IMES DISCUSSION PAPER SERIES

スマートフォンによる顔認証のセキュリティ:
ディープフェイクによる脅威と対策

宇根正志（うねまさし）

Discussion Paper No. 2024-J-5

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

日本銀行金融研究所

〒103-8660 東京都中央区日本橋本石町 2-1-1

日本銀行金融研究所が刊行している論文等はホームページからダウンロードできます。
https://www.imes.boj.or.jp

IMES Discussion Paper Series 2024-J-5
2024 年 3 月

スマートフォンによる顔認証のセキュリティ：ディープフェイクによる脅威と対策

宇根正志（うねまさし）*

要　　旨

スマートフォンを使用した金融サービスにおけるオンラインでの顧客の認証方法として、スマートフォンのカメラで取得した顔の動画や静止画を用いた手法が広く使われている。被認証者となる金融機関の顧客は、自分の属性情報、それを証明するクレデンシャル、スマートフォンで撮影した顔の動画や静止画を金融機関に送信し、金融機関はこれらを用いて認証を行う。こうした認証のセキュリティを考えるうえで、近年注目されているディープフェイクの脅威について考慮する必要がある。具体的には、機械学習によって合成された顔の動画や静止画をスマートフォンのカメラに提示して本人になりすますなどの攻撃の増加が懸念される。また、一部のクラウドによって提供されている顔認証のシステムでは、合成された動画や静止画を誤って受け入れる事象が実験で観察されている。これらを踏まえると、合成された動画や静止画によるなりすましのリスクが高まっているとみられる。スマートフォンによる顔認証を行っている金融機関は、こうした攻撃によるリスクを評価したうえで、対策手法の動向をフォローし、適切に対応していく必要がある。また、リスク抑制と利便性向上の両立に向けて、合成画像の作成・検知手法を含め、オンラインでの顔画像による認証の研究の更なる進展が望まれる。

キーワード：顔認証、機械学習、スマートフォン、ディープフェイク、なりすまし、リスク
JEL classification: G21、O33

* 日本銀行金融研究所参事役（E-mail: masashi.une@boj.or.jp）

本稿は 2024 年 1 月 31 日時点の情報に基づいて作成した。本稿の作成に当たっては、大木哲史准教授（静岡大学）から有益なコメントを頂いた。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

目　　次

１．はじめに

スマートフォンはリテール金融サービスにおいて顧客との重要なチャネルの 1 つとなっている。スマートフォンを使用する代表的な金融サービスとしては、スマートフォン向けのアプリによるモバイル・バンキング、複数の金融取引の実績に関する情報を集約・管理するフィンテック・サービス、バーコードから取引情報を取得して決済するバーコード決済サービスなどが挙げられる。

こうしたサービスを金融機関やフィンテック企業が顧客に提供する際に、顧客の認証（ユーザ認証）をオンラインで実施するケースが一般的であり、スマートフォンに標準装備されているセンサーを用いて生体認証を実施することができるようになっている。例えば、スマートフォン搭載のカメラで撮影した顔の静止画や動画（以下、まとめて画像）を用いる方式や指紋センサーで取得した指紋パターンを用いる方式がよく知られている。このような方式を金融サービスで活用する事例としては、FIDO（Fast IDentity Online）[1]に準拠したユーザ認証を行うアプリにおいて生体認証の方式を選択するケースや、銀行口座開設時のオンラインでの本人確認（eKYC：electronic Know Your Customer）の際に、スマートフォンで撮影した顔や本人確認書類の画像を用いるケースが挙げられる。

顔画像を用いたユーザ認証（顔認証）に焦点を当てると、被認証者から提示される顔画像が登録済みの顔画像（テンプレート）と一致するか否かを判定するさまざまな手法が提案・実用化されてきた。近年、深層学習による画像認識の精度が著しく向上し、深層学習による手法も採用されるようになってきている（国立研究開発法人科学技術振興機構研究開発戦略センター［2023］）。

深層学習を活用して精巧な顔画像を合成する手法の研究も進んでいる。深層学習を含む機械学習によって合成された画像はディープフェイクとも呼ばれている（笹原［2023］、Gaur [2023]、Xu, Frahm, and Monrosec [2016]）。こうした合成画像のなかには、スマートフォン搭載のカメラを用いた顔認証において相応の確率で本人と誤判定されるものも存在する（Ming *et al.* [2020]）。

近年、SNS（Social Networking Service）の普及などによって個人の顔の画像が属性情報とともに公開されるケースがある。SNS から得た情報を用いて精巧な顔画像を合成できるとすれば、スマートフォンによる顔認証のセキュリティが低下し、その結果、なりすましのリスクが高まることが想定される。合成画像によるリスクの状況を評価するためには、最新の研究動向をフォローし、合成画像の脅威や顔認証システムの脆弱性を把握することがまず必要である。

こうした問題意識に基づき、本稿では、機械学習によって合成された顔画像をスマートフォンで撮影して提示する攻撃（提示攻撃〈presentation attack〉）に焦点

[1] FIDO は FIDO Alliance, Inc.の登録商標である。

を当てて、想定される攻撃手順やセキュリティ要件を導出するとともに、顔画像の合成手法や対策手法に関する研究動向を紹介し、今後の課題を考察する。

2 節では、スマートフォンによる顔認証システムの構成を示す。3 節では、提示攻撃と攻撃者に関する想定を説明し、4 節においてそれに対抗するためのセキュリティ要件を示す。5、6 節では、顔画像の主な合成手法と検知手法をそれぞれ紹介する。7 節では顔認証システムの評価実験の事例を紹介し、8 節では合成画像によるなりすましのリスクと今後の課題を考察する.

２．スマートフォンによる顔認証システムの構成

（１）エンティティ

一般的な顔認証システムのエンティティとして、サービス提供者、アプリ、サービス利用者、スマートフォンをそれぞれ以下のとおりとする。

- サービス提供者：スマートフォン用のアプリを提供し、そのアプリによってサービス利用者に特定のサービスを提供する組織（金融機関など）。
- アプリ：スマートフォンにインストールされ、スマートフォンの画面などを介してサービス利用者とやり取りしつつサービスを提供するソフトウエア。
- サービス利用者：サービス提供者の顧客（個人）。
- スマートフォン：カメラを装備し、サービス利用者の顔画像の取得、サービス利用者の属性情報の処理、サービス提供者との通信、サービス利用者とアプリとの間のインタフェースの機能を実現する端末。

また、攻撃者を、特定の顧客になりすましてサービスを不正に受けることを試みる個人または組織とする。

（２）本人確認と当人確認

顔認証は、サービス登録時に行われる本人確認と、サービスを受ける都度実行される当人確認のタイミングでそれぞれ行われると想定する[2]。

イ．サービス登録時の本人確認

サービス登録時の本人確認は、サービス提供者とサービス利用者が対面で行

[2] 本人確認（identity proofing）は、サービス提供者が特定の個人をサービス利用者として登録するか否かを決定するために既存のクレデンシャル（運転免許証やパスポートなど）を用いて個人とその属性の関係を確認するプロセスである。身元確認と呼ばれることもある。当人確認（authentication）は、サービスの利用を求める主体がサービス利用者として既に登録されているか否かをサービス提供者が確認するプロセスである。

うケースとオンラインのケースが想定される。ここでは eKYC のようにオンラインの場合を想定し、以下の流れで処理が行われるとする（図 1 を参照）。

A) サービス利用者は、自分のスマートフォンでアプリを起動し、属性情報（氏名、生年月日など）と、それを証明するクレデンシャル（属性情報が記載・格納された身分証明書など）の情報を入力する。ここで、クレデンシャルは顔画像を含むものとする。また、カメラで顔画像を撮影する。

