Ministère de l'enseignement Supérieur et de la recherche Scientifique
وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculté de Technologie

Département Informatique

جامعة باجي مختار – عنابـــة

كـليـة التكنـولـوجيـا

قســم الاعلام الالـي

Thèse

Présentée pour obtenir le diplôme de

Doctorat En-Sciences

Spécialité : Informatique

Par :
Yamina Bordjiba

Thème :

Animation du Visage à base de Reconnaissance des Expressions Faciales

Thèse soutenue le 11 Octobre 2023 devant le jury composé de :

N°	Nom et prénom	Grade	Etablissement	Qualité
01	Hakim Bendjenna	Prof.	Université Larbi Tebessi - Tebessa	Président
02	Hayet Farida Merouani	Prof.	Université Badji Mokhtar -Annaba	Rapporteur
03	Akila Djebbar	MCA	Université Badji Mokhtar -Annaba	Examinateur
04	Nawel Zemmal	MCA	Université Mohamed-Chérif Messaadia - Souk Ahras	Examinateur
05	Nabiha Azizi	Prof.	Université Badji Mokhtar -Annaba	Invitée

A la mémoire de mon père

A ma mère

A mon mari et à mes enfants

A mes frères et sœurs

A mes amis

Je dédie ce travail

'تحريك الوجه على أساس التعرف على تعبيرات الوجه'

ملخـــــص

تلعب تعابير الوجه دورًا مهمًا في الرسوم المتحركة للوجه لأنها تتيح توصيل المشاعر والنوايا بوضوح ودقة. يعيدون الحياة إلى الشخصيات من خلال السماح لهم بإظهار مجموعة من المشاعر وردود الفعل الواقعية والمعقولة. قد يحتاج رسامو الرسوم المتحركة إلى إعطاء تعبيرات الوجه يدويًا إلى الشخصيات الافتراضية.

ومع ذلك، فإن هذه المهمة تستغرق وقتًا طويلاً وقد يفوت صانع الرسوم المتحركة التعبيرات الدقيقة.

في الرسوم المتحركة للوجه، يمكن استخدام العديد من الأساليب، وهي توليف الرسوم المتحركة ، ونقل الرسوم المتحركة للوجه ، والرسوم المتحركة لالتقاط الحركة ، والأفاتار أو إنشاء الوجه الاصطناعي والرسوم المتحركة.

في هذه الأطروحة، تمت دراسة مشكلة نقل تعبيرات الوجه ، وهو نهج أساسي في الرسوم المتحركة للوجه. يتم التعرف على التعبيرات أولاً على وجه حقيقي ثم يتم نقلها إلى وجه اصطناعي تم إنشاؤه بواسطة نظام التعلم العميق.

الهدف من هذه الأطروحة هو اقتراح نظام قائم بالكامل على التعلم العميق ، مما يسمح بإنشاء وجه اصطناعي يدرك تعبيرات الوجه المنبعثة من الوجه الحقيقي في الإدخال. في هذا المنظور، تمت دراسة محورين رئيسيين على نطاق واسع ، وهما التعرف على تعبيرات الوجه وتوليد الوجه الاصطناعي التعبيرية.

يعد التعرف على تعبيرات الوجه خطوة حاسمة في نظام إنشاء الوجوه التعبيرية المقترح. تم تطوير نموذج مزدوج الفروع، من أجل ضمان تنوع البيانات والتغلب على قيود حجم مجموعة البيانات في أنظمة FER. حقق هذا النموذج دقة 63.80% لقاعدة بيانات Fer-2013 و99.32% لقاعدة بيانات CK+

يتمثل نقل تعبيرات الوجه من وجه إلى آخر في توليد وجه اصطناعي يحقق تعبير وجه المصدر. هذه المهمة ذات أهمية كبيرة في مختلف المجالات، مثل الاتصال والترفيه الرقمي وتحريك الوجه. تم اقتراح واختبار نظام تكييف StyleGAN2. النتائج التي تم الحصول عليها مشجعة.

الكلمات المفتاحية: تعبيرات الوجه - تكوين الوجه - التعرف على تعبيرات الوجه - التعلم العميق - نقل التعلم - StyleGAN2.

« Face animation based on facial expression recognition »

Abstract

Facial expressions play a crucial role in facial animation as they allow emotions and intentions to be communicated clearly and accurately. They bring characters to life by allowing them to show a range of realistic and believable emotions and reactions. Animators may need to manually give facial expressions to virtual characters. However, this task is time consuming and the animator may miss subtle expressions.

In face animation, several approaches can be used, namely animation synthesis, face animation transfer, motion capture animation, avatar or synthetic face generation and animation.

In this thesis, the problem of facial expression transfer, which is an essential approach in face animation, is studied. The expressions are first recognized on a real face and then transferred to a synthetic face created by a deep learning system.

The objective of this thesis is to propose a system entirely based on deep learning, allowing to generate a synthetic face realizing the facial expression emitted by the real face in input. In this perspective, two main axes have been extensively studied, which are facial expression recognition and expressive synthetic face generation.

Facial expression recognition is a crucial step in the proposed expressive face generation system. A dual-branch model has been developed, in order to ensure data diversity and to overcome the limitations of dataset size in FER systems. It achieved an accuracy of 63.80% for the Fer-2013 database and 99.32% for the CK+ database.

The transfer of a facial expression from one face to another consists in generating a synthetic face realising the expression of the source face. This task is of great importance in various fields, such as communication, digital entertainment and facial animation. A system for conditioning the StyleGAN2 has been proposed and tested. The results obtained are encouraging.

Keywords: Facial expression - Face generation - Facial expression recognition - Deep learning - Transfer learning - StyleGAN2.

« Animation du Visage à base de reconnaissance des expressions faciales »

Résumé

Les expressions faciales jouent un rôle crucial dans l'animation faciale car elles permettent de communiquer les émotions et les intentions de manière claire et précise. Elles donnent vie aux personnages en leur permettant de montrer une gamme d'émotions et de réactions réalistes et crédibles. Les animateurs peuvent être amenés à donner manuellement des expressions faciales aux personnages virtuels. Cependant, cette tâche prend du temps et l'animateur peut manquer des expressions subtiles.

Dans l'animation de visage, plusieurs approches peuvent être utilisées, à savoir la synthèse d'animation, le transfert d'animation de visage, animation par capture de mouvement, la génération et l'animation d'avatar ou de visage synthétique.

