TSUBAME 共同利用　令和 7 年度　学術利用　成果報告書

利用課題名:　合成メディアの編集履歴を説明可能に解析するフォレンジック技術
英文：Explainable forensic analysis of editing history in synthetic media

利用課題責任者　越前 功
Isao ECHIZEN

所属: 国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://scholar.google.com/citations?user=P-tAbagAAAAJ&hl=en

邦文抄録（300 字程度）

本プロジェクトでは，生成 AI 等によって作成または編集された合成メディアを対象とし，画像がどのような編集操作を受けて生成されたかという編集履歴を説明可能な形で解析するフォレンジック技術の開発を目的とした．特に，意味的な編集，色や明るさの調整，および回転や拡大縮小といった幾何的変換という，代表的な三種類の編集操作に着目し，それぞれの編集の影響が分かりやすく現れる参照用ウォータマークを画像に埋め込む手法を構築した．さらに，編集後に抽出されたウォータマークを解析することで，どのような編集がどの程度行われたかを推定する説明的解析手法を設計した．実験の結果，提案手法は画質への影響を抑えつつ，編集内容を高い精度で推定できることを確認した．

英文抄録（100 words 程度）

This project aims to develop an explainable forensic method for analysing the editing history of synthetic media. We focus on three typical types of image editing, namely semantic editing, photometric adjustment and geometric transformation. Dedicated watermark patterns are embedded into images so that the effects of different edits can be observed in an interpretable manner. By analysing the extracted watermarks after editing, the types and parameters of editing operations can be estimated. Experimental results show that the proposed method achieves high image quality and enables reliable and explainable analysis of editing histories in synthetic media.

Keywords: 合成メディア，フォレンジック，編集履歴解析，説明可能 AI，ウォータマーク

背景と目的

近年，生成 AI の発展により，画像の一部を生成・置換する編集や，色調補正，幾何的変換などを組み合わせた高度な編集が容易に行えるようになっている．その結果，合成メディアがどのような編集過程を経て作成されたかを事後的に把握することが極めて困難になっている．

従来の合成メディアフォレンジックの多くは，画像中の統計的特徴や不自然な痕跡を用いて真偽判定を行う方式であり，どのような編集が行われたかを分かりやすく説明することは必ずしも得意ではなかった．

本プロジェクトでは，合成メディアの生成・編集過程をより分かりやすく解析することを目的とし，画像と同じ編集操作を受けて変化するウォータマークをあらかじめ埋め込むことで，編集内容を説明可能な形で推定できるフォレンジック技術の構築を目指した．

編集履歴の推定を，編集後の観測結果から編集操作を逆に推定する問題として扱い，説明可能な解析を実現することを本研究の目的とする．

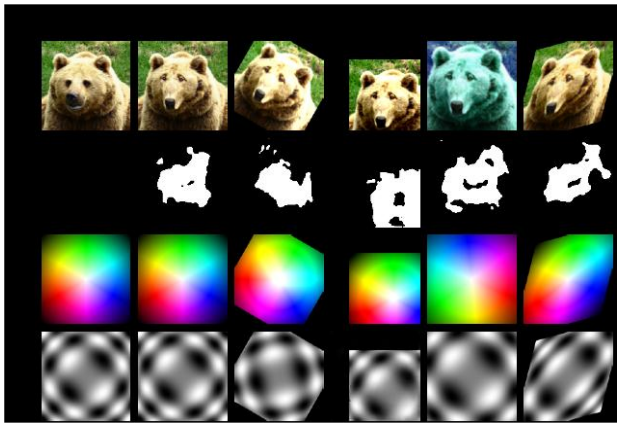

図 1 異なる編集操作の組合せにおける，説明可能ウォータマークの変化例。

概要

本プロジェクトでは，合成メディアに対して代表的に用いられる編集操作を，意味的編集，光学的編集，および幾何的編集の三種類に分類し，それぞれに対応したウォータマークを設計した．

意味的編集に対しては，画像のどの領域が編集されたかが分かる参照パターンを用いた．光学的編集に対しては，色相や明るさの変化が把握しやすいカラーパターンを用いた．幾何的編集に対しては，回転や拡大縮小，せん断などの変形が視覚的に表れやすい波状パターンを用いた．

これらのウォータマークは，ニューラルネットワークにより画像に埋め込みおよび抽出が行われ，編集操作後も編集の影響がウォータマークに反映されるよう学習されている．

編集後に抽出されたウォータマークを解析することで，編集操作の種類やパラメータを推定し，編集履歴を説明可能な形で復元する．

結果および考察

実験により，提案手法は画質への影響が極めて小さく，平均 PSNR は約 48.8 dB，SSIM は約 0.998，LPIPS は約 0.0002 と高い画質保持性能を示した．さらに，編集後に抽出されたウォータマークと，理想的に変換された参照ウォータマークとの平均誤差は小さく，編集操作とウォータマークの変化が高い同期性をもって対応していることを確認した．また，意味編集領域の推定においては，編集領域と推定結果の一致度（IoU）が平均して約 0.6〜0.7 の範囲に達し，光学的および幾何的編集に対しては，各変換パラメータを高い精度で推定可能であることを確認した．さらに，提案手法を用いた合成メディア検出では，各種編集条件下において総合判定精度が 90%以上となり，高い編集履歴推定能力と実用性を有することを確認した．

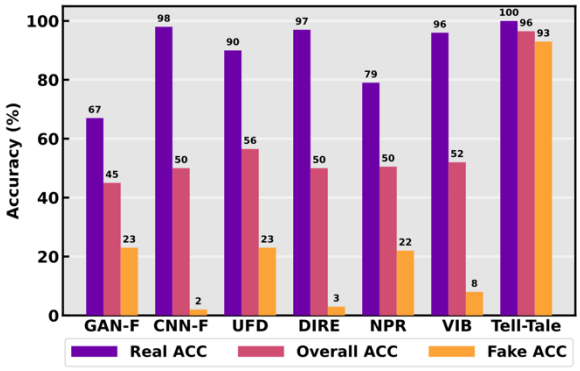

図 2 最先端ベースライン手法との比較における合成メディア検出精度。

まとめ、今後の課題

本プロジェクトでは，合成メディアに対して，編集内容を分かりやすく解析することを目的とした説明可能フォレンジック技術を構築した．

編集操作に対応したウォータマークを用いることで，従来手法では困難であった編集履歴の推定を可能とし，編集の種類およびその程度を説明可能な形で示せることを確認した．

今後の課題としては，対象とする編集操作や生成モデルの種類をさらに拡張すること，編集操作の順序に制約を設けない一般的な編集履歴推定への対応，および大規模データや実運用環境における計算効率の向上が挙げられる．