asse-part*o*ut bi*o*métriques

Tanguy Gernot

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

Passe-partout biométriques

Présentée et soutenue par
TANGUY GERNOT

Thèse soutenue le 28/11/2022 devant le jury composé de		
M. PATRICK BAS	Directeur de recherche, ECOLE CENTRALE LILLE	Rapporteur du jury
M. GILDAS AVOINE	Professeur des universités, INSA de Rennes	Membre du jury
MME SANDRA CREMER	Docteur, Thales Digital Identity and Security	Membre du jury
MME CAROLINE FONTAINE	Directeur de recherche au CNRS, École normale supérieure Paris-Saclay	Membre du jury
MME MARINE MINIER	Professeur des universités, Université de Lorraine	Président du jury
M. PATRICK LACHARME	Maître de conférences HDR, ENSICAEN	Directeur de thèse

Thèse dirigée par PATRICK LACHARME (Groupe de recherche en informatique, image, automatique et instrumentation)

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

MIIS

GREYC
Laboratoire de recherche en sciences du numérique

Remerciements

Durant cette thèse, j'ai pu m'initier à la recherche puis aboutir à mes premières contributions. Je tiens à remercier le laboratoire GREYC qui m'a accueilli, et plus précisément l'équipe SAFE, dont la gaieté et la démocratie m'inspirent.

Mes premiers remerciements vont évidemment à mon directeur de thèse, Patrick Lacharme, qui a su me guider tout en me laissant une large indépendance, ainsi que pour ses attentives relectures. Sa disponibilité et sa rigueur ont permis aux contributions de ce manuscrit d'aboutir.

Je remercie les membres de l'équipe, dont Jean-Marie Le Bars pour ses relectures et discussions sur la thèse, Christophe Rosenberger et Christophe Charrier pour leurs visions complémentaires de la biométrie. Je remercie particulièrement Emmanuel Giguet qui m'a accompagné durant tout mon parcours universitaire et m'a dirigé vers le monde de la recherche, ainsi que pour sa détermination à résoudre des challenges. J'espère continuer à bénéficier de nos échanges, jusqu'à ce que la retraite nous sépare.

Je remercie le ministère et l'université de Caen qui ont financé ces 3 années de thèse, ainsi que l'Ensicaen qui me permet de poursuivre grâce à un contrat d'ATER.

Je remercie les services administratifs, dont Sophie, Arielle et Gaëlle au GREYC, mais aussi Marie à l'école doctorale. Enfin, je remercie les membres de mon comité de suivi individuel de thèse, Gaétan Richard et Amine Nait-ali.

Je remercie tous les membres de mon jury, Marine Minier et Patrick Bas d'avoir accepté de rapporter minutieusement cette thèse, ainsi que Gildas Avoine, Caroline Fontaine, et Sandra Crémer d'avoir pris le temps de l'examiner.

Je remercie mes amis, notamment Tristan et Guillaume qui m'ont accompagné pendant une partie de mes études. Nos parties de jeux les ont enjolivées !

Enfin, je remercie toute ma famille, à commencer par toi, Papy, qui n'aura pas vu ce beau projet se conclure, mais dont la fierté persiste. Je remercie finalement ma femme, Mélina, dont le soutien permanent m'enjoie au quotidien. Merci à vous deux, mes filles, Lola et Romy. Vos sourires et éclats inspirent mes journées.

À vous trois.

Résumé

L'informatique est un domaine de plus en plus utilisé au quotidien et il est désormais incontournable dans une grande partie de nos activités. La sécurité informatique s'impose pour protéger ces activités et nos données privées.

La biométrie nous permet à tous de contribuer à cette sécurité en limitant les contraintes pour l'usager, mais les données biométriques sont à caractère personnel. Cette thèse s'inscrit dans la sécurité des données biométriques, qui passe notamment par des transformations paramétrées par des graines. Nous avons initialement attaqué la propriété de non-inversibilité de ces transformations en construisant des préimages proches et réutilisables à l'aide d'un algorithme génétique. Des variantes de ces préimages nous ont amenés à vouloir construire des passe-partout biométriques, permettant d'usurper une large partie des utilisateurs d'un système biométrique. Enfin, nous avons voulu orienter le choix des graines et donc des transformations pour permettre à un passe-partout fixé d'usurper l'intégralité des utilisateurs d'un système biométrique. Nous avons étendu ce concept à un individu passe-partout, pour lequel les futures acquisitions biométriques persistent à usurper les utilisateurs.

Nous avons validé ces concepts à l'aide de différentes données biométriques de visages, d'empreintes digitales et d'électrocardiogrammes. En outre, nous avons comparé nos algorithmes à d'autres existants.

Abstract

Computer science is a field that is used more and more in our daily lives and is now essential to a large part of our activities. Cybersecurity is necessary to protect these activities and our private data.

Biometrics allows us to contribute to this cybersecurity by limiting the constraints for the user, but biometric data is personal. This thesis deals with the security of biometric data, which is notably achieved through transformations parameterized by seeds. We initially attacked the non-invertibility property of these transformations by building long-lived nearby-template preimages using a genetic algorithm. Variations of these preimages led us to build biometric masterkeys, allowing spoofing a large part of the users of a biometric system. Finally, we wanted to choose the seeds to allow a fixed masterkey to spoof all the users of a biometric system. We have extended this concept to multiple masterkeys person, for which future biometric acquisitions persist in spoofing users.

We validated these concepts using different biometric data of faces, fingerprints and electrocardiograms. In addition, we compared our algorithms to other existing ones.

Liste des abréviations, sigles et symboles

$3D$ Trois dimensions

$\#hd$ Nombre de comparaisons en distance de Hamming

Ω Enfants

Φ Individu

Ψ Chromosome

ψ Allèle

Ξ Population

ξ Parents

ADN Acide désoxyribonucléique

$CNIL$ Commission Nationale de l'Informatique et des Libertés

ECG Électrocardiogramme

EEG Électroencéphalogramme

EER Taux d'erreur égale

EMG Électromyogramme

FMR Taux de fausses acceptations

$FNMR$ Taux de faux rejets

FVC Concours de vérification des empreintes digitales

GS Gram-Schmidt

hd_{max} Nombre maximum de comparaisons en distance de Hamming autorisé

it Nombre d'itérations

$JL1$ Première méthode de construction de matrice proposée par Achlioptas

$JL2$ Seconde méthode de construction de matrice proposée par Achlioptas

LFW Visages réels labélisés

LSH Hachage localement sensible

max Valeur maximale d'une série

min Valeur minimale d'une série

pop Taille de la population

PPR Préimage proche et réutilisable

PSO Optimisation par essaims particulaires

PTB Institut fédéral de physique et de technologie

$Q1$ Premier quartile

$Q2$ Deuxième quartile

$Q3$ Troisième quartile

TCO Taille de couverture optimale

TOD Taille optimale de dictionnaire

WAP Probabilité d'attaque par des loups

Table des matières

1 Introduction **1**
1.1 Contexte 1
1.2 La biométrie 2
1.3 Sécurité biométrique 2
1.4 Contributions 2
1.5 Perspectives 3

2 État de l'art et définitions **5**
2.1 La biométrie 5
2.1.1 Différentes modalités biométriques 6
2.1.2 Utilisation de la biométrie 8
2.1.3 Sensibilité des données biométriques 11
2.2 Base de données biométriques 13
2.2.1 Les captures de modalités biométriques 13
2.2.2 Extraction de caractéristiques 14
2.2.3 Comparaison de caractéristiques 15
2.3 Système biométrique 16
2.3.1 Schéma d'identification biométrique 16
2.3.2 Performance du schéma d'identification 17
2.4 Sécurité de données biométriques 18
2.4.1 Stockage centralisé ou décentralisé 18
2.4.2 Protection des données stockées 19
2.4.3 Schéma d'identification avec transformation 23
2.4.4 Projection 23
2.5 Données biométriques utilisées 24
2.5.1 Empreintes digitales : FVC2002 25
2.5.2 Visages : LFW 26

2.5.3 Électrocardiogramme : PTB . . . 27
2.5.4 Interprétations . . . 29
2.6 Individus à fort potentiel d'usurpation . . . 30
2.6.1 Classification . . . 30
2.6.2 Dans les bases utilisées . . . 30
2.7 Algorithme génétique . . . 33
2.8 Méthode d'escalade . . . 35
2.9 Attaques existantes sur les transformations . . . 35
2.9.1 Fuites d'informations depuis les gabarits . . . 36
2.9.2 Différentes méthodes d'attaques . . . 36
2.9.3 Attaques avec algorithmes génétiques . . . 39
2.10 Conclusion . . . 40

3 Préimage proche et réutilisable 43
3.1 Introduction . . . 43
3.2 Préimage 44
3.2.1 ... proche 44
3.2.2 ... et réutilisable . . . 45
3.3 Comparaison avec d'autres algorithmes . . . 46
3.3.1 Choix parmi l'existant . . . 46
3.3.2 Construction aléatoire . . . 48
3.3.3 Construction par escalade . . . 50
3.4 Construction de préimage avec un algorithme génétique . . . 52
3.4.1 Fonction d'évaluation . . . 54
3.4.2 Taille de la population et nombre d'itérations . . . 54
3.4.3 Étape de sélection . . . 57
3.4.4 Étape de mutation . . . 59
3.4.5 Étape de croisement . . . 60
3.4.6 Paramètres optimaux . . . 64
3.5 Variantes de choix de gabarits . . . 68
3.6 Préimage universelle . . . 71
3.6.1 Passe-partout . . . 71
3.6.2 Construction d'un passe-partout pour une base de données biométriques révocables . . . 72
3.6.3 Cas d'usage . . . 74
3.7 Conclusion . . . 75

4 Passe-partout biométrique 77

4.1 Introduction 77
4.2 Transformations utilisées 79
4.2.1 Lemme de Johnson-Lindenstauss 79
4.2.2 Projections proposées par Achlioptas 79
4.2.3 Gain de performance sur le coût de génération des matrices . . 80
4.2.4 Analyse de ces projections 80
4.3 Passe-partout : construction d'une base de données biométriques révocables 83
4.3.1 Recherche de graines pour un vecteur de caractéristiques . . . 84
4.3.2 Résultats des expériences 85
4.4 Complexité des deux scénarios 88
4.5 Extension à un individu passe-partout 89
4.5.1 Recherche de graines pour un ensemble de vecteurs de caractéristiques 89
4.5.2 Résultats des expériences pour l'ensemble de recherche et pour l'ensemble de test 90
4.5.3 Corrélation entre ces ensembles 92
4.6 Cas d'usage 94
4.7 Conclusion 95

5 Perspectives **97**
5.1 Introduction 97
5.2 Passe-partout réutilisable 98
5.2.1 Contexte 98
5.2.2 Résultats 98
5.2.3 Conclusion 100
5.3 Passe-partout sur d'autres bases 100
5.3.1 Base multimodale 100
5.3.2 Base issue d'oreilles 103
5.3.3 Conclusion 105
5.4 Conclusion et perspectives 106
5.4.1 Préimage proche et réutilisable 106
5.4.2 Construction d'un passe-partout pour une base de données biométriques révocables 107
5.4.3 Construction d'une base de données biométriques révocables pour un passe-partout 107
5.4.4 Perspectives communes 108

6 Conclusion **109**
6.1 Contexte 109
6.2 Problématique 109
6.3 Contributions 110
6.3.1 Préimage proche et réutilisable 110
6.3.2 Passe-partout biométriques 111
6.4 Perspectives 113

7 Publications de l'auteur **115**

Bibliographie **117**

Table des figures **125**

Liste des tableaux **128**

Chapitre 1

Introduction

Résumé : *Ce court chapitre introduit l'évolution de l'informatique et la nécessité de sécuriser ce domaine. La biométrie et sa sécurité, formant le point central dans cette thèse, ainsi que deux contributions, sont introduites, avant de donner les perspectives.*

1.1 Contexte

L'informatique est un domaine qui a révolutionné nos vies dans de nombreuses facettes. Son utilisation est aujourd'hui massive et encore croissante. La communication, la téléphonie, internet, les véhicules et même nos compteurs d'électricité sont informatisés et communicants.

Cet usage massif et croissant envahit notre quotidien et permet souvent de le simplifier, de l'accélérer, et de faire jaillir de nouveaux usages.

Néanmoins, cette technologie pénètre de plus en plus dans notre sphère privée, avec à la clé la connaissance d'informations sensibles. C'est pourquoi, comme pour notre maison, imbibée de notre histoire et de nos biens, nous la souhaitons inviolable.

La sécurité informatique est alors apparue, avec pour objectif de protéger les systèmes informatiques en empêchant des vols de données, des contournements de système, et en garantissant leur fonctionnement continu. Comme pour notre maison, nous voulons être protégés des intrusions tout en voulant aller et venir ou recevoir temporairement des invités.

Comme pour notre maison, ce n'est pas inviolable : finalement, la sécurité réside entre la durée acceptable pour un utilisateur légitime d'aller et venir - ouvrir la porte - et une durée bien plus importante pour un attaquant - casser la porte, ou la baie vitrée, l'objectif étant de rentrer - nous laissant le temps d'intervenir.

1.2 La biométrie

Cette thèse en sécurité informatique s'intéresse à la biométrie. Ce domaine consiste à reconnaître quelqu'un, sans qu'il ne possède une clé, sans qu'il ne connaisse un mot de passe. Dans notre quotidien, nous sommes au contact de la biométrie avec nos téléphones, équipés de capteurs d'empreintes digitales ou de reconnaissance faciale. Cette technologie est proposée en alternative aux mots de passe, aux codes à 4 chiffres, aux schémas de déverrouillage. Elle permet un accès plus rapide, et plus discret, contrairement aux codes qu'un tiers peut observer puis rejouer.

Tel le jeu du chat et de la souris, cet usage de la biométrie permet de protéger nos données personnelles, le contenu du téléphone, par une autre donnée personnelle, notre empreinte digitale ou notre visage. En effet, on ne veut pas qu'un attaquant ayant un accès au système de déverrouillage puisse obtenir notre empreinte digitale, pourtant nécessaire au dit système pour la reconnaissance de l'individu légitime.

1.3 Sécurité biométrique

Des méthodes de protection de ces informations biométriques sont ainsi apparues. L'objectif est de conserver une version protégée de l'empreinte digitale, ne permettant pas de la retrouver, mais conservant l'efficacité de la reconnaissance.

Dans les travaux de cette thèse, nous avons étudié la sécurité de ces protections de données biométriques. Nous commençons par décrire précisément ce qu'est la biométrie ainsi que la nécessaire protection qu'elle implique, puis nous présentons dans ce manuscrit deux contributions.

Nous utilisons des données biométriques sous forme de vecteurs à valeurs réelles, dont la protection consiste à transformer ce vecteur en le projetant en un plus petit vecteur à valeurs binaires, dit gabarit. Cette transformation est paramétrée par une graine considérée publique.

1.4 Contributions

Première contribution :

Une protection de données biométriques doit satisfaire différentes propriétés, dont la non-inversibilité. C'est cette propriété que nous avons attaquée en essayant d'inverser la projection. L'objectif de l'attaque est de construire un vecteur depuis un gabarit corrompu, permettant de s'authentifier. Par exemple, nous avons en notre

possession un téléphone depuis lequel nous avons extrait le gabarit stocké. Nous devons construire un vecteur que nous proposons au système de déverrouillage.

Nous utilisons un algorithme génétique dont les paramètres ont été choisis pour optimiser le résultat. Cette attaque d'inversion de gabarit construisant un vecteur permettant de s'authentifier existait déjà.

Nous avons apporté la construction de préimage proche et réutilisable. Dans ce concept, nous construisons un vecteur depuis deux gabarits avec un algorithme génétique, et ce vecteur permet de s'authentifier avec le premier gabarit, mais aussi avec le second gabarit issu du même individu. Par exemple, nous avons en notre possession le téléphone et l'ordinateur d'un individu, et nous avons réussi à extraire le gabarit du téléphone et celui de l'ordinateur. Nous construisons un unique vecteur permettant de déverrouiller le téléphone et l'ordinateur.

Deuxième contribution :

Nous avons étendu ce concept de préimage proche et réutilisable au passe-partout biométrique.

Dans un premier scénario, nous possédons des gabarits issus de plusieurs individus. Par exemple, une entreprise a 50 téléphones avec protection biométrique, et chaque téléphone est protégé avec le gabarit de l'employé le possédant. À partir de ces gabarits, nous construisons un unique vecteur permettant de déverrouiller l'ensemble des téléphones.

Dans un second scénario, nous avons à nouveau cette entreprise avec les 50 téléphones, et nous souhaitons que le directeur de l'entreprise puisse tous les déverrouiller. Nous n'avons donc pas le choix du vecteur passe-partout. Pour chaque téléphone, nous choisissons la graine paramétrant la projection du vecteur de l'employé en un gabarit, de telle sorte que le vecteur du directeur déverrouille le téléphone avec ce gabarit. Nous avons ensuite étendu ce scénario pour permettre à de futurs vecteurs du directeur de déverrouiller les téléphones.

1.5 Perspectives

Nous finissons cette thèse par les perspectives et les nouvelles problématiques qu'elle apporte. Nous apportons des indicateurs sur la réutilisabilité d'un passe-partout, puis nous mettons en avant les différences de performances de choix de graines pour un passe-partout en fonction des données biométriques utilisées. Enfin, nous détaillons les perspectives pour chaque contribution.

Chapitre 2

État de l'art et définitions

Résumé : *Ce chapitre introduit la biométrie, avec les objectifs, cas d'usages, et données utilisées. Les risques d'attaques sont introduits ainsi que les nécessaires protections. Les fuites de données des transformations, permettant des attaques de construction de préimages, notamment avec des algorithmes génétiques, sont présentées.*

Mots-clés : *biométrie, protection, identification, attaque, algorithme génétique.*

2.1 La biométrie

La biométrie est une science se basant sur des caractéristiques physiques, comportementales et biologiques. On utilise ces caractéristiques pour reconnaître un individu parmi un ensemble d'individus. Cette gestion de l'identité est importante dans notre société, avec les enquêtes criminelles, les contrôles aux frontières, déverrouiller un appareil informatique comme un téléphone, ouvrir une porte, payer, ou même se connecter sur internet.
La biométrie est une alternative à d'autres méthodes d'identification d'un individu. Dans le cas des téléphones ou des sites internet, le mot de passe est le moyen le plus utilisé. Dans un contexte plus physique comme le contrôle aux frontières, la possession d'une carte d'identité permet notre authentification.
La biométrie est une méthode d'identification ou d'authentification décrivant *ce que l'on est*, en alternative à *ce que l'on sait* et *ce que l'on possède*. Une vue générale de la biométrie, des différentes modalités utilisées et de ses usages se trouve dans le *Handbook of biometrics* [Jain et al., 2007]. Ces informations peuvent être

inaltérables, c'est-à-dire qu'elles ne changent pas au cours de l'existence de l'individu. D'autres en revanche évoluent au fur et à mesure de la vie, comme le visage. L'objectif étant de reconnaître un individu, de telles caractéristiques doivent être discriminantes et avoir une certaine unicité propre à l'individu qui les porte.

Dans cette section, nous commençons par décrire brièvement différentes modalités biométriques, incluant celles qui seront utilisées dans cette thèse, puis nous décrivons les principes d'identification et d'authentification biométriques avant de terminer par une description des enjeux de sécurité engendrés par la biométrie.

2.1.1 Différentes modalités biométriques

Nous allons décrire trois catégories de modalités biométriques en y donnant des exemples puis en fournissant quelques indicateurs. Nous décrivons ensuite la multibiométrie, qui allie plusieurs modalités biométriques.

Biométrie physique

La biométrie peut utiliser des caractéristiques physiques. On y trouve notamment les empreintes digitales, les veines des doigts et des mains, l'iris, le visage, et les oreilles. Les caractéristiques physiques permettent une bonne performance de reconnaissance [Jain et al., 1999a].

Biométrie biologique

On peut utiliser des caractéristiques biologiques. On retrouve l'ADN, unique sauf dans le cas de vrais jumeaux (qui sont eux discriminables par leurs empreintes digitales), et permanent.
On utilise différents signaux biologiques comme l'électrocardiogramme (ECG), l'électroencéphalogramme (EEG), et l'électromyogramme (EMG). La capture de telles modalités biométriques biologiques est difficile, et nécessite du matériel spécifique.

Biométrie comportementale

La biométrie comportementale se base sur les manières de faire qu'a un individu pour effectuer une certaine tâche. On y trouve la démarche, la dynamique de frappe au clavier, de signature, de souris, le regard, la voix, et l'utilisation d'un écran tactile. Les données comportementales ne sont pas utilisées dans cette thèse. Parmi les récentes études sur le sujet, on peut se reporter vers l'article qui l'aborde dans le cadre de l'authentification avec un téléphone de [Alzubaidi and Kalita, 2016].

Par exemple la dynamique de frappe au clavier utilise les intervalles de temps entre pression et relâchement des touches du clavier. L'étude de [Teh et al., 2013] synthétise les travaux existants et conclut que cette modalité biométrique peut difficilement être utilisée pour de l'identification, mais peut être utilisée comme complément. De leur côté, [Giot et al., 2015] analysent les disparités entre les processus d'acquisition de différentes données dans les bases publiques, ce qui complique leur analyse.

Intérêt des différentes modalités

Maintenant que nous connaissons plusieurs modalités biométriques, nous donnons des indicateurs quant à l'utilisation de certaines modalités, issus des travaux de [Jain et al., 1999a], un livre de référence. On y trouve notamment des indicateurs d'unicité, de facilité d'acquisition, du caractère permanent de la modalité, de performance, et d'acceptabilité , dont certains sont reportés dans le tableau 2.1.

Modalité	Acquisition	Unicité	Permanent	Acceptabilité	Performance
Visage	Facile	Faible	Moyen	Fort	Faible
Digitale	Moyen	Forte	Forte	Moyen	Forte
Rétine	Difficile	Forte	Moyen	Faible	Forte
Signature	Facile	Faible	Faible	Forte	Faible
Vocale	Moyen	Faible	Faible	Forte	Faible

TABLE 2.1 – Comparaison de modalités biométriques [Jain et al., 1999a]

On remarque qu'une unicité importante implique une performance importante. Il faut garder à l'esprit que cette évaluation date de 1999, où les technologies d'extraction de caractéristiques notamment par réseaux de neurones profonds n'étaient pas avancées. Ceci explique la faible performance du visage par exemple.

Multibiométrie

Les différentes modalités biométriques précédemment décrites peuvent être utilisées ensemble dans le cadre de la multibiométrie que nous détaillons maintenant. Dans l'objectif d'améliorer les performances d'identification à partir de caractéristiques biométriques, on peut être amené à utiliser plusieurs modalités biométriques en même temps. Par exemple, lors d'un contrôle d'accès, poser la main sur un capteur qui récupère les empreintes digitales et le trajet des veines, et avoir une caméra prenant une photo du visage. On peut combiner de la biométrie physique et comportementale. Cela nous permet d'améliorer les performances de l'identification, en multipliant des caractéristiques discriminatoires, et ainsi pallier divers défauts comme un visage blessé, un doigt abimé, ou le changement avec l'âge ou le temps.

Un des premiers schémas de multibiométrie exploitant plusieurs modalités a été proposé par [Brunelli and Falavigna, 1995]. Un tel système doit alors fusionner les résultats, ce qui peut se faire de plusieurs manières. Ainsi, [Ross and Jain, 2003] décrivent les 3 niveaux de fusion suivants :

- Fusion de caractéristiques : les caractéristiques obtenues depuis les différentes captures de modalités biométriques sont utilisées pour calculer un unique vecteur de caractéristiques. Une manière classique est de concaténer les vecteurs de caractéristiques obtenues depuis les différentes captures.
- Fusion de scores : les caractéristiques sont extraites indépendamment depuis chaque capture, puis sont comparées indépendamment aux autres vecteurs de caractéristiques de la base pour produire un score, comme une distance. Les distances obtenues sont fusionnées. Pour cela, les auteurs proposent d'utiliser de la régression logistique pour combiner les scores, telle que décrite par [Jain et al., 1999b]. L'objectif est de minimiser le taux de faux rejets pour un taux de fausses acceptations fixé.
- Fusion de décisions : les captures ont été indépendamment extraites, puis comparées afin d'obtenir des distances, décidant l'acceptation ou le rejet. La fusion de décisions récupère les différentes décisions d'acceptations et de rejets pour finalement décider de l'acceptation ou du rejet. Un schéma de vote à la majorité, tel que décrit dans [Zuev and Ivanov, 1999], peut être utilisé.

Il y a eu beaucoup de travaux sur la biométrie multimodale : pour plus de détails, le lecteur peut se référer à la récente étude [Singh et al., 2019].

Nous avons introduit différentes modalités biométriques, données en plusieurs catégories, ainsi que des indicateurs quant à leur utilisation. Nous allons désormais décrire leur utilisation globale.

2.1.2 Utilisation de la biométrie

La biométrie est principalement utilisée selon deux modes : l'identification et l'authentification. Les concepts de ces modes sont décrits avec des exemples et nous discutons de l'usage actuel de la biométrie.

Identification ou authentification

Il faut impérativement poser la différence entre identification et authentification. Tout d'abord, l'individu doit s'enregistrer dans la base, cette étape est la même que l'on utilise l'identification ou l'authentification.

Enregistrement

Cette étape permet l'enregistrement d'un individu et précède nécessairement l'authentification ou l'identification. Il s'agit d'enregistrer l'individu dans une base de données, avec son identité et une donnée permettant de le reconnaître. On parle aussi d'enrôlement.

Authentification

L'authentification est considérée comme une vérification. Un individu déclare qui il est, et fournit une donnée permettant de vérifier qu'il est bien la personne qu'il prétend être. On effectue alors une comparaison, entre la donnée fournie et la donnée stockée pour l'utilisateur ciblé. C'est une comparaison $1 - 1$.

Exemple : notre téléphone portable, dans lequel nous avons enregistré une empreinte digitale à l'initialisation, vérifie à chaque futur déverrouillage que l'empreinte digitale nouvellement acquise correspond à celle de la phase d'enregistrement.

Identification

Un individu fournit une donnée, et il faut chercher dans l'ensemble des N individus de la base qui il est. On fait alors autant de comparaisons qu'il y a de personnes dans la base. L'identification est plus coûteuse que l'authentification, c'est une comparaison $1 - N$.

Exemples : dans la cantine d'un lycée, un élève pose son doigt sur un capteur pour être facturé. Le système doit, uniquement avec l'empreinte digitale acquise, trouver quel est l'élève pour le facturer. Une variante : un employé souhaitant rentrer dans les locaux de l'entreprise pose son doigt sur un capteur, le système doit déduire, uniquement avec l'empreinte digitale acquise, si l'employé est autorisé à rentrer, sans désigner précisément un individu.

Une version faible de l'identification consiste à vérifier si l'individu est dans la base, sans trouver qui et sans qu'il indique qui il est, comme dans le second exemple.

Contrôle d'accès

Le contrôle d'accès est omniprésent dans le monde actuel. Deux grandes catégories coexistent : le contrôle d'accès physique, à des bâtiments, à des bureaux, à des locaux techniques, et le contrôle d'accès numérique, comme se connecter sur un ordinateur ou lire une base de données. Ces deux catégories de contrôle coexistent et peuvent utiliser les mêmes informations pour autoriser ou non une demande d'accès. Un tel contrôle doit être adapté à la diversité des terminaux utilisés : badgeuse,

téléphone, ordinateur, applications. Il doit aussi être adapté pour de nombreux utilisateurs dont les permissions varient ou expirent. Pour demander une autorisation d'accès, les utilisateurs peuvent utiliser une carte d'accès, un mot de passe, mais aussi de la biométrie. La CNIL, dans son règlement relatif à l'accès par authentification biométrique sur les lieux de travail [CNIL, b], fournit les finalités éligibles à un système biométrique : il s'agit de locaux, ou d'applications limitativement identifiés de par leur sensibilité, nécessitant un haut niveau de protection par biométrie. Quatre principaux types de modèles de contrôle d'accès sont proposés dans la récente étude [Parkinson and Khan, 2022] :

— Contrôle d'accès basé sur les rôles : ce modèle prend l'hypothèse que des utilisateurs ayant le même rôle ont besoin des mêmes ressources.
Exemples de rôles *:* gestionnaire des ressources humaines, comptables, administrateurs système, agents d'entretien.

— Contrôle d'accès discrétionnaire : un individu hérite des droits d'accès de son groupe. Il peut aussi donner une autorisation d'accès qu'il possède à un autre individu.
Exemple *:* le système de droits lecture-écriture-exécution de Linux avec les groupes et les utilisateurs.

— Contrôle d'accès obligatoire : dans ce modèle, une autorité impose les droits d'accès. Ce modèle est utilisé des cadres strictes, militaire par exemple.
Exemple *:* l'autorité contrôle les accès, et non le propriétaire d'une ressource.

— Contrôle d'accès basé sur les attributs : ce modèle utilise les attributs d'un individu pour définir les autorisations. Une majorité des récentes recherches porte sur ce modèle, considéré comme étant la prochaine génération de contrôle d'accès, en permettant une importante flexibilité et une faible granularité.
Exemple *:* un responsable des ressources humaines peut accéder au dossier des employés de son site. S'il change de site, ses droits évoluent automatiquement.

Les travaux de cette thèse, et particulièrement ceux de notre second scénario décrit dans la section 4.3, contribuent à résoudre des problématiques de contrôle d'accès.

Une utilisation croissante

L'usage de la biométrie est croissant dans la société. Nos empreintes digitales ont historiquement été utilisées dans le cadre d'enquêtes judiciaires, suivies de l'analyse ADN lorsque les technologies l'ont permis. La confirmation d'une identité par empreinte digitale est largement déployée aux frontières avec les passeports biométriques. Outre ces cas d'usage régalien, la biométrie prend une place croissante dans notre quotidien. Elle est désormais massivement intégrée dans le déverrouillage de

téléphone et d'ordinateur, dans le contrôle d'accès à un trousseau de mots de passe, dans le contrôle d'accès à un disque dur. Ces usages du quotidien utilisent principalement l'empreinte digitale et la reconnaissance faciale. Les capteurs pour acquérir ces données sont peu onéreux, et leur usage est de plus en plus accepté avec confiance dans la société.

Nous avons globalement décrit l'identification et l'authentification à titre de compréhension. Une description technique est donnée dans la section 2.3. Nous abordons désormais des problématiques relatives aux risques de la biométrie.

2.1.3 Sensibilité des données biométriques

Nous allons discuter des problématiques de vie privée émergeant avec l'usage croissant de la biométrie. Nous introduirons ensuite la biométrie douce, en réaction à ces problématiques.

Vie privée

Les données biométriques sont des données à caractère personnel, dont la protection fait l'objet d'études variées depuis plusieurs décennies [Prabhakar et al., 2003]. En effet, elles permettent l'identification d'un individu. De plus, ces données possèdent une propriété non négligeable : on ne peut pas en changer. Contrairement à une adresse email, une adresse postale, un numéro de téléphone, une plaque d'immatriculation : une caractéristique biométrique ne peut pas être remplacée simplement. Dans le cadre des empreintes digitales, si un doigt est corrompu, on peut en utiliser un autre. Mais ça ne fonctionne que dix fois. La législation est donc stricte en ce qui concerne le stockage des ces données et leur protection.

Cas de la biométrie douce

La biométrie douce permet de catégoriser des groupes d'individus, mais ne permet pas l'identification d'un individu parmi ce groupe. Elle se base sur des informations peu discriminantes, comme la taille, la couleur des yeux ou des cheveux, le poids, le genre.
[Reid et al., 2013] décrivent l'identification avec la biométrie douce, qui ne nécessite pas forcément de capteurs précis, et mettent en avant le manque de traits distinctifs et la difficulté à identifier quelqu'un sans autre information et sans restreindre l'identification à un petit ensemble d'individus.

Protection des données

Les données biométriques sont sensibles et massivement acquises, stockées, et utilisées. La thématique de la protection de ces données est importante. De multiples contributions ont été apportées pour tenter de trouver un compromis entre l'efficacité en temps et en précision de l'identification, l'utilisabilité au quotidien, le coût, et la protection de ces données. Ces avancées sont données dans la section 2.4.

La sensibilité de ces données donne de l'attrait aux attaques sur ces systèmes biométriques, que nous introduisons désormais.

Attaque sur un système biométrique

Le périmètre d'attaque sur un système biométrique est vaste. [Jain et al., 2000] décrivent un système biométrique générique pour lequel 8 points d'attaques génériques sont proposés par [Ratha et al., 2001a].

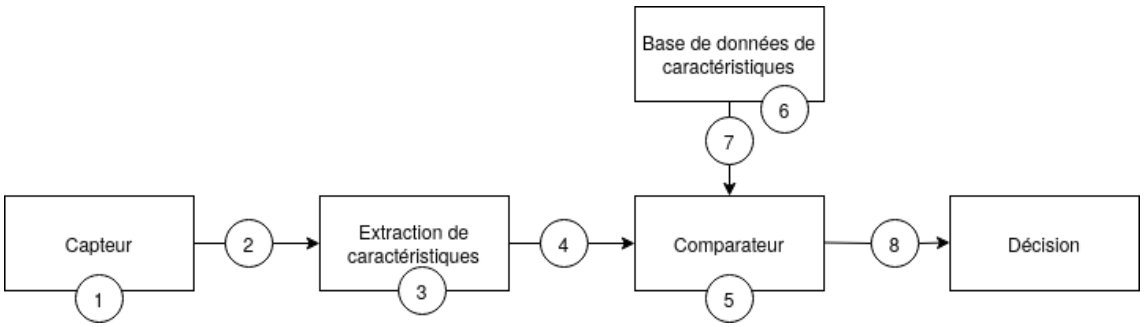


FIGURE 2.1 – Points d'attaques génériques dans un système biométrique [Ratha et al., 2001a]

La figure 2.1 représente les 8 points d'attaques possibles. Le point 1 correspond à une attaque par présentation : on présente au capteur qui acquiert la modalité biométrique un faux doigt ou un visage imprimé en 3D. Le point 2 correspond à une attaque par rejeu, on renvoie une capture déjà capturée que l'attaquant a interceptée. L'attaque du point 3 correspond à une compromission du module effectuant l'extraction de caractéristiques : l'attaquant peut par exemple fixer le vecteur de caractéristiques transmis par le module, quelle que soit la capture donnée en entrée. Il peut transmettre ce vecteur préétabli seulement si c'est son doigt qui a été acquis et dont la capture a été transmise. L'attaque du point 4 consiste à corrompre les caractéristiques transmises, de manière similaire au point 2. L'attaque au point 5 consiste en la modification du score de proximité entre les caractéristiques comparées. L'attaque au point 6 correspond à une corruption de la base de données de caractéristiques. L'attaquant peut avoir modifié les caractéristiques d'un individu

par les siennes afin de l'usurper. L'attaque au point 7 est similaire aux points 2 et 4, avec la modification des caractéristiques transmises au comparateur. L'attaque au point 8 peut inverser la décision d'identification au moment où elle est transmise.

2.2 Base de données biométriques

Pour reconnaître un individu parmi un ensemble d'individus, nous allons travailler avec des données biométriques. Dans cette section, nous allons tout d'abord détailler les captures de modalités biométriques. Cette première étape nous permet de récupérer des informations biométriques pour des individus. Dans un second temps, nous souhaitons travailler avec ces captures. Pour cela, nous allons en extraire des informations sous un format normalisé. Enfin, pour reconnaître un individu, il nous faut le comparer. C'est dans un dernier temps que nous détaillons comment comparer des données biométriques extraites.

2.2.1 Les captures de modalités biométriques

Pour effectuer des travaux de recherche, évaluer des algorithmes utilisant des données biométriques, comparer différentes modalités biométriques, ou comparer différents capteurs, nous avons besoin de lots de données biométriques. Il s'agit généralement de captures de modalités biométriques de plusieurs personnes, idéalement plusieurs par personne.
Par exemple, plusieurs photos de visages de plusieurs personnes. Pour une même personne, on peut avoir quelques photos distinctes, en fonction de différents éclairages, de différents angles par rapport au capteur, ou de différentes coiffures.
Dans le cadre des empreintes digitales, l'image du doigt varie entre autre en fonction de la pression effectuée, de la sueur, de la rotation.
Évidemment, les images restituées dépendent du capteur utilisé. Dans le cadre des visages, cela dépend de la qualité des lentilles de l'appareil photo, la distance, la résolution. Pour les empreintes digitales, on peut avoir une photo de doigt, une photo d'empreinte d'encre laissée par un doigt, une image restituée à partir d'un capteur pyroélectrique ou à ultrasons.
Il est nécessaire de vérifier la qualité d'une capture de modalité biométrique. Comme l'indiquent [Jain et al., 2011], divers éléments peuvent fortement influer la qualité de la capture et donc la fiabilité du système biométrique. Dans le cas d'une capture jugée de qualité insuffisante, il convient d'effectuer une nouvelle capture jusqu'à atteindre la qualité attendue.

Dans cette thèse, nous considérons une base de captures de modalités biométriques telle que définie dans 2.2.1.

Définition 2.2.1. *Soit $M = \{c_i\}_{i=1,...,n}$ une base de n captures de modalités biométriques. Elle est composée pour chaque individu d'une unique capture d'une modalité biométrique.*

Après avoir défini ce qu'est une base de captures de modalités biométriques, nous nous intéressons à l'extraction de caractéristiques depuis ces captures de modalités biométriques.

