研究紹介 FOCUS ON

PRESS RELEASE

岡山大学 OKAYAMA UNIVERSITY

岡山大学記者クラブ　御中

令和４年５月２７日
岡　　山　　大　　学

フェイクコンテンツの真偽判定技術 ~AIによるウソ発見器~

◆発表のポイント

・加工・編集が加えられた不正な動画を判別するための新しいフレームワークを提案しました。
・悪用される可能性がある元のコンテンツに識別情報や制御信号を忍ばせることで、利用制限を与える方法です。
・音声と映像との関連性に注目して信号処理することで、不正な加工の痕跡を解析できます。

AI技術を用いた信号処理技術の発達に伴って、マルチメディアコンテンツの加工・編集のレベルが年々向上しています。正常な映像や音声などのコンテンツを多数集めてAIシステムを学習させれば、自然に動作および発言している動画コンテンツを創り出すことが可能となっています。そのため、個人攻撃のために創造されるウソの発言や言動の動画だけでなく、社会に混乱を招くプロパガンダへの利用目的で作成されるフェイクコンテンツの流布にも繋がる危険性が指摘されています。

本研究では、社会的に重要な役割りを担う人物に関する動画を対象として、加工・編集の有無を確認するために、公開前に識別情報や制御信号を忍ばせる新しいフレームワークを考案しました。公式な場での発言に対して、口の動きと音声を精巧に加工して、特定の発言に置き換えたフェイクコンテンツに本技術を用いれば、不正な加工の存在を解析可能であります。今後はコンテンツの出処を誰でも確認できるようなフレームワークへと拡張させたいと考えています。

■発表内容

＜導入＞

AI技術を用いた画像・映像・音響処理技術によって、人工的に作成されたマルチメディアコンテンツが極めて精巧にできるようになっています。エンターテインメントにおいては、実際には撮影困難な危険を伴う映像シーンを、コンピュータグラフィックス技術に代わってAI技術により人工的に制作することも可能となっており、その活用の可能性は広がりつつあります。大量の映像シーンを収めた動画データセットを用いてAIシステムを学習させることで、本人に成り代わって、特定の動作や発言をするコンテンツを制作することができます。一方で、この技術を悪用すれば、完全にでっち上げとして動画を創造することも可能となり、誹謗中傷や名誉棄損となるコンテンツを動画配信サイトやソーシャルネットワークサービスを通して拡散されることが問題として挙がっています。

＜背景＞

アナログの時代より合成写真に代表されるウソの情報によって、プロパガンダなどの情報操作が問題となっています。このウソの情報を創造するために AI 技術が悪用されれば、悪意のある者がフェイクコンテンツを作成し、ソーシャルメディアサービスを介して発信することで広く拡散される可能性があります。従来の合成写真の場合、高度に経験を積んだ専門家が光や影の位置、不自然な境目の存在などを目視によりチェックして、確認を行っていました。図 1 に示すように、AI 技術により人工的に作成されたフェイクコンテンツの場合、人の視覚や聴覚だけでは見破ることが困難となりつつあります。このような問題に対応すべく、合成技術とは反対の方向として「人工的に創造されたコンテンツ」と「正常に撮影されたコンテンツ」を判定する技術（我々は“AI によるウソ発見器”と呼んでいます）の開発が急務となっております。

