Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.2, Apr. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.2.201

프라이버시 보호를 위한 얼굴 인증이 가능한 비식별화 얼굴 이미지 생성 연구*

이 정 재,[1†] 나 현 식,[2] 옥 도 민,[2] 최 대 선[3‡]
[1,2,3]숭실대학교 (학생, 대학원생, 교수)

De-Identified Face Image Generation within Face Verification for Privacy Protection*

Jung-jae Lee,[1†] Hyun-sik Na,[2] To-min Ok,[2] Dae-seon Choi[3‡]
[1,2,3]Soongsil University (Student, Graduate student, Professor)

요 약

딥러닝 기반 얼굴 인증 모델은 높은 성능을 보이며 많은 분야에 이용되지만, 얼굴 이미지를 모델에 입력하는 과정에서 사용자의 얼굴 이미지가 유출될 가능성이 존재한다. 얼굴 이미지의 노출을 최소화하기 위한 방법으로 비식별화 기술이 존재하지만, 얼굴 인증이라는 특수한 상황에서 기존 기술을 적용할 때에는 인증 성능이 감소하는 문제점이 있다. 본 논문에서는 원본 얼굴 이미지에 다른 인물의 얼굴 특성을 결합한 뒤, StyleGAN을 통해 비식별화 얼굴 이미지를 생성한다. 또한, HopSkipJumpAttack을 활용해 얼굴 인증 모델에 맞춰 특징들의 결합 비율을 최적화하는 방법을 제안한다. 우리는 제안 방법을 통해 생성된 이미지들을 시각화하여 사용자 얼굴의 비식별화 성능을 확인하고, 실험을 통해 얼굴 인증 모델에 대한 인증 성능을 유지할 수 있음을 평가한다. 즉, 제안 방법을 통해 생성된 비식별화 이미지를 사용하여 얼굴 인증을 할 수 있으며, 동시에 얼굴 개인정보 유출을 방지할 수 있다.

ABSTRACT

Deep learning-based face verificattion model show high performance and are used in many fields, but there is a possibility the user's face image may be leaked in the process of inputting the face image to the model. Although de-identification technology exists as a method for minimizing the exposure of face features, there is a problem in that verification performance decreases when the existing technology is applied. In this paper, after combining the face features of other person, a de-identified face image is created through StyleGAN. In addition, we propose a method of optimizing the combining ratio of features according to the face verification model using HopSkipJumpAttack. We visualize the images generated by the proposed method to check the de-identification performance, and evaluate the ability to maintain the performance of the face verification model through experiments. That is, face verification can be performed using the de-identified image generated through the proposed method, and leakage of face personal information can be prevented.

Keywords: De-identification, Face verification, Face privacy, Generative Adversarial Network, Decision-based attack

Received(01. 26. 2023), Accepted(02. 15. 2023)
* 본 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1014813)
† 주저자, dlwjdwo00701@gmail.com
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

딥러닝 기반 얼굴 인증 모델은 사용자의 신원을 인증하기 위해 다양한 분야에서 상용화되고 있다. 특히 의료 서비스, 직원 인증, 스마트 기기 잠금 해제, 출입국 관리 등 다양한 분야에서 하나의 인증 수단으로 사용될 수 있다[1, 2]. 최근에는 딥러닝 기술의 기하급수적인 발달로 인해 얼굴 인증의 기술의 신뢰도가 더욱 높아져, 금융 서비스에서도 점차 상용화되고 있는 추세이다[3]. 얼굴 인증은 스마트 기기를 기반으로 빠르고 편리하게 사용자의 신원을 확인할 수 있다는 장점이 있다. 또한, 여러 보안 소프트웨어와 호환이 될 수 있으며, 서비스를 쉽게 통합할 수 있다. 하지만 얼굴 인증을 통한 사용자 신원 확인 서비스의 경우 사용자의 얼굴 프라이버시가 침해될 가능성이 있다.

스마트 기기 내부에 얼굴 인증 모델이 저장된 경우가 아닌 검증을 위한 기등록된 사용자 얼굴 이미지와 얼굴 인증 모델이 중앙 서버에 저장된 시스템의 경우, 중앙 서버와 연결된 네트워크를 통해 스마트 기기는 사용자의 얼굴 이미지를 전송해 얼굴 인증 모델에 입력하는 과정이 존재한다. 이때, 외부 공격자가 접근하여 ①사용자에게 데이터 입력에 대한 거짓 경로를 제공하여 사용자의 얼굴 이미지를 탈취하는 경우, ②네트워크 내 경로를 따라 전송되는 데이터를 도청하는 경우[4], ③얼굴 인증 시스템 또는 데이터 저장소에 접근하여 내장된 얼굴 이미지를 수집하는 경우가 발생할 수 있다. 또한, 얼굴 인증 시스템을 관리하는 관리자가 무단으로 ④데이터 저장소에 접근해 얼굴 이미지 데이터를 금전적인 이익을 취하기 위해 타인에게 판매 및 유출하는 경우, ⑤다양한 목적으로 얼굴 이미지를 데이터 세트로 공개 및 사용하여 사용자의 얼굴 이미지를 노출하는 경우가 발생할 수 있다. 얼굴 이미지는 사용자를 특정할 수 있는 민감한 개인정보로써 적절한 보안 대책이 마련되지 않으면 심각한 프라이버시 침해 문제를 초래할 수 있다. 따라서, 얼굴 인증 모델의 이용에 있어서 사용자의 개인정보 침해 가능성을 방지하기 위한 효과적인 방법이 필요하다.

