2024 年度年次報告書
信頼される AI システムを支える基盤技術
2020 年度採択研究代表者

越前 功

国立情報学研究所 情報社会相関研究系
教授

インフォデミックを克服するソーシャル情報基盤技術

主たる共同研究者:
笹原 和俊（東京科学大学 環境・社会理工学院 教授）
馬場口 登（大阪大学 D3 センター 特任教授）

研究成果の概要

本研究課題は，AI により生成されたフェイク映像，フェイク音声，フェイク文書などの多様なモダリティによるフェイクメディア（FM）を用いた高度な攻撃を検出・防御する一方で，信頼性の高い多様なメディアを積極的に取り込むことで人間の意思決定や合意形成を促し，サイバー空間における人間の免疫力を高めるソーシャル情報基盤技術を確立することを目的とする．具体的には，(1)多様なモダリティによる高度な FM 生成技術，(2)FM 検出・防御技術，(3)FM 無毒化技術，(4)インフォデミックを緩和し多様な意思決定を支援する情報技術の確立を目標としている．

2024 年度の主だった成果は以下の通りである．(1)多様なモダリティによる高度な FM 生成技術では，プロパガンダ型 FM 生成に向けたデータセットを構築し，10 個のコミュニケーション技術を特定した上で，その検出手法を提案するとともに，プロパガンダ型 FM 生成のための基礎的検討を始めた． (2) FM 検出・防御技術では，自己教師ありビジョントランスフォーマーのディープフェイク検出への活用を検討し，ViTがディープフェイク検出の特徴検出器として有効であることを示した．また，フェイク顔映像検出プログラム（SYNTHETIQ VISION）の社会実装を進めた．(3)FM 無毒化技術では，顔映像に復元情報を知覚できないように混入することで，Deepfake による顔の置き換えを経てもオリジナルの顔映像を高精度で復元する手法を確立した．(4)インフォデミックを緩和し多様な意思決定を支援する情報技術の確立に向けてでは，偽情報の拡散に Bot が動員されている実態や YouTube が偽情報の主戦場になっている状況を明らかにした。さらに，ディープフェイク検出技術の有効性を検証する XFinch 実験の準備を進め，予備実験を行った．

【代表的な原著論文情報】

1) H. Liu, Y. Nakashima, and N. Babaguchi, “Paladin: Understanding Video Intentions in Political Advertisement Videos,” 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 8239-8248, 2025
2) H. H. Nguyen, J. Yamagishi and I. Echizen, “Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis,” 2024 IEEE International Joint Conference on Biometrics (IJCB), Buffalo, NY, USA, 2024, pp. 1-10, doi: 10.1109/IJCB62174.2024.10744497.
3) Y. Furuhashi, J. Yamagishi, X. Wang, H. H. Nguyen and I. Echizen, “Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection,” 2024 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2024, pp. 1-5, doi: 10.1109/BIOSIG61931.2024.10786748.
4) W. Xu, K. Sasahara, J. Chu, B. Wang, W. Fan, and Z. Hu, “Social Media Warfare: Investigating Human-Bot Engagement in English, Japanese and German during the Russo-Ukrainian War on Twitter and Reddit,” EPJ Data Science 14(10), 2025
5) K. Miyazaki, T. Uchiba, H. Kwak, J. An, and K. Sasahara, “The Influence of Toxic Trolling Comments on Anti-vaccine YouTube Videos,” Scientific Reports 14 (5088), 2024