2.2.2 Extraction de caractéristiques

Pour pouvoir exploiter les données biométriques, il faut extraire des caractéristiques depuis les captures de modalités biométriques. Dans un premier temps, on peut effectuer un prétraitement sur la capture brute, pour retirer du bruit et mettre en avant la modalité biométrique utilisée.
Il faut ensuite décrire la capture à l'aide de caractéristiques extraites. Dans le cadre des empreintes digitales, on peut extraire l'ensemble des minuties. Une minutie est une singularité dans l'empreinte digitale, comme une terminaison de ligne papillaire, ou une bifurcation. On décrit cette singularité selon un modèle défini, pour finalement avoir un ensemble de descriptions de singularités. Néanmoins, on peut avoir besoin que les caractéristiques extraites soient sous une forme donnée de taille fixe, comme un vecteur composé de t valeurs réelles de b bits. Ainsi, quelles que soient la modalité capturée et la taille de la capture, on est certain d'avoir cette structure de représentation des caractéristiques. On les appelle des vecteurs de caractéristiques.

Dans nos travaux, une donnée biométrique est représentée sous la forme d'un vecteur de caractéristiques tel que précédemment décrit. L'algorithme permettant l'extraction de caractéristiques, depuis la capture de modalité biométrique, sous forme de vecteur, diffère selon la modalité biométrique utilisée. Pour un type de modalité biométrique, différents algorithmes sont possibles. Une formalisation générale de l'algorithme d'extraction est donnée en 2.2.2 et sa schématisation en 2.2.

Nous stockons ces vecteurs de caractéristiques, extraits avec l'algorithme d'extraction, dans une base de données biométriques définie en 2.2.3.

Définition 2.2.2. *Soit $E_t(.)$ un algorithme d'extraction de caractéristiques prenant en entrée une capture de modalité biométrique et restituant en sortie un vecteur de caractéristiques de taille t.*

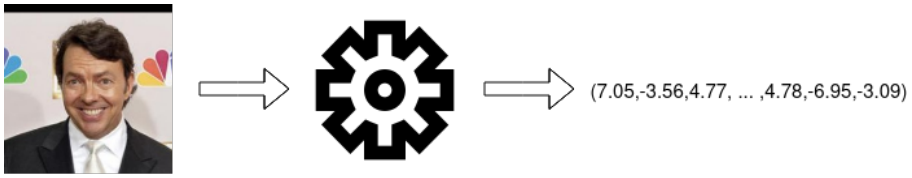


FIGURE 2.2 – Algorithme d'extraction de caractéristiques

Définition 2.2.3. *Soit $B = \{x_i\}_{i=1,...,n}$ une base de données biométriques composées de vecteurs de caractéristiques. Elle a été construite depuis la base de capture de modalités biométriques M définie dans 2.2.1. On a $x_i = E_t(c_i)$ avec $x_i \in B$, $c_i \in M$, et $E_t(.)$ défini dans 2.2.2.*

Ces caractéristiques sont à valeur dans un espace métrique, généralement réel ou binaire. Dans la suite, on définit les distances à notre disposition pour comparer de tels vecteurs de caractéristiques.

2.2.3 Comparaison de caractéristiques

On rappelle que l'objectif de la biométrie est d'identifier un individu parmi une base connue, initialement la base de captures de modalités biométriques M telle que définie dans 2.2.1. Les caractéristiques extraites de cette base M ont permis la construction de la base de données biométriques B 2.2.3. Un individu souhaitant s'identifier se soumet à une capture de modalité biométrique, la même que celle utilisée pour son enregistrement dans la base M. On note cette capture destinée à l'identification C^*. L'objectif est de comparer le vecteur de caractéristiques extrait $x^* = E_t(C^*)$ 2.2.2 avec les vecteurs de caractéristiques dans la base B, d'évaluer les similarités, de mettre en avant l'individu le plus proche, et enfin de décider si la proximité est telle que c'est bien la même personne.
Dans notre cas, avec une représentation sous forme de vecteurs de caractéristiques, la comparaison se fait en calculant une distance telle que définie dans 2.2.4 et 2.2.5.

Définition 2.2.4. *Soit x, x' deux vecteurs de taille N, à valeurs réelles.*
Leur comparaison s'effectue avec la distance euclidienne :
$D_A(x, x') = \sqrt{\sum_{i=1}^{N} (x_i - x'_i)^2}$.

Définition 2.2.5. *Soit u, u' deux vecteurs de taille M, à valeurs binaires.*
Leur comparaison s'effectue avec la distance de Hamming :
$D_B(u, u') = \sum_{i=1}^{M} u_i \oplus u'_i$, *avec $\oplus$ le XOR.*

Nous savons désormais comment extraire les caractéristiques depuis une capture de modalité biométrique, puis comment comparer ces caractéristiques.

Nous allons désormais formaliser la comparaison dans un objectif d'authentification et d'identification.

2.3 Système biométrique

Dans cette section, nous décrivons comment réaliser un schéma d'identification ou d'authentification à partir de ces données biométriques. Nous finissons en détaillant l'évaluation de performance d'un schéma d'identification biométrique.

2.3.1 Schéma d'identification biométrique

La finalité du système biométrique est d'identifier un individu à partir d'une nouvelle capture de modalités biométriques, parmi un ensemble d'individus. Cet ensemble est stocké dans une base de données. Un schéma biométrique générique s'effectue en deux phases.

1. La première phase s'appelle l'enrôlement. Il s'agit d'inscrire un nouvel utilisateur. Pour cela il faut capturer une ou plusieurs fois une ou plusieurs modalités biométriques que l'on va enregistrer dans la base de données biométriques.
2. La seconde phase est l'identification. L'utilisateur préalablement enregistré effectue une nouvelle capture de modalités biométriques, que l'on va comparer à celles de la base de données biométriques.

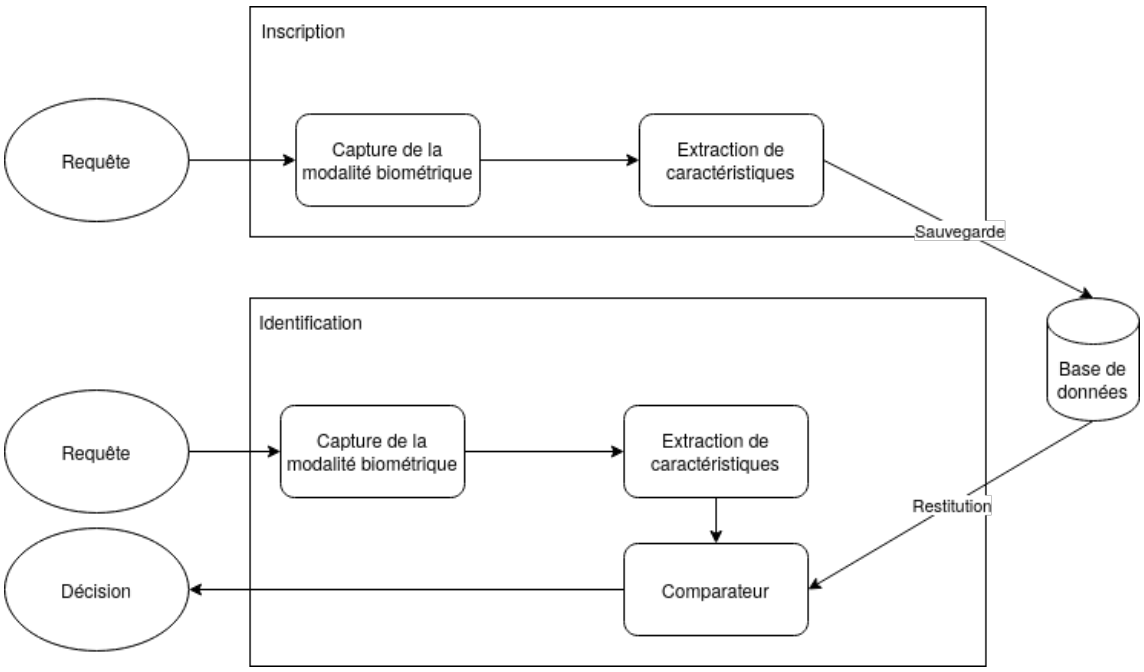


FIGURE 2.3 – Système biométrique générique [Jain et al., 2000]

Nous discutons ici des parties comparateur et décision telles que présentées dans la figure 2.3. L'objectif est de comparer un individu qui effectue une capture avec une capture présente dans la base, et de déterminer si les caractéristiques sont *assez proches* pour valider l'identification. Le terme *assez proche* est définie par un seuil τ fixant une limite de différences acceptables.

On commence par définir un schéma d'authentification 2.3.1 puis on généralise en définissant un schéma d'identification 2.3.2.

Définition 2.3.1. *Soit $(\mathcal{M}_A, D_A)$ un espace métrique. Un schéma d'authentification biométrique pour deux captures de modalité biométrique b et b' est une paire d'algorithmes (E, V) avec :*

- *E prend une capture d'une modalité biométrique b comme entrée, et retourne un vecteur de caractéristiques $x \in \mathcal{M}_A$.*
- *V prend deux vecteurs de caractéristiques $x = E(b)$, $x' = E(b')$, et un seuil τ_A comme entrées, et retourne $Vrai$ si $D_A(x, x') < \tau_A$, et $Faux$ sinon.*

Remarque 1. *Il est attendu que deux vecteurs de caractéristiques x et x' issus de deux captures d'une même modalité biométrique d'une personne soient assez proches pour vérifier $D_A(x, x') < \tau_A$, avec D_A la distance euclidienne définie dans 2.2.4.*

Définition 2.3.2. *Soit $B = \{x_i\}_{i=1,\dots,n}$ une base de données biométriques composées de vecteurs de caractéristiques. Un schéma d'identification pour B est un algorithme qui prend un vecteur de caractéristiques x et un seuil τ_A comme entrées et retourne $Vrai$ s'il existe i entre 1 et n tel que $V(x, x_i, \tau_A) = Vrai$ et retourne $Faux$ sinon. Dans le premier cas, x est identifié avec succès comme x_i.*

Ainsi, pour identifier un utilisateur représenté par son vecteur de caractéristiques x au sein de la base de données biométriques, on récupère un par un les i vecteurs de caractéristiques stockés, notés x_i, et on teste si la distance $D(x, x_i)$ est inférieure au seuil τ. Si c'est le cas, l'utilisateur est identifié comme étant l'individu i.

2.3.2 Performance du schéma d'identification

La performance du schéma d'identification est estimée avec différents indicateurs [Jain et al., 2004].

Définition 2.3.3. *Le taux de fausses acceptations (FMR) donne le pourcentage d'imposteurs acceptés par le système.*

Définition 2.3.4. *Le taux de faux rejets (FNMR) donne le pourcentage d'utilisateurs légitimes rejetés par le système.*

Définition 2.3.5. *Les taux FMR et FNMR dépendent du seuil τ utilisé. Lorsqu'un seuil τ permet l'égalité de ces deux taux, on note ce seuil $\tau@EER$ et $EER = FMR@\tau = FNMR@\tau$.*

Le taux τ est fixé en fonction des exigences de sécurité et d'utilisabilité. Dans un cadre de contrôle d'accès strict, on souhaite qu'aucun imposteur ne soit accepté. Ainsi, on fixe τ lorsque $FMR@\tau = 0$.
Si l'on souhaite un système convivial, on souhaite qu'aucun utilisateur légitime ne se voie refuser l'accès. Dans ce cas, on fixe τ lorsque $FNMR@\tau = 0$.
Lorsque l'on souhaite un système intermédiaire, alliant sécurité et convivialité, on fait un compromis en fixant τ de telle sorte que $FMR@\tau = FNMR@\tau$. Dans ce cas, la sécurité et la convivialité dépendant fortement de la base et de son EER.

Nous avons formalisé l'identification et décrit les performances d'un tel schéma. La prochaine section aborde la sécurité de ces données biométriques, et formalise leur protection avec un nouveau schéma.

2.4 Sécurité de données biométriques

Les données biométriques sont sensibles de par leur possibilité d'identification et leur caractère permanent. C'est pourquoi les captures de modalités biométriques et les vecteurs de caractéristiques ne sont pas stockés et utilisés sous ces formes. Une compromission du stockage impliquerait un vol irréparable de ces données que l'on ne peut ni révoquer ni remplacer.
Cette section aborde l'impact du mode de stockage, puis décrit les protections de données biométriques existantes avec les propriétés qu'elles doivent apporter. Un nouveau schéma biométrique est proposé ainsi qu'une formalisation des projections.

2.4.1 Stockage centralisé ou décentralisé

Le lieu de stockage de ces données biométriques influe fortement sur l'impact potentiel d'une compromission. S'il s'agit du stockage local sur un téléphone ou une carte d'accès contenant uniquement les données biométriques de son propriétaire, l'impact n'est pas le même que celui de la compromission d'un centre de stockage centralisé de données biométriques. Un tel centre peut contenir toutes les données biométriques des employés d'une entreprise, et dans un cas plus large les données biométriques des citoyens d'un pays.

Un système d'authentification permet que la donnée biométrique soit conservée par l'individu qui la porte, alors qu'un système d'identification implique une base de données biométriques centralisée et stockée par un tiers.

2.4.2 Protection des données stockées

La protection des données stockées est importante, pour minimiser l'impact d'une compromission du stockage. Dans le cadre des mots de passe, ils sont généralement stockés sous une forme transformée par une fonction de hachage. Les fonctions de hachage protègent d'une inversion grâce à différentes propriétés, comme l'utilisation d'un sel pour éviter les tables de hachage, une utilisation importante de temps de processeur et de taille de mémoire nécessaire pour éviter les attaques par force brute. Un mot de passe est confirmé uniquement si celui saisi est *exactement* le même que celui enregistré. Dans le cadre des données biométriques, cette propriété n'existe pas : deux captures successives d'une même modalité biométrique avec le même capteur ne produiront pas les mêmes captures ni les mêmes vecteurs de caractéristiques.

[Ratha et al., 2001b] mettent en avant les implications sur la vie privée de la biométrie et proposent la distorsion volontaire de ces données pour limiter ces impacts. [Jain et al., 2008] décrivent les deux grandes catégories permettant de sécuriser les données biométriques dans la figure 2.4.

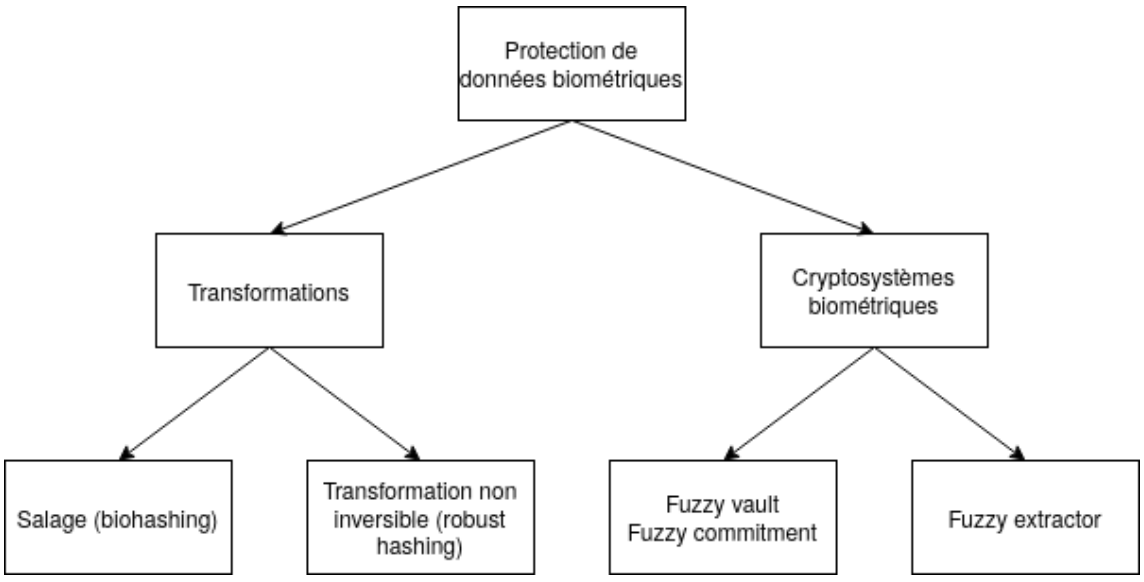


FIGURE 2.4 – Schémas de protection biométrique [Jain et al., 2008]

Ces schémas de protection prennent notamment en entrée un vecteur de caractéristiques x pour fournir en sortie un gabarit u. Gabarit est le terme français utilisé pour *template* par la CNIL [CNIL, a]. Ils doivent respecter plusieurs propriétés :

— Indistinguabilité : soit un vecteur de caractéristiques x protégé avec un schéma une première fois en un gabarit u_1, puis ce même vecteur x protégé une seconde fois en un gabarit u_2. Un attaquant possédant u_1 et u_2 ne doit pas pouvoir déduire qu'ils viennent du même individu. Notamment, un attaquant ne doit pas déterminer que deux gabarits distants sont issus du même individu, mais il ne doit pas non plus déterminer que deux gabarits proches ne sont pas issus du même individu.
— Révocabilité : soit un vecteur de caractéristiques x protégé avec un schéma une première fois en un gabarit u_1. Si u_1 est compromis, on doit pouvoir révoquer u_1, son utilisation ne permet plus une identification, et on doit pouvoir construire un nouveau gabarit u_2 permettant l'identification.
— Non-inversibilité : soit u le gabarit issu du vecteur de caractéristiques x grâce à un schéma de protection. Si un attaquant obtient u, il doit être difficile de construire x à partir de u.
— Performance : l'ajout d'un schéma de protection ne doit pas significativement dégrader les performances, notamment en temps, mais aussi en précision d'identification.

Les cryptosystèmes biométriques, introduits par [Davida et al., 1998], ne sont pas étudiés dans ce manuscrit. [Jain et al., 2008] les divisent en deux catégories : les fuzzy commitment introduits par [Juels and Wattenberg, 1999] et les fuzzy vaults introduits par [Juels and Sudan, 2006]. Nous orientons le lecteur vers les revues [Rathgeb and Uhl, 2011] [Sadhya et al., 2016] pour plus de détails.

Les transformations perturbent ou projettent un vecteur. La projection peut être paramétrée par une graine, ce qui facilite la propriété de révocabilité : on supprime la projection corrompue, et on en calcule une nouvelle avec une autre graine. Les méthodes de projection sont différentes du hachage généralement connu, pour les mots de passe ou le calcul d'empreintes numériques. Dans ce cas général, on souhaite qu'une petite différence dans la donnée à hacher engendre un haché complètement différent. Pour être utilisable dans un cadre d'identification biométrique, il faut que deux vecteurs de caractéristiques proches engendrent deux projections proches afin de pouvoir évaluer la distance qui sépare les vecteurs dans cet espace transformé.

La figure 2.6, telle que décrite dans [Patel et al., 2015], classe en différentes modalités des schémas biométriques révocables . On trouve deux grandes catégories : les schémas nécessitant un comparateur spécifique pour évaluer leurs transforma-

tions, et ceux utilisables avec des comparaisons génériques comme des calculs de distances. On trouve deux sous-catégories, ceux nécessitant une étape d'inscription, et ceux pouvant s'en passer. Enfin, chaque schéma prend en entrée soit directement la capture de modalités biométriques sous forme de signal, soit les caractéristiques extraites comme un vecteur de caractéristiques. Nous fournissons cette figure pour constater la variété de comparateurs disponibles, mais aussi les différences sur le format des données biométriques utilisées et la présence d'une étape d'enregistrement ou non orientant le choix de l'algorithme de biométrie révocable utilisé.

Dans nos travaux, nous nous intéressons à des schémas biométriques révocables utilisables avec un comparateur générique et basés sur des caractéristiques, comme les permutations et le biohashing. En effet, un algorithme de biométrie révocable prenant en entrée un vecteur de caractéristiques nous permet d'utiliser n'importe quelle modalité biométrique tant qu'on a un algorithme d'extraction permettant d'en obtenir un vecteur de caractéristiques. Cette protection est schématisée en 2.5.

Ces transformations ont été introduites par [Ratha et al., 2001b]. Les transformations biométriques révocables sont généralement basées sur le hachage localement sensible [Charikar, 2002, Andoni and Indyk, 2006, Wang et al., 2018]. Celles basées sur une projection aléatoire paramétrée avec un jeton, éventuellement suivie d'une étape de binarisation, ont été introduites par [Teoh et al., 2004, Teoh et al., 2008] avec l'algorithme du biohashing. D'autres projections ont ensuite été proposées [Feng et al., 2010, Pillai et al., 2011, Wang and Plataniotis, 2010].

Nous allons désormais décrire un schéma d'identification biométrique utilisant une protection de ces données.

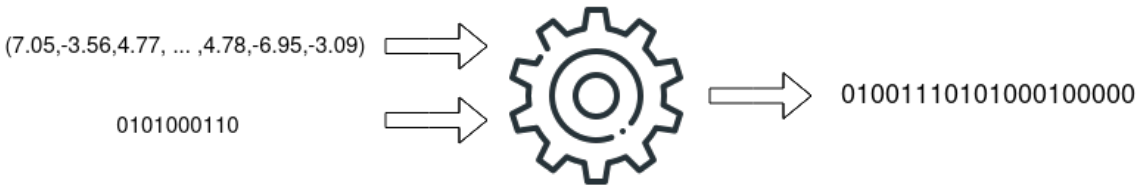


FIGURE 2.5 – Protection avec une graine du vecteur de caractéristiques en un gabarit

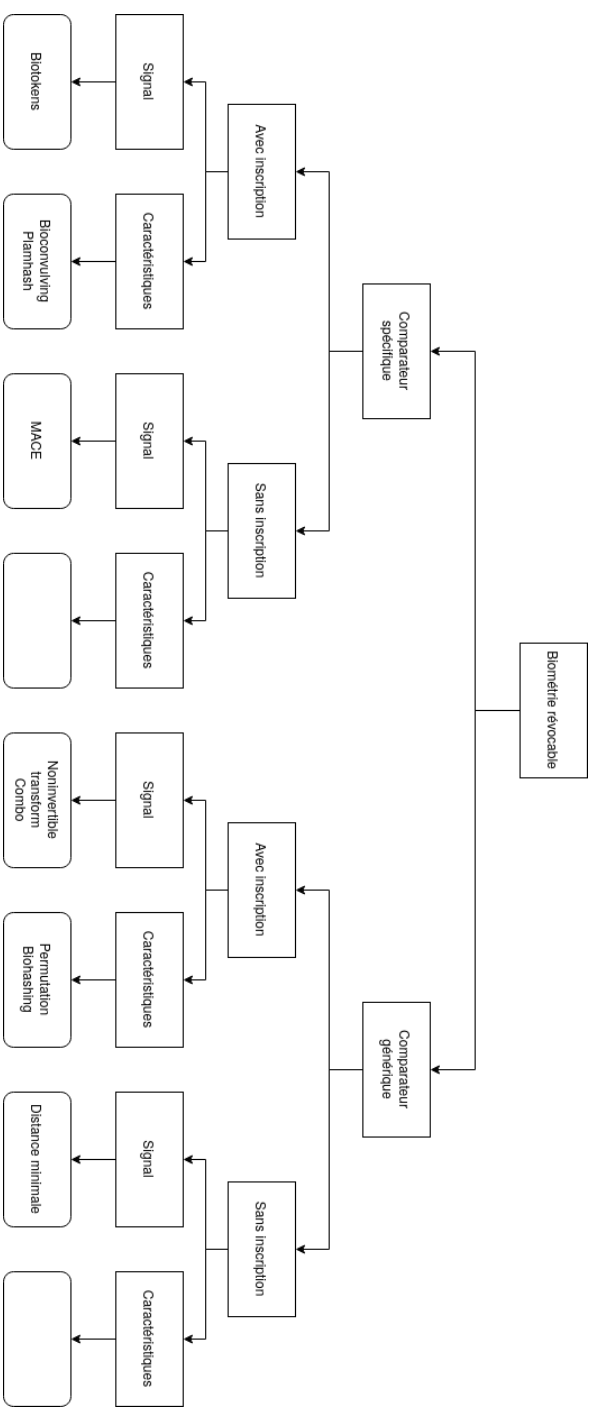


FIGURE 2.6 – Biométrie révocable [Patel et al., 2015]

2.4.3 Schéma d'identification avec transformation

Comme pour la section 2.3.1, nous détaillons les étapes de comparaison et de décision. On commence par définir un schéma d'authentification de biométrie révocable 2.4.1 puis on généralise en définissant un schéma d'identification de biométrie révocable 2.4.2.

Définition 2.4.1. *Soit $\mathcal{K}$ un espace de jetons (graines) et $(\mathcal{M}_B, D_B)$ un espace métrique. Un schéma d'authentification de biométrie révocable est une paire d'algorithmes $(\mathcal{T}, \mathcal{V})$, avec*

- *$\mathcal{T}$ prend un jeton secret $s \in \mathcal{K}$, et un vecteur de caractéristiques $x \in \mathcal{M}_A$ comme entrée, et retourne un gabarit biométrique $u = \mathcal{T}(s, x) \in \mathcal{M}_B$.*
- *$\mathcal{V}$ prend deux gabarits biométriques $u = \mathcal{T}(s, x)$, $u' = \mathcal{T}(s, x')$, et un seuil τ_B comme entrées, et retourne $Vrai$ si $D_B(u, u') < \tau_B$, et $Faux$ sinon.*

Remarque 2. *Il est attendu que la transformation biométrique révocable $\mathcal{T}$ n'affaiblisse pas significativement les performances de la base de données biométriques originale. Ainsi, pour deux gabarits révocables $u, u' \in \mathcal{M}_B = \{0,1\}^M$ issus de la même personne, nous voulons vérifier $D_B(u, u') < \tau$, avec D_B la distance de Hamming telle que définie dans 2.2.5.*

Définition 2.4.2. *Soit $D = \{u_i\}_{i=1,\dots,n}$ une base de données biométriques révocables, composée de gabarits biométriques $u_i = \mathcal{T}(s_i, x_i)$. Un schéma d'identification biométrique révocable pour D est un algorithme qui prend un gabarit biométrique $u = \mathcal{T}(s, x)$ et un seuil τ_B comme entrées, et retourne $Vrai$ s'il existe i entre 1 et n tel que $s = s_i$ et $\mathcal{V}(u, u_i, \tau_B) = Vrai$ et retourne $Faux$ sinon. Dans le premier cas, x est identifié avec succès comme étant l'individu i.*

Remarque 3. *Le schéma d'identification biométrique révocable ne retourne pas la distance entre les deux gabarits, mais retourne seulement $Vrai$ ou $Faux$. En effet, retourner le score procure une information supplémentaire permettant des attaques décrites en 2.9.*

2.4.4 Projection

Nous décrivons ici les projections 2.4.3, un des schémas biométriques révocables. Le biohashing 4 est une projection couramment utilisée qui nous permet de donner des résultats indicatifs. D'autres projections seront utilisées dans la suite de ce manuscrit.

Définition 2.4.3. *Soit M_s une matrice à valeurs réelles avec N lignes et M colonnes, générées pseudoaléatoirement depuis un jeton s.*
x est un vecteur de caractéristiques de taille N. La projection de x par M_s est calculée par xM_s. Cette projection peut être binarisée. Pour cela nous allons utiliser $D : \mathbb{R}^M \to \{0,1\}^M$, définie par $D(t_1, \ldots, t_M)$=$(u_1, \ldots, u_M)$ avec

$$u_i = \begin{cases} 0 & si \quad t_i < 0 \\ 1 & si \quad t_i \geq 0 \end{cases}$$

La transformation biométrique révocable avec binarisation $\mathcal{T}$ est définie par

$$\mathcal{T}(x, s) \;=\; D(xM_s)$$

.

Remarque 4. *Dans le cas de l'algorithme du biohashing, la matrice pseudoaléatoire m est orthonormalisée avec l'algorithme de Gram-Schmidt.*

Remarque 5. *Dans ces travaux, sauf mention contraire, nous utilisons des gabarits de 128 bits.*

La sécurité des schémas de biométrie révocable est généralement étudiée avec le modèle du jeton volé, où les jetons sont considérés comme publics. Un attaquant a connaissance de s, et donc de la matrice M_s. Cette sécurité est globalement faible et non prouvée. Une partie des travaux de la thèse est justement d'utiliser la faiblesse de ces transformations.

La prochaine section décrit les données biométriques que nous utilisons dans cette thèse, avec des indicateurs de performance avec et sans protection.

2.5 Données biométriques utilisées

Dans le cadre de nos expériences, nous validons nos hypothèses à l'aide de bases de données biométriques issues de captures de modalités réelles. Nous présentons dans les sous-sections ci-dessous les différentes bases de captures de modalités utilisées, les modalités capturées, l'algorithme d'extraction de vecteurs de caractéristiques depuis la modalité, et les performances de la base de données biométriques construites. Nous donnons aussi les performances des bases de données biométriques révocables construites depuis ces bases de données biométriques avec comme transformation le biohashing 4 paramétrée par des graines de 128 bits.

Les performances des bases de données biométriques et des bases de données biométriques révocables sont représentées de deux manières.

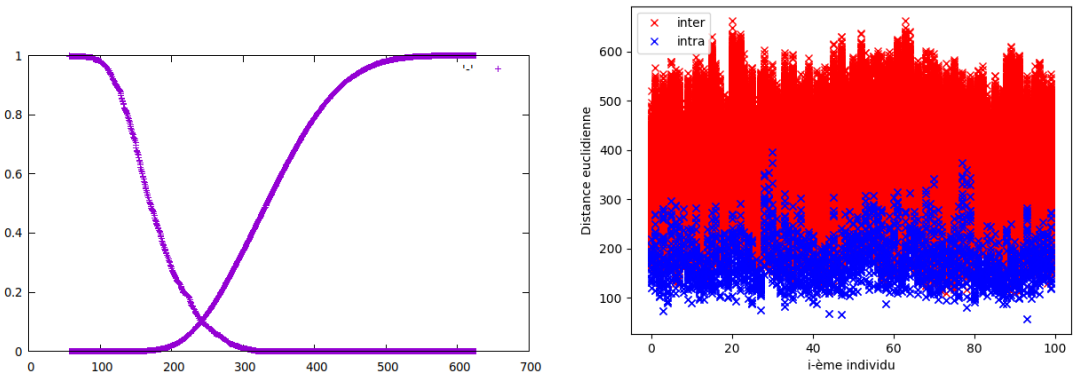
La première manière est sous la forme de courbes *FMR* et *FNMR*. Dans ces figures, l'abscisse correspond au seuil (en termes de distance) et l'ordonnée correspond au taux (en termes de pourcentages de fausses acceptations ou de faux rejets donnés entre 0 et 1). L'EER correspond à l'ordonnée du point de croisement de ces deux courbes, et le taux $\tau@EER$ à son abscisse.

La seconde manière est sous la forme de nuages de points, représentant la distance intraclasse et la distance interclasse. L'abscisse correspond à un individu, l'ordonnée à la distance, euclidienne pour les bases de données biométriques, de hamming pour les bases de données biométriques révocables. Un point bleu correspond à la distance entre deux vecteurs de caractéristiques (ou gabarits) du même individu (en abscisse), représentant la distance intraclasse. Un point rouge correspond à la distance entre un vecteur de caractéristiques (ou gabarit) de l'individu en abscisse et un vecteur de caractéristiques (ou gabarit) d'un autre individu, représentant la distance interclasse.

2.5.1 Empreintes digitales : FVC2002

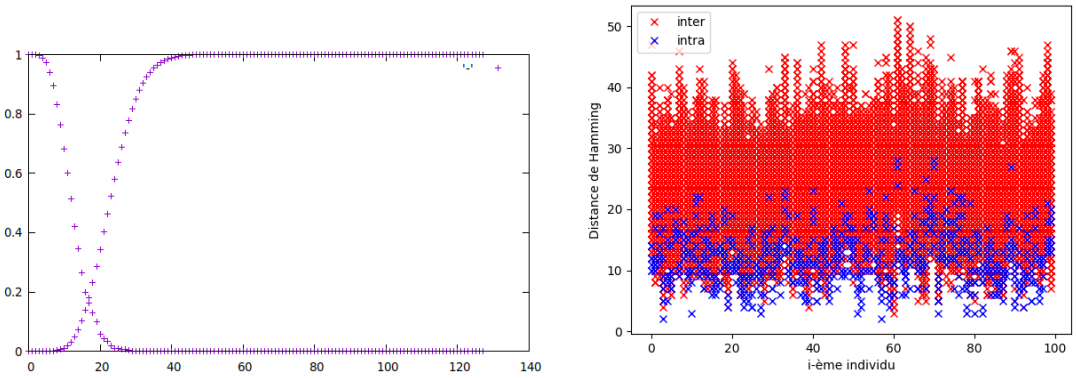
FIGURE 2.7 – Exemple de signal de la base FVC

La base d'empreintes digitales FVC2002 DB2 [Maio et al., 2002] contient $t = 8$ images d'empreintes digitales de $n = 100$ personnes.
Les vecteurs de caractéristiques ont été extraits des images avec des filtres de Gabor [Belguechi et al., 2016]. Chaque vecteur de caractéristiques est composé de $N = 512$ valeurs réelles. L'EER de la base de données biométriques est d'environ 10% avec un seuil $\tau_A = 240.7$.
La base de données biométriques révocables a un EER d'environ 16.5% avec un seuil $\tau_B = 17$.



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.8 – Indicateurs de performance de la base FVC sans transformation



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.9 – Indicateurs de performance de la base FVC avec transformation (biohashing)

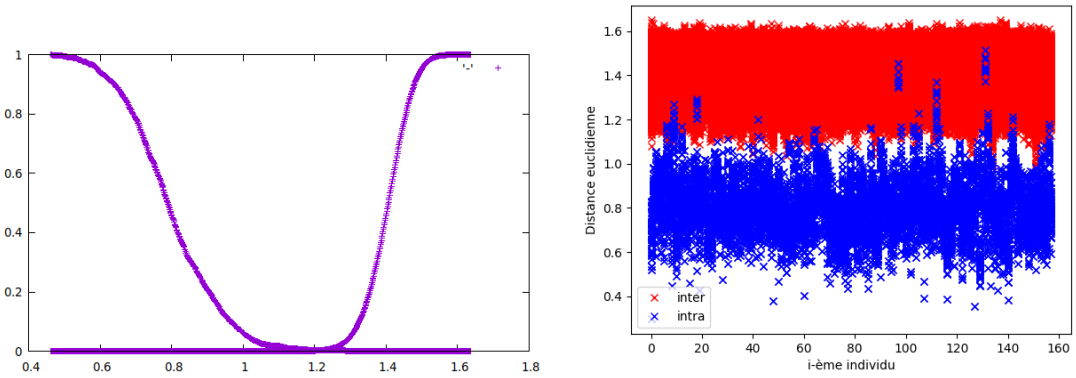
2.5.2 Visages : LFW

La base de visages LFW [Huang et al., 2008] utilisée dans les expériences de [Dong et al., 2019b], contient $t = 10$ vecteurs de caractéristiques de $n = 158$ personnes obtenues depuis leurs photos de visage. La base des données biométriques est issue de [Dong et al., , Dong et al., 2019c] depuis [Jin,].

Dans ce manuscrit, on note aussi cette base complète LFW10. Nous avons extrait une sous-base, noté LFW8, utilisant les 8 premiers vecteurs de caractéristiques des 100 premières personnes pour comparer avec la base FVC2002.

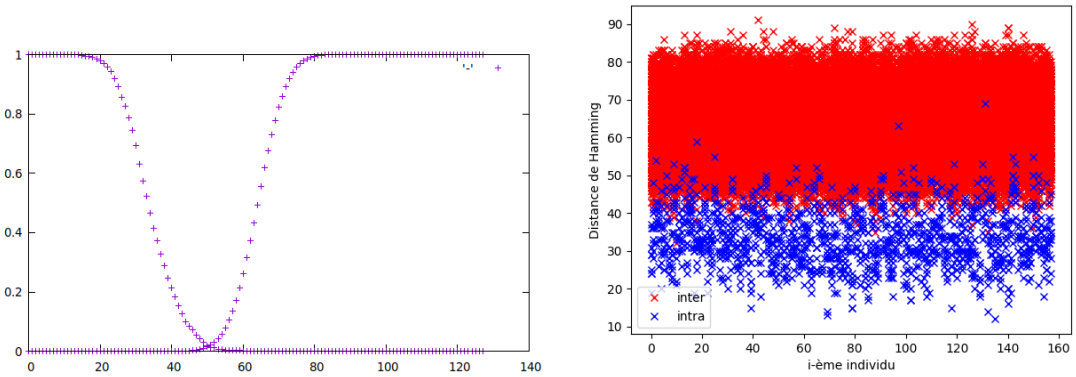
Les vecteurs de caractéristiques ont été obtenus depuis les images de visage grâce au

réseau profond InsightFace [Deng et al., 2019]. Chaque vecteur de caractéristiques est composé de $n = 512$ valeurs réelles. L'EER de cette base de données biométriques est 0.2% avec un seuil $\tau_A = 1.227$.
La base de données biométriques révocables a un EER d'environ 1.9% avec un seuil $\tau_B = 51$.