B) アプリは、クレデンシャルの顔画像と自撮りの顔画像から、顔の特徴を表すデータ（特徴データ）をそれぞれ生成する。

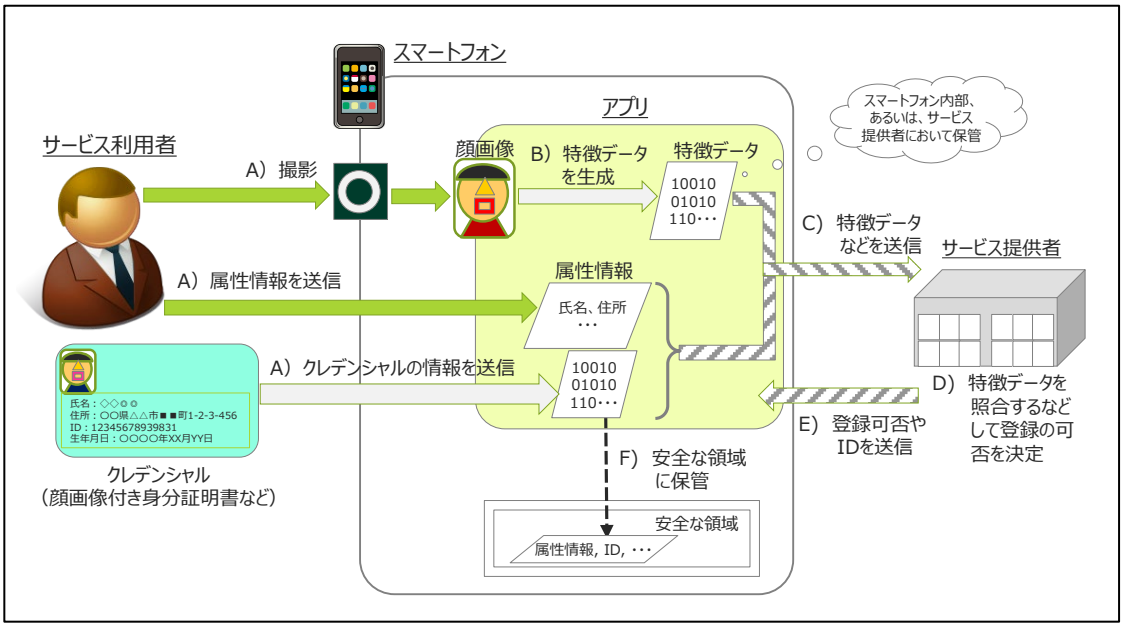
C) アプリは、特徴データを他の属性情報やクレデンシャルの情報とともにサービス提供者（または、サービス提供者から作業を受託した第三者）に送信する。

D) サービス提供者は、クレデンシャルの情報を用いて属性情報を確認するとともに、特徴データを照合して一致するか否かを確認し、それらの結果を踏まえて登録の可否を決定する。登録可の場合、サービス利用者の ID を生成する。

　ここで、当人確認の際に顔画像の照合をサービス提供者が実施するケースでは、テンプレートを ID とともに自分のデータベースに保管する。

E) サービス提供者は登録可否をアプリに送信する。登録可の場合は、サービス利用者の ID をアプリに送信する。

図 1　オンラインでの本人確認における処理のイメージ



ここで、当人確認の際に顔画像の照合をスマートフォンで実施するケースでは、サービス提供者はテンプレートもアプリに送信する。

F) アプリは、登録可の場合、登録成功をサービス利用者に通知する。また、ID、属性情報、クレデンシャルの情報をスマートフォン内部の安全な領域[3]に保管する。

当人確認の際に顔画像の照合をスマートフォンで実施するケースでは、アプリはテンプレートも安全な領域に保管する

ロ．サービス利用時の当人確認

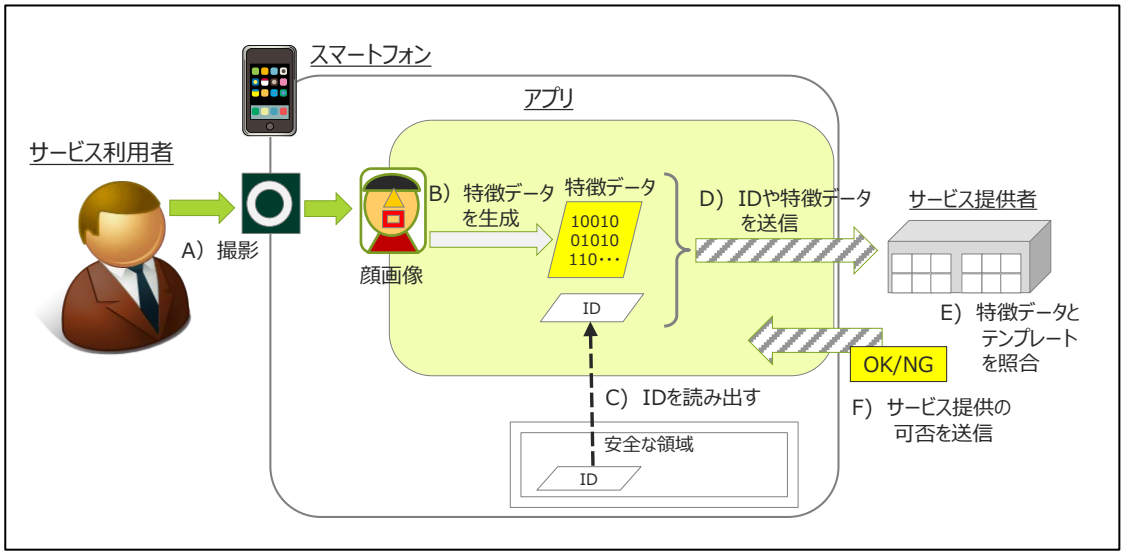
当人確認の処理フローは、顔画像の照合をサービス提供者において実施するケースとスマートフォンで実施するケースで異なる。それぞれの処理フローは以下のとおりである。

（イ）顔画像の照合をサービス提供者において実施するケース（図 2 参照）

A) サービス利用者は、アプリを起動してスマートフォンのカメラで顔画像を撮影する。

B) アプリは取得した顔画像から特徴データを生成する。

図２　当人確認の処理のイメージ：照合をサービス提供者において行う場合



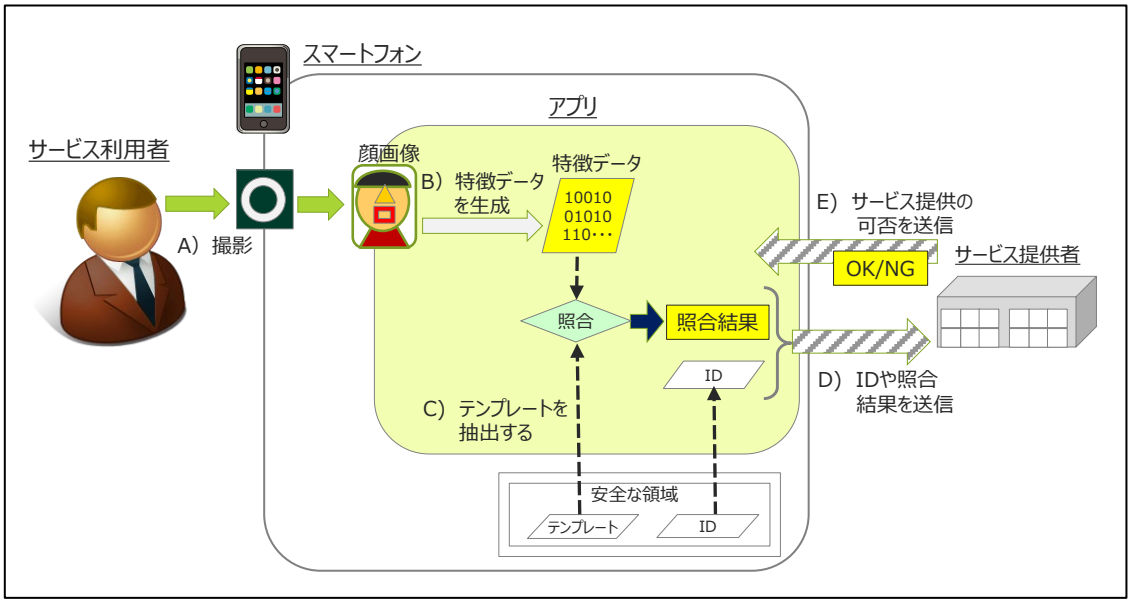
[3] こうした領域として、例えば、トラスティッド・エグゼキューション・エンバイロメント（Trusted Execution Environment）で動作するアプリのみがアクセスできるメモリー領域が挙げられる。トラスティッド・エグゼキューション・エンバイロメントは、通常の処理やデータと比べて高いセキュリティが求められるもの（例えば、暗号処理や認証処理）を取り扱うための実行環境であり、多くのスマートフォンに標準装備されている。

