Sorbonne Université

EDITE – ED 130

EURECOM / Sécurité numérique

Robustesse aux Attaques en Authentification Digitale par Apprentissage Profond

Par **Anis TRABELSI**

Thèse de doctorat - **« Images et Vision »**

Dirigée par **Jean-Luc DUGELAY**

Co-encadrée par **Marc PIC**

Présentée et soutenue publiquement le **28 novembre 2022**

Devant un jury composé de :

M. Florent RETRAINT	PROFESSEUR	Président
M. Vincent NOZICK	MAÎTRE DE CONFERENCES - HDR	Rapporteur
M. Touradj EBRAHIMI	PROFESSEUR	Rapporteur
Mme Antitza DANTCHEVA	CHARGÉE DE RECHERCHE	Examinateur
M. Abdenour HADID	PROFESSEUR	Examinateur
M. Jean-Luc DUGELAY	PROFESSEUR	Directeur de Thèse
M. Marc PIC	DOCTEUR	Encadrant

À mes grands-parents, Mes parents, Ma femme.

Remerciements

Je tiens à remercier toute l'équipe de doctorants et postdoctorants du Prof. Jean-Luc DUGELAY pour leur accueil et d'avoir partagé des conseils précieux pour ma thèse. Je remercie également toute l'équipe digitale au sein de SURYS qui m'a accueilli. Ce fut un réel plaisir de travailler à leur côté au sein de SURYS. Je remercie particulièrement Amine OUDDAN, Keltoum OULAHOUM et Gaël MAHFOUDI pour m'avoir conseillé et aidé dans mes travaux.

J'ai beaucoup appris auprès de Marc PIC, mon encadrant et de Jean-Luc DUGELAY, mon directeur de thèse. En 2017, j'ai eu la chance d'effectuer mon stage au sein de SURYS. Marc PIC, m'a ensuite offert l'opportunité de réaliser cette thèse. Je le remercie chaleureusement pour sa confiance, son soutien et pour tous les conseils qu'il m'a prodigués. Je remercie Jean-Luc DUGELAY, pour son suivi durant cette période. Son expérience et ses enseignements ont été d'une grande aide pour moi.

Je voudrais enfin remercier Vincent NOZICK et Touradj EBRAHIMI d'avoir accepté d'être rapporteurs. Merci aussi à Florent RETRAINT, Antitza DANTCHEVA et Abdenour HADID d'avoir bien voulu faire partie des examinateurs.

Avis au lecteur

Cette thèse est rédigée en français, cependant certains termes présents dans le manuscrit sont difficiles à traduire ou rendent la lecture plus difficile. Par conséquent, nous avons choisi de ne pas donner la traduction française de certains mots techniques anglais. Par exemple, deux traductions françaises du mot *deepfake* ont été proposées. La Commission d'enrichissement de la langue française a proposé de traduire *deepfake* par vidéotox alors que le Grand Dictionnaire terminologique québécois préfère utiliser le terme hypertrucage. À notre sens, ces deux traductions alourdissent la lecture et nous préférons conserver le terme original. Lorsque c'est le cas, ces mots sont indiqués en italique.

Certaines des illustrations donnent un aperçu d'une vidéo. Il est difficile de décrire une vidéo avec une seule illustration. Pour remédier à ce problème, nous accompagnons les illustrations représentant une vidéo avec un QR Code qui renvoie à ladite vidéo.

Sommaire

Table des illustrations

Table des tableaux

Table des acronymes

ANN	Artificial Neural Network
ANSSI	Agence nationale de la sécurité des systèmes d'information
AUC	Area Under (ROC) Curve
CFA	Color Filter Array
CNN	Convolutional Neural Networks
DCT	Discrete Cosine Transform
DFDC	DeepFake Detection Challenge
eIDAS	Electronic Identification, Authentification and Trust Services
E-KYC	Electronic Know Your Customer
F2F	Face2Face
FAR	False Acceptance Rate
FRR	False Rejection Rate
GAN	Generative Adversarial Network
HOG	Histogram of Oriented Gradients
ICAO	International Civil Aviation Organization
ISO	International Organization for Standardisation
JPEG	Joint Photographic Experts Group
KYC	Know Your Customer
LBP	Local Binary Patterns
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
OACI	Organisation de l'aviation civile internationale
OCR	Optical Character Recognition

PRNU Photo Response Non-Uniformity

PVID Prestataires de vérification d'identité à distance

RNN Recurrent Neural Network

ROC Receiver Operating Characteristics

SVM Support Vector Machine

VAE Variational Auto-Encoder

1. Introduction

1.1 Préambule

L'identification et l'authentification des personnes sont deux processus importants dans un système de contrôle d'accès physique ou numérique.

L'identification est l'action d'attribuer une identité à une personne. C'est un élément clé de l'interaction sociale humaine, car elle permet aux gens de se différencier et d'interagir les uns avec les autres.

L'authentification est le procédé qui consiste à vérifier l'identité d'une personne. L'authentification est importante pour s'assurer que les personnes sont bien celles qu'elles prétendent être et pour empêcher tout accès non autorisé dans un système.

Depuis les premières civilisations, les populations ont eu besoin de distinguer et d'authentifier les autres personnes. Au sein des premières sociétés, ce besoin était souvent réalisé par le recours à des signes distinctifs ou à d'autres indicateurs physiques. Au fur et à mesure que les civilisations se sont développées, ces indicateurs n'étaient plus suffisants et les documents écrits sont devenus plus courants. L'authentification a évolué pour inclure des éléments de preuves écrites comme les sceaux et les signatures. Aujourd'hui, ces éléments de preuves écrites reposent sur des documents d'identités comme la carte d'identité ou le passeport.

L'authentification est devenue un enjeu majeur sécuritaire à la suite des attentats du 11 septembre 2001. Depuis cette date, les nations ont investi vers les systèmes biométriques pour remplir les tâches d'authentification.

La biométrie fait référence à la mesure et à l'analyse de caractéristiques physiques ou comportementales utilisées à des fins de vérification d'identité. Les caractéristiques biométriques les plus courantes sont les empreintes digitales, l'iris et la reconnaissance faciale. Ces caractéristiques sont très stables dans la vie d'une personne, il ne change pas ou peu au cours du temps. De plus, les caractéristiques biométriques sont uniques à chacun.

Le secteur de la biométrie était déjà en pleine croissance avant la pandémie de COVID-19, mais l'apparition du virus a accéléré son déploiement et en particulier le développement des

systèmes d'authentification à distance basée sur les mesures *Know Your Customer* (KYC). *Know Your Customer* est un processus utilisé par les institutions financières ainsi que d'autres secteurs pour vérifier l'identité de leurs clients. Lorsque ce processus est réalisé à distance, on parle d'*electronic Know Your Customer* (e-KYC).

L'e-KYC repose le plus souvent sur deux preuves pour authentifier une personne. D'une part une photo ou une vidéo d'une pièce d'identité et d'autre part une photo ou une vidéo du visage de l'utilisateur. Une comparaison est ensuite effectuée entre la photo présente sur le document d'identité et le visage de l'utilisateur. Grâce au progrès des algorithmes biométriques et de la miniaturisation des capteurs, comme les capteurs photographiques, l'authentification à distance est le plus souvent réalisée sur des smartphones.

De nombreux secteurs d'activité se tournent vers des solutions d'authentification à distance. Des institutions bancaires utilisent l'e-KYC pour permettre à des clients d'ouvrir de nouveaux comptes ou traiter les demandes de prêt, des organismes gouvernementaux emploient l'e-KYC pour le renouvellement des pièces d'identité, etc.

Cependant, cette augmentation de l'utilisation de l'e-KYC s'accompagne d'une multiplication des risques de sécurité. L'une des plus grandes menaces de l'e-KYC est l'usurpation d'identité.

1.2 Problématique

Dans un système e-KYC, l'usurpation d'identité peut se réaliser de plusieurs façons. Par exemple, un fraudeur peut utiliser un faux document d'identité ou falsifier des parties du document d'identité. Un autre moyen courant d'usurpation d'identité consiste à présenter une preuve biométrique falsifiée au système. Dans le cas d'un système e-KYC, cette preuve biométrique est le visage de l'utilisateur. Donc il peut s'agir, par exemple, d'une photo imprimée d'un visage ou d'une image numérique falsifiée.

Les systèmes d'authentification à distance sont équipés d'algorithmes capables de distinguer les vrais visages et les vrais documents des faux avec un haut degré de précision. En plus de ces algorithmes, il est souvent demandé à l'utilisateur d'effectuer une action spécifique, comme cligner des yeux ou sourire. Ces actions font partie d'un processus appelé test de *liveness* permettant de s'assurer que l'utilisateur est authentique.

Les fraudeurs peuvent avoir beaucoup de difficultés à imiter les opérations d'un test de *liveness*. Cependant, le développement des techniques génératives basées sur l'apprentissage profond a permis de mettre au point des méthodes permettant de synthétiser ou falsifier des images et des vidéos, capables de tromper aussi bien les humains que les systèmes d'authentification.

Un exemple récent est la création des *deepfake*. Les *deepfakes* sont des fausses vidéos réalistes qui sont fabriquées à l'aide d'algorithmes d'apprentissage profond. Le plus souvent la falsification permet d'échanger un visage dans une vidéo de façon très réaliste. Au départ, il était très coûteux et très long de réaliser un *deepfake* de quelques secondes. Cependant la technologie s'est grandement améliorée. Aujourd'hui il est possible de réaliser un *deepfake* en temps réel. De plus, la technologie est implémentée dans plusieurs applications sur smartphone qui permettent à n'importe qui de créer un *deepfake* sans connaissance technique.

Ces technologies sont aujourd'hui considérées comme une grande menace contre les systèmes d'eKYC. Par conséquent, il est essentiel de développer de nouvelles méthodes pour détecter ce type de contenu, tout en considérant les anciennes attaques. L'objectif de cette thèse est d'étudier les vulnérabilités des systèmes d'authentification à distance et de proposer des méthodes basées sur l'apprentissage profond pour rendre les systèmes plus robustes.

1.3 Contributions

Articles de conférence

- M. Pic, G. Mahfoudi et A. Trabelsi, "*Remote KYC: Attacks and Counter-Measures*" 2019 European Intelligence and Security Informatics Conference (EISIC), 2019.
- A. Trabelsi, M. Pic et J-L. Dugelay, "*Improving Deepfake Detection by Mixing Top Solutions of the DFDC*", 2022 30th European Signal Processing Conference (EUSIPCO). 2022.
- A. Trabelsi, M. Pic et J-L. Dugelay, "*OLED vs. LCD Recaptured Image Detection*", 2022 4th International Conference on Video, Signal and Image Processing (VSIP), 2022. [UNDER REVIEW].

Chapitre de livre

- M. Pic, G. Mahfoudi, A. Trabelsi et J-L. Dugelay, "*Face Manipulation Detection in Remote Operational Systems*". Handbook of Digital Face Manipulation and Detection. Ed. by C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, C. Busch, Springer Cham, 2022.

Brevet

- A. Trabelsi, G. Mahfoudi et M. Pic. "*Method for automatically detecting facial impersonation*". Pat. WO2020099400. 2020.

Présentations

- J-L. Dugelay et A. Trabelsi, "*Malevolent Schemes in Face Recognition Applications*", 2019 16th Int.l Summer School for Advanced Studies on Biometrics for Secure Authentication: BIOMETRICS AND FORENSIC SCIENCE IN THE DEEP LEARNING ERA, 2019. (Donnée par J-L. Dugelay).
- J-L. Dugelay et A. Trabelsi, "*Malicious Facial Image Processing and Counter-Measures: A review.*", 2019 GDR-ISIS (Groupement de Recherche Information Signal Image viSion). 2019. (Donnée par J-L. Dugelay).
- A. Trabelsi, M. Pic et J-L. Dugelay, "*Improving Deepfake Detection by Mixing Top Solutions of the DFDC*", 2022 GDR-ISIS (Groupement de Recherche Information Signal Image viSion). 2022.
- A. Trabelsi, M. Pic et J-L. Dugelay, "*Recapture Detection to fight Deep Identity Theft*", 2022 Optical & Digital Document Security (ODDS). 2022. (Donnée par M. Pic).

Communications

- A. Trabelsi, M. Pic et J-L. Dugelay, "*A countermeasure against new attacks on facial recognition systems*", 2020 EURECOM Scientific Council, 2020.
- J-L. Dugelay, A. Trabelsi et S. Husseini, "*Deepfake*", World AI Cannes Festival (WAICF), 2022.

1.4 Organisation

Cette thèse est organisée en six chapitres. Nous étudions les attaques auxquelles les systèmes d'authentification à distance sont vulnérables ainsi que les méthodes automatisées basées sur l'apprentissage profond pour les détecter.

Dans le **chapitre 1 - "Introduction"**, nous introduisons le sujet en expliquant le contexte et les motivations de la thèse. Dans un deuxième temps, nous présentons les contributions scientifiques qui ont été réalisées au cours des travaux de cette thèse.

Dans le **chapitre 2 - "Positionnement"**, nous dressons un examen détaillé de l'état actuel de la littérature dans le domaine de la criminalistique des images et des vidéos numériques appliquées au système d'authentification basé sur la reconnaissance faciale. Pour mieux comprendre les enjeux actuels, nous commençons par présenter l'histoire et l'évolution de l'identification et de l'authentification des personnes. Ensuite, nous expliquons le principe de l'authentification à distance qui représente notre principal sujet d'étude. Enfin, nous détaillons les deux catégories d'attaques existantes (physiques et numériques) contre ces systèmes ainsi que les méthodes de détection disponibles.

Dans le **chapitre 3 - "Robustesse aux attaques par présentations"**, nous abordons la première catégorie d'attaques contre un système d'authentification à distance : les attaques physiques. Nous présentons notre méthode de détection de ce type d'attaque et les résultats obtenus lors de nos différentes expérimentations.

Dans le **chapitre 4 - "Robustesse aux attaques par recaptures"**, nous nous intéressons au cas particulier des attaques physiques contre un système e-KYC basé sur l'utilisation de la présentation vidéo. Après avoir expliqué le contexte et les motivations des attaques par présentation vidéo, nous présentons une nouvelle base de données d'images recapturées sur différents types d'écrans. Nous expliquons ensuite notre méthode de détection des images recapturées basée sur l'apprentissage profond.

Dans le **chapitre 5 - "Robustesse aux attaques par *deepfakes*"**, nous proposons une analyse des meilleures solutions de détection du *DeepFake Detection Challenge* (DFDC). L'objectif de cette compétition était d'obtenir des méthodes généralisables pour détecter les *deepfakes*. Pour ce faire, l'une des plus grandes bases de données de *deepfakes* a été proposée.

Aucune des méthodes proposées n'a pu être généralisable. Nous avons donc étudié l'assemblage et la complémentarité de ces solutions. Les résultats que nous avons obtenus démontrent qu'un assemblage judicieux peut améliorer à la fois les résultats et la capacité de généralisation des méthodes.

Enfin, dans le **chapitre 6 - "Conclusion"**, nous concluons cette thèse et nous abordons les perspectives de recherche futures.

2. Positionnement

Ce premier chapitre fournit tous les éléments nécessaires pour comprendre les problèmes de sécurité d'un système de vérification d'identité à distance. Ce chapitre est divisé en 5 parties.

Dans la première partie, nous apportons des éléments historiques sur l'identification et l'authentification des personnes. Dans une deuxième partie, nous expliquons le principe de fonctionnement des systèmes biométriques, plus particulièrement la reconnaissance faciale et son utilisation dans les systèmes de vérification d'identité à distance. Dans la troisième partie, nous dressons un état de l'art sur les menaces existantes contre un système de vérification d'identité à distance. Dans la quatrième partie, nous présentons les méthodes permettant de détecter et de développer des contre-mesures contre ces menaces. Enfin, la cinquième et dernière partie décrit l'apprentissage profond, son importance et ses usages dans le cadre de notre sujet.

2.1 Historique

La reconnaissance des personnes est devenue un processus courant dans nos sociétés modernes. Aujourd'hui, elle repose principalement sur les systèmes biométriques et la présentation de documents d'identité. Bien que les systèmes actuels d'identification et d'authentification exploitent des technologies très avancées, ces concepts ne sont pas nouveaux. Il est intéressant de dresser un aperçu historique des fondements de l'identification pour comprendre les enjeux actuels et comment les besoins ont évolué au fil du temps.

2.1.1 Reconnaissance interpersonnelle

L'identité est une notion que nous partageons tous, mais qui est souvent difficile à définir. Toutefois, il est admis que l'identité est considérée comme une composition de notre apparence physique, de notre personnalité et de nos expériences. C'est ce qui nous rend uniques et nous distingue des autres. Les êtres humains ont par nature une capacité à reconnaître l'identité des autres personnes due à notre système visuel très développé, qui nous permet de percevoir des indices physiques subtils sur chaque individu. Reconnaître l'identité d'une personne est le sens du terme identification. L'identification visuelle des personnes basée uniquement sur des caractéristiques physiques est la forme la plus ancienne d'identification. Il

existe plusieurs caractéristiques physiques essentielles qui sont utilisées pour identifier une personne. Parmi ces caractéristiques nous pouvons citer la taille, la corpulence, la couleur des yeux, la couleur des cheveux et les traits du visage. Ces caractéristiques évoluent au cours du temps et ne sont donc pas les plus pertinentes.

Certaines caractéristiques physiques ne changent pas au cours d'une vie et constituent des identifiants physiques uniques. Il s'agit de signe inscrit sur la peau de certaines personnes. On peut citer les cicatrices et les tâches de naissances. Prenons l'exemple de l'Odyssée d'Homère qui raconte l'épopée mythologique d'Ulysse. Après un long périple de plusieurs années, Ulysse retourne à Ithaque déguisé en mendiant afin de cacher son identité. Ulysse utilise ce déguisement pour obtenir des informations sur la situation de son royaume et pour préparer sa vengeance contre les prétendants qui ont courtisé sa femme, Pénélope. Euryclée, la nourrice d'Ulysse durant son enfance, parvient à reconnaître Ulysse grâce à la cicatrice qu'il porte à la jambe (Figure 1). Ulysse avait reçu cette cicatrice d'une blessure infligée par un sanglier alors qu'il chassait dans sa jeunesse.

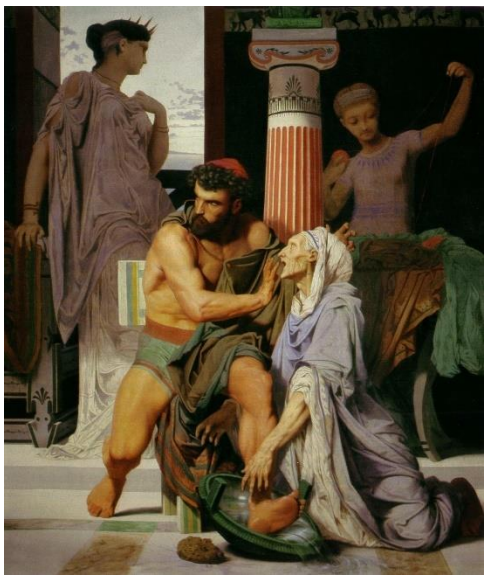

Figure 1 : Ulysse est identifié par Euryclée grâce à sa cicatrice.
Source : Gustave Boulanger (1849)

Cependant, très peu d'individus possèdent un trait physique unique et une identification basée uniquement sur des caractéristiques physiques qui évoluent avec le temps est insuffisante. D'autres méthodes d'identification ont alors été développées en plus des caractéristiques physiques. L'une des premières méthodes d'identification a probablement été l'utilisation de noms. Dans de nombreuses cultures, un nom est donné à un individu à sa naissance et ce nom

est utilisé pour l'identifier tout au long de sa vie. À l'époque de l'Empire romain, les parents étaient tenus de donner trois noms (tria nomina) à leurs enfants. Un praenomen (prénom), un nomen (nom de famille) et un cognomen (surnom).

À cette époque, le besoin d'identification était déjà multiplié. L'identification était utilisée à des fins de comptabilité ou de mesure, par exemple pour recenser une population ou une armée. Dans d'autres cas, l'identification était destinée à assurer la sécurité de la cité, en contrôlant l'accès aux portes. À mesure que la population de l'Empire romain augmentait, l'utilisation du trio nomina diminuait, car de plus en plus de personnes partageaient le même nom et les opérations d'identification de sécurité devenaient vulnérables. Le domicile et les vêtements étaient également investis comme moyen d'identification. Ces deux notions indiquaient de l'origine et du statut de l'individu. Cependant les limites sont évidentes, en particulier lorsqu'il s'agit d'identifier une personne par ses vêtements. Il suffit de porter d'autres vêtements pour dissimuler son identité et même se faire passer pour quelqu'un d'autre. C'est ainsi que, malgré tous ces moyens d'identification, de nombreux cas d'usurpation d'identité ont été recensés au cours de l'histoire.

L'affaire Martin Guerre est l'un des premiers cas d'usurpation d'identité recensés en France [1]. Martin Guerre était un paysan français qui a quitté son foyer en 1548 après avoir été accusé de vol. Il revient huit ans plus tard, mais n'est pas reconnu par sa femme et sa famille. Un homme ressemblant à Martin Guerre et connaissant de nombreux détails sur sa vie avait pris sa place. Les villageois, ses frères et même sa femme pensaient qu'il était le vrai Martin Guerre. Après une longue procédure judiciaire, l'imposteur est démasqué et condamné à mort. Ce célèbre cas d'imposture illustre les limites des témoignages basés uniquement sur la parole des personnes pour identifier les individus.

Au fur et à mesure que de plus en plus de personnes étaient en mesure d'écrire et de lire, de nouveaux moyens d'identification par écrit ont été instaurés. C'est ainsi qu'au XVIIe siècle, « l'ordonnance de Saint-Germain-en-Laye » propose de remplacer les preuves par témoins par des registres écrits. Alors que toute identification était basée sur une reconnaissance impersonnelle, le document écrit s'est peu à peu imposé comme nouveau moyen d'identification.

2.1.2 Les premiers documents d'identités

La naissance des sociétés modernes impose un changement dans la nécessité d'identifier les personnes. Les populations sont de plus en plus nombreuses et deviennent mobiles. Afin de contrôler les déplacements des peuples, de lutter contre l'usurpation d'identité et d'identifier des criminels, la reconnaissance des personnes est encadrée par des écrits délivrés par les États. Cela est notamment lié à la popularisation du papier au niveau européen et de l'invention de l'imprimerie en 1450. Les premiers documents d'identité délivrés par les états sont les sauf-conduits et les passeports.

Les sauf-conduits tirent leurs origines dans les « Paiza ». Les tablettes Paiza étaient utilisées par les fonctionnaires de haut rang et les nobles sous l'Empire mongol. Elles étaient faites de jade ou d'autres matériaux précieux, et portaient le nom et le titre du propriétaire. Les Paiza permettaient à leurs détenteurs de voyager librement dans tout l'empire, et leur donnaient accès aux commerces et ressources impériales. Marco Polo les a décrites comme des "tablettes d'autorité" qui étaient "portées par les personnes en mission spéciale".

Un sauf-conduit est un document officiel qui permet à une personne étrangère de voyager en toute sécurité dans un territoire défini (Figure 2). Il contient généralement le nom et des informations d'identification physique du porteur. Les sauf-conduits sont délivrés par les gouvernements ou les commandants militaires et sont généralement valables pour une période déterminée.

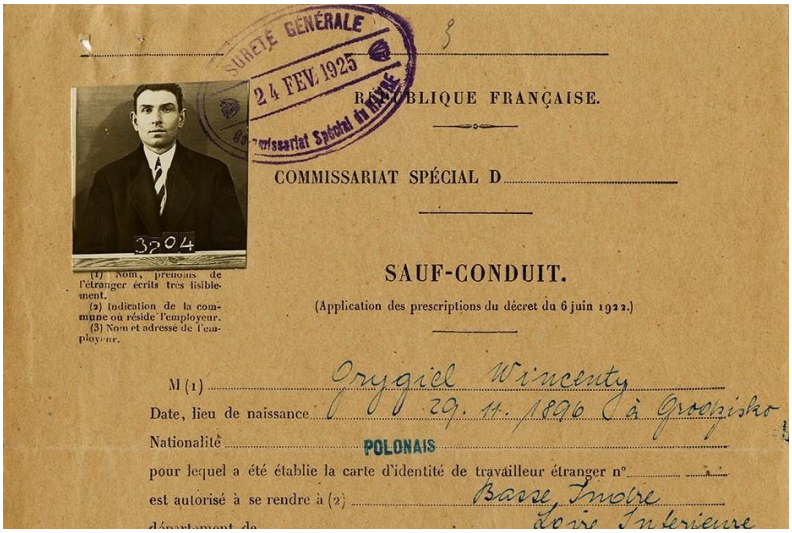

Figure 2 : Exemplaire d'un sauf-conduit délivré en 1925.

Les passeports sont similaires aux sauf-conduits dans la mesure où ils permettent à leur détenteur de voyager en toute sécurité en territoire ennemi (Figure 3). Cependant, les passeports sont généralement délivrés par les autorités civiles et sont valables pour une période plus longue. Le mot "passeport" vient des deux termes "pass" et "port", qui signifient le passage d'une porte, et donc il représente un document qui permet de traverser les portes d'une ville et de pouvoir se déplacer dans un territoire. Le passeport est un document qui sera rapidement adopté par de nombreux pays dans le monde. Tout comme un sauf-conduit, de nombreuses caractéristiques physiques sont inscrites sur les passeports. Ces informations comprennent la taille, la couleur des yeux et la couleur des cheveux. Reprenant ainsi les premières caractéristiques d'identification qui ont été utilisées dans le passé. Ces données sont utilisées pour aider à identifier les détenteurs de passeport lorsqu'ils voyagent.

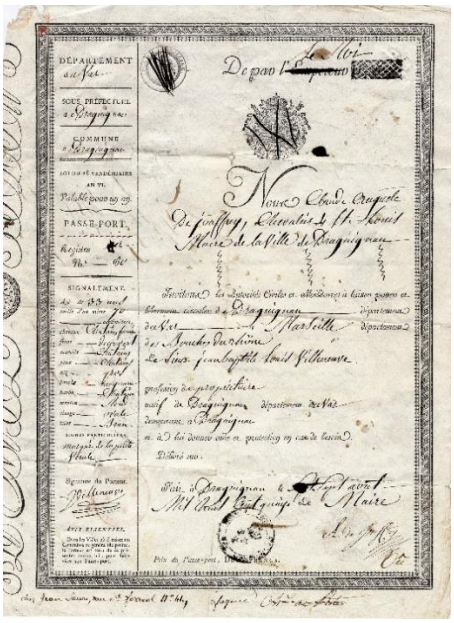

Figure 3 : Exemplaire d’un passeport délivré en 1815.

Les cartes nationales d'identités (CNI) vont être introduites après la Première Guerre mondiale. Les CNI sont un document qui permet aux individus de prouver leur identité. Dans de nombreux pays, les cartes d'identité sont délivrées par le gouvernement et sont utilisées pour voter, demander un permis de conduire et avoir accès aux prestations publiques.

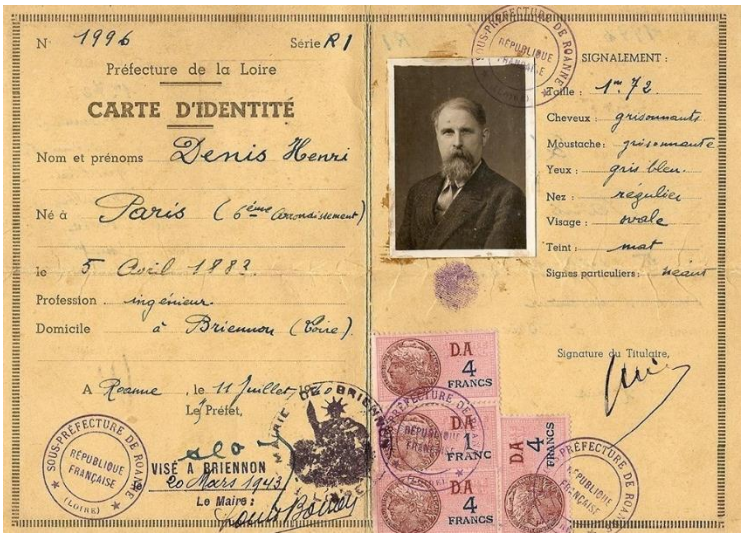

Figure 4 : Exemplaire d'une carte d'identité française délivrée en 1940.

L'identification des personnes sur la base de documents d'identité ne signifie pas qu'elle remplace les anciennes méthodes. Au contraire, cette méthode d'identification coexiste avec la reconnaissance interpersonnelle. Cependant, la reconnaissance interpersonnelle couplée aux documents d'identité ne met pas fin à la fraude et au vol d'identité. C'est même plutôt l'inverse qui se produit. Les documents écrits donnent lieu à de nouveaux types de fraude, car les faussaires trouvent le moyen de les falsifier.

2.1.3 L'identité numérique

Avec les progrès technologiques et la popularisation d'Internet, de nouveaux besoins d'identification vont émerger. Internet a particulièrement modifié les interactions entre individus. Comme nous l'avons expliqué, l'identification était historiquement basée sur des caractéristiques physiques et des documents écrits. Cependant, ces moyens sont difficilement exploitables en ligne. Un nouveau type d'identité a alors été développé : l'identité numérique.