딥러닝 기반 얼굴 인증 모델에 입력되는 이미지가 유출되어도, 공격자나 제 3자가 해당 입력 이미지로부터 사용자를 식별 및 특정할 수 없게 만든다면 사용자의 개인정보를 보호할 수 있다. 이때, 비식별화는 사용자 얼굴의 노출을 최소화하는 방법으로 효과적으로 사용될 것이다. 하지만 기존 비식별화 기술[5, 6, 7, 8]은 실제 얼굴의 노출을 최소화하기 위해 높은 수준의 익명화에만 초점을 두었다. 이는 얼굴 인증 시스템에 적용을 상정하지 않았기 때문이다. 따라서 비식별화와 얼굴 인증이 함께 필요한 상황에서 기존 기술을 적용하기에는 한계가 있고, 이를 보완하기 위한 연구가 필요하다. 사용자의 프라이버시를 지키기 위한 기술인 비식별화와 사용자의 신원을 확인하기 위한 얼굴 인증 기술의 목적은 서로 상충하지만 적절한 수준에서 두 기술이 혼합된다면, 유용하게 작용할 것이다. 이에 따라, 우리는 얼굴 인증 모델의 인증 성능을 유지하면서 입력 이미지의 충분한 비식별화를 할 방법을 찾는 것에 초점을 두었다.

본 논문에서는 얼굴 특징 벡터를 기반으로 이미지에 따른 다양한 스타일을 생성하여 입력해 얼굴 이미지를 생성할 수 있는 StyleGAN[12]을 활용한 새로운 비식별화 방법을 제안한다. 이 방법은 사용자의 얼굴 특징 정보에 목표로 하는 특정 인물의 얼굴 특징 정보를 결합해 StyleGAN에 입력하여 새로운 얼굴 이미지를 생성하며, 이때 얼굴 인증 모델의 인증 성능을 유지하는 동시에 두 얼굴 특징 정보의 결합 비율을 최적화하여 비식별화 강도를 최대화한다. 최적화를 위해 타겟 얼굴 인증 모델에 반복적인 질의를 시도해 공격하는 Decision-based attack의 일종인 HopSkipJumpAttack[14]을 활용한다.

우리는 제안방법을 통해 생성된 비식별화 이미지의 시각화를 통해 원본 이미지로부터 사용자의 얼굴이 충분히 변조됨을 확인한다. 또한, 원본 이미지와의 유사도를 측정하여 얼굴 인증 모델의 인증 성능이 유지됨을 평가한다. 또한, 각 얼굴 인증 모델별로 두 얼굴 특징 결합의 최적화 과정에 따른 비식별화 얼굴 이미지를 확인하며, 생성 얼굴 이미지가 목표로 하는 특정 인물의 얼굴 특징에 가까워지는 것을 확인한다.

II. 관련 연구

2.1 얼굴 비식별화

최근 사용자 개인 프라이버시의 중요성이 커지면서 얼굴 이미지의 중요성이 함께 강조되고 있으며, 초기에 여러 연구에 있어 사용자의 얼굴 이미지에 특정 이미지를 덧씌우는 방식으로 비식별화 기술들이 제안됐다[5, 6]. 얼굴 이미지에 검은색의 가려진 모자이크 이미지를 더하는 기술[5]과 얼굴 주위를 픽셀화해 모자이크하는 비식별화 기술[6]은 얼굴 이미지에 빠

르게 적용할 수 있지만, 역으로 복원할 수 있고 얼굴의 형태를 지우는 문제점이 있다. 이를 보완하기 위해 비식별화 얼굴 이미지 생성에 있어 인공지능 기술 GAN(Generative Adversarial Networks)을 적용하는 연구들이 진행되었다[7, 8]. 얼굴의 특징을 담고 있는 랜드마크 영역을 잘라낸 뒤, 잘라낸 부분에 다른 인물의 얼굴 이미지를 GAN을 통해 덧씌워 생성하는 비식별화 기술[7]과 두 개의 GAN 모델을 통해 얼굴 특징과 공간적 특징을 분리해 학습해 두 얼굴의 특징을 추출해 합성해 얼굴을 생성하는 비식별화 기술[8]이 있다.

2.2 얼굴 인증 모델 공격 기술

얼굴 인증 모델들이 발달하면서 함께 얼굴 인증 모델을 공격하는 여러 기술이 연구되었다[9, 10, 11]. 해당 공격 기술들은 얼굴 이미지의 인물이 시각적으로는 다르게 보이지만 얼굴 인증 모델의 오분류를 일으킨다는 점에 있어서 제 3자에게 얼굴 특징을 숨기는 것이 목적인 비식별화와 유사하다고 볼 수 있다. 여기에는 목표 얼굴 인증 모델에 대해서 질의를 하거나[9], 역전파를 통한 Adversarial attack[10]으로 기존 얼굴 이미지에 노이즈를 주입하는 방식으로 얼굴 이미지를 변형해 얼굴 인증 모델이 오분류를 일으키도록 하는 공격 방식이 있다. 또한, 얼굴 인증 모델에 대해 사용자의 얼굴에 다른 얼굴을 합성하는 공격인 Morphing Attack을 GAN에 적용하는 방식[11]이 있다.

