多媒体伪造检测十年：全面回顾

Tao Luan, Joan P. Lazaro*

Graduate School, University of the East, 2219 Recto Ave, Sampaloc, Manila, Metro Manila, Philippines

【摘要】数字媒体的迅速普及和易于操纵，需要强大的伪造检测技术来保持多媒体的可信度。这篇评论文章全面概述了过去十年伪造检测技术的进步，重点介绍了传统方法、基于机器学习的方法和基于深度学习的方法。传统技术涉及水印、签名和统计属性分析，而基于机器学习的方法则采用监督学习进行自动伪造分类。基于深度学习的方法利用卷积神经网络（CNN）从原始像素数据中学习分层特征，在检测高级操纵方面表现出色。尽管取得了这些进步，但挑战依然存在，包括标记数据的可用性有限、对抗性攻击、跨不同伪造技术的泛化以及实时检测。应对这些挑战对于提高数字媒体的可信度和维护数字景观的完整性至关重要。这篇评论文章旨在全面了解多媒体伪造检测的现状，并启发未来的研究方向以应对剩余的挑战。

【关键词】伪造检测；多媒体；机器学习；深度学习；数字媒体

【收稿日期】2024 年 10 月 22 日 【出刊日期】2024 年 11 月 20 日 【DOI】10.12208/j.aiml.20240001

A Decade of Multiple-media Forgery Detection: A Comprehensive Review

*Tao Luan, Joan P. Lazaro**

Graduate School, University of the East, 2219 Recto Ave, Sampaloc, Manila, Metro Manila, Philippines

【**Abstract**】The rapid proliferation of digital media and ease of manipulation necessitate robust forgery detection techniques to maintain multimedia trustworthiness. This review paper offers a comprehensive overview of the advancements in forgery detection techniques over the past decade, focusing on traditional, machine learning-based, and deep learning-based approaches. Traditional techniques involve watermarking, signatures, and statistical property analysis, while machine learning-based methods employ supervised learning for automatic forgery classification. Deep learning-based methods utilize convolutional neural networks (CNNs) to learn hierarchical features from raw pixel data, demonstrating exceptional performance in detecting advanced manipulations. Despite these advancements, challenges persist, including limited availability of labeled data, adversarial attacks, generalization across different forgery techniques, and real-time detection. Addressing these challenges is crucial for enhancing the trustworthiness of digital media and preserving the integrity of the digital landscape. This review paper aims to provide a thorough understanding of the current state of multiple-media forgery detection and inspire future research directions to tackle remaining challenges.

【**Keywords**】Forgery Detection; Multiple-media; Machine learning; Deep learning; Digital Media

介绍

数字媒体的普及极大地改变了人们交流、社交和获取信息的方式。随着技术的飞速进步，对数字内容的篡改也变得越来越复杂，使得确定多媒体文件的真实性变得愈加困难。然而，一些恶意篡改给社会带来了许多影响，例如奥巴马的虚假演讲：2018 年，一种名为“Deepfake”的基于人工智能的视频篡改工具被用来制作美国前总统奥巴马的虚假视频。视频中奥巴马发表了他从未发表过的演讲，引发了人们对深度伪造技术可能被滥用来传播虚假信息的担忧。另一个是佩洛西的篡改视频：2019 年，美国众议院议长佩洛西的一段视频被篡改，使其在一场

*通讯作者：Joan P. Lazaro

注：本文于 2023 年发表在 Advances in Computer and Communications 期刊 4 卷 3 期，为其授权翻译版本。

公开活动中出现醉酒和含糊不清的表情。这段被篡改的视频在社交媒体上迅速传播，说明了被篡改的视频可以轻易地被分享，并可能影响舆论。因此，对高效伪造检测技术的需求变得至关重要。这篇评论文章重点介绍了过去十年多媒体伪造检测领域取得的进展，重点介绍了最重要的进步和挑战。它旨在提供对该领域的全面了解并启发未来的研究方向。

近年来，数字取证技术的兴起在打击多媒体伪造方面发挥了至关重要的作用。例如，图像伪造检测的一个重要进步是使用深度学习算法。这些算法在大量数据集上进行训练，可以自动学习被篡改图像的显著特征。一个显著的例子是卷积神经网络（CNN）的发展，它可以通过分析像素级细节和不一致性来准确检测图像中的篡改区域。通过识别不匹配的颜色或类似故障的伪影等异常，这些先进的算法极大地提高了数字媒体的真实性和可信度。

多媒体伪造检测的挑战依然存在，特别是在检测被称为深度伪造的复杂操纵方面。深度伪造涉及使用人工智能技术创建高度逼真但伪造的视频或图像。这些被操纵的媒体甚至可以欺骗训练有素的专业人士，对公众人物、声誉管理甚至国家安全构成威胁。因此，打击深度伪造需要不断研究和创新新技术来识别和验证媒体内容。