L'identité numérique est un ensemble d'informations qui définit un individu dans le monde numérique et permet de vérifier son identité. Ces informations sont le plus souvent représentées par une association d'un nom d'utilisateur et d'un mot de passe. Il existe d'autres moyens d'identification, comme la date de naissance ou les numéros d'identification personnels (PIN). Le nom d'utilisateur est souvent l'adresse électronique de l'individu. Le mot de passe est un code secret que seul l'individu connaît. En utilisant uniquement des identifiants fixes comme le mot de passe et l'email, il n'y a aucun moyen de déterminer avec précision l'identité d'une

personne dans le monde numérique, car ces informations peuvent être piratées ou devinées par des personnes qui pourront usurper l'identité de l'utilisateur.

L'identité numérique est employée principalement pour deux processus : l'identification et l'authentification des utilisateurs en ligne. L'authentification est le processus de vérification de l'identité d'un utilisateur à partir d'un facteur d'authentification, c'est-à-dire une preuve vérifiée. Il existe trois facteurs d'authentification :

- Le facteur de **connaissance** : une information que l'utilisateur connait ;
- Le facteur de **possession** : une information que l'utilisateur possède ;
- Le facteur **d'appartenance** : une information liée au corps de l'utilisateur.

Le facteur de connaissance regroupe les éléments comme les mots de passe et les codes PIN. Afin d'accéder à une ressource en utilisant ce facteur, une personne doit prouver qu'elle connaît les bonnes informations.

Le facteur de possession comprend les objets tels que les documents d'identité et les cartes d'accès. Afin d'accéder à une ressource en employant ce facteur, une personne doit posséder le bon objet physique.

Le facteur d'appartenance désigne principalement les caractéristiques biométriques comme les empreintes digitales, le visage et la voix. Afin d'accéder à une ressource en recourant à ce facteur, une personne doit présenter une preuve physique pour prouver qu'elle est bien celle qu'elle prétend être.

Chaque facteur a ses propres avantages et inconvénients. L'avantage du facteur de connaissance est que les informations sont sous le contrôle de leur propriétaire. Cependant, ces données peuvent être oubliées ou piratées. Ce type de facteur d'authentification est le plus utilisé sur Internet, cependant il est vulnérable au *pishing* et à d'autres attaques d'ingénierie sociale. Le facteur de possession est quant à lui plus simple d'utilisation, car il n'y a pas besoin de retenir une information particulière, cependant il peut être perdu ou volé. Enfin, le facteur d'appartenance est à la fois simple d'utilisation et ne peut pas être perdu, cependant des moyens existent pour falsifier les caractéristiques biométriques.

Pour pallier les inconvénients de chaque catégorie de facteur, il est courant d'avoir recours à des mesures d'authentification forte comme l'authentification à deux facteurs.

L'authentification à deux facteurs exige que l'utilisateur dispose de deux facteurs d'authentification plutôt qu'un seul. La forme la plus courante d'authentification à deux facteurs est l'utilisation d'un facteur de connaissance avec un facteur de possession. Par exemple une combinaison d'un nom d'utilisateur et d'un mot de passe avec un code à usage unique qui est envoyé sur le smartphone de l'utilisateur. Ce code doit être saisi pour pouvoir accéder à la ressource. Les mots de passe à usage unique sont une autre forme d'authentification à deux facteurs. Ces mots de passe ne peuvent être utilisés qu'une seule fois et sont généralement générés par une application d'authentification. Les mots de passe à usage unique sont plus sûrs que les mots de passe traditionnels, car ils ne peuvent pas être réutilisés.

L'essor des services numériques alourdit les processus d'authentification. Pour avoir un niveau de sécurité suffisant, il est nécessaire d'avoir au moins recours à une authentification à deux facteurs, cependant les utilisateurs sont soumis à une multitude de demandes d'authentification, devoir se souvenir d'un mot de passe pour chaque compte devient fastidieux, car on le rappelle, il est important d'avoir un mot de passe différent pour chaque service. C'est pourquoi le facteur d'appartenance s'est progressivement imposé comme la meilleure manière de s'authentifier de façon numérique.

2.1.4 Les prémices de la biométrie

La biométrie comme moyen de vérifier l'identité des personnes a commencé à se développer bien avant l'apparition des besoins d'authentification numérique. La biométrie a pour origine l'anthropométrie. L'un des pionniers dans le domaine de l'anthropologie est Alphonse Bertillon.

Alphonse Bertillon était un criminologue et anthropologue français à qui l'on attribue la création de la science de l'identification criminelle. Il est surtout connu pour son travail sur les empreintes digitales et les mesures faciales. En 1879, il rejoint la police parisienne en tant que commis. Il a rapidement commencé à utiliser ses connaissances en mathématiques et en statistiques pour aider la police à résoudre des crimes.

En 1882, Bertillon a développé un système d'identification des criminels. Ce système, connu sous le nom de bertillonnage, a été utilisé par les forces de police du monde entier pendant de nombreuses années. Son système était basé sur l'anthropométrie. Bertillon a également été l'un des premiers à utiliser les empreintes digitales pour l'identification des

criminels. Dans [2], il décrivait comment prendre et classer les empreintes digitales. Son travail sur les empreintes digitales a conduit à leur adoption généralisée par les forces de police.

L'anthropométrie est l'étude des mesures du corps humain. En plus de l'identification des personnes, l'anthropométrie peut être utilisée pour évaluer la santé des populations et pour comprendre comment différents groupes de personnes diffèrent dans leurs caractéristiques physiques. Les données anthropométriques peuvent être collectées par le biais d'enquêtes ou de mesures directes. Les données d'enquête sont généralement recueillies à l'aide de questionnaires, tandis que les mesures directes sont effectuées à l'aide d'instruments calibrés tels que des mètres rubans, des stadiomètres ou des analyseurs de composition corporelle.

Le système de Bertillon s'appuyait sur plusieurs mesures de la tête et du corps, quatorze au total, pour créer un identifiant unique pour chaque individu. Les mesures étaient ensuite notées sur une fiche (Figure 5).

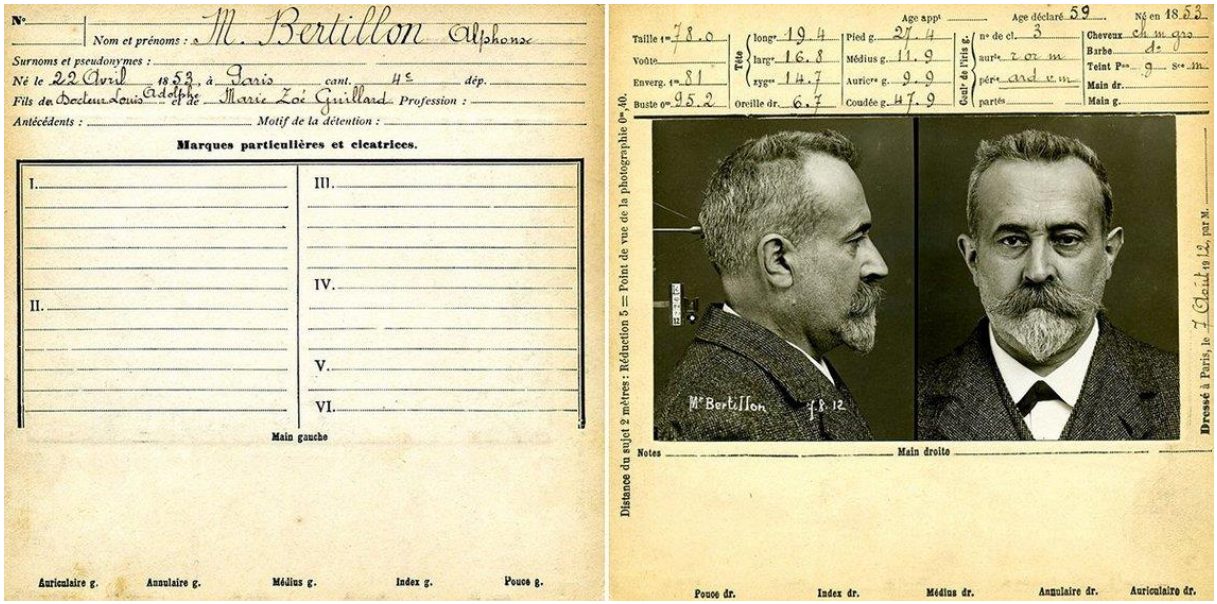

Figure 5 : La fiche anthropométrique d'Alphonse Bertillon.

Ce système a été utilisé pendant de nombreuses années, mais a fini par tomber en désuétude en raison de sa dépendance à des caractéristiques physiques qui pouvaient changer avec le temps.

2.2 Authentification d’identité à distance

L’authentification d’identité à distance est principalement basée sur une authentification biométrique. Contrairement au système de Bertillon, une authentification biométrique repose sur des caractéristiques physiques qui n’évoluent pas ou peu au cours du temps.

2.2.1 Authentification biométrique

Le terme biométrie est tiré des mots grecs "bio" (vie) et " métrie " (mesure). Les caractéristiques biométriques sont des éléments physiques ou comportementaux uniques qui peuvent être utilisés pour identifier un individu. La biométrie est utilisée depuis des siècles, mais ce n'est que récemment que la technologie est devenue suffisamment sophistiquée pour être utilisée à des fins d'identification. La première utilisation enregistrée de la biométrie remonte à la Chine ancienne, où les empreintes digitales étaient utilisées pour identifier des documents.

Au début du 21e siècle, la biométrie a commencé à être utilisée plus fréquemment à des fins de sécurité. Les attaques terroristes du 11 septembre 2001 ont mis en évidence la nécessité d'améliorer les mesures de sécurité, et la biométrie a été considérée comme une solution potentielle. Depuis ce jour, des systèmes d'identification biométrique ont été installés dans les aéroports, les bâtiments gouvernementaux et d'autres lieux de haute sécurité.

Il existe de nombreuses caractéristiques biométriques et de nouvelles sont régulièrement étudiées. Voici une liste de quelques traits biométriques généralement utilisés :

- Les empreintes digitales ;
- L'iris ;
- Le visage ;
- La voix ;
- La géométrie de la main ;
- La dynamique de la signature ;
- La dynamique de la frappe au clavier ;
- La démarche ;
- Etc.

Une récente étude a démontré que le trait biométrique le plus populaire chez les utilisateurs était en 2021 le visage (Figure 6).

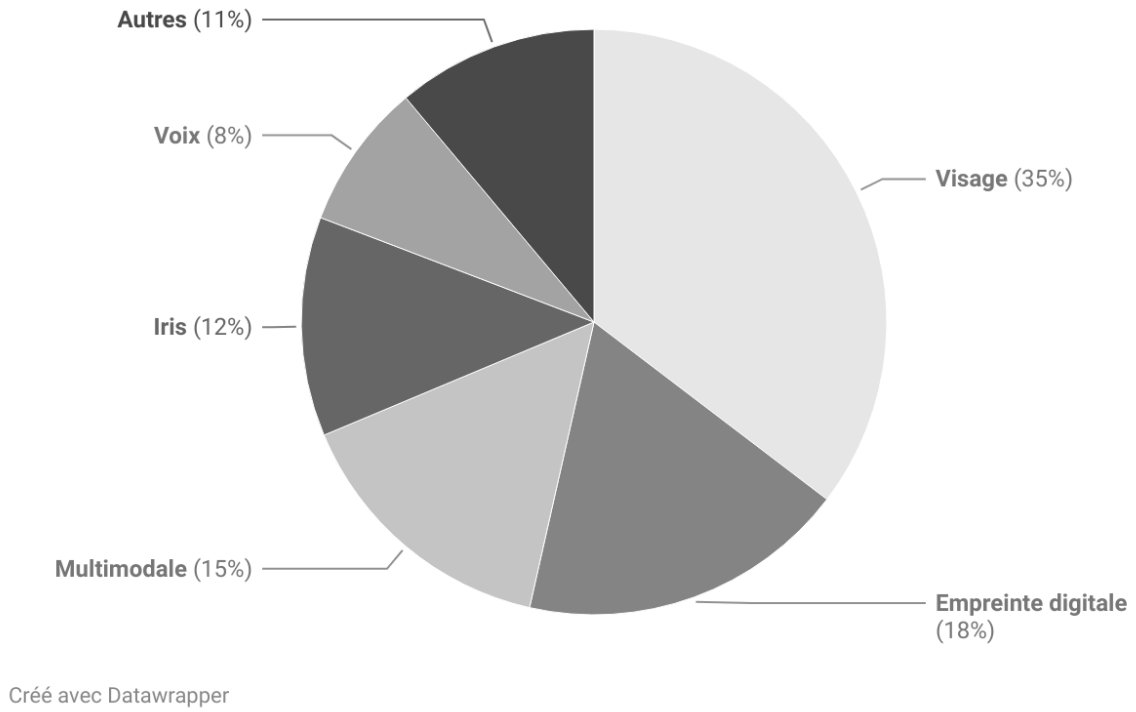

Figure 6 : Les traits biométriques les plus populaires en 2021.
Source : FindBiometrics

Chaque trait biométrique a ses propres avantages et inconvénients, il est donc important de choisir celui qui convient le mieux à chaque application. Par exemple, les empreintes digitales sont fréquemment utilisées car elles sont relativement faciles à collecter et à comparer. Cependant, elles peuvent être falsifiées par de fausses empreintes fabriquées à partir de matériaux comme le latex ou la résine de bois. La reconnaissance de l'iris est plus sécurisée, mais elle nécessite un équipement spécial et un sujet qui coopère. La caractéristique biométrique qui est finalement adoptée dépend de chaque application et des exigences de sécurité. Un certain nombre de facteurs doivent être pris en compte lors du choix d'un système biométrique [3].

- **Universalité** : ce critère détermine si tous les êtres humains possèdent la caractéristique biométrique. Il détermine si le système peut être utilisé par l'ensemble de la population ou seulement par une partie d'entre elle ;

- **Unicité :** il s'agit de la probabilité que deux individus partagent le même trait biométrique et donc détermine le niveau de différenciation entre chaque personne ;
- **Permanence** : la permanence définit le degré d'évolution au cours du temps de la caractéristique biométrique. Le vieillissement d'une personne a impact sur cette modalité ;
- **« Collectabilité » :** il s'agit du niveau de facilité avec laquelle la biométrie peut être acquise ou mesurée ;
- **Acceptabilité :** fait référence à l'acceptabilité sociétale, culturelle, religieuse et éthique de l'utilisation d'une biométrie particulière.
- **Performance :** il s'agit de la précision avec laquelle la technologie biométrique peut identifier ou vérifier les individus.
- **Robustesse :** la capacité de la technologie biométrique à résister aux facteurs environnementaux et physiques qui peuvent dégrader la qualité des données biométriques.

Il n'existe pas de trait biométrique qui réponde à toutes ces exigences. Cependant, certains d'entre eux peuvent être considérés comme supérieurs aux autres. [3] ont proposé un tableau (Tableau 1) permettant de comparer les différents traits biométriques à partir des facteurs qui viennent d'être listés.

Tableau 1: comparaison des traits biométriques les plus courants

(★ = Bas, ★★ = Moyen, ★★★ = Élevé)

Trait biométrique	Univ.	Unic.	Perm.	Coll.	Acc.	Perf.	Rob.
Visage	★★★	★	★★	★★★	★★★	★	★
Empreinte digitale	★★	★★★	★★★	★★	★★	★★★	★★★
Iris	★★★	★★★	★★★	★★	★	★★★	★★★
Voix	★★	★	★	★★	★★★	★	★

On observe que la voix remplie le moins de critères en comparaissant au visage, à l'empreinte digitale et à l'iris. Ceci s'explique, car la voix fait partie des biométries comportementales alors que les autres traits cités font partie des biométries physiologiques. La biométrie comportementale est considérée moins sécurisé que la biométrie physiologique. D'ailleurs, l'Organisation de l'Aviation Civile International (OACI), ne reconnait que trois traits

biométriques comme étant suffisant sécurisé pour être utilisé dans une application d'embarquement. Les trois traits biométriques conformes pour l'OACI sont l'empreinte digitale, l'iris et le visage (Figure 7).

Figure 7 : Un exemple d'une empreinte digitale, d'un iris et d'un visage.

L'OACI est une agence spécialisée des Nations unies qui s'emploie à gérer le secteur mondial de l'aviation civile et établit des normes internationales pour les avions et les aéroports. L'une des principales missions de l'OACI est l'élaboration de normes biométriques destinées à être utilisées dans le secteur de l'aviation civile.

Une empreinte digitale est un élément d'identification biométrique qui repose sur l'analyse des motifs inscrits sur le bout des doigts. Ces motifs sont appelés des crêtes et des vallées. Ils sont uniques à chaque individu et n'évoluent pas au cours du temps. L'iris est un autre trait biométrique basée sur des motifs uniques. L'iris est l'anneau coloré qui entoure la pupille de l'œil et qui contient de nombreuses caractéristiques propres à chaque individu. Enfin, le visage est une caractéristique biométrique qui évolue au cours du temps, mais qui reste assez stable et ne perturbe pas les systèmes de reconnaissance faciale actuels.

2.2.2 Reconnaissance faciale

Le visage est l'une des caractéristiques les plus importantes pour l'identification entre les êtres humains. Notre visage est unique et nous permet de nous distinguer des autres. Le visage est souvent la première information que nous percevons chez une autre personne. De nombreux éléments du visage sont exploités à des fins d'identification, notamment la forme, la taille et les traits du visage. Le visage est également un indicateur efficace des émotions d'une personne. Lorsque nous interagissons avec des tiers, nous nous servons souvent des expressions faciales pour transmettre notre état émotionnel. Nous n'en sommes pas forcément conscients, mais notre visage peut trahir nos véritables sentiments dans une situation donnée. Cela peut se

révéler efficace pour tromper les autres, car nous pouvons utiliser les expressions faciales pour leur faire croire que nous ressentons quelque chose que nous ne ressentons pas.

Lorsque nous croisons un visage qui nous est familier, notre cerveau le reconnaît automatiquement et stocke les informations le concernant dans notre mémoire. Ce processus est connu sous le nom de reconnaissance faciale. La reconnaissance faciale est un mécanisme complexe qui mobilise plusieurs régions du cerveau. Il est considéré comme étant largement contrôlé par l'amygdale, qui est responsable des réactions émotionnelles. La capacité à reconnaître les visages est importante pour les interactions sociales, car elle nous permet d'entrer en contact avec d'autres personnes. Elle a également des implications importantes pour la sécurité, car elle nous aide à identifier les menaces potentielles.

Certaines personnes sont capables de reconnaître des personnes qu'elles ont déjà rencontrées, même si elles ne les ont vues qu'une fois ou si elles ont changé d'apparence. Ces personnes sont appelées les super-reconnaisseurs. Si la plupart d'entre nous peuvent se souvenir du visage d'une personne que nous connaissons bien, les super-reconnaisseurs sont capables de se souvenir de visages qu'ils n'ont vus que brièvement ou même une seule fois. Ils peuvent reconnaître un visage dans une foule, ou repérer un visage sur une photo. Les super-reconnaisseurs contribuent à résoudre des crimes en identifiant les auteurs à partir de séquences de vidéosurveillance. Ils peuvent également travailler dans le domaine de la sécurité, en utilisant leurs compétences pour repérer les fauteurs de troubles connus ou les personnes qui ne sont pas censées se trouver dans une certaine zone. Afin d'automatiser ces tâches de reconnaissance, de nombreuses recherches ont été menées au cours de l'histoire.

La reconnaissance faciale est une technologie qui existe depuis des dizaines d'années. En 1964, les services de renseignement américains ont commencé à recourir à la reconnaissance faciale pour identifier les dirigeants communistes en Chine. En 1967, la reconnaissance faciale a été utilisée à l'aéroport international de Montréal pour identifier un criminel recherché. Et en 1968, un logiciel de reconnaissance faciale a été utilisé par le FBI pour identifier un suspect dans l'assassinat de Martin Luther King. Ces premiers systèmes étaient facilement déjoués par les changements d'expression faciale ou de coiffure. Dans les années 1970, des algorithmes plus sophistiqués ont été développés pour prendre en compte ces changements. Cependant, ces systèmes n'étaient toujours pas très performants. Dans les années 1990, les systèmes de reconnaissance faciale sont devenus plus précis grâce à des techniques d'intelligence artificielle.

La reconnaissance des visages est une tâche complexe qui est rendue difficile par un certain nombre de facteurs (Figure 8). L'un des facteurs les plus importants pour réussir la détection des visages est le type de visage à détecter. Les algorithmes doivent être capables de gérer une grande variété de types de visages, notamment des visages d'hommes et de femmes, des visages jeunes et vieux, et des visages avec différentes teintes de peau. Un autre facteur important est les conditions d'éclairage. Les visages peuvent être éclairés sous de nombreux angles différents, et certaines conditions d'éclairage (comme le contre-jour) peuvent rendre les visages plus difficiles à détecter. Les visages peuvent aussi être masqués par différents objets qui peuvent aussi interférer avec les algorithmes de détection des visages. Les visages sont souvent partiellement masqués par les cheveux, les chapeaux ou les lunettes. De plus, les algorithmes de détection des visages doivent être capables de gérer une variété d'orientations du visage, y compris les vues de dessus et de profil.

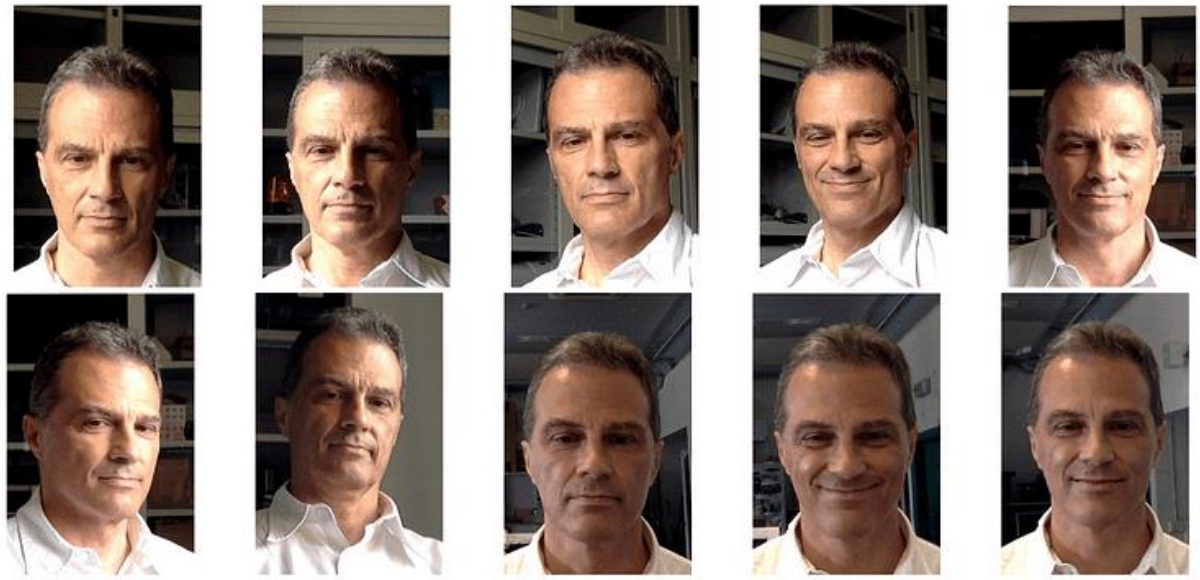

Figure 8 : variétés de différentes conditions d'acquisitions d'un même visage.

Malgré ces contraintes, les algorithmes de reconnaissance des visages ont fait de grands progrès ces dernières années et sont désormais capables de détecter les visages dans diverses conditions avec une grande précision. Avec le développement de l'apprentissage automatique et de l'apprentissage profond, la reconnaissance faciale est désormais supérieure aux capacités humaines. Le meilleur système de reconnaissance faciale atteint une *accuracy* de 99.85% [4]. La reconnaissance faciale est la technologie biométrique privilégiée dans un système de vérification d'identité à distance.

2.2.3 Documents biométriques

Les documents d'identités ont également évolué pour inclure des informations biométriques. Un document biométrique est un document physique qui contient des informations sur les caractéristiques biométriques d'une personne. Ces informations peuvent inclure des éléments comme les empreintes digitales ou le visage. Afin de garantir l'exactitude des informations contenues dans les documents biométriques, ils sont généralement produits par des organismes autorisés à l'aide d'un équipement spécialisé. Cet équipement capture des images de haute qualité des éléments biométriques d'un individu, qui sont ensuite stockées dans une base de données sécurisée. Lorsqu'une personne doit fournir ses données biométriques à des fins d'identification dans une application e-KYC, elle peut le faire en présentant son document biométrique.

Cette évolution fait directement suite aux recommandations de l'OACI au début des années 2000. L'OACI a proposé un nouveau format pour les passeports et les cartes d'identité : le passeport électronique et la carte d'identité électronique. Ces nouveaux documents intègrent une puce à circuit intégré qui stocke à la fois les informations textuelles du document et des informations biométriques du détenteur (photo du visage et empreintes digitales). La plupart des pays ont adoptés les documents d'identité biométriques en suivant les recommandations de l'OACI. En France, c'est le cas de la nouvelle carte nationale d'identité (CNI) qui est rentrée en vigueur en 2021 (Figure 9). Chaque pays de l'Union Européenne doit avoir adopté le nouveau format de carte d'identité depuis 2021 [5].

Figure 9 : exemple d'une CNI française. Recto (à gauche) et verso (à droite).

Les pièces d'identité biométrique sont conçues pour rendre plus difficile l'utilisation de méthodes de falsifications par des imposteurs et pour accélérer le processus de contrôle des documents aux frontières. Les normes de l'OACI exigent également que le document d'identité

soit lisible par une machine. C'est pourquoi on retrouve en bas de chaque document biométrique une zone de lecture automatique (ZLA) permettant d'identifier et vérifier le document.

Les passeports biométriques (Figure 10) avaient déjà été adoptés en France depuis 2009. Les passeports biométriques sont utilisés dans des systèmes automatisés de contrôle aux frontières.

Figure 10 : exemple de page d'informations d'un passeport biométrique.

Un système automatisé de contrôle aux frontières est une solution de sécurité qui recourt à une authentification biométrique pour vérifier l'identité des voyageurs et simplifier le processus de traversée des frontières. Les dispositifs de contrôle automatisé des frontières utilisent également des documents biométriques, pour contrôler l'identité. Ces systèmes se composent généralement d'un kiosque de contrôle automatisé des passeports où les voyageurs peuvent scanner leurs propres documents. Les systèmes automatisés de contrôle aux frontières sont de plus en plus répandus dans les aéroports et autres lieux de passage dans le monde entier. Ils offrent un certain nombre d'avantages par rapport aux processus manuels traditionnels de contrôle des frontières, notamment une réduction des délais de traitement, une amélioration de la sécurité et une meilleure efficacité.

En France, le système « passage automatisé rapide aux frontières extérieures » (PARAFE) est un exemple de système automatisé de contrôle aux frontières. Le premier appareil PARAFE a été inauguré en 2009 dans le but de simplifier les voyages aériens et de les rendre plus pratiques pour les passagers. Ce sont des kiosques en libre-service permettant aux

passagers de passer les contrôles de sécurité des aéroports sans avoir besoin d'un agent humain. Le système PARAFE utilise la technologie de reconnaissance faciale pour faire correspondre les données biométriques d'un voyageur aux informations de son passeport. Cela permet une vérification rapide et facile de l'identité, sans qu'il soit nécessaire de procéder à des contrôles manuels.

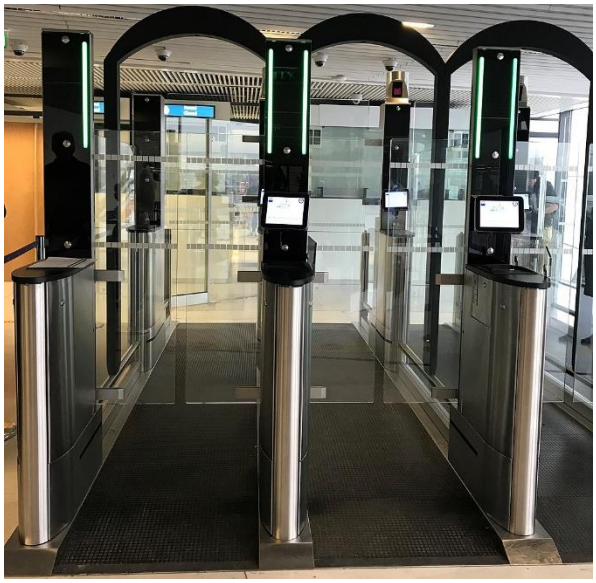

Figure 11 : système PARAFE à l'aéroport de de Paris-Charles-de-Gaulle.

2.2.4 Authentification à distance

Depuis le début de la pandémie du covid, le recours à la technologie de reconnaissance faciale a connu un essor considérable. La reconnaissance faciale permet de déterminer les personnes susceptibles d'avoir été touchées par le virus. La technologie de reconnaissance faciale est utilisée de différentes manières pour lutter contre la propagation du covid. Par exemple, la reconnaissance faciale est employée pour identifier les personnes qui ne portent pas de masque. Elle peut également servir à tracer les déplacements des personnes qui ont été contaminées par le virus. Le principal avantage de la reconnaissance faciale est qu'elle est sans contact et peut se faire à distance, ce qui est capital pour prévenir la propagation du Covid-19.

