信頼される AI システムを支える基盤技術　　2022 年度
2020 年度採択研究代表者　　年次報告書

越前 功

情報・システム研究機構 国立情報学研究所
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（東京工業大学 環境・社会理工学院 准教授）
馬場口 登（大阪大学 データビリティフロンティア機構 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．

2022 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，動画中の歩行者領域を架空の人物像 FM へと違和感なく置換することにより歩容の個人性を匿名化する手法を提案した．また，プロパガンダ型 FM 生成に向けたデータセット構築を実施し，1 万点以上の映像に対するアノテーションが得られた．(2) FM 検出・防御技術では，フェイク顔映像検出のための WebAPI からなるプログラム群（SYNTHETIQ VISION）を開発し，企業による実利用が決定した．また，フェイクメディア検知に有用なアテンションマップを利用した説明可能な AI 技術を検討した．(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検討した．(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では，インフォデミックにおいて偽情報やヘイトの拡散が社会的分断へ与える影響の調査や，SNS ユーザが誤情報を自発的にファクトチェックする行動の特徴を計算社会科学の手法を用いて解明し，それらの知見をプラットフォーム技術に応用するための検討を行った．

【代表的な原著論文情報】

1) P. Ghasiya and K. Sasahara, Rapid Sharing of Islamophobic Hate on Facebook: The Case of the Tablighi Jamaat Controversy, Social Media + Society 8(4), 2022

2) K. Miyazaki, T. Uchiba, K. Tanaka, J. An, H. Kwak, and K. Sasahara, “This is Fake News”: Characterizing the Spontaneous Debunking from Twitter Users to COVID-19 False Information, The 17th International AAAI Conference on Web and Social Media (ICWSM 2023) (accepted)

3) Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi, Anonymization of Human Gait in Video Based on Silhouette Deformation and Texture Transfer, IEEE Transactions on Information Forensics and Security, vol.17, pp.3375-3390, September 2022.

4) B. Wang, L. Li, Y. Nakashima, and H. Nagahara, Learning Bottleneck Concepts in Image Classification. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023 (accepted).

5) H. H. Nguyen, T.-N. Le, J. Yamagishi, and I. Echizen, Analysis of Master Vein Attacks on Finger Vein Recognition Systems, IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023), January 2023