Dans cette thèse, la problématique du transfert d'expressions faciales, qui est une approche essentielle dans l'animation de visages, est étudiée. Les expressions sont d'abord reconnues sur un visage réel, puis transférées sur un visage synthétique créé par un système d'apprentissage profond.

L'objectif de cette thèse est de proposer un système entièrement basé sur l'apprentissage profond, permettant de générer un visage synthétique réalisant l'expression faciale émise par le visage réel en entrée. Dans cette perspective, deux axes principaux ont été profondément étudiés, à savoir la reconnaissance d'expressions faciales et la génération de visage synthétique expressif.

La reconnaissance des expressions faciales est une étape cruciale dans le système de génération de visages expressifs proposé. Un modèle à double branche a été développé, afin d'assurer la diversité des données et de surmonter les limitations de la taille des ensembles de données des systèmes FER. Il a atteint une précision de 63,80 % pour la base de données Fer-2013 et de 99,32 % pour la base de données CK+.

Le transfert d'une expression faciale d'un visage à un autre consiste à générer un visage synthétique réalisant l'expression du visage source. Cette tâche est d'une grande importance dans divers domaines, tels que la communication, les loisirs numériques et l'animation faciale. Un système permettant de conditionner le StyleGAN2 a été proposé et testé. Les résultats obtenus sont encourageants.

Mots-clés : Expression faciale – Génération de visage – Reconnaissance des expressions Faciales – Apprentissage profond – Transfert d’apprentissage – StyleGAN2

Table des matières

Table des figures

Liste des tableaux

Chapitre 1 : Introduction générale

1.1 Contexte du travail

Le visage est la partie la plus importante du corps humain dans l'interaction entre les personnes et l'interaction Homme-Machine. Chaque jour, le visage est utilisé pour recevoir et envoyer des informations, qu'elles soient volontaires ou involontaires. Ces dernières, qu'on appelle des expressions faciales, rendent les humains des experts dans la détection des défauts et des comportements contre-nature dans les créations à l'apparence humaine comme les marionnettes et les modèles en 3D, d'où la nécessité de modèle tridimensionnels et d'animation faciale réaliste.

L'animation faciale est l'un des thèmes de recherche les plus importants en infographie. Depuis les années 1970, et les premiers travaux de F. Parke (Parke, 1972), le pionnier de la discipline, de nombreuses recherches ont été menées, mais reste toujours une tâche difficile. C'est grâce à l'apport des principes de vision par ordinateur qui utilisent des techniques basées sur la performance pour créer des expressions faciales à partir d'une performance faciale humaine, que les plus importants progrès sont enregistrés.

L'animation de visages vise à synthétiser des images successives de visages à partir d'une seule image source, en fonction d'un ensemble de mouvements conditionnels du visage, et à produire des animations expressives et plausibles de visages. Les recherches actuelles se concentrent sur trois domaines fondamentaux pour accroître la réalité : les expressions faciales subtiles, le trucage d'un modèle de visage et le transfert d'expression d'un humain(Ekmen & Ekenel, 2019).

La recherche dans ce domaine a toujours eu pour objectif d'obtenir une animation réaliste en sollicitant le moins possible l'intervention humaine. Avant la percée de l'apprentissage profond, et la révolution qui en découle, les meilleurs résultats étaient obtenus en s'appuyant sur l'utilisation de déformations géométriques, qui prennent généralement en compte la forme et les déformations propres à la physiologie et aux expressions d'une personne, et de manipulations d'images modélisant les propriétés de réflexion de la peau et des cheveux du visage pour obtenir des détails à petite échelle difficiles à modéliser par la manipulation géométrique seule (Noh & Neumann, 1998a).

1.2 Problématique et contribution

Dans le cadre de cette thèse, nous nous intéressons à l'animation du visage par transfert des expressions faciales d'un visage humain réel, à partir d'une seule image, à un autre visage synthétique par l'utilisation de méthodes basées sur l'apprentissage profond. Nous avons concentré nos recherches principalement sur deux axes importants :

- Premièrement sur la reconnaissance des expressions faciales.
- Deuxièmement sur le transfert des expressions faciales d'un visage réel à un visage synthétique.

Les méthodes traditionnelles sont basées sur l'utilisation d'un mélange de points de repère ou de formes pour établir une correspondance entre le visage réel d'une personne et le visage virtuel. Cependant, ces méthodes nécessitent un processus de modélisation lourd, et leurs performances dépendent de l'expérience des experts en modélisation(J. Zhang et al., 2020).

Bien que plusieurs travaux aient été proposés dans ce domaine, peu d'entre eux ont pu aboutir à un résultat satisfaisant. Une des principales causes est la rareté de bases de données de visages accessibles au public, et qui sont limitées en termes de taille et de variabilité des échantillons(Moschoglou et al., 2020).

Plusieurs propositions récentes se basent sur l'utilisation des nouvelles et puissantes méthodes de l'apprentissage profond (Deep Learning DL). L'approche DL, parfois appelée apprentissage universel, peut être adoptée dans presque tous les domaines d'application. Dans ces approches, les caractéristiques optimales sont automatiquement apprises pour la tâche à accomplir. Ainsi, elles offrent une certaine robustesse aux variations naturelles des données d'entrée. L'approche d'apprentissage par transfert, qui peut être utilisée dans différentes applications ou avec différents types de données, est l'une des techniques populaires d'apprentissage direct et est souvent utile lorsque le problème ne dispose pas de suffisamment de données.

Ainsi, plusieurs problèmes sont à résoudre : comment modéliser un visage, laquelle des approches de reconnaissance des expressions faciales faut-il choisir ? Comment transférer ces expressions à un autre visage ? Quel modèle d'apprentissage profond serait le plus approprié au processus de transfert ?

Pour fixer nos objectifs, nous avons essayé de répondre aux problématiques soulevés ci-dessus.

Ainsi, l'objectif de ce travail est de proposer un système basé entièrement sur l'apprentissage profond, permettant de générer un visage synthétique réalisant l'expression faciale exprimée par le visage réel en entrée.

Dans cette thèse, nous avons exploré deux approches, une approche conventionnelle basée sur les points caractéristiques du visage, et une approche basée sur l'utilisations des techniques de l'apprentissage profond et la reconnaissance des expressions faciales.

- ***L'approche de transfert basée sur les points caractéristiques,*** qui est une approche qui utilise les fonctions à bases radiales permettant de transférer des expressions faciales d'un visage source (visage réel 2D) à un visage cible (objet virtuel 3D).
- ***L'approche de transfert basée sur la reconnaissance des expressions faciales,*** qui consiste tout d'abord à reconnaître et à extraire les expressions faciales d'un visage réel, puis à les transférer en générant d'un nouveau visage synthétique réalisant l'expression du visage réel, cette approche se base sur les techniques de l'apprentissage profond.