Le COVID-19 a donc accéléré le développement d'un processus qui était déjà en cours : l'adoption des systèmes de vérification d'identité à distance. Ces systèmes reposent sur les principes du *Know Your Customer* (KYC). Le terme *Know Your Customer* désigne le processus par lequel les banques et autres institutions financières vérifient l'identité de leurs clients. Cela

est fait afin de se conformer aux réglementations sur la lutte contre le blanchiment d'argent. Le KYC à distance (e-KYC pour *electronic* KYC) est le processus de vérification de l'identité d'un client par le biais de canaux numériques. Cette vérification peut se faire par le biais de différentes méthodes, mais principalement par la vérification de documents et l'authentification biométrique. Il s'agit généralement d'utiliser une pièce d'identité délivrée par un État, comme une CNI ou un passeport, et de la soumettre à un prestataire de services. Le prestataire de services vérifie alors la pièce d'identité dans une base de données pour s'assurer de sa validité.

Contrairement aux processus traditionnels de KYC qui reposent sur des interactions en personne, le KYC à distance permet aux entreprises de vérifier l'identité d'un client de n'importe où dans le monde. Le KYC à distance permet d'accélérer le processus d'embarquement des voyageurs et de réduire les coûts associés aux processus KYC traditionnels. La vérification de l'identité à distance contribue également à améliorer la satisfaction des clients en leur offrant un moyen plus pratique de vérifier leur identité. Le processus classique d'un système de vérification d'identité à distance est illustré à la Figure 12.

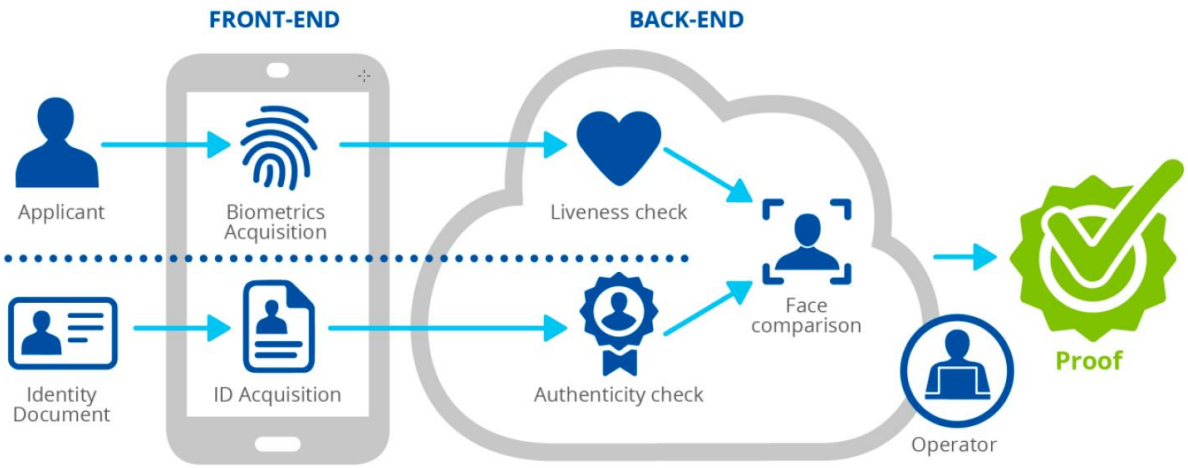

Figure 12 : Schéma d'un processus de vérification d'identité à distance.
Source : *European Union Agency for Cybersecurity*.

Un système de vérification d'identité à distance demande deux types de preuves pour authentifier un individu. L'utilisateur doit d'une part fournir une première preuve à partir d'une pièce d'identité officielle, d'autre part il doit apporter une preuve biométrique. Pour acquérir les preuves, l'utilisateur utilise le plus souvent un smartphone. Il utilise le capteur photographique de son appareil pour capturer une image de son visage et une image de sa pièce d'identité.

Une fois cette première étape réalisée, les deux images vont être analysées pour les valider. Pour l'image du visage de l'utilisateur, un test de *liveness* est souvent demandé pour s'assurer de l'authenticité de l'utilisateur. Dans le cas de l'image de la pièce d'identité, des algorithmes vérifient les éléments de sécurités présents sur le document pour s'assurer de son authenticité. L'image du visage sera ensuite comparée au portrait du document d'identité de l'utilisateur. Si les deux images correspondent, l'utilisateur pourra poursuivre le processus de vérification. Si les images ne correspondent pas, l'utilisateur sera invité à réessayer. En cas de doute durant le processus, un opérateur humain peut intervenir pour vérifier les informations transmises.

2.3 Attaques contre un système de vérification à distance

Pour authentifier un utilisateur et un document d'identité, il est demandé de capturer une preuve visuelle sous forme d'une image ou d'une vidéo du visage de l'utilisateur et de la page informative de son document d'identité. Comme le processus se déroule à distance, des fraudeurs ont la liberté de falsifier les preuves visuelles pour tromper le système et usurper l'identité d'une personne. Dans la suite de ce chapitre, une personne qui attaque un système e-KYC est appelé un **attaquant**.

2.3.1 Sécurité d'un système de vérification à distance

Comme un système e-KYC utilise des technologies de reconnaissance faciale dans son processus d'authentification, le système a les mêmes vulnérabilités qu'un système biométrique. Ratha et al. [6] divisent les vulnérabilités d'un système biométrique en 8 catégories (Figure 13).

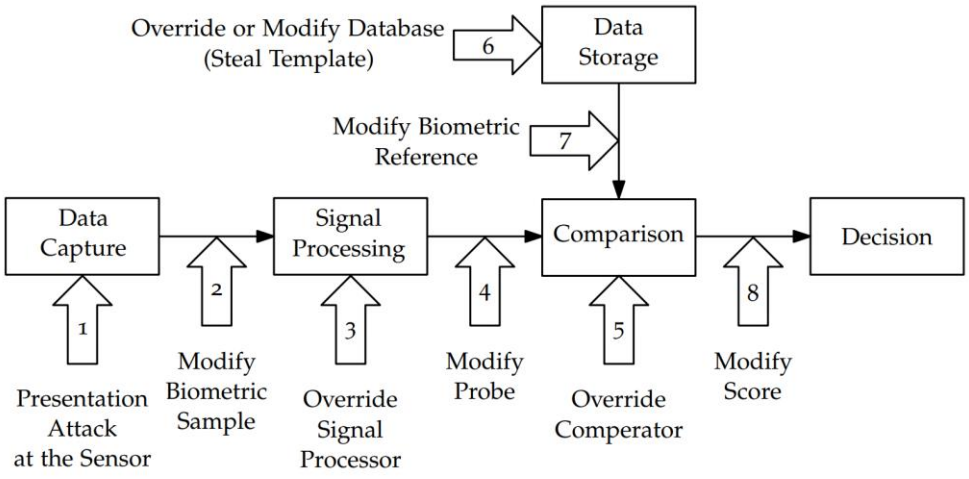

Figure 13 : Les points de vulnérabilité d'un système biométrique.

1. **Attaque par présentation au niveau du capteur :** Cela se produit lorsqu'un attaquant utilise une fausse caractéristique biométrique d'un utilisateur pour tromper le capteur. L'objectif est de contourner le système d'authentification en présentant une preuve biométrique falsifiée au lieu de la preuve biométrique authentique.
2. **Modification de l'échantillon biométrique :** C'est lorsqu'un attaquant modifie un échantillon biométrique authentique. Il s'agit principalement des cas d'injections d'une preuve biométrique falsifiée.
3. **Substitution du signal :** Un attaquant interfère avec le signal afin de modifier le modèle biométrique en cours de création. Par exemple, un attaquant pourrait essayer d'ajouter du bruit à une image pour qu'elle ne corresponde pas au modèle enregistré.
4. **Falsification du signal calculé** : Un attaquant modifie le signal en sortie de l'extracteur de caractéristique. Ce type d'attaque est difficile à réaliser, car l'attaquant doit connaître la méthode employée dans l'extracteur de caractéristique.
5. **Contournement du comparateur :** l'attaquant interfère avec le processus de comparaison entre le signal extrait de l'extracteur de caractéristique et le modèle de référence. Par exemple, un attaquant peut essayer de modifier la valeur du seuil pour qu'elle soit plus basse que d'habitude, ce qui provoque plus de faux positifs.
6. **Attaque de la base de données :** un attaquant essaie de modifier la façon dont la base de données biométriques est gérée afin qu'elle ne fonctionne pas correctement. Par exemple, un attaquant pourrait essayer d'injecter un nouveau modèle de référence dans la base de données.
7. **Modification du modèle de référence biométrique :** l'attaquant modifie le modèle de référence pour qu'il ne corresponde pas au signal calculé par l'extracteur de caractéristiques.
8. **Falsification du score de décision :** un attaquant falsifie le score qui est utilisé pour décider si une correspondance biométrique est authentique ou non.

Dans cette thèse, nous étudions principalement les menaces au niveau du capteur (1) ainsi qu'au point de communication entre le capteur et l'extracteur de caractéristiques (2). Ces deux attaques sont réalisées dans deux environnements différents. L'attaque (1) est catégorisée comme une attaque directe réalisée dans le monde physique. L'attaque (2) est définie comme une attaque indirecte mise en œuvre dans le monde numérique.

2.3.2 Acquisition d’image numérique

Il existe plusieurs types de capteurs photographiques qui peuvent être utilisés dans un système e-KYC. Un capteur photographique est un dispositif permettant de capturer des images. La plupart du temps le capteur photographique utilisé est un capteur RGB présent sur tous les smartphones modernes. La Figure 14 illustre le pipeline d’acquisition d’une image numérique.

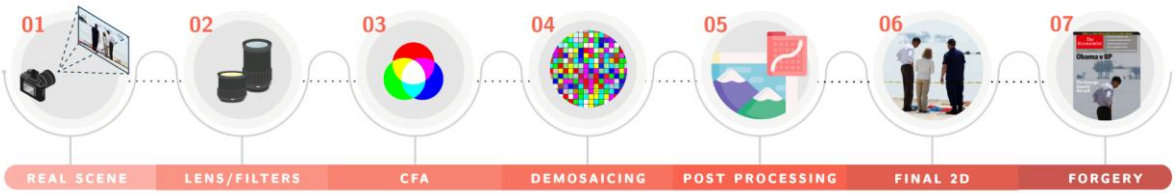

Figure 14 : pipeline d’acquisition d’une image numérique.

Le capteur photographique contient un objectif qui concentre la lumière sur des millions de cellules photoélectriques. Lorsque la lumière frappe ces cellules, ceux-ci la convertissent en un signal électrique. Ce signal est ensuite traité par un processeur interne qui converti le signal en une image numérique.

Plus précisément, une image numérique est acquise lorsque la lumière traverse l'objectif d'un capteur photo et atteint une matrice de filtres colorés ou CFA (*Color Filter Array*). La matrice contient plusieurs capteurs qui convertissent la lumière en signaux électriques. La matrice de filtres colorés d'un capteur photo est une grille constituée d'éléments filtrants placés sur le capteur d'images. Le CFA permet à l'appareil de capturer des informations sur les couleurs de la scène photographiée. Chaque pixel du CFA est doté d'un élément filtrant qui ne laisse passer qu'une certaine longueur d'onde de lumière vers le capteur d'images. Le capteur d'images de l'appareil photo convertit ensuite la lumière en un signal électrique qui est utilisé pour créer l'image. Il existe de nombreux types de CFA, mais le plus répandu est le CFA de Bayer [7]. Une matrice de Bayer est composée de 50% de filtres vers, 25% de filtres bleus et 25% de filtres rouges.

Le CFA de Bayer code une image de façon à ce que chaque pixel ne contienne qu'un seul canal de couleur. Afin de générer une image complète, les informations de couleur manquantes pour chaque pixel doivent être estimées via un processus appelé « dématriçage ». Les algorithmes de dématriçage utilisent généralement une combinaison d'interpolation et/ou

de moyennes des pixels voisins afin d'estimer les valeurs des canaux de couleur manquants. Des étapes de post-traitement telles que la correction des couleurs, la balance des blancs et la correction gamma sont souvent appliquées après le dématriçage afin d'améliorer la qualité de l'image finale.

Une fois l'image numérique acquise, il est nécessaire de la stocker. L'image est au départ en format RAW. Le format RAW est un conteneur pour les données d'images brutes. L'image est souvent très lourde. Pour réduire la taille de l'image, l'image numérique passe par un processus de compression. Deux types de compressions existent : compression avec et sans perte. La compression sans perte ne perd aucune donnée d'image au cours du processus de compression. Il est donc toujours possible de décompresser l'image et d'obtenir une copie identique à l'original. L'inconvénient de la compression sans perte est qu'elle ne réduit pas la taille du fichier autant que la compression avec perte. La compression avec perte, quant à elle, perd certaines données d'image pendant la compression. Cela signifie que lorsque l'on décompresse l'image, nous n'obtiendrons pas une copie exacte de l'original. Cependant, la compression avec perte permet de réduire la taille des fichiers beaucoup plus que la compression sans perte, ce qui la rend idéale pour stocker des images.

Le format d'image de compression avec perte le plus populaire est JPEG. JPEG est une norme de compression d'image largement utilisée dans les fichiers d'images numériques. La compression JPEG fonctionne en réduisant la quantité de données d'image stockées dans un fichier. Cela peut se faire en éliminant les données redondantes ou en utilisant moins de données pour représenter l'image. L'un des principaux aspects du format JPEG est qu'il utilise un espace couleur YCbCr. Cet espace couleur sépare une image en ses composantes de luminance (Y) et de chrominance (Cb et Cr). La composante de luminance représente la luminosité d'une image, tandis que les composantes de chrominance représentent la couleur. L'image est divisée en blocs de 64 pixels (8 x 8). Chaque bloc est ensuite transformé à l'aide de la Transformée Cosinus Discrète (TCD). La DCT crée un ensemble de coefficients qui représentent l'intensité des différentes fréquences de l'image. Ces coefficients sont ensuite quantifiés, les coefficients quantifiés sont ensuite codés à l'aide d'un algorithme de compression sans perte.

Il est aujourd'hui très courant que les internautes publient leurs photos personnelles sur les réseaux sociaux. Le partage de photos se fait le plus souvent directement à partir de la caméra avant du smartphone. Lorsqu'une personne utilise la caméra avant de son smartphone, il s'agit le plus souvent de prendre en photo son visage qui se retrouve ensuite à disposition du

public une fois posté sur les réseaux sociaux. Un imposteur peut alors facilement se procurer du matériel pour attaquer un système e-KYC pour usurper l'identité d'un individu. Les systèmes de vérification d'identité à distance sont sécurisés et donc il est nécessaire pour l'imposteur de falsifier l'image qu'il récupère pour tromper le système.

2.3.3 Falsification du visage

Depuis le développement de la photographie numérique, les falsifications de photos sont devenues encore plus courantes. Grâce à Photoshop et à d'autres logiciels de retouche, il est relativement facile de truquer des photos très crédibles. Au fur et à mesure que la technologie a progressé, la capacité à créer de fausses images crédibles a augmenté de manière significative. Aujourd'hui, il existe des sites Internet et des logiciels dédiés à la création de fausses images de visages.

Attaques physiques

Lorsqu'il s'agit de falsifier physiquement un visage pour attaquer un système e-KYC, l'attaquant va présenter au capteur photo un faux visage. Ce type d'attaque représente les attaques par présentation. Les attaques par présentation peuvent être réalisées de différentes manières, notamment en utilisant une impression photographique du visage de la cible, en portant un masque qui ressemble au visage de la cible ou en rejouant une vidéo d'un visage (Figure 15). Ces attaques peuvent être difficiles à détecter et aboutir à accorder l'authentification à l'attaquant.

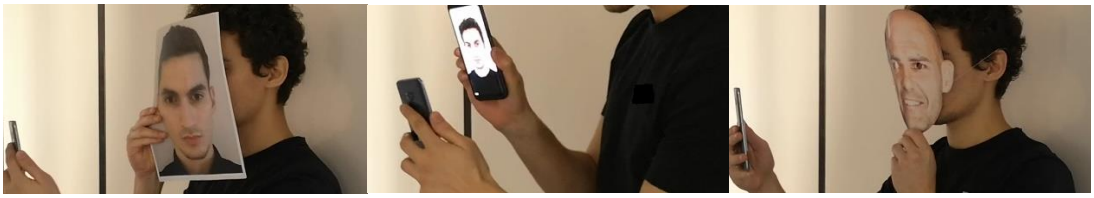

Figure 15 : attaques de présentation faciale, en présentant une photo imprimée (à gauche), en présentant une vidéo (au centre) et en présentant un masque (à droite).

Le moyen le plus simple d'attaquer un système de reconnaissance faciale est d'imprimer une photo de la personne que l'on souhaite usurper et de la présenter à la caméra.

Dans une attaque par lecture d'une vidéo, l'attaquant falsifie une vidéo de la personne qu'il veut faire passer pour une autre, puis la lit devant le système de reconnaissance faciale. Il

suffit pour cela de pointer un téléphone ou une autre caméra vers un écran diffusant la vidéo, ou d'utiliser un dispositif spécial qui projette la vidéo sur le visage.

Les masques 3D sont des reproductions réalistes de visages humains. Ils peuvent être fabriqués à l'aide d'imprimantes 3D ou d'autres techniques. Les masques 3D sont de plus en plus réalistes, et ils sont de plus en plus difficiles à détecter. Ils constituent donc une menace sérieuse pour les systèmes de reconnaissance faciale.

Il existe deux autres types d'attaques physiques moins courantes. La première consiste à porter du maquillage de telle sorte à ressembler à quelqu'un d'autre. Les attaques par maquillage peuvent être difficiles à réaliser, car le maquillage doit modifier suffisamment les traits de la personne sans la rendre complètement méconnaissable. La deuxième façon est d'utiliser des accessoires antagonistes. Ces accessoires sont fabriqués à l'aide d'un bruit antagoniste préalablement générée. Des autocollants intégrant un bruit antagoniste peuvent être apposés sur le visage ou l'attaquant peut porter des lunettes antagonistes. Il est difficile de se défendre contre ce type d'attaque, car ces accessoires sont souvent difficiles à distinguer d'objets ordinaires comme des lunettes de vues.

Attaques numériques

À mesure que les attaques numériques deviennent plus perfectionnées, les systèmes de reconnaissance faciale sont de plus en plus vulnérables à une éventuelle exposition. En particulier, les systèmes de reconnaissance faciale basés sur l'apprentissage profond peuvent être trompés par des images numériques qui ont été manipulées pour ressembler à une autre personne. Il existe quatre catégories de falsifications d'images numériques de visages :

- L'échange de visage ;
- Le *reenactment* de visage ;
- Les visages antagonistes

La figure 16 illustre chacune de ces catégories d'attaques. La première ligne représente un exemple d'échange de visage, la deuxième ligne est un exemple de reconstitution de visage, et la dernière ligne représente un exemple de visage antagoniste.

Figure 16 : les trois types d’attaques de falsifications faciales numériques, à gauche l’image source, au centre l’image cible et à droite l’image falsifiée.

L’échange de visage est traditionnellement réalisé à partir des points de repère faciaux. Il s'agit de prendre une image numérique ou une vidéo de la victime, puis d'utiliser un programme pour détecter ses points de repère faciaux. Le fraudeur réalise ensuite une nouvelle image ou vidéo numérique avec son propre visage sur le corps de la cible. Il existe des falsifications par échange de visage qui se font automatiquement à l'aide de programmes basés sur l’apprentissage profond.

Le *reenactment* de visage est une attaque numérique qui se développe de plus en plus. Cette méthode repose sur l'utilisation de l'apprentissage profond pour créer une copie du visage d'une personne et faire rejouer les traits du visage source sur celui du visage cible.

Les visages antagonistes sont générés en ajoutant des perturbations à l'image originale d'un visage. La perturbation est un bruit antagoniste qui est calculée à partir de l’image source et de l’image cible.

2.3.4 Falsification du document

Dans le cas de la falsification de documents d'identités, l'attaquant peut falsifier différentes zones du document. Un document d'identité contient généralement trois zones (Figure 17) :

- Le **portrait**, où figure la photo du visage du détenteur du document ;
- La **zone d'information textuelle**, qui contient les données personnelles du titulaire ;
- La **zone de lecture automatique** (ZLA), qui contient des données codées pouvant être lues par un appareil.

La zone d'inspection visuelle (VIZ) est la zone du document où se trouvent la photo du porteur et ainsi que les informations textuelles.

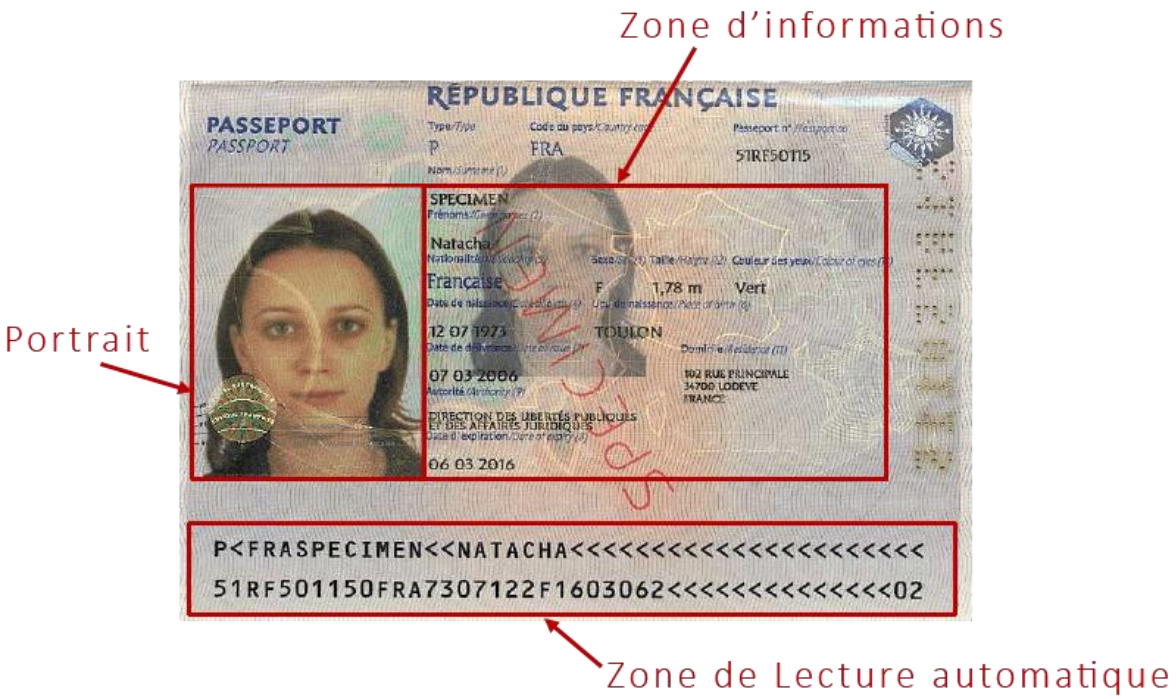

Figure 17 : les trois zones d'un document d'identité.

Attaques physiques

Il existe trois catégories de falsifications physiques de documents d'identité : les documents altérés, les documents contrefaits et les documents vierges volés. Les documents altérés sont des documents authentiques qui ont été modifiés pour changer les informations qu'ils contiennent. Les documents contrefaits sont des faux documents qui ont été créés de

toutes pièces pour ressembler à des documents authentiques. Les documents vierges volés sont des documents authentiques qui ont été volés puis complétés avec des fausses informations.

Attaques numériques

Les falsifications numériques d'un document d'identité peuvent avoir lieu sur chacune des zones décrites. La falsification la plus courante est la falsification par copie-déplacement. C'est une méthode de falsification d'image qui consiste à copier et coller une partie d'une image à un autre endroit de la même image. En général, le falsificateur va copier et coller des lettres ou des chiffres. Le *splicing* est une autre méthode de falsifications similaire au copie-déplacement. C'est un processus qui consiste à ajouter des informations à partir d'un autre document d'identité et à les combiner en un seul document.

Lorsqu'un attaquant souhaite falsifier la photo du document d'identité, il aura recours le plus souvent à des méthodes de *face morphing*. Le *face morphing* est une technique qui permet de créer un nouveau visage en combinant les caractéristiques de deux visages existants. Cette technique est relativement facile à mettre en œuvre et ne nécessite que la maîtrise d'un logiciel de retouche photo de base. Avec le *face morphing*, les fraudeurs peuvent créer une image composite du visage d'une personne qui est suffisamment réaliste pour tromper les logiciels de reconnaissance faciale. Ils peuvent ainsi créer un faux passeport avec le visage d'une autre personne [8].

Trois étapes sont généralement suivies pour réaliser un *face morphing.* La première étape est la correspondance. La correspondance permet d'identifier les points de références des deux images puis de les aligner. La déformation est la deuxième étape. C'est un processus qui consiste à déplacer les points de références. La triangulation de Delaunay est la méthode la plus courante pour déformer les images. Cela revient à diviser les points de références en triangles, puis à déplacer les sommets des triangles vers de nouvelles positions. La dernière étape est le mélange. Le mélange est le processus qui consiste à combiner deux images ensemble. Le mélange est utilisé pour créer une transition douce entre les deux images. La Figure 18 illustre un exemple de *face morphing*.

Figure 18 : un exemple de *face morphing*. Le visage morphé est au centre.

2.4 Contre-mesures et détection

Les méthodes d'attaque décrites dans la section précédente peuvent représenter un défi pour les systèmes de vérification d'identité à distance. Pour cette raison, diverses contre-mesures et méthodes de détection ont été conçues et intégrées dans les systèmes e-KYC.

2.4.1 Contre-mesures

Lors du processus d'authentification du document d'identité, il est intéressant d'analyser les éléments de sécurité présents sur le document. Il existe différents éléments de sécurité qui sont utilisés pour sécuriser un document d'identité (Figure 19).

Figure 19 : les différents éléments de sécurité pour empêcher ou rendre difficile la falsification des documents d'identités.

Les éléments de sécurité sont soit des éléments tactiles ou soit des éléments visuels. Comme le système d'e-KYC demande une image ou une vidéo numérique du document d'identité, les éléments de sécurité tactiles ne peuvent pas être utilisés. Certains éléments visuels ne sont visibles que dans certaines conditions. Par exemple les microtextes sont de très petite impression qui est difficile à copier ou à falsifier. Lorsqu'on regarde la micro-impression à l'œil nu, elle n'est pas visible. Cependant, lorsqu'elle est agrandie, la micro-impression peut être lue. Les éléments de sécurité imprimés pour réagir aux UV ne peuvent être visibles que sous une lampe qui diffuse des UV. Comme le processus d'acquisition se fait via le capteur photo d'un smarphone, ces éléments de sécurités ne peuvent pas être analysés sur l'image ou la vidéo du document.

Il existe de nombreux autres éléments de sécurités qui sont exploitables pour authentifier un document d'identité à partir d'une image ou d'une vidéo numérique.

Le premier élément est la zone de lecture automatique (ZLA). La zone de lecture automatique est une série de lignes au bas de la page de données qui peut être lue par une machine et qui contient les informations du titulaire du passeport. Une correspondance est alors réalisée entre la ZLA et les informations textuelles du document. Une correspondance peut être aussi faite entre le portrait et la photo fantôme. La photo fantôme est une seconde photographie du visage du détenteur qui est imprimé sur le document. En général l'image est imprimée avec un procédé différent du portrait principal.

L'utilisation d'hologrammes de sécurité est l'un des moyens d'ajouter une couche supplémentaire de sécurité aux documents. Un hologramme de sécurité est un type particulier d'hologramme qui contient des éléments de sécurité qui le rendent difficile à falsifier. Les hologrammes de sécurité contiennent généralement un logo ou une image de sécurité qui ne peut être observé que sous certains angles. Il est donc difficile de reproduire l'hologramme de sécurité sans avoir accès à l'original. Différentes technologies existent. La technologie DOVID est un hologramme qui peut être utilisé pour projeter des images 2-D ou 3D. Les hologrammes DOVID sont créés en utilisant un laser pour encoder une image dans un matériau photosensible. Ce matériau est ensuite éclairé par un autre laser, ce qui permet de projeter l'image codée sous forme d'hologramme tridimensionnel. La technologie DID permet d'obtenir un hologramme dont les couleurs permutent lorsque l'on applique une rotation à 90 degrés (Figure 20). Ce type d'hologramme permet de facilement authentifier le document par un examen visuel humain ou

automatiquement à l'aide d'un algorithme. Cette technologie est développée par l'entreprise française, SURYS.

Figure 20 : exemple d'un hologramme DID® sous deux angles de vision.

Concernant les contre-mesures qui sont utilisées pour les visages, elles reposent principalement sur des tests de *liveness*. Un test de *liveness* facial est un moyen de vérifier qu'un utilisateur est physiquement présent et n'utilise pas une photographie ou une autre méthode d'usurpation pour contourner les mesures de sécurité. Les systèmes de reconnaissance faciale sont souvent trompés par les attaques de présentation, mais un test de *liveness* peut aider à garantir que l'utilisateur est bien celui qu'il prétend être. Il existe différentes méthodes pour réaliser un test de *liveness*, mais elles consistent toutes essentiellement à s'assurer que l'utilisateur participe activement au processus et ne se contente pas de présenter une image statique.

Une méthode courante pour réaliser un test de *liveness* facial consiste à demander à l'utilisateur d'effectuer une série d'actions spécifiques, comme cligner des yeux ou sourire. Le système utilise ensuite ses algorithmes de reconnaissance faciale pour vérifier que l'utilisateur effectue réellement l'action demandée.