C) アプリは安全な領域から ID などを読み出す。
D) アプリはサービス利用者の ID や特徴データをサービス提供者に送信する。
E) サービス提供者は、サービス利用者の ID に対応するテンプレートを自分のデータベースから抽出して特徴データと照合する。
F) サービス提供者は、照合結果に基づいてサービス提供の可否を決定し、それを示すメッセージをアプリに送信する。

（ロ）顔画像の照合をスマートフォンで実施するケース（図 3 参照）

A) サービス利用者は、アプリを起動してスマートフォンのカメラで顔画像を撮影する。
B) アプリは、取得した顔画像から特徴データを生成する。
C) アプリは、安全な領域からテンプレートを抽出して特徴データと照合する。
D) アプリは、安全な領域からサービス利用者の ID などを抽出し、照合結果とともにサービス提供者に送信する [4]。
E) サービス提供者は、照合結果などからサービス提供の可否を決定し、それを示すメッセージをアプリに送信する。

図３　当人確認の処理のイメージ：照合をスマートフォンで実施する場合



[4] 照合結果をサービス提供者に送信する代わりに、アプリにおける照合が成功すると、アプリがスマートフォン内部に格納されている認証用の情報（公開鍵暗号の秘密鍵など）を用いてサービス提供者との間で認証を実行するケース（例えば、FIDO における認証）もある。

３．なりすましを目的とした攻撃

（１）攻撃の目的

本稿では、攻撃の目的として以下を想定する。

- サービス登録時のなりすまし：攻撃者が、特定の個人（被攻撃者）になりすましてサービスに登録し、金銭を得ようとしたり別の不正行為を試行したりする。例えば、銀行口座を別の個人の名義で不正に開設し、それを第三者に転売したり、資金洗浄に使用したりするケースが考えられる。
- サービス利用時のなりすまし：被攻撃者によってサービス登録が正常に完了している状況のもとで、攻撃者が、被攻撃者になりすましてサービスを悪用し、金銭を得ようとする。例えば、顔画像による当人確認を用いるオンライン・バンキングを被攻撃者が利用していた場合、攻撃者は、被攻撃者になりすまして送金を行い、被攻撃者から金銭を盗取するケースが考えられる。

（２）顔画像を提示する攻撃

攻撃者が被攻撃者になりすます方法として、スマートフォンのカメラに何らかの被写体を提示するケース（提示攻撃）に焦点を当てる。提示攻撃の実行には、被攻撃者のスマートフォンの解析や改変、サービス提供者への事前の不正行為などが不要である。そのため、攻撃実施のハードルが低く、サービス提供者がなりすまし対策を検討するうえでベースラインとなる攻撃といえる。

静止画による認証の場合、提示攻撃として、被攻撃者の写真をカメラに提示する攻撃や、被攻撃者の顔の特徴を再現したマスクを装着して頭部を提示する攻撃がよく知られている。動画による認証の場合には、被攻撃者の顔の動画を提示する攻撃が知られている。機械学習によって合成した画像をディスプレイなどに表示してカメラに提示する攻撃も提示攻撃の一種である。

（３）攻撃者の想定

現実の状況を考慮して、攻撃者に関する以下の想定を置く。

- 被攻撃者のスマートフォン（アプリをインストール済み）を攻撃実行時に一時的に使用する[5]。
- 被攻撃者のスマートフォン内部に不正な仕掛け（マルウェア・アプリのインストールなど）を行わない。また、サイドチャネル攻撃[6]をスマートフォン

[5] 例えば、攻撃者がスマートフォンを盗取するケース（被攻撃者が盗難に気づくまでの間に攻撃を実行）や、被攻撃者が眠っている間に攻撃者がスマートフォンを操作するケースが想定される。

[6] サイドチャネル攻撃は、暗号アルゴリズムなどの処理をパソコンやスマートフォンで動作させ

に仕掛けることもしない。
- 特徴データやテンプレートの生成や照合の方法を知らない（ブラックボックスの想定）。
- 被攻撃者の顔画像やサービス登録に必要な属性情報（氏名、住所など）を SNS などから入手する。
- 被攻撃者のクレデンシャルを入手できないほか、サービス提供者が本物と誤認するクレデンシャルを偽造することもできない。
- 被攻撃者のスマートフォンとサービス提供者との間の通信の盗聴・改変、サービス提供者への不正行為（サーバへの侵入など）を行うことができない。

（４）攻撃の手順

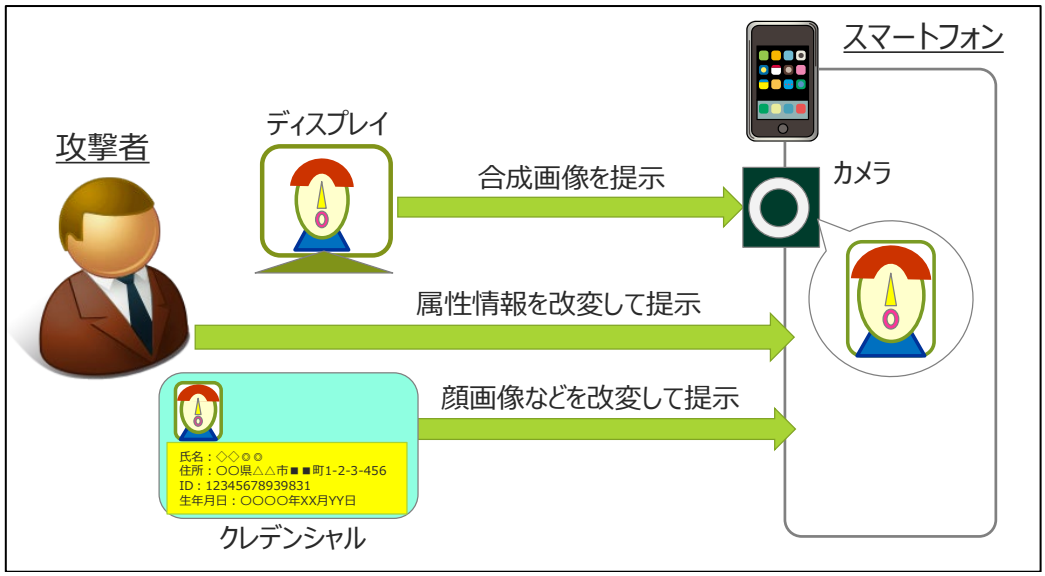
イ．サービス登録時における本人確認での攻撃

個人がサービスに登録する際には、サービス提供者と対面で本人確認を行うケースと、オンラインで行うケース（eKYC など）が想定される。

対面での本人確認の場合、攻撃者はサービス登録時にクレデンシャルをサービス提供者に提示する必要があるが、本節（3）で示した想定により、本物のクレデンシャルを提示できず攻撃は成功しないことになる。

オンラインでの本人確認では、顔画像の合成による攻撃として以下の攻撃が想定される（図 4 を参照）。

図４　サービス登録時における本人確認への攻撃のイメージ



た際に、予期せぬチャネル（サイドチャネル）から漏れる情報（消費電力パターン、処理時間パターン、漏洩電磁波パターンなど）を用いて秘密情報（パソコンなどの内部に格納されている暗号鍵など）やアルゴリズムの構造を効率的に推定するタイプの攻撃である。

- 被攻撃者または架空の個人の顔画像（合成したもの）をディスプレイに表示し、それを自分のスマートフォンのカメラで撮影する。
- 属性情報については被攻撃者のものを提示する。
- 自分のクレデンシャルの情報のうち、顔画像については、被攻撃者または架空の個人の顔画像（合成したもの）に改変して提示する。その他の情報は被攻撃者のものに改変して提示する。

ロ．サービス利用時の当人確認での攻撃

攻撃者は、被攻撃者の顔画像を合成したうえで、被攻撃者の ID やテンプレートが内部に保管されている被攻撃者のスマートフォンを一時的に盗取する。そして、そのスマートフォンのアプリを起動し、合成した顔画像をカメラで撮影する。