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.10 – Indicateurs de performance de la base LFW sans transformation



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.11 – Indicateurs de performance de la base LFW avec transformation (biohashing)

2.5.3 Électrocardiogramme : PTB

La base d'ECG PTB [Bousseljot et al., 1995, Goldberger et al., 2000] est composée d'un nombre variable de données biométriques issues de 290 personnes. Nous utilisons les $t = 7$ premières captures des $n = 158$ dernières personnes de la base.

Les vecteurs de caractéristiques sont extraits par délimitation des ondes ECG telle que proposée par [Martínez et al., 2004] en utilisant la librairie python *NeuroKit2* [Makowski et al., 2021]. Chaque vecteur de caractéristiques est composé de $n = 990$ valeurs réelles.
Un battement de cœur est capturé par plusieurs électrodes appliquées sur la peau. Ces différentes électrodes nous permettent d'enregistrer plusieurs signaux d'un même battement, dans notre cas il y a 15 signaux enregistrés en parallèle. On découpe pour chaque signal chaque battement en détectant le début, le pic, et la fin de chaque vague : P, Q, R, S, et T. Ce découpage est visualisé dans la figure 2.12. Les caractéristiques sont construites avec les durées séparant les phases des vagues. On obtient 66 intervalles par signaux, soit 990 intervalles pour les 15 signaux réunis pour chaque battement.

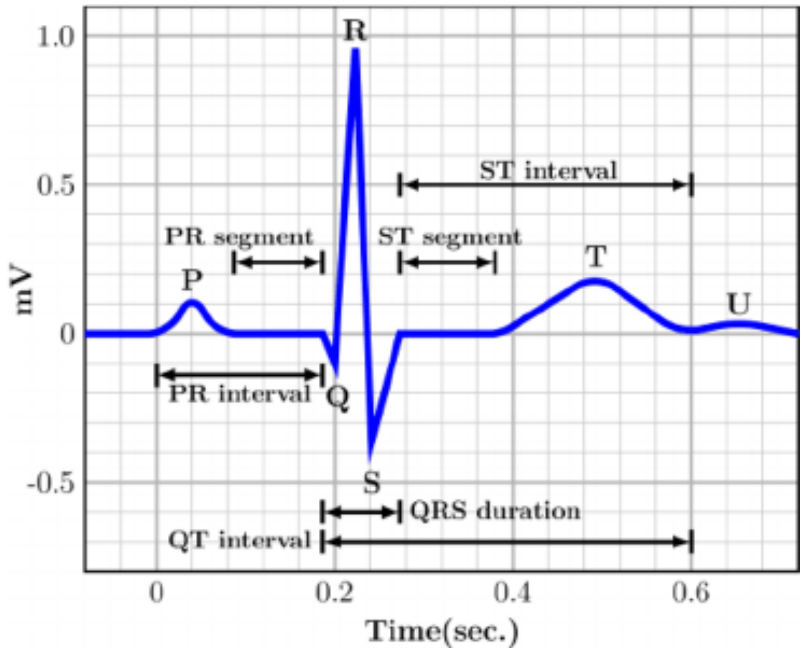
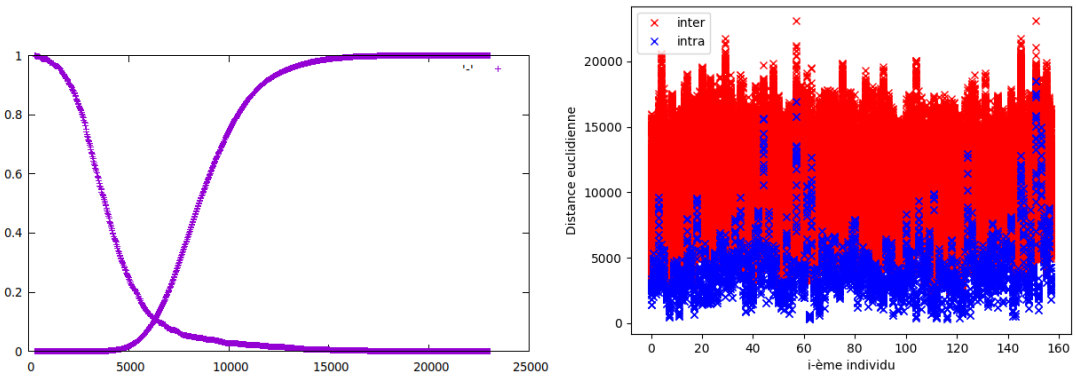


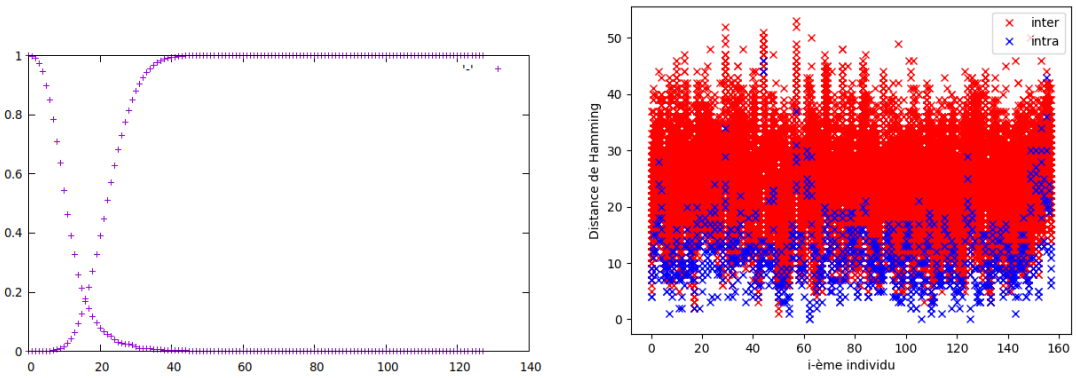
FIGURE 2.12 – Ensemble PQRST d'un signal ECG [Chen et al., 2014]

L'EER de la base de données biométriques est environ de 10.8% avec un seuil $\tau_A = 6321$ (un EER similaire est proposé dans [Pinto and Cardoso, 2019] pour cette même base).
La base de données biométriques révocables a un EER d'environ 17% avec un seuil $\tau_B = 16$.



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.13 – Indicateurs de performance de la base PTB sans transformation



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 2.14 – Indicateurs de performance de la base PTB avec transformation (biohashing)

2.5.4 Interprétations

Nous constatons que la base LFW a un meilleur EER que les bases FVC et PTB. Pour les trois bases, on constate une sensible dégradation de l'EER suite à la transformation avec l'algorithme du biohashing (avec binarisation).

Les nuages de points représentant les distances interclasses et intraclasses montrent bien deux ensembles globalement séparés avec la base LFW. Ces deux ensembles se superposent légèrement dans le cas des deux autres bases, FVC et PTB, moins performantes.

2.6 Individus à fort potentiel d'usurpation

Certains vecteurs de caractéristiques d'une base de données biométriques influent fortement sur ces performances (FMR/FNMR/EER). [Doddington et al., 1998] trient les individus en 4 catégories, les moutons, les chèvres, les agneaux et les loups, en fonction de leur performance au sein de la base de données biométriques. Dans cette section, après avoir défini ces différents types d'individus, nous étudions les trois bases de données biométriques précédentes avec cette classification.

2.6.1 Classification

Les loups sont naturellement capables d'usurper plusieurs individus, causant de fausses acceptations. Ces individus ont des caractéristiques génériques.

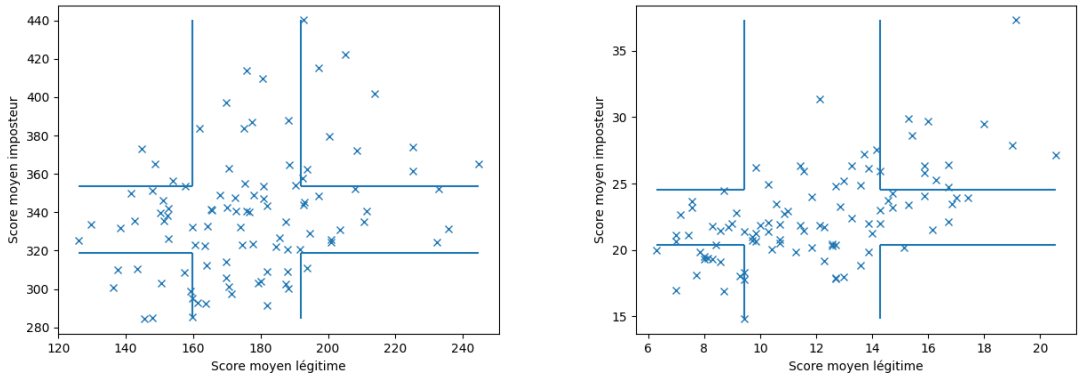
[Yager and Dunstone, 2010] ont trié les individus en 4 catégories d'animaux, en se basant sur la relation entre le taux de fausses acceptations et le taux de faux rejets : les vers, les caméléons, les fantômes et les colombes. Les caméléons profitent d'un important taux de fausses acceptations, et d'un faible taux de faux rejets. Ils sont définis comme obtenant des scores appartenant au premier quart des meilleurs scores. Ces travaux démontrent l'existence de différentes catégories d'individus dans une base, dont certains possèdent des propriétés impactant les performances de la base.

[Une et al., 2007] définissent un loup comme un vecteur ayant une probabilité de fausses acceptations supérieure au FMR de la base. Ils évaluent la sécurité d'une base par la notion de probabilité d'attaque par des loups (WAP). [Inuma et al., 2009] ont mis en avant des contre-mesures, améliorées et testées sur les empreintes par [Murakami et al., 2012].

2.6.2 Dans les bases utilisées

Nous avons analysé les bases de données biométriques utilisées décrites dans la section 2.5. Pour cela, nous nous sommes intéressés, pour chaque individu, au score moyen obtenu en authentification légitime et au score moyen obtenu en étant imposteur. Les résultats sont restitués dans les figures 2.15, 2.16 et 2.17. Comme dans [Yager and Dunstone, 2010], un individu est représenté par un point dont l'abscisse est la distance moyenne obtenue dans une authentification légitime, et l'ordonnée est la distance moyenne obtenue dans une authentification en tant qu'imposteur. Il s'agit de distances euclidiennes avec vecteurs de caractéristiques non transformés, et de distances de Hamming avec des gabarits issus du biohashing. Pour aider à l'inter-

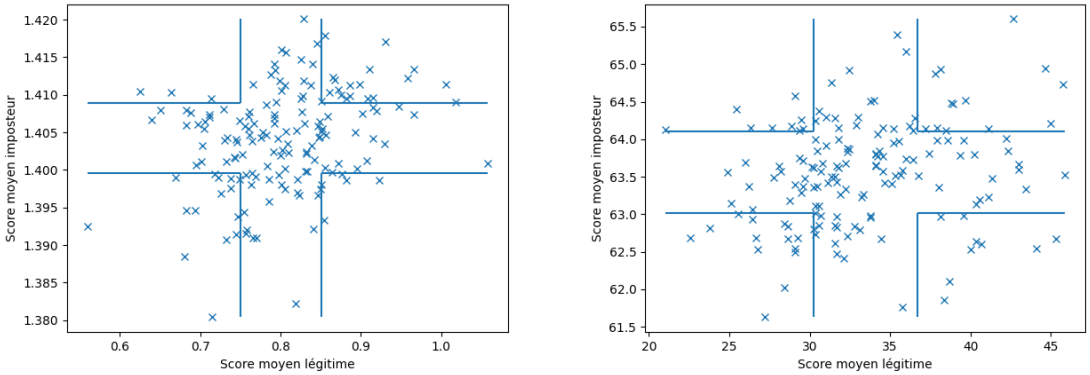
prétation, des lignes verticales et horizontales représentent les premiers et troisièmes quartiles. Les $\tau@EER$ sont fournis pour interpréter les scores de chaque point. Ainsi les loups sont ceux ayant un score imposteur bas, c'est-à-dire en bas des figures. Les caméléons, alliant un score imposteur bas, et donc un haut taux de fausses acceptations, et un score légitime bas, et donc un faible taux de faux rejets, se trouvent en bas à gauche des figures.



(a) Sans transformation (b) Biohasing

FIGURE 2.15 – Ménagerie FVC

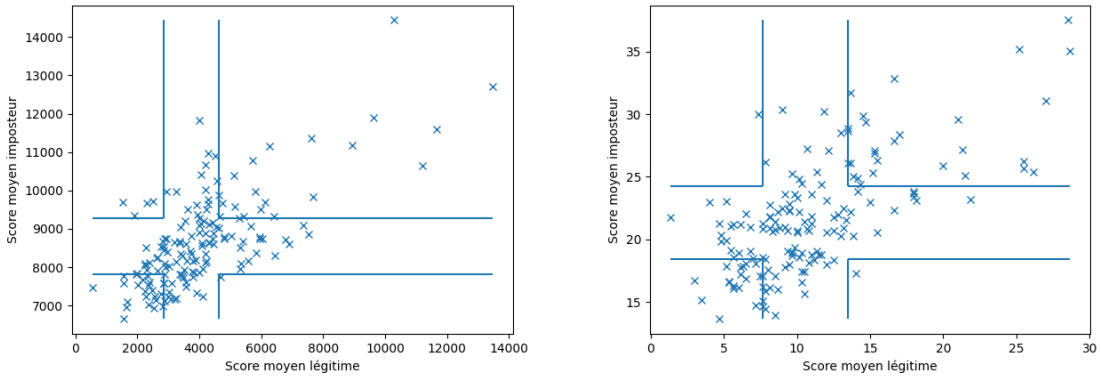
Pour la base FVC, on a sans transformation $\tau@EER = 240$ et avec biohashing $\tau@EER = 17$. On constate que sans transformation, quelques loups ont une moyenne de distance imposteur à environ 285, tout en ayant une excellente moyenne de distance légitime. Dans le cas du biohashing, on a un loup ayant une moyenne de distance imposteur à 15, sous le taux à l'EER.



(a) Sans transformation (b) Biohasing

FIGURE 2.16 – Ménagerie LFW

Pour la base LFW, on a sans transformation $\tau@EER = 1.22$ et avec biohashing $\tau@EER = 51$. On remarque que les loups ont une distance imposteur moyenne éloignée du taux.



(a) Sans transformation (b) Biohasing

FIGURE 2.17 – Ménagerie PTB

Pour la base PTB, on a sans transformation $\tau@EER = 6321$ et avec biohashing $\tau@EER = 16$. Les distances imposteurs moyennes des loups sont proches du taux, et même en dessous dans le cas du biohashing. Pour cette base, on remarque que les loups qui ont une distance imposteur moyenne basse ont aussi une distance légitime basse. Cela a pour conséquence que les loups sont souvent aussi des caméléons.

Dans la table 2.2, on trouve le taux d'usurpation du loup qui usurpe un maximum d'individus, avec et sans transformation.

Base	Sans transformation	Biohashing
FVC	40%	50%
LFW	3%	11%
PTB	55%	51%

TABLE 2.2 – Taux d'usurpation maximum

On constate que les bases FVC et PTB ont au moins un loup avec une importante capacité d'usurpation, environ la moitié des individus.

2.7 Algorithme génétique

Dans cette section, nous introduisons les algorithmes génétiques, qui sont des algorithmes d'optimisation. Ils sont utilisés dans les travaux de cette thèse. Ils permettent de minimiser la valeur d'une fonction d'évaluation f en construisant son paramètre. Leur fonctionnement est inspiré de la reproduction naturelle. Nous avons une population initiale de taille n, qui évolue sur t générations, convergeant vers des solutions minimisant la valeur de la fonction f. Il s'agit d'une métaheuristique évolutionniste.
Nous utilisons les notations de [Mitchell, 1998] qui est un ouvrage de référence pour les algorithmes génétiques. Nous détaillons ensuite la structure globale d'un algorithme génétique et les différents paramétrages possibles.

Une population Ξ est composée de n individus Φ : $\Xi = \{\Phi_i\}_{i=1,\dots,n}$.
Chaque individu Φ_i est représenté par un chromosome Ψ_i, instancié sous la forme d'un vecteur de taille m, $\Psi_i = (\psi_i^1, \dots, \psi_i^m)$.
Un indice j d'un vecteur Ψ_i code un gène, en le fixant à un allèle particulier, la valeur ψ_i^j.

À chaque itération, des individus sont sélectionnés pour persister dans la génération suivante, complétée par leurs enfants. L'individu issu d'une reproduction de ses deux parents possède un mélange des allèles de ces deux derniers obtenu par croisement, perturbé par des mutations aléatoires. Les individus qui ne se sont pas sélectionnés comme parents disparaissent. La génération suivante est composée des enfants issus d'une reproduction ainsi que de leur parent.

Ainsi, un algorithme génétique comporte plusieurs étapes :

1. La génération d'un ensemble d'individus formant la population de l'itération initiale 0, noté Ξ_0. Pour cela, on génère aléatoirement n vecteurs.
2. On effectue les t itérations, $k = 1, \dots, t$:
   a) On part de la population de l'itération k noté Ξ_k.
   b) La sélection est une étape qui élit les $\frac{n}{2}$ individus de notre population Ξ_k qui vont se reproduire et persister dans la population de l'itération suivante Ξ_{k+1}. Les autres disparaitront. On note ξ_{k+1} les $\frac{n}{2}$ chromosomes sélectionnés comme parents reproducteurs.
   c) Le croisement est l'étape qui va mélanger les allèles de deux parents pour former deux chromosomes enfants. À partir des $\frac{n}{2}$ parents, on obtient $\frac{n}{2}$ enfants, notés Ω_{k+1}

d) La mutation applique des changements aléatoires dans les chromosomes des enfants nouvellement créés $\omega \in \Omega_{k+1}$, en changeant des allèles.

e) L'itération $n+1$ est formée des parents sélectionnés pour la reproduction, et des enfants issus de ces reproductions par croisement puis mutés : $\Xi_{k+1} = \xi_{k+1} \cup \Omega_{k+1}$ avec $\xi_{k+1} \subset \Xi_k$.

3. On récupère finalement le chromosome $\Psi^* \in \Xi_t$ qui minimise f.

Remarque 6. *Il est à noter que la taille de la population est constante au fur et à mesure des itérations.*

Remarque 7. *On peut remplacer le nombre d'itérations à effectuer par une condition à remplir, par exemple avoir un chromosome Ψ^* tel que $f(\Psi^*) < \epsilon$, avec ϵ notre objectif de score pour s'arrêter.*

Tous les paramètres et toutes les fonctions utilisés dans un algorithme génétique peuvent influer sur sa performance. L'idéal est de converger rapidement vers une solution optimale globale, en peu de temps et en peu d'espace.

Dans un premier temps, la taille de population n et le nombre d'itérations t semblent être deux paramètres influant sur la taille utilisée et le temps nécessaire à l'exécution de l'algorithme.

L'étape de sélection, la manière avec laquelle on sélectionne les parents reproducteurs ξ, influe sur la vitesse de convergence du résultat, mais une précipitation peut nous emmener sur une solution optimale locale.

La mutation, en rajoutant de l'aléatoire, peut participer à cette recherche globale.

Remarque 8. *[Mitchell, 1998] indique que l'idéologie des algorithmes génétiques implique pour fonctionner que des parents ayant de bons scores avec la fonction d'évaluation produisent des enfants ayant eux aussi de bons scores. On assume notamment une corrélation entre le score de deux chromosomes voisins. L'idéal étant que la fonction d'évaluation soit convexe.*

Nous avons vu que les algorithmes génétiques sont des algorithmes d'optimisation inspirés de la reproduction naturelle. Nous allons désormais détailler une autre méthode d'optimisation.

2.8 Méthode d'escalade

La méthode d'escalade permet de produire des algorithmes d'optimisation, comme pour les algorithmes génétiques. Ils sont utilisés à titre comparatif dans nos travaux. L'objectif est de minimiser le score d'une fonction d'évaluation f en construisant son paramètre x. Le terme anglais est *hill climbing*. Un algorithme par escalade a typiquement le déroulé suivant :

1. L'algorithme commence avec un paramètre x initial. Dans nos travaux, x est un vecteur de caractéristiques. Ce vecteur peut être généré aléatoirement ou choisi parmi des vecteurs existants.
2. Une fonction G prenant en entrée un paramètre x produit des voisins. Dans notre cas, chaque voisin est un clone du vecteur x dont on a modifié la valeur d'un seul indice. On peut avoir un ou plusieurs voisins par indice.
3. On évalue chaque voisin construit avec la fonction d'évaluation f. On conserve le voisin minimisant cette fonction, noté v^*. S'il obtient un meilleur score que le paramètre x depuis lequel il a été construit, $f(v^*) < f(x)$, alors on recommence à l'étape 2 avec v^* comme paramètre x. Sinon on va à l'étape 4.
4. Nous n'avons pas réussi à obtenir un voisin ayant un meilleur score que x. L'algorithme retourne comme solution x. C'est une solution optimale globale si le problème, la fonction d'évaluation f, est convexe, sinon x est au minimum une solution optimale locale.

Un tel algorithme par escalade est un algorithme itératif et local. La simplicité de mise en place en fait un algorithme largement utilisé en recherche. Des attaques utilisant une méthode d'escalade sont présentées dans la section suivante.

2.9 Attaques existantes sur les transformations

Dans cette section, nous allons décrire des attaques existantes, leurs spécificités, et les différences avec nos travaux. Nous commençons par présenter des travaux théoriques concernant la fuite d'informations avec les transformations préservant les distances. Nous présentons ensuite des attaques impactant la notion de non-inversibilité décrite en 2.4.2, notamment en construisant des préimages proches. Dans un premier temps nous décrivons les attaques n'utilisant pas d'algorithme génétique. Dans un second temps, nous décrivons séparément les attaques utilisant des algorithmes génétiques, en 2.9.3, de par leur proximité avec les travaux présentés dans ce manuscrit.

2.9.1 Fuites d'informations depuis les gabarits

[Liu et al., 2006] décrivent les transformations préservant les distances, équivalentes à des transformations orthogonales suivies d'une translation. Ils formalisent la probabilité de fuites d'informations et introduisent une attaque par analyse de composantes principales.

[Kaplan et al., 2017] poursuivent ce travail sur les transformations préservant les distances en les généralisant à des transformations préservant les relations de distances. Une transformation préserve les distances si $D(x, y) = D(T(x), T(y))$ et préserve les relations de distances si

$$D(a, b) > D(c, d) \rightarrow D(T(a), T(b)) > D(T(c), T(d))$$

Ils décrivent une attaque permettant de récupérer des informations fuitées depuis ces transformations préservant les relations de distance.

Ces travaux sont théoriques et concernent les transformations préservant les distances, ce qui n'est pas le cadre de nos transformations dont l'objectif est de limiter la perte de performance. Ils sont néanmoins intéressants et démontrent les attaques possibles grâce à de telles propriétés.

2.9.2 Différentes méthodes d'attaques

Nous présentons des attaques n'utilisant pas d'algorithme génétique.

[Adler, 2005] utilise une attaque par escalade pour inverser le cryptosystème biométrique de [Soutar et al., 1998]. Il utilise le score de similarité qu'il exploite comme une fuite d'informations. Cette attaque permet d'estimer la capture de modalité biométrique à l'origine du gabarit. Pour empêcher cette attaque, il ne faut pas retourner le score de similarité, mais uniquement la décision.
Différence avec nos travaux : cette attaque concerne les cryptosystèmes biométriques, et non les transformations révocables.

[Nagar et al., 2010] analysent la sécurité de schémas de protection. Un calcul de préimage est possible depuis le gabarit issu du biohashing. Une modification du biohashing limitant cette attaque en dégradant peu les performances est proposée.
Différence avec nos travaux : l'attaque proposée est une attaque linéaire. Elle est donc spécifique au biohashing et non générique.

[Feng et al., 2014] utilisent l'algorithme perceptron comme classifieur puis une attaque par escalade. Le perceptron permet de construire un vecteur de caractéristiques depuis un gabarit binaire. L'attaque par escalade permet de construire une capture de modalité biométrique depuis un vecteur de caractéristiques. Cette attaque nécessite la connaissance de l'algorithme de transformation. Dans un second scénario, ils utilisent un perceptron multicouche pour modéliser la transformation, puis une attaque par escalade pour construire un vecteur de caractéristiques et la capture correspondante.
Différence avec nos travaux : cette attaque permet de calculer une préimage proche et non d'usurper plusieurs individus.

[Roy et al., 2017] introduisent la construction de caractéristiques d'empreintes digitales partielles usurpant de nombreux individus multienrôlés. Ces travaux ont été étendus par [Roy et al., 2018] à la construction de caractéristiques de plusieurs empreintes digitales partielles permettant d'usurper de nombreuses cibles.
La sélection des captures d'empreintes digitales partielles à la phase d'enrôlement comme contre-mesure à leurs attaques a été étudiée par [Roy et al., 2019].
Enfin, la construction d'images d'empreintes au lieu des caractéristiques a été étudiée par [Bontrager et al., 2018] pour rendre possible cette attaque par présentation.
Différence avec nos travaux : cette attaque est basée sur les empreintes digitales partielles uniquement et n'utilise pas de méthodes de protections de données biométriques.

[Gomez-Barrero and Galbally, 2020] est une étude décrivant les différentes méthodes permettant d'inverser les transformations biométriques dites non inversibles. L'objectif est de reconstruire l'image d'une empreinte digitale depuis un gabarit, puis de reconstruire physiquement cette empreinte afin d'effectuer une attaque par présentation (le point 1 décrit en 2.1.3). Les attaques décrites sont séparées en quatre catégories. Chacune de ces catégories nécessite certaines connaissances pour exécuter l'attaque. Nous décrivons ces 4 catégories, nécessitant un niveau de connaissance croissant.
La première catégorie nécessite de connaître le format d'un gabarit : cela permet de construire un gabarit valide. Le format d'un gabarit peut être connu notamment par une norme utilisée par le système. L'attaquant reconstruit une image d'empreinte digitale depuis un gabarit basé sur les minuties (décrites en 2.2.2).
La deuxième catégorie a comme prérequis de connaître le score de similarité, la distance dans notre cas. L'attaquant soumet une capture de modalité biométrique et le

gabarit qu'il souhaite inverser, et obtient en retour le score. Des techniques par escalade initialisées avec des captures synthétiques sont utilisées, et le score retourné, utilisé comme résultat de la fonction d'évaluation, permet d'affiner la reconstruction. Cette attaque est appliquée à la reconstruction d'iris et de visages.
La troisième catégorie a besoin de connaître le fonctionnement de la fonction d'évaluation, en plus du score retourné. Cette connaissance supplémentaire permet d'expliquer le score et d'orienter la reconstruction par escalade sur des points saillants.
La quatrième catégorie utilise en plus la fonction d'extraction de caractéristiques, permettant d'effectuer de la rétro-ingénierie. Le gabarit stocké peut ainsi être inversé si l'attaquant construit une fonction inverse approchant l'entrée de la fonction d'extraction. Des travaux ont permis l'inversion de gabarit issu d'iris avec des filtres de Gabor.
Différence avec nos travaux : ces travaux permettent de construire une attaque par présentation pour usurper un gabarit. Nos travaux nécessitent de connaître uniquement le format du gabarit et le score de similarité pour construire un passe-partout usurpant de nombreux gabarits.

[Nanwate and Sadhya, 2020] décrivent une attaque avec un algorithme d'optimisation par essaims particulaires (PSO), introduite par [Kennedy and Eberhart, 1995]. Cette attaque nécessite de connaître le score de similarité retourné par la fonction d'évaluation. Un PSO simule, comme pour l'algorithme génétique, une population composée d'individus, initialement aléatoires. À chaque itération, chaque individu est modifié selon le score de ses voisins directs, comme pour l'attaque par escalade, mais aussi selon la position du meilleur voisin global, et donc potentiellement un voisin d'un autre individu de la population. L'attaque prend en entrée le gabarit à inverser, la fonction de transformation, et la fonction d'évaluation. Leur attaque est exécutée avec le biohashing sur la base LFW, et permet d'obtenir des préimages proches dans plus du 30% des cas.
Différence avec nos travaux : ces travaux utilisent un algorithme d'optimisation pour construire des préimages proches. Nous construisons des préimages proches et réutilisables, puis des passe-partout.

2.9.3 Attaques avec algorithmes génétiques

Nous décrivons désormais les attaques utilisant des algorithmes génétiques, comme dans nos travaux.

[Lacharme et al., 2013] se placent dans le cadre de l'authentification. Ils génèrent un vecteur de caractéristiques x^* depuis un gabarit u obtenu à partir d'une graine publique s et du vecteur original non connu x. On note $u = T(s, x)$. Le vecteur x^* est appelé préimage de u et est espéré proche du vecteur original x. L'algorithme génétique utilisé évolue sur 2000 itérations avec une population de taille 10000. Les vecteurs de caractéristiques sont issus de la base FVC2002, et deux méthodes d'extraction de caractéristiques ont été utilisées. Le score de la fonction d'évaluation cesse de s'améliorer entre 400 et 800 itérations, rendant plus de la moitié des itérations inutiles pour l'amélioration du score.
Dans une seconde partie, une fois la préimage x^* construite depuis u et s et proche de x, ils construisent deux nouveaux gabarits avec une nouvelle graine s', un avec le vecteur original et un avec la préimage $u' = T(s', x)$ $u^* = T(s', x^*)$. Cette seconde étape permet d'évaluer la réutilisation de la préimage avec d'autres graines. Ils comparent le score qui est proche, mais moins performant qu'avec la graine originale.
Différence avec nos travaux : nous n'utilisons pas $u = T(s, x)$ et $u' = T(s', x)$, mais $u = T(s, x)$ et $u' = T(s', x')$. Notre préimage est réutilisable pour un gabarit issu d'une autre capture de modalité biométrique.

[Dong et al., 2019a] décrivent l'impact de l'usage d'une fonction de hachage localement sensible (LSH) utilisée pour la biométrie révocable. Une telle fonction conserve globalement les distances entre deux entrées et leurs sorties. Cela permet de lancer une attaque de reconstruction avec un algorithme génétique. Il utilise comme fonction d'évaluation la distance de Hamming entre le gabarit stocké (à inverser) et le gabarit produit depuis un vecteur de caractéristiques construit par l'algorithme génétique. L'objectif est de minimiser cette distance. L'espace des vecteurs de caractéristiques est borné et la population initiale de l'algorithme génétique est générée aléatoirement. La méthode de sélection utilisée est la sélection par rang et l'algorithme génétique s'arrête lorsqu'un vecteur de caractéristiques permettant de générer un gabarit valide, dont la distance avec le gabarit stocké est inférieure à un seuil donné, est trouvé. Il y est montré que plus le gabarit a une grande taille, plus il fait fuiter des informations, et plus la solution construite est précise. Cependant, un gabarit de petite taille rend possibles des attaques par force brute. Il faut donc utiliser une taille de gabarit intermédiaire. Les expériences ont notamment été effec-

tuées avec le biohashing pour la base LFW, et permettent d'obtenir une préimage proche pour 80% des gabarits de la base avec une taille de gabarit de 500 bits, et 5% pour des gabarits de tailles 16 bits. Ils utilisent aussi plusieurs gabarits, jusqu'à 3, pour calculer des préimages. Ces expériences sont effectuées avec une autre transformation de biométrie révocable non utilisée dans ce manuscrit, le filtre de Bloom, avec une base issue d'iris. Ils constatent une augmentation importante concernant la précision des préimages construites.
Différence avec nos travaux : ils construisent des préimages proches uniquement, alors que nous construisons des préimages proches et réutilisables, mais aussi des passe-partout. Nous utilisons une taille de gabarit permettant de minimiser l'EER sans être trop importante. Nos gabarits sont de taille 128 bits, alors que leurs travaux vont de 16 bits, avec une attaque à faible efficacité, jusqu'à 500 bits, avec une attaque très efficace. Ils utilisent plusieurs gabarits pour affiner la construction d'une préimage proche quand nous utilisons plusieurs gabarits pour permettre à un individu passe-partout d'usurper une partie importante des utilisateurs, et ce même avec de futures captures de modalité biométrique.

2.10 Conclusion

Dans ce premier chapitre, nous avons introduit ce qu'est la biométrie. Tout d'abord, nous avons décrit les grandes catégories de modalités biométriques existantes, leurs utilisations, et leurs dangers, notamment sur la vie privée. Nous avons ensuite défini les captures de modalités biométriques, l'extraction de caractéristiques avec la formation de bases de données biométriques, et la comparaison de telles données. Nous avons décrit la structure classique d'un système biométrique, ainsi que le schéma générique d'identification. La différence essentielle entre identification et authentification a été posée. Nous avons introduit la sécurité des données biométriques, notamment les opportunités d'attaques en fonction de la centralisation ou non de leur stockage, mais aussi les protections des données biométriques stockées, pour les sécuriser et limiter l'impact en cas de vols. Pour cela, nous avons formalisé les projections, et l'impact sur le schéma d'identification. Nous avons précisément décrit les bases de captures de données biométriques utilisées, avec des indicateurs de tailles, de performances, et la méthode d'extraction et donné des indicateurs de qualité . La notion de classification d'individus en animal dans une base de données biométriques, et notamment les loups, a été introduite. Nous donnons des indicateurs de présence et de performance de tels loups dans les bases biométriques que nous utilisons.

Nous avons constaté que pour les bases FVC et PTB, des loups usurpant environ 50% des individus sont présents. Deux algorithmes d'optimisation ont été introduits, avec objectifs et caractéristiques de ces derniers : les algorithmes génétiques et la méthode par escalade. Enfin, nous avons présenté les attaques existantes dans le cadre de construction de préimages, avec différents algorithmes de protection de données biométriques, en détaillant les attaques utilisant des algorithmes génétiques, ainsi que leurs différences avec les contributions de cette thèse.

Chapitre 3

Préimage proche et réutilisable

Résumé : *Ce chapitre décrit ce qu'est une préimage proche et réutilisable. Il explique les performances de recherche de préimages et le fonctionnement de différents algorithmes pour cet objectif : algorithme génétique, recherche aléatoire, recherche par choix, méthode d'escalade. Il compare les performances de différents paramétrages de l'algorithme génétique, en comparaison avec les autres modes de recherche, afin d'optimiser les résultats des expériences. Enfin, des variantes de préimage sont proposées, aboutissant au concept passe-partout qui y est décrit et expérimenté.*

Mots-clés : *préimage, attaque par escalade, algorithme génétique, passe-partout.*

3.1 Introduction

Ce chapitre décrit notre première contribution. Nous présentons la construction de préimages proches puis réutilisables telles que définies en 3.2.1 et 3.2.2, permettant d'usurper deux gabarits d'un individu.

Ces préimages sont construites avec un algorithme génétique, pour lequel nous avons effectué de nombreuses expériences afin d'optimiser son paramétrage. Ces travaux sont décrits dans la section 3.4.

Pour positionner les performances de ces travaux, nous les comparons avec d'autres méthodes de construction décrites en 3.3. Nous y décrivons le choix de préimage parmi les vecteurs existants 3.3.1, la recherche aléatoire de préimage 3.2, et finalement la construction de préimage avec une méthode par escalade 3.3.3.

Nous introduisons des variantes de choix de gabarits pour lesquels nous cherchons une préimage en section 3.5. Ces variantes nous ont permis de nous orienter vers de nouveaux travaux décrits en 3.6. Nous y calculons des préimages universelles, que l'on nomme passe-partout. Un passe-partout est construit avec un algorithme génétique dans l'objectif d'être une préimage proche de chaque gabarit d'une base de données biométriques révocables.

3.2 Préimage ...

Nous souhaitons construire des préimages respectant certaines propriétés. Tout d'abord, nous définissons ce qu'est une préimage proche usurpant un gabarit en 3.2.1, avant d'étendre ce concept à une préimage proche et réutilisable usurpant un second gabarit défini en 3.2.2.

Ces constructions de préimages sont possibles grâce aux attaques basées sur la similarité décrites en 2.9 exploitant la propriété de non-inversibilité décrite en 2.4.2.

Pour rappel, notre attaque utilise un modèle où la graine qui paramètre la transformation est publique.

3.2.1 ... proche ...

Le concept de préimage proche permettant de s'authentifier avec un gabarit et une graine est formalisé en 3.2.1 et schématisé en 3.1.

Définition 3.2.1. *Soit $x \in \mathcal{M}_A$ un vecteur de caractéristiques, $u = \mathcal{T}(s, x) \in \mathcal{M}_B$ le gabarit obtenu avec la graine $s \in \mathcal{K}$ et τ un seuil. Une préimage proche de u avec s est un vecteur de caractéristiques x^* tel que $V(u, \mathcal{T}(s, x^*), \tau) = Vrai$.*

Remarque 9. *Il est à noter que nous travaillons avec des préimages dans le cadre d'un gabarit issu d'une transformation. Ces travaux ne s'appliquent pas hors transformation d'un vecteur paramétrée par une graine.*

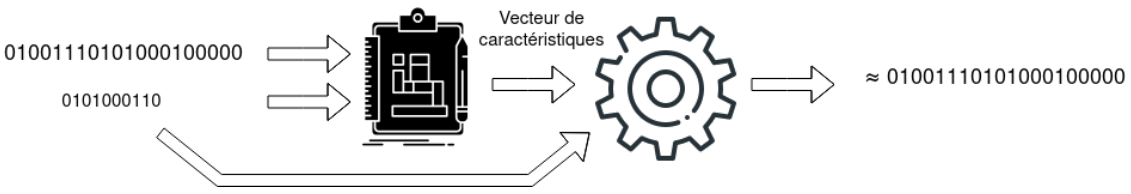


FIGURE 3.1 – Préimage proche d'un gabarit

3.2.2 ... et réutilisable

Le concept de préimage proche étant posé, nous l'étendons en lui ajoutant une propriété de réutilisabilité. Une telle préimage proche et réutilisable, notée PPR, permet à nouveau de s'authentifier avec un gabarit, mais aussi de s'authentifier avec un autre gabarit issu de la même modalité. Cela est formalisé dans la définition 3.2.2 et schématisé en 3.2.