2.4.2 Méthodes de détection

Les méthodes de détection sont différentes des contre-mesures. Les contre-mesures ont pour objectif de faciliter ou permettre la détection de falsification. Nous présentons dans cette section les méthodes de détection du champ d'études de la criminalistique des images numériques. D'autres méthodes plus spécifiques seront détaillées dans les chapitres suivants.

Il existe un grand nombre de méthodes pour détecter les falsifications d'une image numérique (Figure 21).

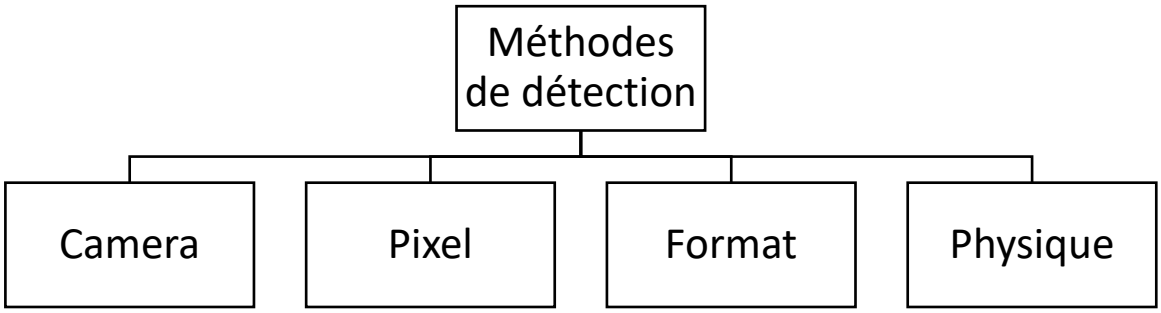

Figure 21 : diagramme des différentes méthodes de détection.

Méthode de détection basée sur la caméra

La détection par caméra est une technique criminalistique d'image numérique qui utilise la signature unique du capteur de la caméra pour identifier si une image a été altérée. La signature unique du capteur d'une caméra est générée par la façon dont le capteur capture la lumière. Chaque caméra a une façon différente de capturer la lumière, ce qui entraîne une signature différente. Cette signature peut être utilisée pour déterminer si une image a été modifiée de quelque façon que ce soit. Il existe plusieurs façons de détecter si une image a été altérée à l'aide de la détection par caméra. Une méthode consiste à comparer les signatures de deux images. Si les signatures correspondent, alors il est probable que les images soient authentiques. Si les signatures ne correspondent pas, alors il est probable que les images ont été trafiquées. Une autre méthode consiste à examiner les motifs de bruit dans une image. Les motifs de bruit sont générés par le capteur de la caméra et sont uniques à chaque caméra. En examinant les motifs de bruit dans une image, il est possible de déterminer si l'image a été modifiée.

Méthode de détection basée sur la caméra

Les méthodes basées sur les pixels sont des algorithmes criminalistiques d'images numériques qui analysent une image au niveau des pixels. Ces méthodes sont souvent utilisées pour détecter les falsifications par copie-déplacement, c'est-à-dire lorsqu'une partie d'une image est copiée et collée dans une autre partie de la même image. Les méthodes basées sur les pixels fonctionnent en examinant les différences de valeurs de couleur et d'intensité entre les pixels voisins. Si un grand groupe de pixels a des valeurs identiques ou similaires, il est probable qu'ils

ont été copiés d'un autre endroit de l'image (ou d'une autre image entièrement). Il existe de nombreux algorithmes de détection basés sur les pixels, chacun ayant ses propres forces et faiblesses. Parmi les algorithmes les plus populaires, on trouve les méthodes basées sur la DCT, les méthodes basées sur la DWT et les méthodes basées sur les ondelettes. Le meilleur algorithme pour détecter une contrefaçon dépend souvent du type de contrefaçon tentée. Par exemple, les méthodes basées sur la DCT sont plus efficaces pour détecter les falsifications par copie-déplacement, tandis que les méthodes basées sur les DWT sont plus efficaces pour détecter les falsifications par *morphing*.

Méthode de détection basée sur le format

La méthode basée sur le format utilise la structure des images numériques pour détecter les falsifications. Cette méthode permet d'analyser l'en-tête du fichier, le format et les artefacts de compression pour déterminer si l'image a été manipulée. Cette méthode peut être utilisée pour détecter les manipulations simples, comme les *splicing* et le redimensionnement, ainsi que les falsifications plus complexes. Cependant, cette méthode n'est pas infaillible, car certains faussaires savent comment manipuler les images d'une manière qui ne laisse aucun signe révélateur. Malgré tout, la méthode basée sur le format est souvent utilisée en combinaison avec d'autres méthodes, comme l'analyse statistique ou l'inspection visuelle, pour augmenter les chances de détecter les falsifications d'images.

Méthode de détection basée sur la physique

Les méthodes basées sur la physique sont des techniques criminalistiques d'images numériques qui utilisent les propriétés physiques des images pour détecter les falsifications. Cette approche est souvent utilisée pour identifier les photos retouchées ou manipulées, ainsi que pour déterminer la caméra source d'une image. Ce type de technique repose sur le fait que chaque caméra possède une "empreinte digitale" unique basée sur les caractéristiques de son objectif et de son capteur. En analysant ces propriétés physiques, il est possible de détecter si une image a été modifiée ou si elle a été prise avec un type particulier de caméra. Une méthode courante de détection basée sur les propriétés physiques est appelée analyse du niveau d'erreur (ELA). Cette approche utilise des algorithmes de compression avec perte pour trouver les zones d'une image qui ont été modifiées. La théorie est que lorsqu'une image est modifiée, les algorithmes de compression produisent des résultats différents pour les zones modifiées et pour

les zones non modifiées. En comparant les deux, il est possible de détecter les modifications et même d'estimer le degré de modification.

2.4.3 Régulations

Dans le but de normaliser les documents d'identités biométriques et de sécuriser les systèmes de vérification d'identité à distance, plusieurs organismes officiels ont établis des réglementations et des normes à suivre.

Le règlement eIDAS

L'eIDAS est le cadre de l'Union européenne pour l'identification électronique, l'authentification et les services de confiance. L'eIDAS vise à faciliter l'utilisation transfrontalière des services électroniques par les entreprises, les citoyens et les autorités publiques. Le cadre définit des règles communes pour les e-ID, les signatures électroniques et les sceaux électroniques, et fournit un mécanisme de reconnaissance pour les e-ID délivrés par les pays de l'UE. L'eIDAS établit également un cadre pour les fournisseurs de services d'e-KYC. Les documents d'identité électroniques délivrés dans le cadre d'eIDAS sont basés sur des normes internationales et sont interopérables dans toute l'UE. Les documents d'identité électroniques conformes à eIDAS peuvent être utilisés pour accéder à des services en ligne dans d'autres pays de l'UE. Le règlement eIDAS est entré en vigueur le 1er juillet 2016. Ce règlement s'inscrit dans le cadre de la stratégie du marché unique numérique de l'UE, qui vise à faciliter l'accès des entreprises et des citoyens aux services en ligne dans l'UE. L'eIDAS propose des principes de bases pour la sécurité des systèmes e-KYC. D'autres réglementations plus détaillées au niveau national ont été récemment proposées.

Prestataires de vérification d'identité à distance (PVID)

En France, L'ANSSI a produit un ensemble d'exigences pour les Prestataires de vérification d'identité à distance (PVID) afin d'assurer un haut niveau de confiance dans l'identité des individus interagissant avec les services d'identification en ligne. Ces exigences tiennent compte à la fois des éléments technologiques et organisationnels, et sont applicables indépendamment du fait que l'identité de l'individu soit vérifiée par voie électronique ou par d'autres moyens. L'une des contre-mesures du PVID est la demande d'une vidéo comme moyen de vérifier l'identité d'un individu dans un système e-KYC. La vidéo est une preuve plus

sécurisée qu’une simple image. Cette solution aide à renforcer la sécurité en rendant plus difficile pour les acteurs malveillants de se faire passer pour des utilisateurs authentiques.

L'Office fédéral de la sécurité de l'information (BSI)

Une démarche similaire a été proposée en Allemagne par l'Office fédéral de la sécurité de l'information (BSI). Le BSI fournit des conseils sur les meilleures pratiques en matière de sécurité de l'information, certifie des produits et services et mène des recherches sur les nouvelles technologies de sécurité. Ils ont proposé une certification pour les produits et services d'authentification d'identité à distance.

L'Organisation internationale de normalisation (ISO)

L'Organisation internationale de normalisation (ISO) est une organisation qui élabore et publie des normes pour un grand nombre de produits et de services. La certification ISO est une démarche volontaire qui apporte la preuve de l'engagement d'une entreprise en matière de qualité et de satisfaction des clients.

L'un des domaines où la certification ISO peut être particulièrement bénéfique est celui de la technologie de reconnaissance faciale. En obtenant la certification ISO, les entreprises attestent que leurs systèmes de reconnaissance faciale répondent aux normes internationales les plus strictes en matière de précision et de fiabilité. Les consommateurs peuvent ainsi avoir confiance dans l'utilisation de ces technologies.

L’ISO normalise également les documents d’identités. Les caractéristiques des CNI vont être standardisé en 1985 pour faciliter leur authentification. C'est la norme ISO/IEC 7810 qui définit les caractéristiques physiques des cartes d'identité. Elle est définie également des normes pour d'autres documents d'identité comme le permis de conduire et le passeport. La norme est élaborée par l’ISO et la Commission électrotechnique internationale (CEI).

En France, c'est l'imprimerie nationale qui est chargée d'imprimer les documents d'identité gouvernementaux. L'imprimerie nationale a une longue histoire de production de produits imprimés de haute qualité. Au XVIe siècle, l'entreprise était chargée d'imprimer les documents officiels du royaume de France. Aujourd'hui, l'entreprise continue de produire une large gamme de produits imprimés ainsi que les documents d'identité d'état (passeports, CNI, permis de conduire). L'entreprise est également connue pour ses solutions d'impression

sécurisée. Elle fournit des services aux gouvernements et aux entreprises du monde entier qui ont besoin de protéger leurs documents contre la contrefaçon et d'autres formes de fraude.

2.5 Apprentissage profond

L'apprentissage profond est un sous-ensemble de l'apprentissage automatique qui s'inspire de la structure et du fonctionnement du cerveau. Les algorithmes d'apprentissage profond sont capables d'apprendre des données d'une manière similaire à la façon dont les humains apprennent. Ce type d'apprentissage est utilisé pour résoudre des problèmes complexes qui sont difficiles à résoudre par les algorithmes d'apprentissage automatique traditionnels.

2.5.1 Vision par ordinateur

Le système visuel humain est l'un des systèmes sensoriels les plus complexes et les plus sophistiqués du corps humain. Il nous permet de voir le monde qui nous entoure avec des détails considérables et peut même nous fournir des informations sur le monde au-delà de notre champ de vision. Le système visuel est composé d'un certain nombre de parties distinctes, chacune jouant un rôle essentiel pour nous permettre de percevoir. L'œil humain est l'organe de la vue, mais il ne constitue qu'une partie du système visuel. La fonction de l'œil est de capter la lumière et de la convertir en signaux électriques qui sont envoyés au cerveau. Le cerveau interprète ensuite ces signaux et crée une image que nous pouvons distinguer.

Le cortex visuel primaire est responsable du traitement des informations de base sur une image, telles que sa couleur, sa taille et sa forme. De là, les informations sont envoyées à d'autres parties du cerveau. Ces zones du cerveau sont responsables d'un traitement plus complexe, comme l'identification d'objets et le souvenir de leur apparence.

La vision par ordinateur est un domaine de l'informatique qui traite de l'extraction d'informations de haut niveau à partir d'images numériques. C'est l'une des technologies de base de la reconnaissance faciale et d'autres applications technologiques ayant attiré l'attention ces dernières années.

Les techniques d'apprentissage profond se sont révélées efficaces pour la localisation automatique de documents. Dans cette tâche, l'objectif est de localiser automatiquement des documents dans une image. Étant donné une image d'entrée, l'objectif est de détecter et de classer le texte dans l'image, et de produire les coordonnées des boîtes de délimitation du texte.

Un réseau de neurones convolutif profond (CNN) est utilisé pour détecter et classer le texte dans les images naturelles. Le CNN est entraîné sur un grand ensemble de données d'images avec du texte, et peut apprendre à généraliser à de nouvelles images. Nous utilisons ensuite le réseau CNN pour détecter le texte dans les nouvelles images et fournir les coordonnées des boîtes de délimitation du texte. Cette approche s'est avérée efficace pour la localisation automatique de documents, et peut être utilisée pour localiser du texte dans une variété de langues.

L'apprentissage profond peut être utilisé pour la vérification de l'authenticité des documents. En analysant les structures des documents, l'apprentissage profond peut fournir un verdict sur la probabilité que le document soit authentique ou non. Cela peut être utile dans les cas où l'on soupçonne l'authenticité du document lors de la vérification de l'identité d'une personne. Il peut également être utilisé pour la reconnaissance optique de caractères (OCR). L'OCR est le processus de conversion d'images de texte en texte numérique. Elle peut être utilisée pour convertir des documents numérisés, des images de texte manuscrit ou même des photos de texte en texte numérique éditable et consultable. L'apprentissage profond est bien adapté aux tâches d'OCR car il peut apprendre à reconnaître des modèles de caractères dans les images.

La vision est un domaine d'intérêt depuis des décennies, mais ce n'est qu'à la fin du XXe siècle que la vision par ordinateur a commencé à se développer en tant que discipline à part entière. Les travaux de David Marr [9], qui a défini dans les années 1970 un cadre permettant de comprendre le fonctionnement de la vision au niveau informatique, ont constitué l'un des premiers fondements. Les idées de Marr ont été développées dans les années 1980 par un groupe de chercheurs japonais qui les ont appliquées au problème de la reconnaissance faciale. Ces travaux ont posé les bases des systèmes actuels de reconnaissance faciale.

2.5.2 Origine des réseaux de neurones

Les origines de l'apprentissage profond remontent aux réseaux de neurones artificiels, qui sont un type d'algorithme d'apprentissage automatique. Les réseaux de neurones artificiels ont été développés pour la première fois dans les années 1950, mais ils n'ont pas connu un grand développement à cette époque. Dans les années 1980 et 1990, les réseaux de neurones artificiels ont connu un regain d'intérêt et de nouveaux algorithmes ont été développés pour apprendre

plus efficacement à partir des données. Ces nouveaux algorithmes ont servi de base au développement de l'apprentissage profond.

Le neurone artificiel est un modèle mathématique d'un neurone biologique (Figure 22). Il a été introduit pour la première fois en 1943 par Warren McCulloch et Walter Pitts, dans [10]. Les neurones artificiels ont ensuite été très utilisés en intelligence artificielle, en apprentissage automatique et dans d'autres domaines. En 1949, Donald Hebb a proposé un modèle permettant aux neurones artificiels d'apprendre par l'expérience [11].

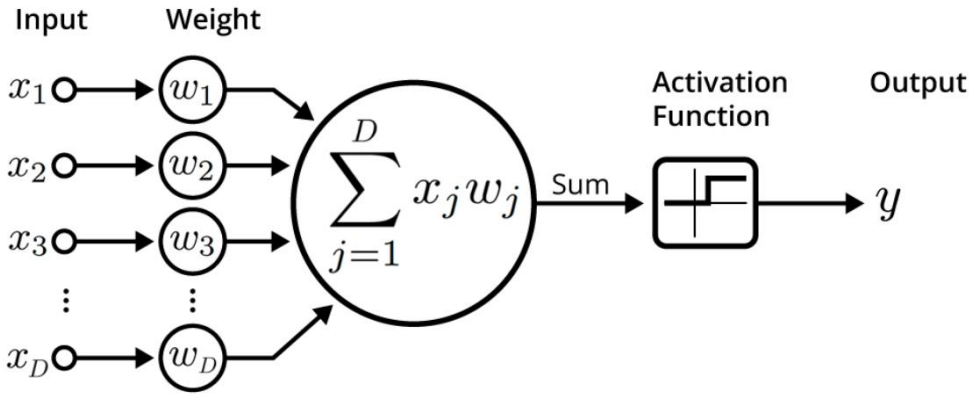

Figure 22 : Exemple d'un neurone artificiel.

Dans les années 1950 et 1960, les réseaux de neurones artificiels ont été développés en utilisant une approche d'algèbre linéaire. Cette approche était limitée par le fait qu'elle ne pouvait apprendre que des relations linéaires. Afin de surmonter cette limitation, des réseaux neuronaux artificiels ont été développés pour apprendre des relations non linéaires. Ces réseaux de neurones artificiels sont connus sous le nom de perceptrons multicouches (MLP).

Le premier réseau de neurones artificiels à apprendre avec succès des relations non linéaires est le réseau de neurones artificiels à rétropropagation. L'algorithme de rétropropagation a été développé par Rumelhart, Hinton et Williams en 1986 [12]. Cet algorithme est utilisé pour calculer l'erreur à chaque couche du réseau neuronal artificiel, puis ajuste les poids du réseau de neurones artificiels en conséquence.

Yann Lecun est un informaticien français qui a contribué de manière significative au domaine des réseaux de neurones artificiels. Il est considéré comme l'un des fondateurs des réseaux de neurones modernes, et ses travaux ont contribué à façonner le domaine tel que nous le connaissons aujourd'hui.

Les premiers travaux de Lecun étaient axés sur la construction de dispositifs capables d'apprendre à reconnaître des caractères manuscrits [13]. Ses travaux dans ce domaine ont conduit au développement du système commercial de reconnaissance de l'écriture manuscrite appelé UNLV System, qui est largement répandu. Au début des années 1990, Lecun a commencé à explorer l'utilisation des réseaux de neurones pour la reconnaissance des images. Ses travaux dans ce domaine ont abouti au développement du réseau de neurones LeNet-5 [14]. L'architecture du réseau de neurone est illustrée à la Figure 23.

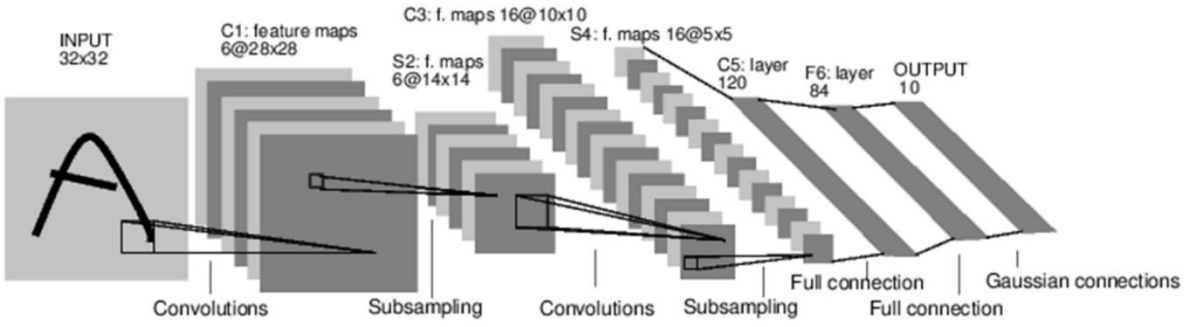

Figure 23 : architecture du CNN LeNet-5

2.5.3 Réseaux de neurones convolutifs

Dans les années 1980, le chercheur japonais Kunihiko Fukushima a mis au point un réseau de neurones appelé néocognitron [15]. Le néocognitron s'inspire de la structure du système visuel humain et est capable d'apprendre automatiquement des caractéristiques à partir de données. Il s'agissait d'une percée majeure dans la recherche sur les réseaux neuronaux, qui a ouvert la voie au développement des réseaux CNN.

LeCun s'est inspiré des travaux de Fukushima sur le néocognitron développer LeNet-5. Les CNN sont depuis devenus l'un des types de réseaux de neurones les plus populaires pour la reconnaissance d'images et sont utilisés dans une variété d'applications dont la reconnaissance faciale et la détection des falsifications d'images numériques.

Les réseaux de neurones convolutifs sont un type de réseau de neurones artificiels qui sont utilisés pour reconnaître des modèles dans les données. Les CNN sont similaires à d'autres types de réseaux neuronaux, mais ils ont une structure destinée à tirer parti de la configuration spatiale des données.

Les CNN sont constitués d'une série de couches, chacune d'entre elles étant composée d'un ensemble de neurones. La première couche d'un CNN est généralement une couche convolutive. Cette couche est destinée à extraire les caractéristiques des données. La couche suivante est généralement une couche de *pooling*, qui est chargé de sous-échantillonner les données. Enfin, la dernière couche d'un CNN est généralement une couche entièrement connectée, qui sert à faire des prédictions basées sur les caractéristiques extraites par les couches convolutionnelles.

Il existe de nombreux types de réseaux neuronaux. Parmi les architectures CNN les plus populaires, on peut citer :

- AlexNet [16]
- VGG [17]
- ResNet [18]
- Inception [19]
- Xception [20]
- DenseNet [21]

Alexnet a été développé par Alex Krizhevsky. L'objectif de ses travaux consistait à améliorer l'état de l'art existant en matière de classification d'images. L'approche de Krizhevsky consistait à augmenter la taille et la profondeur du réseau, qui était auparavant limité par la quantité de données et la puissance de calcul disponibles. Alexnet est un réseau beaucoup plus grand que LeNet, avec cinq couches convolutionnelles et trois couches entièrement connectées. Il utilise également une fonction d'activation différente, appelée ReLU, qui a démontré qu'elle permettait d'améliorer la vitesse d'apprentissage.

Les résultats de l'Alexnet ont été importants : il a remporté le 2012 ImageNet Large Scale Visual Recognition Challenge avec une marge de plus de 10 %. Il s'agissait d'un accomplissement considérable, et cela a démontré que les réseaux de neurones pouvaient être employés pour des tâches bien plus importantes que la simple classification d'images.

Le réseau VGG est composé d'une série de couches convolutionnelles et de *pooling*, qui extraient les caractéristiques des images, et d'une série de couches entièrement connectées, qui utilisent ces caractéristiques pour classer les images. Le réseau VGG a été utilisé pour la

première fois pour remporter ImageNet en 2014, et a été largement utilisé dans les recherches ultérieures.

La rétropropagation a permis d'utiliser les réseaux neuronaux pour des tâches plus complexes, comme la reconnaissance d'images. Cependant, ces systèmes étaient encore limités par la quantité de données à partir desquelles ils pouvaient apprendre. Cela a changé en 2015 avec l'introduction de Resnet, un nouveau type de réseau de neurones qui apprend beaucoup mieux à partir des données. Resnet a été conçu par une équipe de chercheurs de Microsoft Research, et il s'est avéré beaucoup plus efficace pour apprendre à partir des données que les architectures de réseaux de neurones précédentes.

2.5.4 Réseaux de neurones génératifs

Les réseaux de neurones ne sont pas utilisés seulement pour automatiser des tâches complexes comme la reconnaissance d'images. Il existe des architectures permettant de générer de nouvelles données permettant d'attaquer numériquement les systèmes de vérification d'identité à distance. Il existe deux architectures principales de réseaux génératifs : les auto-encodeurs et les réseaux de neurones antagonistes (GANs).

Le premier auto-encodeur a été proposé par Hinton et Salakhutdinov en 2006 [22]. Les auto-encodeurs ont depuis été beaucoup utilisés dans de nombreux domaines tels que la vision par ordinateur, le traitement du langage naturel et les systèmes de recommandation. Les auto-encodeurs sont des réseaux de neurones qui sont utilisés pour apprendre des représentations efficaces des données. L'objectif d'un auto-encodeur est d'apprendre une représentation (codage) des données en entrée qui est compressée, tout en préservant autant d'informations que possible. La partie encodeur de l'auto-encodeur apprend à comprimer les données d'entrée, tandis que la partie décodeur apprend à reconstruire les données d'entrée à partir de la représentation comprimée. Les auto-encodeurs sont soit non supervisés, soit supervisés. Les auto-codeurs non supervisés sont entraînés en utilisant uniquement les données d'entrée, tandis que les auto-codeurs supervisés sont entraînés en utilisant à la fois les données d'entrée et les labels correspondants. Les auto-encodeurs sont souvent utilisés comme étape de pré-entraînement pour les réseaux de neurones profonds. En entraînant au préalable un auto-encodeur, le réseau de neurones peut apprendre à mieux reconnaître les modèles dans les données. Les auto-encodeurs peuvent également être utilisés pour la réduction de la dimension. En apprenant une représentation comprimée des données, les auto-encodeurs peuvent réduire

la dimension des données tout en préservant les informations les plus importantes. Les auto-encodeurs ont de nombreuses applications dans différents domaines. En vision par ordinateur, les auto-encodeurs sont souvent utilisés pour le débruitage et la super-résolution des images. Les auto-encodeurs sont un outil puissant pour apprendre des représentations efficaces des données et ont de nombreuses applications dans différents domaines. L'architecture d'un auto-encodeur est illustré Figure 24.

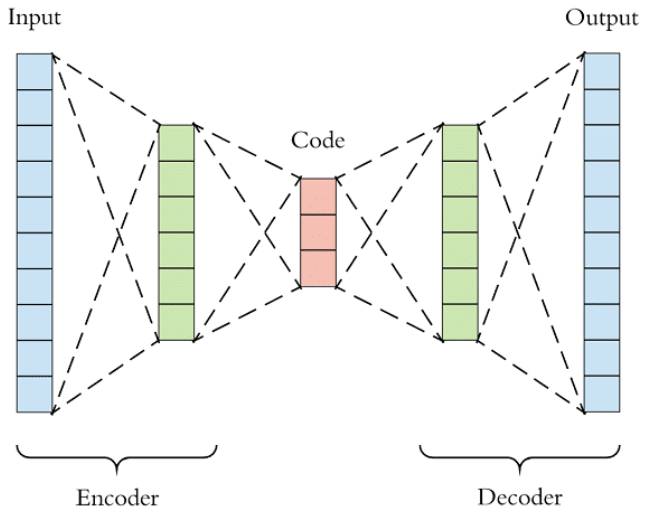

Figure 24 : architecture du CNN LeNet-5

Les GAN sont un type de réseau de neurones qui sont utilisés pour générer de nouvelles données. Les GAN sont exploités pour générer des images, des vidéos et même du texte. Les GAN ont été introduits pour la première fois par Ian Goodfellow en 2014 [23]. Goodfellow a été inspiré par un article qui montrait comment générer des images à partir de bruit [24]. Il a réalisé que si l'on pouvait générer des images à partir de bruit, alors on pouvait aussi générer de nouvelles données. Les GAN sont constitués de deux réseaux de neurones : un générateur et un discriminateur. Le générateur génère de nouvelles données, tandis que le discriminateur tente de distinguer les données réelles des fausses. Les deux réseaux de neurones sont entraînés en même temps et, au fur et à mesure de leur entraînement, le générateur s'améliore dans la génération de nouvelles données. Les GAN ont permis de générer des images réalistes d'objets, de personnes et de visages. Ils ont aussi servi à concevoir des vidéos et du texte. Les GAN sont un outil puissant pour générer de nouvelles données, et ils ne font que s'améliorer au fur et à mesure de leur développement (Figure 25).

Figure 25 : progrès des GANs pour la génération de visages.

Le processus d'entraînement des GANs peut être considéré comme un jeu entre le générateur et le discriminateur. Le générateur essaie de générer des échantillons de données suffisamment réalistes pour tromper le discriminateur, tandis que le discriminateur essaie de distinguer les échantillons de données réels des faux. Le réseau de neurones du générateur est généralement un réseau de neurones convolutif (CNN) ou un réseau de neurones entièrement connecté (FCN). Le réseau de neurones discriminateur est habituellement un CNN ou un FCN. La Figure 26 illustre l'architecture de base d'un GAN.

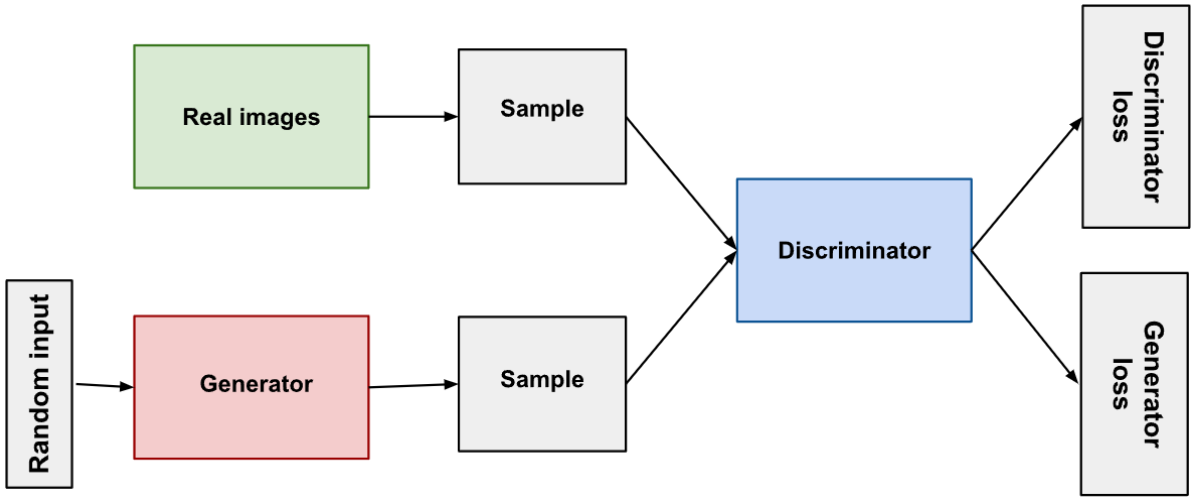

Figure 26 : architecture d'un GAN.