2.3 종래 연구 적용의 문제점

기존의 비식별화 연구들의 제안 초점은 얼굴 이미지의 특징을 감추고자 높은 익명화에 맞춰졌다. 그렇기에 얼굴 인증이 되는 수준까지의 비식별화 정도를 정할 방법이 없어, 앞서 말한 얼굴 인증 시스템에서는 적용할 수 없다. 앞선 연구들의 비식별화 기술을 적용해 얼굴 이미지를 생성하더라도 얼굴 인증 모델은 사용자를 식별하는 데 한계가 있다.

얼굴 인증 모델 공격 기술 중 노이즈를 주입하는 방식의 경우 시각적으로 노이즈가 보여, 얼굴 이미지처럼 보이지 않을 우려가 있다. 또한 [11]의 Morphing Attack의 경우 GAN을 통해 완전한 얼굴을 생성하지만 합성하는 얼굴의 대상 또한 사용자의 얼굴과 동시에 얼굴 인증 모델이 같은 인물이라 식별해야 하기에 우리의 목적인 사용자에 대해서만 얼굴 인증이 되지만 사용자의 얼굴 특징을 최소화하는 비식별화에는 맞지 않는다.

III. 배경 지식

3.1 StyleGAN[12]

StyleGAN은 다양한 종류의 이미지를 생성할 수 있는 GAN 기반 생성 모델이다. StyleGAN은 18개의 층으로 이루어져 있으며, 눈, 코, 입의 형태와 얼굴의 모양 등을 나타내는 1 × 512 크기의 스타일 W 벡터를 Generator의 18계층에 복사해 이미지를 생성한다. 해당 모델의 구조도는 Fig. 1.과 같다. StyleGAN은 각 계층에서 입력된 스타일을 Convolution 계층을 거치며 업샘플링된 이미지들에 AdanIN 연산을 통해 덧씌워가며 고해상도의 이미지를 생성한다. 얼굴 이미지 생성에 대해 학습된 StyleGAN의 경우 18개 계층에서 앞의 4계층은 얼굴의 형태 생성에 관여한다. 그 다음 4계층은 눈, 코, 입의 위치에 관여한다. 나머지 10계층은 머리 색깔, 피부색 등에 관여한다. 이를 통해, 입력되는 스타일 벡터에 따라 다양한 현실 인물 사진과 같은 얼굴 이미지의 생성이 가능하다. 고해상도 이미지 생성 및 이미지의 정교함을 개선하기 위해 임의의 노이즈를 추가적으로 이

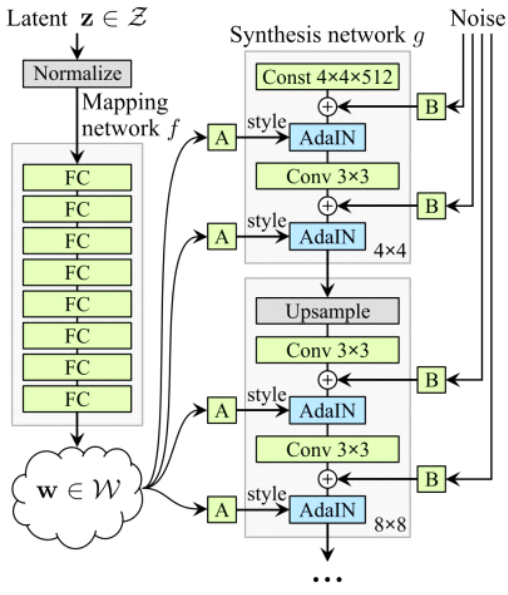

Fig. 1. Architecture of StyleGAN Generator[12]

미지 생성 과정에 주입할 수 있다.

3.2 Pixel2Style2Pixel (PSP)[13]

PSP는 StyleGAN의 Encoder로써, 입력되는 얼굴 이미지로부터 StyleGAN의 Generator에 사용되는 스타일 벡터를 얻을 수 있다. 기존의 1 × 512의 크기로 18개로 복사되는 것이 아닌 18계층의 맞는 18 × 512 크기의 각 계층에 입력되는 W^+ 벡터를 독립적으로 추출해서 정교한 스타일을 얻을 수 있으며, 이로 인해 기존 StyleGAN에 적용될 수 있는 다른 Encoder보다 성능이 뛰어나다. 해당 벡터를 이용해 얼굴 이미지를 생성한다면, 원본 이미지와 매우 유사하게 생성되며, W^+ 벡터를 조작하면서 생성 모델 사용자가 원하는 모습 및 형태의 얼굴 이미지를 생성할 수 있다.

3.3 HopSkipJumpAttack (HSJA)[14]

HSJA는 입력 이미지를 변조해 타겟 분류 모델에 질의한다. 해당 과정의 분류 결과를 이용해 기존 입력 이미지가 시각적으로는 다르게, 변형하고자 하는 목표 이미지처럼 보이지만 기존의 분류 결과를 유지해 오분류를 유도하는 Decision-based attack 공격 방법의 하나며, 최신 관련 공격 기술들과 비교하여 적은 입력 횟수를 통해 목표 이미지에 가까워진다. 모델의 역전파를 통해 계산된 기울기를 이용해 이미지에 노이즈를 주입해 변형하는 다른 공격과 달리 무작위로 변형을 조정하며 공격해가며, 공격하고자는 타겟 모델의 내부 파라미터의 접근이 불가한 블랙박스 상황에서도 적용할 수 있다.