En ce qui concerne nos contributions, nous pouvons les résumer en trois axes :

1) Après avoir dressé un état de l'art des approches d'animation faciale existantes, nous les avons classées en deux catégories ; les Approches conventionnelles et les approches basées sur l'apprentissage profond.
2) L'étude de différentes méthodes de reconnaissance des expressions faciales et La proposition de l'utilisation d'un réseau de neurones à convolutions classiques amélioré pour la reconnaissance des expressions faciales, puis, l'exploration de l'approche de transfert d'apprentissage, et enfin la proposition d'un modèle à double branches, combinant les deux approches précédemment proposées et testées. Les trois modèles sont évalués sur deux bases de données de référence.
3) La proposition d'une approche de génération de visage synthétique expressif pour l'animation faciale par un réseaux antagoniste génératif. Notre proposition permet de générer un visage synthétique avec l'expression faciale réalisée par une personne réelle, et ce à partir d'une seule image.

Les atouts forts de cette approche sont :

- Des visages générés plus réalistes.
- Aucun point de repère (ou point de caractéristique) n'est nécessaire.
- Aucune nécessité d'intervention humaine.

1.3 Organisation du document

Ce document de thèse se compose de huit chapitres, structurés en deux parties.

Le présent chapitre constitue une introduction générale qui spécifie le contexte de l'étude, la problématique de la recherche, les contributions apportées par la thèse ainsi que la structure du manuscrit.

La première partie constitue un état de l'art du domaine, elle présente tous d'abord les notions de base de l'apprentissage profond, les techniques conventionnelles de l'animation faciale et enfin les travaux connexes ayant traité l'animation faciale par apprentissage profond, cette partie est structuré en trois chapitres, comme suit :

Le chapitre 2 est consacré à l'apprentissage profond et à son évolution. Ainsi, les différents concepts, principes et algorithmes de l'apprentissage profond, depuis les réseaux de neurones jusqu'aux modèles les plus récents proposés dans la littérature, sont présentés.

Le chapitre 3 présente un état de l'art des techniques conventionnelles utilisées dans l'animation faciale et celle basées sur l'apprentissage profond. Tout d'abord, les concepts de base du domaine sont présentés, puis une revue de l'état de l'art des techniques conventionnelles d'animation faciale est faite, pour chacune des techniques, une brève présentation est proposée, suivie d'une discussion sur leurs points forts et leurs points faibles. Ensuite, une synthèse détaillée de l'état de l'art des approches appliquant l'apprentissage profond à l'animation faciale est dressée.

La deuxième partie de cette thèse est consacrée à la présentation et l'évaluation des contributions de notre travail. Elle comporte trois chapitres.

Le chapitre 4 traite de l'animation faciale basée sur les points caractéristiques et présente notre contribution, qui peut être résumée en deux parties : premièrement, la synthèse de l'animation à l'aide de la déformation et de la paramétrisation MPEG-4, et deuxièmement, le transfert de l'animation faciale à l'aide des points caractéristiques du visage.

Le chapitre 5 détaille la deuxième contribution de cette thèse, qui est consacrée à la reconnaissance des expressions faciales, où nous proposons un modèle à double branche pour la reconnaissance des expressions faciales, un état de l'art de ce problème est établi, et une description des bases de données de visages les plus utilisées est fournie. Enfin, une évaluation complète est présentée.

Le chapitre 6 est consacré à l'approche proposée pour la génération de visages expressifs basés sur l'expression faciale des entrées, une description détaillée du modèle StyleGan2 est présentée, ainsi que le système proposé, son implémentation et ses expérimentations et tests.

Pour conclure, le chapitre 7 présente une conclusion générale qui résume les contributions de la thèse et expose les perspectives de recherche futures.

Première partie

Etat de l’art

Chapitre 2 : L'apprentissage profond

2.1 Introduction

L'intelligence artificielle est difficile à définir ; elle pourrait être définie comme : le pouvoir de percevoir et de prédire l'avenir immédiat ou lointain, ou la capacité de planifier un ensemble d'actions pour aboutir à un objectif donné, ou la capacité d'apprendre et d'appliquer efficacement ses connaissances.

Selon Yann LeCun(Lecun, 2016) : « l'intelligence artificielle (IA) est un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains et à certains animaux ».

L'IA englobe plusieurs branches qui utilisent diverses techniques pour mimer des comportements précis associés à l'intelligence humaine naturelle. Depuis son apparition, un sous-ensemble de l'intelligence artificielle (IA), appelé apprentissage automatique, a révolutionné de nombreux domaines. Dans cette branche, la machine est en mesure d'apprendre à partir de grands ensembles de données sans être explicitement programmée(Samuel, 1959). Les réseaux neuronaux (NN, **N**eural **N**etwork) sont à leur tour un sous-domaine de l'apprentissage automatique, et c'est ce sous-domaine qui a donné naissance à l'apprentissage profond (DL, **D**eep **L**earning).

Depuis sa création, l'apprentissage profond a provoqué un bouleversement croissant, en montrant un succès remarquable dans différents domaines en raison de leur capacité à résoudre des problèmes complexes, par rapport aux méthodes classiques d'apprentissage automatique. En dépit de ses nombreux avantages, le deep learning est confronté à plusieurs défis relatifs aux problèmes de sur-apprentissage, de dégradation du gradient et de complexité temporelle élevée. Plusieurs modèles performants ont été proposés pour l'apprentissage supervisé et non supervisé afin de réduire ces problèmes.

L'objectif de ce chapitre est de détailler les différents concepts, principes et algorithmes de l'apprentissage profond, des réseaux de neurones aux plus récents modèles proposés dans la littérature.

2.2 De l'intelligence artificielle à l'apprentissage profond

C'est au cours de l'été 1955 que le terme "intelligence artificielle" a été utilisé pour la première fois. Ce sont John McCarthy (Dartmouth College), Marvin L. Minsky (MIT),

Nathaniel Rochester (IBM) et Claude Shannon (Bell Laboratories) qui ont dirigé le projet de recherche estival de Dartmouth sur l'intelligence artificielle(McCarthy et al., 2006).