４．提示攻撃に対抗するためのセキュリティ要件

（１）２つのセキュリティ要件

3 節で示した本人確認と当人確認での攻撃にいずれも対抗するためには、以下のセキュリティ要件をともに満たす必要がある。

A) サービス提供者は、本人確認時に提示される属性情報やクレデンシャルの情報の改変を困難にすること、または、改変を検知すること。

B) サービス提供者は、当人確認時に、合成画像が被攻撃者のテンプレートと一致と誤判定される確率（攻撃が成功する確率）をリスク管理上許容できる水準以下とすること。

イ．セキュリティ要件 A

セキュリティ要件 A は、サービス登録時の本人確認への攻撃に関する要求事項である。

サービス提供者が、属性情報またはクレデンシャルの情報の改変を検知したならば、不審な登録とみなして登録を承認しないようにすることができる。

また、サービス提供者が、カメラに提示された合成画像と、クレデンシャルの情報の一部として提示される顔画像を照合し、攻撃であることを検知するという方法もありうる。ただし、カメラに提示された合成画像とほぼ同一のものがクレデンシャルの情報の一部として提示された場合、サービス提供者が、被攻撃者の顔画像を事前に入手しているといった特別な状況でない限り、攻撃の検知は

困難と考えられる。こうしたことから、顔画像の照合による攻撃検知については考慮しないこととする。

ロ．セキュリティ要件 B

セキュリティ要件 B は、サービス利用時の当人確認におけるセキュリティ上の要求事項である。サービス登録が適切に実施され、被攻撃者の顔画像のテンプレートが登録されているという状況のもとで、合成画像の提示による攻撃を排除するためには、合成画像とテンプレートが一致するという誤判定の確率（攻撃が成功する確率）をリスク管理上許容できる水準以下にすることが求められる。

（２）セキュリティ要件 A を満たす方法

クレデンシャルの情報を改変困難にする方法として、耐タンパー性を有する IC チップにクレデンシャルの情報を格納し、攻撃者がその情報を改変できないようにすることが考えられる。例えば、マイナンバーカードの IC チップに属性情報を格納しておき、マイナンバーカードをクレデンシャルとして使用するといった対応が当てはまる。

クレデンシャルの情報の改変を検知する方法としては、クレデンシャルの情報に（クレデンシャルの発行者による）デジタル署名を付与し本人確認時に検証することが挙げられる。

（３）セキュリティ要件 B を満たす方法

サービス提供者はリスク管理上許容できる誤判定の確率の上限を設定し、攻撃が実行されたとしても、誤判定の確率が上限を超えないような対策手法を採用することが考えられる。この場合の検討の流れとして以下が想定される。

I. なりすましがサービスに与える被害額の期待値（X）、サービス提供者が許容できる被害額の上限（Y）をそれぞれ算出し、Y÷X=Z を計算して、Z を許容できる誤判定の確率の上限とする。
   —— 被害額の期待値（X）は、例えば、一定期間における攻撃の試行回数と、1 回の攻撃成功によって生じる被害額から算出する。このうち、一定期間における攻撃の試行回数については、攻撃者となりうるサービス利用者がどの程度存在するか、また、一定期間においてサービスの登録や利用を何回試行できるかを見積もって算出することが考えられる。
II. 既知の顔画像の合成・提示手法をリストアップする。
III. サービスで使用する予定のスマートフォンやカメラ、顔画像の照合方式

を前提としたときに、リストアップした各手法による攻撃成功確率と攻撃実行に必要なコストを評価する。

—— 攻撃成功確率については、例えば、照合方式における判定しきい値などのパラメータと攻撃成功確率との関係を特定する。

IV. 上記 III の評価を考慮し、実行される可能性が高い手法を特定する。

—— 攻撃者は、実行のハードル（コスト）がなるべく低く、相応の効果（攻撃成功確率）が得られる攻撃を選択すると想定される。そこで、攻撃成功確率が相対的に高い手法を特定するとともに、コストが低い手法についても特定しておく必要がある。

V. 上記 IV で特定した手法に関して、許容できる誤判定の確率の上限よりも攻撃成功確率が小さくなるように、照合方式のパラメータを設定する。こうしたパラメータ設定が困難な場合、他の照合方式（あるいはサービス）の候補を探索し、改めて上記 III の作業から再開する。

上記 V においてパラメータを設定する際には、本人拒否の確率にも配慮する必要がある。一般に、他者を誤って本人と判定する確率を低くしようとすると、本人を他者と誤って判定する確率が高まる。サービス提供に支障が出ないように本人拒否の確率を一定水準以下とする必要がある。

５．顔画像の合成

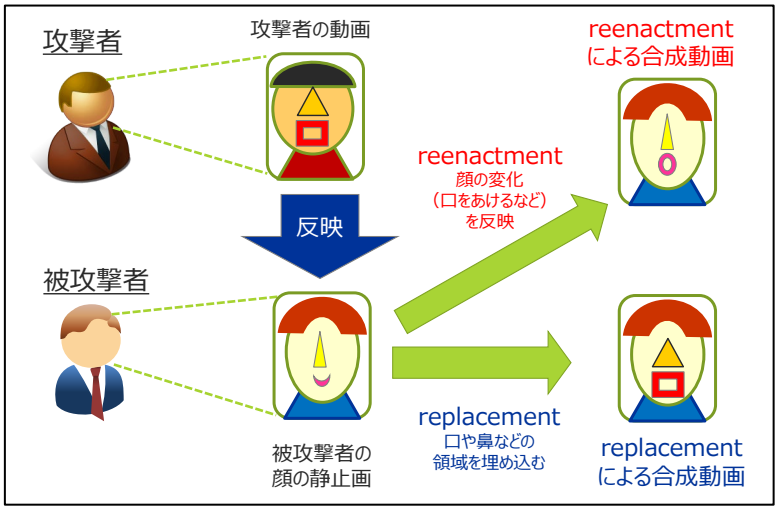
なりすましに使用することができる顔画像の合成手法は、facial reenactment（facial animation とも呼ばれる）と facial replacement（face swapping とも呼ばれる）にわけることができる（Mirsky and Lee [2020]）。

（１）facial reenactment

facial reenactment は、攻撃者の特徴点（目、口、鼻など）の移動や形状の変化を被攻撃者の顔画像に見た目が自然なかたちで反映させるものである（図 5 を参照）。例えば、瞬きによる目の変化や視線の変化、唇の動き、頷きなどによる頭部の上下左右の動きが代表例として挙げられる。

以下では、facial reenactment の主な手法の基本的なアイデアを説明する。以下で取り上げた論文による提案手法を発展させるかたちでさまざまな手法が提案されているが、ここではそれらを網羅的に取り扱うのではなく、画像合成の基本的なアイデアを提案している論文に限定して取り上げることとする。本節（2）および 6 節における説明についても同様である。

図５　reenactment と replacement（イメージ）



- Face2Face

この手法では、攻撃者は事前に被攻撃者の動画を入手する。そして、顔の向きやその輪郭（pose）、特徴点の位置や形状（expression）、顔領域の明度（illumination）などの情報（特徴量）を各フレームから抽出し、被攻撃者の顔の特徴量の分布や頻度を示す確率モデルを生成しておく。動画の合成の際には、攻撃者は自分の顔の動画を撮影し、各フレームから特徴量を抽出したうえで、その特徴量を被攻撃者の確率モデルに適用し、被攻撃者の静止画に時系列的に反映させて動画を合成する。

Thies *et al.* [2016]は、この手法を提案したうえで、市販のウェブ・カメラによって攻撃者の動画を撮影しつつ、提案手法によって被攻撃者の動画を合成した。その結果として、ほぼリアルタイムで合成動画を表示することができた旨を報告している。

- X2Face

この手法は、発言している個人（話者）の顔の動画を合成することを目的としている。まず、被攻撃者の顔の動画から当人の標準的な表情の静止画（例えば、無表情でカメラに正対した顔の静止画）を機械学習モデルによって合成する。そのうえで、攻撃者は、自分の動画の各フレームと被攻撃者の標準的な表情の静止画から、自分の顔の動きが反映された被攻撃者の動画を機械学習モデルによって合成する。動画合成に使用する 2 つのモデルを生成する際には、被攻撃者を含むさまざまな個人の顔の動画から頭部の向きや顔の特徴点の位置などを抽出して訓練データとして用いる。