Définition 3.2.2. *Soit x_1, x_2 deux vecteurs de caractéristiques issues de deux captures différentes d'une même modalité d'une même personne. Soit $u_1 = \mathcal{T}(s_1, x_1) \in \mathcal{M}_B$ et $u_2 = \mathcal{T}(s_2, x_2) \in \mathcal{M}_B$ les gabarits obtenus avec les graines $s_1, s_2 \in \mathcal{K}$ et τ un seuil. Une préimage proche de u_1 avec s_1 est un vecteur de caractéristiques x^* tel que $V(u_1, \mathcal{T}(s_1, x^*), \tau) = Vrai$ et est réutilisable si $V(u_2, \mathcal{T}(s_2, x^*), \tau) = Vrai$*

Remarque 10. *Cette définition de préimage proche et réutilisable est différente de celle utilisée par [Lacharme et al., 2013] dans son attaque décrite en 2.9. En effet, leurs travaux décrivent une préimage proche et réutilisable si elle permet de s'authentifier avec un autre gabarit issu du même vecteur de caractéristiques. Dans notre cas, il s'agit de la même modalité, mais avec une nouvelle capture qui implique un nouveau vecteur de caractéristiques, différent, mais espéré proche.*

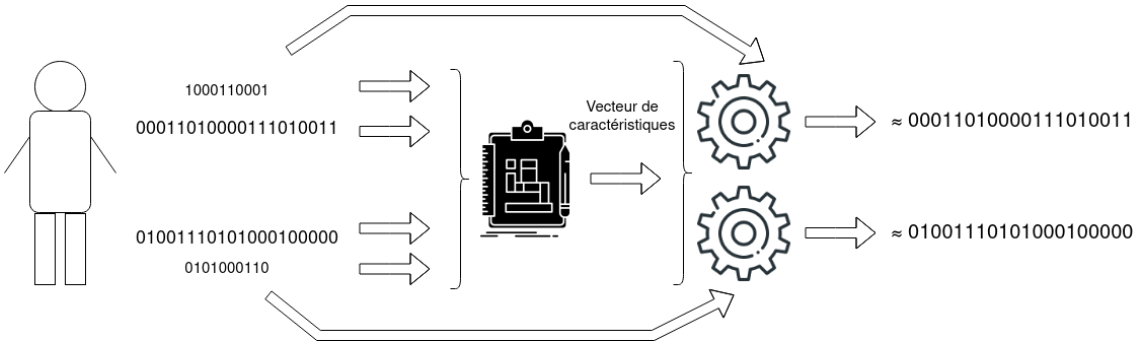


FIGURE 3.2 – Préimage proche et réutilisable

Nous avons décrit les PPR. Nous allons désormais introduire des algorithmes, autres que génétiques, et tester leur performance pour la construction de PPR.

3.3 Comparaison avec d'autres algorithmes

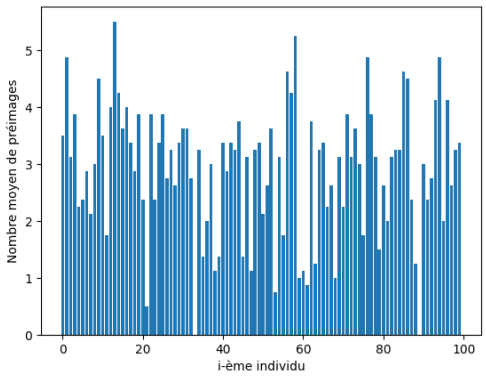
Cette section décrit les autres types d'algorithmes utilisés. Nous avons utilisé trois autres groupes d'expériences pour comparer les résultats issus d'algorithmes génétiques présentés en 3.4. Les expériences basées sur le choix parmi l'existant sont décrites en 3.3.1, celles sur la construction aléatoire en 3.3.2, et celles utilisant de la construction par escalade (*hill-climbing*) en 3.3.3. Les expériences sont effectuées avec le seuil $\tau = \tau@EER$. Chaque expérience utilise un nombre maximum $hd_{max} = 1000000$ fixé de comparaisons de distance entre deux gabarits. Cela permet de comparer les performances à durée égale.

3.3.1 Choix parmi l'existant

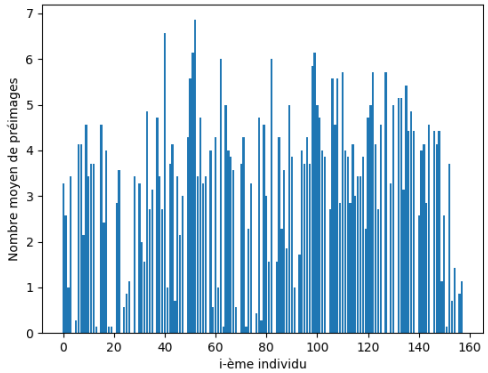
Nous utilisons une stratégie de choix parmi les vecteurs à notre disposition, issus de captures réelles. Une expérience consiste à prendre deux gabarits d'un individu, issus de deux transformations paramétrées par deux graines aléatoires. Nous prenons un par un chaque vecteur de chaque individu autre que celui dont sont issus les gabarits, et nous testons si ce vecteur est une préimage proche pour le premier gabarit et réutilisable pour le second gabarit. Pour les 3 bases, FVC, LFW, et PTB, nous donnons dans le tableau 3.1 le taux de couple de gabarits ayant au moins une préimage proche et réutilisable, ainsi que le nombre moyen de vecteurs de caractéristiques de la base de données biométriques se révélant être une PPR pour un couple de gabarits.

Base	Taux de gabarits ayant une PPR	Nombre moyen de PPR
FVC	96%	2.9
PTB	78%	2.9
LFW	77%	1.2

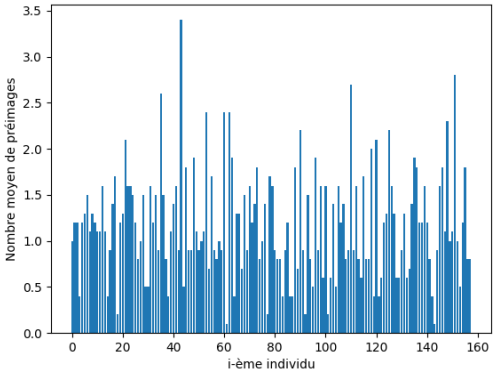
TABLE 3.1 – Indicateurs de préimages proches et réutilisables parmi une base de données biométriques



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.3 – Nombre moyen de PPR par individu

Remarque 11. *Lorsque l'on recherche une préimage pour un couple de gabarits parmi les vecteurs de caractéristiques des autres individus, cette stratégie n'atteint pas hd_{max} comparaisons étant donné la taille de nos bases de données biométriques.*

On trouve dans les figures 3.3 le détail du nombre moyen de préimages par individu pour les 3 bases de données biométriques. On constate une variation importante de ce paramètre. Pour la base FVC, on a entre zéro et plus de cinq préimages en moyenne par individu. Pour PTB, c'est entre zéro et sept, et entre zéro et plus de trois pour LFW.

Remarque 12. *Un gabarit de la base FVC a le choix parmi 792 vecteurs, un gabarit de la base PTB a le choix parmi 1099 vecteurs, et un gabarit de la base LFW a le choix parmi 1570 vecteurs. Cela influe sur le nombre moyen de PPR.*

Dans la suite, cette stratégie est appelée recherche par choix.

Cette stratégie de choix parmi l'existant nous permet de trouver une préimage proche et réutilisable pour une importante majorité des gabarits issus des vecteurs de caractéristiques des trois bases.

3.3.2 Construction aléatoire

Nous définissons une stratégie de construction aléatoire de vecteurs de caractéristiques. Pour construire un vecteur de caractéristiques aléatoires, nous générons une valeur aléatoire réelle bornée pour chaque indice du vecteur. Cette valeur aléatoire est bornée entre une valeur minimale et une valeur maximale déterminées par l'algorithme d'extraction utilisé pour produire les vecteurs de caractéristiques.

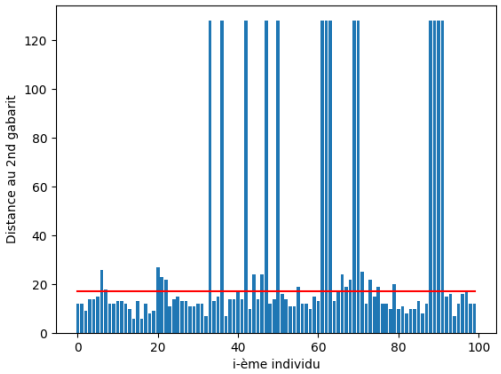
Pour comparer en temps de calcul similaire, nous générons des vecteurs, construisons les gabarits correspondants et calculons les distances au premier et au second gabarit, jusqu'à atteindre hd_{max} comparaisons. Nous conservons le vecteur minimisant la distance au second gabarit, lorsque la distance avec le premier gabarit est inférieure au seuil $\tau = \tau@EER$. Nous fixons $hd_{max} = 1000000$, ce qui permet de générer et de tester comme préimage proche et réutilisable 500000 vecteurs de caractéristiques.

Les résultats présentés dans le tableau 3.2 indiquent le taux de couples de gabarits pour lesquels on a trouvé aléatoirement une préimage proche et réutilisable. On y trouve la moyenne des distances au second gabarit pour tous ces couples, ainsi que la moyenne des distances au second gabarit des couples pour lesquels on a trouvé au moins une PPR. Dans ce dernier cas, la distance au premier gabarit est inférieure au seuil.

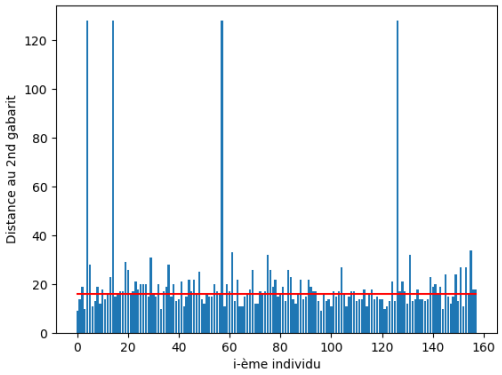
Base	Taux	Distance des candidats	Distance des PPR
FVC	68%	30	12
PTB	42%	20	13
LFW	100%	42	42

TABLE 3.2 – Indicateurs de préimages proches et réutilisables aléatoires

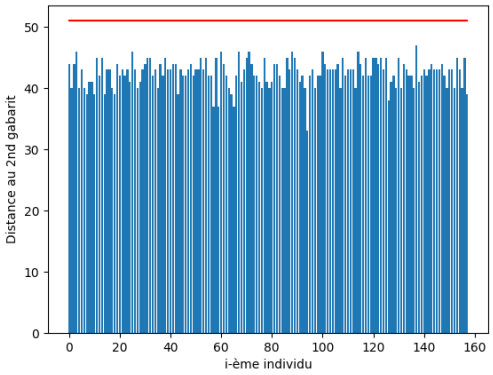
Les figures 3.4 montrent les distances au second gabarit par individu pour les 3 bases. La ligne horizontale rouge indique le seuil $\tau@EER$. Si la distance au second gabarit est en dessous de ce seuil, alors on a une préimage proche et réutilisable pour cet individu.



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.4 – Distance au 2nd gabarit par individu

Dans la suite, cette stratégie est appelée recherche aléatoire.

Cette stratégie de construction aléatoire nous permet de trouver une préimage proche et réutilisable pour environ la moitié des gabarits issus des vecteurs de caractéristiques des bases FVC et PTB, et pour la totalité de ceux issus de LFW. Les performances pour FVC et PTB sont moindres qu'avec le choix parmi l'existant 3.3.1, pour un coût plus élevé.

3.3.3 Construction par escalade

Dans cette section, nous avons défini et testé 3 stratégies d'attaque, inspirées de la méthode d'optimisation par escalade (*hill climbing*). L'algorithme part d'un vecteur de caractéristiques généré aléatoirement noté $x_0 \in \mathcal{M}_A$.
L'algorithme s'exécute ensuite en k itérations jusqu'à atteindre hd_{max} comparaisons de gabarits. Chaque itération utilise un vecteur $x_t \in \mathcal{M}_A$, issue de l'itération précédente t, pour produire un vecteur $x_{t+1} \in \mathcal{M}_A$.
À chaque itération t, l'algorithme génère un certain nombre v de vecteurs dits voisins $V_t = \{c_i \in \mathcal{M}_A\}_{i=1,\dots,v}$ de x_{t-1}, évalue ces vecteurs avec une fonction d'évaluation f tel que décrite en 3.4.1, et produit en sortie un nouveau vecteur x_t.

Nous avons implémenté 3 stratégies d'algorithme par escalade, dont nous décrivons le fonctionnement d'une itération t :

1. Pour chaque indice $i = 1, \dots, n$ du vecteur x_{t-1}, nous générons un voisin $c_i \in V_t$. Ce vecteur c_i voisin de x_{t-1} est construit en clonant x_{t-1} puis en remplaçant l'indice i de c_i par une valeur aléatoire bornée. Le vecteur sortant x_t est le vecteur voisin $c^* \in V_t = \{c_i\}_{i=1,\dots,n}$ minimisant le score de la fonction d'évaluation f.
   Exemple *:* Soit $x_{t-1} = (0.2, 1.3, 0.7)$, pour $i = 2$, $c_i = (0.2, \mathbf{0.9}, 0.7)$.

2. Le vecteur de sortie x_t est construit progressivement. Pour chaque indice $i = 1, \dots, n$ du vecteur x_{t-1}, nous générons v voisins $c_i^j \in V_t$ avec $j = 1, \dots, v$ modifiant une valeur comme dans la première stratégie. Nous insérons dans le $i^{\text{ème}}$ indice du vecteur de sortie x_t la valeur du $i^{\text{ème}}$ indice du vecteur voisin c_i^j minimisant le score de la fonction d'évaluation f.
   Exemple *:* Soit $x_{t-1} = (0.2, 1.3, 0.7)$ et $x_t = (0.4, NaN, NaN)$, pour $i = 1$, $c_i^2 = (0.2, \mathbf{0.9}, 0.7)$ le meilleur voisin, alors $x_t = (0.4, \mathbf{0.9}, NaN)$. *NaN* pour *Not a Number*, ces valeurs n'étant pas définies.

3. Dans cette stratégie, le vecteur de sortie x_t est construit progressivement. On initialise $x_t = x_{t-1}$. Pour chaque indice $i = 1, \dots, n$ du vecteur x_t, nous générons v voisins $c_i^j \in V_t$ avec $j = 1, \dots, v$, modifiant une valeur comme dans la première stratégie, et insérons dans le $i^{\text{ème}}$ indice du vecteur x_t la valeur du $i^{\text{ème}}$ indice du vecteur voisin c_i^j minimisant le score de la fonction f. La différence avec la précédente stratégie est que les vecteurs voisins c_i^j prennent en compte les modifications précédemment faites sur les i premiers indices.
   Exemple *:* Soit $x_{t-1} = (0.2, 1.3, 0.7)$ et $x_t = (0.4, 1.3, 0.7)$ (modifié à la première itération), pour $i = 1$, $c_i^2 = (0.2, \mathbf{0.9}, 0.7)$ le meilleur voisin, alors $x_t = (0.4, \mathbf{0.9}, 0.7)$.

Pour les stratégies 2 et 3, nous posons $v = 50$, permettant d'obtenir les meilleures performances dans nos expériences.

Le tableau 3.3 présente les résultats de ces trois stratégies. Comme pour le tableau 3.2, on indique le taux de couples de gabarits pour lesquels on a trouvé une PPR, la moyenne des distances au second gabarit pour tous ces couples, ainsi que la moyenne des distances au second gabarit des couples pour lesquels on a trouvé au moins une PPR.

Les performances sont médiocres pour la base PTB, composée de vecteurs de caractéristiques de taille 990 contre 512 pour les bases LFW et FVC. La contrainte d'exécution en $hd_{max} = 1000000$ ne permet pas de modifier suffisamment un unique vecteur aléatoire afin d'obtenir une préimage proche et réutilisable.

La deuxième stratégie est la plus performante. Avec cette deuxième stratégie, chaque itération permet de modifier toutes les valeurs du vecteur que l'on fait évoluer, contrairement à la première stratégie qui ne modifie qu'une valeur par itération.

Stratégie	Base	Taux	Distance des candidats	Distance des PPR
Première	FVC	11%	100	13
	PTB	1%	117	13
	LFW	10%	108	46
Deuxième	FVC	31%	61	13
	PTB	3%	113	12
	LFW	54%	71	44
Troisième	FVC	12%	95	12
	PTB	2%	121	14
	LFW	10%	106	48

TABLE 3.3 – Résultats des trois attaques par escalade

Cette stratégie de construction par escalade ne nous permet pas de trouver une préimage proche et réutilisable pour une partie importante des gabarits issus des vecteurs de caractéristiques des trois bases. Les performances de ces stratégies sont bien moindres qu'avec la recherche aléatoire ou par choix.

Nous avons décrit des algorithmes autres que génétiques et allons désormais construire des préimages avec un algorithme génétique.

3.4 Construction de préimage avec un algorithme génétique

Nous décrivons dans cette section la construction de préimage proche et réutilisable telle que définie en 3.2.2 à l'aide d'un algorithme génétique, décrit en 2.7.

L'algorithme a à sa disposition les deux gabarits et les deux graines paramétrant la transformation. À partir de ces données, nous construisons un vecteur de caractéristiques dit préimage. Nous testons s'il est proche du premier gabarit et vérifions s'il est réutilisable avec le second gabarit.

L'algorithme génétique est configuré avec différents paramètres. La fonction d'évaluation est un élément central et est décrite en 3.4.1.

Nous présentons ensuite l'impact de différentes combinaisons de paramètres, avec la taille de la population et le nombre d'itérations en 3.4.2, la stratégie de sélection en 3.4.3, les probabilités de mutation en 3.4.4, et le type de croisement en 3.4.5. Par exemple, lorsque nous présentons les résultats du type de croisement, nous avons une série de distances au second gabarit avec croisement simple, et une autre avec croisement double. Dans chacune de ces séries, seul ce paramètre de croisement est fixé, et on trouve donc des résultats d'expériences utilisant toutes les probabilités de mutation et tous les réglages possibles des autres paramètres.

À chaque fois, nous présentons les résultats de performances en fonction des paramètres sous forme de distance au second gabarit définie en 3.4.1.

Définition 3.4.1. *Soit x_1, x_2 deux vecteurs de caractéristiques. Soit $u_1 = \mathcal{T}(s_1, x_1) \in \mathcal{M}_B$ et $u_2 = \mathcal{T}(s_2, x_2) \in \mathcal{M}_B$ les gabarits obtenus avec les graines $s_1, s_2 \in \mathcal{K}$ et τ_B un seuil. Soit x^* le vecteur de caractéristiques candidat pour être une préimage proche et réutilisable telle que définie en 3.2.2. La distance au second gabarit est*

$$\delta(x^*, u_1, s_1, u_2, s_2) = \begin{cases} M & si\ V(u_1, \mathcal{T}(s_1, x^*), \tau_B) = Faux \\ D_B(u_2, \mathcal{T}(s_2, x^*)) & sinon. \end{cases}$$

avec D_B la distance de Hamming définie en 2.2.5 et M la taille d'un gabarit.

Une expérience est caractérisée par la personne pour laquelle on construit la préimage représentée par les deux gabarits u_1, u_2 et les deux graines s_1, s_2, et la configuration de l'algorithme génétique en termes de mode de sélection, de probabilité de mutation, de taille de population et de nombre d'itérations, ainsi que du type de croisement. Chaque expérience utilise un nombre maximum $hd_{max} = 1000000$ fixé de comparaisons de distance entre deux gabarits. Cela permet d'évaluer les performances sur un temps égal, exprimé en nombre de comparaisons noté $\#hd$.

Ainsi, l'algorithme génétique d'une expérience est relancé plusieurs fois jusqu'à atteindre cette limite hd_{max}. Le taux utilisé est $\tau = \tau@EER$. À la fin d'une expérience, on conserve le vecteur de caractéristiques x^* minimisant la distance au second gabarit $\delta(x^*, u_1, s_1, u_2, s_2)$.

Les résultats ont été agrégés en fonction de différents paramètres et sont présentés dans les sous-sections suivantes.

Pour chaque analyse d'un paramètre, nous présentons les résultats des expériences sous différentes formes.

Les diagrammes en boîte représentent la distribution des distances au second gabarit obtenues à la fin de chaque expérience. Les boîtes sont triées de bas en haut avec un ordre croissant sur la médiane. La moyenne est représentée par un triangle vert. Un diagramme en boite, aussi appelé boite à moustaches, représente des indicateurs statistiques d'une série, ici les distances au second gabarit obtenues avec des expériences partageant un même paramétrage. La boite représente le premier quartile *Q1*, la médiane (le deuxième quartile *Q2*), et le troisième quartile *Q3*. Les moustaches s'étendent jusqu'à 1.5 fois l'espace interquartile, soit $Q3+1.5*(Q3-Q1)$ et $Q1 - 1.5 * (Q3 - Q1)$

Des indicateurs clés sont fournis dans les tableaux. On y trouve notamment les valeurs des quartiles :

— Le premier quartile *Q1* : valeur en dessous de laquelle se trouve un quart des valeurs de la série.
— Le deuxième quartile *Q2*, autrement appelé la médiane : valeur en dessous de laquelle se trouve la moitié des valeurs de la série, et au-dessus de laquelle se trouve l'autre moitié des valeurs de la série.
— Le troisième quartile *Q3* : valeur au-dessus de laquelle se trouve un quart des valeurs de la série.

D'autres indicateurs y sont donnés :

— La moyenne.
— La valeur minimale de la série (min).
— La valeur maximale de la série (max).

Les courbes cumulées croissantes permettent de visualiser en tout point le positionnement d'une stratégie par rapport aux autres.

La transformation utilisée est le biohashing telle que définie en 2.4.3 avec la remarque 4. Les expériences ont été effectuées sur les trois bases de données biométriques, décrites en 2.5.1, 2.5.2 et 2.5.3.

3.4.1 Fonction d'évaluation

La fonction d'évaluation f est l'élément qui oriente la recherche. L'algorithme génétique cherche à minimiser le score de cette fonction d'évaluation en construisant son paramètre. Dans le cas présent, nous souhaitons avoir un vecteur qui s'authentifie avec le premier gabarit, puis qui minimise la distance au second gabarit. La fonction d'évaluation utilisée est :

$$D_B(u_1, \mathcal{T}(s_1, x^*)) \; + \; D_B(u_2, \mathcal{T}(s_2, x^*))$$

avec x^* le vecteur à évaluer.
D'autres fonctions d'évaluation ont été testées, mais elles n'ont pas permis la construction de préimage proche ou réutilisable, ou avec de moindres performances.

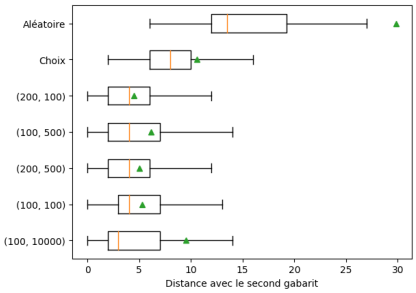
Nous allons désormais analyser les performances de chaque paramètre en utilisant cette fonction d'évaluation.

3.4.2 Taille de la population et nombre d'itérations

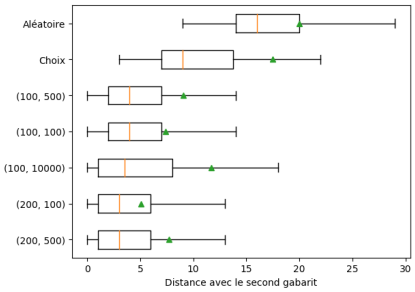
Nous nous intéressons à l'impact de la taille de la population et du nombre d'itérations sur les performances de l'algorithme génétique. Dans notre cas, les n vecteurs manipulés à chaque itération composent la population de taille n. Nous utilisons comme condition d'arrêt l'exécution de t itérations, et on relance l'algorithme jusqu'à atteindre hd_{max} comparaisons.

On constate dans les tableaux 3.4 et figures 3.5 et 3.6 que les meilleures médianes sont obtenues avec $pop = 200$ / $it = 500$ et $pop = 200$ / $it = 100$. Les valeurs minimales pour les bases FVC et PTB sont 0 quel que soit le paramétrage. Le paramétrage minimisant le score pour la base LFW est obtenu avec $pop = 200$ / $it = 500$. L'algorithme génétique est systématiquement plus performant que la recherche par choix 3.3.1 ou la recherche aléatoire 3.4.

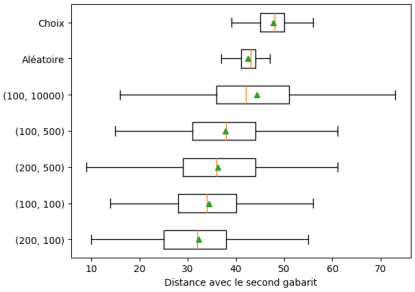
Nous conservons $pop = 200$ / $it = 500$ comme paramètre à utiliser.



(a) FVC

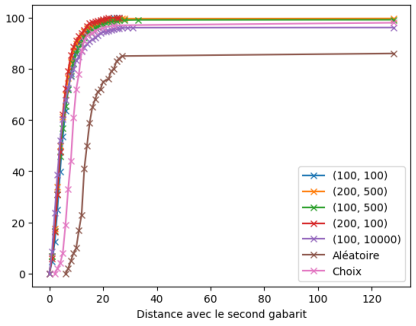


(b) PTB

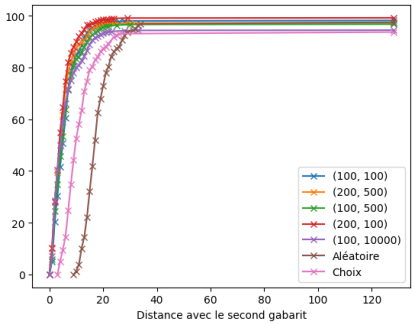


(c) LFW

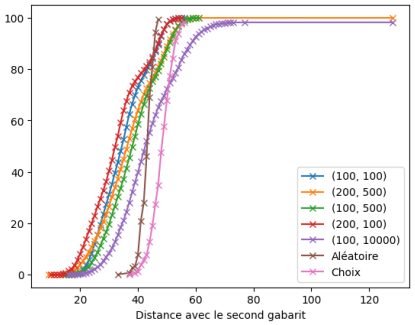
FIGURE 3.5 – Diagrammes en boîte en fonction de la taille de la population et du nombre d'itérations



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.6 – Courbes cumulées croissantes en fonction de la taille de la population et du nombre d'itérations

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	6	12.0	13.5	29.8	19.2	128
Choix	-	2	6.0	8.0	10.6	10.0	128
Population Itérations	(100, 100)	0	3.0	4.0	5.3	7.0	26
	(200, 500)	0	2.0	4.0	5.0	6.0	128
	(100, 500)	0	2.0	4.0	6.1	7.0	128
	(200, 100)	0	2.0	4.0	4.5	6.0	26
	(100, 10000)	0	2.0	3.0	9.5	7.0	128

PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	9	14.0	16.0	20.0	20.0	128
Choix	-	3	7.0	9.0	17.5	13.8	128
Population Itérations	(100, 100)	0	2.0	4.0	7.4	7.0	128
	(200, 500)	0	1.0	3.0	7.7	6.0	128
	(100, 500)	0	2.0	4.0	9.1	7.0	128
	(200, 100)	0	1.0	3.0	5.1	6.0	128
	(100, 10000)	0	1.0	3.5	11.7	8.0	128

LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	33	41.0	43.0	42.5	44.0	47
Choix	-	37	45.0	48.0	47.7	50.0	56
Population Itérations	(100, 100)	14	28.0	34.0	34.4	40.0	56
	(200, 500)	9	29.0	36.0	36.2	44.0	128
	(100, 500)	15	31.0	38.0	37.8	44.0	61
	(200, 100)	10	25.0	32.0	32.2	38.0	55
	(100, 10000)	16	36.0	42.0	44.3	51.0	128

TABLE 3.4 – Indicateurs des distances au second gabarit selon la taille de la population et le nombre d'itérations

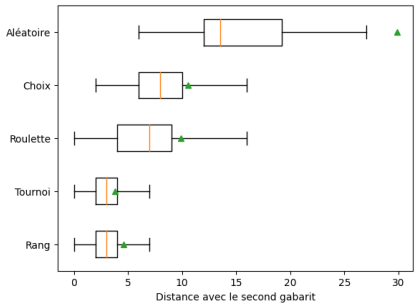
3.4.3 Étape de sélection

Nous présentons maintenant l'impact de l'étape de sélection. Dans cette étape, nous avons comme entrée une population $\Xi = \{\Phi_i\}_{i=1,\dots,n}$, composée d'un ensemble de n individus représentés par des vecteurs Φ_i. L'objectif est de sélectionner un sous-ensemble de cette population qui persiste dans l'itération suivante et sert d'entrée à l'étape de reproduction. Nous disposons d'une fonction d'évaluation f prenant en entrée un vecteur et donnant en sortie un score.
Nous évaluons trois stratégies [Hsu, 2002] :

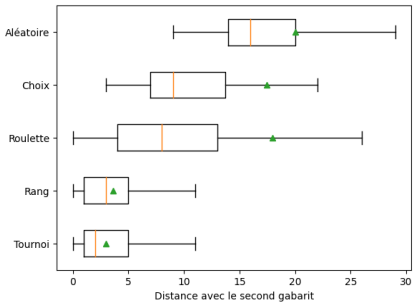
— Sélection par roulette : la probabilité d'un vecteur Φ_i d'être sélectionné est proportionnelle à sa note donnée par la fonction d'évaluation. $P(\Phi_i) = \frac{f(\Phi_i)}{\sum_{j=1}^{n} f(\Phi_j)}$
— Sélection par tournoi : on pioche aléatoirement deux vecteurs de la population pour une confrontation. Avec une probabilité p, on sélectionne celui qui a la meilleure note donnée par la fonction d'évaluation et on l'ajoute à la population suivante. On recommence jusqu'à avoir sélectionné le nombre de vecteurs souhaité. Cette probabilité p peut varier au cours du processus. Elle peut être initialement faible pour augmenter la probabilité de sélectionner le mauvais candidat et limiter la convergence rapide vers un optimum local, puis accroître cette probabilité au fur et à mesure du processus pour finalement converger vers une bonne solution globale.
— Sélection par rang : on trie les vecteurs en fonction de leur score donné par la fonction d'évaluation, et on les sélectionne dans l'ordre.

On constate dans les tableaux 3.5 et figures 3.7 et 3.8 que les meilleures médianes sont obtenues avec les modes de sélection par rang et tournoi. La sélection par roulette a des performances significativement en recul des deux autres méthodes de sélection. Elle est même moins performante que la recherche aléatoire avec la base LFW. Pour les bases FVC et PTB, on obtient à nouveau une distance minimale de 0 avec les trois modes de sélection. Pour la base LFW, la sélection par tournoi minimise la distance obtenue, mais la sélection par rang obtient la meilleure médiane. La sélection par tournoi permet d'obtenir les meilleures moyennes, médianes, et distances au second gabarit minimales.

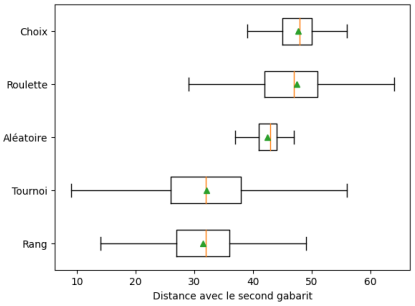
Nous conservons la méthode de sélection par rang comme paramètre à utiliser. Elle obtient des scores similaires à la sélection par tournoi, mais est déterministe.



(a) FVC

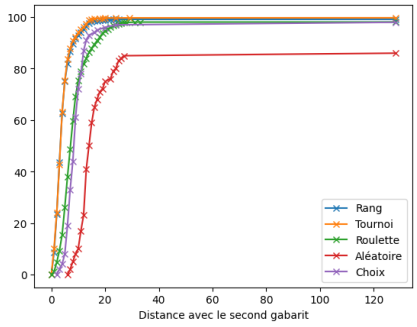


(b) PTB

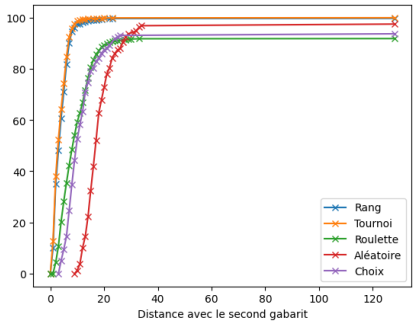


(c) LFW

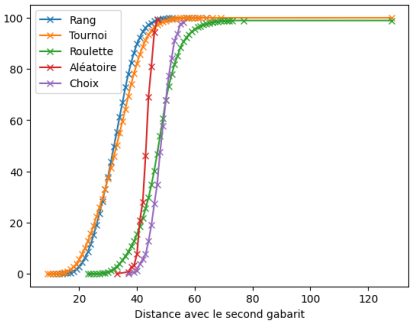
FIGURE 3.7 – Diagrammes en boîte en fonction de la méthode de sélection



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.8 – Courbes cumulées croissantes en fonction de la méthode de sélection

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	6	12.0	13.5	29.8	19.2	128
Choix	-	2	6.0	8.0	10.6	10.0	128
Mode de sélection	Rang	0	2.0	3.0	4.6	4.0	128
	Tournoi	0	2.0	3.0	3.8	4.0	128
	Roulette	0	4.0	7.0	9.9	9.0	128

PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	9	14.0	16.0	20.0	20.0	128
Choix	-	3	7.0	9.0	17.5	13.8	128
Mode de sélection	Rang	0	1.0	3.0	3.6	5.0	128
	Tournoi	0	1.0	2.0	3.0	5.0	128
	Roulette	0	4.0	8.0	18.0	13.0	128

LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	33	41.0	43.0	42.5	44.0	47
Choix	-	37	45.0	48.0	47.7	50.0	56
Mode de sélection	Rang	11	27.0	32.0	31.5	36.0	64
	Tournoi	9	26.0	32.0	32.1	38.0	128
	Roulette	23	42.0	47.0	47.4	51.0	128

TABLE 3.5 – Indicateurs des distances au second gabarit selon la méthode de sélection

3.4.4 Étape de mutation

Nous analysons l'étape de mutation, qui permet d'appliquer des erreurs en perturbant le vecteur. Pour chaque vecteur Ψ_i, indice par indice $j = 1, \ldots, m$, nous perturbons la valeur ψ_i^j avec une probabilité p.

Nous évaluons différentes probabilités p d'ajouter une erreur. Soit it le nombre d'itérations écoulées, pour chaque valeur d'un vecteur, nous le perturbons avec une probabilité p en lui ajoutant (ou en lui retirant) $(10 * (1 - \frac{it}{itMax})^2)\%$ de l'écart entre la valeur et la borne maximale (ou minimale). Cela permet d'avoir une mutation importante lors des premières itérations afin d'éviter les optimums locaux, puis de faibles perturbations aux dernières itérations pour converger vers une bonne solution globale.

Les résultats sont donnés en fonction de $P_{mutation}$, la probabilité p donnée sous forme de pourcentage. On constate dans les tableaux 3.6 et figures 3.10 et 3.11 que les meilleures médianes sont obtenues avec $P_{mutation} = 20$ et $P_{mutation} = 10$.

Pour les bases FVC et PTB, la probabilité de mutation influence peu la médiane, sauf pour $P_{mutation} = 0$ qui donne des performances médiocres. De même pour les valeurs minimales, les performances sont très proches. Pour la base LFW, l'impact de la probabilité de mutation sur les médianes est plus important, avec globalement une performance décroissante pour des probabilités croissantes.

Pour rappel, on trouve en ordonnée des figures la probabilité de mutation et la comparaison avec la recherche par choix et la recherche aléatoire. Les ordonnées sont triées en fonction de la médiane obtenue, d'où le fait que ça ne soit pas par ordre de probabilité de mutation croissant.

Nous conservons $P_{mutation} = 20$ comme paramètre à utiliser.

3.4.5 Étape de croisement

Finalement, nous analysons l'impact de l'étape de croisement, qui permet de mélanger deux vecteurs dits *parents* pour générer deux vecteurs dits *enfants*. On découpe ainsi les vecteurs parents en plusieurs morceaux, que l'on affecte aux enfants. Ce découpage des parents peut être fait en un point, auquel cas on fait un croisement simple. Les parents peuvent avoir plusieurs points de découpage, auquel cas on fait un croisement multiple. La figure 3.9 représente un croisement simple et un croisement multiple en deux points, sous forme de chromosomes. Ces deux différents croisements sont évalués pour voir si une méthode est plus efficace qu'une autre.