Concrètement, les applications de l'IA incluent les systèmes experts, la reconnaissance vocale et la vision par ordinateur, le contrôle, la prédiction et divers autres domaines. La discipline de l'IA ne s'est pas toujours limitée au rôle essentiel de l'apprentissage dans l'intelligence. Le premier système intelligent pour jouer aux échecs était un programme écrit à la main basé sur la méthode de recherche par arbre, et la reconnaissance de caractères imprimés se faisait par comparaison avec des images prototypes. Pour établir un système d'aide au diagnostic médical à partir de symptômes, les chercheurs se sont appuyés sur la déduction logique de règles écrites par des experts (LeCun, 2016).

Bien que ces progrès aient semblé importants au début, ces méthodes s'avérèrent très difficiles à appliquer à diverses tâches telles que l'identification d'objets dans des scènes, la traduction de textes ou même la reconnaissance faciale, car concevoir un système performant dans toutes les situations est pratiquement impossible. C'est à ce stade que l'apprentissage automatique devient indispensable. Cette technologie est au cœur de nombreux secteurs de la vie contemporaine : des recherches sur le web au filtrage du contenu sur les réseaux sociaux en passant par les recommandations sur les sites de commerce électronique, et elle est de plus en plus présente dans les produits de consommation tels que les appareils photo et les smartphones. Un bref historique des réseaux neuronaux présentant les événements clés est présenté à la Figure 2-1.

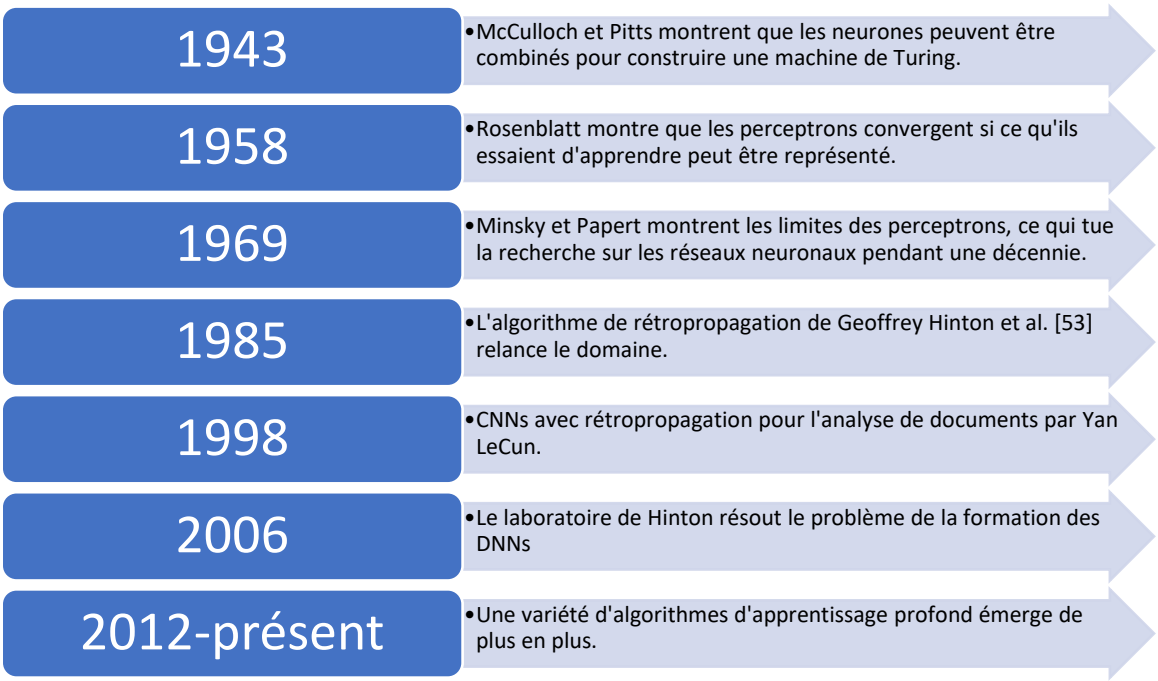


Figure 2-1: L'histoire du développement de l'apprentissage profond.

L'apprentissage automatique s'appuie sur la richesse des données pour construire ou apprendre automatiquement des algorithmes(Bishop, 2016). La possibilité de traiter

des données naturelles sous leur forme brute était limitée par les techniques conventionnelles d'apprentissage automatique. La mise au point d'un système de reconnaissance des formes ou d'apprentissage automatique a requis pendant des décennies une ingénierie minutieuse et une expertise considérable du domaine pour concevoir un outil d'extraction des caractéristiques qui transformait les données brutes (telles que les valeurs des pixels d'une image) en une représentation interne appropriée ou un vecteur de caractéristiques permettant au sous-système d'apprentissage, souvent un classificateur, de détecter ou de classer des formes dans les données d'entrée(LeCun et al., 2015).

En d'autres termes, l'apprentissage automatique repose sur l'extraction de caractéristiques pertinentes et exploitables (données, connaissances) à partir de grands volumes de données à l'aide d'algorithmes d'apprentissage et de bases de données. Toutefois, la mise en œuvre de ce processus est souvent assez coûteuse et est liée au contexte. De plus, une mauvaise extraction de caractéristiques entraîne des performances d'apprentissage très insuffisantes(Hardy, 2019). En revanche, l'aspect essentiel des méthodes d'apprentissage profond consiste à apprendre automatiquement les représentations et la structure interne des données d'entrée brutes (les caractéristiques), en utilisant plusieurs couches de traitement superposées(LeCun et al., 2015).

L'apprentissage profond est issu de la branche de l'apprentissage automatique, qui utilise des modèles informatiques composés de plusieurs couches de traitement pour apprendre des représentations de données avec plusieurs niveaux d'abstraction. Ces couches de traitement sont souvent des transformations linéaires, des activations de non-linéarité ou d'autres modules tels que la mise en commun. L'ensemble du modèle est conventionnellement appelé un réseau de neurones (neural network ou neural net NN)(Z. Liu, 2021).

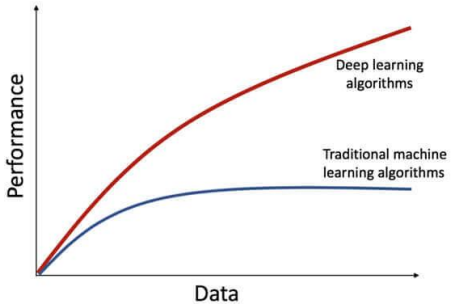


Figure 2-2 : Les performances de l'apprentissage profond en fonction du volume de données(Alom et al., 2019).