1 多媒体伪造检测技术的演变

过去十年，多媒体伪造检测技术取得了显著进展。早期方法主要依赖人工分析，既耗时又容易出错。机器学习和计算机视觉算法的引入彻底改变了这一领域，使自动检测伪造品成为可能，而且准确率和效率都有所提高。

1.1 传统技术

传统的伪造检测技术大致可分为主动和被动方法。主动技术涉及将水印、签名或其他形式的隐藏信息嵌入原始内容中，之后可以验证其真实性[1]而被动技术则依赖于数字内容的固有特性，如噪声模式、压缩伪影和统计特性[2]这些技术已被广泛应用于检测图像和视频伪造，如复制、移动、拼接和篡改[3]。

1.2 基于机器学习的技术

随着机器学习算法的出现，伪造检测技术变得更加复杂。这些技术采用监督学习，在标记的数据集上训练模型，使其能够对新的、未见过的实例进行分类[4]。基于机器学习的技术在检测各种形式的伪造方面表现出良好的前景，例如图像拼接、视频操纵和深度伪造生成[5]。

1.3 基于深度学习的技术

深度学习和卷积神经网络（CNN）的出现进一步改变了伪造检测领域。CNN 在图像和视频分析任务中表现出色[6]。这些模型可以从原始像素数据中自动学习分层特征，使其特别适合伪造检测任务[2]基于深度学习的技术已成功应用于检测图像和视频伪造，包括深度伪造、生成对抗网络（GAN）和其他先进的操纵技术[7]。

深度学习技术还促进了深度伪造视频的检测，深度伪造视频是使用人工智能算法创建的高度逼真的合成视频。通过分析面部动作和表情以及视频中视觉元素的整体一致性，CNN 可以识别出表明存在深度伪造操作的异常。这些模型可以有效地检测出光照、阴影、像素级不一致和其他可能表明篡改的细微线索。

1.4 图像取证技术的进步

图像取证技术的进步在改进多媒体伪造检测技术方面发挥了重要作用。一个重要的进步领域是检测复制移动伪造，即将图像的一部分复制并粘贴到另一个区域。最初，复制移动检测依赖于手动方法，例如目视检查或搜索重复模式。然而，随着伪造技术的复杂性增加，开发了自动算法来准确识别重复区域，即使它们被调整大小、旋转或被噪声覆盖[8]。这些算法利用基于块的比较、关键点匹配和特征提取等技术来定位和分析图像中的重复区域[9]。

图像取证的另一个进步领域是检测图像中的篡改或拼接。传统技术依赖于检测像素级属性中的异常或检查光照、颜色或纹理中的不一致。最近，基于深度学习的方法来分析像素级细节并学习被操纵图像的显著特征。这些技术在识别图像中的细微变化和操纵方面表现出了惊人的准确性[10]。例如，在大量真实图像和被操纵图像数据集上训练的深度学习模型可以根据像素值、噪声模式或其他视觉伪影的差异准确检测被篡改的区域[11]。

1.5 视频取证进展

除了图像取证，视频取证领域也取得了重大进展。由于连续帧的性质以及时间操纵的可能性，视频伪造检测带来了独特的挑战。传统的视频取证技术涉及分析帧级属性（例如运动矢量、帧速率或压缩伪影）以检测不一致性。然而，这些方法在检测微

妙或复杂的操纵方面的能力有限。

深度学习的出现增强了视频取证能力。基于循环神经网络（RNN）和长短期记忆（LSTM）网络的复杂算法已被开发出来，用于分析时间依赖性并准确识别视频篡改。这些模型可以检测各种视频伪造技术，包括删除帧、插入帧和更改帧速率。此外，基于深度学习的技术已用于视频伪造定位，其目标是识别被篡改的特定帧或片段。

2 多媒体伪造检测的挑战和未来方向

近年来，多媒体伪造检测领域取得了许多进展。然而，仍然存在一些挑战，未来研究人员可以研究几个方向来提高伪造检测技术的有效性和效率。本节讨论其中一些挑战和潜在的未来方向。

（1）伪造的语境理解伪造检测的主要挑战之一是理解伪造发生的语境。传统的伪造检测技术通常侧重于检测特定类型的操作，例如复制移动或拼接。然而，伪造可能更加复杂，可能涉及多种操作或不同技术的组合。因此，未来的研究应致力于开发能够分析图像或视频的整体语境一致性的技术，以更有效地检测复杂的伪造。

正在努力鼓励研究人员和组织之间共享和协作数据集。这有助于解决标记数据稀缺的问题，并确保伪造检测领域能够共同进步。开放式挑战和竞赛（例如由学术机构和行业领导者组织的挑战和竞赛）也促进了使用标准化数据集开发和评估伪造检测模型。这些举措有助于创建基准数据集，可作为评估不同检测技术性能的参考。