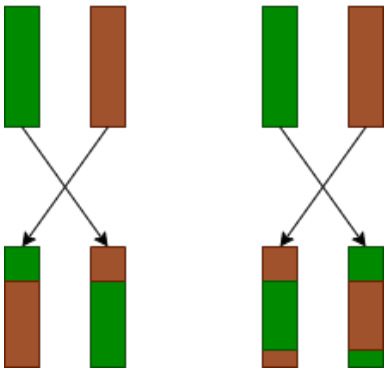
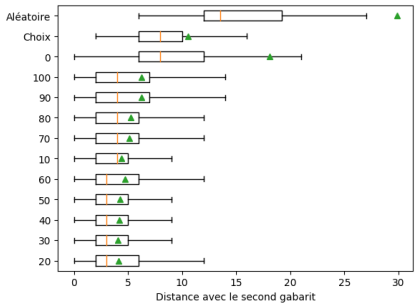


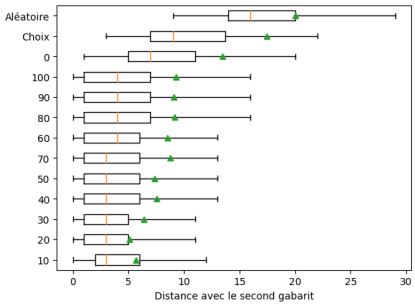
FIGURE 3.9 – Croisements simple et double

On constate dans les tableaux 3.7 et figures 3.12 et 3.13 que les performances sont similaires entre les deux types de croisement. Néanmoins, le croisement double permet de sensibles améliorations de moyennes et de médianes.

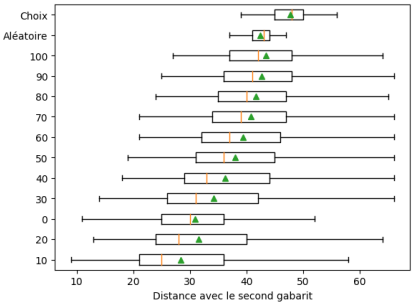
Nous conservons le croisement double comme paramètre à utiliser.



(a) FVC

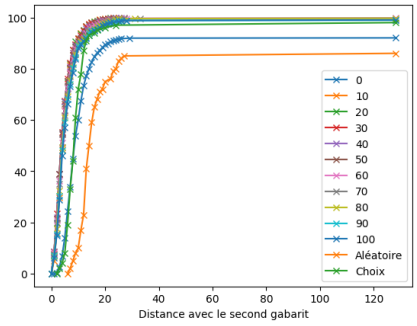


(b) PTB

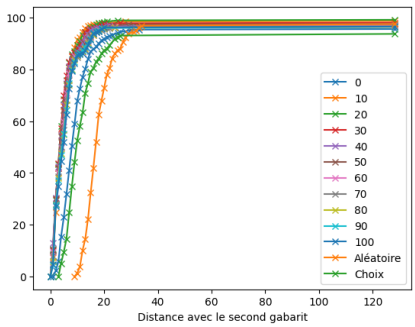


(c) LFW

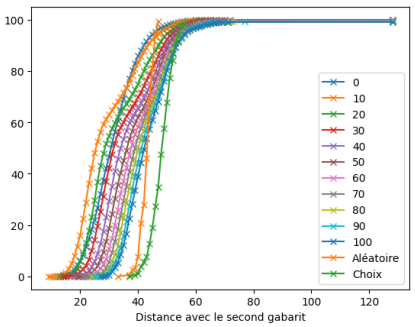
FIGURE 3.10 – Diagrammes en boîte en fonction de la probabilité de mutation



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.11 – Courbes cumulées croissantes en fonction de la probabilité de mutation

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	6	12.0	13.5	29.8	19.2	128
Choix	-	2	6.0	8.0	10.6	10.0	128
Probabilité de mutation	0	0	6.0	8.0	18.1	12.0	128
	10	0	2.0	4.0	4.4	5.0	24
	20	0	2.0	3.0	4.1	6.0	22
	30	0	2.0	3.0	4.1	5.0	23
	40	0	2.0	3.0	4.2	5.0	26
	50	0	2.0	3.0	4.3	5.0	27
	60	0	2.0	3.0	4.7	6.0	128
	70	0	2.0	4.0	5.1	6.0	128
	80	0	2.0	4.0	5.2	6.0	128
	90	0	2.0	4.0	6.2	7.0	128
	100	0	2.0	4.0	6.3	7.0	128

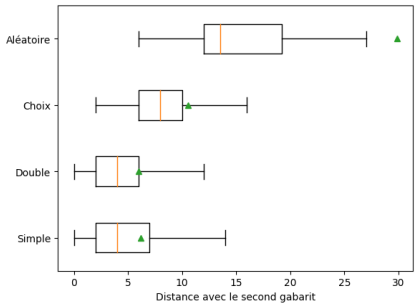
PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	9	14.0	16.0	20.0	20.0	128
Choix	-	3	7.0	9.0	17.5	13.8	128
Probabilité de mutation	0	1	5.0	7.0	13.5	11.0	128
	10	0	2.0	3.0	5.7	6.0	128
	20	0	1.0	3.0	5.1	5.0	128
	30	0	1.0	3.0	6.4	5.0	128
	40	0	1.0	3.0	7.5	6.0	128
	50	0	1.0	3.0	7.3	6.0	128
	60	0	1.0	4.0	8.5	6.0	128
	70	0	1.0	3.0	8.7	6.0	128
	80	0	1.0	4.0	9.2	7.0	128
	90	0	1.0	4.0	9.1	7.0	128
	100	0	1.0	4.0	9.3	7.0	128

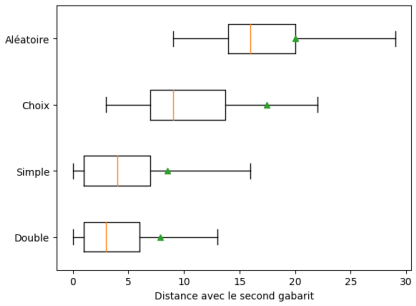
LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	33	41.0	43.0	42.5	44.0	47
Choix	-	37	45.0	48.0	47.7	50.0	56
Probabilité de mutation	0	11	25.0	30.0	30.9	36.0	128
	10	9	21.0	25.0	28.4	36.0	69
	20	13	24.0	28.0	31.5	40.0	128
	30	14	26.0	31.0	34.2	42.0	128
	40	18	29.0	33.0	36.3	44.0	128
	50	19	31.0	36.0	37.9	45.0	128
	60	21	32.0	37.0	39.4	46.0	128
	70	21	34.0	39.0	40.7	47.0	128
	80	24	35.0	40.0	41.7	47.0	128
	90	25	36.0	41.0	42.6	48.0	128
	100	27	37.0	42.0	43.4	48.0	128

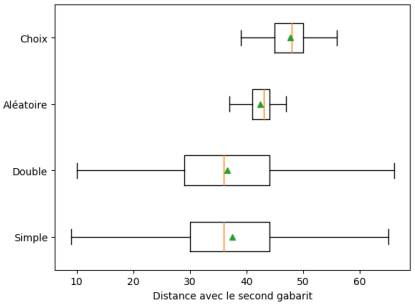
TABLE 3.6 – Indicateurs des distances au second gabarit selon la probabilité de mutation



(a) FVC

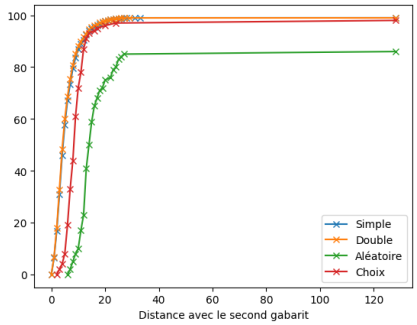


(b) PTB

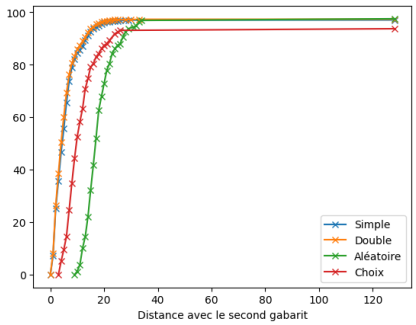


(c) LFW

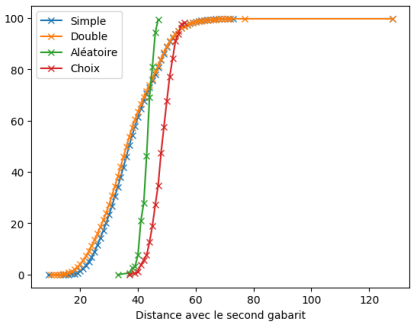
FIGURE 3.12 – Diagrammes en boîte en fonction du type de croisement



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.13 – Courbes cumulées croissantes en fonction du type de croisement

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	6	12.0	13.5	29.8	19.2	128
Choix	-	2	6.0	8.0	10.6	10.0	128
Type de croisement	Simple	0	2.0	4.0	6.2	7.0	128
	Double	0	2.0	4.0	6.0	6.0	128

PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	9	14.0	16.0	20.0	20.0	128
Choix	-	3	7.0	9.0	17.5	13.8	128
Type de croisement	Simple	0	1.0	4.0	8.5	7.0	128
	Double	0	1.0	3.0	7.9	6.0	128

LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	33	41.0	43.0	42.5	44.0	47
Choix	-	37	45.0	48.0	47.7	50.0	56
Type de croisement	Simple	9	30.0	36.0	37.5	44.0	128
	Double	10	29.0	36.0	36.5	44.0	128

TABLE 3.7 – Indicateurs des distances au second gabarit selon le type de croisement

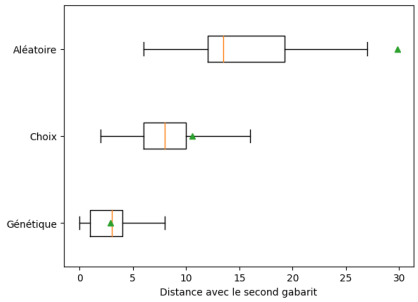
3.4.6 Paramètres optimaux

Les sous-sections précédentes ont permis de dégager des paramètres optimaux. Nous calculons leur performance et nous les comparons aux autres méthodes.

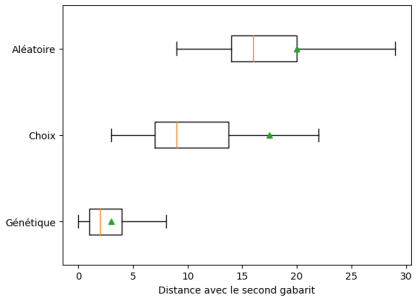
On trouve dans les tableaux 3.8 et figures 3.14 et 3.15 les résultats de la construction de préimage par choix, de la construction de préimage aléatoire, et de la construction de préimage par algorithme génétique. L'algorithme génétique utilise les paramètres précédemment décrits et sélectionnés : une population de taille 200 évoluant sur 500 itérations, utilisant une sélection par rang, combinée avec un croisement double, perturbée par des mutations de probabilité $P_{mutation} = 20$.

On constate que l'algorithme génétique surpasse les deux autres méthodes pour les 3 bases, et permet une amélioration significative de tous les indicateurs.

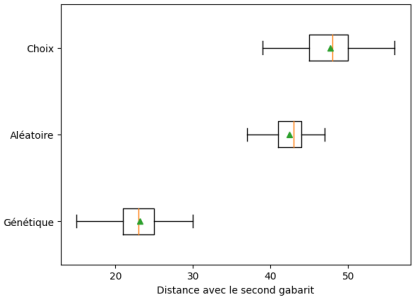
La figure 3.16 représente l'évolution de la distance au second gabarit au cours des itérations. On observe cette évolution pour l'expérience minimisant cette distance pour les trois bases.



(a) FVC

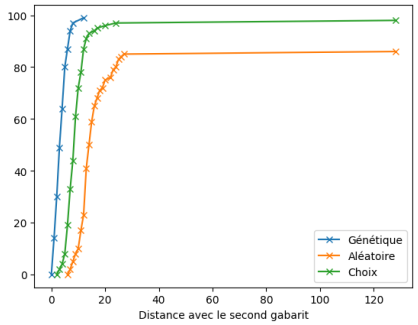


(b) PTB

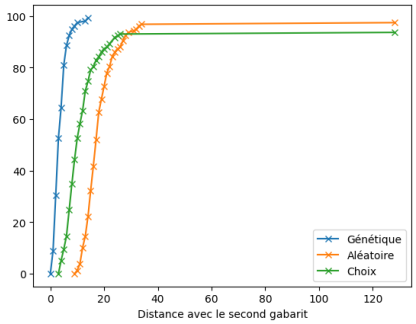


(c) LFW

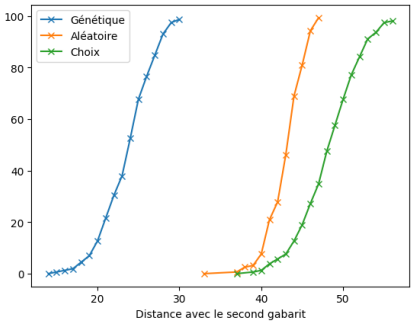
FIGURE 3.14 – Diagrammes en boîte en fonction de l'algorithme



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.15 – Courbes cumulées croissantes en fonction de l'algorithme

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	6	12.0	13.5	29.8	19.2	128
Choix	-	2	6.0	8.0	10.6	10.0	128
Génétique	-	0	1.0	3.0	2.9	4.0	12

PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	9	14.0	16.0	20.0	20.0	128
Choix	-	3	7.0	9.0	17.5	13.8	128
Génétique	-	0	1.0	2.0	3.0	4.0	14

LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Aléatoire	-	33	41.0	43.0	42.5	44.0	47
Choix	-	37	45.0	48.0	47.7	50.0	56
Génétique	-	14	21.0	23.0	23.1	25.0	30

TABLE 3.8 – Indicateurs des distances au second gabarit en fonction de l'algorithme

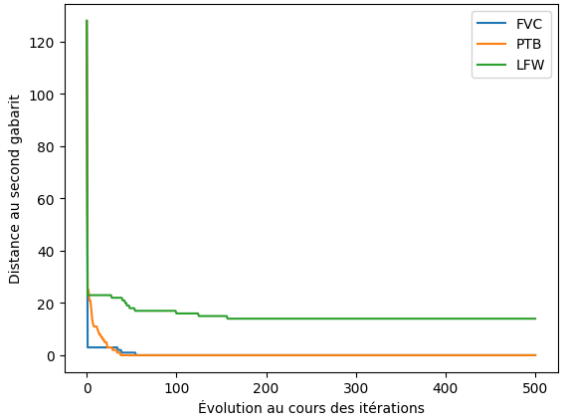


FIGURE 3.16 – Évolution de la distance au second gabarit

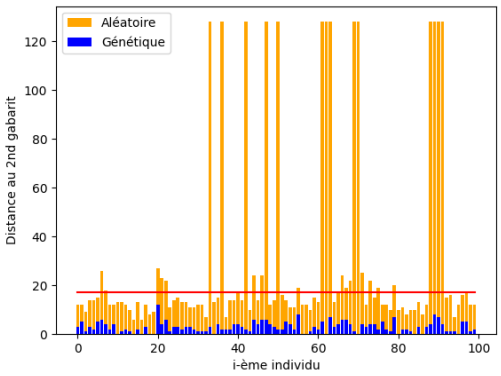
À titre de comparaison avec la recherche aléatoire dont les résultats sont présentés en 3.3.2, nous fournissons dans le tableau 3.9 le taux de couples de gabarits pour lesquels nous avons trouvé une préimage proche et réutilisable avec l'algorithme génétique, ainsi que la moyenne des distances obtenues au second gabarit pour tous ces couples. Nous obtenons désormais une préimage pour 100% des couples de gabarits, et améliorons largement la distance au second gabarit.

Les figures 3.17 montrent les distances au second gabarit par individu pour les 3 bases. La ligne horizontale rouge indique le seuil utilisé $\tau@EER$. Nous reportons les résultats de la recherche aléatoire présentés dans les figures 3.4 pour constater

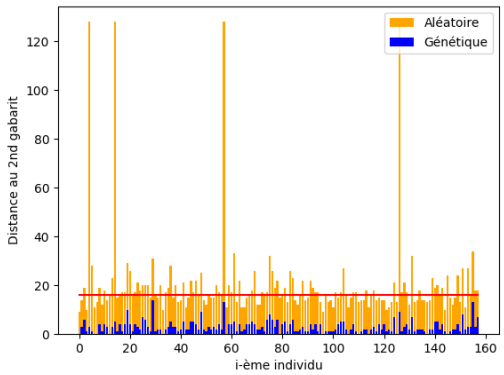
l'amélioration significative des performances obtenues avec l'algorithme génétique par rapport à l'algorithme de recherche aléatoire, les deux se limitant à hd_{max}.

Base	τ	Taux	Distance des PPR
FVC	17	100%	3
PTB	16	100%	3
LFW	51	100%	23

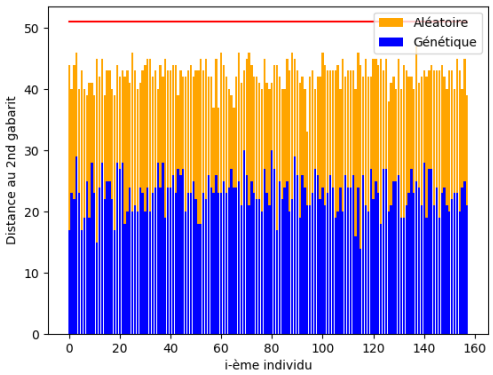
TABLE 3.9 – Indicateurs de préimages proches et réutilisables issues de l'algorithme génétique



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.17 – Distance au 2nd gabarit par individu

L'algorithme génétique ainsi paramétré nous procure de très bonnes performances avec 100% de préimages proches et réutilisables. Nous allons maintenant décrire d'autres variantes de notre problématique, puis vérifier si les performances se maintiennent.

3.5 Variantes de choix de gabarits

Dans l'algorithme génétique utilisé jusqu'ici, nous cherchons une préimage proche et réutilisable telle que définie dans 3.2.2. On note que les deux vecteurs de caractéristiques sont issus de captures de la même modalité de la même personne, par exemple, deux captures d'empreintes digitales du même doigt. Cette stratégie initiale est notée $mode0$.

Nous avons défini deux autres stratégies :

— La deuxième stratégie utilise deux vecteurs issus de captures de personnes différentes. Dans l'exemple des empreintes digitales, on utilise un premier vecteur x_1 issu d'une capture d'un doigt d'une personne, et un second vecteur x_2 issu d'une capture d'un doigt d'une autre personne. Cette deuxième stratégie est notée $mode1$.

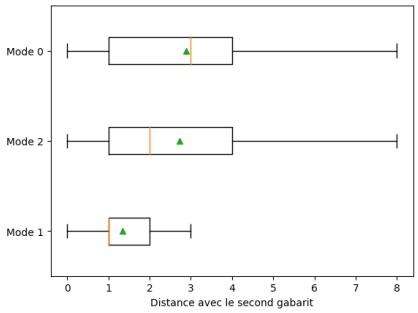
— Nous posons une troisième stratégie qui utilise en vecteurs de caractéristiques x_1, x_2 un unique vecteur v, c'est-à-dire la même capture unique d'un doigt d'une personne, $x_1 = x_2$. Cette stratégie pour la comparaison est notée $mode2$. Il s'agit au final de construire une préimage qui minimise la distance avec un unique gabarit.

Nous avons mené des expériences sur ces 3 modes, avec une recherche génétique paramétrée comme dans la section 3.4.6, avec une recherche aléatoire, et avec une recherche par choix. À nouveau nous relançons ces algorithmes pour chaque personne jusqu'à atteindre notre limite $hd_{max} = 1000000$. La recherche par choix est à nouveau bornée par le nombre de vecteurs de caractéristiques disponibles dans la base, ne permettant pas d'atteindre cette limite.

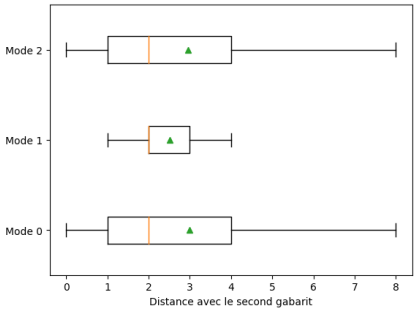
On trouve dans les tableaux 3.10 et figures 3.18 et 3.19 les résultats de ces expériences. Nous pensions intuitivement avoir des performances significativement meilleures avec $mode0$ qu'avec $mode1$, ce qu'infirment les expériences. On obtient des médianes similaires, mais légèrement plus faibles pour $mode1$ que pour $mode0$. Les autres indicateurs sont aussi similaires. Les courbes 3.19 montrent un avantage de performances pour $mode1$ et des performances confondues entre $mode0$ et $mode2$.

Les performances du $mode2$ ne sont en effet pas les meilleures. Cela est expliqué par notre fonction d'évaluation f décrite en 3.4.1 qui dirige l'algorithme génétique avec la moyenne des distances aux deux gabarits. Ainsi, avoir deux gabarits distincts peut procurer un avantage pour minimiser cette moyenne.

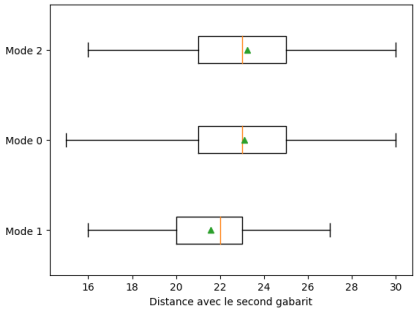
La figure 3.20 permet de suivre l'évolution de la distance au second gabarit au cours des itérations pour les trois bases. On remarque que cette distance converge lentement avec la base LFW.



(a) FVC

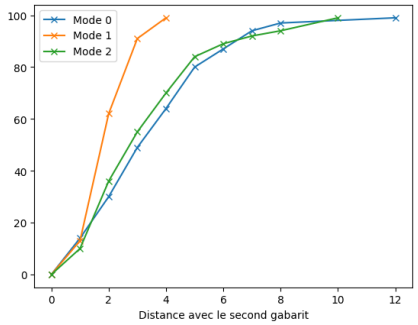


(b) PTB

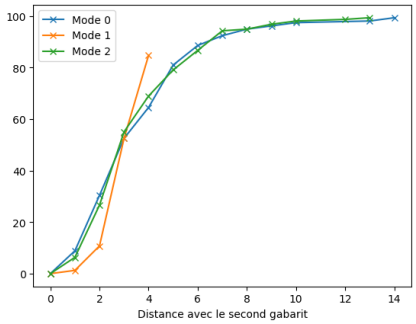


(c) LFW

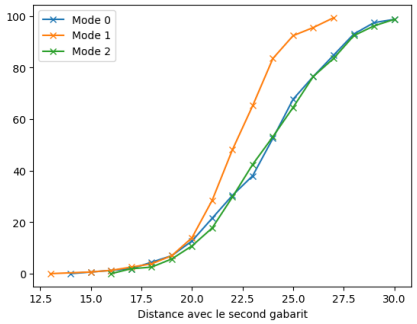
FIGURE 3.18 – Diagrammes en boîte selon la stratégie



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.19 – Courbes cumulées croissantes selon la stratégie

FVC

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Génétique	Mode 0	0	1.0	3.0	2.9	4.0	12
	Mode 1	0	1.0	1.0	1.4	2.0	4
	Mode 2	0	1.0	2.0	2.7	4.0	10

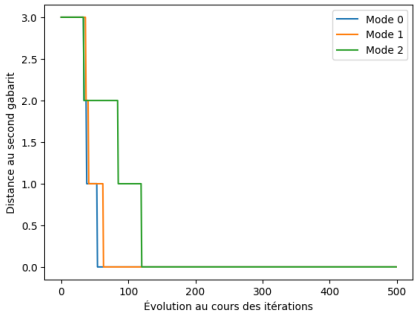
PTB

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Génétique	Mode 0	0	1.0	2.0	3.0	4.0	14
	Mode 1	0	2.0	2.0	2.5	3.0	4
	Mode 2	0	1.0	2.0	3.0	4.0	13

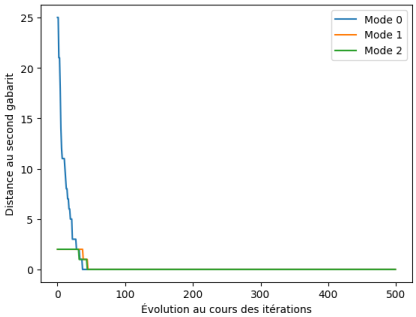
LFW

Paramètre	Valeurs	min	Q1	médiane	moyenne	Q3	max
Génétique	Mode 0	14	21.0	23.0	23.1	25.0	30
	Mode 1	13	20.0	22.0	21.6	23.0	27
	Mode 2	16	21.0	23.0	23.2	25.0	30

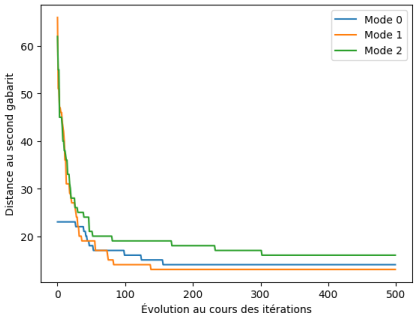
TABLE 3.10 – Indicateurs des distances au second gabarit selon la stratégie



(a) FVC



(b) PTB



(c) LFW

FIGURE 3.20 – Évolution de la distance au second gabarit selon la stratégie

Le *mode2* sert uniquement à titre de comparaison. Le *mode1* construit un vecteur qui permet de s'authentifier soit même, mais aussi pour quelqu'un d'autre. La prochaine section va généraliser le *mode1*, qui obtient de bonnes performances malgré la diversité des gabarits proposés pour lesquels il faut construire une PPR.

3.6 Préimage universelle

Dans cette section, nous abordons le concept de passe-partout biométrique. L'idée nous est venue du constat fait en section 3.5, démontrant la possibilité de rechercher une préimage proche et réutilisable avec deux gabarits issus de deux personnes différentes. Nous allons commencer par définir ce qu'est un passe-partout, avant de décrire la construction d'un passe-partout avec un algorithme génétique pour une base de données biométriques révocables.

3.6.1 Passe-partout

Nous avons souhaité pousser le concept plus loin : avoir un vecteur de caractéristiques dont les gabarits sont proches de tous les gabarits.

Un passe-partout biométrique x est un vecteur de caractéristiques synthétique dont les gabarits biométriques sont proches de nombreux gabarits de la base de données biométriques révocables. Ce concept est schématisé en 3.21.

Définition 3.6.1. *Soit* $D = \{(u_i, s_i)\}_{i=1,\dots,n}$ *une base de données biométriques révocables. Soit* τ *le seuil.* x *est un passe-partout pour* D*, avec* τ*, si quel que soit* i *entre* 1 *et* n*,* $\mathcal{V}(\mathcal{T}(s_i, x), u_i, \tau) = Vrai$*.*

Maintenant que le concept de passe-partout biométrique est formalisé, nous allons détailler les expériences effectuées et les résultats obtenus.

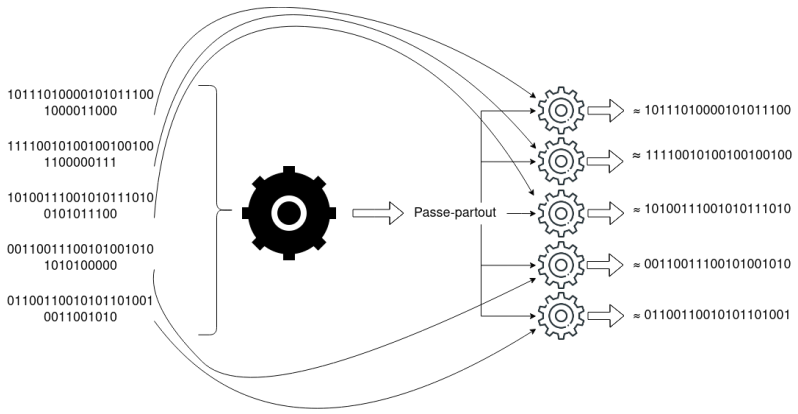


FIGURE 3.21 – Passe-partout depuis une base de données biométriques révocables

3.6.2 Construction d'un passe-partout pour une base de données biométriques révocables

La base de données biométriques révocables $D = \{(u_i, s_i)\}_{i=1,\dots,n}$ est construite et nous connaissons les gabarits et les graines associées. Nous souhaitons construire un vecteur de caractéristiques artificiel x^* qui maximise la taille de $D' \subseteq D$ tel que $\forall\ (u, s)\ \in\ D', \mathcal{V}(\mathcal{T}(s, x^*), u, \tau) = Vrai$.

Évidemment, la recherche d'un passe-partout x^* pour la base de données biométriques révocables n'est pas aisée lorsque la taille de la base est importante. Ainsi, nous introduisons le partitionnement d'une base de données biométriques révocables par un ensemble de r passe-partout, de telle sorte que chaque gabarit de la base soit proche du gabarit correspondant d'au moins un des r passe-partout. Cette notion de couverture d'une base de données biométriques révocables par plusieurs passe-partout est définie en 3.6.2.

Définition 3.6.2. *Soit $D = \{(u_i, s_i)\}_{i=1,\dots,n}$ une base de données biométriques révocables et τ le seuil. D est dit couvert par r passe-partout $\{x^1, \dots x^r\}$ avec τ si quel que soit i entre 1 et n, il existe k entre 1 et r tel que $\mathcal{V}(\mathcal{T}(s_i, x^k), u_i, \tau) = Vrai$.*
Le nombre minimum R tel que D est couvert par R passe-partout est appelé taille optimale de dictionnaire.

Remarque 13. *Dans la définition 3.6.2, chaque passe-partout $x_i \in \{x_1, \dots x_r\}$ est un passe-partout, au sens de la définition 3.6.1, pour un sous-ensemble $D_i \subset D$ et on a $\cup_{i=1}^{r} D_i = D$. Cependant, le nombre de gabarits de chaque sous-ensemble D_i est généralement différent.*

Pour construire les R passe-partout, nous construisons récursivement un passe-partout pour la base de données biométriques révocables $D_i \subset D$ composée des gabarits non usurpés par un des passe-partout déjà construits. Nous poursuivons la construction de nouveaux passe-partout jusqu'à ce que $D_i = \phi$.

Afin d'évaluer le potentiel de couverture d'un passe-partout tel que remarqué en 13, nous définissons le taux de couverture en 3.6.3.

Définition 3.6.3. *Une base de données biométriques révocables $D = \{(u_i, s_i)\}_{i=1,\dots,n}$ est dite ϵ-couverte par un passe-partout x, avec $0 < \epsilon \leq 1$ si il existe un sous-ensemble $D' \subset D$ avec ϵn personnes tel que x est un passe-partout D'. Le taux de couverture optimale de D est le nombre maximum E tel que D est E-couvert par le passe-partout $x \in \mathcal{M}_A$.*

Nous utilisons un algorithme génétique pour rechercher un passe-partout, tel que présenté dans la définition 3.6.1, à partir d'une base de données biométriques révocables. L'algorithme génétique est paramétré comme en 3.4.6. Trois fonctions d'évaluation f ont été investiguées :

— La première est basée sur la moyenne des distances de Hamming de chaque personne.
— La deuxième est basée sur la somme des distances de Hamming des personnes non usurpées par le passe-partout en construction.
— La troisième est basée sur le nombre de personnes usurpées par le passe-partout en construction.

La seconde fonction d'évaluation (*somme*) permet d'obtenir les meilleurs résultats sur les trois bases. Les résultats présentés ont été obtenus avec cette fonction.

Les expériences ont été effectuées avec une base de données biométriques révocables construite avec des graines aléatoires paramétrant la transformation du premier vecteur de caractéristiques de chaque individu.

Le tableau 3.11 donne les meilleurs taux de couverture optimaux (TCO), tels que décrits dans 3.6.3, et les meilleures tailles optimales de dictionnaire (TOD), telles que décrites dans 3.6.2. Pour ces expériences, $\tau = \tau@EER$.

Base	TCO (%)	TOD
FVC	73	5
LFW8	21	12
LFW10	15.2	18
PTB	61	12

TABLE 3.11 – Taux de couverture optimale / Taille optimale de dictionnaire

On voit sur la figure 3.22 que le taux de couverture n'augmente plus à partir de 400 itérations pour les 3 bases.

L'EER des bases FVC et PTB étant important, nous présentons dans le tableau 3.12 les taux de couverture optimaux en utilisant des seuils $\tau \leq \tau@EER$. Cela permet de comparer avec la base LFW en abaissant le taux de fausses acceptations (FMR).

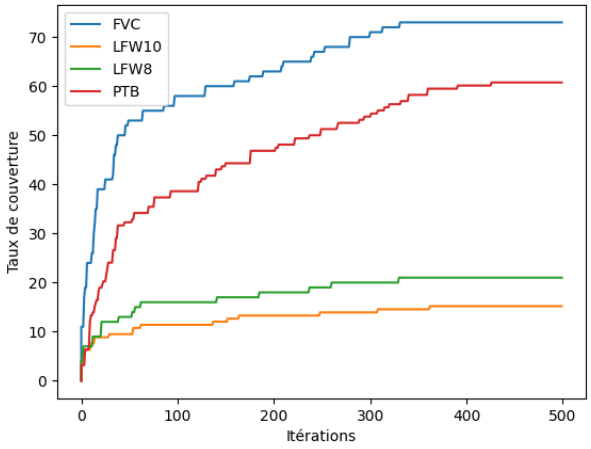


FIGURE 3.22 – Évolution du taux de couverture pendant les itérations

τ	16	15	14	13	12	11	10	9	8	7
FMR (%) for FVC	12.9	9.5	6.7	4.4	2.8	1.7	1	0.5	0.2	0.1
FMR (%) for PTB	17	13.3	9.5	6.4	4.3	2.7	1.6	0.9	0.5	0.2
TCO (%) for FVC	64	58	49	42	38	29	23	19	14	11
TCO (%) for PTB	61	53	46	40	32	25	19	15	11	7

TABLE 3.12 – FMR/TCO en fonction des seuils (FVC et PTB)

La construction de passe-partout pour une base de données biométriques révocables étant performante avec l'algorithme génétique que nous avons optimisé, nous discutons désormais de cas d'usage.

3.6.3 Cas d'usage

Nous présentons un cas d'usage de la construction d'un passe-partout depuis une base de données biométriques révocables. Un attaquant qui a un accès en lecture à une base de données biométriques révocables peut utiliser les gabarits et graines dont il a la connaissance pour construire un vecteur de caractéristiques passe-partout couvrant au mieux cette base de données biométriques révocables. Si cet attaquant peut exploiter l'attaque du point 4 définie en 2.1.3, qui permet d'écrire le vecteur de caractéristiques transmis au comparateur, alors il peut s'identifier sur le système avec le passe-partout construit.

3.7 Conclusion

Dans ce chapitre, nous avons introduit la notion de préimage proche et réutilisable. Une telle préimage permet de s'authentifier avec un premier gabarit d'un individu, mais aussi de s'authentifier avec un second gabarit de cet individu. Nous construisons ces préimages à l'aide d'un algorithme génétique. Afin de vérifier les bonnes performances de cet algorithme, nous comparons les résultats avec d'autres méthodes de construction : la construction aléatoire de préimage, et le choix de préimage parmi des vecteurs de caractéristiques existants. Toutes ces constructions sont effectuées en temps constant, défini en termes de comparaisons de gabarits. Nous avons aussi exploré la construction de préimage par escalade à travers trois stratégies. Néanmoins, en temps constant avec les autres méthodes présentées, les performances sont médiocres.

Nous avons ensuite cherché à optimiser notre algorithme génétique, en cherchant les bons paramètres à utiliser. Nous avons déterminé qu'utiliser une population de taille 200 évoluant sur 500 itérations permet de meilleures performances. Sélectionner les individus se reproduisant à l'aide d'une sélection par rang permet de bien meilleurs résultats que la sélection par roulette et est déterministe contrairement à la sélection par tournoi. La probabilité de mutation influe sur la précision des préimages, et la fixer à 20% permet d'optimiser les résultats. Enfin, un double croisement des parents pour la construction d'enfants permet d'améliorer sensiblement les performances. Finalement, un algorithme génétique paramétré tel que décrit en 3.4.6 permet de construire des préimages significativement meilleures qu'aléatoirement ou par choix.

Nous avons testé cette construction de préimage avec un algorithme génétique sur d'autres scénarios. Nous observons que la construction d'une préimage proche, s'authentifiant pour un gabarit d'un individu, et réutilisable, en s'authentifiant pour un gabarit d'un autre individu, est aussi réalisable. De plus, la construction dans ce scénario obtient des performances similaires.

Nous avons élargi le concept de préimage proche et réutilisable au concept de passe-partout, une préimage s'authentifiant pour de nombreux gabarits d'une base de données biométriques révocables. Nous avons formalisé la taille optimale de dictionnaire et le taux de couverture optimal, et avons construit un passe-partout usurpant 73% des gabarits de la base FVC. Seulement 5 passe-partout permettent de

couvrir intégralement cette base. Les expériences ont été effectuées sur les bases LFW et PTB, mais aussi avec des seuils τ plus faibles.

