République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid – Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études
pour l'obtention du diplôme de Master en Informatique

Option : *Modèle Intelligent et Décision (M.I.D)*

Thème

Système d'Identification Biométrique Multimodal

Réalisé par :

- MEZOUAR Yassine

Présenté le 04 Juillet 2022 devant le jury composé de :

- *Mr. BENTAALAH Amine* *(Président)*
- *Mr. BENAZZOUZ Mourtada* *(Encadreur)*
- *Mr. BERRABAH Sid-Ahmed* *(Examinateur)*

Année universitaire : 2021 – 2022

Remerciements

En premier lieu, Je remercie ALLAH le tout puissant qui m'a aidé et m'a donné le courage, la patience, la force et la volonté pour la réalisation de ce travail.

Mes plus vifs remerciements à tous mes professeurs qui ont contribué à ma formation le long de mon cursus universitaire et qui ont bien voulu par leur grande générosité partager leur savoir avec leurs étudiants.

J'aimerais aussi présenter mes sincères remerciements à mon encadrant, Mr Mourtada BENAZZOUZ, pour son aide inestimable au cours de ce travail, et qui surtout, m'a donné le courage d'achever ce travail.

Ce travail n'aurait pas été possible sans le soutien de ma famille et de mes amis, je tiens à remercier mes proches pour tous les encouragements, ainsi que mes amis, pour leur aide dans l'achèvement de ce travail.

Dédicace

Je dédie ce modeste travail

À mes chers parents pour leur patience, leur amour, leur soutien et leurs encouragements,

À mon frère et mes sœurs,

À tous mes proches,

À tous mes amis.

Merci

Résumé

Aujourd'hui, l'intelligence artificielle est utilisée presque partout dans notre vie, l'une de ses applications se trouve être la biométrie, permettant l'identification des personnes à travers de nouveaux traits complexes, l'amélioration des ceux utilisés auparavant, ainsi qu'une meilleure facilité d'utilisation. La biométrie n'est cependant pas infaillible, et peut échouer dans certains rares cas.

Le Deep Learning est l'un des sous-domaines de l'intelligence artificielle, basé sur des réseaux de neurones artificiels, il permet l'accomplissement de tâches complexes comme la reconnaissance d'objets, les prédictions d'évènements futurs, et bien d'autres, avec une très haute performance. D'où son utilisation déjà présente dans la biométrie, aidant à identifier les personnes à travers des traits complexes comme le visage, la démarche, la voix, et plusieurs autres traits.

Dans ce travail, nous allons explorer et tester l'utilisation du Deep Learning dans la détection biométrique multimodale, à savoir, l'utilisation de plus d'un trait simultanément. On a utilisé la base de données SWAN-Idiap ainsi qu'un réseau de neurones convolutifs modifié à partir d'un autre modèle. Commençant par tester les performances d'un système unimodal sur les deux modalités du visage et de la région périoculaire, puis on passera au multimodal, qui utilisera une fusion de ces deux-là, tout en collectant et en comparant les données résultantes au long du travail.

Mots clés : Système biométrique multimodal, apprentissage profond, réseaux de neurones convolutifs, reconnaissance du visage, reconnaissance périoculaire.

Abstract

Today, artificial intelligence is used almost everywhere in our life, one of its applications is biometrics, allowing the identification of people through new complex traits, improving the ones used before, as well as a better usability. However, biometrics are not infallible, and can fail in some rare cases.

Deep Learning is one of the subfields of artificial intelligence, based on artificial neural networks, it allows the accomplishment of complex tasks such as object recognition, predictions of future events, and many others, with a very high performance. Hence its use in biometrics, helping to identify people through complex features such as face, gait, voice, and several other features.

In this work, we will explore and test the use of Deep Learning in multimodal biometric recognition, namely, the use of more than one trait simultaneously. We used the SWAN-Idiap database as well as a convolutional neural network modified from another model. Starting by testing the performance of a unimodal system on the two modalities of the face and the periocular region, then moving on to the multimodal, which will use a fusion of these two, while collecting and comparing the resulting data throughout the work.

Keywords: Multimodal biometric system, deep learning, convolutional neural networks, face recognition, periocular recognition.

ملخص

في الوقت الحاضر، يتم استخدام الذكاء الاصطناعي تقريبا في كل مكان في حياتنا، ويصادف أن يكون أحد تطبيقاته القياسات الحيوية، التي تسمح بتحديد الأشخاص من خلال سمات معقدة جديدة، وتحسين السمات المستخدمة من قبل، وكذلك سهولة الاستخدام بشكل أفضل. ومع ذلك، فإن القياسات الحيوية ليست معصومة من الخطأ، ويمكن أن تخفق في بعض الحالات النادرة.

التعلم العميق هو أحد الحقول الفرعية للذكاء الاصطناعي، ويعتمد على الشبكات العصبية الاصطناعية، ويسمح بإنجاز المهام المعقدة مثل التعرف على الأشياء، والتنبؤ بالأحداث المستقبلية، وغيرها كثيرا، بأداء عالي. ومن ثم فإن استخدامه موجود بالفعل في القياسات الحيوية، مما يساعد على تحديد الأشخاص من خلال السمات المعقدة مثل الوجه والمشية والصوت والعديد من السمات الأخرى.

في هذا العمل، سوف نكتشف ونختبر استخدام التعلم العميق في الكشف عن القياسات الحيوية متعددة الوسائط، أي استخدام أكثر من سمة في وقت واحد. استخدمنا قاعدة بيانات SWAN-Idiap بالإضافة إلى شبكة عصبية التفافية معدلة من نموذج آخر. بدءًا من اختبار أداء من نظام أحادي الوسائط على طريقتي الوجه والمنطقة المحيطة بالعين، سننتقل بعد ذلك إلى الوسائط المتعددة، والتي سيستخدم اندماجا بين هذين الاثنين مع جمع ومقارنة البيانات الناتجة طول العمل.

الكلمات المفتاحية : نظام القياسات الحيوية متعدد الوسائط، التعلم العميق، الشبكات العصبية الالتفافية، التعرف على الوجه، التعرف على ما حول العين.

Table des matières

Table des figures

Liste des tables

Liste des abréviations

IA - Intelligence Artificielle

SVM - Support Vector Machines

GPU - Graphics Processing Unit (Processeur graphique)

NN - Neural Network

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

LSTM - Long Short-Term Memory (Longue mémoire à court terme)

GRU - Gated Recurrent Units (Unités récurrentes clôturées)

CPU - Central Processing Unit (Processeur central)

RAM - Random Access Memory (Mémoire vive)

OS - Operating System (Système d'exploitation)

ReLu - Rectified Linear Unit (Unité linéaire rectifiée)

RMSprop - Root Mean Squared Propagation

Introduction Générale

Contexte

La biométrie regroupe l'ensemble des techniques informatiques permettant de reconnaître automatiquement un individu à partir de ses caractéristiques physiques, biologiques, voire comportementales. [1]

Durant les dernières années, la biométrie est devenue de plus en plus pertinente dans nos vies, commençant par faciliter l'identification des individus au niveau national et international (Carte d'identité, passeport…), intégration dans les lieux professionnels (pointeuse biométrique…), et aujourd'hui, étant intégré à nos smartphones, elle est présente quotidiennement.

Même en Algérie, elle est présente dans plusieurs secteurs, facilitant l'identification et l'authentification des personnes pour les services publics, avec comme récente addition, le permis biométrique.

Problématique

Malgré sa grande supériorité aux anciennes méthodes, la biométrie n'est pas parfaite, en effet, l'un des systèmes d'authentification les plus répondus dépends de l'empreinte digitale de l'individu, celle-ci n'étant pas sans faille, d'autres systèmes ont vu le jour : reconnaissance du visage, de la voix, etc… Aucun de ses systèmes n'est 100% infaillible, d'ailleurs, dû à l'erreur humaine, aucun ne le sera.

Afin de se rapprocher de la perfection, une des techniques utilisées est de combiner plusieurs traits biométriques dans l'authentification : Biométrie Multimodale. En effet, en combinant deux ou plusieurs traits, le taux de reconnaissance augmente, et certains cas où l'identification était impossible, deviennent possibles. La sécurité est aussi rehaussée, en minimisant les failles de l'unimodal, et en ajoutant des couches de sécurité qui ne sont possibles que dans les systèmes multimodaux.

Contribution

Dans ce travail, nous allons discuter de la façon dont nous avons appliqué le Deep Learning dans la conception d'un système d'identification biométrique multimodal, ainsi qu'une comparaison des performances avec les systèmes unimodaux.

Nous présenterons la façon dont nous avons utilisé la base de données SWAN-Idiap, les traitements et les préparations effectués afin d'obtenir une grande quantité d'images traitables et homogène, la création d'un modèle de réseau de neurones convolutifs dérivé de VGG16, puis nous passerons aux apprentissages unimodaux, en commençant par la modalité du visage, puis celle de la région périoculaire. Après cela, on passera à la multimodalité qui sera une fusion du visage et de la région périoculaire.

Enfin, on terminera par l'interprétation des résultats obtenus, des performances générales, une comparaison entre l'unimodal et le multimodal et les éventuelles améliorations qui pourraient être apportées.

Plan du Mémoire

Ce mémoire est divisé en trois chapitres distincts qui présentent des informations générales sur le domaine de la biométrie et de l'apprentissage profond, le travail effectué et ses différentes étapes, une discussion sur les résultats obtenus ainsi qu'une comparaison entre la viabilité d'un système unimodal et multimodal.

Le chapitre 1 résume la biométrie unimodale et multimodale en général, les différences entre-elles, les différentes possibilités de son intégration, ainsi que ses bénéfices, et les éventuels problèmes qu'elle peut présenter.

Le chapitre 2 aborde le Deep Learning, contenant une brève description de son fonctionnement, des algorithmes qui y sont utilisés, ainsi que les techniques spécifiquement utilisées dans le secteur biométrique multimodal.

Le chapitre 3 montre les préparations préalables au travail concernant la base de données, la conception du modèle de réseau de neurones convolutif utilisé, les améliorations apportées, ainsi que les résultats obtenus, avec des comparaisons aux systèmes unimodaux.

Enfin, le mémoire sera clôturé par une conclusion générale, et des perspectives pour d'éventuels travaux futurs.

Chapitre 1 : Etat de l’Art

1.1 Introduction

Dans ce chapitre, on abordera la biométrie en général, les différents traits qui peuvent être utilisés, leurs caractéristiques, ainsi que les différentes combinaisons qui pourraient être utilisées dans la biométrie multimodale.

1.2 La biométrie

La biométrie (*littéralement « mesure du vivant »*) représente l'ensemble des caractéristiques biologiques et comportementales d'un individu à partir de laquelle des caractéristiques biométriques distinctives et répétables peuvent être extraites à des fins de reconnaissances biométriques.[2]

1.3 Système Biométrique et Fonctionnement

Un système biométrique est un système qui détecte les traits uniques d'une personne, ces traits peuvent être de type physiologique, comme le visage, ou comportementale, comme les habitudes d'une personne. Ces traits sont analysés, puis utilisés dans l'identification ou l'authentification d'une personne.

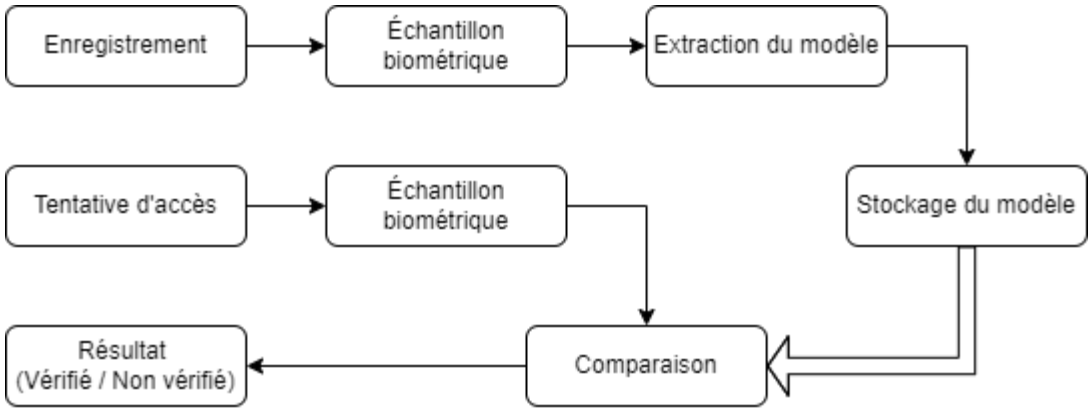

Figure 1-1 - Fonctionnement général d'un système biométrique

Il existe deux façons d'appliquer ces systèmes, en premier, l'identification, où l'échantillon est comparé à plusieurs modèles stockés (1:N), en second, la vérification, où l'échantillon est simplement comparé à un seul modèle (1:1).

Les systèmes biométriques utilisent les caractéristiques du corps humain d'un individu qui ne changent normalement pas avec le temps, comme le visage, l'iris, les empreintes digitales et les

empreintes de main. L'identification implique l'enregistrement de ces caractéristiques dans une base de données à des fins de reconnaissance future, comme le montre la figure 1. En outre, les individus peuvent être vérifiés ou identifiés sur la base de caractéristiques comportementales, par exemple la signature, la parole, le rythme de frappe et la démarche.[3]

1.4 Les types de mesures biométriques

Ces mesures sont séparées en deux types principaux : Physiologique, et comportementale. Chaque type comprend plusieurs approches qui sont continuellement améliorées.