IV. 제안 방법

4.1 얼굴 인증 시스템 설정

본 논문에서 제안한 얼굴 인증이 가능한 비식별화 얼굴 이미지 생성 방법의 전체적인 구조는 Fig. 2.와 같다. 우리는 개인의 스마트 기기가 아닌 중앙 서버에 얼굴 인증 모델(Face Recognition Model)과 검증을 위한 X_{system}(기등록 얼굴 이미지)이 사전에 등록 및 저장되어 있으며, 사용자는 스마트 기기의 카메라를 통해 자신의 얼굴인 X_{source}(원본 얼굴 이미지)를 캡처한 뒤, 중앙 서버와 연결된 네트워크를 통해 모델에 입력해 인증 결과를 확인하는 일반적인 얼굴 인증 시스템을 가정한다. 또한 Decision-based attack을 기반으로 하기에, 비식별화를 진행하기 이전에 X_{source}와 X_{system}은 얼굴 인증을 성공해야 한다.

4.2 비식별화율 조정 변수 λ 설정

우리는 사용자의 얼굴 이미지인 X_{source}와 사용자

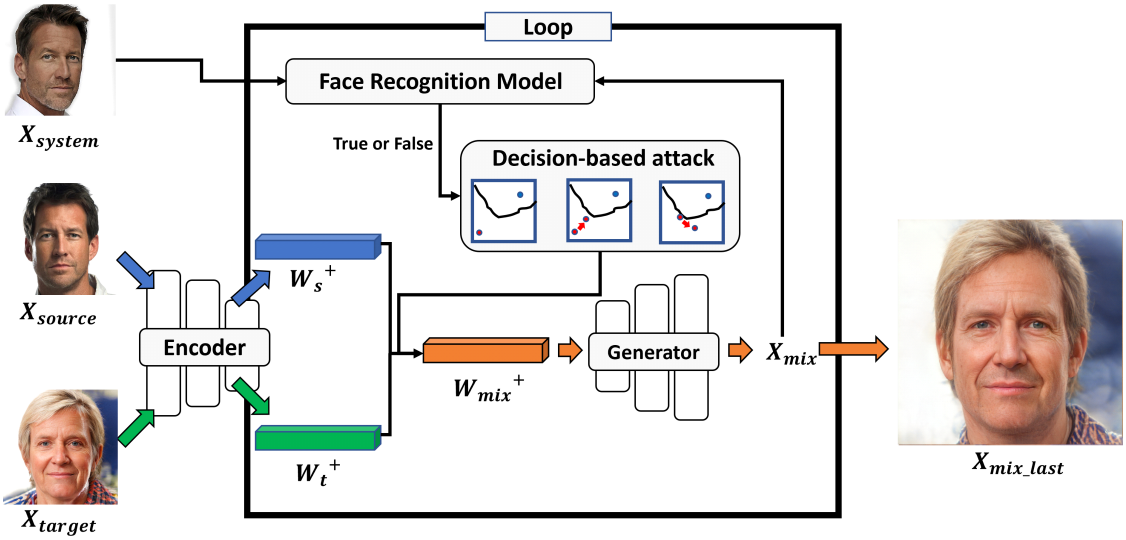

Fig. 2. Architecture of generating de-indentification with face verification

가 아닌 다른 인물의 얼굴 이미지인 X_{target}(주입 얼굴 이미지)를 PSP(Encoder)를 통해 각각의 이미지 속 인물의 얼굴 특징을 포함하고 있는 스타일 벡터인 W_s^+와 W_t^+를 생성한다. 두 스타일 벡터의 크기는 StyleGAN의 Generator의 18계층에 맞춰 18 × 512이다. 스타일이 StyleGAN 모델에 입력되어 얼굴 이미지를 생성하기 이전에, 최적의 비식별화 정도를 찾기 위해서 비식별화율 조정 변수인 $\lambda \in [0,1]$를 설정했다. λ는 각 스타일 벡터의 원소 간 결합 비율의 역할을 하기 때문에 0과 1 사이의 범위를 갖는 소수로 범위를 제한한다. 변수 λ의 크기 또한 스타일 벡터들과 동일하게 18 × 512 크기이다. 설정한 λ를 비율로서 이용해 선형 결합식인 수식(1)을 통해 W_s^+와 W_t^+를 결합한다.

$$W_{mix}^+ = (1-\lambda) \times W_s^+ + \lambda \times W_t^+ \quad (1)$$

λ의 값이 모두 0일 경우 X_{xource}의 얼굴 특징인 $W_{mix}^+ = W_s^+$를 StyleGAN에 입력해 얼굴 이미지를 생성하기 때문에 X_{source}와 같은 얼굴을 생성 및 복원한다. 반대로 λ의 값이 모두 1일 경우 $W_{mix}^+ = W_t^+$로 X_{target}과 같은 얼굴을 생성한다. λ의 값이 모두 0.5일 경우, X_{source}의 얼굴 특징 절반과 X_{target}의 얼굴 특징 절반으로 생성된 얼굴 이미지를 얻을 수 있을 것이다. 즉, 최적의 λ가 형성될 수 있도록 반복적으로 업데이트하여 X_{mix}를 생성한다면 얼굴 인증이 되지만 충분한 비식별화가 되는 얼굴 이미지를 생성할 수 있다. StyleGAN Generator의 18계층 중 상위 4계층은 입력되는 스타일로부터 얼굴의 형태와 모습을 이미지화하는데, 본 논문에서는 X_{source}의 얼굴 특징만을 온전히 비식별화하고자 λ의 상위 4 × 512의 값을 0으로 고정해주고, 나머지 하위 14 × 512의 값만을 업데이트한다.