À la lecture de la Figure 2-2, il apparaît clairement que les performances des approches traditionnelles d'apprentissage automatique sont meilleures pour les petites quantités de données d'entrée (Alom et al., 2019). Au-delà d'une certaine quantité de données, les performances des approches traditionnelles d'apprentissage automatique

deviennent stables. D'autre part, les performances des approches d'apprentissage profond augmentent avec la quantité de données.

2.3 Les types d'apprentissage

Ces dernières années, l'apprentissage profond s'est développé rapidement et a été appliqué à la plupart des domaines d'application traditionnels, ainsi qu'à certains nouveaux domaines qui présentent plus d'opportunités, il a démontré des performances de pointe par rapport aux approches traditionnelles.

Différentes méthodes ont été proposées sur la base de différentes catégories d'apprentissage, principalement distinguées par leur objectif, notamment l'apprentissage supervisé, semi-supervisé et non supervisé, ainsi que l'apprentissage par renforcement (RL), qui est souvent abordé dans le contexte des approches d'apprentissage semi-supervisé ou parfois non supervisé. Ces différents types d'apprentissage peuvent être appliqués dans des contextes différents, mais ils peuvent aussi être combinés dans un même système.

2.3.1 Apprentissage supervisé

L'apprentissage supervisé est de loin la forme la plus fréquente de l'apprentissage automatique, qu'il soit profond ou non. Ce type d'apprentissage utilise un ensemble de données dans lequel une étiquette, une classe ou une catégorie (target, label) est associée à chaque échantillon(I. Goodfellow et al., 2016).

La base de données comporte un ensemble d'entrées et de sorties correspondantes (X_t, Y_t). Ainsi, pour l'entrée X_t, le modèle prédit $\widehat{Y_t} = f(X_t)$, puis une valeur de perte est calculée $l(Y_t, \widehat{Y_t})$. Le modèle est ensuite modifié de manière itérative (les paramètres du réseau) pour obtenir une meilleure approximation des sorties souhaitées. Après un apprentissage réussi, le modèle sera capable d'obtenir des prédictions correctes.

Diverses approches d'apprentissage supervisé existent pour l'apprentissage profond(Alom et al., 2019), parmi lesquelles les réseaux de neurones profonds (DNN), les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN), y compris les mémoires à court et à long terme (LSTM), et les unités récurrentes à déclenchement (GRU).

2.3.2 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est un apprentissage basé sur des ensembles de données partiellement étiquetés, autrement dit, seul un petit sous-ensemble d'échantillons possède des étiquettes de classe correspondantes(Kingma et al., 2014).

Parmi les approches existantes, l'algorithme le plus simple pour l'apprentissage semi-supervisé est basé sur un schéma d'auto-apprentissage(Rosenberg et al., 2005) où

le modèle est amorcé avec des données étiquetées supplémentaires obtenues à partir de ses propres prédictions très fiables ; ce processus est répété jusqu'à ce qu'une condition de fin soit atteinte.

Dans certains cas, l'apprentissage profond par renforcement et les réseaux adversaires génératifs (GAN) sont utilisés comme techniques d'apprentissage semi-supervisé(Alom et al., 2019). En outre, les RNN, y compris les LSTM, sont également utilisés pour l'apprentissage semi-supervisé.

2.3.3 Apprentissage non supervisé

Dans l'apprentissage non supervisé, les données d'entrainement ne sont pas étiquetées. L'algorithme tente donc de trouver et d'apprendre la représentation interne ou les caractéristiques importantes pour découvrir des relations ou des structures inconnues dans les données d'entrée, afin de les différencier et de révéler l'existence de classes ou de groupes.

Les techniques de regroupement, de réduction de la dimensionnalité, les techniques génératives ainsi que les auto-encodeurs sont souvent considérées comme des approches d'apprentissage non supervisé(Larochelle, 2008).

2.3.4 Apprentissage par renforcement

L'objectif de l'apprentissage par renforcement est d'entraîner un modèle à se comporter de manière intelligente dans un environnement donné. Le modèle apprend à agir par essais et erreurs, il va donc interagir avec l'environnement en choisissant, à chaque instant, d'exécuter une action parmi un ensemble d'actions autorisées(Larochelle, 2008). Chaque action sera récompensée par un signal de renforcement, qui sera utilisé pour améliorer son comportement. L'objectif du modèle est de maximiser la récompense totale qu'il reçoit au fil du temps.

Le recuit et les méthodes d'entropie croisée en sont des exemples d'algorithme d'apprentissage par renforcement, mais le plus utilisé est le Q-Learning.

2.4 Les réseaux de neurones

Les réseaux neuronaux artificiels s'inspirent du fonctionnement des réseaux neuronaux biologiques du cerveau. Ce dernier contient un grand nombre de neurones biologiques massivement connectés, qui échangent des messages par le biais de signaux transmis par des synapses. Comparativement aux systèmes artificiels, ces systèmes biologiques se caractérisent par leur grande complexité en raison du grand nombre de neurones massivement interconnectés qui ne peuvent être traités par une machine.

2.4.1 Perceptron

Le perceptron ou le neurone, élément de calcul de base, est appelé un nœud (ou unité) qui reçoit des entrées de sources externes et possède certains paramètres internes (y compris les poids et les biais appris pendant la formation) qui produisent des sorties. Cette unité est appelée un perceptron. En d'autres termes, sa fonction fondamentale est de recevoir des *entrées* multiples x_1, x_2, ... et de calculer une somme pondérée z pour ces entrées en utilisant les *poids* w_1, w_2, ... La somme pondérée z est *une transformation linéaire* des entrées du neurone. *Le biais b* est ajouté à cette somme, et le résultat passe par *une fonction d'activation* non linéaire σ, ce qui donne la sortie finale a. Par conséquent, le neurone effectue une transformation non linéaire sur ses entrées en calculant d'abord une transformation linéaire, et en appliquant en plus un biais et une fonction d'activation non linéaire, avec comme résultat :

$$a = \sigma\left(\sum_i x_i . w_i + b\right) = \sigma(z) \qquad (2.1)$$

Il est possible de reformuler cette formule en une notation vectorielle plus pratique, grâce au produit scalaire des poids vectoriels w et des entrées vectorielles x, comme suit :

$$a = \sigma(x^T . w + b) = \sigma(z) \qquad (2.2)$$

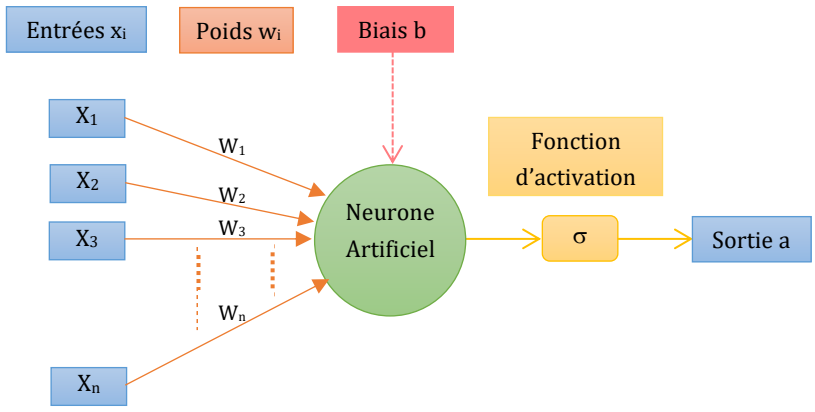


Figure 2-3 : Schéma d'un neurone artificiel.

