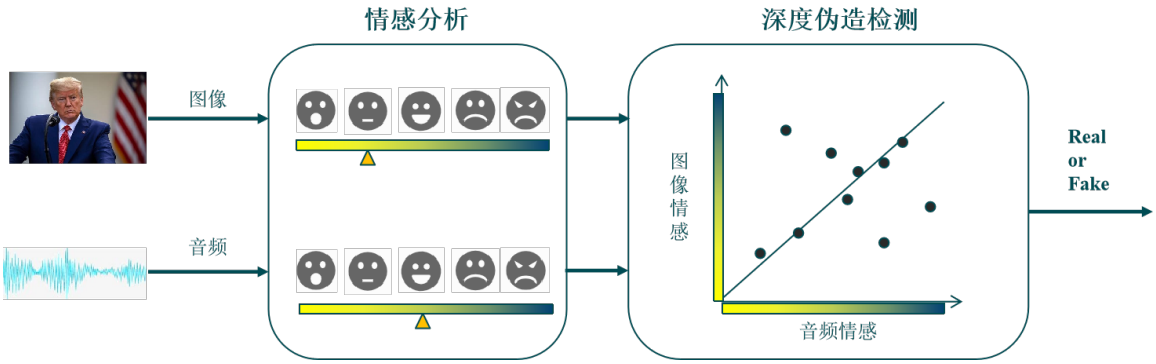

图 2.2 基于情感识别的音画一致性判别流程

3 未来工作计划

3.1 未来工作内容

目前未完成的工作如下：

- 测试数据集的构建。
- 完善基于密度聚集的特征域划分算法的实验测试。
- 设计音画情感一致性判别算法，设计相应实验测试对应准确率及性能。
- 实现基于特定人物细粒度建模以及音画情感一致性判别算法的深度伪造视频检测方法，并设计实验测试方法的效能。

3.2 未来工作计划

未来的任务计划安排如下所示：

- 第 7 周：构建测试数据集。
- 第 8 周：完成特征提取以及基于密度聚集的特征域划分算法的实验测试。
- 第 9-10 周：复现情感识别模型，完成音画情感一致性判别算法的细节及代码实现，并设计实验测试。
- 第 11 周：实现原型系统，并设计实验测试基于特定人物细粒度建模以及音画情感一致性判别算法的深度伪造视频检测方法的效能。
- 第 12 周及以后：完善各模块的实验，优化各算法，完成毕设论文的写作。

参 考 文 献

[1] Rossler A., Cozzolino D., Verdoliva L., et al. Faceforensics++: Learning to detect manipulated facial images[A]. Proceedings of the IEEE/CVF International Conference on Computer Vision[C]. .[S.l.]: [s.n.] , 2019:1–11.

[2] Dolhansky B., Howes R., Pflaum B., et al. The deepfake detection challenge (dfdc) preview dataset[J]. arXiv preprint arXiv:1910.08854, 2019.

[3] He Y., Gan B., Chen S., et al. Forgerynet: A versatile benchmark for comprehensive forgery analysis[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. .[S.l.]: [s.n.] , 2021:4360–4369.

[4] Li H., Li B., Tan S., et al. Detection of deep network generated images using disparities in color components. arXiv 2018[J]. arXiv preprint arXiv:1808.07276.

[5] LIY C. M., InIctuOculi L. ExposingAICreated FakeVideosbyDetectingEyeBlinking[A]. 2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE[C]. .[S.l.]: [s.n.] , 2018.

[6] Güera D., Delp E. J. Deepfake video detection using recurrent neural networks[A]. 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)[C]. .[S.l.]: [s.n.] , 2018:1–6.

[7] Li Y., Lyu S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.

[8] Sabour S., Frosst N., Hinton G. E. Dynamic routing between capsules[J]. Advances in neural information processing systems, 2017, 30.

[9] Hinneburg A., Gabriel H.-H. Denclue 2.0: Fast clustering based on kernel density estimation[A]. International symposium on intelligent data analysis[C]. .[S.l.]: [s.n.] , 2007:70–80.

[10] Muda L., Begam M., Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. arXiv preprint arXiv:1003.4083, 2010.

[11] Torfi A., Dawson J., Nasrabadi N. M. Text-independent speaker verification using 3d convolutional neural networks[A]. 2018 IEEE International Conference on Multimedia

and Expo (ICME)[C]. .[S.l.]: [s.n.] , 2018:1–6.

[12] Hosler B., Salvi D., Murray A., et al. Do Deepfakes Feel Emotions? A Semantic Approach to Detecting Deepfakes via Emotional Inconsistencies[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. .[S.l.]: [s.n.] , 2021:1013–1022.