4.3 변수 λ 최적화

목적에 맞게 비식별화율 조정 변수 λ의 최적화하기 위해 HSJA를 활용해 얼굴 인증 모델에 일종의 공격을 진행한다. λ는 X_{source}의 얼굴 특징을 갖는 얼굴 이미지를 생성할 수 있는 $\lambda_{start} = [[0,0\ldots]\ldots]$를 초깃값으로 하며 X_{target}의 이미지를 생성할 수 있는 $\lambda_{target} = [[1,1\ldots]\ldots]$를 목표로 고정한 뒤, 이를 향해 조정해준다. 최적의 λ를 업데이트하기 위해 사용되는 HSJA은 Fig. 3.에서처럼 총 3단계가 반복된다. 경계를 통해 나눠진 아래 부분은 얼굴 인증 모델이 입력되는 이미지를 사용자라고 인증하는 영역이며, 윗 부분은 X_{target}의 해당 클래스 영역이다. 먼저, λ_{start}에서 시작한 λ를 λ_{target} 향해 이진 탐색을 통해 이동시킨다. 이진 탐색의 기준은 이동시킨 λ를 통해 생성한 X_{mix}의 얼굴 인증 성공 여부이다. 이 때, 성공 여부는 True와 False로만 표현된다. 이를 통해 얼굴 인증 모델이 X_{mix}를 사용자로 인증하는 경계에 위치를 알 수 있다. 다음으로, λ에 여러번 무작위의 변조를 추가하여 생성한 X_{mix}을 얼굴 인증 모델에 반복적으로 질의한다. 파란색 화살표는 변조된 λ로 생성한 얼굴 이미지 X_{mix}가 사용자의 얼굴로 인증하는 영역의 포함된 것이고, 빨간색 화살표의 경우 인증하

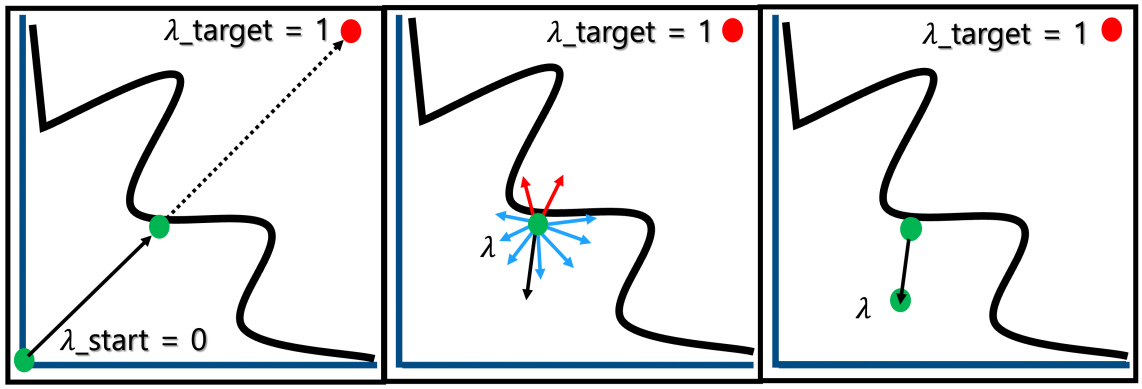

Fig. 3. Step per HSJA in our method

지 않는 영역에 포함된 것이다. 해당 질의에 따른 인증 결과를 이용해 얼굴 인증이 가능한 영역을 향한 방향 벡터를 생성한다. 마지막으로, 생성한 방향 벡터를 통해 λ를 움직여준다. 해당 3단계는 모든 값이 1인 λ_{target}를 향해 최대한 가까워지는 횟수에 맞춰 반복된다. 반복이 끝난 마지막 횟수에서 첫 번째 과정인 이진 탐색을 수행해 λ가 λ_{target}에 최대한 가까워질 수 있다. 해당 최적화 방식으로 얼굴 인증이 가능한 최적의 비식별화 얼굴 이미지인 X_{mix_last}를 생성할 수 있다.

V. 실 험

5.1 비식별화 확인

비식별화 확인을 위한 실험을 위해 FFHQ[15] 데이터셋으로 학습된 StyleGAN Generator를 이용했으며, Encoder로 PSP를 사용했다. 타겟 얼굴 인증 모델로 VGG-Face[16], Facenet[17], ArcFace[18], Facenet512[19]로 총 4개의 모델을 설정해 진행했다. Celeb-HQ[20]의 경우 고화질의 유명인들의 얼굴 이미지 데이터셋으로 동일인에 대한 여러 이미지로 구성되어 있어 해당 데이터셋으로 X_{source}와 X_{system}를 설정했다. X_{target}은 StyleGAN에 무작위로 생성한 스타일 벡터를 주입해 얼굴 이미지를 생성했다. 각 모델별로 400개의 X_{source}와 X_{system}쌍에 대해 2,246번의 질의를 통해 얼굴 인증이 가능한 비식별화 이미지를 생성했다.