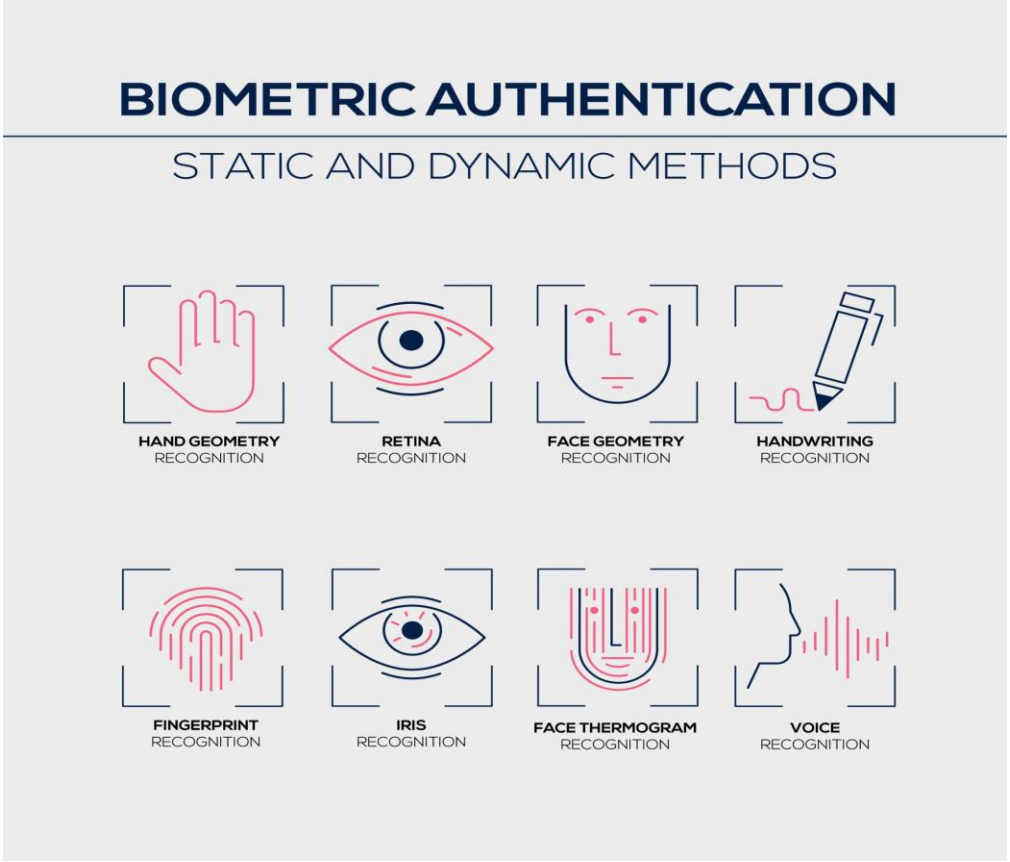

Figure 1-2 - Quelques mesures biométriques [5]

1.4.1 Mesures physiologiques

Établis sur les caractéristiques physiques du corps humain, chacune de ces mesures a ses avantages ainsi que ses inconvénients, elles comprennent les mesures suivantes :

- **Reconnaissance d'ADN :** Une personne partage 99,7 % de son ADN (acide désoxyribonucléique) avec ses parents biologiques, et les 0,3 % restants sont des codes

répétitifs variables. C'est sur ce codage répétitif que travaille la biométrie ADN par le biais du profilage génétique ou de l'empreinte génétique, où des régions d'ADN répétitives uniques sont isolées et identifiées.[4] Cependant, l'échantillonnage de l'ADN est actuellement assez intrusif et nécessite une forme de tissu, de sang ou d'un autre échantillon corporel, en ajoutant à cela, les techniques actuelles ne sont pas assez automatisées, et requièrent plus de temps que les autres méthodes pour finaliser l'analyse. Ces points ont conduit à une utilisation restreinte de cette mesure, étant principalement présente dans la détection de crimes et autres domaines des forces de l'ordre.

- **Reconnaissance faciale :** C'est l'une des mesures la plus simple à utiliser coté utilisateur. La reconnaissance faciale utilise les attributs du visage pour trouver l'identité de l'individu, tels que les yeux, le nez, les lèvres et autres traits, ainsi que la relation entre ceux-là, comme la distance entre eux, etc. Cette mesure se fait au biais d'images ou de vidéos. Bien qu'elle semble parfaite, elle est susceptible à échouer dans quelques rares cas.
- **Empreintes digitales :** Les empreintes digitales sont formées par les crêtes papillaires en relief qui courent à la surface de la peau des doigts. Cette mesure est l'une des plus utilisées à travers les années dû à sa simplicité ainsi que sa haute précision. Par le passé, l'empreinte était extraite au biais d'impression par encre sur du papier, puis, grâce à l'évolution des technologies, des capteurs simples nous fournissent une image avec bien plus de détails, par un simple contact du doigt. Dû à ce fait, une réduction de la précision du capteur, ainsi que des problèmes d'hygiène, peuvent survenir.
- **Scan de l'iris :** L'iris est la partie circulaire de l'œil, lorsqu'on voit une image en haute de résolution de l'iris, on constate que la géométrie est très complexe, étant constitué principalement de muscles et de nerfs, la disposition de ceux-là crée une forme unique à l'individuel, d'ailleurs, même l'iris des vrais jumeaux sont différents. Nécessitant un niveau élevé de détails, cette mesure est devenue beaucoup plus accessible récemment grâce aux améliorations des technologies des caméras, permettant de créer des systèmes où l'utilisateur n'a plus besoin de coller son œil au scanner. Un autre point pour l'iris est son adaptation à la lumière, ce mouvement ainsi que la géométrie qui en résulte peut servir comme mesure biométrique additionnelle, fortifiant ainsi la précision.

- **Géométrie de la main :** Cette mesure se base sur certains traits de la main, telle que la longueur et la largeur des doigts, la surface de la main, et même son épaisseur. Bien que l'utilisation de cette mesure soit plus présente aujourd'hui, elle reste beaucoup moins utilisée dû aux capteurs nécessaires qui sont relativement plus compliqué.

Plusieurs autres mesures moins connues existent, comme : les battements de cœur, la géométrie des doigts, des oreilles, des veines, et même une mesure olfactive basée sur des caractéristiques chimiques uniques. Ces mesures sont cependant moins populaires, et beaucoup moins utilisés par rapport aux autres que ce soit dû au manque de recherche, ou bien à des problèmes de convenance.

1.4.2 Mesures comportementales

Ces mesures sont basées sur la répétition de certains comportements durant une action, elles sont moins utilisées dû à la variance présente dans certains cas, on y trouve :

- **Reconnaissance vocale :** Mesure basée sur la voix d'un individu, cette mesure peut être définie comme partiellement physiologique (morphologie des cordes vocales, de la bouche, des lèvres, et autres facteurs physiques), elle reste néanmoins comportementale dû au fait que chaque personne parle différemment, mais aussi aux changements qui peuvent influencer la voix, comme l'âge de la personne, son état mental, les maladies, et bien d'autres. Par contre, ces mêmes paramètres peuvent rendre l'authentification plus difficile, limitant ainsi l'utilisation de cette mesure dans certains cas.
- **Signature :** L'une des mesures les plus anciennes, elle s'appuie sur l'écriture d'une personne et de la façon dont elle signe. Cette mesure est aussi influencée par quelques traits physiques (géométrie de la main), et, tout comme la reconnaissance vocale, l'influence comportementale de la personne est supérieure à ces traits. Aujourd'hui, cette mesure est plus précise, grâce à l'utilisation de capteurs électroniques qui captent aussi la vitesse ainsi que la pression du stylet. Certaines personnes sont susceptibles d'échouer la reconnaissance si leurs signatures ne sont pas consistantes (problème qui est allégé par le point précédent).
- **Frappe de clavier :** En effet, différentes personnes utilisent le clavier par différentes manières. Certaines personnes qui sont habituées à l'utilisation des ordinateurs tapent très vite, les personnes qui utilisaient des machines à écrire peuvent avoir pris l'habitude de

taper avec une plus forte force. Un autre point est la durée de l'appui, la différence entre la durée de l'accentuation et celle de la relâche, etc…

Autres mesures : démarche (caractéristiques de la marche d'une personne).

1.5 La biométrie multimodale

Après que de nombreux problèmes aient fait surface dans les systèmes unimodaux (Données bruyantes, variation interclasse, similarités interclasses, non-universalité, usurpation d'identité, etc. [6]), des systèmes utilisant deux ou plusieurs mesures, biométriques ont commencé à voir le jour, ces systèmes sont plus performants, avec un taux de succès supérieur, et naturellement plus sécurisé.

1.5.1 Composition

Dans un système biométrique multimodal, il peut y avoir de nombreuses caractéristiques et composantes différentes [8]. Des systèmes existent où il y a :

- Mesure biométrique unique avec plusieurs capteurs.
- Mesure biométrique unique avec plusieurs classificateurs.
- Mesure biométrique unique, collecté sur plusieurs sources (Plusieurs enregistrements vocaux, capturer le visage depuis plusieurs angles, etc…)
- Regroupement de plusieurs mesures biométrique. (Visage et iris, visage et empreinte digitale, visage et démarche, etc…)

Dans ce travail, nous nous intéressons aux systèmes utilisant un regroupement de plusieurs mesures biométriques.

1.5.2 Niveaux de fusion

Les systèmes multimodaux sont basés sur le même fonctionnement que les unimodaux (Figure 1.1), avec comme addition, une étape de fusion pour réunir les données de plusieurs capteurs ou classificateurs.

Cette fusion peut être positionnée sur l'une des étapes suivantes [7] :

- **Pendant l'extraction des échantillons :** Cela se fait en groupant les données récupérées auprès de plusieurs capteurs d'une même mesure biométrique (ex : Seulement les données d'empreintes digitales), ou bien plusieurs données d'un même capteur. Les

données sont ensuite combinées directement pour créer une information brute qui représentera notre individu. Cette technique est aussi être utilisé pour améliorer les données acquises lors de l'enregistrement d'une personne, par exemple pour obtenir toutes les informations d'une empreinte digitale, et afin de faciliter les prochaines utilisations et permettre l'utilisateur de s'authentifier même en utilisant le côté du doigt.

- **Pendant la comparaison des échantillons (Au niveau des caractéristiques) :** Pour la fusion à cette étape, les données provenant de différents capteurs (tels qu'un microphone ou une caméra) sont d'abord traité et des vecteurs de caractéristiques sont créés indépendamment ; puis, en utilisant un algorithme de fusion spécifique, les vecteurs de caractéristiques sont combinés pour former un vecteur de caractéristiques composite, qui sera ensuite utilisé pour représenter les données de l'individu. Comme cette méthode utilise plusieurs traits biométriques, une étape de normalisation est nécessaire pour le traitement correct des données.
- **Au niveau du score d'appariement :** Il s'agit de joindre des scores identiques produits par un module de correspondance pour chaque vecteur d'échantillon en entrée et les modèles stockés dans la base de données. Les caractéristiques sont traitées séparément, et un score de correspondance individuel est obtenu.[9]
- **Pendant la prise de décision :** La fusion à ce niveau est l'approche la plus couramment discutée dans la littérature biométrique, principalement en raison de la facilité d'accès et de traitement des scores de correspondance (par rapport aux données biométriques brutes ou à l'ensemble des caractéristiques extraites des données). Les méthodes de fusion à ce niveau peuvent être classées en trois grandes catégories : les schémas basés sur la densité, les schémas basés sur la transformation et les schémas basés sur les classifications.[10]

D'autres façons moins connues de fusion existent aussi, comme la fusion à partir de la qualité des données [11].

1.5.3 Méthodes de fusion

Il existe trois catégories principales méthodes pour la fusion :

- **Méthodes basées sur les règles :** Ces méthodes sont également appelées méthodes de fusion non supervisées, car il n'y a pas de processus de formation [12]. Ces méthodes

fusionnent les données acquises à partir des classificateurs en utilisant certaines règles, avec les plus connus étant de somme, produit, min, et max, mais aussi d'autres, comme la fusion linéaire pondérée, la règle du vote majoritaire et la règle définie par l'utilisateur, cette dernière approche n'utilise pas de méthodes statistiques, contrairement aux autres, mais plutôt une approche basée sur des règles de production, où chaque entré est définie dans son contexte d'utilisation, qui est après déterminé sur la base des événements d'entrées préalablement reconnus au même utilisateur.[13]

- **Méthodes basées sur la classification :** Ces méthodes utilisent des techniques pour classifier l'observation multimodale dans des classes prédéfinies, ces méthodes incluent l'inférence bayésienne, qui peut être appliquée à la fois pendant l'extraction des échantillons, et au niveau de la prise décision où les observations dérivées de divers classificateurs sont combinées et une inférence de la probabilité conjointe d'une décision est obtenue [14]. Les autres méthodes comprennent la théorie de Dempster-Shafer [15], les réseaux de neurones (NN), les réseaux bayésiens dynamiques [16], et les Machines à Vecteur Support (SVM).
- **Méthodes basées sur l'estimation :** Ces méthodes sont utilisées lorsqu'il y a un besoin de reconnaissance en temps réel, avec des personnes en mouvement (tracking), et cela pour mieux connaitre la position de la personne (ex : Identifier des personnes dans un supermarché rempli de personnes, et enregistrer leurs habitudes). Ces méthodes incluent entre autres le filtre de Kalman, le filtre de Kalman étendu et les méthodes de fusion par filtre à particules.[17]

1.6 Conclusion

En considérant la simplicité d'utilisation des technologies biométriques, il n'y a pas de surprises lorsqu'on voit son adoption dans le monde, et dans plusieurs domaines, que ce soit dans les locaux gouvernementaux, publics, ou privés, et même pour les utilisations personnelles à travers nos objets du quotidien. Ces technologies restent néanmoins susceptibles à être exploitée et abusée par des personnes malfaisantes, ne sont pas toujours parfaites, et sont limité par le type utilisé.

C'est pourquoi, la biométrie multimodale constitue une direction préférable pour l'évolution de ces technologies, fournissant un grand nombre de solutions pour les lacunes et les problèmes des systèmes unimodaux, tout en restant relativement simple à utiliser.

Dans ce chapitre, nous avons vu ce qu'est la biométrie, son utilisation et son fonctionnement, la différence entre unimodal et multimodal, ainsi que les différentes façons d'implémenter un système multimodal. Le prochain chapitre parlera du Deep Learning.

Chapitre 2 : Deep Learning

2.1 Introduction

L'apprentissage automatique, appelé Machine Learning en anglais (et de ce fait, l'intelligence artificielle) est l'un des piliers modernes de notre société moderne, il passe inaperçue pour la plupart, fournissant des avancées technologiques impressionnantes dans ce vingt-et-unième siècle. Le ML peut être appliqué et intégré dans pratiquement chaque domaine, qu'il soit médical, mathématique, et même artistique. Un de ses sous domaines est l'apprentissage profond (Deep learning en anglais), basé sur la structure du cerveau, il est retrouvé derrière les projets intelligents les plus ambitieux, comme la conduite automatique, les systèmes de diagnostic médical, et la création de robots simulant le comportement humain.