3. Robustesse aux attaques par présentation

3.1 Contexte

Les attaques de présentation constituent la première catégorie de menaces envers un système e-KYC. Il s'agit de méthodes qui attaquent le système physiquement. L'attaquant présente des preuves physiques falsifiées à la caméra du système. Les principales méthodes d'attaques par présentation du visage ont déjà été identifiées dans la littérature. Trois catégories d'attaques sont jugées principales :

- Attaques par photo
- Attaques par relecture de vidéo
- Attaques par masque 3D

Avec le développement technologique, d'autres moyens d'attaquer les systèmes e-KYC ont été mis au point. Ces attaques constituent un deuxième groupe, que l'on peut qualifier de secondaire, car l'utilisation de ces attaques est moins fréquente, dans la mesure où elles sont plus difficiles à mettre en œuvre. Nous identifions deux catégories d'attaques secondaires :

- Attaques par maquillage
- Attaques physiques antagonistes

Dans le cas spécifique du document d'identité requis pour l'authentification, un dernier type d'attaque est à envisager : le morphing du visage.

Un système e-KYC est susceptible d'être la cible de toutes ces attaques. L'objectif pour l'attaquant est d'usurper le visage d'un autre utilisateur. Comme l'authentification se déroule à distance sans aucune supervision humaine, l'attaquant a la liberté d'essayer de tromper le système plusieurs fois. De plus, il peut utiliser de nombreux outils performants pour élaborer une attaque de présentation réaliste. Une imprimante lui permet de matérialiser une attaque par photo, les écrans des smartphones lui permettent d'attaquer le système en diffusant une vidéo et les imprimantes 3D lui permettent de fabriquer des masques 3D.

De nombreuses méthodes de détection ont été proposées dans la littérature. La plupart des algorithmes traditionnels sont basés sur des tests de *liveness*. On demande à l'utilisateur

d'effectuer une action avec son visage, de cligner des yeux, de sourire, de tourner la tête, etc. D'autres méthodes sont basées sur une analyse physiologique ou sur des descripteurs classiques (LBP, SIFT, HOG etc.).

Dans ce chapitre, nous étudions l'impact des attaques de présentation principale contre un système e-KYC sur un appareil mobile. Nous proposons une méthode de détection de ces attaques basée sur la corrélation des métriques faciales avec le positionnement dans le temps du smartphone de l'utilisateur. Plutôt que de demander à l'utilisateur d'effectuer uniquement une action avec son visage, nous lui demandons également de déplacer son smartphone autour de son visage. L'avantage de cette méthode est qu'elle peut être implémentée en temps réel, ce qui est difficile à réaliser avec les méthodes basées sur les descripteurs.

3.2 État de l'art

Les attaques de présentation sont réalisées dans le monde réel, c'est-à-dire de manière physique. En reprenant le diagramme représentant un système e-KYC, les attaques de présentation ont lieux au niveau du capteur d'acquisition, c'est-à-dire la caméra. Ce type d'attaque est également qualifié d'attaque directe (Figure 27).

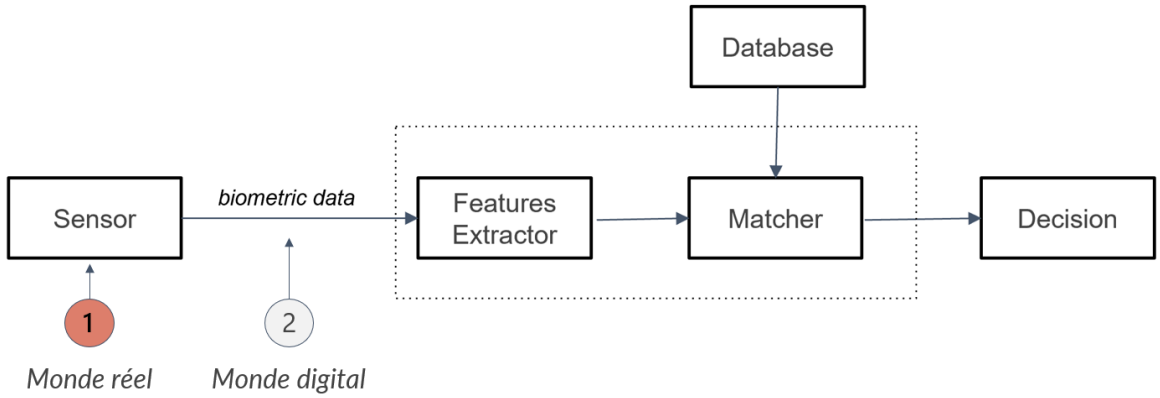

Figure 27 : localisation des attaques de présentation sur un système e-KYC.

3.2.1 Description

La popularisation des réseaux sociaux a favorisé le partage de photos de visages d'individus sur Internet. Il est alors très simple pour un attaquant d'obtenir la photo d'une personne dont il veut usurper l'identité [1]. Il lui suffit de la présenter sur un support physique sans fournir un effort quelconque.

L'attaque la plus facile à réaliser est l'attaque par présentation de photos [2] [3]. Bien que ce type d'attaque puisse sembler négligeable à première vue, les nouvelles imprimantes permettent d'imprimer des visages en haute résolution et avec beaucoup de détails. Si les attaquants ne disposent pas d'une imprimante, ils peuvent aussi simplement afficher la photo sur l'écran d'un smartphone [4] [5]. Une version plus avancée de ce type d'attaque consiste à créer un masque 2D [6]. Plutôt que de simplement imprimer la photo complète d'un visage, l'attaquant procédera à des modifications de celle-ci. Il découpera le visage en suivant ses contours pour supprimer l'arrière-plan. Il réalisera également des trous au niveau des yeux et éventuellement de la bouche pour tromper les tests de *liveness* les plus courants (clignement des yeux et sourire). Ce type d'attaque a un coût très faible, il suffit à l'attaquant d'avoir accès à une imprimante ou de posséder un smartphone.

Ces dernières années, une nouvelle technologie d'imprimante a été développée et est accessible au grand public : L'impression 3D. Les imprimantes 3D permettent aux attaquants de réaliser des masques tridimensionnels qui, contrairement aux masques bidimensionnels, incluent des niveaux de profondeur et des zones d'ombre [7] [8]. Des algorithmes sont utilisés pour convertir une image bidimensionnelle en un modèle tridimensionnel. C'est à partir de ce modèle tridimensionnel que le masque est imprimé. Le matériau le plus souvent utilisé est le plastique (résine, polyamide ou ABS). D'autres matériaux plus proches de la texture de la peau humaine existent (latex, silicone) mais nécessitent des ressources financières plus importantes et ne sont donc pas accessibles à tous. Une variante de ce type d'attaque propose de projeter la vidéo d'un visage en temps réel sur un masque 3D [9]. Ce type d'attaque est difficile à réaliser, car il nécessite un équipement spécialisé qui n'est pas accessible au grand public.

L'attaque par relecture vidéo représente la dernière grande catégorie d'attaques par présentation. L'attaquant diffuse une vidéo du visage de la victime sur un écran placé devant la caméra du système [10] [11]. La plupart des écrans de smartphones modernes ont une résolution proche de la résolution 2K et parfois 4K, ce qui apporte un haut niveau de détails et de réalisme à la vidéo. La principale difficulté de cette attaque est d'obtenir une vidéo de la victime qui puisse remplir les conditions des tests de *liveness*. Il est difficile, voire impossible, pour un attaquant d'obtenir ce type de vidéo sur Internet. Cependant, des méthodes open-source et en temps réel existent pour échanger le visage d'une personne dans une vidéo. Le coût de cette attaque est relativement faible puisque 83% de la population mondiale possède un smartphone [12] et donc un écran mobile pour afficher une vidéo.

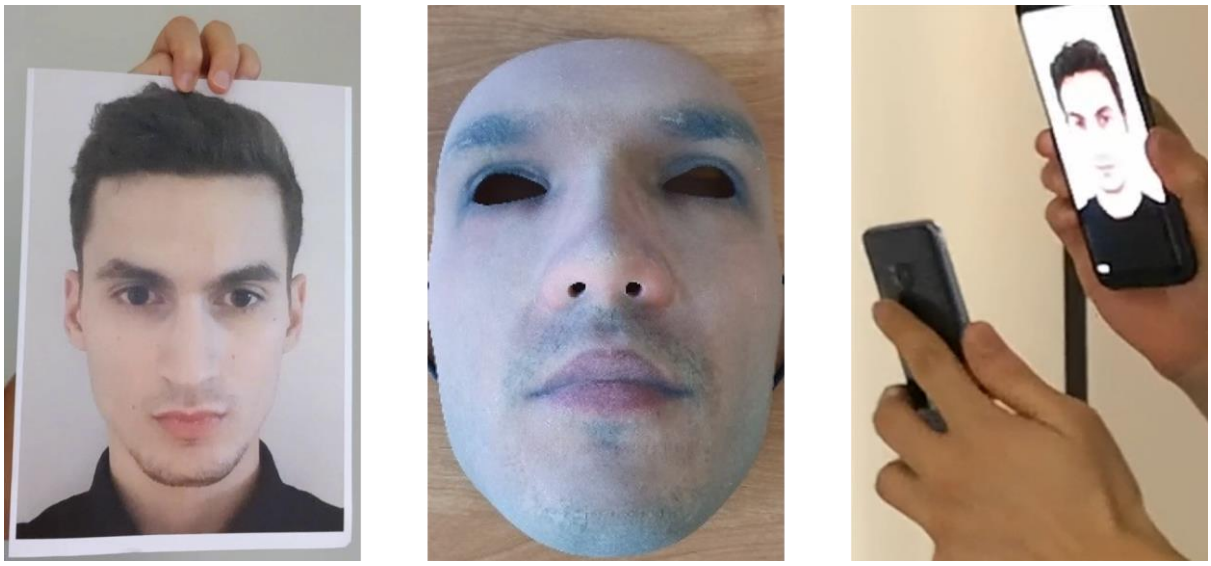

Figure 28 : les trois catégories principales d'attaques de présentation, par photo (à gauche), par masque 3D (au centre), par vidéo (à droite).

3.2.2 Méthodes de détection

Pour être robustes aux différentes attaques de présentations existantes, des méthodes de détection et des contre-mesures ont été développées et intégrées dans les systèmes e-KYC. Trois catégories se distinguent :

- Les solutions basées sur un matériel
- Les solutions basées sur un logiciel
- Les solutions basées sur l'apprentissage profond

Les solutions basées sur le matériel impliquent généralement l'utilisation d'appareils spéciaux qui peuvent capturer des images de haute qualité des visages. L'imagerie infrarouge est souvent utilisée dans les solutions matérielles, car elle permet d'obtenir une image claire du visage d'une personne, même dans des conditions de faible luminosité. Pavlidis et al. [13] ont proposé un système qui utilisait deux caméras infrarouges pour prendre des photos du visage d'une personne sous différents angles, puis utilisait un algorithme informatique pour comparer les images et rechercher des similitudes. L'imagerie thermique est un autre type de matériel qui peut être utilisé pour l'anti-spoofing des visages. Les images thermiques contiennent des informations sur la chaleur émise par un objet, qui peuvent être utilisées pour distinguer les vrais visages des faux. L. Sun, W. Huang et M. Wu [14] comparent les images de visages dans le spectre infrarouge et la lumière visible. En examinant la corrélation entre les deux images, ils sont capables de détecter si un visage est réel ou non. Les solutions basées sur des caméras

3D sont également de plus en plus courantes car elles peuvent fournir une représentation plus précise du visage d'une personne. Ces types de systèmes utilisent souvent les informations de profondeur pour distinguer les vrais et faux visages [15].

Les solutions logicielles utilisent généralement des méthodes statiques ou dynamiques [16]. Les méthodes statiques analysent des caractéristiques de l'image acquise. De nombreuses méthodes basées sur l'analyse de descripteurs d'images ont démontré de bons résultats de détections. Dans [17] J. Li et al. présentent une méthode pour détecter les visages réels en analysant des spectres de Fourier 2D. Dans [18] les auteurs ont proposé un modèle de détection de la *liveness* d'un modèle discriminatoire bilinéaire pour apprendre le sous-espace latent qui sépare les visages usurpés des vrais visages. Les auteurs de [19] ont utilisé des motifs binaires locaux (LBP) pour extraire les caractéristiques de texture des images de visage capturées. D'autres méthodes proposent d'utiliser les composants spéculaires [20], les Ondelettes de Gabor [21], ou de fusionner plusieurs descripteurs [22] [23]. Les méthodes dynamiques utilisent des informations récupérées à partir d'actions de l'utilisateur. Dans [24] [25] les auteurs proposent des méthodes de détection des attaques de présentation basées sur l'analyse du clignement des yeux. [26] et [27] utilisent une approche de *liveness* qui demandent à l'utilisateur de tourner la tête à droite ou à gauche. Une autre façon de détecter ces attaques consiste à utiliser les méthodes *eulerian video magnification* [28]. *Eulerian video magnification* est une technique permettant de révéler des détails subtils dans les vidéos. Elle peut être utilisée pour mettre en valeur des caractéristiques qui sont autrement difficiles à voir, comme de petits mouvements ou des changements de couleur permettant d'estimer le rythme cardiaque d'un visage.

Lorsque les réseaux de neurones convolutifs ont démontré de grandes performances pour les tâches de reconnaissances d'images, ils ont été employés également pour la détection d'attaques de présentation et ont démontré également des très bonnes performances. Des méthodes hybrides proposent d'utiliser des réseaux de neurones pour classifier des caractéristiques extraites de l'image [29] [30] [31]. Des méthodes entraînent des CNN sur une base de données de visages réels et de faux visages et passent en entrée du réseau directement une image ou un frame de vidéo [32] [33] [34].

3.2.3 Bases de données existantes

À mesure que l'utilisation de la technologie de reconnaissance faciale se répand, le risque d'attaques par usurpation d'identité augmente et de nombreuses bases de données ont été proposées pour aider à développer des méthodes de détection. Il existe un certain nombre de bases de données open source contre les attaques de présentation. Des *datasets* sont construites à l'aide d'une caméra RGB alors que d'autres utilisent des caméras spécialisées [35] [36] [37]. Nous couvrirons uniquement les bases de données qui ont été construites à partir de caméra RGB puisque c'est le type de caméra qui est utilisé dans les applications e-KYC. Le tableau 2 regroupe les dix *datasets* les plus cités dans la littérature.

NUAA-Spoof [38] est l'une des premières bases de données dédiées à la détection des attaques de présentation. Cette base contient uniquement des attaques de présentation par photo. YALE *Recaptured* [39] est une autre base de données qui contient uniquement des attaques de présentation par photo. Il contient près de 2 000 images de visages dans diverses conditions (par exemple, avec et sans lunettes, dans différentes conditions d'éclairage, etc.). CASIA-MFSD [40] est la première base de données qui inclue des exemples d'attaques de présentations par relecture vidéo en plus des attaques de présentation par photo. Les images sont collectées sur le Web et varient en termes de résolution, de qualité et d'arrière-plan. La base de données REPLAY-ATTACK [41] est composée de 1 000 enregistrements vidéo de 50 visages différents. Les vidéos ont été enregistrées à l'aide d'un capteur Kinect, et elles couvrent un large éventail de scénarios, des environnements intérieurs aux environnements extérieurs. La base de données MSU-MFSD [42] contient 210 vidéos de 35 sujets. Les types d'attaques de présentation de cette base comprennent les attaques par photo, les attaques de relecture et les attaques de masque 3D. Le *dataset* HKBU-MARs V2 [43] se compose de 504 séquences vidéo de haute qualité d'attaque de présentation par masque 3D en résine. MSU USSA [44] contient des attaques de présentations par photo et par relecture vidéo. Les sujets exécutent différentes expressions faciales, comme sourire ou froncer les sourcils. La base de données OULU-NPU [45] est une grande base de données composée de 2880 vidéos d'attaques de présentations par photo et par relecture vidéo. Enfin, SiW [46] et SiW-M [47] sont des *datasets* d'attaques de présentation par photo et relecture vidéo (pour la première version) et avec en plus des exemples d'attaques par masque 3D et maquillage (pour sa deuxième version).

Tableau 2: résumé des dix *datasets* publics les plus cités dédiés à la détection des attaques de présentations (I : Image, V : Vidéo).

Dataset	Année	Sujet	Réel	Faux	Photo	Vidéo	Masque 3D
NUAA	2010	15	5105 (I)	7509 (I)	✓		
Yale *Recaptured*	2011	10	640 (I)	1920 (I)	✓		
CASIA-MFSD	2012	50	150 (V)	450 (V)	✓	✓	
REPLAY-ATTACK	2012	50	200 (V)	1000 (V)	✓	✓	
MSU-MFSD	2014	35	70 (V)	210 (V)	✓	✓	
HKBU-MARs V2	2016	12	504 (V)	504 (V)			✓
MSU USSA	2016	1140	1140 (I)	9120 (I)	✓	✓	
OULU-NPU	2017	55	720 (V)	2880 (V)	✓	✓	
SiW	2018	165	1320 (V)	3300 (V)	✓	✓	
SiW-M	2019	493	660 (V)	968 (V)	✓	✓	✓

3.3 Méthode proposée

Les méthodes d'anti-spoofing ont démontré de bonnes performances dans la littérature pour détecter les attaques de présentation. Cependant, leurs faisabilités sur les appareils mobiles ne sont pas optimales, car elles ne sont pas en temps réel. Dans ce chapitre, nous proposons une nouvelle approche anti-spoofing de visage qui exploite les propriétés des capteurs de positions des smartphones avec des mesures biométriques du visage. Nous utilisons ces données pour entraîner un SVM capable de détecter les attaques de présentation avec une grande précision. La méthode repose également sur le mouvement du smartphone autour du visage de l'utilisateur (Figure 29). Nous évaluons notre méthode sur une base de données que nous avons nous-mêmes construite.

Figure 29 : mouvement du smartphone qui est demandé à l'utilisateur.

3.3.1 Extraction des points de références du visage

La première partie de notre méthode est basée sur l'analyse des points de référence du visage. Il existe une variété de méthodes de détection des points de repère du visage qui peuvent être utilisées pour identifier les points clés d'un visage. La méthode la plus courante consiste à utiliser un modèle qui adapte un masque de points de références par rapport à la forme d'un visage. Pour estimer les points de référence d'un visage, nous avons utilisé la bibliothèque dlib [48]. Dlib est une bibliothèque de vision par ordinateur populaire qui fournit un certain nombre d'algorithmes de l'état de l'art. C'est un outil performant pour détecter et analyser les visages dans les images. Il peut être utilisé pour de nombreuses applications. Le détecteur de visages de dlib est une implémentation du descripteur de caractéristiques *Histogram of Oriented Gradients* (HOG) combiné à un classificateur linéaire de machine à vecteurs de support (SVM). L'algorithme de détection des visages utilisé dans dlib est très robuste et peut gérer une grande variété d'images, y compris dans des conditions de faible éclairage et de contre-jour. Il est également capable de détecter les visages partiellement occultés ou de profil. En plus de la détection des visages, dlib peut aussi être utilisé pour la détection des points de repère. Les points de repère sont des points d'intérêt sur un visage, comme les yeux, le nez et la bouche. Dlib propose un modèle pré-entraîné permettant de détecter 68 points de référence d'un visage (Figure 30). Dlib détecte un visage et calcule les 68 points de références en une milliseconde, ce qui rend l'algorithme adapté à une implémentation en temps réel sur smartphone.

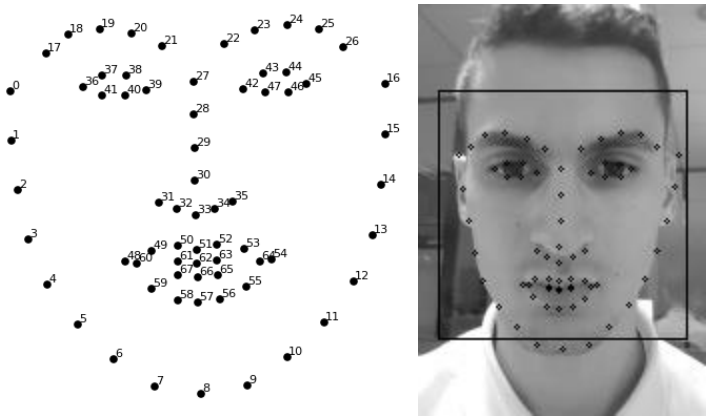

Figure 30 : les 68 points de références du visage de Dlib.

3.3.2 Extraction des données des capteurs de position du smartphone

La deuxième partie de notre méthode repose sur l'analyse des capteurs de mouvement du smartphone tout au long du mouvement qui est demandé à l'utilisateur. Les capteurs de mouvement sont de plus en plus populaires dans les smartphones. À l'intérieur de chaque smartphone se trouvent un accéléromètre et un gyroscope. Le principe d'un accéléromètre est de mesurer la quantité de force (F_s) appliquée à un appareil (A_D) pour estimer l'accélération de l'appareil. Cette relation est représentée par l'équation suivante :

$$A_D = -\left(\frac{1}{masse}\right)\sum F_s \qquad (1)$$

La force exercée par la gravité de la Terre a un impact sur l'accéléromètre, c'est pourquoi il est nécessaire de la prendre en compte dans l'équation (1). L'équation devient alors :

$$A_D = -g - \left(\frac{1}{masse}\right)\sum F_s \qquad (2)$$

Où g = 9.81 m/s²

Le gyroscope est un appareil qui mesure la vitesse angulaire. Les gyroscopes peuvent être utilisés pour mesurer la vitesse de rotation autour d'un axe. Les gyroscopes sont basés sur le principe de la conservation du moment angulaire. Lorsqu'une masse est mise en rotation autour d'un axe, elle subit une force proportionnelle à la vitesse de rotation. Cette force peut être utilisée pour mesurer la vitesse angulaire. Dans un smartphone moderne, le gyroscope mesure la rotation en rad/s autour des axes x, y et z (Figure 31).

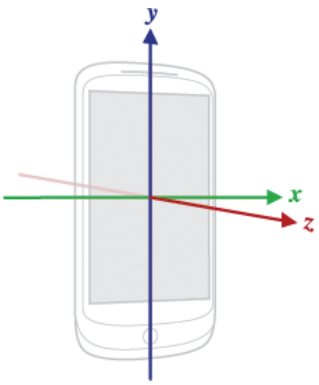

Figure 31 : les trois axes de rotation d’un smartphone.

3.3.3 Classification

Notre solution doit être exécuté sur un smartphone en temps réel sans passer par une étape de calcul sur serveur. Il est donc nécessaire d’utiliser une méthode de classification qui ne demande pas un temps de calcul élevé. Nous avons donc décidé d’utiliser une méthode d’apprentissage automatique et particulièrement l’algorithme de Machine à vecteurs de support (SVM). Une machine à vecteur de support (SVM) est un algorithme d'apprentissage automatique supervisé qui peut être utilisé pour les tâches de classification et de régression.

L'algorithme SVM est basé sur le concept de la recherche d'un hyperplan qui sépare le mieux un *dataset* en deux classes. En d'autres termes, l'algorithme est conçu pour trouver une ligne (ou plus précisément, un hyperplan) qui peut être utilisé pour diviser une base de données en deux groupes de telle sorte que les points de chaque groupe soient aussi éloignés que possible de l’hyperplan. Une fois que l'hyperplan a été trouvé, l'algorithme SVM peut alors être utilisé pour faire des prédictions sur de nouveaux points de données. Pour les tâches de classification, l'algorithme prédit la classe d'un nouveau point en fonction du côté de l'hyperplan où il se trouve. Pour les tâches de régression, l'algorithme prédit la valeur d'un nouveau point en fonction de sa distance par rapport à l'hyperplan. Un exemple d’hyperplan qui sépare deux classes est illustré à la Figure 32.

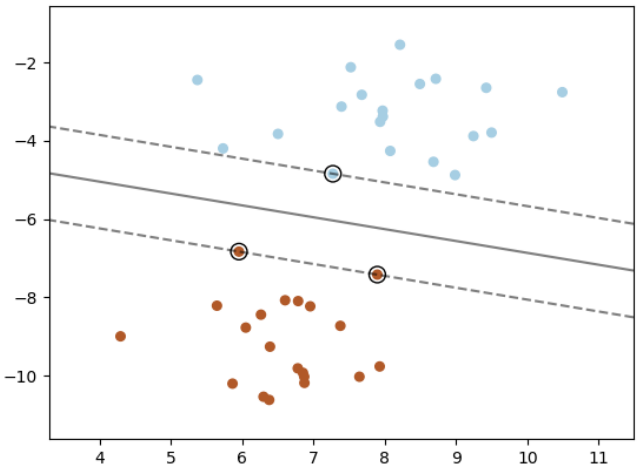

Figure 32 : exemple d'un hyperplan séparant deux classes.

3.4 Expériences et résultats

Dans cette partie nous allons présenter les différentes expériences et les résultats que nous avons obtenus de notre méthode de détection des attaques de présentation.

3.4.1 Base de données

Dans le but de pouvoir entraîner notre SVM et pour pouvoir évaluer notre méthode, la première étape a été de construire une base de données de visage réel et de différentes attaques de présentation. Le mouvement de déplacement du smartphone autour du visage ne se fait pas de la même façon pour chaque utilisateur et les données renvoyées par les capteurs de mouvement du smartphone seront donc différentes. C'est pourquoi nous avons demandé à plusieurs sujets d'effectuer des acquisitions en suivant le scénario imposé. Pour pouvoir collecter les données des capteurs de mouvements du téléphone ainsi que les points de références du visage, nous avons développé une application sur smartphone qui enregistre une vidéo durant le mouvement du smartphone ainsi que les données renvoyées par les capteurs de mouvement du smartphone. Nous avons répété ce processus d'acquisition sur des attaques de présentation par photo, par relecture vidéo et par masque 2D.

3.4.2 Expériences

Dans le but de pouvoir différencier les visages réels des attaques de présentations de notre *dataset*, nous avons commencé par chercher la meilleure métrique possible à partir des

68 points de références calculés par dlib. Nous avons évalué l'évolution des métriques au niveau des yeux ou de la mâchoire (Figure 33), mais ces informations ne permettaient pas d'obtenir des résultats satisfaisants.

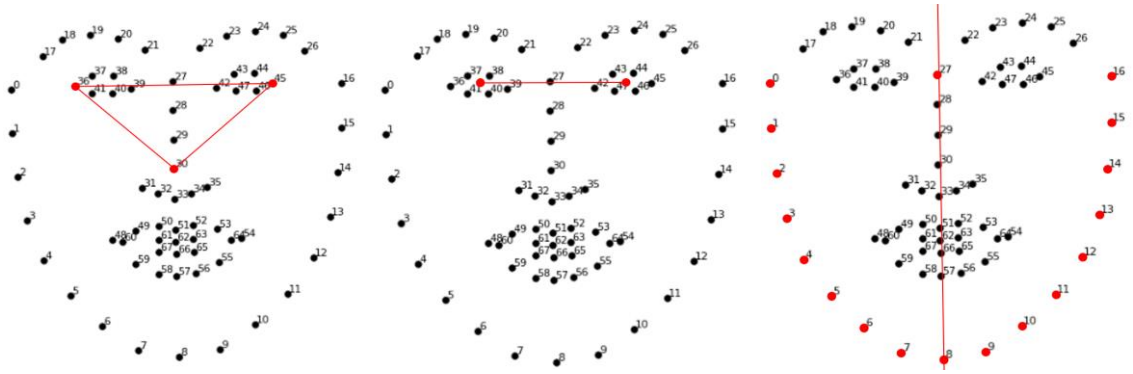

Figure 33 : différentes métriques expérimentées pour différentier les visages authentiques des attaques de présentation.

Nous avons ensuite recherché une métrique sur les points de référence du nez (Figure 34).

Figure 34 : points de référence du nez.

Après plusieurs expérimentations, nous avons déterminé que l'angle de l'arête du nez était une métrique intéressante pour détecter les attaques de présentation (Figure 35).