Wiles, Koepke, and Zisserman [2018]は、この手法を提案したうえで、動画デー

タセット VoxCeleb からのフレーム（約 90 万件）を訓練データとして各モデルを生成し、それによって生成した合成動画を既存手法によるものと比較している。GAN を用いた手法（GAN については補論を参照）による合成動画と比較したところ、顔の特徴点の位置の正確性、フレーム間での背景や頭髪の形態の整合性などの点で提案手法が相対的に優れていたと評価している。

- Neural Talking Head Model

この手法は、X2Face と同じく、話者の顔の動画を合成することを目的としている。ただし、被攻撃者の顔の動画を用いて機械学習モデルを訓練する代わりに、被攻撃者以外のさまざまな個人の動画を用いて汎用的な事前学習済みモデルを GAN の手法によってまず生成し、その後、被攻撃者の少数の動画によって事前学習済みモデルを被攻撃者向けにカスタマイズするという方法を採用している。

事前学習では、話者（被攻撃者以外）の動画の各フレームから、頭部の形状、特徴点の位置や形状、頭部や特徴点の状態に依存しない情報（画像の色相、明度など）を訓練データとしてそれぞれ抽出し、生成モデルに入力して静止画を合成する。合成した静止画を分類モデルに入力し、出力された類似度に基づいて生成モデルをアップデートする。カスタマイズでは、被攻撃者の動画から少数のフレームを抽出して事前学習済みモデルに入力し、事前学習と同じプロセスで再学習する。再学習の結果として得られた生成モデルに攻撃者の動画の各フレームを入力して、被攻撃者の合成動画を得る。

Zakharov *et al.* [2019]は、この手法を提案したほか、話者の動画のデータセット VoxCeleb1（256 件の動画）と VoxCeleb2（224 件の動画）を用いて合成動画を生成し、X2Face などの既存手法による合成動画と比較している。人間による主観的な評価（表情の自然さや動きの整合性を評価）において既存手法より高い評価であった旨を報告している。

- FSGAN（Face Swapping GAN)

この手法は face swapping による動画合成を目的とする手法であるが、その途中段階で facial reenactment を実行している。facial reenactment を行う機械学習モデルは GAN の手法で生成される。

生成モデルは、ある個人（被攻撃者以外）の 2 つの異なるフレーム A、B が入力されたとき、A の特徴点などの情報に基づいて、B に類似した静止画を合成するように訓練される。具体的には、顔の（複数の）特徴点の位置の重心を算出し、A の重心が B の重心に近づくように静止画 A’を生成する。A’と B は分類モデルに入力され、出力された類似度に基づいて生成モデルがアップデートされる。攻撃者は、こうして得た生成モデルに被攻撃者の静止画（A）と自分の動画の各フ

レーム（B）を入力してそれぞれフレームを合成する。

Nirkin, Keller, and Hassner [2019]は、この手法を提案したうえで、顔動画のデータセット IJB-C から選択した動画（5,500 件）を訓練データとして生成モデルを準備し動画を合成した。それらを一部の既存手法（Face2Face）による合成動画と比較したところ、既存手法よりも自然な表情を再現できた場合があったものの、顔の向きによっては画像がより不鮮明になる場合があったと報告している。

- FOMM（First Order Motion Model）

この手法では、まず、顔の特徴点の位置や動きに関する情報（motion）と各領域の色相や明度などの情報（appearance）を動画の各フレームからそれぞれ抽出し、攻撃者の特徴点の動きを被攻撃者の特徴点に反映させて（motion のみからなる）フレームを構成する。そのうえで、被攻撃者のフレームの色相や明度などを追加して最終的な合成フレームを完成させる。特徴点などの抽出、特徴点の動きの予測、最終的な合成フレームの生成には、それぞれ別々の機械学習モデルが使用され、それらのモデルの生成には GAN が用いられている。

Siarohin *et al.* [2019]は、この手法を提案したうえで、VoxCeleb の動画（約 12,000 件）などを用いて機械学習モデルをそれぞれ生成し、合成した動画を既存の手法（提案チームによる手法や X2Face）によるものと比較した。その結果、特徴点の位置の正確さを平均ユークリッド距離などで比較すると、既存の手法よりも正確であったと評価している。

- ICface（Interpretable and Controllable Face)

この手法の特徴は、合成画像における頭部の動きや表情（目と鼻の動き）をそれぞれ別々に操作できるという点である。これは、攻撃者の動画のフレームから抽出した顔の特徴点（複数の要素から構成される）のうち、被攻撃者の静止画に反映したい特徴点のみを攻撃者のものに設定して機械学習モデルに入力することによって行われる。機械学習モデルは、まず、被攻撃者の静止画からニュートラルな顔（無表情で正面を向いたもの）の静止画を生成し、攻撃者が設定した特徴点の情報がその静止画に反映されるように合成する。機械学習モデルの生成には GAN の手法が用いられる。

Tripathy, Kannala, and Rahtu [2020]は、この手法を提案したうえで、VoxCeleb の動画を訓練データ（サンプル数は不明）としてモデルを生成し、複数の動画を合成した。こうした合成動画に関して、既存の手法（X2Face など）による合成動画と見た目による主観的な比較を行ったところ、攻撃者の動画の頭部や表情の変化がより正確に反映されたと評価している。

（２）facial replacement

facial replacement は、攻撃者の顔の静止画（または動画の各フレーム）における特徴点の領域を、被攻撃者の顔画像に自然な風合いで埋め込むというものである。facial replacement の主な手法としては、facial reenactment としても使用できる FSGAN や FaceShifter が挙げられる。

- FSGAN

FSGAN を facial replacement に使用する場合、facial reenactment における生成モデルを使用する。まず、攻撃者の動画のフレームと被攻撃者の静止画を生成モデルに入力し、合成フレームを得る。合成フレームは、背景、顔の形状・輪郭・特徴点が被攻撃者のもの、特徴点などの動きが攻撃者のものとなっている。そこで、この合成フレームに攻撃者の顔の特徴点を機械学習モデルによって埋め込む。機械学習モデルは GAN の手法によって生成される。

Nirkin, Keller, and Hassner [2019]は、この手法を提案したうえで、提案手法によって合成した動画を、提案者らのチームによる既存手法による合成動画と比較した。合成動画のフレームと被攻撃者の静止画の特徴点間のユークリッド距離を比較して、特徴点間の位置の正確性を評価すると、提案手法の方が短く正確であった旨を報告している。

- FaceShifter

この手法では、被攻撃者と攻撃者の動画のフレームからそれぞれ特徴点の情報を特徴量として抽出する際に、フレームの解像度を変化させながら複数の特徴量セットを取得する点が特徴である。これによって、動画のフレームからより多くの情報を抽出・利用することが可能となる。ある解像度で取得した被攻撃者と攻撃者の動画の特徴量からフレーム X を合成し、次に、別の解像度で取得した被攻撃者と攻撃者の動画の特徴量と合成フレーム X から新たな合成フレーム X'を生成する。これを一定の回数繰り返して最終的な合成フレームを生成する。特徴量の抽出やフレームの合成はそれぞれ機械学習モデルによって行い、そのモデルは GAN を用いて生成される。

Li *et al.* [2020]は、この手法を提案したうえで、複数の動画データセット（CelebA-HQ、FFHQ、VGGFace）を用いて提案手法のモデルを生成し、動画を合成して他の手法（FSGAN など）のものと比較した。その結果、顔の輪郭の形状、顔領域の色相や明度などに関して、より忠実に攻撃者の動画を再現することができたと説明している。

６．合成画像の提示への対策手法

被認証者によって提示されている被写体が生体か否かを確認する方法として、頭部や表情を変化させるように被認証者にランダムに指示するなどのチャンレンジ・レスポンスによる方法がよく知られている。もっとも、リアルタイムで動画を合成するタイプの攻撃に対しては十分な対策とならない可能性があることから、チャレンジ・レスポンス以外の手法についても考慮する必要がある。

スマートフォン搭載の RGB カメラ（深度や赤外光を使用しない、可視光による汎用カメラ）による撮影を前提とすると、合成画像の提示に対抗しうる手法が提案されている。以下では、主な手法の基本的なアイデアを説明する。