Au cours de ce chapitre, on apprendra ce qu'est le DL, son histoire, son fonctionnement, ainsi que son utilisation dans la biométrie multimodale.

2.2 Deep Learning

L'apprentissage profond est un sous-domaine du ML, alors que le ML traditionnel utilise des concepts relativement simples, le DL fonctionne à la base de neurones artificiels, d'où l'imitation du cerveau, mais y ajoute plusieurs autres couches de neurones, ce qui nous donne des réseaux de neurones profonds. Ce type de réseau permet alors de lui faire apprendre des concepts complexes que seuls les humains peuvent faire, et d'autres fonctions qui nous sont nous-même inaccessible.

Ce type d'apprentissage devient de plus en plus populaire, et son utilisation continue de croitre chaque jour, provoquant un changement immense dans notre société, sans que la plupart ne le remarquent.

2.3 Développement du DL/DNN

La première apparition substantielle de l'apprentissage automatique était en 1935, lorsque Alan Turing, un mathématicien britannique, a proposé son idée de "machine à apprendre", dotée d'une intelligence artificielle [20]. Puis, en 1997, le système de jeu d'échecs contrôlé par ordinateur, et développé par IBM, a battu le champion du monde en titre, Garry Kasparov, dans une série de six matchs. [20]

En 2001, un algorithme d'apprentissage automatique appelé Adaboost a été développé pour détecter les visages dans une image en temps réel. L'algorithme marchait en filtrant les images par le biais d'ensembles de décisions tels que "l'image présente-t-elle un point lumineux entre des taches sombres, ce qui pourrait indiquer l'arête d'un nez ?". Lorsque les données se déplaçaient plus loin dans l'arbre de décision, la probabilité de sélectionner le bon visage dans une image augmentait.[21]

Puis, dans les années suivantes, et grâce aux avancements technologiques, mais aussi grâce à l'apparition d'ImageNet, une base de données immense d'images étiquetées [22], les réseaux de neurones ont enfin pu connaitre leur vrai potentiel. Ajouté à ces deux points (Evolution de la puissance de calcul, et augmentation de la quantité de données), une troisième cause viendra aider l'apprentissage profond : le développement de nouveaux algorithmes pour le DL, et la sortie de bibliothèques comme TensorFlow et autres. [21]

2.4 Fonctionnement

Les DNN peuvent être visualisés comme plusieurs couches de neurones artificiels, divisé en une couche d'entré, des couches cachées, et enfin la couche de sortie. Chaque couche contient des nœuds de neurones, ces nœuds sont eux-mêmes connectés aux nœuds des couches adjacentes, formant un réseau de neurones complexe et 'profond', similaire à ceux trouvés dans les cerveaux.

Ces neurones reçoivent des signaux provenant des neurones adjacents, et, en dépendant du seuil d'activation de ce neurone, un nouveau signal sera émis vers les neurones des couches suivantes, l'importance (l'effet) de ce signal, dépendra de son poids (neurone émetteur). Cela va continuer vers les couches suivantes, jusqu'à ce que le signal soit transmis à la dernière couche, où toute les entrés seront compilé, et une valeur finale en sortira.

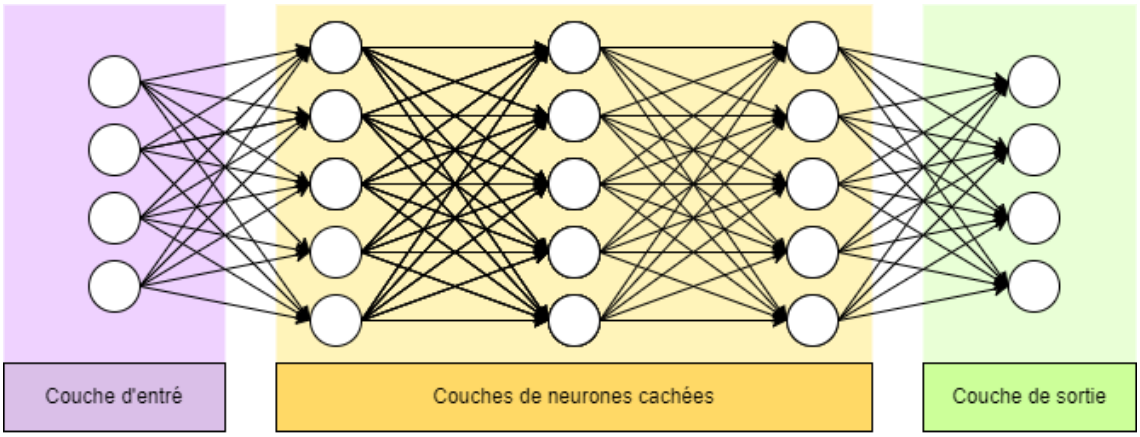

Figure 2-1 – Réseau de neurones multicouches.

Ces réseaux de neurones requièrent une très grande quantité de données pour être entrainés, et donc, une grande puissance de calcul, tout en prenant un plus grand temps à être entrainés ; d'ailleurs, l'un des exemples de ce type de DNN est OpenAI Five, le premier système d'IA à battre les champions du monde dans un jeu vidéo (Dota 2), l'entrainement a duré en total dix mois, avec une utilisation constante de 770±50 Pétaflop/s de puissance de calcul[18] (Pour comparaison, une carte graphique professionnel moderne de très haute performance, NVIDIA A100, est doté d'environ 312-624 Téraflop/s[19], environ 0.04%~0.09% du total).

2.5 Utilité et applications du DL

L'apprentissage profond est activement utilisé dans beaucoup de domaines, que ça soit scientifique, technologique, ou autre. Quelques utilisations populaires incluent :

- **Divertissement numérique et média :** Les systèmes de recommandations pour les vidéos, les chansons, les articles, ainsi que les publicités utilisent du DL pour fournir une expérience optimale à l'utilisateur. Le système étudie les habitudes de l'utilisateur et son historique, afin de prédire le type de contenu pertinent pour celui-ci. D'autres utilisations sont aussi possibles, par exemple reconnaitre les paroles des chansons, ou bien générer automatiquement un sous-titrage correct.
- **Secteur de la santé :** La détection des maladies et le diagnostic assisté par ordinateur ont tous deux étés rendus possibles grâce au DL. Il est largement utilisé pour la recherche médicale, la découverte de médicaments et le diagnostic de maladies potentiellement mortelles telles que le cancer et la rétinopathie diabétique grâce à l'imagerie médicale [23].

Durant la pandémie du COVID-19, la Chine a utilisé un scanner supplémenté par une IA comme méthode principale de diagnostic précoce du COVID-19, réduisant le temps de diagnostic de trente minutes, à quelques secondes seulement.[24]

- **La robotique :** Une application évidente du DL est la création d'IA pour les robots. Ces robots peuvent être entrainés à faire des taches complexes, que seul l'humain pouvait faire auparavant. Comme exemple, Boston Dynamics, une société qui se spécialise dans les robots, a développé une IA pour le robot 'Atlas', visant à répliquer les activités physiques dont les humains sont capables, comme courir sur divers types de terrain, traverser des obstacles, et même faire de l'acrobatie [25].

2.6 Les types d'algorithmes de DL

Plusieurs types d'algorithmes d'apprentissage profond existent, certains peuvent traiter des problèmes généraux, d'autres sont plus spécifique, et donc, plus adaptés à certains problèmes.
Nous allons voir les deux algorithmes les plus populaires du Deep Learning :

2.6.1 Réseaux de neurones convolutifs

Appelé Convolutional Neural Network en anglais (abrév. ConvNet), les CNN sont souvent utilisés pour le traitement des images, la détection des objets et les problèmes de classification. La convolution est un processus unique de filtrage d'image, qui a pour but d'évaluer chaque élément qui la compose, à travers leurs poids attribués (automatiquement). Les CNN s'appuient sur les principes de l'algèbre linéaire, en particulier la multiplication matricielle, pour identifier les motifs dans une image. Bien qu'ils soient plus rapides et pratiques que les méthodes utilisées traditionnellement pour la détection d'objets, ils requièrent une grande puissance de traitement, nécessitant l'utilisation de GPUs.[26]
Les CNN utilisent un système semblable à un perceptron multicouche conçu pour réduire les exigences de traitement, et sont composés de :

- **Couche de convolution :** Considérée comme composante principale, cette couche a pour rôle d'analyser l'entrée et de détecter les caractéristiques (features), ses composantes sont les données d'entrées, un filtre, et une carte de caractéristiques (qui est la sortie). La convolution est un type spécialisé d'opération linéaire utilisé pour l'extraction de

caractéristiques, où un tableau de nombres, appelé kernel ou filtre, est appliqué sur l'entrée, qui est aussi un tableau de nombres (tensor).

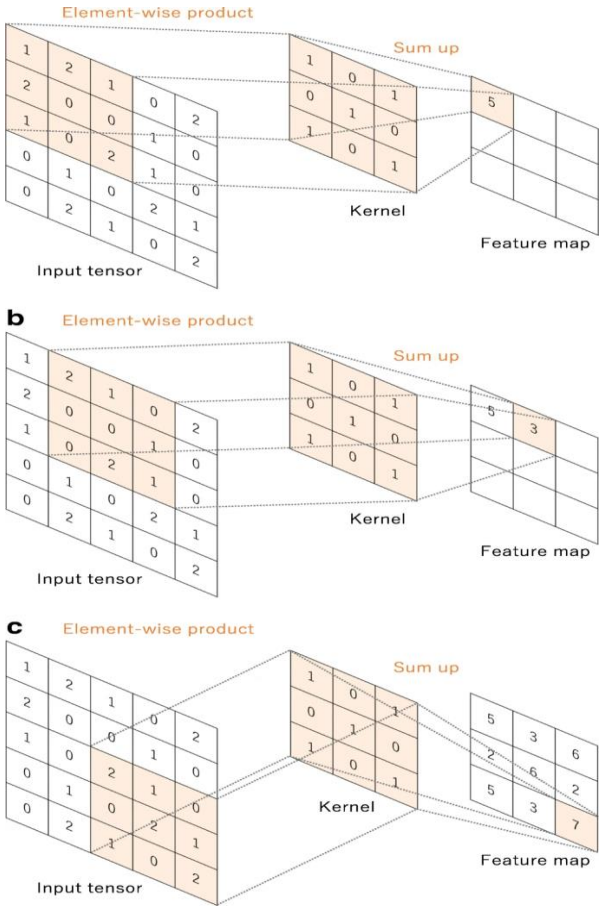

Figure 2.6.1 – Exemple de convolution [27]

- **Couche de pooling :** Cette couche a pour but de diminuer la dimensionnalité, en réduisant le nombre de paramètres d'entrées, cela permet de réduire la puissance de calcul nécessaire au traitement des données. Elle sert aussi à extraire les caractéristiques dominantes qui sont invariantes en termes de rotation et de position, ce qui permet de maintenir le processus d'apprentissage efficace du modèle [28]. Il existe deux principaux types de pooling :
  - **Max pooling :** La plus utilisée, lorsque le filtre se déplace sur l'entrée, il sélectionne le pixel ayant la valeur maximale à envoyer au tableau de sortie. [26]
  - **Average pooling :** Lorsque le filtre se déplace sur l'entrée, il calcule la valeur moyenne dans le champ réceptif pour l'envoyer au tableau de sortie. [26]
- **Couche Fully-Connected (FC) :** Une fois que les caractéristiques extraites par les couches de convolution et sous-échantillonnées par les couches de pooling sont créées,

elles sont mises en correspondance par un sous-ensemble de couches FC avec les sorties finales du réseau, telles que les probabilités de chaque classe dans les tâches de classification. Alors que les couches convolutives et de pooling ont tendance à utiliser des fonctions ReLu (Unité Linéaire Rectifiée), les couches FC s'appuient généralement sur une fonction d'activation softmax pour classer les entrées de manière appropriée, en produisant une probabilité entre 0 et 1. La couche finale FC possède généralement le même nombre de nœuds de sortie que le nombre de classes.[28]

Plusieurs architectures de CNN existent, quelques-unes sont énumérées ci-dessous :

- LeNet-5 (Fameux CNN utilisé pour identifier et reconnaître des modèles dans une série de codes postaux écrits à la main [29])
- AlexNet [30]
- VGGNet (Visual Geometry Group) [31]
- Network-in-network [32]
- GoogLeNet [33]
- ResNet [34]
- ZFNet [35]

2.6.2 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (RNN) sont un type de réseau neuronal dans lequel la sortie des entrées précédentes sont utilisées comme entrée additionnelle pour l'entrée actuelle. Dans les réseaux de neurones traditionnels, toutes les entrées et sorties sont indépendantes les unes des autres, mais dans des cas où le contexte est requis, comme par exemple prédire le mot suivant d'une phrase ou résumer un texte, les mots précédents sont requis, et il est donc nécessaire de s’en souvenir.

C'est ainsi qu'est né le RNN, qui a résolu ce type de problème à l'aide d'une couche cachée. Le point le plus important qui distingue les RNN des autres types est cet état caché, qui lui permet de se souvenir des informations acquises précédemment.

Pour faire simple, un RNN se réinjecte sa sortie à l'étape suivante, formant une boucle et transmettant les informations nécessaires.

