氏　　名　NGUYEN Hong Huy

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2323 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Improving Security in Facial Biometrics: Views from both Attacker Side and Defender Side

論 文 審 査 委 員　主　査　越前　功
情報学専攻　教授
山岸　順一
情報学専攻　教授
池畑　諭
情報学専攻　助教
佐藤　真一
国立情報学研究所　コンテンツ科学研究系　教授
鄭　銀強
東京大学　大学院情報理工学系研究科　准教授
伊藤　康一
東北大学　大学院情報科学研究科　准教授

(Form 3)

Summary of Doctoral Thesis

Name in full: NGUYEN Hong Huy

Title: Improving Security in Facial Biometrics: Views from both Attacker Side and Defender Side

Biometric authentication has become popular, especially on handheld devices. The face is one of the most popular biometric identifiers because of its contactless property. Thanks to the recent development of technologies in both hardware and software, it has been on the way to replacing the traditional inconvenient password authentication. However, advanced technologies can also be misused by attackers to attack such authentication systems. One example is the superiority of deep neural networks, which can generate realistic images, video, and speech. Synthetic and manipulated images and videos created in this way are called deepfakes. They can be used to break face authentication systems (by attacking the integrated face recognition systems) besides making fake news, impersonating, and harassing. As a result, dealing with deepfakes is a vital task in facial biometrics. There are countless battles between attackers and defenders in which they have been continuously improving themselves to become stronger. This fighting philosophy is the principal motivation of this thesis. By standing on both attacker and defender sides, we can simultaneously uncover some crucial problems in facial biometrics and propose some solutions to make it more secure.

From the attacker side, we assert the robustness of face recognition systems under wolf attack using generative adversarial images. A wolf sample is an input that could match with multiple enrolled user templates. This kind of attack was introduced in 2007 on fingerprints and from then on, it has been a popular attack on fingerprint- and finger-vein-based authentication systems. We are the first in the literature to demonstrate the existence of the master faces that can match with multiple faces from different identities by the face recognition systems. The master face attack is stronger than the face morphing attack since it does not require knowledge from the victim. By improving the latent variable evolution algorithm used in the master print attack and using a powerful facial generation model (StyleGAN), we can generate high-quality master faces that can attack several face recognition systems in white-box, gray-box, and even black-box scenarios. Our experiments confirmed that even with limited resources and using only pre-trained models available on the Internet, attackers can initiate master face attacks. Another contribution is that, by analyzing the distributions of the face embedding (identity) spaces, we point out the limitations of some current face recognition systems and suggest some improvements.

From the defender side, we first propose a detector that works with various kinds

of computer-generated and manipulated images and videos, commonly known as “deepfakes”. The proposed detector uses a capsule network, which is an upgraded version of the traditional convolutional neural networks (CNNs). For traditional CNNs, their performance can be improved by increasing their depth and/or their width, adding more internal connections, or fusing several features or predicted probabilities from multiple CNNs. Consequently, they become bigger, consume more memory and computation power, and require more training data. Besides the dynamic routing algorithm, with the addition of the feature extractor to address the small data issue (common in this task) and the statistical pooling layers (which works well with deepfake artifacts), the proposed Capsule-Forensics network has high detection performance without sacrificing computational resources and memory. To further understand the Capsule-Forensics, we visualize the activation of the components of the Capsule-Forensics, heading towards its explainability.

Second, locating manipulated regions (*i.e.*, performing segmentation) is also important when dealing with fake images and videos, improving the explanability of the results. We design a CNN that uses the multi-task learning approach to simultaneously detect manipulated images and videos and locate the manipulated regions for each query. The information gained by performing one task is shared with the other task and thereby enhance the performance of both tasks. A semi-supervised learning approach is used to improve the network's generalizability. The network includes an encoder and a Y-shaped decoder. Activation of the encoded features is used for binary classification. The output of one branch of the decoder is used for segmenting the manipulated regions while that of the other branch is used for reconstructing the input, which helps improve overall performance. With this design, fine-tuning the network using just a small amount of data enables it to deal with unseen attacks.

In the appendices, we introduce three additional contributions. The first one is about asserting the possibility of enhancing computer-generated (CG) facial images to fool the spoofing detectors, which are usually integrated into the face authentication systems. Unlike the traditional viewpoint of computer graphics which focuses on the rendering phase, we enhanced the rendered images by proposing an enhancer using a CNN called H-Net. It can perform black-box attacks and successfully degraded the accuracies of three spoofing detectors. The second contribution is about discriminating between CG images and photographic images. We build a modular discriminator and devise a probabilistic patch aggregation strategy to deal with high-resolution images. This proposed method outperformed a state-of-the-art method and achieved accuracy up to 100%. The final contribution is about the proposed correction algorithms to correct adversarial images and their labels. Recently, there are adversarial attacks targeting deepfake detection and segmentation methods, hence detecting and correcting them is important. Our proposed method had a promising performance when correcting nearly 90% of adversarial images while only having a little effect on bonafide images.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　NGUYEN Hong Huy