본 논문에서 제안하는 방법의 경우 얼굴 인증 모델의 내부 파라미터와 인증 방식을 확인할 필요가 없는 블랙박스 상황에서 진행되지만, 평가를 위해 이를 확인한다. 얼굴 인증 모델의 경우 두 얼굴 이미지를 학습된 모델을 입력해 벡터화시킨 뒤, 두 벡터의 코사인 거리(2)를 계산해 해당 값이 특정 임계값보다 낮게 측정될 경우 동일인으로, 높을 경우 다른 인물로 판단한다. 실험에 있어 각 모델별로 설정된 임계값은 VGG-Face : 0.4, Facenet : 0.4, ArcFace : 0.68, Facenet512 : 0.3이다. X_{source}와 동일인의 얼굴 이미지인 X_{system}과 X_{source}에 X_{target}를 주입한 최종 비식별화 이미지인 X_{mix_last}의 코사인 거리를 비교해 인증 성능을 계산한다.

$$cosine\ \ distance = 1 - \frac{X \cdot Y}{\|X\|_2 \cdot \|Y\|_2} \quad (2)$$

Fig. 4.의 X_{mix_last}는 Facenet512 모델에 대해 제안 방법의 실험으로, 최종적으로 구한 λ_{mix_last}를 이용해 생성한 이미지들이다. 시각적으로 보았을 때 원본 이미지인 X_{source}로부터 X_{mix_last}는 X_{system}를 파악하기 어려울 정도로 충분한 비식별화가 되었음을 확인할 수 있다. 또한 여성-여성, 남성-남성의 동성의 쌍인 X_{source}와 X_{target}의 경우 시각적으로 비식별화 정도가 비교적 떨어지는 것으로 파악되었다. 여성-남성, 남성-여성의 이성의 쌍일 때 시각적 비식별화 정도가 큰 것으로 분석된다.

Table 1.은 각 모델별로 생성한 400장의 X_{system}과 X_{mix_last}를 모델에 입력해 출력된 벡터를 코사인 거리를 통해 유사도를 측정한 결과이다. 400장에 대한 평균 코사인 거리가 임계값에 매우 가까워지고, 또한 모든 코사인 거리가 모델의 입계값 미만에 위치해 인증 성공 비율이 100%가 되어 제안 방법으로 생성한 이미지가 충분한 비식별화가 되지만 얼굴 인증 모

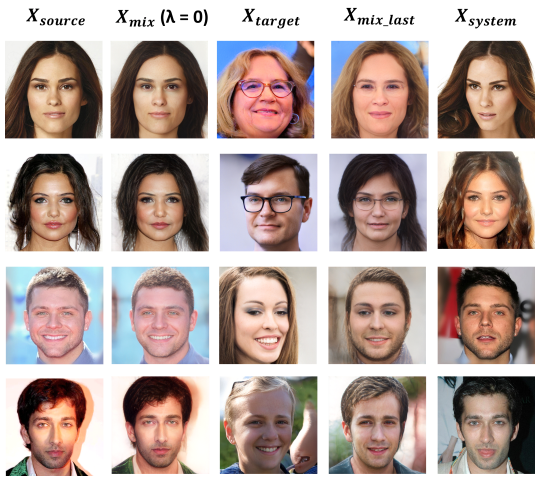

Fig. 4. Result for de-identifiaction with Facenet512

Table 1. Cosine distance per face recognitior model

Model	Distance	Rate
VGG-Face	0.3989760	100 %
Facenet	0.3998697	100 %
ArcFace	0.6798958	100 %
Facenet512	0.2998695	100 %

델의 인증 성능이 유지됨을 확인할 수 있었다.

5.2 모델별 이미지 변화

추가 실험을 통해 모델별로 동일한 X_{source}, X_{system}와 X_{target} 쌍에 대한 비식별화 이미지 최적화에 있어, 제안 방법으로 생성한 얼굴 이미지인 X_{mix}의 얼굴 특징이 질의 횟수에 대비해 어떻게 달라지는지 확인했다. 실험 설정의 경우 5.1과 동일하게 설정해 진행했다.

Fig. 5.는 제안방법에서 활용되는 HSJA의 3단계를 한 step으로서, 생성한 얼굴 이미지를 시각화한 것이다. 각 모델에 대한 얼굴 인증 성능의 경우 Facenet512, Facenet, ArcFace, VGG-FACE의 순서로 좋다. 시각화된 결과를 확인해보면 제일 성능이 떨어지는 VGG-FACE와 ArcFace 모델의 경우 1~2 step에서부터 X_{target} 특징이 잘 주입되는 것을 확인할 수 있지만 Facenet512와 Facenet의 경우 6~7 step이 되어야 비식별화의 기능으로서 X_{target}의 얼굴 특징이 주입되는 것을 확인할 수 있었다.