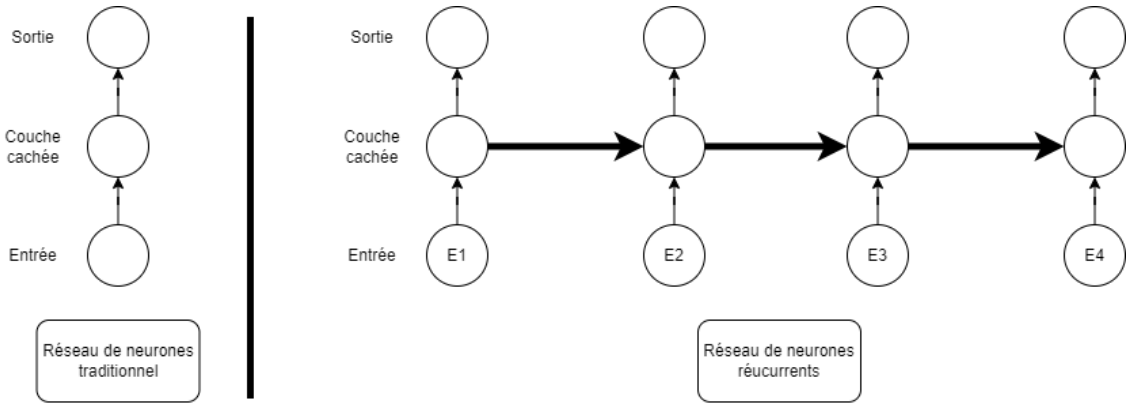

Figure 2.6.2 – NN traditionnel / RNN

Apparu dans les années quatre-vingts, les RNN sont devenus les NN les plus utilisés pour diverses tâches impliquant la notion de données séquentielles, telles que la reconnaissance vocale, la modélisation du langage, la traduction, le sous-titrage d'images, etc. Néanmoins, cette capacité de conserver les informations des entrées passées ont ouvert de nouveaux domaines de problèmes aux réseaux neuronaux, les plus courants étant les problèmes de disparition du gradient et d'explosion. Les gradients font référence aux erreurs commises lors de la formation du réseau de neurones. Si les gradients commencent à exploser, le réseau neuronal devient instable et incapable d'apprendre à partir des données de formation [36].

Ces problèmes dépendent de la taille du gradient, qui est la pente de la fonction de perte le long de la courbe d'erreur. Lorsque le gradient est trop petit, il continue à diminuer, mettant à jour les paramètres de poids jusqu'à ce qu'ils deviennent insignifiants, c'est-à-dire 0. Lorsque cela se produit, l'algorithme n'apprend plus. Les gradients explosifs se produisent lorsque le gradient est trop important, ce qui donne des poids trop importants au modèle, créant un modèle instable. Une solution à ces problèmes consiste à réduire le nombre de couches cachées dans le réseau de neurones, éliminant ainsi une partie de la complexité du modèle RNN. [37]

Il existe différents types de RNNs, avec des architectures variées. Quelques exemples sont présentés ci-dessous :

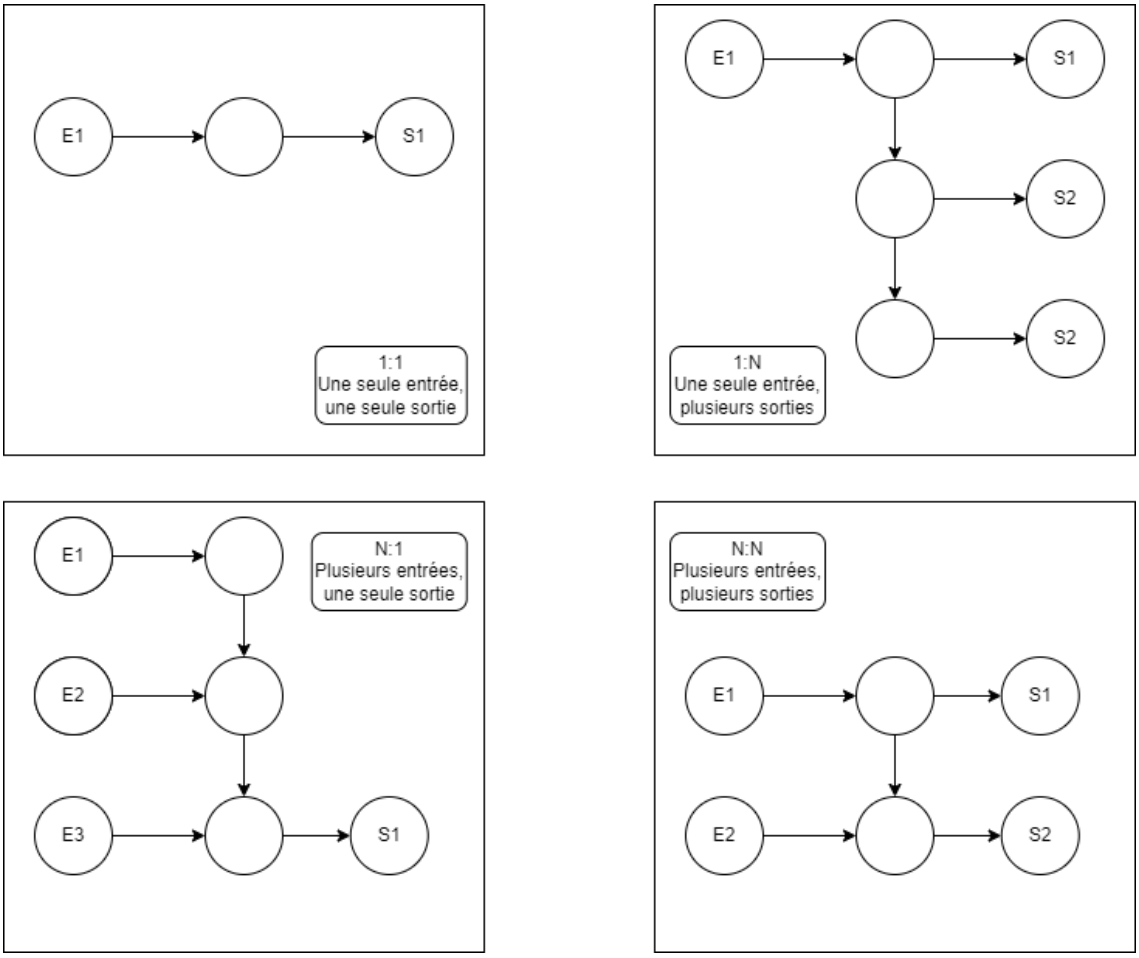

Figure 2.6.24 – Quelques types de RNN

Les réseaux de neurones récurrents ont aussi plusieurs variations d'architectures, les plus utilisées sont :

- **Long Short-Term Memory (LSTM) :** L'un des inconvénients des RNN standard étant la disparition du gradient, et vu l'inefficacité d'y remettre par un réglage optimal des paramètres dans les premières couches (Temps perdu et un grand coût en termes de puissance calcul), les réseaux à mémoire à long court terme (LSTM) sont apparus.

  Inventés par les informaticiens Sepp Hocheriez et Jurgen Schmidhuber en 1997, ils constituent une solution à ce problème. Les RNN construits avec des unités LSTM classent les données dans des cellules de mémoire à court et à long terme. Cela permet aux RNN de déterminer quelles données sont importantes et doivent être mémorisées et réinjectées dans le réseau. Cela permet également aux RNN de déterminer quelles données peuvent être oubliées, octroyant au modèle une performance encore meilleure.[38]

- **Gated Recurrent Units (GRU) :** Similaire aux LSTM, c'est une autre architecture visant à corriger le problème de mémoire à court terme. Les GRU ont une porte de réinitialisation et de mise à jour. Ces portes contrôlent la quantité et le type d'informations à conserver.
- **Bidirectional Recurrent Neural Networks (BRNN) :** Contrairement au RNNs standards qui évoluent en apprenant que dans une seule direction, les BRNNs utilisent simultanément les directions avant et arrière du flux d'informations pour apprendre, d'où la bidirectionnalité.

2.7 Conclusion

Le plus grand atout du Deep Learning est qu'il peut apprendre par lui-même à travers des données non structurées, et réagir de manière proactive à de nouvelles informations et situations, sans intervention humaine. Bien que le DL existe depuis plusieurs années, la tendance ne s'est vraiment accélérée qu'au cours des trois ou quatre dernières années. La raison est, entre autres, de meilleures ressources matérielles, des algorithmes plus sophistiqués et des réseaux de neurones mieux optimisés. L'apprentissage profond n'est pas une nouvelle approche, mais un développement de l'approche plus ancienne des réseaux neuronaux artificiels.

On a vu dans ce chapitre ce qu'est le Deep Learning, son utilité et ses diverses applications, ainsi que son fonctionnement. On a aussi vu que plusieurs architectures et plusieurs types d'algorithmes existent, leurs usages dépendants du type de traitement souhaité.

Le prochain chapitre concernera l'implémentation de l'un de ces algorithmes pour créer et tester un système biométrique multimodal.

Chapitre 3 : Contribution

3.1 Introduction

On a vu dans les chapitres précédents les notions de la biométrie, les implémentations multimodales, ainsi qu'une introduction au Deep Learning et ses diverses applications.
Dans ce chapitre, nous allons discuter la méthode proposée et appliqué pour tester le DL dans la biométrie multimodale, incluant les étapes, les divers outils utilisés, ainsi que les résultats obtenus.

3.2 Environnement et outils utilisés

Afin de réaliser notre travail, une variété de logiciels, bibliothèques, et autres outils ont été utilisés, dans un environnement qui n'a pas changé du début à la fin.

3.2.1 Matériel informatique

Ce travail a été accompli à l'aide d'un ordinateur fixe personnel, les spécifications de celui-ci sont listées ci-dessous :

CPU	Intel® Core™ i5-9400F @2.9GHz
GPU	GeForce GTX 1650 SUPER (4GB)
Taille RAM	16 Go
OS	Windows 11 Pro (Version 21H2)

Tableau 3-1 - Système utilisé

3.2.2 Logiciels et bibliothèques

- **Jupyter Notebook :** Il s'agit d'une application web client-serveur, doté d'un environnement de calcul interactif, dans lequel on peut combiner l'exécution de code, du texte riche, des mathématiques, des graphiques et autres contenus médias. Composé principalement d'un noyau (kernel) et un tableau de bord, cette application est l'environnement où notre développement se fera.
- **TensorFlow :** TensorFlow a été développé par l'équipe Google Brain et a été introduit au public pour la première fois en 2015. TensorFlow est une bibliothèque open source pour les calculs numériques et l'apprentissage automatique à grande échelle. Regroupant un ensemble de modèles et d'algorithmes d'apprentissage automatique et d'apprentissage profond, en les rendant utiles par le biais de métaphores de programme communes.

TensorFlow a été conçu pour fonctionner sur plusieurs processeurs ou GPU et même sur des systèmes d'exploitation mobiles, et il dispose de plusieurs implémentations dans plusieurs langages comme Python, C++ ou Java.[39]

- **TFLearn :** TFLearn est une bibliothèque d'apprentissage profond modulaire et transparente construite au-dessus de Tensorflow. Elle a été conçue pour fournir une API de plus haut niveau à TensorFlow afin de faciliter et d'accélérer les expérimentations, tout en restant totalement transparente et compatible avec ce dernier. L'API de haut niveau supporte actuellement la plupart des modèles d'apprentissage profond récents, tels que les CNN, LSTM, BiRNN, BatchNorm, PReLU, les réseaux résiduels, les réseaux génératifs... [40]
- **OpenCV :** OpenCV-Python est une bibliothèque Python conçue pour résoudre les problèmes de vision par ordinateur. OpenCV-Python utilise Numpy, qui est une bibliothèque hautement optimisée pour les opérations numériques avec une syntaxe de type MATLAB. Toutes les structures des tableaux d'OpenCV sont converties en tableaux Numpy et à partir de ceux-ci. Cela facilite également l'intégration avec d'autres bibliothèques qui utilisent Numpy, comme SciPy et Matplotlib.[41]

3.3 Base de données

La base de données utilisée est SWAN-Idiap[42]. C'est une BDD biométriques multimodales (visage, voix et périoculaire) acquises à l'aide d'un smartphone. Elle comprend 60 sujets capturés au cours de six sessions différentes reflétant des scénarios réels d'authentification assistée par smartphone. L'une des caractéristiques uniques de cet ensemble de données est qu'il est collecté dans quatre lieux géographiques différents, représentant une population et une ethnicité diverse. Les protocoles d'acquisition et la diversité des sujets des données recueillies dans différents lieux géographiques permettent de tester et de développer efficacement des algorithmes pour la biométrie unimodale ou multimodale.

Le tout comprend un ensemble d'environ 104 Go de photos et de vidéos, avec environ 9600 fichiers au total, chaque fichier est nommé de la même façon, à savoir :

<Site>_<identité>_<sexe>_<session>_<enregistrement>_<appareil>_<biométrie>

Où :

- ***Site :*** 1 caractère ('1' pour NTNU, '2' pour UiO, '3' pour MORPHO, '4' pour IDIAP, 5' pour HDA).
- ***Identité :*** Numéro entier de 5 caractères (allant de 00001 à 99999) associé aux participants à la collecte de données.
- ***Sexe :*** 1 caractère ('m' pour homme ou 'f' pour femme).
- ***Session :*** 2 caractères représentant un nombre entier (de 01 à 99) associé à la session (01, 02, 03, 04, 05, 06).
- ***Enregistrement :*** 2 caractères représentant un nombre entier (allant de 01 à 99) correspondant à l'enregistrement à l'intérieur de la session.
- ***Appareil :*** 1 caractère ('p' pour téléphone ou 't' pour tablette).
- ***Biométrie :*** 1 caractère ('1' pour le visage, '2' pour la voix, et '3' pour l'œil) pour les modes biométriques.

Par exemple, le fichier « **4_00016_m_04_01_p_3.mp4 »** : une vidéo ("**mp4**") de l'œil ("**3**") d'un homme ("**m**") d'IDIAP ("**4**") capturée avec un téléphone ("**p**") pendant la session "**04**" et durant l'enregistrement "**01**".

Pour les données biométriques de l'œil, contrairement aux données habituelles de l'iris, on a la biométrie périoculaire, celle-ci représente l'œil en sa totalité ainsi que son entourage.