Les vulnérabilités des transformations révocables avec jetons publics ont déjà été étudiées. Les préimages proches ont déjà été construites selon diverses méthodes. Néanmoins, c'est la première fois que ces attaques sont étendues à l'entièreté de la base avec la construction de passe-partout. Notre construction utilise un algorithme génétique et non une faiblesse spécifique du biohashing. Par exemple, les attaques avec un algorithme d'optimisation linéaire ont de bons résultats dans le calcul de préimages utilisant du biohashing, mais dépendent de cette transformation.

Chapitre 4

Passe-partout biométrique

Résumé : *Ce chapitre introduit un second scénario de passe-partout. Nous choisissons les graines, projetant la base de données biométriques en une base de données biométriques révocables couverte par un vecteur de caractéristiques passe-partout connu. Une stratégie de force brute pour trouver les graines est utilisée. Cette stratégie nous amenant à construire de nombreuses matrices de projection, nous introduisons deux nouvelles constructions de matrice de projection en alternative à l'utilisation de l'algorithme de Gram-Schmidt. Nous développons une extension d'individu passe-partout, en utilisant plusieurs vecteurs de caractéristiques d'un individu pour la recherche de graine, permettant d'améliorer la couverture d'autres vecteurs issus du même individu. La complexité des deux scénarios de passe-partout est formalisée.*

Mots-clés : *passe-partout, projection, attaque, complexité, corrélation.*

4.1 Introduction

Dans cette partie, nous travaillons sur un nouveau scénario du passe-partout présenté en 3.6.2. Dans ce premier scénario, la base de données biométriques révocables est utilisée pour construire avec un algorithme génétique un vecteur de caractéristiques passe-partout couvrant au maximum la base.

Dans le second scénario que nous introduisons dans ce chapitre, la base de données biométriques est connue. Sans perte de généralité, cela fonctionne aussi si on connait les données biométriques au fur et à mesure. Nous fixons un vecteur de caractéristiques x^* que nous choisissons comme candidat passe-partout.

L'objectif est de créer une base de données biométriques révocables, en choisissant les graines paramétrant la transformation en gabarits, de telle sorte que les gabarits générés avec x^* et les différentes graines s'authentifient avec un maximum de gabarits de la base nouvellement construite. Cette stratégie permet d'avoir un meilleur taux de couverture que dans le premier scénario.

Ce second scénario est formalisé dans la définition 4.1.1.

Définition 4.1.1. *Soit $B = \{x_i\}_{i=1,\dots,n}$ une base de données biométriques telle que définie dans 2.3.2. Soit x^* un vecteur de caractéristiques fixé, réel ou artificiel. L'objectif est de trouver un ensemble de graines $S = \{s_i\}_{i=1,\dots,n}$ permettant de créer la base de données biométriques révocables $B' = \{(u_i, s_i)\}_{i=1,\dots,n}$ avec $u_i = \mathcal{T}(s_i, x_i)$. Les graines choisies doivent permettre de maximiser la taille de $B^* \subseteq B'$ tel que $\forall\ (u, s) \in B^*$, $V(u, \mathcal{T}(s, x^*), \tau) = Vrai$.*

Remarque 14. *Nous utilisons t bases de données biométriques $B_j = \{x_i^j\}_{i=1,\dots,n}$, avec $j = 1, \dots, t$. Ces t bases sont composées d'un vecteur de caractéristiques original de i individus. Ces t bases permettent uniquement de multiplier les données expérimentales.*

Ce scénario nécessite l'essai de nombreuses graines et donc la construction de nombreuses matrices de projection depuis ces graines. Nous présentons dans la section 4.2 les transformations que nous allons utiliser, permettant une construction de matrice plus efficace que dans le biohashing avec l'algorithme de Gram-Schmidt. Les performances de ces nouvelles projections sont analysées, notamment leur impact sur l'EER, mais aussi leur gain en efficacité de construction de matrice.

La section 4.3 présente en détail notre second scénario de construction d'une base de données biométriques révocables pour un passe-partout, ainsi que les résultats des expériences.

Nous détaillons la complexité théorique des deux scénarios dans la section 4.4

Nous introduisons une extension de ce second scénario dans la section 4.5. Les graines sont choisies en prenant en compte plusieurs vecteurs de caractéristiques d'un individu. Le concept d'individu passe-partout y est détaillé, notamment le choix des vecteurs pour rechercher la graine et le choix des vecteurs pour tester les performances en 4.5.1. Les résultats des expériences de ces individus passe-partout sont présentés en 4.5.2, et nous détaillons la corrélation entre la couverture des vecteurs utilisés pour la recherche de graine, et la couverture des autres vecteurs.

Un cas d'usage éthique et une attaque sont présentés dans la section 4.6.

Finalement, nous concluons dans la section 4.7.

4.2 Transformations utilisées

Dans cette section, nous décrivons les transformations que nous avons choisies pour ce chapitre. Contrairement au premier scénario, où les graines sont déjà fixées et pour lesquelles on génère les matrices de projection en nombre limité, ce second scénario nécessite l'essai de nombreuses graines, et donc la génération de nombreuses matrices de projection depuis ces graines. Le biohashing avec l'orthogonalisation de la matrice avec l'algorithme de Gram-Schmidt rend cette étape coûteuse, et nous avons cherché à limiter ce coût. Cette section introduit nos nouvelles transformations, et analyse leurs propriétés. En effet, dans ce second scénario, nous nous permettons de choisir la graine et la transformation pour la création du passe-partout.

4.2.1 Lemme de Johnson-Lindenstauss

Nous analysons un schéma, basé sur la projection aléatoire avec le lemme de Johnson-Lindenstauss [Johnson and Lindenstrauss, 1984]. Leur performance est analysée dans [Dasgupta and Gupta, 2003]. Il est établi que pour tout $0 < \epsilon < 1$, il existe une fonction $f : \mathbb{R}^N \to \mathbb{R}^M$ telle que pour tout x, y d'un sous-ensemble de n points de $\mathbb{R}^N$ avec $M \geq O(\epsilon - 2 \log n)$

$$(1 - \epsilon) \parallel x - y \parallel_2^2 \leq \parallel f(x) - f(y) \parallel_2^2 \leq (1 + \epsilon) \parallel x - y \parallel_2^2 .$$

Dans la partie suivante, nous proposons deux projections analysées avec ce lemme.

4.2.2 Projections proposées par Achlioptas

Deux projections aléatoires proposées par [Achlioptas, 2003] sont utilisées en alternative à Gram-Schmidt. Ce sont deux matrices $N \times M$.

Dans le premier cas, noté JL1, les coefficients de la matrice de projection M_s sont

$$\begin{cases} 1/\sqrt{M} \text{ avec probabilité } 1/2 \\ -1/\sqrt{M} \text{ avec probabilité } 1/2 \end{cases}$$

Dans le second cas, noté JL2, les coefficients de la matrice de projection M_s sont

$$\begin{cases} \sqrt{3/M} \text{ avec probabilité } 1/6 \\ 0 \text{ avec probabilité } 2/3 \\ -\sqrt{3/M} \text{ avec probabilité } 1/6 \end{cases}$$

4.2.3 Gain de performance sur le coût de génération des matrices

La rapidité de génération de ces matrices de projection est le principal avantage de ces méthodes en comparaison de l'algorithme de Gram-Schmidt.
Le gain de temps de construction de matrice de projection entre la méthode du biohashing avec une matrice aléatoire orthogonalisée par Gram-Schmidt et les deux matrices proposées est décrit dans le tableau 4.1.

Nous y indiquons cette accélération en termes de cycle d'horloge processeur avec deux tailles de vecteurs de caractéristiques, correspondant à celles des bases de données biométriques utilisées. Nous avons effectué dix-mille générations de matrice $N \times M$, pour chaque méthode et pour chaque taille de vecteur N. M est la taille des gabarits, fixée à 128.

Taille des vecteurs	$\frac{GS}{JL1}$	$\frac{GS}{JL2}$
512	33	25
990	31	26

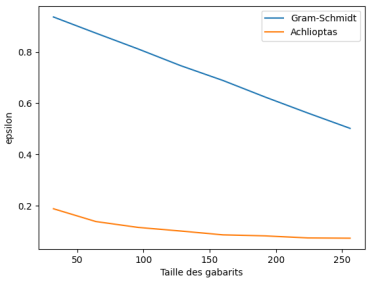
TABLE 4.1 – Accélération entre Gram-Schmidt et les 2 projections proposées

L'utilisation de ces projections nous permet d'aller environ 30 fois plus vite pour l'étape de génération de matrice depuis une graine.

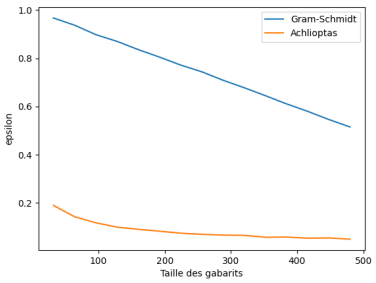
4.2.4 Analyse de ces projections

La valeur d'ϵ de l'inégalité du lemme de Johnson-Lindenstauss décrit en 4.2.1 est détaillée dans la figure 4.1, avec les trois bases, pour $N = 512$ pour FVC et LFW, $N = 990$ pour PTB, et M, la taille des gabarits, entre 32 et 256.

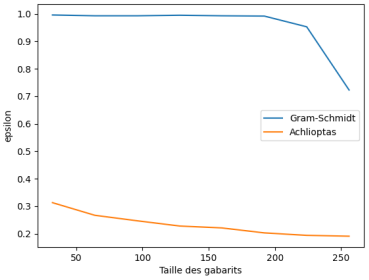
Deux transformations sont représentées : une courbe pour la fonction générant une matrice aléatoire orthonormalisée avec l'algorithme de Gram-Schmidt, où ϵ décroit de 1 à 0.5, et une courbe pour les deux transformations précédentes d'Achlioptas qui procurent des courbes similaires ($\epsilon \simeq 0.1$ pour $M \geq 128$). La valeur d'epsilon est en ordonnée et la taille des gabarits est en abscisse.



(a) FVC

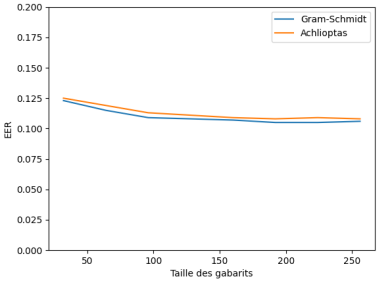


(b) PTB

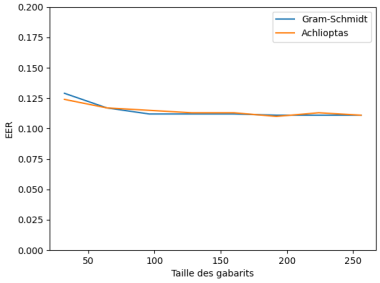


(c) LFW

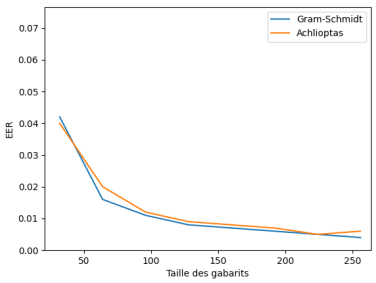
FIGURE 4.1 – Estimation d'ϵ pour différentes tailles de gabarits



(a) FVC



(b) PTB



(c) LFW

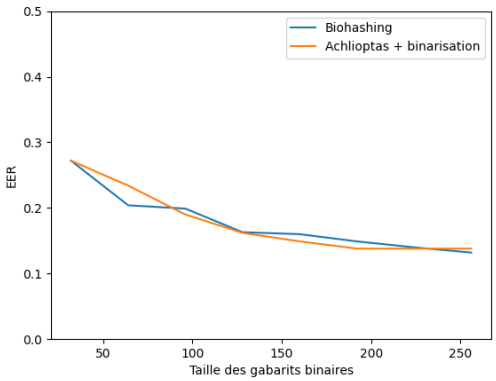
FIGURE 4.2 – EER (distance euclidienne)

Les précédentes expériences pour ϵ impactent peu l'EER des bases de de données biométriques révocables qui reste proche des bases de données biométriques originales, comme montré dans la figure 4.2, avec $M \geq 128$.

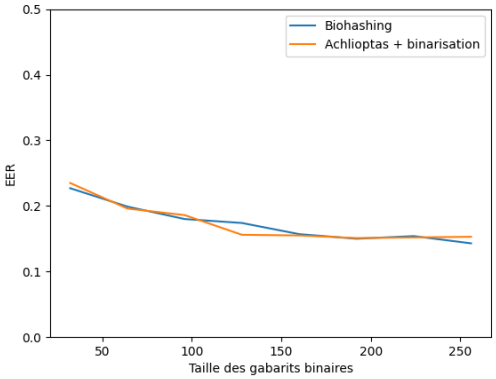
L'EER de ces transformations dans l'espace euclidien ne change pas significativement l'EER par rapport aux bases de données biométriques originales, si M n'est pas trop petit. Pour $M = 128$, l'EER est d'environ 11% contre 10% pour FVC et PTB, environ 1% contre 0.2% pour LFW.

Néanmoins, le temps de calcul pour la recherche de graine n'est pas négligeable dans l'espace euclidien, particulièrement pour certains vecteurs de caractéristiques. Nous allons donc étudier comment se comportent ces transformations avec une étape de binarisation en complément.

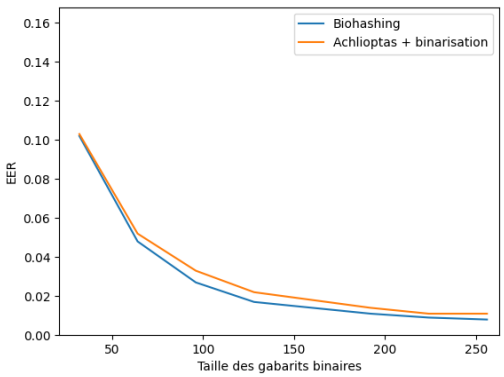
L'EER de la base de données biométriques révocables obtenue avec les deux transformations d'Achlioptas et l'étape de binarisation est légèrement augmenté par rapport à l'EER original : environ 16.5% pour FVC et 2.1% pour LFW, tel que présenté dans la figure 4.3, avec $M = 128$.



(a) FVC



(b) PTB



(c) LFW

FIGURE 4.3 – EER (distance de Hamming)

La recherche de passe-partout pour une base de données biométriques révocables est plus intéressante dans l'espace euclidien qui dégrade moins l'EER. L'espace de Hamming est le plus utilisé dans la suite des travaux pour des raisons de performance.

Nous fixons pour la suite des travaux $M = 128$. En effet, l'EER décroit peu pour $M \geq 128$, mais le temps de calcul pour la recherche de graines augmente lorsque l'on augmente la taille M des gabarits, la matrice à construire étant de taille $N \times M$.

Remarque 15. *Le second scénario décrit dans ce chapitre n'utilise pas la transformation biométrique comme une couche de sécurité, mais pour permettre la construction efficace de passe-partout.*
Si une transformation non sécurisée est utilisée pour obtenir cette propriété, alors il faut utiliser une autre couche de sécurité pour le stockage des gabarits.

Nous avons indiqué la valeur de la borne ϵ de l'inégalité du lemme de Johnson-Lindenstauss. Nous constatons que ces nouvelles projections obtiennent un EER similaire qu'en utilisant l'algorithme de Gram-Schmidt, et que cet EER ne réduit plus lorsque le gabarit est au minimum de taille $M = 128$, valeur que nous conservons pour la suite des travaux. Les transformations JL1 et JL2 sont utilisées du fait de leur rapidité à performance similaire.

Maintenant que les transformations utilisées dans les travaux de ce chapitre sont décrites, nous allons introduire notre second scénario de passe-partout et présenter les résultats des expériences.

4.3 Passe-partout : construction d'une base de données biométriques révocables

Dans cette partie, nous cherchons des graines pour construire la base de données biométriques révocables pour un passe-partout. Nous effectuons les expériences pour le second scénario tel que défini dans 4.1.1, avec les transformations $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^M$ ou $\mathcal{T} : \mathbb{R}^N \to \{0,1\}^M$, définies par Achlioptas, en terminant ou non par une étape de binarisation, avec $M = 128$. Les matrices aléatoires sont générées à partir de graines de 16 octets comme précédemment.

Une stratégie de force brute est utilisée pour trouver les graines permettant que le passe-partout couvre toute la base construite. Ce choix des graines ne doit pas dégrader les performances de la base de données biométriques révocables par rapport à l'usage de graines aléatoires.

4.3.1 Recherche de graines pour un vecteur de caractéristiques

Dans cette partie, nous transformons la base de données biométriques B en t bases. Soit $B_j = \{x_i^j\}_{i=1,\dots,n,j=1\dots,t}$ les t bases de données biométriques, composées de vecteurs de caractéristiques. Au final, nous avons n personnes avec t vecteurs par personne. Dans les expériences, un candidat passe-partout x_j^* est un vecteur de caractéristiques choisi parmi cette base B_j. Nous cherchons des graines afin de construire une base de données biométriques révocables conforme pour le candidat passe-partout x_j^*, tel que défini en 4.3.1.

Définition 4.3.1. *Soit x^* un vecteur de caractéristiques, τ le seuil, et $B = \{x_i\}_{i=1,\dots,n}$ une base de données biométrique. La base de données biométriques révocables D est dite conforme pour x^* avec τ s'il existe $s_1, \dots s_n$ tel que $D = \{\mathcal{T}(s_i, x_i)\}$ et x^* est un passe-partout pour D avec τ.*

Pour chaque base B_j, nous prenons un par un les $(n-1)$ j^{eme} vecteurs de caractéristiques de chaque i^{eme} personne. Pour chacun de ces vecteurs, noté x, nous cherchons une graine s, comme schématisé en 4.4, telle que

$$D(\mathcal{T}(s,x), \mathcal{T}(s,x_j^*)) \ \leq \ \tau$$

Les graines sont générées aléatoirement depuis $\{0,1\}^{128}$ jusqu'à ce que le candidat passe-partout soit authentifié avec succès.

Les expériences utilisent l'algorithme suivant 1 et donnent en sortie un tableau de $(n-1)$ graines pour chaque candidat passe-partout. On utilise le seuil $\tau = \tau@EER$.

La recherche de graine est limitée à c_{max} essais. Si une graine permettant l'authentification du candidat passe-partout n'est pas trouvée après c_{max} essais, on enregistre *Echec*. Dans la plupart des expériences, c_{max} n'est pas atteint.

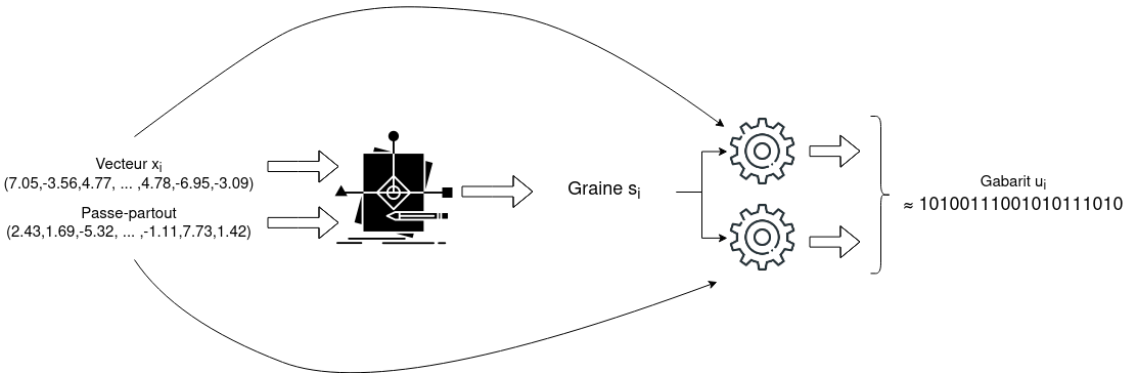


FIGURE 4.4 – Choix d'une graine pour un passe-partout

Algorithme 1 Recherche de graines pour un passe-partout

Nécessite : B_j les bases de données biométriques, τ le seuil, c_{max}
  SeedT = []
  pour $i = 1$ à n et $j = 1$ à t **faire**
    SeedL = []
    pour $k = 1$ à n, avec $k \neq i$ **faire**
      $cpt \leftarrow 0$, $s_k \leftarrow aleatoire(1, 2^{128})$
      tant que $\mathcal{V}(\mathcal{T}(s_k, x_i^j), \mathcal{T}(s_k, x_k^j), \tau) = Faux$ et $cpt < c_{max}$ **faire**
        $s_k \leftarrow aleatoire(1, 2^{128})$
      fin tant que
      si $\mathcal{V}(\mathcal{T}(s_k, x_i^j), \mathcal{T}(s_k, x_k^j), \tau) = Vrai$ **alors**
        SeedL.ajout(s_k)
      sinon
        SeedL.ajout($Echec$)
      fin si
    fin pour
    SeedT.ajout(SeedL)
  fin pour
  retourne *SeedT*

La recherche de graines permettant d'obtenir un gabarit usurpable par le passe-partout est un succès pour tous les vecteurs de caractéristiques des bases FVC, LFW et PTB, avec binarisation. Ces expériences ont été effectuées avec les deux transformations d'Achlioptas, JL1 et JL2, avec et sans étape de binarisation.

4.3.2 Résultats des expériences

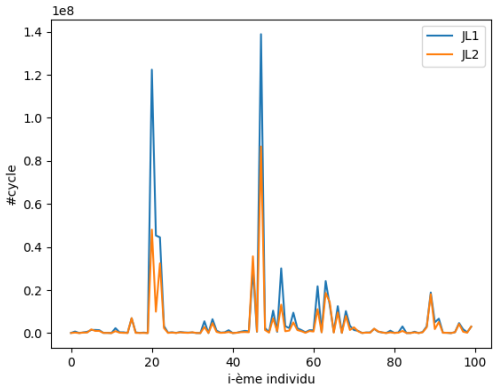
La figure 4.5 présente le temps nécessaire moyen exprimé en cycle d'horloge processeur pour trouver les graines satisfaisantes pour les t vecteurs de caractéristiques de chaque individu. On y trouve les résultats pour les deux transformations pour chaque base. L'abscisse représente le temps moyen pour trouver les graines pour les t vecteurs du i^{eme} individu. L'ordonnée représente le temps nécessaire.

On constate que pour les bases FVC et PTB, certains individus nécessitent un temps de recherche bien plus important que pour les autres individus, jusqu'à 10^8 cycles d'horloge processeur, ce qui est nettement moins le cas pour la base LFW, avec des durées autour de 10^6 cycles d'horloge processeur.

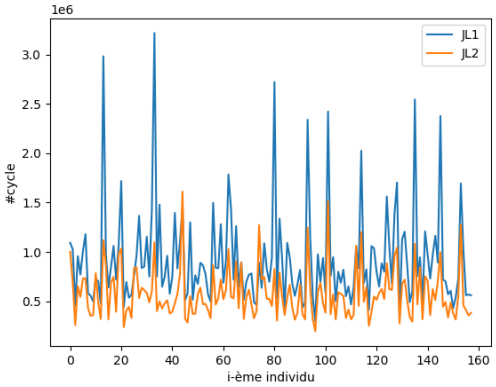
On remarque aussi que la seconde transformation JL2 obtient des résultats un peu plus rapides que JL1.

La recherche de graines est globalement très efficace. La recherche d'une graine s pour un vecteur x et un passe-partout x^*, satisfaisant $\mathcal{V}(\mathcal{T}(s, x), \mathcal{T}(s, x^*), \tau) = Vrai$, prend seulement quelques secondes pour la majorité des cas.

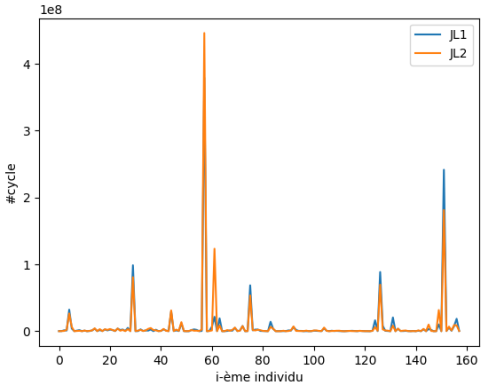
La moyenne des EER de ces $(n * t)$ bases de données biométriques révocables, pour lesquelles on choisit les graines, est autour de 17% pour FVC, 17.5% pour PTB, et 2.4% pour LFW. Ces EER moyens sont similaires pour les deux transformations JL1 et JL2. Ils sont proches des EER des bases de données biométriques révocables construites avec des graines aléatoires, à savoir 16.5% pour FVC, 17% pour PTB, et 2.1% pour LFW.



(a) FVC



(b) LFW



(c) PTB

FIGURE 4.5 – Nombre moyen de cycles d'horloge processeur (avec étape de binarisation)

Des expériences ont été effectuées sans finir avec une étape de binarisation, en restant dans l'espace euclidien. Le temps de recherche de graines augmente de manière significative pour les bases FVC et PTB, empêchant les expériences de se finir avec succès dans un temps raisonnable.

Dans le cas de la base LFW, la durée de recherche n'augmente pas significativement par rapport aux expériences avec étape de binarisation, et les expériences se finissent avec succès.

La figure 4.6 présente le nombre moyen, pour les t vecteurs de caractéristiques de chaque personne de la base LFW, de cycles d'horloge processeur nécessaires pour trouver une graine satisfaisante dans l'espace euclidien. La seconde transformation JL2 est significativement plus efficace que la première transformation JL1.

L'EER moyen de ces $(n*t)$ bases de données biométriques révocables construites avec graines choisies est autour de 1.3%, ce qui est proche de l'EER de bases de données biométriques révocables construites avec graines aléatoires pour la base LFW qui est de 1% pour les deux transformations sans binarisation.

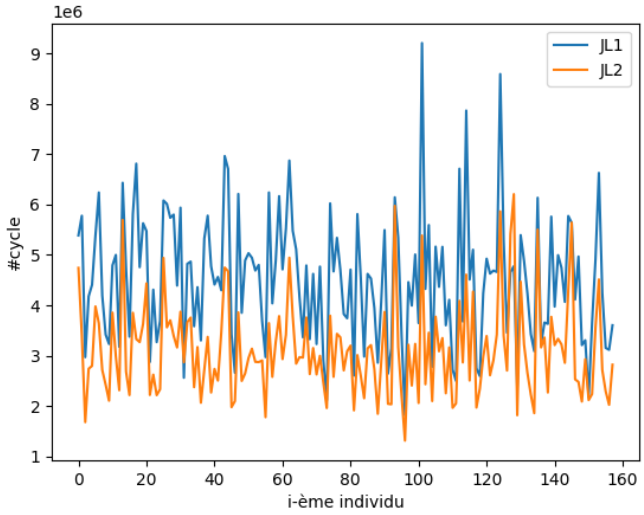


FIGURE 4.6 – Nombre moyen de cycles d'horloge processeur (LFW sans étape de binarisation)

Nous avons efficacement construit des bases de données biométriques révocables pour LFW dans l'espace euclidien, sans étape de binarisation, ainsi que pour les trois bases, avec étape de binarisation.

Néanmoins, la recherche de graine n'est pas efficace pour les bases FVC et PTB dans l'espace euclidien. Il semble pertinent d'utiliser les transformations avec étape de binarisation. Le temps étant variable d'une donnée à l'autre, une transformation sans étape de binarisation pourrait être acceptable avec de bonnes acquisitions.

4.4 Complexité des deux scénarios

Dans ce paragraphe, nous considérons la complexité théorique de trouver des passe-partout dans les deux scénarios.

On suppose que l'espace métrique $(\mathcal{M}_B, D_B)$ est $\{0,1\}^M$ avec la distance de Hamming, et la transformation biométrique $\mathcal{T}$ est pseudoaléatoire. Ayant un gabarit u, un vecteur de caractéristiques x et un seuil τ, la probabilité p telle que la distance $D_B(\mathcal{T}(x,s),u) < \tau$ pour une graine aléatoire s est $p = \frac{1}{2^M}\sum_{i=0}^{\tau-1}\binom{M}{i}$.

Complexité du premier scénario : construire un passe-partout pour une base de données biométriques révocables

Soit $D = \{u_i\}_{i=1,\dots,n}$ une base de données biométriques révocables avec n gabarits binaires. Le vecteur de caractéristiques x, tel que D est ϵ-couvert par x, est espéré après $1/p^{\epsilon n}$ essais.

Complexité du second scénario : construire une base de données biométriques révocables pour un passe-partout

Soit $B = \{x_i\}_{i=1,\dots,n}$ une base de données biométriques avec n vecteurs de caractéristiques et x un vecteur de caractéristiques. Un ensemble de n graines $s_1, \dots s_n$, tel que la base de données biométriques révocables D, construite depuis ces graines, est conforme à x tel que défini en 4.3.1, est espéré après n/p essais.

Remarque 16. *La construction d'une base de données biométriques révocables pour un passe-partout est clairement plus facile que la construction d'un passe-partout pour une base de données biométriques révocables.*

On constate que le second scénario est significativement plus rapide que le premier scénario.

Nous décrivons dans la prochaine section une extension de ce second scénario dans laquelle nous avons un individu passe-partout, composé de multiples vecteurs de caractéristiques passe-partout.

4.5 Extension à un individu passe-partout

Dans cette section, nous présentons une extension du scénario précédemment décrit, dans laquelle nous construisons une base de données biométriques révocables en choisissant les graines. Nous avons décrit la construction d'une base de données biométriques révocables en choisissant les graines pour qu'un passe-partout connu couvre toute cette base.

Le passe-partout est un vecteur de caractéristiques issu d'un individu. Nous étudions ici la performance de couverture d'autres vecteurs issus du même individu pour la base construite.

Nous formalisons dans la définition 4.5.1 la notion d'individu passe-partout.

Définition 4.5.1. *Soit $B = \{x_i\}_{i=1,\dots,n}$ une base de données biométriques telle que définie dans 2.3.2. Soit $V^* = \{x_1^*, \dots, x_t^*\}$ des vecteurs de caractéristiques fixés issus d'une unique modalité d'un même individu V^*. Soit $D = \{(u_i, s_i)\}_{i=1,\dots,n}$ la base de données biométriques révocables, avec $u_i = \mathcal{T}(s_i, x_i)$ et les graines $S = \{s_i\}_{i=1,\dots,n}$ choisies telles que décrites dans 4.1.1 avec x_1^* le passe-partout. V^* est un individu passe-partout pour D si $\forall\ (u, s)\ \in\ D$, $\forall\ x^*\ \in\ V^*$ on a $V(u, \mathcal{T}(s, x^*), \tau) = Vrai$.*

Remarque 17. *Ces vecteurs $x_1^*, \dots x_t^*$ sont issus du même individu, ce qui n'assure pas que les gabarits issus de ces vecteurs aient une distance deux à deux inférieure au seuil.*

Dans le second scénario précédemment étudié, le candidat passe-partout pour lequel on cherche des graines est un unique vecteur de caractéristiques x^*.

Pour travailler dans cette extension du second scénario, nous avons besoin de plusieurs vecteurs de caractéristiques par individu. En effet, nous allons aussi utiliser T vecteurs de caractéristiques d'un même individu, $\{x_1^*, \dots, x_T^*\}$, parmi ces t vecteurs pour choisir les graines, au lieu d'un seul.

Une telle base de données biométriques est décrite en 4.5.2

Définition 4.5.2. *Soit $B = \{x_i^j\}_{i=1,\dots,n, j=1\dots,t}$ une base de données biométriques. Cette base est composée de t vecteurs de caractéristiques pour n individus.*

4.5.1 Recherche de graines pour un ensemble de vecteurs de caractéristiques

Dans cette extension du second scénario, nous avons une base de données biométriques $B = \{x_i^j\}_{i=1,\dots,n, j=1\dots,t}$. Nous notons les vecteurs d'un candidat individu passe-partout $V^* = \{x_1^*, \dots, x_t^*\}\ \subset\ B$, les vecteurs de caractéristiques d'un i^{eme}

individu. Soit deux ensembles, $V_{recherche} = x_1^*, \ldots, x_T^*$ l'ensemble de recherche, et $V_{test} = x_{T+1}^*, \ldots, x_t^*$, l'ensemble de test, avec $V_{recherche} \cap V_{test} = \emptyset$ et $V_{recherche} \cup V_{test} = V^*$.

Dans un premier temps, nous recherchons des graines avec la connaissance des T vecteurs de l'ensemble de recherche : $\forall\ k \in \{1, \ldots, t\}$, nous avons la sous-base $B_k = \{x_i^j\}_{i=1,\ldots,n, j=k}$, $\forall\ x \in B_k$, nous cherchons une graine s telle que $\forall\ x^* \in V_{recherche}$ on ait $\mathcal{V}(\mathcal{T}(s,x), \mathcal{T}(s,x^*), \tau) = Vrai$. Si nous n'y arrivons pas en c_{max} essais, on conserve la graine minimisant la pire distance.

Dans un second temps, nous regardons la couverture de la base de données biométriques révocables D_k, construite avec les graines choisies depuis les T vecteurs de $V_{recherche}$, par les $t - T$ vecteurs de V_{test}.

Les gabarits correspondants de $x_1^*, \ldots, x_t^*$, notés $u_1^*, \ldots, u_t^*$, forment les gabarits de l'ensemble de recherche et de l'ensemble de test.

4.5.2 Résultats des expériences pour l'ensemble de recherche et pour l'ensemble de test

Les expériences ont été faites uniquement avec la seconde transformation JL2, permettant les meilleures performances. Elles ont été effectuées avec et sans binarisation pour la base LFW, et uniquement avec binarisation pour les bases FVC et PTB, avec $T = 1$ et $T = 4$.

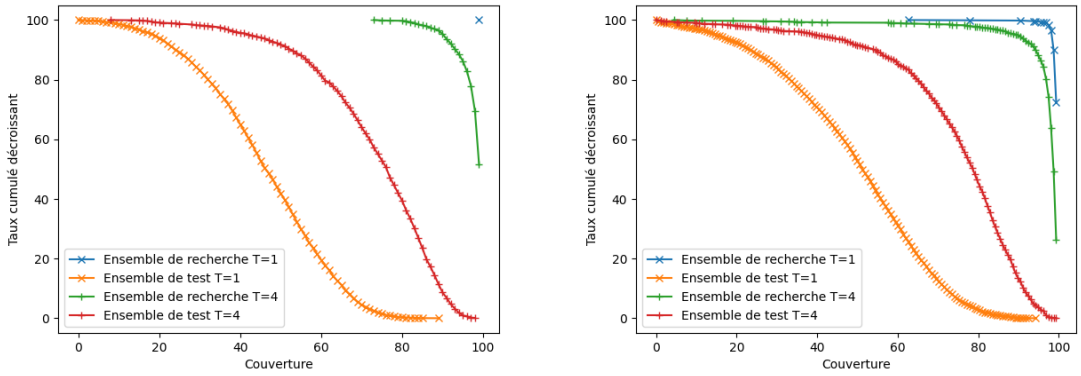
Remarque 18. *Dans le cas de $T = 1$, nous sommes dans la version originale du second scénario, et nous évaluons simplement les performances de l'ensemble de test. Dans le cas de $T = 4$, nous avons dû rechercher des graines pour cet ensemble de recherche, avant d'évaluer les performances de l'ensemble de test.*

Pour des raisons pratiques, nous utilisons une limite de temps de recherche par graine de 5 minutes au lieu de c_{max} essais. Les taux de couverture de l'ensemble de recherche et de l'ensemble de test pour $T = 1$ et $T = 4$ avec binarisation sont présentés dans les figures 4.7 pour FVC et PTB, sous forme de courbes cumulées décroissantes. Les figures 4.8 présentent les résultats pour LFW avec et sans binarisation.

Par exemple, avec $T = 1$ pour FVC, environ 60% des gabarits de l'ensemble de test usurpent au moins 40% des gabarits de la base de données biométriques révocables construite avec les graines choisies. Lorsque $T = 4$, 60% des gabarits de l'ensemble de test usurpent cette fois au moins 75% des gabarits de la base.

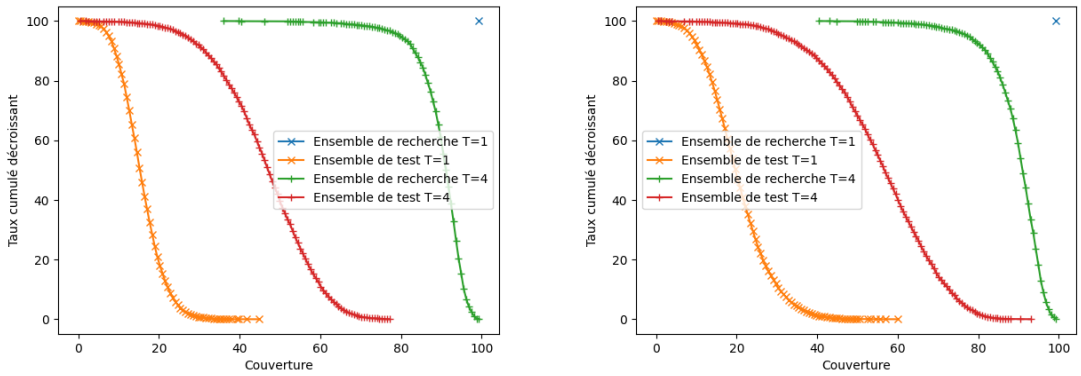
Remarque 19. *Les courbes qui ne se résument qu'à un point (bleu dans ces figures) signifient que 100% des gabarits de l'ensemble de recherche couvrent 100% de la base.*

Tous les gabarits de l'ensemble de recherche pour PTB avec $T = 1$ ne couvrent pas toute la base à cause de la limite de 5 minutes par recherche d'une graine.