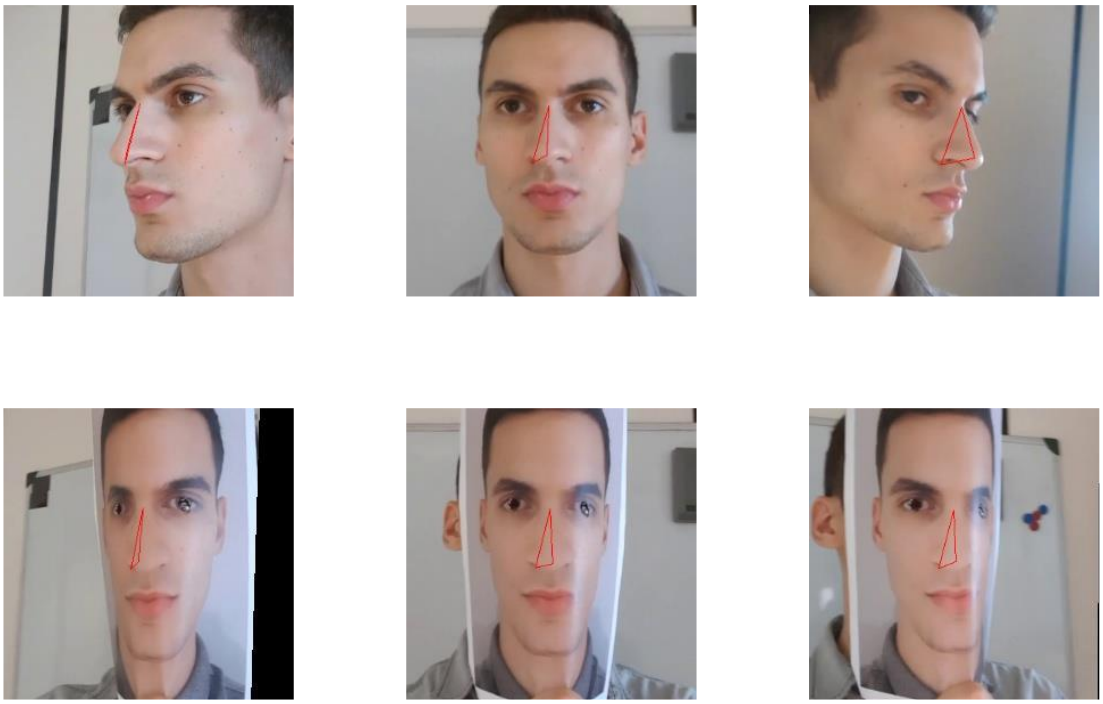

Figure 35 : évolution de l'angle de l'arête du nez sur un vrai visage (en haut) et sur un masque en 2D (en bas).

Dans le cas d'un vrai visage, l'angle de l'arête du nez évolue de manière constante durant le mouvement horizontal du smartphone. Alors que dans le cas d'une photo d'un visage imprimée ou d'un masque 2D, l'angle n'évoluera pas. Nous avons ensuite cherché à déterminer la position du smartphone (et donc de la caméra) durant le mouvement. L'idée est de vérifier si la position de la caméra est cohérente avec l'angle du nez qui est calculé. Il est possible d'estimer la position de la caméra en intégrant deux fois l'accélération. Intégrer l'accélération donne le vecteur vitesse au cours du temps $v(t)$ (Équation 3). Intégrer le vecteur vitesse donne la position au cours du temps $x(t)$ (Équation 4).

$$v(t) = \int A_D(t) \tag{3}$$

$$x(t) = \int v(t) \tag{4}$$

Nous avons donc estimé la position de la caméra au cours du mouvement du smartphone. Cependant, les accéléromètres intégrés dans les smartphones ne sont pas assez performants pour retourner des valeurs précises. Nous intégrons donc au cours du temps du bruit qui fausse la position du téléphone. Nous avons donc cherché à exploiter les informations

retournées par le gyroscope. Plutôt que d'estimer la position du téléphone au cours du temps, nous cherchons à estimer la rotation sur l'axe y du smartphone au cours du temps. Lorsqu'un utilisateur déplace horizontalement son smartphone autour de son visage, on observe une anti-corrélation entre l'évolution de l'angle du nez et l'angle du smartphone. En revanche, lors d'une attaque de présentation, cette corrélation est inexistante (Figure 36).

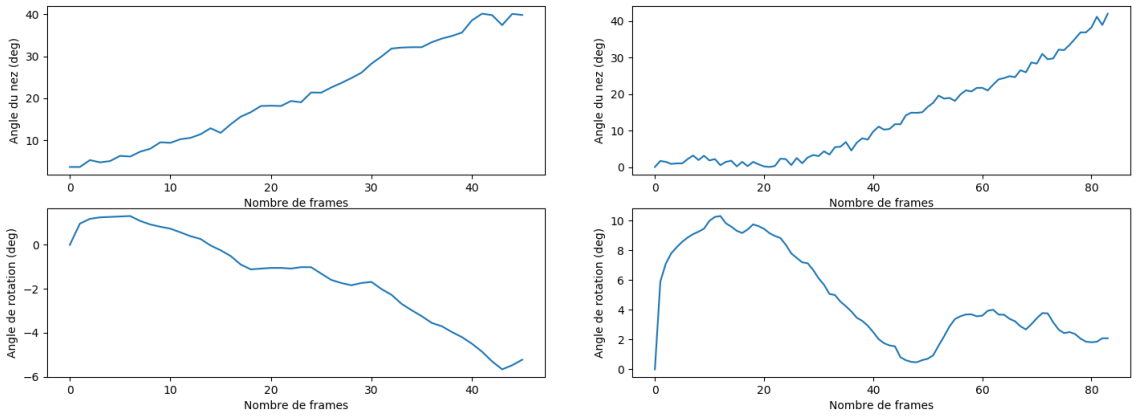

Figure 36 : graphiques de l'évolution de l'angle du nez et de l'angle du smartphone pour un visage authentique (à gauche) et une attaque de présentation par relecture vidéo (à droite).

Pour évaluer la corrélation entre les valeurs d'angle du nez et d'angle du smartphone nous employons la métrique du Coefficient de corrélation de Pearson (PCC). Le PCC est une métrique de la corrélation linéaire entre deux variables X et Y. Sa valeur est comprise entre +1 et -1, où 1 correspond à une corrélation linéaire positive complète, 0 à une absence de corrélation linéaire et -1 à une corrélation linéaire négative complète. Dans notre cas, plus le PCC est proche de -1 est plus cela indique que le visage est authentique.

3.4.3 Résultats

Pour évaluer notre méthode, nous avons entraîné un SVM sur les PCC calculés sur notre *dataset*. La pipeline de détection est illustré à la Figure 37. Dans un premier temps, une vidéo de l'utilisateur effectuant le mouvement du smartphone autour de son visage est récupérée. Ensuite, dlib est utilisé pour détecter et recadrer les visages de chaque frame de la vidéo. Dans le même temps, dlib est utilisé pour calculer les points de références du visage. L'étape suivante est de calculer pour chaque frame l'angle du nez et l'angle de rotation du smartphone. Chacune des deux valeurs est ajoutée dans un vecteur distinct. Enfin, le PCC est calculé à partir des deux vecteurs précédemment obtenus, et il est passé dans le SVM entraîné pour obtenir une décision.

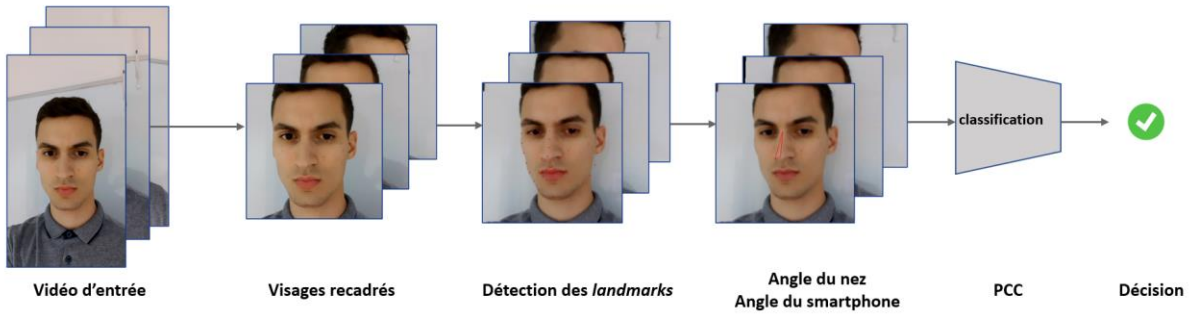

Figure 37 : pipeline de décision d'une vidéo.

Pour évaluer les performances de notre modèle, nous calculons l'aire sous la courbe (AUC). Les résultats sur notre *dataset* sont indiqués à la Figure 38. L'AUC obtenue est de plus de 0.99.

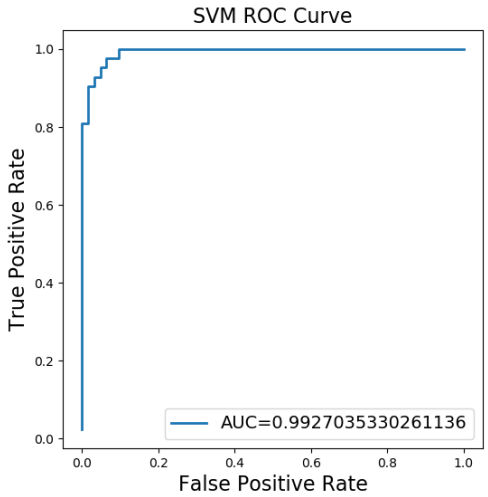

Figure 38 : AUC calculée du SVM.

3.4.4 Limites

Deux limites de cette méthode peuvent être établies. La première concerne les attaques de présentation par masque 3D. Ce type d'attaque fait échouer notre méthode de détection, puisque basée sur la profondeur du visage calculée à partir de l'angle du nez. La deuxième limite concerne les attaques de présentation par relecture vidéo. Bien que notre méthode parvienne à détecter les attaques par relecture vidéo, il est possible pour un attaquant de tromper facilement le système. L'attaquant doit au préalable préparer une vidéo en suivant le mouvement horizontal imposé, puis lors de l'authentification, il doit simuler de nouveau ce mouvement (Figure 39). Si le mouvement effectué est en correspondance avec le mouvement de vidéo rejouée, la tentative d'authentification sera accordée.

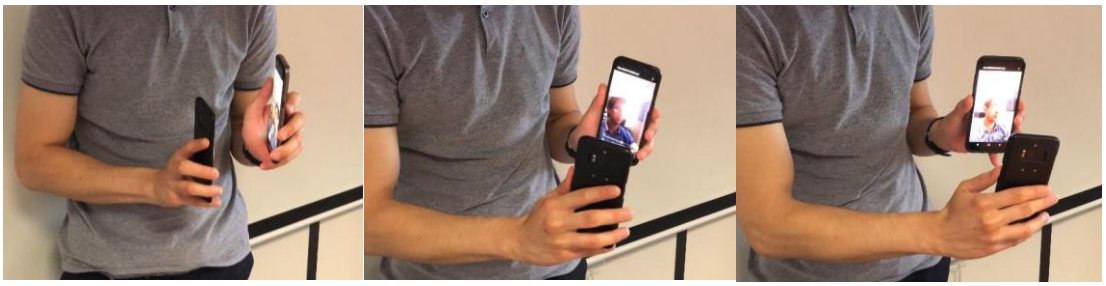

Figure 39 : mouvement du smartphone avec une attaque de relecture vidéo préparée en amont.

3.5 Conclusion

Dans ce chapitre, nous avons présenté les attaques de présentation auxquelles les systèmes d'authentification basés sur la reconnaissance faciale sont vulnérables. Nous avons également discuté des méthodes de détection et des contre-mesures existantes. Une nouvelle contre-mesure basée sur l'analyse des points de référence biométriques du visage et l'angle de rotation du smartphone est présentée.

Cette nouvelle méthode est principalement destinée aux smartphones. Il est demandé à l'utilisateur de déplacer son appareil autour de son visage en suivant un mouvement horizontal sans bouger son visage. De cette manière, il est possible d'analyser le comportement des points de référence sur le visage pendant le mouvement et ainsi d'évaluer la profondeur du visage présenté.

L'approche est principalement basée sur les points de référence du nez. Nous récupérons également les données gyroscopiques du téléphone pour estimer l'orientation de l'appareil pendant le mouvement. Il est ainsi possible de corréler les valeurs biométriques et les valeurs gyroscopiques pour vérifier l'authenticité du visage en temps réel. En effet, lorsqu'un utilisateur présente une photo fixe dans ce scénario, une grande différence existe entre l'analyse des points biométriques du nez et du gyroscope. Il est donc possible de détecter 99% des tentatives d'attaques avec ce type de support. De même, il est possible de détecter certains types de masques.

Cependant, la précision des capteurs gyroscopiques des téléphones portables ne permet pas d'estimer l'orientation du téléphone suffisamment pour être robuste aux attaques par présentation de masque 3D réaliste et de vidéo. Cependant, il existe des méthodes efficaces

basées sur les capteurs infrarouges ou l'analyse de texture pour surmonter les attaques de présentation par masque 3D. Dans le cas d'une attaque par vidéo, l'attaquant peut présenter une vidéo préparée à l'avance et simuler le mouvement du téléphone afin de déjouer le système. Dans le prochain chapitre, nous présenterons une méthode de détection des attaques par présentation vidéo qui est complémentaire à la méthode que nous venons de décrire.

4. Robustesse aux recaptures

4.1 Contexte

Parmi les attaques de présentation contre un système e-KYC, les attaques par relecture vidéo sont devenues les plus fréquentes. Cette situation est liée à différents facteurs.

Le premier facteur est que de plus en plus de systèmes e-KYC exigent une vidéo comme preuve d'authentification. Pendant la capture de la vidéo, l'utilisateur est invité à effectuer plusieurs actions pour vérifier son authenticité. La présentation d'une simple image imprimée n'est donc plus une attaque réalisable. L'attaquant devra nécessairement utiliser un écran pour diffuser une vidéo et la présenter au système pour tenter de le déjouer.

Le deuxième facteur est que les performances et la qualité des smartphones ont considérablement évolué au fil des ans. C'est notamment le cas des technologies d'écran et de caméra qui sont utilisées dans les smartphones. Les caméras des smartphones sont très performantes et parviennent à capturer des scènes du monde réel avec beaucoup de détails. Aujourd'hui, les nouveaux smartphones possèdent même plusieurs caméras. Par exemple, le dernier smartphone, présenté en septembre 2022 par Apple, embarque trois capteurs de photos (dans sa version haut de gamme). En outre, les écrans des smartphones ont une qualité d'affichage de plus en plus détaillée. Il est habituel qu'un smartphone ait une résolution comprise entre 2560 x 1440 pixels et 3840 x 2160 pixels. Aujourd'hui, les deux technologies les plus souvent utilisées sur les écrans de smartphones sont le LCD (*Liquid Crystal Display*) et l'OLED (*Organic Light-Emitting Diode*). L'OLED est la technologie la plus récente et offre un meilleur niveau de luminosité, un meilleur niveau de contraste et de meilleurs angles de vision que les LCD [1].

Le troisième facteur concerne les nouvelles technologies de falsification des vidéos. Des outils puissants et faciles à utiliser sont disponibles pour manipuler numériquement un visage dans une vidéo. Ces algorithmes peuvent être utilisés pour échanger le visage d'une personne avec celui d'une autre ou pour manipuler les expressions et les mouvements du visage d'une personne en temps réel.

C'est grâce à ces facteurs qu'un imposteur aura le plus de chances d'utiliser un écran pour attaquer un système de reconnaissance faciale. Il suffit à l'attaquant d'utiliser un deuxième

smartphone pour la modification du visage, d'afficher le résultat sur son écran haute résolution et de le présenter à la caméra du premier smartphone où s'exécute l'application KYC à distance (Figure 40).

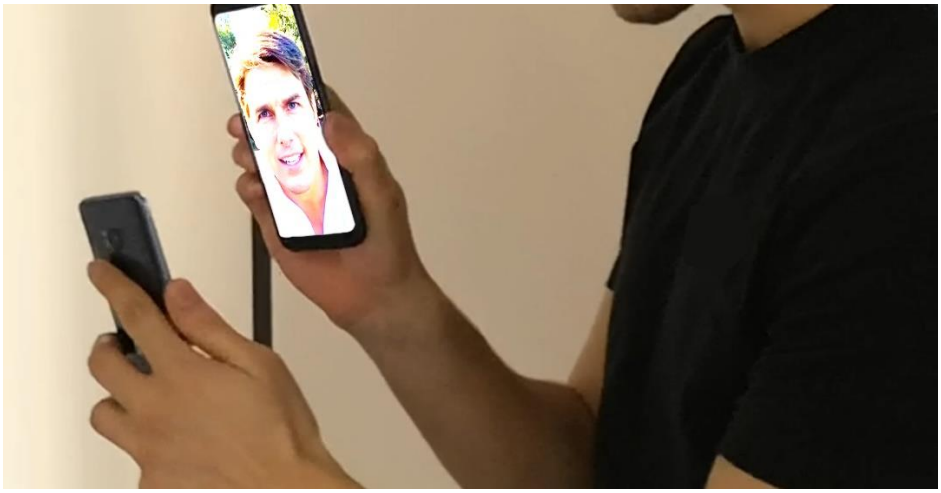

Figure 40 : Attaque de présentation avec un écran. L'un des smartphones affiche une image d’un visage pour attaquer une application KYC à distance exécutée sur le deuxième smartphone.

Les smartphones sont très répandus et sont plus faciles à manipuler que d'autres types d'écrans tels que les écrans d'ordinateur ou de télévision. Il est donc nécessaire de prendre en compte ce type d'attaque et de proposer des méthodes pour détecter la présentation des vidéos rejouées sur les écrans de smartphones. En examinant le flux de la caméra, nous pouvons analyser en détail la nature de l'image pour détecter ce type d'attaque. Dans ce chapitre, nous nous concentrons sur les attaques de présentation qui utilisent un écran affichant une image/vidéo.

4.2 État de l’art

4.2.1 Description

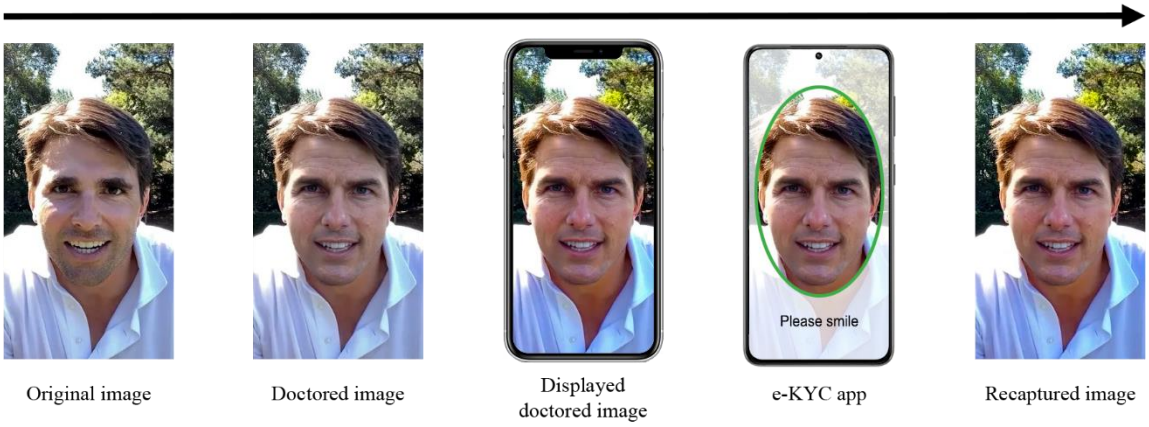

Figure 41 : Le processus habituel suivi par les imposteurs pour attaquer un système de reconnaissance faciale en présentant une image falsifiée affichée sur un écran.

Source : Image originale et falsifiée @vfxchrisume

Lorsque la caméra du système d’e-KYC reçoit une vidéo, elle capture en réalité une série d’images. Lorsque ces images sont affichées sur un écran, il s’agit d’images recapturées.

Une image recapturée est une image qui a été capturée par un premier capteur photo, puis affichée sur une surface (impression papier ou écran numérique) et enfin capturée par un autre capteur photo. Il est très difficile de distinguer visuellement l'image recapturée de l'image naturelle pour l'œil humain [2].

Par conséquent, de nombreux travaux ont été réalisés dans la littérature pour détecter automatiquement les images recapturées, en particulier les images affichées depuis un écran. Cependant, tous ces travaux se sont concentrés sur la détection des images recapturées à partir d'un écran LCD. Selon une étude d'IHS Market, les fabricants de smartphones équiperont plus de 50 % des écrans de leurs appareils avec la technologie OLED d'ici 2023. Les images recapturées à partir d'un écran LCD peuvent laisser des traces (aliasing, flou, bruit). Ces traces sont beaucoup moins présentes sur les images recapturées à partir d'un écran OLED (Figure 42).

Figure 42 : Comparaison entre une image recapturée à partir d'un écran LCD (à gauche) et une image recapturée à partir d'un écran OLED (à droite). On distingue des motifs au niveau du ciel sur l'image recapturée à partir d'un écran LCD.

De nombreuses méthodes ont vu le jour pour détecter les images recapturées en se basant sur les traces laissées pendant le processus de recapture. Comme dans de nombreux autres domaines de la criminalistique des images numériques, de nouvelles approches basées sur l'apprentissage profond sont également apparues ces dernières années pour détecter les images recapturées des écrans.

4.2.2 Méthodes traditionnelles de détection des images recapturées

Les techniques traditionnelles de détection des images recapturées des écrans LCD reposent sur une multitude de méthodes de traitement du signal. Les images recapturées des écrans LCD présentent souvent des motifs de moiré. Un motif de moiré est un exemple d'aliasing dû au chevauchement des grilles numériques du capteur de l'appareil qui entraîne un bruit à haute fréquence dans l'image. Muammar et al. ont proposé dans [3] une étude sur l'aliasing et les motifs de moiré à partir d'images recapturées d'écrans LCD. Ils ont proposé une méthode qui permet de détecter les images recapturées en analysant la présence de motifs de moiré.

Une idée similaire a été proposée par Mahdian et al. [4]. Les auteurs détectent l'aliasing en se basant sur la présence de motifs cyclostationnaires. Un motif cyclostationnaire est un motif qui a une propriété de périodicité. Ce type de motif est également présent dans les images recapturées d'un écran LCD. Après une étape de prétraitement pour améliorer l'image et obtenir des caractéristiques plus fortes, l'image est convertie dans le domaine spectral. Une image est

considérée comme recapturée si une forte corrélation est trouvée entre la version améliorée et sa version dans le domaine spectral.

Les auteurs de [5] ont proposé de détecter les images recapturées à l'aide de plusieurs caractéristiques. Les caractéristiques qu'ils ont utilisées comprennent la caractéristique de bruit du motif du capteur, la caractéristique de texture et les informations de couleur. Afin de classer les images recapturées, ils ont entraîné un algorithme de classification binaire en utilisant une machine à vecteurs de support (SVM).

4.2.3 Méthodes de détection des images recapturées basées sur l'apprentissage profond

Dans [6], les auteurs proposent un réseau neuronal convolutif laplacien (L-CNN) pour détecter les images recapturées. La partie d'apprentissage des caractéristiques (premières couches) a intégré un filtre laplacien afin d'améliorer le rapport signal/bruit introduit par le processus recapturé. Ce qui est un défi dans cet article, c'est que les images recapturées sont petites. Les images ont une taille de 64x64 pixels.

Plusieurs autres articles pertinents décrivent des méthodes d'apprentissage profond pour détecter les images recapturées des écrans LCD, comme dans [7] où les auteurs ont proposé de combiner des réseaux neuronaux convolutionnels (CNN) avec un réseau neuronal récurrent (RNN). Ils utilisent une étape de prétraitement composée d'une opération convolutive. Après la formation, les caractéristiques extraites du modèle CNN entraîné ont été introduites dans un réseau neuronal récurrent pour classer les images.

Dans [8], Abraham a proposé de détecter les motifs de moiré à l'aide d'un réseau neuronal convolutif. Après avoir utilisé la décomposition en ondelettes de Haar sur l'image d'entrée, il transmet les fréquences résultantes comme entrée pour le réseau. Il utilise ensuite l'une des fréquences résultantes comme paramètre de poids pour les autres fréquences pendant l'entraînement pour estimer la propagation des motifs moirés dans l'image.

4.2.4 Bases de données existantes

En ce qui concerne les bases de données existantes, il existe quatre grandes bases de données d'images recapturées d'un écran. Dans [9], les auteurs présentent une base de données d'images recapturées acquises uniquement par des caméras de smartphones. Les images

recapturées comprennent des photos de scènes du monde réel ainsi que des images recapturées correspondantes. Du papier imprimé et un écran d'ordinateur ont été utilisés pour le processus de recapture. Les conditions de prise de vue non contrôlées lors du processus de recapture font apparaître différents artefacts, tels que les artefacts de dégradation, des artefacts de texture et des artefacts de réflectance. Agarwal et al. ont proposé dans [10], une grande base de données d'images recapturées composée de 14500 échantillons. Quatre types de supports de recaptures sont présents : photographie d'une image imprimée, scan d'une image imprimée, photographie d'une image affichée sur un écran et capture d'écran de l'image. Une grande diversité d'appareils ont été utilisés pour les recaptures, incluant notamment 234 écrans. La base de données NTU-ROSE [2], contient 2 700 images recapturées à partir de 3 écrans LCD. Elles ont été enregistrées à l'aide de trois appareils photo dans différentes conditions d'éclairage. Enfin, ICL [11] est une base de données d'images recapturées obtenu à partir d'un seul écran LCD et neuf caméras différentes. Au total 1035 images originales ont été utilisées pour obtenir 2520 recaptures. Le tableau 3 regroupe les quatre *datasets* que nous venons de décrire.

Tableau 3: résumé des *datasets* publics dédiés à la détection des images recapturées.

Dataset	Année	Réel	Faux	Type d'écran	Nombre d'écrans	Nombre de caméras
NTU-Rose	2010	300	2700	LCD	3	3
ASTAR	2010	2240	1765	LCD	3	3
ICL	2015	1035	2520	LCD	1	8
LS-D	2018	14500	14500	LCD	234	180

Ces bases de données sont très utiles pour le développement de méthodes de détection des images recapturées. Cependant, nous remarquons que parmi toutes ces bases de données, il n'y a que des écrans LCD. Aucune des bases de données ne comprend d'écrans OLED. Comme nous l'avons dit, il est très important de prendre également en compte la technologie OLED. La plupart des méthodes de référence sont basées sur l'analyse des traces visuelles présentes dans les images recapturées d'un écran LCD. Comme ces artefacts ne sont pas souvent visibles sur les images recapturées à partir d'écrans OLED, ces méthodes peuvent ne plus fonctionner sur ce type d'écran.

4.3 Contributions

4.3.1 Base de données proposée

Afin d'évaluer la performance de notre détecteur, il est nécessaire de construire une base de données d'images recapturées d'écrans OLED. Comme mentionné précédemment, les bases de données existantes ne comprennent que des images recapturées d'écrans LCD. Cette base de données sera utilisée pour entraîner et tester différents réseaux neuronaux. Il est important que la base de données d'images recapturées ne contienne pas de biais d'entraînement. Pendant le processus de recapture, nous avons donc fait varier différents paramètres pour éviter tout biais d'entraînement. Ces paramètres sont les suivants :

- La distance entre la caméra et l'écran ;
- L'angle de vue entre la caméra et l'écran ;
- Le contenu des images affichées ;
- Les niveaux de luminosité de l'écran ;
- Le modèle de la caméra qui collecte les images.

L'un des paramètres les plus importants est le contenu de l'image. Les images recapturées doivent être suffisamment variées. Le choix de la base de données d'images naturelles (c'est-à-dire les images utilisées pour la recapture) est donc crucial. Nous avons décidé d'utiliser la base de données d'images MS-COCO [12]. La base de données MS-COCO est une grande base de données d'images d'objets du monde réel. C'est une base de données qui est adaptée à notre problématique car le contenu des images est très varié (Figure 43).

Figure 43 : Exemples d'images de notre base de données. Première ligne : Les images originales de MS-COCO, deuxième ligne : Les images recapturées correspondantes.

Les images originales de la base de données MS-COCO ont été affichées sur un écran OLED intégré à un ordinateur portable Gigabyte AERO de 15 pouces. Les images ont ensuite été capturées avec les caméras frontales de deux smartphones différents : un Samsung Galaxy S6 (5 mégapixels) et un Samsung Galaxy S8 (8 mégapixels). Pour supprimer les bords de l'écran des images recapturées, nous avons recadré toutes les images. Comme nous l'avons indiqué, nous avons fait varier différents paramètres durant le processus de recaptures (Tableau 44). La base de données est composée de 25 000 images (12 500 images originales et 12 500 images recapturées).

Figure 44 : De gauche à droite et de haut (en haut) : Image originale, fond noir, éclairage nocturne de l'écran, inclinaison horizontale de la caméra, inclinaison verticale de la caméra. De gauche à droite et de haut (en bas) : éclairage sombre, luminosité maximum, luminosité minimum, caméra éloignée de l'écran, caméra proche de l'écran.

4.3.2 Méthode de détection proposée

Pour pouvoir détecter automatiquement une image recapturée, nous avons décidé de développer une approche basée sur l'apprentissage profond en entraînant un réseau neuronal convolutif (CNN). Un CNN est un réseau neuronal bien adapté pour travailler avec des images, et cette architecture a montré d'excellents résultats dans les tâches de classification d'images. En théorie, les réseaux neuronaux sont capables de trouver une fonction pour classer deux classes si une différence significative existe. Pour notre problème, c'est approprié car cette distinction existe entre une image originale et une image recapturée.

Nous avons choisi d'entraîner deux architectures de référence sur notre base de données afin de sélectionner celle que nous utiliserons. Les deux architectures utilisées sont :

- ResNet50 [13] ;
- EfficientNet [14].

EfficientNet est l'architecture qui a donné les meilleurs résultats d'entraînement par rapport à Xception et ResNet50 (Tableau 4). Nous avons entraîné chaque architecture durant 50 *epochs* et avec un taux d'apprentissage de 10^{-3}.