Cette BDD contient également un ensemble de données multimodales d'attaque de présentation (Presentation Attack) ou d'usurpation d'identité utilisant des instruments d'attaque de présentation (Presentation Attack Instruments) à faible coût tel que les attaques par impression et affichage électronique. Ces données ne seront pas utilisées dans notre travail, mais peuvent être utilisés dans d'éventuels travaux futurs pour tester et prévenir des failles de sécurité, améliorant le résultat final. Des informations supplémentaires sont disponibles directement en contactant le centre de recherche IDIAP. [43][44]

3.4 Préparation des données

Afin de faciliter l'utilisation des données, chaque type a été séparé en dossier, à l'aide de python, un script a été écrit pour parcourir les dossiers et copier tout fichier catégorisé comme biométrie du visage ('1'), que ce soit une vidéo, ou une simple image. À la fin, on a un total de 1800 fichiers de la biométrie du visage, comprenant des vidéos sans son ainsi que des images

PNG, 2400 fichiers vidéo pour la biométrie de la voix, et 4200 fichiers vidéo sans son et images PNG pour la biométrie de la région périoculaire.

Les images en format PNG du visage étaient en format paysage tourné à 90° dans le sens inverse des aiguilles, pour remédier à cela une rotation de -90° a été effectué sur ces images au biais de la bibliothèque imutils (Bibliothèque de fonctions utiles pour le traitement d'images).
Pour les vidéos de la biométrie type visage, elles avaient généralement une longueur de 5 à 6 secondes, avec un nombre d'IPS (images par seconde) variant, une extraction a été faite en lisant la vidéo image par image, puis en extrayant une image, chaque 30 images.
Afin d'augmenter le nombre de données pour le visage, et de rendre utile les données biométriques de la voix (qui ne seront pas directement utilisés), l'opération précédente (extraction à partir de vidéos) a été appliqué sur les données biométriques du type voix, cela nous a fournis non seulement un bien plus grand nombre d'images (augmentation de 24000+ images), mais aussi des données plus variées (visages au moment de parler).
Pour les deux types (PNG et MP4), l'image finale a été réduite au format 125x125 pixels, afin de réduire le cout du processus ainsi que le temps d'entrainement et de tests, toutes ces opérations ont été accomplies grâce à la bibliothèque OpenCV. Au final, le nombre total d'images acquises a atteint 33910 images.
Pour notre deuxième modalité biométrique, qui est la région périoculaire, un traitement similaire a été appliqué, à savoir, rotation de -90°, capture des images à partir des vidéos, et redimensionnement en 125x125 pixels, cependant, une étape de recadrage a été incluse pour ne prendre que la partie périoculaire gauche du visage, pour cela, un calcul approximatif a été fait en utilisant un pourcentage des dimensions de l'image. Pour cette modalité, 28261 images ont été extraites.
Enfin, pour les données multimodales, une fusion 'au niveau des capteurs' a été faite, parcourant la liste des images pour la biométrie du visage, chaque élément a reçu une image périoculaire de la même personne, jusqu'à ce que toutes les images périoculaires soient fusionnées. D'abord, en essayant de trouver une image équivalente de la même session, en remplaçant le caractère de la biométrie, par '3' (œil), exemple :

'*4_00001_m_01_01_p_1_vid660*' → '*4_00001_m_01_01_p_3_vid660*'

Ou bien en prenant un élément au hasard dans une liste précédemment créé et trié, contenant le nom de toutes les images périoculaires d'une personne. Le nombre total d'images acquises fusionné est de 28261 aussi.

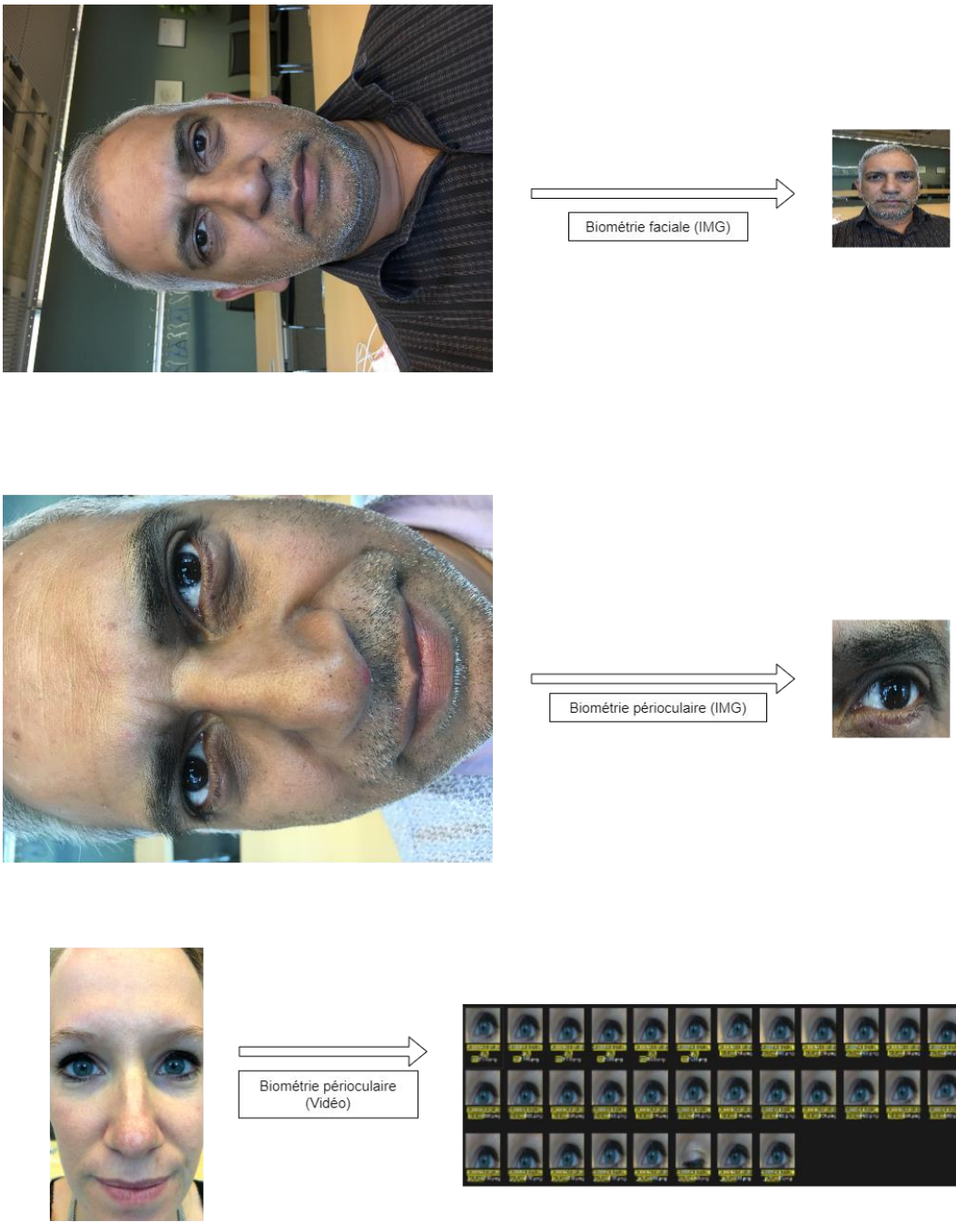

Figure 3-1 – Exemple de traitement des données

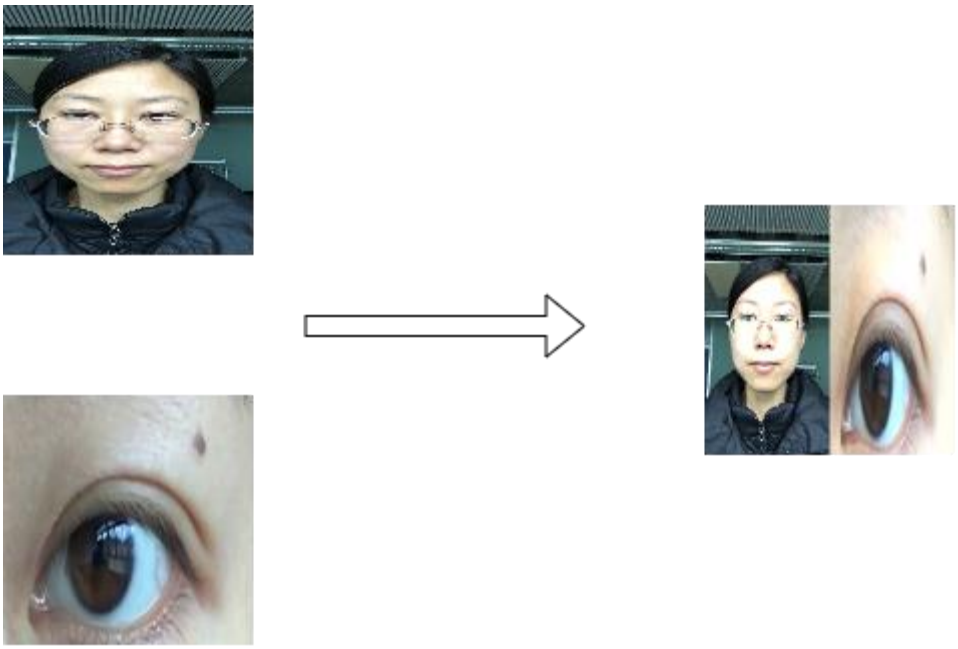

Figure 3-2 – Fusion des images

3.4 Création du modèle

Pour faire notre apprentissage et tester le modèle résultant, le modèle VGG16 a été utilisé comme base et modifié pour être moins couteux en termes de ressources et prendre moins de temps.

De base, VGG16 prends en entrée une image 224x224 pixels avec 3 canaux pour les couleurs RGB, l'image passe par les deux premières couches de convolution de la très petite taille réceptive de 3 x 3, suivie par des activations ReLU. Chacune de ces deux couches contient 64 filtres. Le pas de convolution est fixé à 1 pixel, et le padding à 1 pixel. Cette configuration préserve la résolution spatiale, et la taille de la carte d'activation de sortie est la même que les dimensions de l'image d'entrée. Les cartes d'activation sont ensuite passées dans un pooling spatial max sur une fenêtre de 2 x 2 pixels, avec un stride de 2 pixels. Cela divise par deux la taille des activations. Ainsi, la taille des activations à la fin de la première pile est de 112 x 112 x 64. Les activations passent ensuite par une deuxième pile similaire, mais avec 128 filtres contre 64 dans la première. Par conséquent, la taille après la deuxième pile devient 56 x 56 x 128. Vient ensuite la troisième pile avec trois couches convolutives et une couche de max pooling. Le nombre de filtres appliqués ici est de 256, ce qui fait que la taille de sortie de la pile est de 28 x

28 x 256. Elle est suivie de deux piles de trois couches convolutives, chacune contenant 512 filtres. La sortie à la fin de ces deux piles sera de 7 x 7 x 512.

Les piles de couches convolutives sont suivies de trois couches entièrement connectées, entre lesquelles se trouve une couche d'aplatissement. Les deux premières ont 4 096 neurones chacune, et la dernière couche entièrement connectée sert de couche de sortie et a 1 000 neurones correspondant aux 1 000 classes possibles pour le jeu de données ImageNet. La couche de sortie est suivie par la couche d'activation Softmax utilisée pour la classification catégorielle.

Pour ce travail, le modèle a reçu quelques modifications, en commençant par le type d'entrés, les images qu'il traitera seront 125x125 pixels, les 3 canaux de couleurs sont conservés. Toutes les couches convolutives suivantes avant d'arriver à la couche entièrement connectée ont vu leurs nombres de filtres réduit par 50%, à savoir les deux premières passent de 64 à 32, les deux d'après passent de 128 à 64, et ainsi de suite (256 > 128, 512 > 256). Les deux couches entièrement connectées qui avaient 4096 neurones ont été réduites à une seule, et le nombre de neurones a été réduit à 1024. La couche d'après est une couche de 'dropout', fixé à 0.5 (50%), le dropout va nous aider à combattre un éventuel overfitting, forçant les nœuds d'une couche à assumer de manière probabiliste plus ou moins de responsabilité pour les entrées, ce qui va nous mener à un modèle plus robuste.

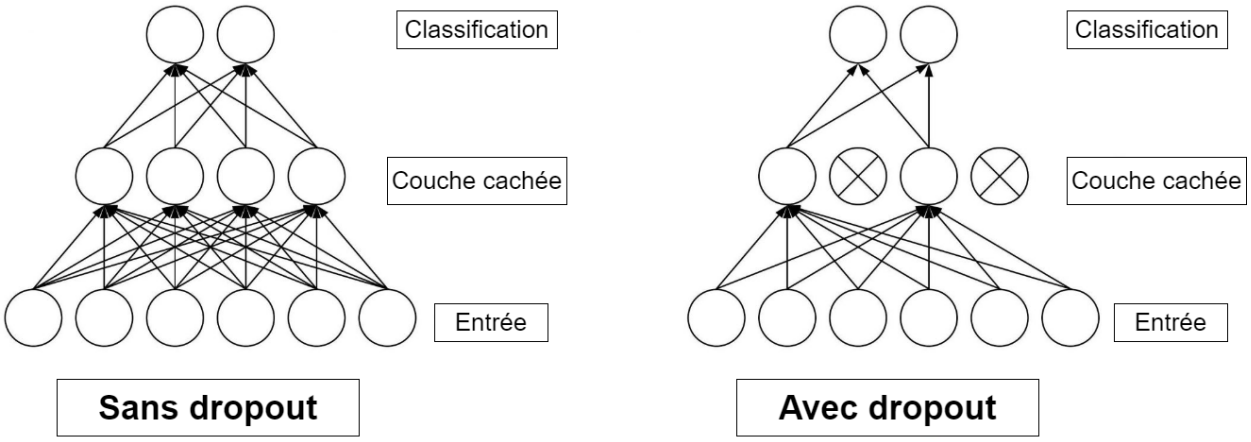

Figure 3-3 – Fonctionnement du dropout

Enfin pour la dernière couche entièrement connectée, le nombre de neurones a été réduit de 1000 à 60, afin de différencier nos 60 personnes dans la BDD. Ces changements serviront

principalement à rendre le processus d'apprentissage moins couteux, et mieux adapté à notre matériel.

La taille 3x3 a été conservée pour le noyau, ainsi que la fonction d'activation ReLu (Rectified Linear Unit), ce qui va nous aider à réduire le temps d'entrainement et d'évaluation due au calcul simple de ReLu.