(a) FVC (b) PTB

FIGURE 4.7 – Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour FVC et PTB avec binarisation



(a) Binaire (b) Euclidien

FIGURE 4.8 – Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour LFW avec et sans binarisation

Les résultats montrent que si les graines sont choisies en considérant un espace de recherche de 4 vecteurs de caractéristiques, la moyenne de couverture des vecteurs de caractéristiques de l'ensemble de test s'améliore de manière importante.

En effet, si la graine est choisie en ne considérant qu'un seul vecteur de caractéristiques en ensemble de recherche, la couverture moyenne de l'ensemble de test se dégrade fortement par rapport à la couverture de l'ensemble de recherche.

Sur les courbes, cela se constate par un glissement de la courbe représentant la couverture moyenne de l'ensemble de test vers la droite avec $T = 4$ par rapport à $T = 1$, indiquant une augmentation de cette couverture moyenne.

Un utilisateur souhaitant être un individu passe-partout a donc intérêt à procéder à de multiples captures pour la recherche de graines, s'il veut ensuite utiliser de nouvelles captures ayant de bonnes couvertures sur la base.

Dans la sous-section suivante, nous étudions si l'efficacité de la recherche de graine avec $V_{recherche}$ fournit un passe-partout plus performant avec V_{test}.

4.5.3 Corrélation entre ces ensembles

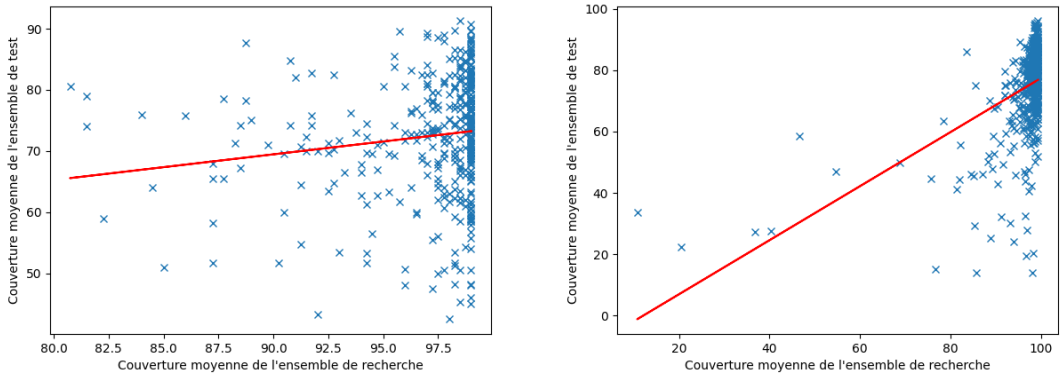
Nous étudions la corrélation entre la performance de couverture moyenne de l'ensemble de recherche et la performance de couverture moyenne de l'ensemble de test.

Les résultats sont présentés dans les figures 4.9 et 4.10. Ces résultats sont donnés pour les expériences avec $T = 4$. Pour $T = 1$, cette notion de corrélation n'a pas de sens, la couverture de $V_{recherche}$ étant de 100% pour FVC et LFW, et de légèrement moins pour PTB.

Un point correspond à une expérience où l'on a cherché des graines pour un candidat individu passe-partout. L'abscisse correspond à la couverture moyenne des vecteurs de caractéristiques de l'ensemble de recherche. L'ordonnée correspond à la couverture moyenne des vecteurs de caractéristiques de l'ensemble de test.

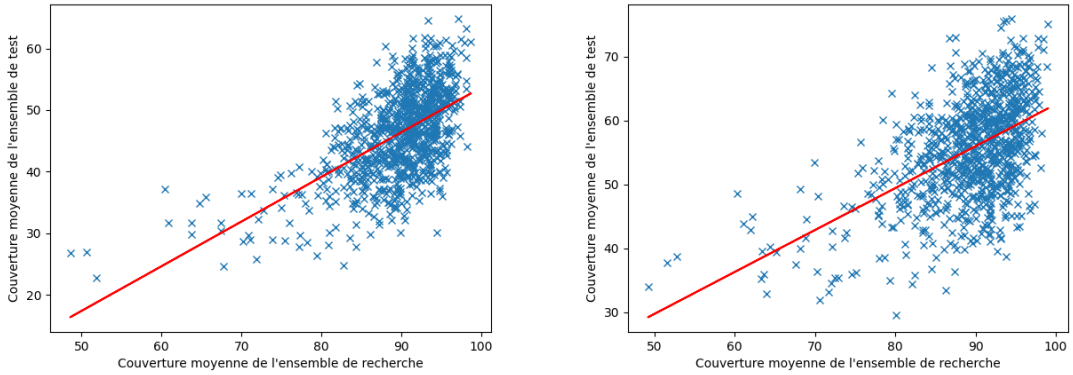
Les courbes rouges correspondent à une régression linéaire des points. La régression n'est pas exploitable pour FVC et PTB, le taux de couverture de l'ensemble de recherche étant à 100% de la base dans de nombreuses expériences, ce qui place une partie importante des points à droite des figures.

Pour la base LFW, on constate une corrélation importante entre les deux indicateurs de couverture.



(a) FVC (b) PTB

FIGURE 4.9 – Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour FVC et PTB avec binarisation



(a) Binaire (b) Euclidien

FIGURE 4.10 – Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour LFW avec et sans binarisation

Finalement, il faut que $V_{recherche}$ soit composé de vecteurs de caractéristiques ayant de bonne performance de couverture pour optimiser les performances de couverture de V_{test}.

Maintenant que l'extension du second scénario a été décrite, et sa faisabilité démontrée, nous allons décrire des cas d'usage.

4.6 Cas d'usage

Nous allons décrire deux cas d'usage de ce second scénario.

Le choix de graines pour la construction d'une base de données biométriques révocables permet que quelqu'un ait le même circuit d'accès que les utilisateurs sans ajouter un autre canal d'authentification. Ce cas d'usage n'est pas une attaque, mais une démarche éthique. C'est un nouveau paradigme d'authentification biométrique avec la création de bases de données biométriques révocables procurant des droits d'accès spécifiques. Par exemple, dans le cadre de compte bancaire, un utilisateur A peut donner procuration à un individu B qui peut s'authentifier comme A. De manière similaire, le directeur d'établissement bancaire peut s'authentifier comme chacun de ses clients. Dans un autre domaine, comme le contrôle d'accès par empreinte à un domicile, l'habitant peut autoriser le voisin à s'authentifier comme lui pour débloquer l'accès.

Dans un second cas d'usage, un attaquant actif dans la phase d'enregistrement peut construire une porte dérobée dans le système d'authentification, sans ajouter de code suspicieux ou rendre la base de données biométriques révocables suspectes, telle que définie en 4.6.1. Par exemple, à nouveau dans le cadre bancaire, un salarié corrompu peut offrir une clé d'accès physique contenant la graine, comme gage de sécurité, alors qu'il aurait choisi la graine pour lui permettre d'être authentifié sans autorisation.

Définition 4.6.1. *Soit $B = \{x_i\}_{i=1,\dots,n}$ une base données biométriques et $D = \{\mathcal{T}(s_i, x_i)\}$ une base de données biométriques révocables conforme pour le passe-partout x^*. x^* est une porte dérobée s'il n'existe pas d'algorithme polynomial distinguant $s_1, \mathcal{T}(x_1, s_1), \dots s_n, \mathcal{T}(x_n, s_n)$ de $s'_1, \mathcal{T}(x_1, s'_1), \dots s'_n, \mathcal{T}(x_n, s'_n)$, où $s'_1, \dots s'_n$ sont des graines générées aléatoirement avec x^* non connu et $s_1, \dots s_n$ sont des graines choisies pour x^* connu.*

4.7 Conclusion

Dans ce chapitre, nous avons proposé un nouveau concept de passe-partout. Nous utilisons la graine paramétrant la projection d'un vecteur de caractéristiques pour obtenir une nouvelle propriété.

Le premier scénario de passe-partout décrit dans le chapitre précédent proposait une attaque construisant un vecteur de caractéristiques dit passe-partout pour une base de données biométriques révocables déjà existante.

Dans ce second scénario, nous fixons le vecteur de caractéristiques x^* pour choisir les graines en connaissance de x^*, afin de projeter la base de données biométriques en une base de données biométriques révocables couverte par x^*.

Nous avons introduit plusieurs transformations dans la section 4.2. Nous constatons que ces deux projections permettent une génération de matrice de projection environ 30 fois plus performante.

La section 4.3 décrit le second scénario, avec la construction d'une base de données biométriques révocables en choisissant les graines pour un vecteur de caractéristiques. Les résultats détaillés des expériences sont présentés dans la section 4.3.2. Nous réussissons à construire des bases de données biométriques révocables, à gabarit binaire, complètement couvertes par le vecteur de caractéristiques choisi. Pour la base LFW, nous avons effectué ces expériences avec succès dans l'espace euclidien.

La section 4.4 décrit les complexités théoriques des deux scénarios. Le second scénario est bien moins complexe que le premier scénario.

La section 4.5 présente une extension de ce second scénario. Nous analysons la couverture de futures captures de l'individu passe-partout. L'objectif est que de futures captures de cet individu continuent à couvrir la base. Pour atteindre cet objectif, il faut plusieurs vecteurs de caractéristiques issus d'un même individu pour rechercher des graines. De bonnes couvertures des vecteurs de l'ensemble de recherche sont essentielles, comme le montre les corrélations décrites en 4.5.3.

Nous présentons dans la dernière section 4.6 deux cas d'usage de ce second scénario.

Chapitre 5

Perspectives

Résumé : *Ce chapitre revient sur les deux scénarios du passe-partout. Nous étudions la réutilisabilité d'un passe-partout du premier scénario sur d'autres données biométriques. Nous avons expérimenté deux nouvelles bases, une multimodale et l'autre issue d'oreilles, pour construire des passe-partout du second scénario. Enfin, nous donnons les perspectives de nos contributions.*

Mots-clés : *réutilisabilité, multimodalité, oreille.*

5.1 Introduction

Dans ce chapitre, nous présentons des résultats complémentaires et explorons des perspectives de recherche aux précédents travaux décrits dans ce manuscrit.

Dans une première section 5.2, nous étudions la réutilisabilité des passe-partout de notre premier scénario. En effet, les passe-partout sont construits avec la connaissance de toute la base de données biométriques révocables, et leur couverture est évaluée sur cette même base. Afin d'avoir des indicateurs correspondant mieux à un cas d'usage réaliste, nous avons construit un passe-partout à partir de données biométriques révocables connues, par exemple issues d'une intrusion, et nous regardons la performance de couverture sur d'autres données biométriques révocables.

Dans une deuxième section 5.3, nous étudions les performances des passe-partout de notre second scénario sur d'autres bases : une première base multimodale issue d'une fusion, et une seconde base originale issue d'images d'oreilles.

Dans une dernière section 5.4, nous résumons ces nouvelles expériences avant de conclure avec les perspectives offertes par ces travaux.

5.2 Passe-partout réutilisable

5.2.1 Contexte

Dans cette section, nous revenons sur la construction de passe-partout avec un algorithme génétique depuis une base de données biométriques révocables, telle que décrite en 3.6.2. Dans ce premier scénario, nous construisons le passe-partout depuis une base de données biométriques révocables et nous constatons la couverture du passe-partout pour cette même base. Si l'on souhaite effectuer une attaque dans ce scénario, cela implique d'avoir pu récupérer l'ensemble de la base, puis d'attaquer cette même base.

Nous avons souhaité construire un passe-partout depuis une sous-partie de la base de données biométriques révocables (la première moitié des individus), puis tester la couverture de ce passe-partout pour l'autre sous-partie de la base (la seconde moitié des individus), non utilisée pour la construction du passe-partout. Un scénario d'attaque implique désormais de ne connaître qu'une sous-partie de la base pour construire le passe-partout, puis on évalue l'efficacité de l'attaque, à savoir la couverture du passe-partout, sur une sous-partie inconnue de la base.

5.2.2 Résultats

Dans ces expériences, les bases de données biométriques révocables sont générées depuis des graines aléatoires. Les passe-partout sont construits avec un algorithme génétique paramétré comme dans 3.6.2. Le seuil utilisé est $\tau = \tau@EER$.

Le tableau 5.1 présente le taux de couverture optimale obtenu avec chaque base pour la seconde sous-partie. On constate une couverture importante de cette seconde sous-partie, malgré qu'elle ne soit pas utilisée pour la construction du passe-partout. Pour la base LFW, on couvre encore quelques individus.

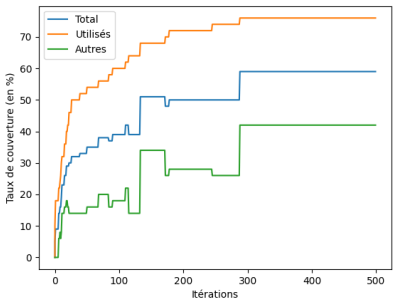
Les figures 5.1 présentent l'évolution de la couverture du passe-partout pour 3 ensembles. Le premier ensemble est la sous-base utilisée pour la construction du passe-partout, il obtient les meilleures couvertures, et le taux de couverture est croissant. Cet ensemble est noté *Utilisés*. Le deuxième ensemble est la sous-base non utilisée pour la construction du passe-partout, il obtient une couverture moindre. Ce taux de couverture est calculé au cours de l'exécution de l'algorithme génétique, mais cet ensemble n'est pas utilisé pour la construction du passe-partout, et donc pour le calcul de score. Cet ensemble est noté *Autres*. Le troisième ensemble est la

base complète, composée des deux sous-bases précédentes. Cet ensemble est noté *Total*. L'évolution de la couverture du deuxième ensemble et du troisième ensemble n'est pas croissante, car non prise en compte par l'algorithme génétique, qui n'optimise donc pas ces indicateurs.

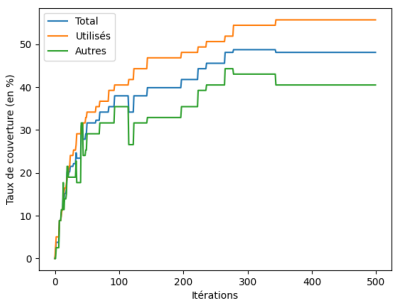
Il est à noter que le taux de couverture est en pourcentage de l'ensemble. Si la base complète est de taille N pair, les deux sous-bases sont de taille $N/2$.

Base	TCO (%)
FVC	42
LFW	6.3
PTB	44.3

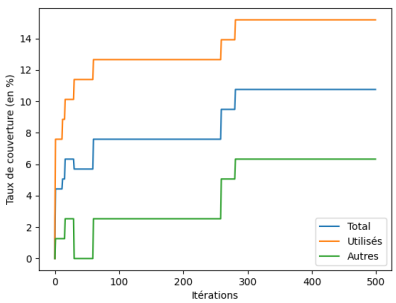
TABLE 5.1 – Taux de couverture optimale pour les données biométriques révocables non utilisées



(a) FVC



(b) PTB



(c) LFW

FIGURE 5.1 – Évolution du taux de couverture du passe-partout

5.2.3 Conclusion

Nous avons démontré que la construction de passe-partout depuis une base de données biométriques révocables permet d'obtenir une bonne couverture pour une autre base de données biométriques révocables issues de captures de modalités biométriques du même type, ici empreintes digitales, visages, et ECG. Pour ce faire, nous avons divisé en deux parts égales les bases à notre disposition. La couverture de la sous-base non utilisée pour la construction du passe-partout est plus faible que pour la sous-base utilisée, mais reste significative.

5.3 Passe-partout sur d'autres bases

Dans cette section, nous avons expérimenté le second scénario de passe-partout, décrit en 4.3, avec l'extension à l'individu passe-partout, décrite en 4.5, sur d'autres bases de données biométriques. Ces nouvelles expérimentations ont produit des performances distinctes des 3 bases de données biométriques présentées en 2.5. Nous présentons en 5.3.1 une nouvelle base de données biométriques multimodale issue d'une fusion entre les bases FVC et LFW. Dans la section 5.3.2, nous présentons une base de données biométriques issue d'une nouvelle modalité biométrique : les oreilles.

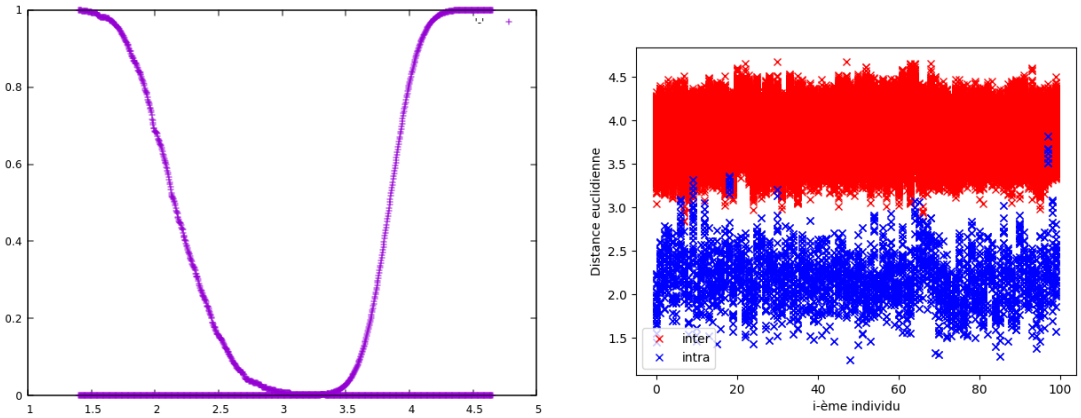
5.3.1 Base multimodale

Nous souhaitons étudier le comportement de passe-partout sur des bases de données biométriques multimodales. Les schémas de biométrie multimodale ont été décrits dans la section 2.1.1. Nous allons utiliser la fusion de caractéristiques. En effet, nos passe-partout sont appliqués au niveau des vecteurs de caractéristiques. Dans nos travaux, la fusion de caractéristiques consiste à concaténer des vecteurs de caractéristiques issus de captures de différentes modalités biométriques. L'ordre de grandeur des valeurs composant un vecteur de caractéristiques dépendant de l'algorithme d'extraction, nous normalisons chaque valeur des vecteurs de caractéristiques originaux entre 0 et 1.

Dans cette section, nous avons effectué une fusion d'empreintes digitales issues de la base FVC et de visages issus de la base LFW. La base LFW étant plus grande que la base FVC, nous avons utilisé uniquement les 100 premiers individus et les 8 premières captures par individu. Nous avons ainsi 2 bases de tailles identiques, et nous avons construit une base fusionnée constituée de 100 individus avec 8 vecteurs chacun. Chaque vecteur est composé de la concaténation entre le vecteur de la base

FVC avec les valeurs normalisées et le vecteur de la base LFW avec les valeurs normalisées. Les indicateurs de cette base sont donnés dans la figure 5.2.

Chaque vecteur de caractéristiques est composé de $N = 1024$ valeurs réelles. L'EER de la base de données biométriques est 0.28% avec un seuil $\tau_A = 3.32$.
La base de données biométriques révocables avec binarisation a un EER d'environ 17% avec un seuil $\tau_B = 11$, et environ 0.7% avec un seuil $\tau_A = 3.12$ sans binarisation.



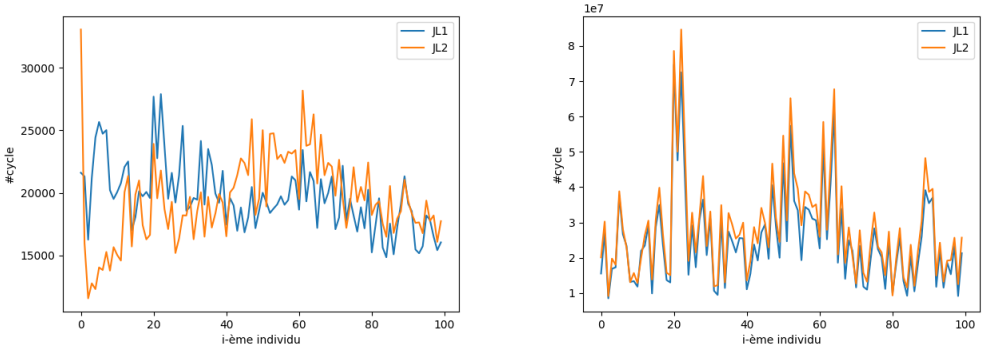
(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 5.2 – Indicateurs de performance de la base fusionnée sans transformation

La recherche de graines pour cette base a été très efficace en temps et en performance. On constate dans la figure 5.3 que les durées moyennes par individu pour trouver des graines permettant au passe-partout de s'authentifier avec les gabarits sont plus rapides que pour les autres bases.

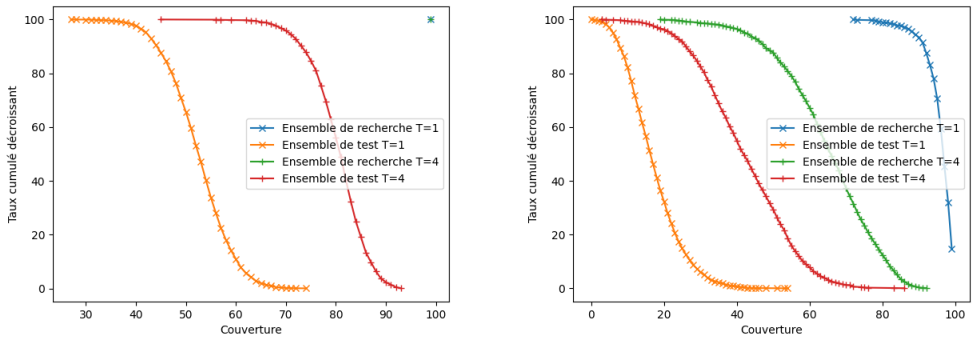
Nous constatons que même pour $T = 4$, les graines rapidement trouvées permettent au passe-partout de s'authentifier pour tous les gabarits utilisés dans l'ensemble de recherche. Cela s'observe dans la figure 5.4 par deux courbes qui ne sont qu'un point, pour $T = 1$ et pour $T = 4$. La recherche dans l'espace euclidien est aussi rapide et performante, mais ne permet pas d'obtenir une totale couverture. Les performances de couverture de l'ensemble de test sont significativement améliorées avec $T = 4$, et particulièrement dans l'espace binaire.

La figure 5.5 représente la corrélation entre la performance de couverture de l'ensemble de recherche et de l'ensemble de test. On ne peut pas constater de corrélation dans l'espace binaire, car la couverture de l'espace de recherche est totale même pour $T = 4$. Ainsi, chaque vecteur de l'espace de recherche couvre les 99 autres individus de la base. Pour l'espace euclidien, on constate une forte corrélation de performance de couverture de ces deux ensembles.



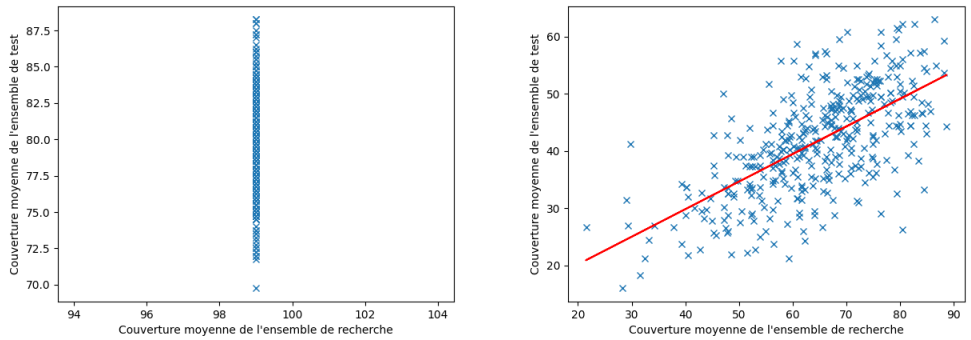
(a) Binaire (b) Euclidien

FIGURE 5.3 – Nombre moyen de cycles d'horloge processeur pour la base fusionnée



(a) Binaire (b) Euclidien

FIGURE 5.4 – Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour la base fusionnée



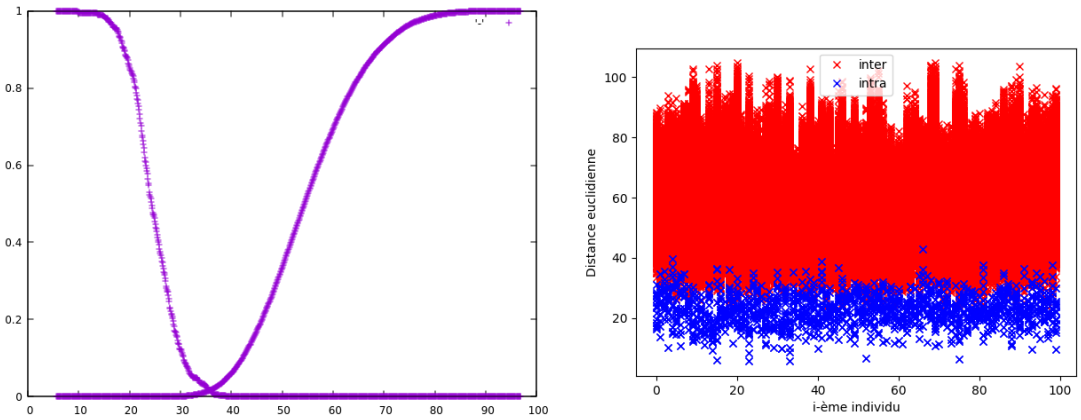
(a) Binaire (b) Euclidien

FIGURE 5.5 – Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour la base fusionnée

5.3.2 Base issue d'oreilles

Nous avons souhaité valider notre concept de passe-partout avec une modalité biométrique moins utilisée. Cette base de données est issue d'images d'oreilles de la base de captures de modalité biométrique AMI [Gonzalez et al., 2012]. Nous disposons de $t = 8$ captures de modalité biométrique pour $n = 100$ individus. Les vecteurs de caractéristiques ont été obtenus grâce au réseau profond AlexNet proposé par [Krizhevsky et al., 2012], et sont issus de la 23ème couche sur les 25 couches du réseau. Chaque vecteur de caractéristiques de cette base est composé de $N = 100$ valeurs réelles. On constate que les vecteurs de caractéristiques sont de plus petite taille que pour les autres bases. L'EER de la base de données biométriques est 1.7% avec un seuil $\tau_A = 35.9$.
La base de données biométriques révocables avec binarisation a un EER d'environ 3% avec un seuil $\tau_B = 40$. La base de données biométriques révocables sans binarisation a un EER d'environ 2% avec un seuil $\tau_A = 35.5$.

Les indicateurs de cette base sont donnés dans la figure 5.6.



(a) FNMR/FMR (b) Distance intraclasse et interclasse

FIGURE 5.6 – Indicateurs de performance de la base d'oreilles sans transformation

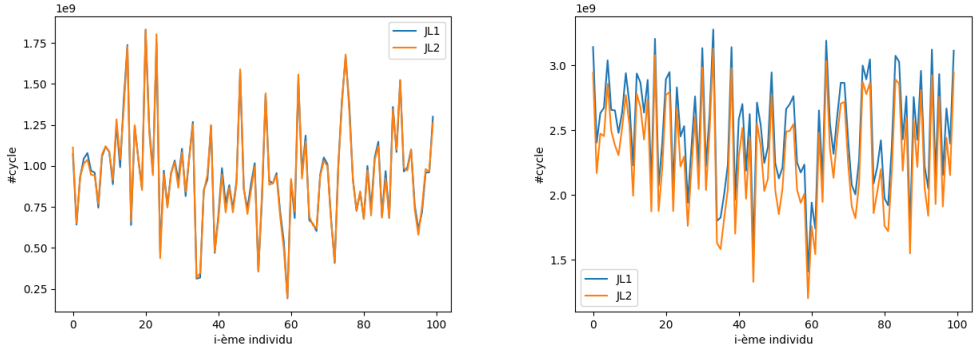
La recherche de graines pour cette base est plus longue que pour les autres bases comme on peut le constater dans la figure 5.7. Cela s'explique par les faibles performances de couverture. Pour chaque vecteur de caractéristiques, de nombreuses graines sont essayées. La limite de 5 minutes de recherche est régulièrement atteinte, car une graine satisfaisante n'est pas trouvée.

En effet, même pour $T = 1$, nous ne pouvons pas avoir un passe-partout couvrant l'intégralité des gabarits. Comme on peut le constater dans la figure 5.8, les passe-partout couvrent au moins 40% de la base, et dans la moitié de nos expériences

ils couvrent au moins 70% de la base. Dans l'espace euclidien, la moitié des passe-partout couvre plus de 30% de la base.

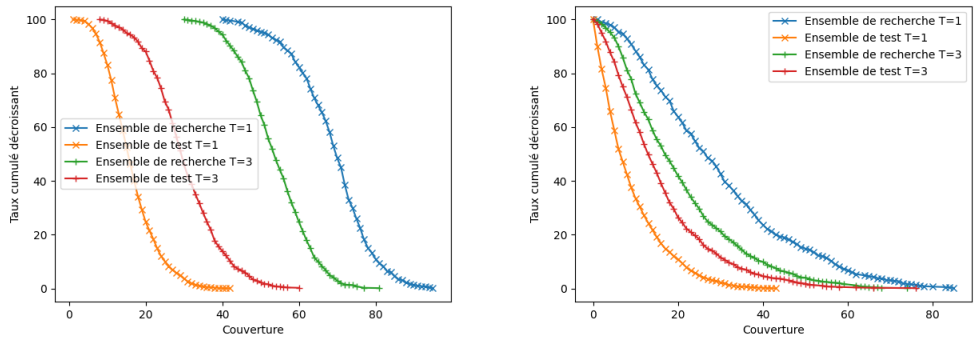
Même avec de moindres performances, les expériences valident l'amélioration de la couverture de l'ensemble de test avec $T = 3$. Nous utilisons cette fois $T = 3$ au lieu de $T = 4$ car pour cette base nous avons $t = 6$ captures par individu.

La figure 5.9 représente la forte corrélation constatée entre les performances de couverture de l'ensemble de recherche et de l'ensemble de test.



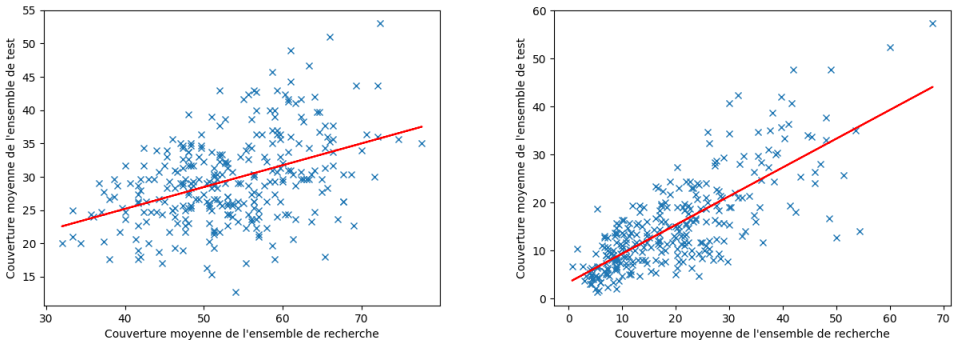
(a) Binaire (b) Euclidien

FIGURE 5.7 – Nombre moyen de cycles d'horloge processeur pour la base d'oreilles



(a) Binaire (b) Euclidien

FIGURE 5.8 – Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 3$ pour la base d'oreilles



(a) Binaire (b) Euclidien

FIGURE 5.9 – Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour la base d'oreilles

5.3.3 Conclusion

Dans cette section, nous avons répété les expériences du second scénario des passe-partout avec de nouvelles bases de données biométriques.

Dans un premier temps, nous avons construit une base de données biométriques issue de la fusion entre les bases FVC et LFW. Pour cette base, nous avons obtenu d'excellentes performances, particulièrement dans l'espace binaire. L'amélioration des performances d'un individu passe-partout en utilisant $T = 4$ au lieu de $T = 1$ a été validée, et nous avons constaté une importante corrélation entre les performances de couverture de l'ensemble de recherche et de l'ensemble de test dans l'espace euclidien.

Dans un second temps, nous avons utilisé une base de données biométriques issue d'une nouvelle modalité : des images d'oreilles. Les performances de la recherche de graines pour cette base sont moindres que pour les autres bases, mais permettent d'obtenir un passe-partout couvrant une partie non négligeable des individus. À nouveau, nous validons qu'il faut utiliser plusieurs vecteurs de caractéristiques pour le choix de graines, ici $T = 3$, ce qui permet d'augmenter significativement les performances de couverture d'autres vecteurs. Nous avons aussi constaté une importante corrélation entre les performances de couverture de l'ensemble de recherche et de l'ensemble de test dans l'espace binaire et dans l'espace euclidien, ce qui confirme la nécessité d'utiliser plusieurs captures de qualité pour rechercher des graines.

5.4 Conclusion et perspectives

Dans ce chapitre, nous avons expérimenté les passe-partout sous d'autres problématiques et avec d'autres données.

Dans la première section, nous sommes revenus sur le premier scénario où l'on construit un passe-partout pour une base de données biométriques révocables. Cette fois, nous avons construit le passe-partout depuis une sous-partie de cette base et nous avons testé ses performances sur l'autre sous-partie non utilisée pour la construction. Nous constatons que le passe-partout conserve une capacité de couverture importante pour des individus non pris en compte lors de sa construction. Il couvre encore plus de 40% des individus des bases d'empreintes digitales et de visages.

Dans la seconde section, nous avons étudié le second scénario où l'on construit une base de données biométriques révocables pour un passe-partout. Dans un premier temps, nous avons utilisé une base de données biométriques multimodale issue de la fusion entre les bases FVC et LFW. Les performances de couverture sont plus importantes qu'avec ces bases seules, et nous pouvons couvrir tous les gabarits de l'ensemble de recherche même pour $T = 4$. La corrélation de performances de couverture des deux ensembles est importante et l'utilisation de plusieurs vecteurs pour le choix de graines améliore significativement les performances de futurs vecteurs. Dans un second temps, nous avons utilisé une base avec pour modalité biométrique des images d'oreilles. Les performances de couverture sont moindres, mais la corrélation est à nouveau importante ainsi que l'amélioration des performances de couverture de l'ensemble de test avec $T = 3$. La taille des vecteurs de caractéristiques est plus de cinq fois plus petite que pour les autres bases.

Nous allons désormais donner des perspectives de ces travaux.

5.4.1 Préimage proche et réutilisable

Nous avons construit des préimages proches et réutilisables pour 2 gabarits binaires issus de transformations par biohashing. Les expériences ont été effectuées pour 3 bases de données biométriques avec succès pour 100% des couples de gabarits.

Nous souhaitons étudier l'existence théorique de PPR selon le seuil utilisé et les 2 gabarits, avec leur taille et la distance de Hamming les séparant. Nous pouvons ensuite estimer la difficulté de construction d'une PPR avec un seuil et un couple de gabarits, avant de proposer une contre-mesure à cette attaque. Nous pouvons agir sur les choix des graines afin d'obtenir des gabarits limitant l'efficacité d'une attaque

par construction de préimage proche et réutilisable. Ce choix peut ne pas limiter la construction d'une préimage proche, mais peut empêcher une réutilisabilité avec un autre gabarit.

Finalement, nous avons construit un vecteur de caractéristiques, et nous souhaitons évaluer la génération d'une capture de modalité biométrique synthétique dont l'extraction permet d'obtenir ce vecteur de caractéristiques. Cette étape supplémentaire permet ensuite une attaque par présentation, correspondant au premier point des attaques génériques définies en 2.1.3. Ces travaux peuvent ensuite être généralisés aux transformations basées sur les vecteurs de caractéristiques, sans se limiter au biohashing basé sur une projection paramétrée par une graine.

5.4.2 Construction d'un passe-partout pour une base de données biométriques révocables

Dans ce premier scénario, nous avons construit des passe-partout pour des bases de données biométriques révocables. Nous avons constaté des taux de couverture optimaux différents selon les bases, de même pour les tailles optimales de dictionnaire.

Les bases FVC et LFW8 sont de tailles identiques, mais ont des propriétés différentes, par exemple en termes d'EER. Nous souhaitons analyser l'impact des indicateurs de qualité d'une base de données biométriques révocables sur le TCO et la TOD.

Nous souhaitons étudier l'impact de la taille de la base de données biométriques révocables sur ces deux indicateurs, le TCO et la TOD, en déterminant s'il est plus facile de couvrir 70% d'une base de 100 individus que de couvrir 70% d'une base de 1000 individus. Cette propriété dépend notamment de la distribution des gabarits binaires de taille M dans $\{0,1\}^M$.

Comme pour les PPR, nous souhaitons déterminer des contre-mesures, comme le choix de graines pour limiter la couverture d'un passe-partout. À nouveau, la génération d'une capture de modalité biométrique synthétique depuis le vecteur de caractéristiques du passe-partout permet une attaque par présentation plus réalisable.