Tableau 4: résultats du détecteur appliqué à différents jeux de test

Architecture	Taille des images	TPR	TNR	ACC
EfficientNet-B0	224 x 224	0.9709	0.8832	0.9266
ResNet50	224 x 224	0.8605	0.9703	0.9160

EfficientNet est un réseau neuronal convolutionnel (CNN) qui a été développé par Google. Dans le but de réduire le nombre de paramètres tout en maintenant un haut niveau de précision, les auteurs de EfficientNet ont étudié l'impact de trois paramètres sur la performance d'un CNN. Ces trois paramètres sont la profondeur du réseau profondeur du réseau, la largeur du réseau et la résolution de l'image d'entrée. Ils ont pu proposer une méthode qui améliore les performances d'un CNN en utilisant un coefficient de redimensionnement pour les trois paramètres. Ils ont également proposé huit variantes de leur solution allant de EfficientNet-B0 à EfficientNet-B7. EfficientNet-B0 est la variante la plus légère en termes de nombre de paramètres et aussi la moins précise. EfficientNet-B7, quant à lui, est la configuration avec le plus grand nombre de paramètres et la meilleure précision. Cependant, l'entraînement d'EfficientNet-B7 nécessite beaucoup de puissance de calcul. Avec EfficientNet-B0, nous obtenons une précision de 0,916. La sensibilité (TPR) du modèle est de 0,860, et la spécificité (TNR) du modèle est de 0,9703. L'architecture d'EfficientNet-B0 est décrite Figure 46.

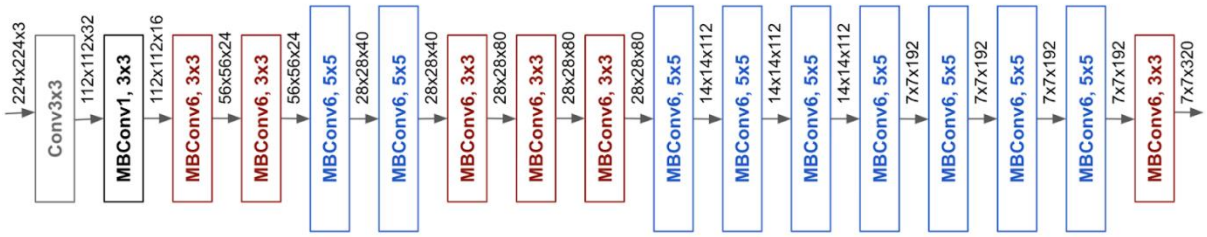

Figure 45 : Architecture d'EfficientNet-B0

4.4 Expériences et résultats

Pour pouvoir évaluer notre solution, nous avons construit plusieurs ensembles de test différents. Nous avons besoin d'évaluer si notre modèle peut détecter des contenus d'images recapturées qui ne sont pas présents dans la base de données. Nous avons donc construit un jeu de test d'images recapturées de visages et de documents (Figure 46). Une partie des images de visages provient de la base de données Face Recognition Grand Challenge (FRGC), et l'autre partie est d'une base de données que nous avons nous-mêmes acquis. Les images de documents proviennent également d'une base de données que nous avons construit.

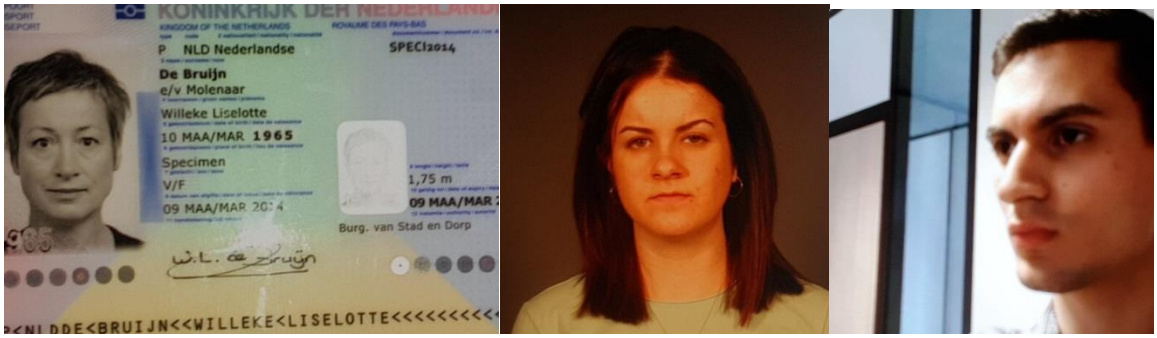

Figure 46 : Exemples d'images de notre base de test.

Il est également nécessaire de tester notre modèle sur des images recapturées sur des écrans qui ne sont pas présents dans la base de données. Pour ce faire, nous avons recapturé de nouvelles images sur deux autres écrans OLED provenant d'un Samsung Galaxy S6 et d'un Samsung Galaxy S8. Nous avons également testé notre détecteur sur des images recapturées d'un écran LCD Samsung S24D390H. Les résultats sont détaillés dans le tableau 5.

Tableau 5: résultats du détecteur appliqué à différents jeux de test

Jeu de test	Nombre d'images	Precision	Accuracy	F1-score
Visage + Document (OLED)	1000	0.996	0.948	0.950
MS-COCO (OLED)	1000	0.956	0.919	0.921
MS-COCO (LCD)	1000	0.604	0.743	0.701

Nous avons évalué notre méthode sur l'ensemble de test FRGC pour calculer le taux de fausse acceptation (FAR) et le taux de faux rejet (FRR). Le FAR est de 0,14 et le FRR de 0,018. Les résultats montrent que notre méthode réduit le taux de réussite des attaques de 86%. De plus, sans aucune image recapturée d'un écran LCD dans la base de données d'entraînement,

notre modèle est capable de détecter les images recapturées avec une précision de 74%. Notre technologie ne dépend donc pas du type d'écrans des images recapturées.

4.4.1 Comparaison avec une méthode de l'état de l'art

Dans cette section, nous comparerons nos résultats avec une méthode qui atteint une bonne précision sur les images recapturées des écrans LCD. Nous allons comparer notre méthode avec la méthode décrite dans [8]. Nous avons choisi cette méthode car elle utilise à la fois le motif de moiré comme caractéristiques et un réseau convolutif comme tâches de classification. Nous pouvons donc vérifier notre hypothèse en appliquant cette méthode aux images recapturées d'un écran OLED. Pour cette expérience, nous avons implémenté la méthode fournie par l'auteur. Lorsque la méthode est appliquée sur une base de données composée d'images naturelles et d'images recapturées à partir d'un écran LCD, la précision est d'environ 95 %. Mais, lorsque nous utilisons notre base de données avec des images recapturées d'un écran OLED, la méthode n'est pas capable de détecter plus d'un tiers des images recapturées (Tableau 6).

Tableau 6: résultats du détecteur [8]

Jeu de test	Nombre d'images	Precision	Accuracy	F1-score
Visage + Document (OLED)	327	0.9240	0.8590	0.9510
MS-COCO (OLED)	500	0.9490	0.7050	0.7070

4.5 Conclusion

Le développement des technologies de falsification des vidéos en temps réel génère un risque important pour tous les processus de KYC à distance. L'une des clés pour éviter leur diffusion trop rapide est de détecter les recaptures à partir des écrans et notamment des écrans de smartphones. Cependant, les algorithmes précédemment conçus étaient en grande partie basés sur des phénomènes d'interférence tels que le moiré, qui sont plus faciles à détecter à sur des écrans LCD. Les technologies des écrans ont également évolué, avec la popularisation des écrans OLED qui affichent des images avec très peu de dégradations.

Nous avons conçu un détecteur basé sur l'apprentissage profond capable de prendre en compte ces nouveaux écrans. Nous les avons comparés aux algorithmes précédents sur les anciens et les nouveaux écrans et avons montré un net écart de performance sur la détection des

recapturés des écrans OLED. De plus, comme illustré dans ce chapitre, notre algorithme est aussi capable de s'adapter entre les générations d'écrans et nous pensons qu'il fonctionnera sur les futures technologies d'écran.

Le contenu vidéo prendra une place croissante dans les années à venir dans le partage de la confiance entre les peuples et entre les entreprises. La détection des recaptures, qui est un indice de falsification, sera une exigence majeure pour augmenter le niveau de confiance de notre future société de personnes connectées à distance.

5. Robustesse aux *deepfakes*

5.1 Contexte

Les techniques de falsifications des vidéos en temps réels évoquées dans le chapitre précédent concernent principalement les visages. La manipulation des visages dans une image numérique n'est pas un problème nouveau. Il existe de nombreuses méthodes efficaces pour détecter tous les types de falsification, comme le morphing et l'échange de visages. Cependant, de nouvelles méthodes de falsification sont apparues au cours des dernières années. Ces nouvelles méthodes peuvent être appliquées aux vidéos et la détection de ces falsifications est beaucoup plus difficile. La technologie la plus populaire est appelée *deepfakes*.

Un *deepfake* est une méthode qui permet d'échanger de manière rapide, automatique et réaliste un visage dans une vidéo. Depuis l'apparition de la première vidéo *deepfake* fin 2017, cette technologie est devenue de plus en plus réaliste et il est désormais très difficile pour un humain d'identifier une vidéo *deepfake* [1].

Par conséquent, les *deepfakes* peuvent être utilisés comme un outil pour diffuser des *fake news* en falsifiant des vidéos et en les partageant sur Internet. La figure 47 montre un exemple de *deepfake*. Le visage de la vidéo originale (à gauche) a été remplacé par un autre visage (à droite).

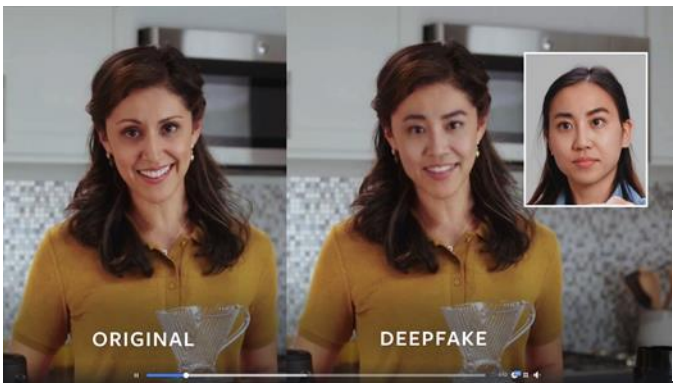

Figure 47 : Un exemple de vidéo *deepfake*.
Source : Facebook

Les *deepfakes* constituent également une menace sérieuse dans le domaine de la biométrie. En effet, il est aujourd'hui courant d'utiliser une caractéristique biométrique comme moyen d'authentification et le visage est la caractéristique biométrique la plus utilisée. De nombreux systèmes d'authentification par reconnaissance faciale exigent une vidéo du visage de l'utilisateur comme preuve biométrique. La technologie *deepfake* peut alors être utilisée comme moyen d'attaque afin de tromper un système. Outre le haut degré de réalisme d'une vidéo *deepfake*, le plus grand danger est qu'aucune compétence technique particulière n'est nécessaire pour produire ce type de falsification. Il n'est pas non plus nécessaire de maîtriser un logiciel compliqué. Aujourd'hui, n'importe qui peut fabriquer un *deepfake*. Pour toutes ces raisons, il est essentiel de combattre les *deepfakes* en développant des méthodes de détection puissantes qui évaluent automatiquement l'authenticité d'une vidéo et repèrent les *deepfakes*.

Entre fin 2019 et début 2020, un concours international appelé *DeepFake Detection Challenge* (DFDC) a été organisé. L'objectif de ce concours était d'obtenir des méthodes généralisables pour détecter les *deepfakes*. Des milliers de participants ont proposé leurs propres méthodes. Malgré des résultats prometteurs, aucune méthode n'a été capable de détecter tous les types de *deepfakes* en circulation.

Dans ce chapitre, nous proposons une analyse des solutions victorieuses du défi de détection des *deepfakes*. En particulier, nous étudions l'assemblage de ces solutions et la complémentarité entre elles. Nous testons différents ensembles avec diverses stratégies pour fusionner les scores et nous montrons qu'un assemblage bien choisi peut améliorer considérablement les résultats.

5.2 État de l’art

Le *deepfake* pourrait être considéré comme un sujet de recherche en soi. C'est un sujet qui est largement étudié dans la littérature. Chaque année, le nombre d'articles traitant des *deepfakes* augmente. Cela va de l'étude générale des *deepfakes* et des questions soulevées au développement de méthodes de détection des *deepfakes*, en passant par la publication de bases de données dédiées aux *deepfakes*. De nombreux articles proposent également de nouvelles méthodes pour réaliser des *deepfakes*.

5.2.1 Méthodes de création des *deepfake*

La plupart des méthodes permettant de générer un *deepfake* sont basées sur deux types de réseaux neuronaux : les auto encodeurs et les réseaux antagonistes génératifs (GAN).

Un *deepfake* basé sur les auto-encodeurs consiste à utiliser deux auto-encodeurs et à croiser les décodeurs. Un auto-encodeur est un type de réseau neuronal utilisé pour reconstruire une image à partir d'informations compressées (appelées espace latent) de la même image.

Un *deepfake* basé sur les GAN est composé de deux parties distinctes, un générateur G et un discriminateur D. Dans le cas de la génération de *deepfake*, le rôle du générateur est de synthétiser une vidéo capable de tromper le discriminateur et le rôle du discriminateur est de déterminer si le contenu proposé par le générateur est authentique ou non. De nombreuses variantes de la génération de *deepfake* basée sur les GAN ont été développées : FSGAN [2], StyleGAN [3], PGGAN [4].

Au début, il fallait beaucoup de ressources pour produire un *deepfake* réaliste. Aujourd'hui, ce n'est plus le cas. Le grand public peut créer des *deepfakes* avec un effort limité grâce à des applications faciles à utiliser. La plus populaire de ces applications est FaceApp [5], mais de plus en plus d'autres applications sont publiées chaque année (DeepFaceLab [6], ZAO [7], Faceswap web [8], etc.).

5.2.2 Méthodes de détection des *deepfake*

Compte tenu des nombreuses menaces liées aux *deepfakes*, de nombreuses méthodes de détection ont été proposées. Dans la littérature, il existe principalement trois catégories de méthodes de détection des *deepfakes* : basées sur l'analyse physiologique, basées sur l'analyse de la texture des images et basées sur la détection automatique avec l'intelligence artificielle. Dans le cadre de l'analyse physiologique, Li et al. [9] ont observé certaines incohérences dans le clignement des yeux dans une vidéo *deepfake*. À l'aide d'un réseau convolutionnel récurrent à long terme (LCRN), ils ont réussi à détecter les vidéos *deepfake*.

Dans [10], les auteurs déterminent si une vidéo est un *deepfake* en analysant les incohérences dans la position de la tête. Pour les méthodes de détection basées sur l'analyse d'images ou de textures, les auteurs recherchent principalement des incohérences dans le flux optique [11] ou la présence d'artefacts [12]. Enfin, les approches purement basées sur une

détection utilisant l'intelligence artificielle font passer les images d'une vidéo par des réseaux neuronaux. Les réseaux neuronaux peuvent être des réseaux neuronaux récurrents [13], des réseaux convolutionnels 3D [14] ou un ensemble de ces réseaux.

Malheureusement, et en raison de la diversité importante des différentes façons de générer un *deepfake*, il est très difficile de développer une méthode adaptée pour détecter toutes les vidéos *deepfake*. Il est également important de considérer que les modèles doivent être robustes aux attaques antagonistes. En effet, dans [15], il a été montré qu'il est possible de tromper facilement un détecteur en injectant un bruit antagoniste dans une de leurs vidéos. Pour faire face à tous ces problèmes, des bases de données de *deepfakes* de plus en plus diverses et riches sont mises à disposition.

5.2.3 Bases de données existantes

À notre connaissance, nous dénombrons sept grandes banques de données de *deepfakes* (Tableau 7). Nous pouvons évaluer la "qualité" d'une base de données en fonction du nombre de vidéos *deepfakes*, du nombre de vidéos originales, du nombre d'identités distinctes, du nombre de méthodes utilisées pour créer un *deepfake* et du nombre d'augmentations appliquées.

La base de données qui correspond le mieux à ces critères est la base de données créée par Facebook pour le *DeepFake Detection Challenge*.

Tableau 7 : liste des différentes bases de données existantes de *deepfakes*

Dataset	Années	# Vidéos falsifiées	# Vidéos authentiques	# Identités	# Méthodes	# Augmentation
UADFV [23]	2018	49	49	49	1	-
DeepfakeTIMIT [12]	2018	640	320	43	2	-
FaceForensics++ [21]	2019	4000	1000	?	4	2
Google DFD [18]	2019	3000	363	28	5	-
Celeb-DFD [16]	2019	5639	890	59	1	-
DeeperForensics [9]	2020	1000	59000	100	1	7
DFDC [6]	2020	104500	23654	960	5	19

5.3 Deepfake Detection Challenge

Le *Deepfake Detection Challenge* (DFDC) [16] est une compétition internationale, lancée en décembre 2019 par Facebook, en collaboration avec Microsoft, Amazon Web Services (AWS) et quelques partenaires universitaires également. Une base de données de plus de 100 000 vidéos authentiques et de *deepfakes* a été mise à la disposition des participants.

Facebook a utilisé différentes techniques pour modifier le visage des acteurs présentés dans les vidéos. Les résultats finaux ont été publiés le 12 juin 2020. Au total, 2114 équipes du monde entier ont participé.

La meilleure solution a été proposée par Selim Seferbekov. Il a extrait 32 images dans une vidéo, a détecté le visage et l'a recadré, puis a introduit les visages présents dans ces images dans une architecture composée d'un ensemble de sept EfficientNet-B7. Pendant l'étape d'apprentissage, il utilise une stratégie d'augmentation en supprimant les parties sémantiques du visage.

L'équipe qui a terminé en deuxième position a proposé un ensemble de deux XCeption et un EfficientNet-B3. Ils ont également utilisé une stratégie spéciale d'augmentation des données appelée WS-DAN [17].

Le troisième meilleur modèle a proposé un ensemble de trois EfficientNet-B7. Pendant l'étape d'entraînement, il utilise la stratégie d'augmentation de données mixup. La quatrième équipe a proposé un grand ensemble de différents CNN (EfficientNet-B0, EfficientNetB1, EfficientNet-B3, ResNet-34, Xception et SlowFast).

Enfin, la cinquième solution propose une architecture composée d'un ensemble de CNN 2D et 3D. Ils appliquent la stratégie d'augmentation des données cutmix pendant l'étape d'entraînement.

Toutes les méthodes que nous venons de décrire utilisent une stratégie de fusion simple en appliquant des poids sur les prédictions de chacun des modèles qu'ils utilisent dans leur ensemble. Pour évaluer les soumissions de chaque participant pendant ce concours, les organisateurs ont utilisé la fonction de perte logarithmique (5).

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\widehat{y_i}) + (1 - y_i)\log(1 - \widehat{y_i})] \quad (5)$$

Où n est le nombre de vidéos à prédire, $\widehat{y_i}$ représente la probabilité que la vidéo soit fausse, et y_i est l'étiquette de la vidéo, 0 pour une vraie vidéo et 1 pour une fausse vidéo. Cette métrique est utilisée pour évaluer les prédictions renvoyées par les modèles soumis. Une prédiction erronée avec une confiance élevée sera fortement pénalisée.

Selon Facebook [16], la perte logarithmique du meilleur modèle est de 0,4279 (ce qui correspond à une précision de 82%) sur l'ensemble de test mis à disposition par Facebook. Malheureusement, lorsqu'on évalue le même modèle avec des vidéos récupérées sur Internet, la précision a chuté de plus de 15 %, et le même modèle a atteint une précision de 65 %.

5.4 Expériences et résultats

Dans cette section, nous allons réaliser plusieurs expériences. Tout d'abord, nous allons déterminer si les cinq meilleures solutions peuvent se compléter en déterminant le nombre de faux positifs et de faux négatifs qu'elles ont en commun. Pour y parvenir, nous avons réimplémenté les 5 meilleures solutions du challenge. Ces cinq méthodes ont été rendues open source et les modèles entraînés ont été partagés à la fin du concours. Nous avons réexécuté chaque solution sur le jeu de test public que Facebook a mis en ligne et qui est composé de 5000 vidéos (Tableau 8).

Tableau 8 : score des solutions victorieuses sur le jeu de test publique de Facebook

Résultats [6]	Rang final [6]	Réimplémentation (Log-Loss)	Réimplémentation (Précision %)	Nouveau rang
0.1983	1ère	0.1957	92.68	4ème
0.1787	2ème	0.1790	93.36	1ère
0.1703	3ème	0.1821	93.90	2ème
0.1882	4ème	0.1863	92.58	3ème
0.2157	5ème	0.2158	90.86	5ème

Ce jeu de test ne représente que 50 % du jeu de test total qui a été utilisé pour déterminer le classement final pendant le DFDC. En réalité, le jeu de test complet se compose de 10 000 vidéos. La moitié des vidéos sont des *deepfakes* réalisés par Facebook et des séquences originales, l'autre moitié est composée de *deepfakes* et de vidéos originales récupérées sur Internet. Les vidéos récupérées sur Internet n'ont pas été rendues publiques par les organisateurs. Compte tenu de ce point, nous nous concentrons uniquement sur les 5 000 vidéos publiques fabriquées par Facebook. C'est pourquoi les scores présentés dans le tableau 3 ne correspondent pas exactement au classement final du DFDC. Dans le reste du document, nous faisons référence au nouveau classement que nous obtenons après la réimplémentation.

5.4.1 Faux positifs et faux négatifs en commun

Les différentes solutions ont un taux d'erreur compris entre 7 et 10 % (tableau 3), ce qui correspond aux faux positifs et aux faux négatifs. Afin de déterminer le niveau de similarité entre chaque solution, nous calculons le pourcentage de faux positifs et de faux négatifs en commun entre deux solutions (Tableau 9).

Tableau 9: pourcentage de faux positifs et faux négatifs en commun entre les solutions

	1ère	2ème	3ème	4ème	5ème
1ère	100%	48%	41%	48%	26%
2ème		100%	53%	47%	22%
3ème			100%	49%	27%
4ème				100%	27%
5ème					100%

Nous pouvons observer qu'il y a beaucoup moins de faux positifs et de faux négatifs en commun entre la cinquième solution et les autres. À partir de cette observation, on peut supposer que ce ne sont pas les mêmes éléments qui déterminent la décision de la méthode n°5 par rapport aux autres méthodes. Cela peut s'expliquer par l'architecture particulière de la cinquième solution composée principalement de 3D-CNN contrairement aux autres méthodes qui ont principalement utilisé des 2D-CNN. Au vu de cette observation, nous avons testé tous les différents ensembles de solutions gagnantes pour déterminer si la cinquième méthode peut être complémentaire lorsqu'elle est utilisée avec l'une des autres méthodes et ainsi améliorer les résultats.

5.4.2 Stratégies pour fusionner les scores

Dans la littérature, la réalisation d'un assemblage est un concept qui peut améliorer considérablement les résultats si les modèles ne convergent pas vers les mêmes prédictions. L'inconvénient est qu'en formant un assemblage, nous ajoutons de la complexité et il devient plus difficile de comprendre les décisions.

L'assemblage utilisant une stratégie de vote est l'une des méthodes les plus simples. Il existe deux types de classificateurs de vote. Le vote majoritaire : il s'agit d'un vote sur la prédiction de la classe. Dans nos expériences, nous avons également géré le cas d'une égalité entre les modèles. En cas d'égalité, nous prenons en compte le score de prédiction. Vote par

poids : Nous définissons des poids en fonction de l'importance du rang de chaque modèle (0,3 pour le 1er, 0,25 pour le 2e, 0,2 pour le 3e, 0,15 pour le 4e et 0,1 pour le 5e).

Des méthodes plus sophistiquées sont basées sur l'apprentissage automatique :

Bagging : entrainement de plusieurs sous-modèles sur des portions aléatoires de données du jeu d'entrainement (par exemple, la forêt aléatoire).

Boosting : entraînement de plusieurs modèles les uns après les autres, chaque modèle corrige les erreurs de son prédécesseur (par ex. adaboost).

Stacking : entraîne un modèle pour prédire un score final à partir des prédictions de chacun des modèles (par exemple, le modèle d'ensemble de vote).

5.4.3 Résultats des expériences des différents assemblages

Nous avons réalisé plusieurs assemblages composés de deux, trois, quatre et de tous les modèles parmi les solutions victorieuses en utilisant les stratégies décrites dans la dernière sous-section. Nous avons commencé par utiliser une stratégie de vote car ce sont les méthodes les plus simples. Notre base de données est composée des prédictions de 5000 vidéos de chacun des modèles. Les résultats du meilleur assemblage pour ces expériences sont présentés dans le tableau 10 (Assemblage par n signifie un assemblage utilisant n modèles parmi les modèles).

Tableau 10: score des solutions victorieuses sur le jeu de test publique de Facebook

Type d'assemblage	Meilleur score (Log-loss)	Meilleur score (Précision %)	Stratégies de fusion	Modèles
Seul	0.1790	93.90	-	$1^{ère}$
Assemblage par 2	0.1376	95.24	Vote majoritaire	$1^{ère} + 5^{ème}$
Assemblage par 3	0.1605	95.82	Vote majoritaire	$1^{ère} + 2^{ème} + 5^{ème}$
Assemblage par 4	0.1583	95.64	Vote majoritaire	$1^{ère} + 2^{ème} + 4^{ème} + 5^{ème}$
Assemblage par 5	0.1597	95.42	Vote majoritaire	Tous

C'est toujours la stratégie de fusion par vote majoritaire qui donne les meilleurs résultats sur les différents types d'assemblages. En combinant la première et la cinquième solution, nous parvenons à améliorer la perte logarithmique de 23% et la précision de 1,34% par rapport à la meilleure solution individuelle. Cette combinaison est composée de la meilleure solution (1ère) et de la solution qui avait le moins de faux positifs et de faux négatifs en commun avec les autres (5ème), ce qui vérifie notre hypothèse préliminaire. L'ajout d'autres modèles à

l'assemblage dégrade la perte logarithmique par rapport au meilleur assemblage de deux modèles.

Nous avons ensuite évalué cinq algorithmes d'apprentissage automatique comme méthode d'assemblage sur toutes les combinaisons possibles entre les cinq modèles. Ces algorithmes sont : Régression logistique (LR), Machine à vecteur de support (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost) et Perceptron multicouche (MLP). Ces algorithmes sont mis en œuvre dans la bibliothèque Python open source populaire appelée "scikit-learn". Nous avons utilisé cette bibliothèque pour utiliser ces algorithmes avec les paramètres par défaut.

Nous entraînons chacun de ces algorithmes avec les prédictions des 5 meilleures solutions. Nous avons divisé notre base de données en 75% pour l'entraînement et 25% pour les tests. Les résultats de chaque algorithme sont présentés dans le tableau 11. À partir de ces résultats, nous pouvons faire plusieurs observations.

Tableau 11: résultats des meilleurs assemblages pour chacune des stratégies de fusion sur le jeu de test public de Facebook (perte logarithmique | précision %)

Assemblage	Meilleur score (LR)	Meilleur score (SVM)	Meilleur score (RF)	Meilleur score (AdaBoost)	Meilleur score (MLP)
Assemblage par 2	0.1231 | 95.36	0.1268 | 94.96	0.1888 | 94.48	0.4680 | 95.52	0.1221 | 95.36
Assemblage par 3	0.1063 | 95.52	0.1092 | 95.60	0.1160 | 95.60	0.4608 | 95.68	0.1076 | 95.52
Assemblage par 4	0.1056 | 96.16	0.1080 | 95.92	0.1096 | 96.16	0.4763 | 95.20	0.1065 | 96.00
Assemblage par 5	0.1049 | 95.76	0.1072 | 95.52	0.1233 | 95.92	0.5842 | 90.40	0.1056 | 95.84

Tout d'abord, les résultats sont nettement meilleurs lorsque l'on passe d'une solution individuelle à un assemblage de deux modèles. Une observation similaire peut être faite lorsqu'on passe d'un ensemble de deux méthodes à un ensemble de trois méthodes. Ensuite, l'amélioration est limitée.

Par ailleurs, toutes ces stratégies de fusion sont meilleures que la fusion par vote (sauf pour RF). Le meilleur assemblage est obtenu avec la stratégie MLP avec une log-loss de 0,1221 (ce qui correspond à une amélioration de 31,78% par rapport à la meilleure solution unique).

De plus, le fait de combiner tous les modèles peut améliorer les résultats de plus de 41% par rapport au meilleur modèle unique. Cependant, nous observons que l'assemblage de seulement trois modèles peut déjà améliorer la perte de log de 40% et que, par conséquent, l'ajout de modèles supplémentaires n'est pas pertinent étant donné le compromis entre la

complexité ajoutée et le gain de performance. En termes de précision, nous observons une amélioration pour chaque stratégie de fusion sur chacun des assemblages. Avec la stratégie de fusion LR, la précision est améliorée de 2,26% lors de l'assemblage des 1er + 3e + 4e + 5e modèles. Il n'est pas nécessaire de combiner tous les modèles, car la précision pour chacune des stratégies de fusion diminue par rapport à l'assemblage de quatre modèles.

Enfin, AdaBoost est la seule stratégie qui dégrade les performances pour chaque type d'assemblage. Avec cette stratégie, le meilleur assemblage est celui composé de la 1ère + 2ème + 5ème solution avec une perte logarithmique de seulement 0,4608.

5.4.4 Test sur une base de données inconnue

Nous avons également testé les différents assemblages proposés sur une base de données externe avec des vidéos *deepfakes* inconnues (vidéo générée par un algorithme qui n'a pas été utilisé dans la phase d'entraînement). Ce jeu de données fait partie de [17]. La distribution des vraies/fausses vidéos dans cette base de données est déséquilibrée, il y a beaucoup plus de *deepfakes* que de vidéos originales. Nous avons décidé de sélectionner au hasard 1000 vidéos falsifiées et 1000 vidéos originales pour pouvoir l'évaluer de manière plus équitable.