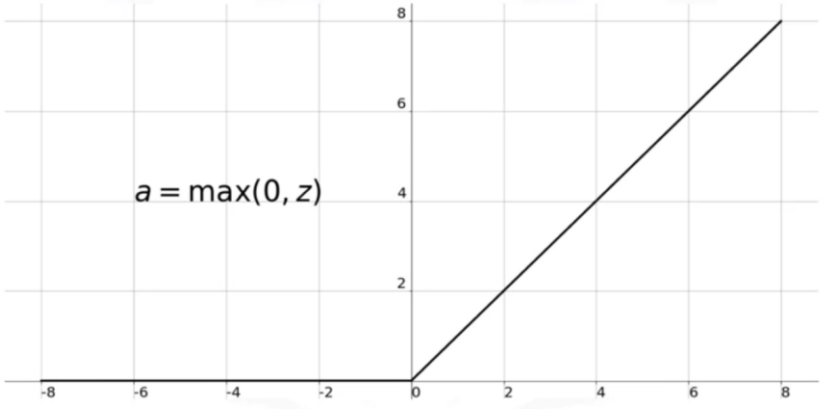

Figure 3-4 – Fonction d'activation ReLu

Pour la dernière couche entièrement connectée de 60 neurones (60 classes, une pour chaque personne), la fonction d'activation est une fonction softmax, une fonction classique utilisée dans la couche de sortie des modèles ayant des problèmes de classification à plus de deux types de classes. Son activation produira une valeur pour chaque neurone de la couche de sortie. Ces valeurs représentent des probabilités et la somme des valeurs est égale à 1.

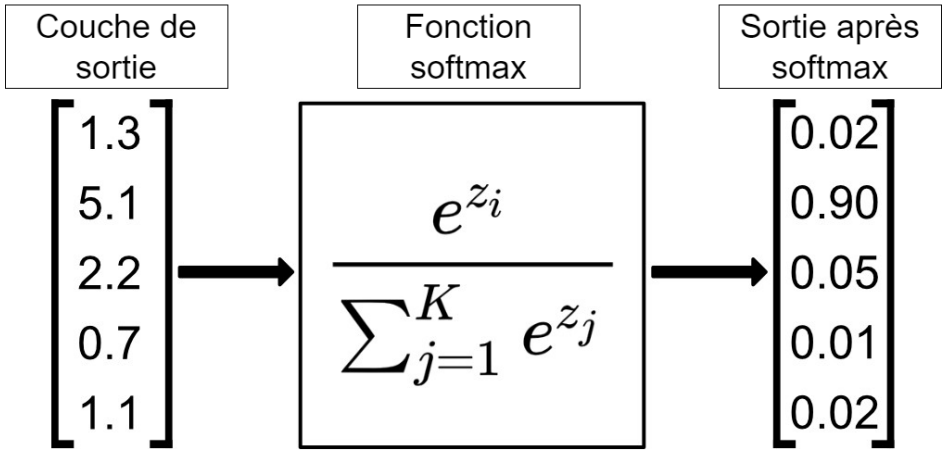

Figure 3-5 – Exemple d'un traitement par softmax

Pour notre optimiseur, RMSprop est utilisé, c'est une technique d'optimisation similaire à la descente de gradient avec élan (momentum), il se concentre principalement sur l'accélération du processus d'optimisation en diminuant le nombre d'évaluations de fonctions pour atteindre les minima locaux. L'algorithme conserve la moyenne mobile des gradients carrés pour chaque poids et divise le gradient par la racine carrée du carré moyen [45]. Enfin pour notre fonction de coût, 'categorical_crossentropy' a été utilisé, qui est bien adapté pour la classification multi-classe :

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{output}\\ \text{size}}} y_i \cdot \log \hat{y}_i$$

3.5 Apprentissage et résultats

Pour tous les apprentissages qu'on a faits, aucun paramètre n'a été changé, les taux d'apprentissages ont été fixé à 0.0001, et le nombre d'époques à 14. Pour le batch size, il a été fixé à 32.

Commençant par la modalité du visage, cet apprentissage s'est fait sur 33910 images, dont 20% ont été utilisé pour la validation, résultant à 27128 images pour l'apprentissage et 6782 pour la validation.

Dans un environnement où ce modèle pourrait être déployé, une caméra sera nécessaire afin de capturer le visage d'une personne, la personne devra soit se positionner de façon à ce que le capteur ne capture que le visage, ou bien notre système prendras des photos complètes, et se chargera de recadrer l'image sur le visage. Le traitement restant sera simplement de redimensionner l'image (125x125).

Le résultat se trouve ci-dessous :

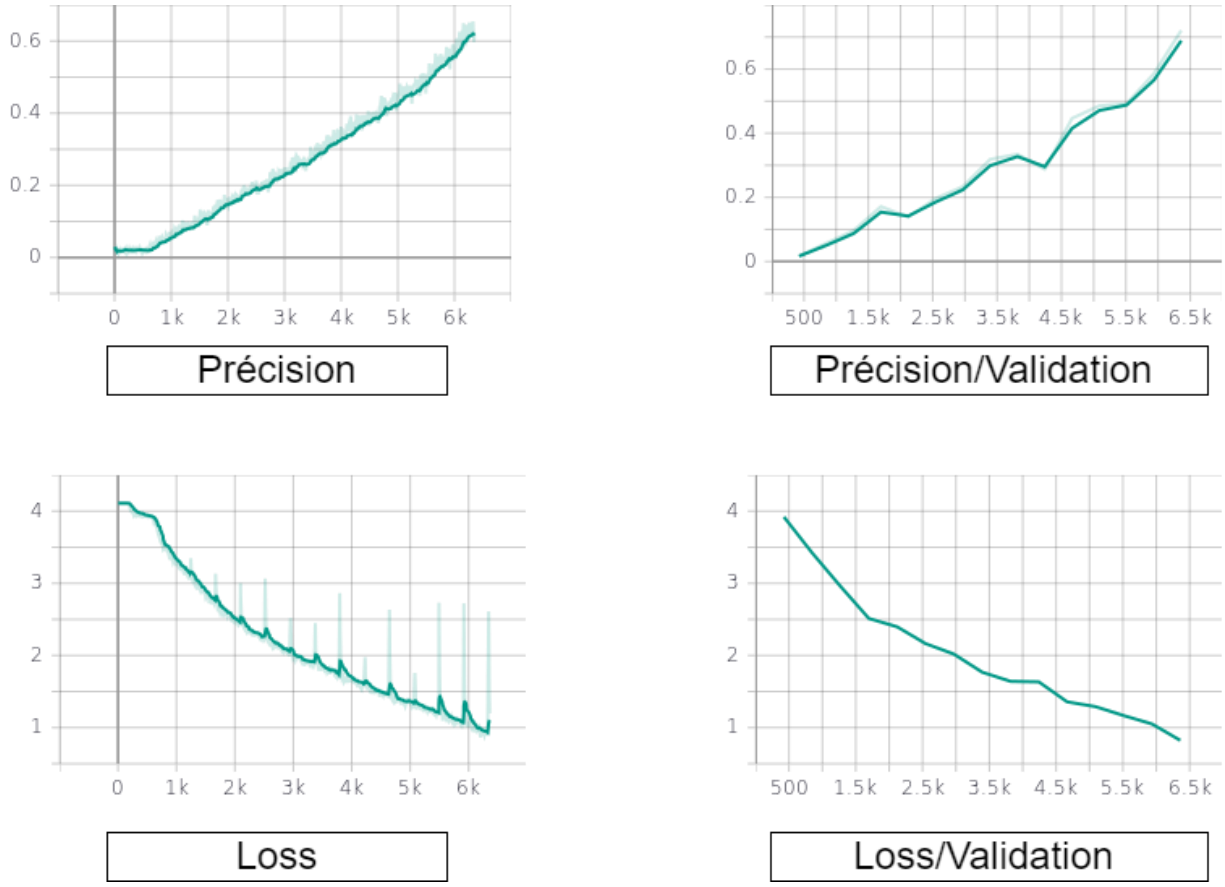

Figure 3-6 – Résultat de l'apprentissage du visage en graphes

L'apprentissage s'est terminé avec une précision finale de 62.35%, et une précision de validation 72.07%, bien qu'elle soit assez faible, on peut voir que notre modèle apprend à un rythme ordonné, les graphes de validation nous montre aussi qu'il n y'a pas eu de sur-apprentissage. On peut en conclure que le choix de réduire la complexité du modèle VGG16 vers un modèle plus simple a été correct. La résolution choisie (125x125x3) ainsi que la quantité de données semblent aussi être adéquats. Le graphe de la fonction de perte semble aussi normal, avec quelques rares pics qui apparaissent brièvement.

Un point qu'on pourrait améliorer cependant est la longueur de l'apprentissage (nombre d'époques) , un point qu'on pourra explorer plus tard.

Pour la suite, on est passé à l'apprentissage de la région périoculaire, cette fois ci une éventuelle mise en place se fera presque de la même façon, la différence étant que l'image finale devra être focalisé sur la partie périoculaire du visage.

Cet apprentissage s'est déroulé sur 28261 images au total, prenant 28261 pour l'apprentissage et 22608 pour la validation, les résultats sont comme suivent :

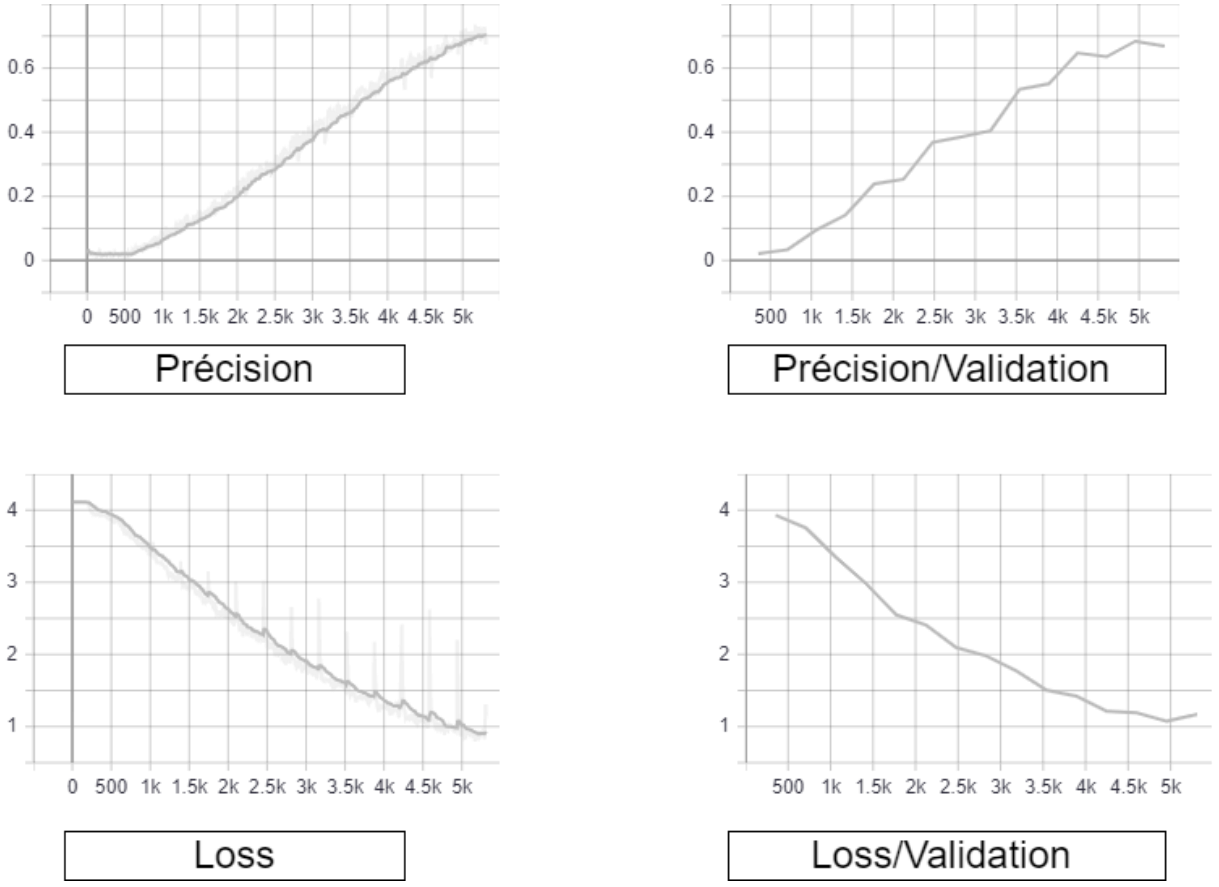

Figure 3-7 – Résultat de l'apprentissage de la région périoculaire en graphes

Similairement à l'apprentissage du visage, notre modèle semble assez bien adapté à traiter nos données, achevant cette fois-ci une précision de 67.39% (66.82% pour la validation), pour les autres points, à savoir, pas d'anomalies entres la précision d'apprentissage et la précision de validation, et un bon graphe pour la fonction de perte (à part les petits pics).

Comparant ces deux résultats :

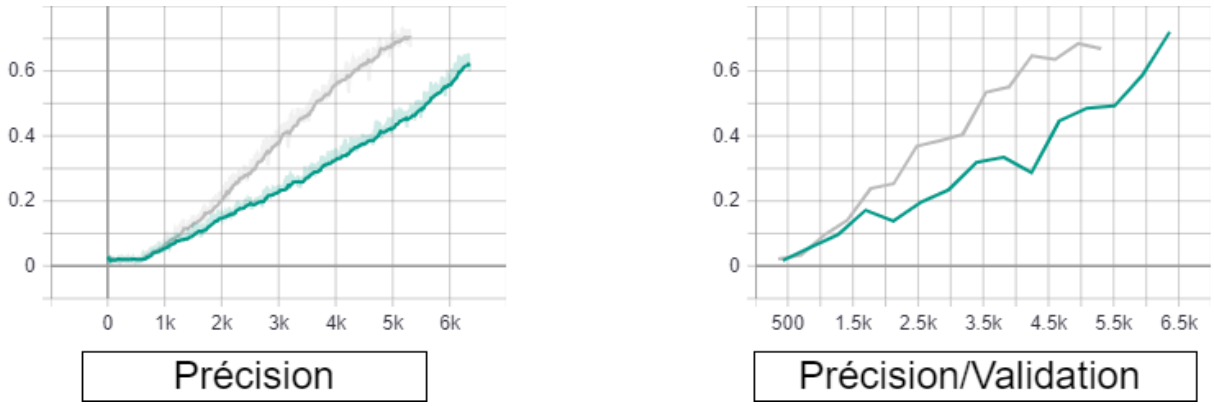

Figure 3-8 – Comparaison entre apprentissage visage/région périoculaire

On constate une nette différence entre la courbe du visage (en bas) et celle de la région périoculaire (en haut), on présume qu'elle est dû à la simplicité relative de cette dernière, ce qui rend son apprentissage plus facile. On voit aussi un petit décalage par rapport aux courbes de l'apprentissage du visage, ceci est dû à la différence entre la quantité des données, pour la région périoculaire, seulement 28261 images ont été traité, cette différence est néanmoins négligeable, grâce à la simplicité de ces données, nécessitant une moindre quantité.