La structure d'un neurone artificiel est illustrée à la Figure 2-3. Les neurones artificiels les plus anciens, les Threshold Logic Units (McCulloch & Pitts, 1943), nécessitaient un réglage manuel des paramètres d'apprentissage, qui consistaient uniquement en des poids et un seuil, et ne calculaient donc qu'une transformation linéaire. Désormais, des méthodes avancées d'apprentissage automatique basées sur le gradient sont utilisées, elles sont décrites plus en détail dans la section 2.5.

Le neurone artificiel, bien qu'il s'agisse d'un modèle très simple, peut être utilisé pour une grande variété de problèmes. Ainsi, le choix de la fonction échelon comme fonction d'activation permet d'utiliser le neurone pour la classification binaire. Ces types de neurones utilisés pour la classification binaire sont appelés *Perceptrons*. En 1958,

Rosenblatt (Rosenblatt, 1958) a proposé le premier algorithme d'apprentissage automatique pour les perceptrons, qui, étant donné des données d'apprentissage linéairement séparables, converge vers une solution. Il était basé sur l'ajustement des poids des perceptrons en fonction de la classe prédite et de la classe réelle de l'ensemble de données d'apprentissage, et représente donc l'un des premiers modèles d'apprentissage supervisé. Cependant, les recherches n'ont repris qu'en 1986, lorsque Rumelhart et al. (Rumelhart et al., 1986)ont vulgarisé l'algorithme de rétropropagation pour l'entraînement des réseaux neuronaux, dans le but de combiner des neurones artificiels pour former des réseaux neuronaux. En combinant plusieurs perceptrons dans un réseau, il est même possible de résoudre des problèmes qui ne sont pas linéairement séparables, comme le problème XOR (I. Goodfellow et al., 2016).

2.4.2 Perceptron multicouche

Les perceptrons multicouches, appelé en anglais Multi-layer Perceptron (MLP), et connu aussi comme les réseaux neuronaux à propagation en avant (Feedforward neural networks), sont des réseaux de neurones composés de multiples neurones interconnectés et structurés en couches successives, ces neurones sont combinés pour former un graphe dirigé sans cycles (acyclique), c'est-à-dire les connexions entre les nœuds ne forment pas de cycle. Ils sont alors plus généraux que le perceptron. Les neurones artificiels de même profondeur dans ce graphe constituent des couches, où chaque couche peut être résumée à une fonction unique constituée de ces multiples neurones.

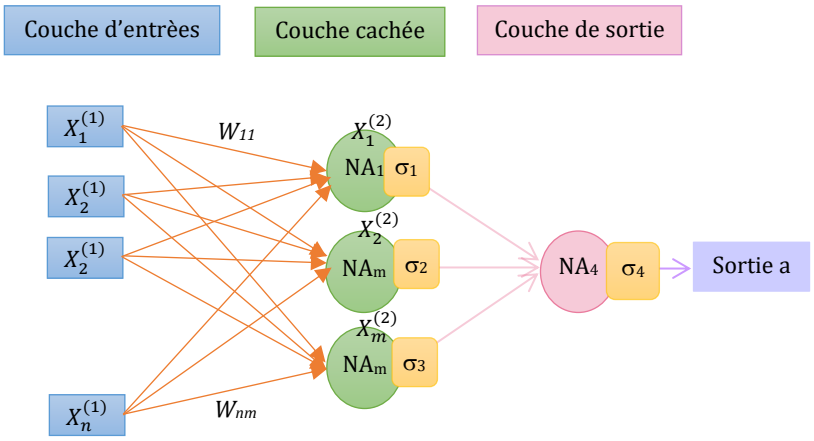


Figure 2-4 : Structure d'un FNN simple.

Pour des raisons de clarté visuelle, les biais sont omis, bien que chaque neurone artificiel AN_i possède également un terme de biais b_i.

La Figure 2-4 illustre un exemple de perceptron multi couches, qui est composé de trois couches : une couche d'entrée, un nombre arbitraire de couches intermédiaires de neurones appelées couches cachées et d'une couche de sortie. Le terme "apprentissage profond" vient également de cette approche basée sur les couches, car la profondeur d'un FNN décrit le nombre de couches dont il est constitué(I. Goodfellow et al., 2016).

Dans ce modèle, chaque nœud est doté d'une valeur x_i et est connecté aux couches adjacentes par des pondérations de poids w_{ij}. Ainsi, l'ensemble des valeurs x_i de la couche courante constitue une entrée pour la couche suivante. La couche d'entrée est composée de plusieurs nœuds, qui dépendent du nombre d'attributs de la base d'apprentissage(Dif, 2020). Pour calculer les valeurs des couches cachées, les x_i sont calculés selon l'équation suivante :

$$x_i^{(k)} = F^{(k)}(\sum_{j=1}^{m} w_{ij} x_j^{(k-1)} + b_j) \qquad (2.3)$$

Où $x_i^{(k)}$ représente la valeur du nœuds i de la couche k, et F représente la fonction d'activation, w_{ij} sont les différents poids associés à x_i, et m représente le nombre de nœuds dans la couche suivante.

2.4.3 Les fonctions d'activation

Les réseaux neuronaux utilisent généralement des fonctions d'activation non linéaires pour convertir un signal d'entrée provenant d'un nœud en un signal de sortie, ces fonctions étant appliquées sur la somme pondérée des entrées. C'est ce qui leur permet d'introduire des propriétés non linéaires afin de pouvoir traiter des données complexes et résoudre des problèmes non linéaires.