Fig. 6.의 경우 각 step별로 조정되는 λ와 모든 값이 1인 목표 비율 값인 λ_{target}간의 거리인 L2 Norm을 측정한 결과를 도식화한 것이다. 비식별화 이미지를 시각화한 것과 마찬가지로 성능이 떨어지는 VGG-FACE와 ArcFace의 경우 step 대비 빠르게 거리가 줄어듦을 확인할 수 있으며, 9~10 step 정도 진행되었을 때 더이상 거리가 가까워지지 않음을 확인할 수 있었다. 즉, 두 얼굴 인증 모델에 대한 인증 경계가 비교적 엄격하게 형성되지 않음을 생각할 수 있다. Facenet512과 Facenet의 경우 두 모델에 비해 12~13 step이 진행되었을 때, 거리가 가까워지지 않으며, 그 값이 비교적 컸다. 다시 말해, 두 모델은 VGG-FACE와 ArcFace보다 인증 경계가 비교적 엄격하다는 것을 관찰할 수 있다.

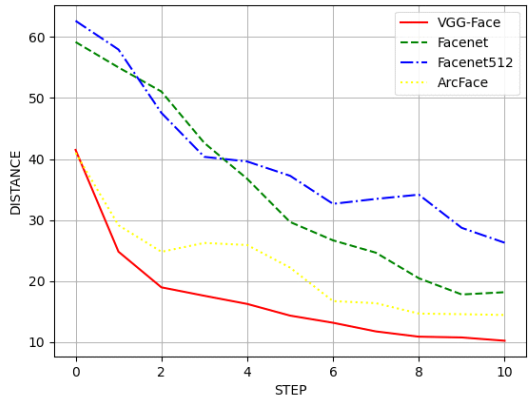

Fig. 6. Distance for step per face recognition model

VI. 고 찰

6.1 적용 방안 및 기대 효과

본 논문에서 제안하는 비식별화 기술은 사용자가 직접 다운로드할 수 있는 스마트폰 애플리케이션을 통해 서비스 형식으로 적용할 수 있다. 만일 사용자가 특정 서비스를 이용할 때 디지털 환경에서 본인 얼굴 이미지를 제공하면서 신원 인증을 해야하는 경우, 본 논문의 제안 기술을 기반으로 한 스마트폰 애플리케

Fig. 5. De-identification images per step according to model

이선을 통해 얼굴 인증이 가능한 비식별화 이미지를 생성하여 활용할 수 있다. 이 때, 애플리케이션에 주입 얼굴 이미지 X_{target} 를 사용자가 원하는 외모로 변경, 여러 개의 비식별화된 이미지를 선택, 비식별화 정도를 설정하는 등 사용자 중심의 다양한 옵션을 제공할 수 있다.

한편, 스마트폰 등 카메라를 통한 신원 인증 시 이미지 수집 과정 중 본 논문의 제안 기술을 내장하여 기업의 관점에서 프라이버시 안전성이 보장된 서비스를 구축할 수 있다. 일반적으로 얼굴 인식을 기반으로 하는 서비스는 실시간 얼굴 이미지 수집, 얼굴 영역 인식, 얼굴 특징 추출, 얼굴 특징 계산 및 분석, 그리고 신원 확인의 절차를 통해 수행된다. 이 때, 사용자가 다양한 서비스에서 실시간 얼굴 인증을 하는 경우, 하나의 스마트폰 카메라를 통해 데이터가 수집되며, 스마트폰 기기 자체에 제안 기술을 기반으로 한 얼굴 인증용 이미지 변환 기술을 내장함으로써 실시간 얼굴 이미지 수집 단계에서 내장된 기술을 통해 자동으로 비식별화 이미지를 생성 및 전송할 수 있다.

그 밖에도, 본 논문의 제안 기술은 공항 내 관광객 출입국 관리 시스템 보호, 스마트폰 애플리케이션 내 본인 인증을 위한 얼굴 이미지 전송 중 유출 후 피해 방지, 연구용 얼굴 이미지 데이터 세트 수집 중 일반인 얼굴 노출 최소화, 사용자의 외모 특징을 유지한 비식별화된 아바타 생성 등 다양한 시나리오를 통해 활용이 될 수 있다. 이를 통해 사용자의 얼굴 프라이버시를 보호할 수 있으며, 더 나아가 메타버스와 같은 가상 세계 내 사용자의 외모를 닮은 아바타 생성 시, 실제 얼굴을 그대로 노출하지 않고 활동할 수 있으며, 아바타를 통한 생체 인증 기술의 발전에 기여할 수 있을 것으로 기대된다.

6.2 한계점 및 향후 계획

본 논문은 서론에서 네트워크 상의 얼굴 인증 모델의 얼굴 이미지 입력과정에서 유출 가능성을 문제점으로 지적했다. 본 논문의 제안 방법으로 얼굴 인증이 가능한 비식별화 얼굴 이미지를 생성할 수 있지만, 최적화 과정의 초기에 비식별화율이 적은 얼굴 이미지로 얼굴 인증 모델에 대해 질의를 하는 과정에서 여전히 원본 이미지의 얼굴 특징의 노출 가능성과 이에 따라 제 3자가 사용자를 특정할 가능성에 대한 한계가 존재한다. 다만 얼굴 인증 시스템의 모델에 대해 유사한 인증 결과를 유추할 수 있는 복제된 얼굴 인증 모델이 있다면 해당 한계를 해결할 수 있을 것이고, 이를 해결할 추가적인 다른 방법이 필요하다.