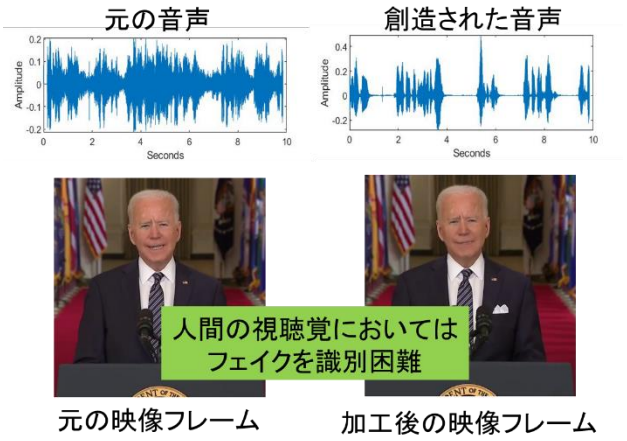

図 1. 精巧に作成されるフェイクコンテンツ

＜研究内容、業績＞

本研究プロジェクト[1]では、音声・映像データの真偽判定技術により、ウソの情報によるプロパガンダの拡散を防ぐことを目指しています。公式な記者会見などの映像は、公開前に内容の確認も含めて悪用されない処理がなされることを前提として考えます。その処理において、原本性を保証するための情報を忍ばせておき、その情報を検証することで加工・編集の有無を確認する方法を考案しています[2]。図 2 に示すように映像中の唇の動きを特徴成分として抽出し、対応する音声信号に忍ばせた情報を検証し、コンテンツ中の不自然な動きが含まれていないかを調べています。二種類の電子透かし技術[3]を組み合わせて頑強な信号と脆い信号の両方を適用することで、二段階の検証を可能としており、映像の加工・編集の有無と、音声との関係性の確認のそれぞれの用途に使い分けています。

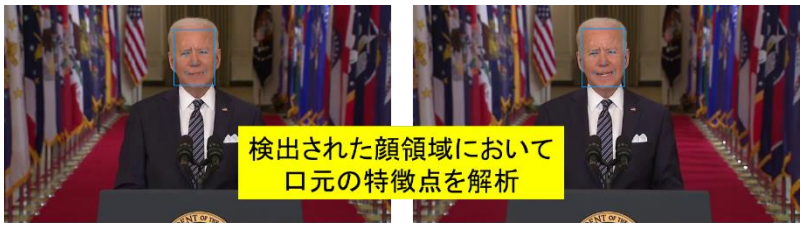

図 2. 顔検出と特徴点領域の解析による真偽判定

正式に公開されるコンテンツは、暗号技術で用いられるような電子署名を付けておけば、加工・編集の有無を確認することは可能であります。しかし、マスコミにおける編集権も考慮して、部分的に切り出した動画は正常な編集権の範囲内であることを認める技術的な解決が必要かと思われます。本手法を用いれば、切り出す開始点や終了点の選択を柔軟に認める編集権の付与を考慮することが可能です。

＜展望＞

本人の音声の一部分を切り取って、同じコンテンツ内の別の映像フレームに移植するような加工の場合、その音声は本人のものであることから個々の時間枠内のみで検証を行うだけでは真偽判定は不十分となります。今後は、音声信号と映像信号との同期にまで着目して、両方が揃っていることまで確認する手法に拡張させることを検討していく予定です。また、並行して進めている技術として、事前の対策がなされておらず受け身的に対応せざるを得ないコンテンツにおいて、加工・編集によって生じた不自然な信号成分を解析するマルチメディアセキュリティ技術にも注目して研究を進めています。

＜略歴＞

1977 年生まれ。神戸大学工学部卒、同大学院自然科学研究科博士課程中退。博士（工学）。専門はマルチメディアセキュリティ。神戸大学助手、助教を経て、2015 年より現職。

■補足・用語説明

[1] 日本ースペインーポーランド間の国際共同研究プロジェクト

JST 戦略的国際共同研究プログラム(SICORP) EIG CONCERT-Japan 「レジリエント、安全、セキュアな社会のための ICT」

"ソーシャルメディアプラットフォームにおけるフェイクニュース検出(DISSIMILAR)"

研究課題番号: JPMJSC20C3 (2021-2024)

日本側研究代表者：栗林稔 (岡山大学)

[2] A. Qureshi, D. Megias, M. Kuribayashi, "Detecting deepfake videos using digital watermarking," Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC 2021), pp.1786-1793, 2021.

[3]電子透かし技術：マルチメディアコンテンツに対して、その品質をあまり損なうことなく副情報を忍ばせることを可能とする技術です。簡単に取り除けない頑強な手法は著作権保護などへの応用、コンテンツの変化に対して脆い手法は改ざん検知などへの応用があります。

＜お問い合わせ＞
岡山大学　学術研究院自然科学学域（工）
准教授　栗林　稔
（電話番号）086-251-8249