● 顔の明度の時系列変化を用いる手法

血管を流れる血液の量は心拍によって時系列的に変化する。その結果、顔の表皮の明度や色相も微妙に変化する。こうした変化を動画から読み取り、提示物が生体か否かを判定する手法が提案されている（Li *et al.* [2016]）。可視光によって顔表面の動画を撮影し、顔の特定の領域における色相（赤、緑、青など）の変化やその周期性を抽出して判定に用いる。判定に機械学習モデルを使用する手法が提案されており、実験の結果、一部の合成画像を高い確率で検知することができることが示されている（Ciftci, Demir, and Yin [2020]）。

● 顔画像の局所的な表面形状の差異を手掛りとする手法

人間の顔の表面の形状は微細で複雑な形状を有している。一方、合成画像を表示した印刷物やディスプレイの表面は比較的滑らかである。こうした表面形状の差は、カメラで撮影した顔画像の色相や明度、濃淡などにも反映される。そこで、顔画像の各領域から得られる情報を局所的な特徴量として抽出し、提示物を判定する。動画の場合、個々のフレームにおける特徴量に加えて、連続したフレーム間における特徴量の変化を手掛りとすることもできる。こうした局所的な特徴量の抽出やフレーム間での特徴量の時系列的な変化の抽出に機械学習モデルを使用する手法が提案されている（Xu, Li, and Deng [2015]）。また、正確な判定に特に寄与する顔画像上の領域を特定し、それらの領域の特徴量を判定時に重視する手法も提案されている（Yang *et al.* [2019]）。

● 頭部の三次元構造を手掛りとする手法

この手法は、提示物の奥行きや深さなどの三次元構造を用いて生体か否かを判定するものである。合成画像の提示がフラットなディスプレイによって行わ

れる場合、奥行きや深さがほとんどなく、人間の頭部の場合と明らかに異なる[7]。可視光によって顔の動画を撮影し、動画のフレームから得られた色相の変化、特徴点の位置の変化、カメラと特徴点との間のアングルとその変化などを測定する。これらの情報から提示物の三次元構造を再構成し、対象物が頭部の三次元構造を有しているか否かを判定する。三次元構造の再構成を機械学習モデルによって行う手法が提案されており、一部の合成画像を高い確率で検知可能であることが示されている（Wang *et al.* [2019]）。

- **合成画像を訓練データとして判定器を生成する手法**

真正な顔画像と合成された顔画像をそれぞれ訓練データとして使用し、両者を識別するように訓練されたモデルを用いる手法である。判定に用いる顔画像の特徴量を人間が事前に定める代わりに、機械学習によって帰納的に決定する。判定器の精度は、主に訓練データと機械学習のアルゴリズムに左右されることになる。

例えば、Yang, Lei, and Li [2014]は、合成画像の訓練データおよびテストデータとして 2 つのデータセット（CASIA〈合成画像 600 件〉と REPLAY-ATTACK〈合成画像 1200 件〉）を使用し、3 層の畳み込みニューラル・ネットワークに基づく判定器を提案した。Afchar *et al.* [2018]は、合成画像のデータセット（7,500 件）などを使用し、4 層の畳み込みニューラル・ネットワークからなる判定器を提案した。そのうえで、この判定器によって特定の合成画像を高い確率で検知可能である旨を示した。

これらのほかにも、近年、非常に多くの手法が提案されている（Khan and Dang-Nguyen [2023]）。ただし、提案手法の効果を評価する際に用いられる合成画像のデータセットや評価の基準が手法によって異なっており、横並びでの評価が難しいという課題が残されている。

このように、さまざまな観点から合成画像を検知するための手法が提案されている。ただし、いずれの対策手法も、既存の合成画像を検知する効果に関して網羅的に評価・検証されているわけではない。同じカテゴリーの複数の手法における効果の比較に加えて、異なるカテゴリー間での効果の比較も今後の課題となっているのが実情である。こうした評価面での課題や画像合成手法の研究の活発化を踏まえると、筆者が知る限り、「決め手」となる対策手法はまだ特定されていない。

---

[7] 攻撃者が、被攻撃者の顔の特徴を再現したマスクを着用した場合には対応できない可能性がある。したがって、合成画像だけでなくマスク着用などの攻撃への対策を考慮する際には、他の手法と組み合わせることが必要となる。

７．実験による顔認証システムの評価

スマートフォンのカメラによる顔認証システムを対象に、提示攻撃に近い状況において実験的に評価する試みも少数ではあるが報告されている。以下では、そうした研究 2 件の概要を紹介する。これらの実験では、本人確認時に被攻撃者のクレデンシャルを使用しており、2 節（3）における攻撃者の想定よりも攻撃者に有利な状況となっている。

（１）独自に開発した顔認証システムの評価実験

川名らは、顔画像と運転免許証（クレデンシャル）を用いた模擬 eKYC システムを評価用に構築した（川名ほか［2021］）。本人確認の際に facial reenactment の手法で合成した動画をディスプレイに表示し、それをスマートフォンのカメラで撮影して提示した際に誤って本人と判定されるか否かを実験した。

イ．本人確認の処理

本人確認の処理は以下の流れで実施された。

① 被認証者は、運転免許証の表面（顔写真付き）、裏面、側面（厚みを確認）をスマートフォンのカメラでそれぞれ撮影する。
② 被認証者は、自分の顔の静止画を正面からスマートフォンで撮影する。
③ 模擬 eKYC システムは、被認証者に対して、頭部を動かす（左を向く、右を向くなど）ように指示する。頭部の動作の内容はランダムに決められる。
④ 被認証者は、指示に従って頭部を動かし、動かした後の状態の静止画をスマートフォンで撮影する。
⑤ 模擬 eKYC システムは、運転免許証の顔写真の画像が②、④の静止画と一致するか否かを判定する。

上記②と④で撮影された顔画像における顔の領域の特定や特徴点の抽出、上記⑤における照合・判定の処理は、それぞれ、機械学習を用いたオープン・ソースのツールによって実装された。

ロ．実験の概要と結果

合成画像を提示する攻撃の実験は次の流れで行われた。

A) 攻撃者は、被攻撃者の運転免許証（本物を使用）を撮影し、模擬 eKYC システムに送信する。

B) 攻撃者は、模擬 eKYC システムによるランダムな指示に応じて自分の顔を動かし、それを撮影する。その動画から顔の特徴点の変化を抽出し、被攻撃者の顔の静止画（事前に別途取得しておく）に反映させた動画を合成してモニターに表示する。この一連の処理をリアルタイムで実施する。
C) 攻撃者は、モニター上の合成動画をスマートフォンで撮影して模擬 eKYC システムに送信する。
D) 模擬 eKYC システムは、上記 A と C で撮影された顔の静止画を照合して同一人物か否かを判定する。

合成動画の生成にはオープン・ソースのツール Avatarify（FOMM をベースとしているもの）が用いられた。

実験の結果、運転免許証の顔写真と合成動画が同一人物であると誤って判定された。この結果は、試行回数やそのうちの成功回数など、詳しい内容は論文に記載されていないものの、合成動画が顔認証のシステムにおいて現実の脅威となりうることを示したものといえる。

（２）クラウドが提供する顔認証サービスの評価実験

Li らは、スマートフォンで取得した動画と静止画をアプリがクラウドに送信して本人確認を行うケースを対象に実験を行った（Li *et al.*[2022]）。実際に運用されている 6 つの顔認証サービス（FLV：facial liveness verification）に対して合成画像を送信し、誤判定が生じるか否かを検証した。

イ．本人確認の処理

各 FLV の本人確認の手順は概ね次のとおりである。

① 被認証者は使用したいアプリを起動する。
② アプリは、被認証者の顔画像や音声をスマートフォンのカメラやマイクで取得し、FLV に送信する。
③ FLV は顔画像や音声を用いて生体検知を実施する。
④ FLV は別途取得していた顔画像と②で取得した顔画像を照合する。
⑤ FLV は照合結果をアプリに送信する。
⑥ アプリは照合結果に基づいてサービス提供の可否を決定する。

ロ．生体検知方法のバリエーション

Li らは、上記③の生体検知の形態を各 FLV について推定した（表 1 を参照。FLV の方式名は F1～F6）。その結果、1 件の FLV（表 1 の F5 の方式）が動画に

表 1　6 つの FLV で用いられる顔画像の形態

FLV の種類	本人確認で使用するデータ			
	顔の静止画 (image)	顔の動画 (video)		
		頭部の動作なし		頭部の動作あり (action)
		数字の発音なし (silence)	数字の発音あり (voice)	
F1	**対応**			
F2	**対応**			
F3	**対応**	未対応		**対応**
F4	**対応**			未対応
F5	未対応	**対応**		未対応
F6	**対応**		未対応	