（2）深度伪造和人工智能生成内容的检测随着深度学习技术的进步，制作逼真且令人信服的深度伪造已成为一个重大问题。深度伪造是指人工生成的媒体，例如看似真实但实际上是使用人工智能算法操纵或合成的图像或视频。开发针对深度伪造和其他人工智能生成内容的强大检测技术对于打击虚假信息和恶意活动的传播至关重要。未来的研究应侧重于探索新方法，例如多模态分析和基于深度学习的算法，以准确识别深度伪造。

（3）推广至各种媒体类型虽然在检测图像和视频伪造方面已经取得了重大进展，但将这些技术扩展到音频和文本等其他形式的媒体，则带来了新的挑战。检测音频伪造（如语音变形或音频操纵）需要专门的算法来分析音频内容的声学特性和模式。同样，检测文本伪造（如文档篡改或抄袭）需要开发能够有效识别被操纵或伪造的文本文档的文本分析技术。未来的研究应致力于将伪造检测技术推广到不同的媒体类型，以确保全面的媒体取证。

（4）实时和可扩展的解决方案随着在线制作和共享的媒体内容量不断增加，对实时和可扩展的伪造检测解决方案的需求变得至关重要。传统技术通常涉及计算量大的过程，可能不适合实时应用。未来的研究应侧重于开发能够实时处理大量数据的高效算法，以便及时检测和缓解伪造行为。

水印和数字签名等传统技术通常用于伪造检测，但它们可能不足以实时处理日益增长的数据量。这些技术通常需要计算量大的过程，这会减慢检测和验证任务的速度。

通过结合这些方法，我们可以开发实时且可扩展的伪造检测解决方案，有效处理日益增长的在线媒体内容。这些解决方案将有助于确保在各种数字平台上共享的信息的可靠性和完整性。

（5）对抗性攻击和对策随着伪造检测技术的进步，攻击者用来逃避检测的技术也在不断进步。对抗性攻击涉及操纵媒体内容，使其能够欺骗伪造检测算法。开发针对对抗性攻击的强大对策对于确保伪造检测技术的可靠性和有效性至关重要。未来的研究应探索检测和减轻对抗性攻击的方法，包括对抗性训练和使用生成模型来创建更具弹性的伪造检测算法。

（6）隐私和道德问题随着伪造检测技术越来越强大，人们对隐私和道德问题的担忧也与日俱增。许多伪造检测技术依赖于对个人媒体内容的分析，例如个人拍摄的图像或视频。因此，解决隐私问题并开发能够确保个人隐私信息得到保护并防止伪造检测技术被滥用的技术至关重要。

3 结论

多媒体伪造检测技术取得了重大进展，但挑战仍然存在。研究人员可以专注于几个未来方向，例如提高语境理解、检测深度伪造、推广到各种媒体类型、开发实时和可扩展的解决方案、应对对抗性攻击以及解决隐私和道德问题。通过应对这些挑战并追求这些方向，多媒体伪造检测领域可以不断发展并增强其打击数字操纵和确保媒体内容真实性的能力。在过去十年中，由于机器学习和深度学习算法的引入，多媒体伪造检测技术取得了重大进展。尽管取得了这些进展，但仍存在一些挑战，包括标

记数据的可用性有限、对抗性攻击、泛化和实时检测。随着技术的不断发展，研究人员必须应对这些挑战，以确保伪造检测系统的持续有效性并维护数字媒体的可信度。

致谢

作者要感谢过去十年来为多媒体伪造检测技术的发展做出贡献的众多研究人员和机构。他们的奉献和创新为该领域的现状奠定了基础，并将继续激励未来的发展。

参考文献

[1] Farid, H. (2009). Digital image forensics. IEEE Signal Processing Magazine, 26(2), 26-37.

[2] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. IEEE Transactions on Information Forensics and Security, 13(5), 1211-1226.

[3] Bayram, S., Sencar, H. T., & Memon, N. (2010). An efficient and robust method for detecting copy-move forgery. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 1053-1056. IEEE.

[4] Zhang, J., Han, B., Lin, Z., & Yang, M. (2018). Image manipulation detection using semi-supervised learning. In Proceedings of the European Conference on Computer Vision (ECCV), 802-816.

[5] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1-11.

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[7] Li, Y., Chang, M. C., & Lyu, S. (2020). In ictu oculi: Exposing AI-generated fake face videos by detecting eye blinking. IEEE Transactions on Information Forensics and Security, 15, 2316-2325.

[8] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2307-2311. IEEE.

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[10] Agarwal, S., Farid, H., Gu, Y., & He, M. (2021). Protecting world leaders against deep fakes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2052-2061.

[11] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2021). Learning rich features for image manipulation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1054-1063.