En général, les fonctions sigmoïde, tangente hyperbolique ou unité linéaire rectifiée sont utilisées pour les couches cachées, tandis que la couche de sortie est basée sur la fonction Softmax ou Sigmoïde, selon le type de classification. Nous décrivons dans ce qui suit les fonctions d'activation couramment utilisées ainsi que leurs propriétés lors de l'optimisation :

2.4.3.1 La fonction sigmoïde

La fonction d'activation sigmoïde a la forme mathématique suivante :

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad (2.4)$$

Elle est illustrée dans la Figure 2-5 :

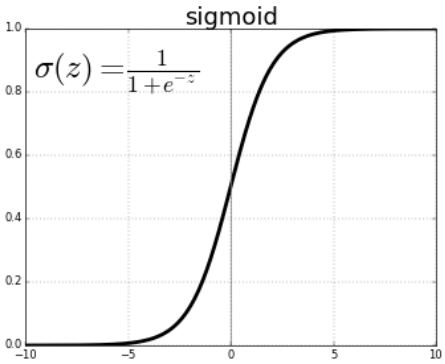


Figure 2-5 : Fonction d'activation sigmoïde dans l'intervalle x = [-10, 10].

Cette fonction sigmoïde est habituellement appliquée dans le cadre des applications de classification binaire(Dif, 2020), car elle permet de convertir l'entrée à valeur réelle z en une probabilité comprise entre 0 et 1. Elle est plus utilisée en apprentissage automatique (notamment pour la régression logistique) qu'en apprentissage profond, ceci est dû à sa propriété de tendre rapidement vers 0 ou 1, induisant ainsi la saturation de certains neurones du réseau et entraînant l'arrêt de l'apprentissage, c'est d'ailleurs son principal inconvénient(I. Goodfellow et al., 2016).

2.4.3.2 La fonction tangente hyperbolique

La fonction d'activation tangente hyperbolique (tanh) est étroitement liée à la fonction d'activation sigmoïde et a la forme mathématique suivante :

$$\tanh(x) = \frac{sinh(x)}{cos(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = 2\sigma(2x) - 1 \qquad (2.5)$$

La Figure 2-6 illustre le graphe de cette fonction d'activation :

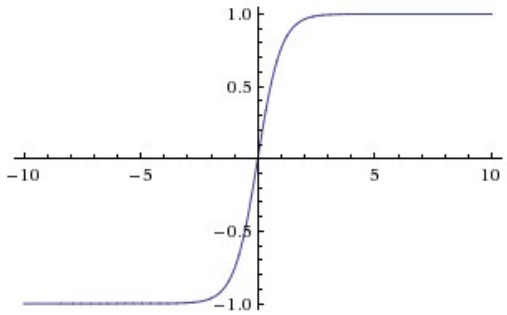


Figure 2-6 : Fonction d'activation tangente hyperbolique dans l'intervalle x = [-10, 10].

Elle est généralement utilisée dans les couches cachées pour l'introduction de la non linéarité. Elle se distingue par sa bonne précision en reconnaissance par rapport à la fonction logistique sigmoïde(Olgac & Karlik, 2011). Comme le montre l'équation 2.5, *tanh* est simplement une version mise à l'échelle de l'activation sigmoïde. Cependant, comme elle est centrée sur 0 (son résultat est compris entre -1 et 1), elle ne souffre pas

de certains des problèmes de la fonction sigmoïde. Ainsi, la fonction *tanh* est presque toujours préférée à la fonction d'activation sigmoïde.

2.4.3.3 La fonction unité de rectification linéaire (ReLU)

C'est aussi une fonction d'activation couramment utilisée pour les couches cachées, qui la forme mathématique suivante :

$$ReLU(w) = \max(0, x) \tag{2.6}$$

Et dont le graphe est illustré dans la Figure 2-7.

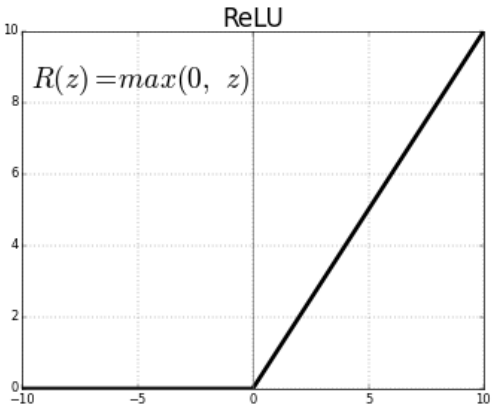


Figure 2-7 : Fonction d'activation ReLU dans l'intervalle z = [-10, 10]

Contrairement aux fonctions précédentes, La fonction ReLU permet au réseau de neurone de converger rapidement (par l'accélération du temps d'apprentissage(Krizhevsky et al., 2012)), et ne provoque pas de saturation, ce qui la rendue la fonction la plus utilisée dans les dernières années en apprentissage profond. Elle est également très efficace à calculer, puisqu'il suffit de seuiller les activations des neurones à zéro. Comme le montre son équation 2.6, elle annule les valeurs négatives pour les ramener à 0.

L'un des principaux inconvénients des ReLU est que les unités peuvent "mourir" pendant la formation. Si, pendant la formation, les poids sont décalés de telle sorte que pour toutes les entrées de données pendant la descente du gradient, l'activation du neurone se trouve sur le plan plat dans la zone z négative, le neurone ne fera que rétropropager un gradient de 0, ce qui rend très improbable la récupération de ce neurone à partir de cet état.

Plusieurs versions améliorées de ReLU ont été proposées, ils offrent une précision encore meilleure par rapport à la fonction ReLU tel que PReLU proposée par Kaiming He et al., et la fonction Leaky ReLU qui est utilisé pour résoudre le problème de 'la mort' des neurone(He et al., 2015).

2.4.3.4 La fonction Sotmax

Elle est appelée aussi fonction exponentielle normalisée, et est généralement utilisée par les couches de sortie. Elle a la forme mathématique suivante :

$$Softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, j \in \{1, \dots, K\} \quad (2.7)$$

C'est une généralisation multidimensionnelle de la fonction sigmoïde, sa sortie peut être considérée comme une distribution de probabilité sur un ensemble fini de résultats, elle est la mieux adaptée aux problèmes de classification multi-classes.

Il convient de noter qu'elle n'est pas appliquée élément par élément mais sur un vecteur entier de "scores". Elle est principalement utilisée comme sortie non linéaire pour prédire les probabilités discrètes sur les catégories de sortie.