Pour notre dernier apprentissage, nous passons à l'apprentissage multimodal, on utilisera dans ce système les deux modalités précédentes : visage et partie périoculaire. Une éventuelle implémentation pourrait se faire de quelques manières différentes. Par exemple, une seule caméra pourrait être mise en place, et une photo serrait prise du visage de la personne, après cela, soit la partie périoculaire sera isolé du visage, créant une deuxième photo à partir de la première, soit une deuxième photo sera prise, se focalisant sur la région périoculaire. Une autre façon serait de mettre en place deux caméras différentes, une pour le visage, et une pour la région périoculaire. Enfin, pour atteindre la multimodalité, ces deux modalités seront ensuite concaténées (d'où la fusion) afin de produire une nouvelle image avec le visage à gauche, et la région périoculaire à droite, avec un redimensionnement final qui nous fournira une image de 125x125 pixels. Cette image sera ensuite traitée par notre système afin de reconnaitre où de vérifier la personne.

Le résultat de cet apprentissage est comme suit :

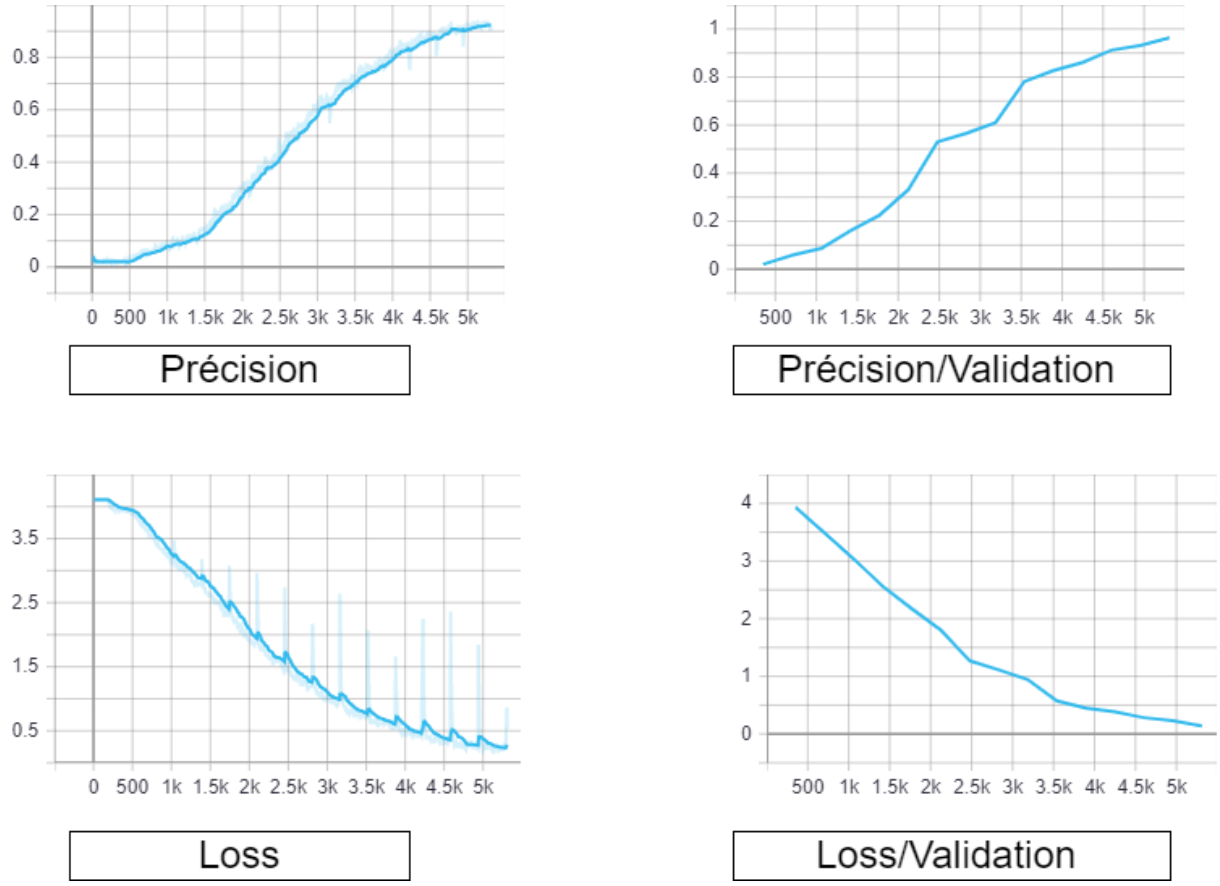

Figure 3-9 - Résultat de l'apprentissage multimodal en graphes

On constate immédiatement une nette amélioration par rapport aux apprentissages à traits unique (unimodaux), atteignant cette fois-ci une respectable précision de 92.12%, avec une précision de validation de 96.60%.

La courbe d'apprentissage semble aussi montrer une tendance de stabilisation vers la fin, ce qui nous laisse à penser qu'on est assez près de l'efficacité maximale atteignable sur ce modèle avec ces données.

Comparant ces résultats aux précédents :

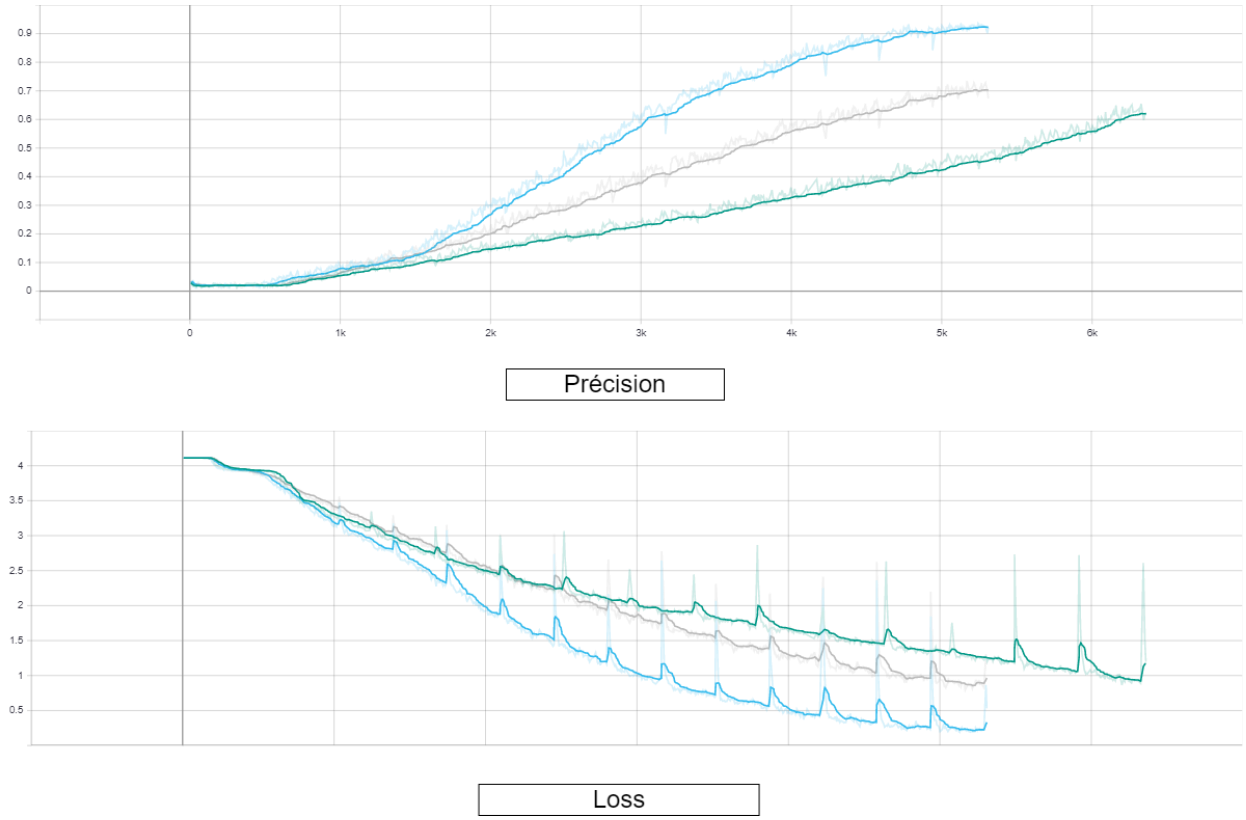

Figure 3-10 – Comparaison entre les différents apprentissages

On voit que durant les premières étapes, l'apprentissage est similaire à celui du visage, puis la courbe se sépare et suit une montée plus prononcée, dépassant celle de la région périoculaire quelque peu après.

Nous pouvons donc voir qu'une implémentation multimodale améliore effectivement l'efficacité de notre modèle. Bien que notre apprentissage se soit passé sans problèmes apparents (à part les pics, pas d'overfitting/underfitting, etc…), il nous semble que l'apprentissage pourrait utiliser un peu plus de temps, on a donc refait le travail, cette fois-ci en augmentant le nombre de d'époques de 14 à 18, une augmentation de 28%~, ce qui nous semble assez pour voir une amélioration, tout en évitant un éventuel sur-apprentissage.

Les résultats obtenus sont affichés ci-dessous :

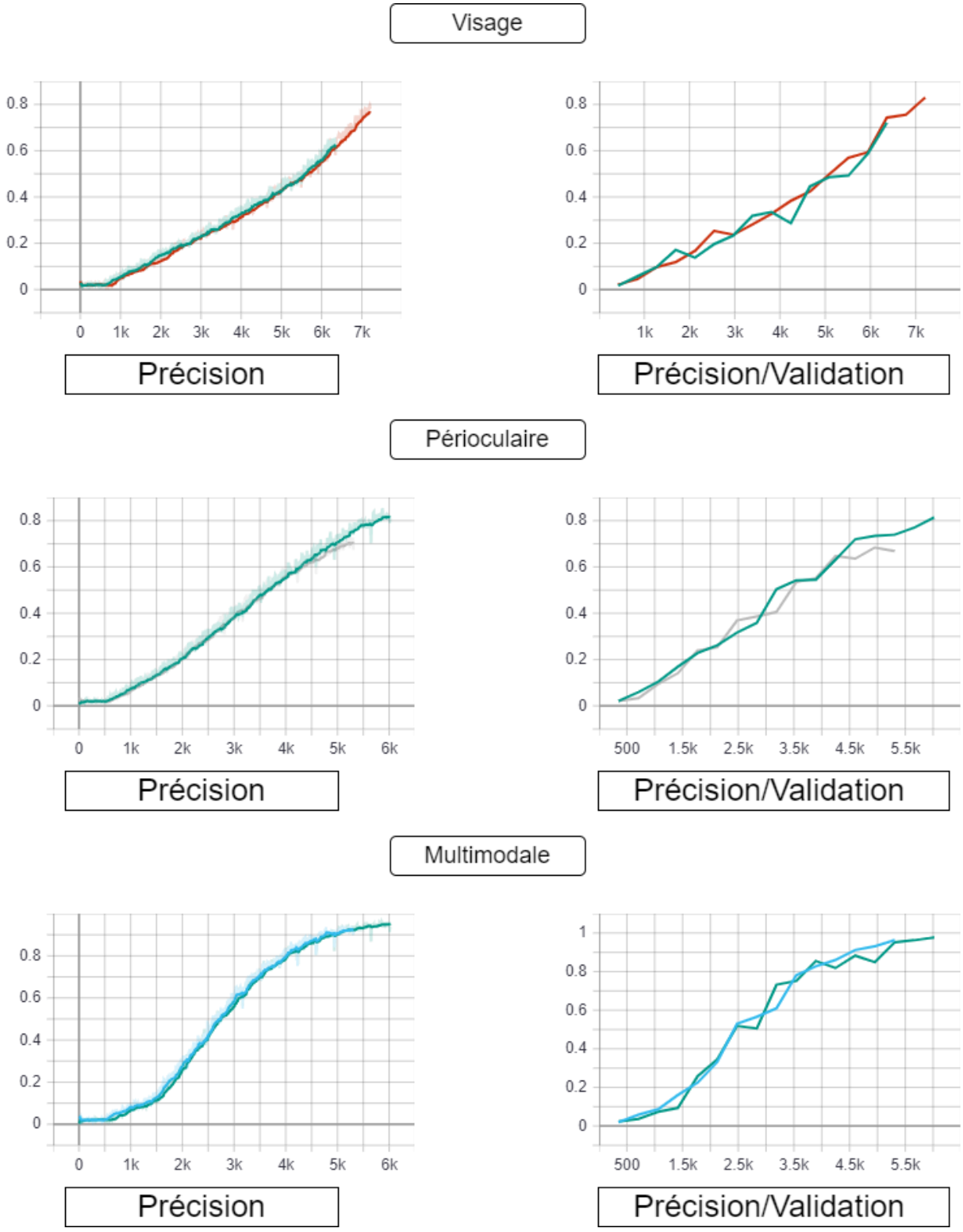

Figure 3-11 – Comparaison entre les résultats sur un nombre d'époques différents

Pour faire court :

Mesure | Nb époques	14	18
Visage	62.35%	78.15%
Périoculaire	67.39%	81.17%
Multimodale	92.12%	93.51%

Tableau 3-2 - Résultats finaux

On constate une grande amélioration de 20%~ et 25%~ pour les mesures unimodales, qui semblent encore améliorables, par contre, la mesure multimodale a reçu une très faible amélioration, en accordance avec notre lecture précédente du graphe.