（資料）Li *et al.* [2022] Table 1

よる生体検知のみを実施しており、残り 5 件は静止画による生体検知と動画による生体検知の両方を実施していた。動画による生体検知の形態には以下のバリエーションがあった。

A) 被認証者に対して頭部を動かすように指示し、その動きを撮影して検証するもの（動作のバリエーションは、瞬きをする、上下左右を向く、口を開けるなど）。表 1 の F1、F2、F3 がこれに相当する。
B) 頭部の動作を指示しないで動画を撮影して検証するもの。以下のバリエーションが存在していた。
   - B-1) スマートフォンの画面に数字を示し、被認証者にそれを発音させ、唇の動きと音声の整合性を検証するもの（提示する数字は 3〜6 桁）。F1、F2、F4、F5 の方式がこれに相当する。なお、音声から得たデータによる本人確認は実施していない。
     —— 特に、F2 と F4 の方式は、唇の動きと音声が同じタイミングで生じているか否かを検証していた（lip language detection）。
   - B-2) 被認証者による発音がないケース。F3 以外の方式が相当する。

このほか、F1 と F2 の方式は、提示された画像が合成されたものか否かを判定して結果を示す機能を有していることも判明した。なお、判定方法に関する記述は論文にはなかった。

ハ．攻撃者の想定

攻撃者に関する想定は主に次の 3 点である。

- 攻撃対象の FLV の内部情報（顔画像の照合モデルなど）を知らない（black-box setting）。
- 被攻撃者の静止画を 1 つ入手する（one-shot setting）。ただし、それを用いて新しい攻撃用モデルを生成することはできない。
- 被攻撃者の静止画からリアルタイムで（FLV のサービスがタイムアウトする前に）合成動画を生成する。

顔画像の合成は facial reenactment と facial replacement をそれぞれ用いて行われた。facial reenactment の手法として、X2Face、ICface、FSGAN、FOMM が用いられたほか、facial replacement の手法として FSGAN と FaceShifter が用いられた。音声の合成は、F2 の FLV が提供している音声合成サービスを用いて別途行われたが、その詳細について論文に記述はない。

ニ．実験結果

実験では、合成画像とレファレンス用静止画を各 FLV に対して直接送信し、生体検知や顔画像の照合を経た結果、最終的に一致していると誤判定されるか否かを検証した。川名らの実験のように、ディスプレイなどに合成画像を表示してそれをスマートフォンのカメラで撮影したわけではない。

（イ）ベースラインの実験

表 1 の F1～F6 を対象とした実験の主な結果は以下のとおりである。

- 顔の静止画を用いる FLV（5 件）は、FaceShifter による合成静止画に対して 50%以上の確率で一致と誤判定した。
- 頭部の動作と発音がともになしの動画を用いる FLV（5 件）は、FSGAN、FOMM、FaceShifter のいずれかの合成動画に対して、少なくとも約 40%の確率で一致と誤判定した。
- 発音ありの動画を用いる FLV（4 件）では、頭部の動作と発音がともになしの動画を用いる FLV に比べ、合成動画に対して一致と誤判定する確率が低くなった。もっとも、一部の方式（F1 の方式）では、誤判定の確率が約 60%と高い値となったほか、lip language detection を行う方式（F4 の方式）についても、唇の動きと音声が同期するように合成した動画に対して誤判定の確率が約 60%となった。
- 頭部の動作がありの動画を用いる FLV（F2 と F3〈F1 は実験時に動作せず対象外〉）では、頭部の動作のバリエーションにそれぞれ対応する動画を合成

した。そして、FLV によって動作の指示があったところで、それに対応する合成動画を FLV に送信した。その結果、FOMM や FaceShifter による合成動画に対して、相応の確率（最大で約 80%）で一致と誤判定した。頭部の動作と音声がともになしの動画を用いるケースと比較すると、頭部の動作のバリエーションが誤判定の確率に与える影響はほとんどなかったといえる。

上記の実験で用いられた合成画像は、1 つの合成手法のみによってそれぞれ生成されていた。Li らは、複数の手法を組み合わせて画像を合成すれば誤判定の確率が高まる可能性があると考察している。

（ロ）別のクラウドの顔認証サービスを対象とする実験

Li らの実験結果は生体検知などが有効に機能しているとは言い難い状況を示したといえる。この結果が他の FLV においても生じるか否かを検証するために、Li らは、別のクラウドの顔認証サービスを選択し、同様の実験を追加的に行った。主な結果は次のとおりである。

- 上記のベースラインの実験と同じく、FLV（4 件）に対して合成画像を直接送信する実験では、いずれの FLV も合成画像に対して 50%以上の確率で一致と誤判定した。
- これらの FLV を実際に使用している組織 4 社（フィンテック会社、航空会社、生命保険会社、政府系機関）の各アプリをスマートフォンにインストールしてユーザ登録を行った後、合成動画を（カメラ撮影ではなく）アプリに挿入して当人確認を実施したところ、いずれの FLV も誤判定が発生した（誤判定の確率は論文に記載がない）。
- 本人確認のデモ用アプリを提供している FLV（4 件）を対象に、デモ用アプリをスマートフォンにインストールし、合成画像を（カメラで撮影するのではなく）デモ用アプリに挿入したところ、3 件のサービスにおいて誤判定が発生した（誤判定の確率は論文に記載がない）。

追加の実験で用いられた画像の合成手法などの詳細は論文に記載されていないものの、この結果は、当初の実験で対象となった顔認証サービス以外のサービスでも合成画像に対して脆弱なものが存在することを示したといえる。

８．考察

（１）なりすましリスクの高まり

合成画像の検知手法について決定打となるような手法が確立していないなか、

7 節で紹介した実験研究で示されたように、高度な合成画像を用いた提示攻撃をオープン・ソース・ツールなどによって実現することが可能になってきている。また、昨今の機械学習に関する研究開発の進展により、そうした合成画像を生成するためのスキル、時間、費用は、今後も低下していく可能性が高い。これらを踏まえると、合成画像によるなりすましのリスクが高まっていると考えられる。

スマートフォンを用いたオンラインでの顔認証システムを提供するベンダーや、そのシステムを使用して顧客の本人確認などを行う金融機関などのユーザ組織は、関連する研究動向をフォローしつつ、合成画像を用いた提示攻撃によるなりすましのリスクを再評価する必要があろう。そのうえで、リスクが許容できるレベルを超えていると判断する場合には、そのリスクに適切に対応することが求められる。

（２）高まるリスクへの対策の検討

なりすましのリスクを再評価した結果、リスクを低減させる必要がある場合には、顧客の利便性に配慮しつつ、追加的な対応を検討することになる。4 節(3)で示したような手順で、より効果的な合成画像の検知手法を選定・採用するという対応が考えられるものの、既存の検知手法を横並びで評価・比較する方法や枠組みが確立していないことから、現時点で適切な検知手法を選択することは容易でない。

このため、既に効果が評価されている他の認証手段を追加して使用する、顔認証が成功した際に顧客に提供するサービスの内容を制限する（取引可能な金額の上限を引き下げるなど）といった対応策が考えられる。

（３）対策手法の評価

イ．評価の枠組みの統一化

合成画像によるなりすましのリスクを軽減させる方法として、合成画像を検知する効果的な手法の開発が今後期待される。検知手法の提案に際して、手法の評価に使用されたデータセット（モデルの訓練・テストデータ、評価用の合成画像など）が統一されているわけではなく、それぞれの提案者が選択して評価を行うケースが多い。また、データセットなどを統一化した評価の枠組みも確立していない。こうしたなかで、ユーザ組織が複数の検知手法を適切に比較することは困難である。

今後、検知手法の提案においては、他の手法の評価結果と比較できるように、共通のデータセットを用いて共通の基準で評価した結果が示されることが望まれる。また、統一化された評価の枠組みを確立するための研究も重要であるといえる。

この点に関して、最近、さまざまな機械学習ベースの検知手法の効果を同一の枠組みで評価する試みとして、例えば、Khan and Dang-Nguyen [2023]、Yan *et al.* [2023]、Le *et al.* [2024]といった成果が発表されている。こうした研究成果が蓄積され、合成画像の検知手法を適切に選択できるようになることが期待される。