Title
論文題目　Improving Security in Facial Biometrics: Views from both Attacker Side and Defender Side

本学位論文は，顔認証システムに対するプレゼンテーションアタックに関する脅威と対策に関するものである．脅威に関しては，(1) 顔認証システムに登録された複数の顔特徴と類似する顔画像（Master Face）の生成手法について述べ，対策に関しては，(2) 近年社会問題となっている機械学習モデルにより合成された顔画像・映像（フェイク顔画像・映像）を検出（真贋判定）する手法（Deepfake Detection），および(3) フェイク顔画像・映像の検出と改ざん領域の推定を同時に行う手法（Joint Detection and Segmentation of Deepfake）について述べている．顔認証システムに対するプレゼンテーションアタックに関して，攻撃側と防御側の双方の観点から考察を加えることで，安全な顔認証を実現する技術の発展に資することを目的としている．

本学位論文は，全 6 章から構成される．第 1 章では，本論文で扱う問題の重要性，位置付けおよび貢献について説明している．最初に，顔認証が適用されているアプリケーション（顔認証システム）を概観し，それらに対するプレゼンテーションアタックの脅威について述べるとともに，機械学習モデルの進展によって Deepfake に代表されるフェイク顔画像・映像の生成手法がプレゼンテーションアタックに及ぼす影響について述べている．次に，顔認証システムに対するプレゼンテーションアタックに対して，攻撃側と防御側の双方の観点から考察を加えることの重要性について述べ，本学位論文の貢献である 3 つの研究課題（3 章:顔認証システムに登録された複数の顔特徴と類似する顔画像（Master Face）の生成手法, 4 章:フェイク顔画像・映像を検出（真贋判定）する手法（Deepfake Detection），5 章 : フェイク顔画像・映像の検出と改ざん領域の推定を同時に行う手法（Joint Detection and Segmentation of Deepfake））について，研究課題の関係性に言及するとともに，概説している．第 2 章では，歩容認識の概説に加えて，関連研究を分析・比較し，本学位論文の研究課題の新規性および有用性について述べている．

第 3 章では,顔認証システムに登録された複数の顔特徴と類似する顔画像（Master Face）の生成手法について述べている．潜在変数から顔全体を生成する StyleGAN と Latent Variable Evolution (LVE)を組み合わせることで，潜在変数を自動更新し，複数の顔特徴と類似する顔画像（Master Face）を生成する手法を考案し，一定の条件のもとで，Master Face による顔認証システムへの攻撃が有効であることを示している．

第 4 章では，フェイク顔画像・映像を検出（真贋判定）する手法（Deepfake Detection）について述べている．コンピュータビジョンのタスクに用いられていた Capsule Network に着目し, 入力された顔画像にアーティファクトがあるか否かを, 複数の primary capsule が異なる観点で判断し, それらの判断結果に基づいて output capsule が入力画像の真贋を

判断する手法を考案している．評価実験により，提案手法は，従来の Deepfake Detection 手法と比較して，モデルのパラメター数を大幅に削減しながら，検出精度の向上を実現することを示している．

第 5 章では，フェイク顔画像・映像の検出と改ざん領域の推定を同時に行う手法（Joint Detection and Segmentation of Deepfake）について述べている．入力された顔画像の真贋判定を行うタスクと改ざん領域の推定タスクを同時に行う multi-task learning を考案し，どの手法によって顔画像が合成されたかを推定可能な手法を考案している．評価実験により，既知の手法に対して，高い検出精度と改ざん領域の推定精度を示すだけでなく，未知の手法によって合成された顔画像が出現しても，少量の生成画像を用いて学習することで，当該手法に対する高精度の検知が可能となることを示している．

第 6 章では，結論として，本学位論文の貢献についてまとめ，同分野における今後の研究課題について述べている．

公開発表会では博士論文の章立てに従って発表が行われ，その後に行われた論文審査会及び口述試験では，審査員からの質疑に対して適切に回答がなされた．質疑応答後に審査委員会を開催し，審査委員で議論を行った．その結果，出願者は情報学分野の十分な知識と研究能力を持つと認められ，また研究内容は学位論文として十分なレベルの新規性や有効性があると認められた．本学位論文は，これまで検討がなされていなかった，機械学習モデルを用いた，顔認証システムのプレゼンテーションアタックに関する脅威と対策に着目し，顔認証システムに登録された複数の顔特徴と類似する顔画像の生成手法，フェイク顔画像・映像を検出（真贋判定）する手法，およびフェイク顔画像・映像の検出と改ざん領域の推定を同時に行う手法を提案し，マルチメディアフォレンジクスやバイオメトリクス分野において重要な貢献をなしたものである．また，本学位論文の成果は，査読付きジャーナル論文 1 編，査読付き国際会議論文 5 編として発表されている．以上の理由により，審査委員会は，本学位論文が学位の授与に値すると判断した．