3.6 Conclusion

Les résultats précédents nous montrent clairement que la fusion (au niveau de capteurs) de la biométrie du visage ainsi que celle périoculaire, apporte une amélioration claire dans les performances du modèle, et cela, avec les mêmes paramètres.

Avec un réglage fin et spécifique à notre type de données, on pense que les résultats seront bien meilleurs, d'ailleurs même notre BDD n'était parfaite, avec plusieurs anomalies éparpillées à travers les données, le modèle crée s'est néanmoins montré capable de déchiffrer le code, et d'achever des résultats assez corrects.

Conclusion Générale

Bien que l'apprentissage profond existe depuis longtemps, ce n'est que durant la dernière décennie que cette technologie a pris son envol, en partant de la simple réplication de quelques taches jusqu'à l'achèvement de taches pensé être impossible, fournissant à l'humanité des avancés spectaculaires dans la plupart des domaines. L'un de ces domaines est la biométrie.

On a exploré à travers ce travail l'utilisation du DL dans ce domaine, plus précisément, dans la biométrie multimodale, une technique moins répandue que l'unimodale, mais qui semble prendre de l'élan grâce aux résultats obtenus jusqu'à maintenant. Pour notre évaluation de cette technique, on a choisi d'utiliser un CNN, en démarrent à partir d'un modèle connu (VGG16), et le modifiant pour obtenir un modèle plus adéquat à notre BDD.

Les résultats obtenus nous ont assez surpris, achevant une efficacité assez passable pour les mesures unimodales, puis voir cette efficacité sauter vers des valeurs bien plus hautes et prometteuses à travers la multimodalité, prouvant ainsi que même des modèles CNN faiblement étudiés peuvent profiter des avantages de la biométrie multimodale.

Ce travail fournit aussi une réponse utile en ce qui concerne l'utilisation du visage ainsi que la région périoculaire comme deuxième modalité. Cette combinaison semble l'une des plus simples, pouvant utiliser le même capteur (caméra) et étant moins intrusive que certaines modalités, son application semble naturelle.

D'autres modalités restent à explorer, sur des niveaux de fusion différents, et même l'utilisation de plus de deux modalités (ex : visage, œil, démarche). Une autre possibilité qui peut d'ailleurs être testé sur cette même base de données est l'ajout de la reconnaissance de la voix comme autre modalité, contrairement aux modalités qui utilisent une caméra comme capteur, la détection de la voix n'est pas limitée par la luminosité, les éventuelles obstructions (masque, lunettes de soleil, etc.) et autres problèmes. D'ailleurs, ce système pourrait combler les lacunes de la voix (environnement bruyant, etc.) avec le visage, améliorant la confidence du résultat final, tout en restant simple à utiliser.

Bibliographie

1. Biométrie | CNIL. (2016). Cnil.fr. https://www.cnil.fr/fr/biometrie. **Consulté le 25 Juin 2022**
2. Biometrics Institute. What is Biometrics? (2020, November 11). https://www.biometricsinstitute.org/what-is-biometrics/. **Consulté le 25 Juin 2022**
3. W. Meng, D. S. Wong, S. Furnell and J. Zhou, "Surveying the development of biometric user authentication on mobile phones", IEEE Commun. Surveys Tut., vol. 17, no. 3, pp. 1268-1293, 3rd Quart. 2015.
4. RecFaces. Biometric Devices — Complete Guide on Technology. (2021, August 4). https://recfaces.com/articles/articles-biometric-devices#9. **Consulté le 25 Juin 2022**
5. Das, R. Hand geometry recognition: pros and cons - Keesing Platform. (2019, December 27. https://platform.keesingtechnologies.com/pros-cons-hand-geometry/. **Consulté le 25 Juin 2022**
6. Sanjekar, Priti & Patil, Jayantrao. An Overview of Multimodal Biometrics. Signal & Image Processing: An International Journal (SIPIJ) Vol. 4. 57-64. 2013.
7. Wenyi Zhao, Rama Chellappa. Multimodal Biometrics: Augmenting Face With Other Cues, Chapter 21. Academic Press, 2006.
8. Smith, C. L., & Brooks, D. J. Integrated Identification Technology. Security Science, 153–175. 2013.
9. A. Ross and A. Jain, "Information fusion in biometrics", Pattern Recognit. Lett., vol. 24, no. 13, pp. 2115-2125, 2003.
10. Ross A., Nandakumar K. Fusion, Score-Level. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. 2009.
11. Poh N. Fusion, Quality-Based. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. 2009.
12. M. O. Oloyede and G. P. Hancke, "Unimodal and Multimodal Biometric Sensing Systems: A Review," IEEE Access, vol. 4, pp. 7532-7555, 2016.
13. Y.-F. Yao, X.-Y. Jing, and H.-S. Wong, ``Face and palmprint feature level fusion for single sample biometrics recognition," Neurocomputing, vol. 70, nos. 7-9, pp. 1582-1586, Mar. 2007.

14. T. Konishi, T. Kubo, K. Watanabe, and K. Ikeda, ``Variational Bayesian inference algorithms for infinite relational model of network data," IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 9, pp. 2176-2181, Sep. 2015.
15. Ullah, I., Youn, J., & Han, Y.-H. Multisensor Data Fusion Based on Modified Belief Entropy in Dempster–Shafer Theory for Smart Environment. IEEE Access, 9, 37813–37822. 2021.
16. C. Smaili, et al, Multi-sensor Fusion Method Using Dynamic Bayesian Network for Precise Vehicle Localization and Road Matching, 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), pp. 146-151. 2007.
17. P. K. Atrey, M. A. Hossain, A. El Saddik and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey", Multimedia Syst., vol. 16, no. 6, pp. 345-379, 2010.
18. Christopher Berner. Et Al. 'Dota 2 with Large Scale Deep Reinforcement Learning', p. 8. December 13, 2019.
19. NVIDIA A100 GPUs Power the Modern Data Center. (2022). https://www.nvidia.com/en-us/data-center/a100/. **Consulté le 25 Juin 2022**
20. Alan Turing and the beginning of AI | Britannica. In Encyclopædia Britannica. 2022. https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI. **Consulté le 25 Juin 2022**
21. Reyes, K. What is Deep Learning and How Does It Works [Explained]. 22 Avril 2020. https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning. **Consulté le 25 Juin 2022**
22. ImageNet: A Pioneering Vision for Computers - History of Data Science. 27 Août 2021. https://www.historyofdatascience.com/imagenet-a-pioneering-vision-for-computers/. **Consulté le 25 Juin 2022**
23. Long, M. Deep Learning in Healthcare and Radiology | Aidoc Blog. 17 Février 2020. https://www.aidoc.com/blog/deep-learning-in-healthcare/. **Consulté le 25 Juin 2022**
24. Chen B, Marvin S, While A. Containing COVID-19 in China: AI and the robotic restructuring of future cities. Dialogues in Human Geography. p. 238-241. 11 Juin 2020.
25. Atlas™ | Boston Dynamics. 2022. https://www.bostondynamics.com/atlas. **Consulté le 25 Juin 2022**
26. IBM Cloud Education. What are Convolutional Neural Networks? 20 Octobre 2020. https://www.ibm.com/cloud/learn/convolutional-neural-networks. **Consulté le 25 Juin 2022**

27. Yamashita, R. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018).
28. Saha, S. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. 5 Décembre 2018. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. **Consulté le 25 Juin 2022**
29. Yann LeCun. Et al. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE. Novembre 1998.
30. Alex Krizhevsky. Et al. ImageNet Classification with Deep Convolutional Neural Networks. NIPS. 2012.
31. K. Simonyan. A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition International Conference on Learning Representations, 2015.
32. Min Lin. Et al. Network in Network. 16 Décembre 2013.
33. Christian Szegedy. Et al. Going Deeper with Convolutions. 17 Septembre 2014.
34. Kaiming He. Et al. Deep Residual Learning for Image Recognition. 10 Décembre 2015.
35. Matthew D Zeiler. Rob Fergus. Visualizing and Understanding Convolutional Networks. 12 Novembre 2013.
36. Laskowski, N., & TechTarget Contributor. Recurrent Neural Networks. 2021. https://www.techtarget.com/searchenterpriseai/definition/recurrent-neural-networks. **Consulté le 25 Juin 2022**
37. IBM Cloud Education. What are Recurrent Neural Networks? 14 Septembre 2020. https://www.ibm.com/cloud/learn/recurrent-neural-networks. **Consulté le 25 Juin 2022**
38. Martin Sundermeyer. Et al. LSTM Neural Networks for Language Modeling. INTERSPEECH 2012 ISCA's 13th Annual Conference Portland, OR, USA. 9-13 Septembre 2012.
39. TensorFlow. https://www.tensorflow.org/. **Consulté le 25 Juin 2022**
40. Damien, A. TFLearn | TensorFlow Deep Learning Library. http://tflearn.org/. **Consulté le 25 Juin 2022**
41. OpenCV: Introduction to OpenCV-Python Tutorials. 2013. https://docs.opencv.org/3.4/d0/de3/tutorial_py_intro.html. **Consulté le 25 Juin 2022**
42. Idiap Research Institute. Artificial Intelligence for Society. website: https://www.idiap.ch/en

43. SWAN-Idiap. Idiap Research Institute, Artificial Intelligence for Society. https://www.idiap.ch/en/dataset/swan. **Consulté le 25 Juin 2022**
44. Ramachandra, et al. (2020). SWAN-Idiap [Data set]. https://doi.org/10.34777/61hv-cm07
45. A Comprehensive Guide on Deep Learning Optimizers. 7 Octobre 2021. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/ **Consulté le 25 Juin 2022**

Résumé

Aujourd'hui, l'intelligence artificielle est utilisée presque partout dans notre vie, l'une de ses applications se trouve être la biométrie, permettant l'identification des personnes à travers de nouveaux traits complexes, l'amélioration des ceux utilisés auparavant, ainsi qu'une meilleure facilité d'utilisation. La biométrie n'est cependant pas infaillible, et peut échouer dans certains rares cas.

Le Deep Learning est l'un des sous-domaines de l'intelligence artificielle, basé sur des réseaux de neurones artificiels, il permet l'accomplissement de tâches complexes comme la reconnaissance d'objets, les prédictions d'évènements futurs, et bien d'autres, avec une très haute performance. D'où son utilisation déjà présente dans la biométrie, aidant à identifier les personnes à travers des traits complexes comme le visage, la démarche, la voix, et plusieurs autres traits.

Dans ce travail, nous allons explorer et tester l'utilisation du Deep Learning dans la détection biométrique multimodale, à savoir, l'utilisation de plus d'un trait simultanément. On a utilisé la base de données SWAN-Idiap ainsi qu'un réseau de neurones convolutifs modifié à partir d'un autre modèle. Commençant par tester les performances d'un système unimodal sur les deux modalités du visage et de la région périoculaire, puis on passera au multimodal, qui utilisera une fusion de ces deux-là, tout en collectant et en comparant les données résultantes au long du travail.

Mots clés : Système biométrique multimodal, apprentissage profond, réseaux de neurones convolutifs, reconnaissance du visage, reconnaissance périoculaire.

Abstract

Today, artificial intelligence is used almost everywhere in our life, one of its applications is biometrics, allowing the identification of people through new complex traits, improving the ones used before, as well as a better usability. However, biometrics are not infallible, and can fail in some rare cases.

Deep Learning is one of the subfields of artificial intelligence, based on artificial neural networks, it allows the accomplishment of complex tasks such as object recognition, predictions of future events, and many others, with a very high performance. Hence its use in biometrics, helping to identify people through complex features such as face, gait, voice, and several other features.

In this work, we will explore and test the use of Deep Learning in multimodal biometric recognition, namely, the use of more than one trait simultaneously. We used the SWAN-Idiap database as well as a convolutional neural network modified from another model. Starting by testing the performance of a unimodal system on the two modalities of the face and the periocular region, then moving on to the multimodal, which will use a fusion of these two, while collecting and comparing the resulting data throughout the work.

Keywords: Multimodal biometric system, deep learning, convolutional neural networks, face recognition, periocular recognition.

ملخص

في الوقت الحاضر، يتم استخدام الذكاء الاصطناعي تقريبا في كل مكان في حياتنا، ويصادف أن يكون أحد تطبيقاته القياسات الحيوية، التي تسمح بتحديد الأشخاص من خلال سمات معقدة جديدة، وتحسين السمات المستخدمة من قبل، وكذلك سهولة الاستخدام بشكل أفضل. ومع ذلك، فإن القياسات الحيوية ليست معصومة من الخطأ، ويمكن أن تخفق في بعض الحالات النادرة.

التعلم العميق هو أحد الحقول الفرعية للذكاء الاصطناعي، ويعتمد على الشبكات العصبية الاصطناعية، ويسمح بإنجاز المهام المعقدة مثل التعرف على الأشياء، والتنبؤ بالأحداث المستقبلية، وغيرها كثيرا، بأداء عالي. ومن ثم فإن استخدامه موجود بالفعل في القياسات الحيوية، مما يساعد على تحديد الأشخاص من خلال السمات المعقدة مثل الوجه والمشية والصوت والعديد من السمات الأخرى.

في هذا العمل، سوف نكتشف ونختبر استخدام التعلم العميق في الكشف عن القياسات الحيوية متعددة الوسائط، أي استخدام أكثر من سمة في وقت واحد. استخدمنا قاعدة بيانات SWAN-Idiap بالإضافة إلى شبكة عصبية التفافية معدلة من نموذج آخر. بدءًا من اختبار أداء من نظام أحادي الوسائط على طريقتي الوجه والمنطقة المحيطة بالعين، سننتقل بعد ذلك إلى الوسائط المتعددة، والتي سيستخدم اندماجا بين هذين الاثنين مع جمع ومقارنة البيانات الناتجة طول العمل.**الكلمات المفتاحية :** نظام القياسات الحيوية متعدد الوسائط، التعلم العميق، الشبكات العصبية الالتفافية، التعرف على الوجه، التعرف على ما حول العين.