5.4.3 Construction d'une base de données biométriques révocables pour un passe-partout

Dans notre second scénario, nous avons choisi les graines pour qu'un passe-partout fixé couvre toute la base de données biométriques révocables. Nous avons

constaté, particulièrement pour les bases FVC et PTB, une importante difficulté à trouver une graine permettant l'authentification par le passe-partout pour certains vecteurs de caractéristiques. Cette difficulté ne s'explique pas par la distance euclidienne séparant le passe-partout et le vecteur de caractéristiques pour lequel on cherche une graine et nous souhaitons explorer d'autres pistes.

Nous avons introduit une extension portant sur les individus passe-partout dans laquelle nous utilisons plusieurs vecteurs de caractéristiques issus d'un individu pour choisir la graine. Nous souhaitons désormais utiliser plusieurs vecteurs de caractéristiques, mais issus d'individus distincts, pour leur permettre à tous d'usurper des utilisateurs.

Cette nouvelle propriété risquant de nécessiter encore plus d'essais de graines, nous souhaitons utiliser une autre méthode que la force brute pour le choix de graines. Il faut une nouvelle méthode de construction de matrice à partir d'une graine, permettant d'orienter la recherche et le choix de cette dernière.

5.4.4 Perspectives communes

Pour tous ces travaux, nous avons utilisé des données biométriques réelles, issues de captures de diverses modalités biométriques. Nous avons constaté des écarts de performances en fonction de la base utilisée ou des vecteurs de caractéristiques utilisés. Nous souhaitons détailler et lier les performances de nos travaux en fonction de différents critères, comme la taille des vecteurs de caractéristiques, la taille de la base de données biométriques, ou les performances de la base de données biométriques.

Afin d'évaluer la performance de nos contributions selon de nombreuses combinaisons de critères, nous avons besoin de lots de données biométriques. C'est dans ce cadre que nous nous sommes intéressés à la génération de données biométriques aléatoires sous forme de vecteurs de caractéristiques. À notre connaissance, il n'y a pas d'importants travaux sur ce sujet. Les différentes méthodes que nous avons essayées jusqu'ici n'ont pas encore permis d'obtenir des indicateurs similaires à une base de données biométriques réelles. Par exemple, nous obtenons un faible EER sans transformation puis une importante dégradation avec le biohashing, ou alors un EER plus faible avec binarisation que sans. Nous souhaitons continuer ces travaux sur la génération de données biométriques aléatoires ayant des indicateurs réalistes afin d'ensuite lier les performances de nos contributions à de nombreuses combinaisons d'indicateurs.

Chapitre 6

Conclusion

Résumé : *Ce court chapitre conclut ce manuscrit sur les passe-partout biométriques. On y rappelle les enjeux de la biométrie avant de revenir sur nos deux contributions et les perspectives.*

6.1 Contexte

Dans ce manuscrit de thèse, nous avons décrit la biométrie comme un domaine dont l'objectif est de reconnaître un individu. Pour cela, nous capturons une modalité biométrique, comme une empreinte digitale, une photo du visage, un enregistrement d'électrocardiogramme, afin d'en extraire un vecteur de caractéristiques.

Nous avons utilisé dans nos travaux trois bases de données biométriques issues de modalités différentes : empreintes digitales, visages, et électrocardiogrammes. Ces trois bases ont des caractéristiques et des performances différentes afin de valider largement nos expériences.

6.2 Problématique

Les données biométriques sont des données à caractère personnel, ce qui impose leur protection. On constate que dans le cadre d'un stockage centralisé, l'impact d'une fuite de ces données biométriques peut être important. C'est pourquoi différentes méthodes de protection de données biométriques ont été présentées. Nous nous sommes particulièrement intéressés à des méthodes de protection génériques permettant de protéger des vecteurs de caractéristiques, utilisables avec de nombreuses modalités biométriques. La protection utilisée dans ce manuscrit consiste en

une transformation du vecteur de caractéristiques par projection depuis une matrice construite avec une graine vers un vecteur binaire appelé gabarit.

Nous avons présenté les attaques existantes sur la propriété de non-inversibilité de ces transformations, en détaillant particulièrement celles utilisant des algorithmes génétiques, pour leur proximité avec les travaux de cette thèse. Un des objectifs de ces attaques est de construire une préimage proche depuis un gabarit et la graine à l'origine de la matrice ayant projeté vers ce gabarit. Une préimage proche est un vecteur de caractéristiques permettant de s'authentifier pour le gabarit.

6.3 Contributions

6.3.1 Préimage proche et réutilisable

Dans une première partie de cette thèse, nous avons travaillé la construction de préimage proche et réutilisable. L'objectif est d'avoir un vecteur de caractéristiques permettant d'usurper deux gabarits, chacun issu d'un vecteur de caractéristiques d'un même individu, ayant été projetés avec deux graines distinctes. Cette approche est différente des travaux existants, qui n'utilisent soit qu'un gabarit, soit deux gabarits, mais issus du même vecteur de caractéristiques. Notre approche correspond mieux à la réalité, où deux gabarits stockés en phase d'enregistrement sont issus de deux captures de modalité biométrique, ayant engendré deux vecteurs de caractéristiques distincts.

Nous avons construit ces préimages proches et réutilisables à l'aide d'un algorithme génétique. Nous avons effectué de nombreuses expériences permettant de paramétrer au mieux l'algorithme génétique, en obtenant la combinaison suivante : une population de taille 200 évoluant sur 500 itérations, utilisant une sélection par rang, combinée avec un croisement double, perturbée par des mutations de probabilité $P_{mutation} = 20$. Cet algorithme paramétré nous permet de construire des préimages proches et réutilisables pour tous les couples de gabarits de toutes nos bases de données biométriques.

Afin de donner du relief à ces performances, nous les avons comparées à d'autres algorithmes donnant des préimages proches et réutilisables. La première alternative prend des vecteurs de caractéristiques parmi ceux existants dans les bases de données biométriques en notre possession. Cette méthode permet d'obtenir de nombreuses préimages proches et réutilisables, mais n'en obtient pas pour tous les couples. De plus, elle nécessite d'avoir à disposition de nombreux vecteurs de caractéristiques. La deuxième méthode construit aléatoirement des vecteurs de caractéristiques. Cette

méthode ne permet pas d'obtenir des préimages proches et réutilisables pour tous les couples de gabarits et elle est moins performante que l'algorithme génétique et que la première alternative. Cependant, elle ne nécessite pas de connaître des vecteurs de caractéristiques existants, mais uniquement de pouvoir borner les valeurs obtenues aléatoirement pour construire le vecteur de caractéristiques. Cela revient à connaître l'algorithme d'extraction permettant d'obtenir un vecteur de caractéristiques depuis une capture de modalité biométrique. La troisième alternative est une autre méthode d'optimisation que l'algorithme génétique : la construction par escalade (*hillclimbing*). Cette méthode procure très peu de préimages proches et réutilisables en temps de calcul similaire par rapport à l'algorithme génétique. Trois méthodes de construction par escalade ont été explorées. La meilleure permet d'obtenir des préimages proches et réutilisables pour la moitié des couples de gabarits issus de la base d'empreintes digitales.

Nous avons exploré une variante de la construction de préimage proche et réutilisable : nous partons de deux gabarits issus d'individus différents, contrairement à précédemment où ils étaient issus de captures distinctes d'une même modalité biométrique d'un unique individu. Nous réussissons à nouveau à construire des préimages proches et réutilisables pour des couples de gabarits issus d'individus différents.

6.3.2 Passe-partout biométriques

Notre deuxième contribution consiste en la construction de passe-partout biométriques. Ces derniers ont été explorés sous deux scénarios différents.

Construction d'un passe-partout pour une base de données biométriques révocables :

Dans le premier scénario, nous partons d'une base de données biométriques révocables : un ensemble de couples de gabarits et graines. Nous construisons avec un algorithme génétique un vecteur de caractéristiques permettant de s'authentifier pour un maximum de gabarits de la base de données biométriques révocables. Nous obtenons des passe-partout permettant de s'authentifier comme 70% des individus avec les empreintes digitales, 60% pour les électrocardiogrammes et 15% pour les visages.

Étant donné que nous n'obtenons pas des passe-partout permettant d'usurper l'intégralité des individus de nos bases de données biométriques, nous avons souhaité partitionner ces bases en plusieurs passe-partout : l'objectif est de construire un minimum de passe-partout tel que pour chaque individu de la base, il y ait au moins un passe-partout qui s'authentifie pour lui. Nous avons réussi à construire un ensemble

de seulement 5 passe-partout permettant de partitionner notre base de données biométriques d'empreintes digitales. Il en faut 12 pour les électrocardiogrammes et 18 pour les visages.

Construction d'une base de données biométriques révocables pour un passe-partout :

Dans ce second scénario, nous construisons une base de données biométriques révocables pour un passe-partout. Le vecteur de caractéristiques formant le passe-partout est connu à l'avance et est fixé, et nous disposons de la base de données biométriques. Pour chaque individu, nous choisissons la graine paramétrant la projection de son vecteur de caractéristiques en un gabarit pour lequel le passe-partout s'authentifie avec succès. La matrice de projection étant générée à partir de la graine, nous utilisons une stratégie de force brute en testant de nombreuses graines jusqu'à ce que le gabarit issu de la projection soit compatible avec le passe-partout.

Cette stratégie de force brute implique la génération de nombreuses matrices de projection orthogonales. Le biohashing est une transformation utilisant l'algorithme de Gram-Schmidt pour orthogonaliser la matrice générée depuis la graine. Cet algorithme est coûteux et nous souhaitons aller plus vite pour cette étape. Nous avons testé d'autres méthodes de projection en utilisant des matrices dont le procédé de construction est proposé par Achlioptas. Nous avons comparé ces projections avec le biohashing en utilisant le lemme de Johnson-Lindenstauss. Elles obtiennent des performances similaires au biohashing à condition d'utiliser une taille de gabarit d'au moins 128 bits. Ces matrices nous permettent de mettre en pratique notre second scénario, car elles génèrent des matrices de projection environ 30 fois plus rapidement qu'avec le biohashing qui utilise l'algorithme de Gram-Schmidt.

Nous avons ensuite mené des expériences pour choisir les graines permettant au passe-partout de s'authentifier pour tous les gabarits de la base de données biométriques révocables. Ce choix de graines est particulièrement efficace pour la base issue de visages, dans l'espace binaire comme dans l'espace euclidien. Pour les 3 bases, nous obtenons un passe-partout couvrant tous les gabarits, sans que le choix de graines ne dégrade les performances en termes d'EER.

Extension à un individu passe-partout :

La construction d'une base de données biométriques révocables en choisissant les graines au lieu de les prendre aléatoirement permet à un passe-partout fixé à l'avance de couvrir toute cette base. Nous avons étendu ce concept à l'individu passe-partout. Nous choisissons les graines avec un vecteur de caractéristiques issu d'une capture de modalité biométrique de cet individu passe-partout. Ce premier vecteur

de caractéristiques couvre toute la base. Nous avons mis en évidence une dégradation importante de la couverture de vecteurs de caractéristiques issus d'autres captures de cette modalité biométrique de l'individu passe-partout. Nous avons mené des expériences pour minimiser cette dégradation de couverture et nous avons constaté qu'il faut utiliser plusieurs vecteurs de caractéristiques de l'individu passe-partout pour choisir les graines. Cela permet aux futurs vecteurs de caractéristiques de mieux couvrir la base en limitant la dégradation de performance.

Pour finir, nous avons étudié la corrélation des performances de couverture entre les vecteurs de caractéristiques utilisés pour le choix de graines et les autres vecteurs de caractéristiques. Pour la base LFW, dont les 4 vecteurs utilisés pour la recherche de graines ne couvrent pas toute la base, nous constations une corrélation importante avec les performances de couverture des autres vecteurs. Pour les autres bases, la performance de couverture des 4 premiers vecteurs étant importante, on ne peut pas observer de corrélation.

6.4 Perspectives

Nous avons fini ce manuscrit par les perspectives et les nouvelles problématiques apportées par ces contributions.

Dans le cadre de notre premier scénario, nous construisons un passe-partout avec pour objectif de maximiser la couverture sur une base de données biométriques révocables connue. Nous avons étudié la réutilisabilité de ce passe-partout, en le construisant avec la connaissance d'une sous-base puis en observant sa couverture sur une autre sous-base, issue de la même modalité et du même algorithme d'extraction et de projection. Les passe-partout construits conservent une importante capacité de couverture pour des données biométriques issues d'individus inconnus à la construction. Nous souhaitons investiguer l'impact de la performance de la base sur le TCO et la TOD, ainsi que l'impact de la taille de la base sur ces indicateurs. Nous voulons étudier la pertinence de contre-mesures comme le choix de graines.

Dans le cadre de notre second scénario, nous avons renouvelé les expériences sur deux autres bases de données biométriques : la première est multimodale, issue de la fusion des bases FVC et LFW, et la seconde est issue d'images d'oreilles. Le choix de graines permet d'excellentes performances sur la première base et de plus modestes performances sur la base issue d'oreilles. Nous souhaitons investiguer les différents paramètres influant sur la performance de recherche de graines. Nous voulons étendre l'individu passe-partout à plusieurs individus passe-partout.

Des travaux ont déjà débuté dans la poursuite des nôtres, notamment dans la récente conférence SECRYPT 2022 :

[Durbet et al., 2022a] décrivent une attaque sur l'extraction de caractéristiques avec un filtre de Sobel suivie d'une transformation par projection aléatoire. Dans un premier scénario, l'attaque consiste à modifier au minimum l'image de l'empreinte digitale d'un imposteur depuis un gabarit et sa graine, afin de s'authentifier avec succès pour ce gabarit depuis l'image modifiée. Ils inversent la projection pour obtenir un vecteur de caractéristiques, puis modifient au minimum l'image de l'imposteur pour obtenir une collision de caractéristiques, et finalement effectuer une attaque par présentation avec cette image modifiée. Dans un second scénario, ils proposent de modifier l'image en fonction de nombreux couples (gabarit, graine) afin que cette image modifiée usurpe tous ces gabarits. Les résultats de leurs expériences sont concluants pour des images synthétiques de petites tailles d'empreintes digitales.

[Durbet et al., 2022b] étudient les collisions dans une base de nombreuses données biométriques et décrivent la génération de passe-partout biométriques usurpant de nombreux utilisateurs d'une base fuitée et potentiellement de futurs utilisateurs. Les caractéristiques sont extraites depuis des images d'empreintes digitales avec un filtre de Gabor. Ils partitionnent la base pour accélérer la génération de passe-partout, décrit comme étant dans l'intersection de boules de Hamming. Pour se prémunir de telles intersections, ils proposent d'avoir un seuil τ inférieur à 10% de la taille du gabarit. Par exemple, pour des gabarits de taille $n = 512$ bits, il faut un seuil d'au maximum $\tau = 51$, tolérant au plus une différence de 51 bits en distance de Hamming.

Chapitre 7

Publications de l'auteur

Journal international

Gernot, T. et Lacharme, P. (2022). Biometric masterkeys. *Computers & Security*, 116 :102642.

École d'été

Gernot, T. et Lacharme, P. (2020). Long-lived nearby-template preimages on biometric transformation with genetic algorithm. *XVII Int.l Summer School on Biometrics*.

Bibliographie

[Achlioptas, 2003] Achlioptas, D. (2003). Database-friendly random projections : Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences (JCSS)*, 66(4) :671–687. 79

[Adler, 2005] Adler, A. (2005). Vulnerabilities in biometric encryption systems. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 1100–1109. Springer. 36

[Alzubaidi and Kalita, 2016] Alzubaidi, A. and Kalita, J. (2016). Authentication of smartphone users using behavioral biometrics. *IEEE Communications Surveys & Tutorials*, 18(3) :1998–2026. 6

[Andoni and Indyk, 2006] Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468. 21

[Belguechi et al., 2016] Belguechi, R., Hafiane, A., Cherrier, E., and Rosenberger, C. (2016). Comparative study on texture features for fingerprint recognition : application to the biohashing template protection scheme. *Journal of Electronic Imaging*, 25(1). 25

[Bontrager et al., 2018] Bontrager, P., Roy, A., Togelius, J., Memon, N., and Ross, A. (2018). Deepmasterprints : Generating masterprints for dictionary attacks via latent variable evolution. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE. 37

[Bousseljot et al., 1995] Bousseljot, R., Kreiseler, D., and Schnabel, A. (1995). Nutzung der EKG-signaldatenbank cardiodat der PTB über das internet. 27

[Brunelli and Falavigna, 1995] Brunelli, R. and Falavigna, D. (1995). Person identification using multiple cues. *IEEE transactions on pattern analysis and machine intelligence*, 17(10) :955–966. 8

[Charikar, 2002] Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *ACM Symposium Theory Computing*, pages 380–388. 21

[Chen et al., 2014] Chen, S.-T., Guo, Y.-J., Huang, H.-N., Kung, W.-M., Tseng, K.-K., and Tu, S.-Y. (2014). Hiding patients confidential datainthe ecg signal viaa transform-domain quantization scheme. *Journal of medical systems*, 38(6) :1–8. 28, 125

[CNIL, a] CNIL. Biométrie : un "gabarit" biométrique, c'est quoi ? 20

[CNIL, b] CNIL. Délibération n° 2019-001 du 10 janvier 2019 portant règlement type relatif à la mise en œuvre de dispositifs ayant pour finalité le contrôle d'accès par authentification biométrique aux locaux, aux appareils et aux applications informatiques sur les lieux de travai. 10

[Dasgupta and Gupta, 2003] Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures and Algorithms*, 22(1) :60–65. 79

[Davida et al., 1998] Davida, G. I., Frankel, Y., and Matt, B. J. (1998). On enabling secure applications through off-line biometric identification. In *Proceedings. 1998 IEEE Symposium on Security and Privacy (Cat. No. 98CB36186)*, pages 148–157. IEEE. 20

[Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface : Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694. 27

[Doddington et al., 1998] Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. A. (1998). Sheep, goats, lambs and wolwes : A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *International Conference on Spoken Language Processing (ICSLP)*. 30

[Dong et al.,] Dong, X., Jin, Z., and Jin, A. T. B. A genetic algorithm enabled similarity-based attack on cancellable biometrics. In *10th IEEE International Conference on Biometrics : Theory, Applications and Systems (BTAS)*. IEEE. 26

[Dong et al., 2019a] Dong, X., Jin, Z., and Jin, A. T. B. (2019a). A genetic algorithm enabled similarity-based attack on cancellable biometrics. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE. 39

[Dong et al., 2019b] Dong, X., Jin, Z., and Teoh, A. B. J. (2019b). A genetic algorithm enabled similarity-based attack on cancellable biometrics. In *IEEE Inter-*

national Conference on Biometrics : Theory, Applications and Systems (BTAS), pages 1–8. 26

[Dong et al., 2019c] Dong, X., Jin, Z., Teoh, A. B. J., Tistarelli, M., and Wong, K. (2019c). On the reliability of cancelable biometrics : Revisit the irreversibility. *arXiv preprint arXiv :1910.07770*. 26

[Durbet et al., 2022a] Durbet, A., Grollemund, P., Lafourcade, P., Migdal, D., and Thiry-Atighehchi, K. (2022a). Authentication attacks on projection-based cancelable biometric schemes. In di Vimercati, S. D. C. and Samarati, P., editors, *Proceedings of the 19th International Conference on Security and Cryptography, SECRYPT 2022, Lisbon, Portugal, July 11-13, 2022*, pages 568–573. SCITEPRESS. 114

[Durbet et al., 2022b] Durbet, A., Grollemund, P., Lafourcade, P., and Thiry-Atighehchi, K. (2022b). Near-collisions and their impact on biometric security. In di Vimercati, S. D. C. and Samarati, P., editors, *Proceedings of the 19th International Conference on Security and Cryptography, SECRYPT 2022, Lisbon, Portugal, July 11-13, 2022*, pages 382–389. SCITEPRESS. 114

[Feng et al., 2014] Feng, Y. C., Lim, M.-H., and Yuen, P. C. (2014). Masquerade attack on transform-based binary-template protection based on perceptron learning. *Pattern Recognition*, 47(9) :3019–3033. 37

[Feng et al., 2010] Feng, Y. C., Yuen, P. C., and Jain, A. K. (2010). A hybrid approach for generating secure and discriminating face template. *IEEE Transactions on Information Forensics and Security*, 5(1) :103–117. 21

[Giot et al., 2015] Giot, R., Dorizzi, B., and Rosenberger, C. (2015). A review on the public benchmark databases for static keystroke dynamics. *Computers & Security*, 55 :46–61. 7

[Goldberger et al., 2000] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., and Stanley, H. (2000). Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals. 27

[Gomez-Barrero and Galbally, 2020] Gomez-Barrero, M. and Galbally, J. (2020). Reversing the irreversible : A survey on inverse biometrics. *Computers & Security*, 90 :101700. 37

[Gonzalez et al., 2012] Gonzalez, E., Alvarez, L., and Mazorra, L. (2012). Normalization and feature extraction on ear images. In *2012 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 97–104. IEEE. 103

[Hsu, 2002] Hsu, W. H. (2002). Introduction to Genetic Algorithms. *CIS 732 : Machine Learning and Pattern Recognition*, page 18. 57

[Huang et al., 2008] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild : A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images : Detection, Alignment, and Recognition*, page 15. 26

[Inuma et al., 2009] Inuma, M., Otsuka, A., and Imai, H. (2009). Theoretical framework for constructing matching algorithms in biometric authentication systems. In *International Conference in Biometrics (ICB)*, pages 806–815. 30

[Jain et al., 1999a] Jain, A., Bolle, R., and Pankanti, S. (1999a). *Biometrics : personal identification in networked society*, volume 479. Springer Science & Business Media. 6, 7, 128

[Jain et al., 2000] Jain, A., Hong, L., and Pankanti, S. (2000). Biometric identification. *Communications of the ACM*, 43(2) :90–98. 12, 16, 125

[Jain et al., 2007] Jain, A. K., Flynn, P., and Ross, A. A. (2007). *Handbook of biometrics*. Springer Science & Business Media. 5

[Jain et al., 2008] Jain, A. K., Nandakumar, K., and Nagar, A. (2008). Biometric template security. *EURASIP Journal on advances in signal processing*, 2008 :1–17. 19, 20, 125

[Jain et al., 1999b] Jain, A. K., Prabhakar, S., and Chen, S. (1999b). Combining multiple matchers for a high security fingerprint verification system. *Pattern Recognition Letters*, 20(11-13) :1371–1379. 8

[Jain et al., 2004] Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1) :4–20. 17

[Jain et al., 2011] Jain, A. K., Ross, A. A., and Nandakumar, K. (2011). *Introduction to biometrics*. Springer Science & Business Media. 13

[Jin,] Jin, Z. Lfw_10samples_insightface. 26

[Johnson and Lindenstrauss, 1984] Johnson, W. B. and Lindenstrauss, J. (1984). Extension of lipschitz mapping into a hilbert space. *Contenporary mathematics*, 26 :189–206. 79

[Juels and Sudan, 2006] Juels, A. and Sudan, M. (2006). A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2) :237–257. 20

[Juels and Wattenberg, 1999] Juels, A. and Wattenberg, M. (1999). A fuzzy commitment scheme. In *Proceedings of the 6th ACM conference on Computer and communications security*, pages 28–36. 20

[Kaplan et al., 2017] Kaplan, E., Gursoy, M. E., Nergiz, M. E., and Saygin, Y. (2017). Known sample attacks on relation preserving data transformations. *IEEE Transactions on Dependable and Secure Computing*, 17(2) :443–450. 36

[Kennedy and Eberhart, 1995] Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE. 38

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25. 103

[Lacharme et al., 2013] Lacharme, P., Cherrier, E., and Rosenberger, C. (2013). Preimage attack on biohashing. In *International Conference on Security and Cryptography (SECRYPT)*, pages 363–370. 39, 45

[Liu et al., 2006] Liu, K., Giannella, C., and Kargupta, H. (2006). An attacker's view of distance preserving maps for privacy preserving data mining. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 297–308. Springer. 36

[Maio et al., 2002] Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., and Jain, A. K. (2002). FVC2002 : Second fingerprint verification competition. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pages 811–814. 25

[Makowski et al., 2021] Makowski, D., Pham, T., Lau, Z. J., and Brammer, J. C. (2021). Neurokit2 : The python toolbox for neurophysiological signal processing. 28

[Martínez et al., 2004] Martínez, J. P., Almeida, R., Olmos, S., Rocha, A. P., and Laguna, P. (2004). A wavelet-based ecg delineator : evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4) :570–581. 28

[Mitchell, 1998] Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press. 33, 34

[Murakami et al., 2012] Murakami, T., Takahashi, K., and Matsuura, K. (2012). Towards optimal countermeasures against wolves and lambs in biometrics. In *International Conference on Biometrics : Theory, Applications and Systems (BTAS)*, pages 69–76. 30

[Nagar et al., 2010] Nagar, A., Nandakumar, K., and Jain, A. K. (2010). Biometric template transformation : a security analysis. In *Media Forensics and Security II*, volume 7541, pages 237–251. SPIE. 36

[Nanwate and Sadhya, 2020] Nanwate, S. and Sadhya, D. (2020). Similarity attack on cancelable biometric templates using particle swarm optimization. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 693–697. IEEE. 38

[Parkinson and Khan, 2022] Parkinson, S. and Khan, S. (2022). A survey on empirical security analysis of access control systems : A real-world perspective. *ACM Computing Surveys (CSUR)*. 10

[Patel et al., 2015] Patel, V. M., Ratha, N. K., and Chellappa, R. (2015). Cancelable biometrics : A review. *IEEE Signal Processing Magazine*, 32(5) :54–65. 20, 22, 125

[Pillai et al., 2011] Pillai, J. K., Patel, V. M., Chellappa, R., and Ratha, N. K. (2011). Secure and robust iris recognition using random projections and sparse representations. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 33(9) :1877–1893. 21

[Pinto and Cardoso, 2019] Pinto, J. R. and Cardoso, J. S. (2019). A end-to-end convolutional neural network for ecg based biometric authentication. In *IEEE International Conference on Biometrics, Theory, Appl. Syst. (BTAS)*, pages 1–8. 28

[Prabhakar et al., 2003] Prabhakar, S., Pankanti, S., and Jain, A. K. (2003). Biometric recognition : Security and privacy concerns. *IEEE security & privacy*, 1(2) :33–42. 11

[Ratha et al., 2001a] Ratha, N. K., Connell, J. H., and Bolle, R. M. (2001a). An analysis of minutiae matching strength. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 223–228. Springer. 12, 125

[Ratha et al., 2001b] Ratha, N. K., Connell, J. H., and Bolle, R. M. (2001b). Enhancing security and privacy in biometrics-based authentication system. *IBM Systems J.*, 37(11) :2245–2255. 19, 21

[Rathgeb and Uhl, 2011] Rathgeb, C. and Uhl, A. (2011). A survey on biometric cryptosystems and cancelable biometrics. *EURASIP journal on information security*, 2011(1) :1–25. 20

[Reid et al., 2013] Reid, D. A., Samangooei, S., Chen, C., Nixon, M. S., and Ross, A. (2013). Soft biometrics for surveillance : an overview. *Handbook of statistics*, 31 :327–352. 11

[Ross and Jain, 2003] Ross, A. and Jain, A. (2003). Information fusion in biometrics. *Pattern recognition letters*, 24(13) :2115–2125. 8

[Roy et al., 2017] Roy, A., Memon, N., and Ross, A. (2017). Masterprint : Exploring the vulnerability of partial fingerprint-based authentication systems. *IEEE Transactions on Information Forensics and Security*, 12(9) :2013–2025. 37

[Roy et al., 2019] Roy, A., Memon, N., and Ross, A. (2019). Masterprint attack resistance : A maximum cover based approach for automatic fingerprint template selection. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE. 37

[Roy et al., 2018] Roy, A., Memon, N., Togelius, J., and Ross, A. (2018). Evolutionary methods for generating synthetic masterprint templates : Dictionary attack in fingerprint recognition. In *2018 International Conference on Biometrics (ICB)*, pages 39–46. IEEE. 37

[Sadhya et al., 2016] Sadhya, D., Singh, S. K., and Chakraborty, B. (2016). Review of key-binding-based biometric data protection schemes. *IET Biometrics*, 5(4) :263–275. 20

[Singh et al., 2019] Singh, M., Singh, R., and Ross, A. (2019). A comprehensive overview of biometric fusion. *Information Fusion*, 52 :187–205. 8

[Soutar et al., 1998] Soutar, C., Roberge, D., Stoianov, A., Gilroy, R., and Kumar, B. V. (1998). Biometric encryption using image processing. In *Optical Security and Counterfeit Deterrence Techniques II*, volume 3314, pages 178–188. SPIE. 36

[Teh et al., 2013] Teh, P. S., Teoh, A. B. J., and Yue, S. (2013). A survey of keystroke dynamics biometrics. *The Scientific World Journal*, 2013. 7

[Teoh et al., 2008] Teoh, A. B. J., Kuan, Y. W., and Lee, S. (2008). Cancellable biometrics and annotations on biohash. *Pattern Recognition*, 41(6) :2034–2044. 21

[Teoh et al., 2004] Teoh, A. B. J., Ngo, D. C. L., and Goh, A. (2004). Personalised cryptographic key generation based on facehashing. *Computers & Security*, 23. 21

[Une et al., 2007] Une, M., Otsuka, A., and Imai, H. (2007). Wolf attack probability : A new security measure in biometric authentication systems. In *International Conference in Biometrics (ICB)*, pages 396–406. 30

[Wang et al., 2018] Wang, J., Zhang, T., Song, J., Sebe, N., and Shen, H. T. (2018). A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9) :769–790. 21

[Wang and Plataniotis, 2010] Wang, Y. and Plataniotis, K. N. (2010). An analysis of random projection for changeable and privacy-preserving biometric verification. *IEEE Transations on Systems, Man, and Cybernetics, Part B*, 40(5) :1280–1293. 21

[Yager and Dunstone, 2010] Yager, N. and Dunstone, T. (2010). The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2) :220–230. 30

[Zuev and Ivanov, 1999] Zuev, Y. A. and Ivanov, S. (1999). The voting as a way to increase the decision reliability. *Journal of the Franklin Institute*, 336(2) :361–378. 8

Table des figures

2.1 Points d'attaques génériques dans un système biométrique [Ratha et al., 2001a] 12
2.2 Algorithme d'extraction de caractéristiques 15
2.3 Système biométrique générique [Jain et al., 2000] 16
2.4 Schémas de protection biométrique [Jain et al., 2008] 19
2.5 Protection avec une graine du vecteur de caractéristiques en un gabarit . 21
2.6 Biométrie révocable [Patel et al., 2015] 22
2.7 Exemple de signal de la base FVC 25
2.8 Indicateurs de performance de la base FVC sans transformation 26
2.9 Indicateurs de performance de la base FVC avec transformation (biohashing) 26
2.10 Indicateurs de performance de la base LFW sans transformation 27
2.11 Indicateurs de performance de la base LFW avec transformation (biohashing) 27
2.12 Ensemble PQRST d'un signal ECG [Chen et al., 2014] 28
2.13 Indicateurs de performance de la base PTB sans transformation 29
2.14 Indicateurs de performance de la base PTB avec transformation (biohashing) 29
2.15 Ménagerie FVC 31
2.16 Ménagerie LFW 31
2.17 Ménagerie PTB 32

3.1 Préimage proche d'un gabarit 44
3.2 Préimage proche et réutilisable 45
3.3 Nombre moyen de PPR par individu 47
3.4 Distance au 2nd gabarit par individu 49
3.5 Diagrammes en boîte en fonction de la taille de la population et du nombre d'itérations 55

3.6 Courbes cumulées croissantes en fonction de la taille de la population et du nombre d'itérations 55
3.7 Diagrammes en boîte en fonction de la méthode de sélection 58
3.8 Courbes cumulées croissantes en fonction de la méthode de sélection . . . 58
3.9 Croisements simple et double 60
3.10 Diagrammes en boîte en fonction de la probabilité de mutation 61
3.11 Courbes cumulées croissantes en fonction de la probabilité de mutation . 61
3.12 Diagrammes en boîte en fonction du type de croisement 63
3.13 Courbes cumulées croissantes en fonction du type de croisement 63
3.14 Diagrammes en boîte en fonction de l'algorithme 65
3.15 Courbes cumulées croissantes en fonction de l'algorithme 65
3.16 Évolution de la distance au second gabarit 66
3.17 Distance au 2nd gabarit par individu 67
3.18 Diagrammes en boîte selon la stratégie 69
3.19 Courbes cumulées croissantes selon la stratégie 69
3.20 Évolution de la distance au second gabarit selon la stratégie 70
3.21 Passe-partout depuis une base de données biométriques révocables 71
3.22 Évolution du taux de couverture pendant les itérations 74

4.1 Estimation d'ϵ pour différentes tailles de gabarits 81
4.2 EER (distance euclidienne) 81
4.3 EER (distance de Hamming) 82
4.4 Choix d'une graine pour un passe-partout 84
4.5 Nombre moyen de cycles d'horloge processeur (avec étape de binarisation) 86
4.6 Nombre moyen de cycles d'horloge processeur (LFW sans étape de binarisation) 87
4.7 Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour FVC et PTB avec binarisation 91
4.8 Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour LFW avec et sans binarisation 91
4.9 Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour FVC et PTB avec binarisation 93
4.10 Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour LFW avec et sans binarisation 93

5.1 Évolution du taux de couverture du passe-partout 99

5.2 Indicateurs de performance de la base fusionnée sans transformation . . . 101
5.3 Nombre moyen de cycles d'horloge processeur pour la base fusionnée . . 102
5.4 Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 4$ pour la base fusionnée 102
5.5 Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour la base fusionnée 102
5.6 Indicateurs de performance de la base d'oreilles sans transformation . . . 103
5.7 Nombre moyen de cycles d'horloge processeur pour la base d'oreilles . . . 104
5.8 Courbes cumulées décroissantes de couverture des vecteurs d'un candidat individu passe-partout avec $T = 1$ et $T = 3$ pour la base d'oreilles 104
5.9 Corrélation entre la couverture de l'ensemble de recherche et de l'ensemble de test pour la base d'oreilles 105

Liste des tableaux

2.1 Comparaison de modalités biométriques [Jain et al., 1999a] 7
2.2 Taux d'usurpation maximum 32

3.1 Indicateurs de préimages proches et réutilisables parmi une base de données biométriques 46
3.2 Indicateurs de préimages proches et réutilisables aléatoires 48
3.3 Résultats des trois attaques par escalade 51
3.4 Indicateurs des distances au second gabarit selon la taille de la population et le nombre d'itérations 56
3.5 Indicateurs des distances au second gabarit selon la méthode de sélection 59
3.6 Indicateurs des distances au second gabarit selon la probabilité de mutation 62
3.7 Indicateurs des distances au second gabarit selon le type de croisement . 64
3.8 Indicateurs des distances au second gabarit en fonction de l'algorithme . 66
3.9 Indicateurs de préimages proches et réutilisables issues de l'algorithme génétique 67
3.10 Indicateurs des distances au second gabarit selon la stratégie 70
3.11 Taux de couverture optimale / Taille optimale de dictionnaire 73
3.12 FMR/TCO en fonction des seuils (FVC et PTB) 74

4.1 Accélération entre Gram-Schmidt et les 2 projections proposées 80

5.1 Taux de couverture optimale pour les données biométriques révocables non utilisées 99

Liste des algorithmes

1 Recherche de graines pour un passe-partout 85

Passe-partout biométriques

L'informatique est un domaine de plus en plus utilisé au quotidien et il est désormais incontournable dans une grande partie de nos activités. La sécurité informatique s'impose pour protéger ces activités et nos données privées.

La biométrie nous permet à tous de contribuer à cette sécurité en limitant les contraintes pour l'usager, mais les données biométriques sont à caractère personnel. Cette thèse s'inscrit dans la sécurité des données biométriques, qui passe notamment par des transformations paramétrées par des graines. Nous avons initialement attaqué la propriété de non-inversibilité de ces transformations en construisant des préimages proches et réutilisables à l'aide d'un algorithme génétique. Des variantes de ces préimages nous ont amenés à vouloir construire des passe-partout biométriques, permettant d'usurper une large partie des utilisateurs d'un système biométrique. Enfin, nous avons voulu orienter le choix des graines et donc des transformations pour permettre à un passe-partout fixé d'usurper l'intégralité des utilisateurs d'un système biométrique. Nous avons étendu ce concept à un individu passe-partout, pour lequel les futures acquisitions biométriques persistent à usurper les utilisateurs.

Biometric masterkeys

Computer science is a field that is used more and more in our daily lives and is now essential to a large part of our activities. Cybersecurity is necessary to protect these activities and our private data.

Biometrics allows us to contribute to this cybersecurity by limiting the constraints for the user, but biometric data is personal. This thesis deals with the security of biometric data, which is notably achieved through transformations parameterized by seeds. We initially attacked the non-invertibility property of these transformations by building long-lived nearby-template preimages using a genetic algorithm. Variations of these preimages led us to build biometric masterkeys, allowing spoofing a large part of the users of a biometric system. Finally, we wanted to choose the seeds to allow a fixed masterkey to spoof all the users of a biometric system. We have extended this concept to multiple masterkeys person, for which future biometric acquisitions persist in spoofing users.

Spécialité informatique. Mots-clés : Sécurité, biométrie, algorithme génétique, contrôle d'accès, projection, passe-partout.

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France