ロ．実際の環境を想定した実験

実験研究に関して、 Li らの実験では、ディスプレイに表示した合成画像をスマートフォンのカメラで撮影して照合していない。通常、カメラで撮影した画像は元の画像よりも劣化する。この点を考慮すると、Li らが実験で使用した合成画像をスマートフォンのカメラで撮影して提示攻撃を行ったならば、誤判定の確率が実験の値よりも低い値となっていた可能性がある。提示攻撃の評価をなるべく厳密に行ううえで、合成画像をカメラで撮影して照合するなど、実際の攻撃が行われる環境に近い状態を再現して実験を行うことが望まれる。

９．おわりに

スマートフォンは、金融サービスにおける有望なチャネルとして浸透している。こうしたスマートフォンによる金融サービスを今後さらに充実させていくためには、オンラインでの認証のセキュリティを維持・向上させていくことが必要である。

本稿では、スマートフォンのカメラで撮影した顔の動画や静止画による認証システムに焦点を当て、機械学習による合成画像を用いた提示攻撃の手順やセキュリティ要件を示した。また、合成画像の生成手法と検知手法の研究動向を紹介したほか、実際の顔認証システムにおけるセキュリティを実験によって評価した研究の事例を紹介した。

これらを踏まえると、合成画像によるなりすましのリスクが高まっているとみられる。顔画像による認証を使用している金融機関をはじめとするユーザ組織は、リスクの高まりにどのように対応するかを検討する必要がある。現状では、合成画像を検知する手法に関して、複数の手法を横並びで評価することが容易でなく、適切な手法を選択することが難しい。したがって、リスクを軽減する方法として、他の認証手段を活用する、顔画像認証によって実行できる取引を制限するといった手段が当面の候補となるが、リスク抑制と利便性向上の両立に向けて、合成画像の作成・検知手法を含め、オンラインでの顔画像による認証の研究の更なる進展が望まれる。

参考文献

川名のん・長沼　健・吉野雅之・太田原千秋・冨樫由美子・笹　晋也・山本恭平、「Deepfake を用いた e-KYC に対するなりすまし攻撃と対策の検討」、第 35 回人工知能学会全国大会論文誌、1F2-GS-10a-02、人工知能学会、2021 年、1～4 頁

国立研究開発法人科学技術振興機構研究開発戦略センター、「人工知能研究の新潮流 2～基盤モデル・生成 AI のインパクト～」、CRDS-FY2023-RR-02、国立研究開発法人科学技術振興機構、2023 年

笹原和俊、『ディープフェイクの衝撃：AI 技術がもたらす破壊と創造』、PHP 研究所、2023 年

Afchar, Daruis, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “MesoNet: A Compact Facial Video Forgery Detection Network,” Proceedings of 2018 IEEE International Workshop on Information Forensics and Security, IEEE, 2018, pp. 1-7.

Ciftci, Umur Aybars, İlke Demir, and Lijun Yin, “FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals,” arXiv:1901.02212v3, 2020.

Gaur, Loveleen, eds., *DeepFakes: Creation, Detection, and Impact*, CRC Press, 2023.

Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen, “Deepfake Detection: A Comparative Analysis,” arXiv:2308.03471v1, 2023.

Le, Binh M., Jiwon Kim, Shahroz Tariq, Kristen Moore, Alsharif Abuadbba, and Simon S. Woo, “SoK: Facial Deepfake Detectors,” arXiv:2401.04364v1, 2024.

Li, Lingzhi, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, “FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping,” arXiv:1912.13457v3, 2020.

Li, Xiaobai, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen, “Generalized Face Anti-Spoofing by Detecting Pulse from Face Videos,” Proceeding of the 2016 23rd International Conference on Pattern Recognition, IAPR, 2016, pp. 4244-4249.

Li, Changjiang, Li Wang, Shouling Ji, Xuhong Zhang, Zhaohan Xi, Shanqing Guo, and Ting Wang, “Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era,” Proceedings of the 31st USENIX Security Symposium, USENIX Association, 2022, pp. 2673-2690.

Ming, Zuheng, Muriel Visani, Muhammad Muzzamil Luqman, and Jean-Christophe Burie, “A Survey on Anti-Spoofing Methods for Facial Recognition with RGB Cameras of Generic Consumer Devices,” *Journal of Imaging*, 6(12), 139, 2020, pp. 1-56.

Mirsky, Yisroel, and Wenke Lee, “The Creation and Detection of Deepfakes: A Survey,” arXiv:2004.11138v3, 2020.

Nirkin, Yuval, Yosi Keller, and Tal Hassner, “FSGAN: Subject Agnostic Face Swapping and Reenactment,” Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, IEEE, 2019, pp. 7184-7193.

Siarohin, Aliaksandr, Stephane Lathuiliere, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, “First Order Motion Model for Image Animation,” Proceedings of 33rd Conference on Neural Information Processing Systems, Curran Associates, Inc., 2019, pp. 7135-7145.

Thies, Justus, Michael Zollhöfer, Marc Staminger, Christian Theobalt, and Matthias Nießner, “Face2Face: Real-time Face Capture and Reenactment of RGB Videos,” Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2387-2395.

Tripathy, Soumya, Juho Kannala, and Esa Rahtu, “ICface: Interpretable and Controllable Face Reenactment Using GANs,” Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2020, pp. 3385-3394.

Wang, Zezheng, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, Guojun Qi, Jun Wan, and Zhen Lei, “Exploiting Temporal and Depth Information for Multi-Frame Face Anti-Spoofing,” arXiv:1811.05118v3, 2019.

Wiles, Olivia, A. Sophia Koepke, and Andrew Zisserman, “X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes,” Proceedings of the 15th European Conference on Computer Vision, LNCS 11217, Springer, 2018, pp. 690-706.

Xu, Yi, Jan-Michael Frahm, and Fabian Monrosec, “Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos,” Proceedings of the 25th USENIX Security Symposium, USENIX Association, 2016, pp. 497-512.

Xu, Zhenqi, Shan Li, and Weihong Deng, “Learning Temporal Features Using LSTM-CNN Architecture for Face Anti-Spoofing,” Proceedings of 2015 3rd IAPR Asian Conference on Pattern Recognition, IAPR, 2015, pp. 141-145.

Yan, Zhiyuan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu, “DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection,” arXiv:2307.01426v2, 2023.

Yang, Jianwei, Zhen Lei, and Z. Li, “Learn Convolutional Neural Network for Face Anti-Spoofing,” arXiv:1408.5601v2, 2014.

Yang, Xiao, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifen Li, and Wei Liu, “Face Anti-Spoofing: Model Matters, So Does Data,” Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 3507-3516.

Zakharov, Egor, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, “Few-Shot Adversarial Learning of Realistic Neural Talking Head Models,” Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, IEEE, 2019, pp. 9458-9467.

補論．GAN による画像合成用モデルの生成

GAN（Generative Adversarial Network）は、画像などを生成するモデル（生成モデル〈generator〉）と、生成されたデータと参照データとを照合・識別するモデル（分類モデル〈discriminator〉）をそれぞれ準備し、2 つのモデルを競争させるようにして双方のモデルを同時に学習させる手法である。

GAN は facial reenactment の手法において使用されることが多い。顔画像を合成する生成モデルを GAN によって生成する際の基本的な処理の流れは次のとおりである（図 A1 を参照）。

① 生成モデルは、被攻撃者の顔の静止画と攻撃者の顔の動画を基に合成動画を生成する。
② 生成モデルは合成動画を分類モデルに渡す。
③ 分類モデルは、被攻撃者の顔の静止画を参考にしながら、生成モデルが出力した顔の動画が被攻撃者の顔と一致するか否かを判定して確信度を出力する。
④ 分類モデルは判定結果を生成モデルにフィードバックする。
⑤ 生成モデルは、分類モデルの判定結果に基づいて自身のパラメータを更新する。
⑥ 上記の①～⑤のプロセスが一定の条件（例えば、分類モデルが 95%以上の確率〈確信度〉で本物と判定するまで）が満たされるまで繰り返され、最終的に得られた生成モデルを顔画像合成用の学習済みモデルとする。

図 A1 GAN による顔画像合成用モデルの生成の流れ（イメージ）