Comme indiqué dans [16], les résultats des modèles entraînés sur la base de données du défi DFDC diminuent sur des vidéos non observées. C'est le cas pour tous les modèles, sauf pour le 4ème modèle. Pour éviter d'inclure un biais potentiel dans l'assemblage, nous n'avons pas inclus ce modèle dans nos expériences. Dans le reste du chapitre, nous utilisons les nouveaux rangs du tableau 12 pour désigner les modèles utilisés.

Tableau 12 : score des solutions victorieuses sur la base de données externe

Rang final [6]	Log-Loss	Précision (%)	Nouveau rang
1ère	0.6802	75.24	4ème
2ème	0.4028	81.04	2ème
3ème	0.7035	70.60	5ème
4ème	0.1527	92.04	1ère
5ème	0.5335	76.56	3ème

Tous les résultats sont présentés dans le tableau 13. La combinaison qui améliore le plus la perte de log (de 21%) et la précision (de 3,44%) est celle qui combine les quatre modèles.

Tableau 13 : résultats des meilleurs assemblages pour chacune des stratégies de fusion sur la base de données externe (perte logarithmique | précision %)

Assemblage	Meilleur score (LR)	Meilleur score (SVM)	Meilleur score (RF)	Meilleur score (AdaBoost)	Meilleur score (MLP)
Assemblage par 2	0.3540 | 81.92	0.3540 | 82.56	0.4752 | 82.08	0.5865 | 83.36	0.5865 | 82.56
Assemblage par 3	0.3489 | 82.24	0.3526 | 82.56	0.3380 | 84.16	0.5729 | 83.36	0.3472 | 83.36
Assemblage par 4	0.3446 | 82.40	0.3502 | 82.40	0.3736 | 84.48	0.5763 | 84.00	0.3303 | 84.00

5.5 Conclusion

Dans ce chapitre, nous avons proposé une étude des solutions victorieuses du DFDC. Nous avons démontré qu'il est possible d'améliorer les résultats en faisant des assemblages appropriés. De cette façon, il est possible d'améliorer la perte de log de 41% et la précision de 2,26% sur le jeu de test public du DFDC. Sur un jeu de données externe, il est possible d'améliorer la perte de log de 21% et la précision de 3,44%.

Nous avons observé que les meilleures méthodes utilisent judicieusement des augmentations de données lors de la phase d'apprentissage. Cependant, cela ne suffit toujours pas à rendre ces méthodes généralisables. En effet, lorsqu'on utilise une autre base de données composée de vidéos *deepfake* non observées pendant l'entraînement, les résultats ne sont pas aussi bons.

L'interprétabilité et l'explicabilité sont des domaines de grand intérêt en IA [1] et pas seulement pour le problème de la détection des *deepfakes*. Nous pensons que ce sont les domaines sur lesquels nous devons à présent travailler.

Nous pensons que pour résoudre un problème tel que la généralisation des méthodes de détection des *deepfakes*, il est nécessaire de comprendre les prédictions des modèles et d'étudier la complémentarité des architectures utilisées.

6. Conclusion

6.1 Résumé

La transformation numérique des entreprises et des services gouvernementaux modifie le regard des utilisateurs face aux services d'identification et d'authentification. La pandémie du COVID-19 à accélérer ce processus, aujourd'hui la vérification des personnes se fait à distance par le biais de systèmes de reconnaissance faciale. De nombreux services transforment leur système d'authentification vers des solutions de vérification d'identité à distance (e-KYC). Ce processus s'effectue le plus souvent par le biais d'un smartphone.

Bien que l'e-KYC offre plusieurs avantages pour le fournisseur de services et l'utilisateur (fluidifie les flux d'authentification, simplification du processus, etc.), des menaces sont induites par la non-supervision de l'utilisateur durant l'authentification. Des personnes ont la liberté de concevoir des tactiques pour usurper l'identité d'autres personnes à des fins malveillantes. Ces dernières années, de nouvelles méthodes de falsifications numériques des vidéos ont été développées. Ces méthodes ont été implémentées dans des applications qui permettent à n'importe qui de manipuler une vidéo. Ces méthodes s'ajoutent à des techniques déjà existantes basées sur des attaques de présentations physiques. Dans cette thèse nous avons proposé des méthodes de détections des attaques physiques et numériques basées sur l'apprentissage profond.

Dans le chapitre 2, nous avons présenté le contexte dans lequel cette thèse s'inscrit. Le fil rouge de cette partie est l'identification et l'authentification des personnes au cours de l'histoire. Nous avons vu que les besoins d'authentification ont toujours existé en particulier pour des besoins de sécurité. Nous avons aussi vu pourquoi et comment les documents d'identité ont été introduits. Ces mêmes documents d'identités qui sont aujourd'hui demandés dans un système e-KYC.

Dans le chapitre 3, nous avons proposé une nouvelle méthode de détection des attaques de présentation physique à destination des smartphones. La méthode est basée sur l'analyse de deux vecteurs de données. Un premier vecteur représentant des métriques biométriques du visage qui ont été obtenues à l'aide de dlib. Un deuxième vecteur qui décrit l'angle de rotation du smartphone de l'utilisateur. La méthode proposée impose à l'utilisateur de déplacer son

smartphone autour de son visage. Les deux vecteurs sont calculés durant toute la durée du mouvement. Une anti-corrélation entre les deux vecteurs est observée dans le cas d'un visage authentique. Ce comportement n'est pas retrouvé dans le cas d'attaques de présentation. La mesure de la corrélation des deux vecteurs est calculée à partir du coefficient de corrélation de Pearson. Enfin, la classification automatique est réalisée avec un SVM entraîné sur les coefficients de corrélation de Pearson calculés d'une base de données que nous avons-nous-mêmes établis.

Dans le chapitre 4, nous avons étudié le cas particulier des attaques de présentation vidéo. La méthode proposée dans le chapitre 3 étant vulnérable à ce type d'attaque si elle est bien réalisée. Après avoir analysé l'état de l'art sur ce sujet, nous nous sommes aperçus que les méthodes de détection de ce type d'attaque ne considéraient seulement qu'un seul type d'écran, à savoir les écrans LCD. De nouveaux types d'écrans utilisent la technologie OLED. Cette technologie permet de reproduire plus fidèlement les images et les vidéos qui sont affichées. Ainsi, lorsqu'un attaquant va présenter une vidéo à un système e-KYC affichée sur un écran OLED, les méthodes de détection existantes se retrouvent impactées. Nous avons donc construit une nouvelle base de données constituées de plusieurs types d'écrans. En particulier des technologies LCD et OLED. Plusieurs caméras ont été utilisées et les conditions d'acquisition (luminosité ambiante, distance entre la caméra et l'écran, etc.) ont été variées. À partir de cette base de données, nous avons entraîné un CNN capable de détecter plusieurs types d'écran.

Dans le chapitre 5, nous avons analysé les nouvelles méthodes de falsifications des visages dans une vidéo. Ce genre de falsification est appelé un *deepfake*. Les *deepfakes* représentent la plus grande menace pour un système d'e-KYC. De nombreuses méthodes ont été proposées dans la littérature, cependant, à cause de la diversité des méthodes de création des *deepfakes* aucune d'entre elles ne s'est montrée généralisable. Nous avons donc proposé différents assemblages de détecteur de *deepfake* pour améliorer les capacités de généralisation des méthodes. Le meilleur assemblage améliore également les performances de détection.

6.2 Perspectives

Les perspectives des travaux de cette thèse reposent principalement sur l’habilité de généralisation des modèles entraînés par apprentissage profond. L’une des limites de ce type de modèle est qu’ils ne sont pas généralisables.

6.2.1 Zero Shot Learning

Le *Zero Shot Learning* est un type d'apprentissage automatique où le modèle est capable d'apprendre et de généraliser à partir de quelques exemples. Cela contraste avec l'apprentissage automatique traditionnel, qui nécessite une grande quantité de données pour bien apprendre et généraliser. Le *Zero Shot Learning* est possible car le modèle est capable d'apprendre à partir des relations entre différents concepts, plutôt que de simplement mémoriser des données d'entraînement. Cela permet au modèle d'être plus flexible et robuste, car il peut gérer de nouveaux concepts qui n'ont pas été vus pendant l'entraînement. Une façon de penser au *Zero Shot Learning* est qu'il permet au modèle « d'apprendre par analogie ». Par exemple, si le modèle a été entraîné sur des images de visages, il peut ensuite appliquer ces connaissances à de nouvelles images de visages qu'il n'a jamais vues auparavant. Ainsi, on peut imaginer une application pour les applications sur les falsifications des visages.

6.2.2 Explication et interprétation

Deux autres notions sont importantes pour arriver à généraliser un jour les méthodes de détection des falsifications : l'explication et l'interprétation.

« L'explication » est le processus qui consiste à fournir une justification ou un argumentaire pour quelque chose. Dans le contexte de l'apprentissage automatique, cela fait référence à la capacité de comprendre comment un réseau neuronal est arrivé à une décision ou une sortie particulière. C'est important car cela nous permet d'identifier les erreurs et d'améliorer la précision du système. Cela nous aide aussi à instaurer la confiance dans ces systèmes, car nous pouvons voir comment ils fonctionnent et pourquoi ils prennent certaines décisions. Il existe différentes approches de « l'explication », mais une méthode courante est appelée « analyse de sensibilité ». Il s'agit d'examiner à quel point la sortie est sensible aux changements des valeurs d'entrée. « L'explication » est une partie importante des réseaux neuronaux, et il existe de nombreuses façons différentes de s'y prendre. En comprenant comment ces systèmes fonctionnent, nous pouvons construire de meilleurs modèles, plus précis et plus fiables.

L'interprétation est le processus qui consiste à comprendre comment un réseau neuronal arrive à ses prédictions. C'est important car cela peut nous aider à comprendre comment le réseau fonctionne et à identifier tout biais potentiel. Il existe de nombreuses techniques différentes qui peuvent être utilisées pour interpréter les réseaux neuronaux, mais elles se

classent toutes dans deux catégories principales : les méthodes basées sur un modèle et les méthodes agnostiques. Les méthodes basées sur un modèle impliquent l'entraînement d'un deuxième réseau neuronal, plus petit, pour imiter le comportement du premier. Ce deuxième réseau est alors plus facile à interpréter car il est beaucoup plus simple que l'original. Les méthodes agnostiques de modèle ne nécessitent pas l'entraînement d'un deuxième réseau. Au lieu de cela, elles utilisent des techniques statistiques pour analyser la sortie du premier réseau. Cela les rend plus généralisables mais aussi plus complexes. La meilleure méthode dépend de l'application particulière. En général, les méthodes basées sur des modèles sont plus précises, mais les méthodes diagnostiques de modèles sont plus largement applicables.

Bibliographie

Bibliographie – Chapitre 2

[1] Davis NZ Guerre M Du Tilh A Carrière Jean-Claude Vigne D. "*Le Retour De Martin Guerre"*. Ed. par Robert Laffont, 1982.

[2] Bertillon A. Identification Anthropométrique Instructions Signalétiques Par Alphonse Bertillon. Nouvelle Édition. Melun: Impr. administrative; 1893.

[3] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology, 14(1) :4–20, 2004.

[4] Guo, Wei et al. "A Master Key backdoor for universal impersonation attack against DNN-based face verification." ArXiv abs/2105.00249 (2021): n. pag.

[5] Règlement (UE) 2019/1157 du Parlement européen et du Conseil du 20 juin 2019 relatif au renforcement de la sécurité des cartes d'identité des citoyens de l'Union et des documents de séjour délivrés aux citoyens de l'Union et aux membres de leur famille exerçant leur droit à la libre circulation

[6] N. K. Ratha, J. H. Connell and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," in IBM Systems Journal, vol. 40, no. 3, pp. 614-634, 2001, doi: 10.1147/sj.403.0614.

[7] Bryce E. Bayer. "Color imaging array". Pat. US3971065A. 1975.

[8] M. Ferrara, A. Franco and D. Maltoni, "The magic passport," IEEE International Joint Conference on Biometrics, 2014, pp. 1-7, doi: 10.1109/BTAS.2014.6996240.

[9] Marr, D. "A Theory for Cerebral Neocortex." Proceedings of the Royal Society of London. Series B, Biological Sciences, vol. 176, no. 1043, 1970, pp. 161–234. JSTOR, http://www.jstor.org/stable/76043.

[10] Warren McCulloch & Walter Pitts, A Logical Calculus of Ideas Immanent in Nervous Activity, 1943, Bulletin of Mathematical Biophysics 5:115-133.

[11] D.O. Hebb, The organization of behavior, New York, Wiley, 1949

[12] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). https://doi.org/10.1038/323533a

[13] LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. Advances in Neural Information Processing Systems 2 (NIPS*89).

[14] LeCun, Y.; Bottou, L.; Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition.Proceedings of the IEEE. 86(11): 2278 - 2324.[1]

[15] Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernetics 36, 193–202 (1980).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. https://doi.org/10.1145/3065386

[17] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[18] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90.

[19] Szegedy, Christian & Liu, Wei & Jia, Yangqing & Sermanet, Pierre & Reed, Scott & Anguelov, Dragomir & Erhan, Dumitru & Vanhoucke, Vincent & Rabinovich, Andrew. (2015). Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1-9. 10.1109/CVPR.2015.7298594.

[20] Chollet, Francois. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 1800-1807. 10.1109/CVPR.2017.195.

[21] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[22] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006 Jul 28;313(5786):504-7. doi: 10.1126/science.1127647. PMID: 16873662.

[23] Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

[24] Liu, M., Breuel, T.M., & Kautz, J. (2017). Unsupervised Image-to-Image Translation Networks. ArXiv, abs/1703.00848.

Bibliographie – Chapitre 3

[1] Li, Yan & Xu, Ke & Yan, Qiang & Li, Yingjiu & Deng, Robert. (2014). Understanding OSN-based facial disclosure against face authentication systems. 10.1145/2590296.2590315.

[2] Abate, Andrea & Nappi, Michele & Riccio, Daniel & Sabatino, Gabriele. (2007). 2D and 3D Face Recognition: A Survey. Pattern Recognition Letters. 28. 1885-1906. 10.1016/j.patrec.2006.12.018.

[3] Jenkins, R & Burton, A. (2008). 100% Accuracy in Automatic Face Recognition. Science (New York, N.Y.). 319. 435. 10.1126/science.1149656.

[4] Tan, Xiaoyang & Liu, yi & Liu, Jun & Jiang, Lin. (2010). Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model. 6316. 504-517. 10.1007/978-3-642-15567-3_37.

[5] Duc, Nguyen Minh. “Your face is NOT your password Face Authentication ByPassing Lenovo – Asus – Toshiba.” (2009).

[6] Kollreider, K. & Fronthaler, Hartwig & Bigun, Josef. (2005). Evaluating liveness by face images and the structure tensor. 75- 80. 10.1109/AUTOID.2005.20.

[7] Jia, Shan & Guo, Guodong & Xu, Zhengquan. (2019). A survey on 3D mask presentation attack detection and countermeasures. Pattern Recognition. 98. 107032. 10.1016/j.patcog.2019.107032.

[8] Köse, Neslihan & Dugelay, Jean-Luc. (2013). On the vulnerability of face recognition systems to spoofing mask attacks. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 2357-2361. 10.1109/ICASSP.2013.6638076.

[9] Kuratate, Takaaki & Pierce, Brennand & Cheng, Gordon. (2011). `Mask-bot"- a life-size talking head animated robot for AV speech and human-robot communication research. 111-116.

[10] Chingovska, I. & Anjos, A. & Marcel, Sébastien. (2012). On the effectiveness of local binary patterns in face anti-spoofing. BIOSIG. 1-7.

[11] Zhang, Zhiwei & Yan, Junjie & Liu, Sifei & Lei, Zhen & Yi, Dong & Li, Stan. (2012). A face antispoofing database with diverse attacks. Proceedings - 2012 5th IAPR International Conference on Biometrics, ICB 2012. 26-31. 10.1109/ICB.2012.6199754.

[12] https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

[13] Pavlidis, Ioannis & Symosek, Peter. (2000). The imaging issue in an automatic face/disguise detection system. 15-24. 10.1109/CVBVS.2000.855246.

[14] Sun, Lin & Huang, WaiBin & Wu, Minghui. (2011). TIR/VIS Correlation for Liveness Detection in Face Recognition. 6855. 114-121. 10.1007/978-3-642-23678-5_12.

[15] Bowyer, Kevin & Chang, Jin (Kyong) & Flynn, Patrick. (2006). A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. Computer Vision and Image Understanding. 101. 1-15. 10.1016/j.cviu.2005.05.005.

[16] Galbally, Javier & Marcel, Sébastien & Fierrez, Julian. (2014). Biometric Antispoofing Methods: A Survey in Face Recognition. IEEE Access. 2. 1530-1552. 10.1109/ACCESS.2014.2381273.

[17] Li, Jiangwei & Tan, Tieniu & Jain, Anil. (2004). Live Face Detection Based on the Analysis of Fourier Spectra. Proceedings of SPIE - The International Society for Optical Engineering. 5404. 296-303. 10.1117/12.541955.

[18] Tan, Xiaoyang & Liu, yi & Liu, Jun & Jiang, Lin. (2010). Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model. 6316. 504-517. 10.1007/978-3-642-15567-3_37.

[19] Chingovska, I. & Anjos, A. & Marcel, Sébastien. (2012). On the effectiveness of local binary patterns in face anti-spoofing. BIOSIG. 1-7.

[20] Bai, Jiamin & Ng, Tian & Gao, Xinting & Shi, Y.Q.. (2010). Is physics-based liveness detection truly possible with a single image?. ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems. 3425-3428. 10.1109/ISCAS.2010.5537866.

[21] Komulainen, Jukka & Hadid, Abdenour & Pietikainen, Matti. (2012). Face spoofing detection from single images using texture and local shape analysis. Biometrics, IET. 1. 3-10. 10.1049/iet-bmt.2011.0009.

[22] Kollreider, K. & Fronthaler, Hartwig & Bigun, Josef. (2008). Verifying liveness by multiple experts in face biometrics. Biometrics. 1 - 6. 10.1109/CVPRW.2008.4563115.

[23] Kim, Gahyun & Eum, Sungmin & Suhr, Jae & Kim, Dong & Park, Kang & Kim, Jaihie. (2012). Face liveness detection based on texture and frequency analyses. Proceedings - 2012 5th IAPR International Conference on Biometrics, ICB 2012. 67-72. 10.1109/ICB.2012.6199760.

[24] Pan, Gang & Sun, Lin & Wu, Z. & Lao, Shihong. (2007). Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera. ICCV. 1-8. 10.1109/ICCV.2007.4409068.

[25] Kollreider, K. & Fronthaler, Hartwig & Bigun, Josef. (2008). Verifying liveness by multiple experts in face biometrics. Biometrics. 1 - 6. 10.1109/CVPRW.2008.4563115.

[26] Bigun, Josef & Fronthaler, Hartwig & Kollreider, K.. (2004). Assuring liveness in biometric identity authentication by real-time face tracking. 104 - 111. 10.1109/CIHSPS.2004.1360218.

[27] Ali, Asad & Deravi, Farzin & Hoque, Sanaul. (2012). Liveness Detection Using Gaze Collinearity. Proceedings - 3rd International Conference on Emerging Security Technologies, EST 2012. 10.1109/EST.2012.12.

[28] Bharadwaj, Samarth & Dhamecha, Tejas & Vatsa, Mayank & Singh, Richa. (2013). Computationally Efficient Face Spoofing Detection with Motion Magnification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 105-110. 10.1109/CVPRW.2013.23.

[29] Feng, Litong & Po, Lai & Li, Yuming & Xu, Xuyuan & Yuan, Fang & Cheung, Terence Chun-Ho & Cheung, Kwok-Wai. (2016). Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation. 38. 10.1016/j.jvcir.2016.03.019.

[30] Li, Lei & Feng, Xiaoyi & Boulkenafet, Zinelabidine & Xia, Zhaoqiang & Li, Mingming & Hadid, Abdenour. (2016). An original face anti-spoofing approach using partial convolutional neural network. 1-6. 10.1109/IPTA.2016.7821013.

[31] Shao, Rui & Lan, Xiangyuan & Yuen, P C. (2018). Joint Discriminative Learning of Deep Dynamic Textures for 3D Mask Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2018.2868230.

[32] Yang, Jianwei & Lei, Zhen & Li, Stan. (2014). Learn Convolutional Neural Network for Face Anti-Spoofing.

[33] Xu, Zhenqi et al. "Learning temporal features using LSTM-CNN architecture for face anti-spoofing." 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) (2015): 141-145.

[34] [34] Liu, Yaojie et al. "Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 389-398.

[35] Raghavendra, Ramachandra et al. "Presentation Attack Detection for Face Recognition Using Light Field Camera." IEEE Transactions on Image Processing 24 (2015): 1060-1075.

[36] [Zhang, Shifeng & Wang, Xiaobo & Liu, Ajian & Zhao, Chenxu & Wan, Jun & Escalera, Sergio & Shi, Hailin & Wang, Zezheng & Li, Stan. (2019). A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing. 10.1109/CVPR.2019.00101.

[37] George, Anjith et al. "Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network." IEEE Transactions on Information Forensics and Security 15 (2020): 42-55.

[38] Tan, Xiaoyang et al. "Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model." ECCV (2010).

[39] Peixoto, Bruno Malveira et al. "Face liveness detection under bad illumination conditions." 2011 18th IEEE International Conference on Image Processing (2011): 3557-3560.

[40] Zhang, Zhiwei et al. "A face antispoofing database with diverse attacks." 2012 5th IAPR International Conference on Biometrics (ICB) (2012): 26-31.

[41] Chingovska, Ivana et al. "On the effectiveness of local binary patterns in face anti-spoofing." 2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG) (2012): 1-7.

[42] Wen, Di et al. "Face Spoof Detection With Image Distortion Analysis." IEEE Transactions on Information Forensics and Security 10 (2015): 746-761.

[43] Liu, Siqi et al. “3D Mask Face Anti-spoofing with Remote Photoplethysmography.” ECCV (2016).

[44] Patel, Keyurkumar et al. “Secure Face Unlock: Spoof Detection on Smartphones.” IEEE Transactions on Information Forensics and Security 11 (2016): 2268-2283.

[45] Boulkenafet, Zinelabinde et al. “OULU-NPU: A Mobile Face Presentation Attack Database with Real-World Variations.” 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (2017): 612-618.

[46] Liu, Yaojie et al. “Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 389-398.

[47] Liu, Yaojie et al. “Deep Tree Learning for Zero-Shot Face Anti-Spoofing.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 4675-4684.

[48] King, Davis E.. “Dlib-ml: A Machine Learning Toolkit.” J. Mach. Learn. Res. 10 (2009): 1755-1758.

Bibliographie – Chapitre 4

[1] Z. Luo et S.-T. Wu, « OLED Versus LCD: Who Wins? », 2015.

[2] H. Cao et A. C. Kot, « Identification of recaptured photographs on LCD screens », in *International Conference on Acoustics, Speech and Signal Processing*, USA, 2010.

[3] H. Muammar et P. L. Dragotti, « An investigation into aliasing in images recaptured from an LCD monitor using a digital camera », in *International Conference on Acoustics, Speech and Signal Processing*, Canada, 2013.

[4] B. M. Adam Novoz, « Detecting Cyclostationarity in Re-Captured LCD Screens », *J. Forensic Res.*, vol. 06, n° 04, 2015.

[5] Y. Ke *et al.*, « Image Recapture Detection Using Multiple Features », *Int. J. Multimed. Ubiquitous Eng.*, vol. 8, n° 5, 2013.

[6] P. Yang *et al.*, « Recapture Image Forensics Based on Laplacian Convolutional Neural Networks », in *Digital Forensics and Watermarking*, vol. 10082, 2017.

[7] H. Li *et al.*, « Image Recapture Detection with Convolutional and Recurrent Neural Networks », *Electron. Imaging*, vol. 2017, n° 7, 2017.

[8] E. Abraham, « Moiré Pattern Detection using Wavelet Decomposition and Convolutional Neural Network », in *Symposium Series on Computational Intelligence (SSCI)*, India, 2018.

[9] X. Gao *et al.*, « A Smart Phone Image Database for Single Image Recapture Detection », in *Digital Watermarking*, vol. 6526, 2011.

[10] S. Agarwal *et al.*, « A Diverse Large-Scale Dataset for Evaluating Rebroadcast Attacks », in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, 2018.

[11] T. Thongkamwitoon *et al.*, « An Image Recapture Detection Algorithm Based on Learning Dictionaries of Edge Profiles », *IEEE Trans. Inf. Forensics Secur.*, vol. 10, n° 5, 2015.

[12] T.-Y. Lin *et al.*, « Microsoft COCO: Common Objects in Context », *ArXiv14050312 Cs*, 2015.

[13] He, Kaiming et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.

[14] Tan, Mingxing and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." ArXiv abs/1905.11946 (2019): n. pag.

Bibliographie – Chapitre 5

[1] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *ArXiv*, abs/1811.00656, 2019.

[2] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019.

[3] Karras, Tero et al. "A Style-Based Generator Architecture for Generative Adversarial Networks." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 4396-4405.

[4] Karras, Tero et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." ArXiv abs/1710.10196 (2018): n. pag.

[5] Faceapp. https://www.faceapp.com/. Dernier accès: 2022-09-01.

[6] Perov, Ivan et al. "DeepFaceLab: Integrated, flexible and extensible face-swapping framework." (2020).

[7] Zao. https://apps.apple.com/cn/app/zao. Dernier accès: 2022-09-01.

[8] Faceswap web. https://faceswapweb.com/. Dernier accès: 2022-09-01.

[9] Li, Yuezun et al. "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking." ArXiv abs/1806.02877 (2018): n. pag.

[10] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.

[11] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and A. Bimbo. Deepfake video detection through optical flow based cnn. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207, 2019.

[12] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *ArXiv*, abs/1811.00656, 2019.

[13] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[14] Yaohui Wang and Antitza Dantcheva. A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 515–519, 2020.

[15] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian CantonFerrer. Adversarial threats to deepfake detection: A practical perspective. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 923–932, 2021.

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *ArXiv*, abs/2006.07397, 2020.

[17] Hu, Tao and Honggang Qi. “See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification.” ArXiv abs/1901.09891 (2019): n. pag.

[18] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess,¨ Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

Robustesse aux Attaques en Authentification Digitale par Apprentissage Profond

Résumé :
L'identité des personnes sur Internet devient un problème de sécurité majeur. Depuis les accords de Bale, les institutions bancaires ont intégré la vérification de l'identité des personnes ou Know Your Customer (KYC) dans leur processus d'inscription. Avec la dématérialisation des banques, cette procédure est devenue l'e-KYC ou KYC à distance qui fonctionne à distance via le smartphone de l'utilisateur. De même, la vérification d'identité à distance est devenue la norme pour l'inscription aux outils de signature électronique. De nouvelles réglementations émergent pour sécuriser cette approche, par exemple, en France, le cadre PVID encadre l'acquisition à distance des documents d'identité et du visage des personnes dans le cadre du règlement eIDAS. Cela est nécessaire, car on assiste à l'émergence d'un nouveau type de criminalité numérique : l'usurpation d'identité profonde.

Grâce aux nouveaux outils d'apprentissage profond, les imposteurs peuvent modifier leur apparence pour ressembler à quelqu'un d'autre en temps réel. Les imposteurs peuvent alors accomplir toutes les actions courantes requises lors d'une inscription à distance sans être détectés par les algorithmes de vérification d'identité. Aujourd'hui, il existe des applications sur smartphone et des outils destinés à un public plus limité qui permettent aux imposteurs de transformer facilement leur apparence en temps réel. Il existe même des méthodes pour usurper une identité à partir d'une seule image du visage de la victime. L'objectif de cette thèse est d'étudier les vulnérabilités des systèmes d'authentification d'identité à distance face aux nouvelles attaques afin de proposer des solutions basées sur l'apprentissage profond pour rendre les systèmes plus robustes.

Mots clés : e-KYC, face anti-spoofing, vol d'identité, deepfakes.

Robustness to Digital Authentication Attacks with Deep Learning

Abstract:
The identity of people on the Internet is becoming a major security issue. Since the Bale agreements, banking institutions have integrated the verification of people's identity or Know Your Customer (KYC) in their registration process. With the dematerialization of banks, this procedure has become e-KYC or remote KYC which works remotely through the user's smartphone. Similarly, remote identity verification has become the standard for enrollment in electronic signature tools. New regulations are emerging to secure this approach, for example, in France, the PVID framework regulates the remote acquisition of identity documents and people's faces under the eIDAS regulation. This is required because a new type of digital crime is emerging: deep identity theft.

With new deep learning tools, imposters can change their appearance to look like someone else in real time. Imposters can then perform all the common actions required in a remote registration without being detected by identity verification algorithms. Today, smartphone applications and tools for a more limited audience exist allowing imposters to easily transform their appearance in real time. There are even methods to spoof an identity based on a single image of the victim's face. The objective of this thesis is to study the vulnerabilities of remote identity authentication systems against new attacks in order to propose solutions based on deep learning to make the systems more robust.

Keywords: e-KYC, face anti-spoofing, identity theft, deepfakes.