향후 연구로 앞선 한계에 대해 X_{source}와 X_{target}을 입력으로, 제안 방법을 통해 생성한 최종 비식별화 조정 변수 λ_{mix_last}를 출력으로 데이터셋을 구축해 신경망 기반 생성 모델을 학습시켜 해당 제안 방법의 한계점을 극복하는 연구를 진행할 계획이다. 또한 상용화되어 실사용되고 있는 얼굴 인증 시스템에 대해서도 제안 방법을 적용하는 실험을 진행할 계획이다.

VII. 결 론

본 논문은 얼굴 인증 모델에 대한 인증 성능의 유지와 비식별화 성능을 동시에 만족할 수 있는 연구의 필요성을 최초로 제시한다. 또한, 우리는 이를 해결하기 위해 StyleGAN의 이미지 생성 과정에서의 비식별화 조정 변수 λ의 설정과 HSJA를 활용한 최적화를 기반으로 비식별화 방법을 제안한다. 실험을 통해 해당 방법이 얼굴 인증 모델에 대한 인증 성능을 유지하면서, 원본 이미지로부터 시각적으로 충분한 비식별화가 되었음을 보였다. 즉, 본 논문의 제안 방법은 얼굴 인증과 비식별화가 함께 필요한 상황인 얼굴 이미지를 얼굴 인증 모델에 입력하는 과정에서 사용되어 사용자의 얼굴 개인정보 침해 가능성을 방지할 수 있다.

References

[1] B. Jeon, B. Jeong, S. Jee, Y. Huang, Y. Kim, G.H Park, J. Kim, M. Wufuer, X. Jin, S.W. Kim, and T.H. Choi, "A Facial Recognition Mobile App for Patient Safety and Biometric Identification: Design, Development, and Validation," JMIR mHealth and uHealth, vol. 7, no. 4, e11472. Apr. 2019.

[2] T. Zhu and L. Wang, "Feasibility study of a new security verification process based on face recognition technology at airport," Journal of Physics: Conference Series, vol. 1510, no. 1, pp. 12-25, May. 2020.

[3] R.K. Rija, G. Muttasher, and A.

Al-Araji, "Payment Systems Based on Face Recognition: A Survey," Journal of Optoelectronics Laser, vol. 41, no. 5, pp. 563-571, May.

[4] B. Thirimachos, Face Recognition Across the Imaging Spectrum, Springer, pp. 165-194, Feb. 2016.

[5] E.M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," IEEE transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 232-243, Feb. 2005.

[6] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-Based Face De-Identification," Conference on Computer Vision and Pattern Recognition Workshop, pp. 161-161, Jun. 2006.

[7] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5447-5456, Jun. 2020

[8] Y. Jeong, J. Choi, S. Kim, Y. Ro, T.H. Oh, D. Kim, H. Ha, and S. Yoon, "FICGAN: Facial Identity Controllable GAN for De-identification," arXiv preprint arXiv:2110.00740, Oct. 2021.

[9] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7714-7722, Jun. 2019.

[10] E. Chatzikyriakidis, C. Papaioannidis, and I. Pitas, "Adversarial face de-identification," 2019 IEEE International Conference on Image Processing, pp. 684-688, Sep. 2019.

[11] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, ''Face recognition systems under morphing attacks: A survey," IEEE Access, vol. 7, pp. 23012-23026, Feb. 2019.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110-8119, Jun. 2020.

[13] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2287-2296, Jun. 2021.

[14] J. Chen, M.I. Jordan, and M.J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," IEEE Symposium on Security and Privacy, pp. 1277-1294, May. 2020.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110-8119, Jun. 2020.

[16] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," Proceedings of the 2015 British Machine Vision Conference, pp. 41.1-41.12, Sep. 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823, Jun. 2015.

[18] J. Deng, J. Guo, N. Xue, and S.

Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690-4699, Jun. 2019.
[19] Github, "Facenet," https://github.com/davidsandberg/facenet, Aug. 2022.
[20] Y. Choi, Y. Uh, J. Yoo, and J.W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188-8197, Jun. 2020.

〈저자소개〉

이 정 재 (Jung-jae Lee) 학생회원
2019년 3월~현재: 숭실대학교 소프트웨어학부 학부과정
〈관심분야〉 AI 보안, 머신러닝

나 현 식 (Hyun-sik Na) 학생회원
2021년 2월: 공주대학교 응용수학과 학사
2021년 3월~현재: 숭실대학교 소프트웨어학부 석박사통합과정
〈관심분야〉 AI 보안, 개인정보보호, 엣지 AI

옥 도 민 (To-min Ok) 정회원
1998년 2월: 동국대학교 전자계산학과 학사
2000년 2월: 서울대학교 컴퓨터공학과 석사
2000년~2004년: ㈜ 투더베스트 기술이사
2004년~2006년: ㈜ 네오위즈 개발실장
2006년~2018년: ㈜ 레드덕 모바일 본부장
2018년: 공주대학교 전임연구원
2018년~2020년: ㈜ 크레더블럭 기술이사
2018년~현재: ㈜ 망고스틴 기술이사
2020년~현재: 숭실대학교 소프트웨어학부 박사과정
〈관심분야〉 온라인 서버 기반 기술, 플랫폼, 네트워크 보안, 머신러닝

최 대 선 (Dae-seon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, AI 보안