2.5 Apprentissage des réseaux de neurones

Depuis les débuts de la reconnaissance des formes(Rosenblatt, 1958), le principal objectif des chercheurs a été de remplacer les caractéristiques conçues à la main par des réseaux multicouches entraînables, mais en dépit de sa simplicité, cette solution n'a pas été largement comprise avant la mi-1980. En fait, les architectures multicouches peuvent être entraînées par une simple descente de gradient stochastique. Tant que les modules sont des fonctions relativement lisses de leurs entrées et de leurs poids internes, les gradients peuvent être calculés en utilisant la procédure de rétropropagation. L'idée que cela pouvait être fait, et que cela fonctionnait, a été découverte indépendamment par plusieurs groupes différents au cours des années 1970 et 1980(Rumelhart et al., 1986).

Cette étape d'entrainement est nécessaire et doit être réaliser avant son utilisation pour réaliser une tâche spécifique, c'est ce qu'on appelle *l'apprentissage.* C'est un problème d'optimisation qui consiste à estimer et à ajuster les paramètres du modèle (les poids et les biais) pour qu'il puisse approximer la sortie désirée. Pour y parvenir, il est nécessaire de définir une métrique permettant de mesurer la qualité du résultat approximé par le réseau. Il s'agit de la *fonction de perte* ou *fonction de coût* $J(\theta)$, où θ décrit les paramètres combinés du réseau (poids, biais).

Étant donné un ensemble de *N exemples d'apprentissage* $x_T = [x_{T_1}, x_{T_2}, \dots, x_{T_N}]$et *les cibles* correspondantes $y = [y_1, y_2, \dots, y_N]$, $J(\theta)$ est généralement calculé comme la moyenne de la fonction de perte par exemple $L(a(x_{T_i};\ \theta), y_i)$, où $a(x_{T_i};\ \theta)$ est la sortie du réseau, étant donné l'exemple d'apprentissage x_{T_i} comme entrée et les paramètres du réseau θ :

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} L\big(a\big(x_{T_i};\ \theta\big), y_i\big) \qquad (2.8)$$

L'une des fonctions de perte les plus utilisées est la fonction de perte MSE (erreur quadratique moyenne). Une plus petite valeur de la fonction de perte équivaut généralement à une meilleure performance du réseau. Par conséquent, le processus d'apprentissage du réseau de neurones peut être formulée comme un problème d'optimisation, où l'objectif est de minimiser la fonction de perte $J(\theta)$ par rapport aux paramètres du réseau θ. Cet objectif est généralement atteint en utilisant une variante de l'algorithme de descente de gradient.

2.5.1 Descente de gradient

La descente de gradient a été utilisée avec succès pour l'entraînement des réseaux de neurones artificiels au cours des deux dernières décennies(Alom et al., 2019). C'est un algorithme d'optimisation du premier ordre utilisé pour trouver les minima locaux de la fonction de perte $J(\theta)$ par rapport aux paramètres du réseau θ.

Si pour des fonctions de perte simples, il est possible de calculer analytiquement leur minimum, pour des fonctions de perte plus complexes à paramètres multiples, comme celles des réseaux de neurones avec des millions de paramètres, ceci s'avère infaisable(I. Goodfellow et al., 2016). La descente de gradient, contrairement au calcul analytique des minima, est une approche numérique qui cherche à minimiser la fonction de coût $J(\theta)$ en modifiant θ, elle consiste à choisir des paramètres initiaux aléatoires et à procéder à des améliorations successives de manière itérative dans la direction de la descente la plus raide, en trouvant ou en s'approchant d'un point stationnaire ou d'un minimum global (recherche d'un point qui annule le gradient de la fonction de coût pour atteindre une configuration stable du réseau). Pour un seul échantillon d'apprentissage x_{T_i}avec la sortie cible correspondante yi, la direction de la descente la plus abrupte est donnée par le calcul du gradient négatif de la fonction de perte par échantillon par rapport aux paramètres θ à la position x_{T_i} et y_i :

$$-g_{\theta_i} = -\nabla_\theta L\big(a\big(x_{T_i};\ \theta\big), y_i\big) \qquad (2.9)$$

Le gradient final de la fonction de perte $J(\theta)$ est alors donné en calculant la moyenne de tous les gradients sur l'ensemble de l'apprentissage x_T:

$$-g_\theta = \frac{1}{N}\sum_{i=1}^{N} -g_{\theta_i} \qquad (2.10)$$

Le *taux d'apprentissage* η, qui est un facteur positif contrôlant la magnitude de la descente de gradient, est utilisé pour définir la règle de mise à jour des paramètres θ du réseau par descente de gradient, comme suit :

$$\theta = \theta + \eta . -g_\theta \qquad (2.11)$$

Dans l'apprentissage par descente de gradient, la méthode de rétropropagation est utilisée pour calculer le gradient de la fonction de coût(Dif, 2020). Cette méthode s'est avérée plus rapide à apprendre que les méthodes précédentes(Rumelhart et al., 1986). La descente du gradient peut, en fonction de l'initialisation des paramètres, trouver un minimum global pour $J(\theta)$, cependant, à moins que $J(\theta)$ soit convexe, cela n'est pas garanti. Il convient de noter que pour assurer la convergence de la descente de gradient, la fonction de perte doit être lisse et fournir des gradients partout(I. Goodfellow et al., 2016). La procédure complète de descente du gradient est décrite dans l'Algorithme 1.

Algorithme 1 : Algorithme de descente de gradient

Initialiser aléatoirement les paramètres $\boldsymbol{\theta} = \boldsymbol{\theta_0}$;

Pour *i = 1* à *nombre d'itérations* faire

Pour l'échantillon d'apprentissage d'entrée $\boldsymbol{x_{T_i}}$ faire

Calculer la sortie du réseau $\boldsymbol{a}(\boldsymbol{x_{T_i}}; \boldsymbol{\theta})$ en fonction de l'entrée $\boldsymbol{x_{T_i}}$ et des paramètres actuels $\boldsymbol{\theta}$;

Calculer le gradient de la fonction de perte pour la sortie du réseau $\boldsymbol{a}(\boldsymbol{x_{T_i}}; \boldsymbol{\theta})$ et la sortie cible $\boldsymbol{y_i}$ en fonction des paramètres $\boldsymbol{\theta}$:

$$g_{\theta_i} = \nabla_\theta L(a(x_{T_i};\ \theta), y_i) \qquad (2.12)$$

Mettre à jour les paramètres $\boldsymbol{\theta}$ en ajoutant le gradient moyen négatif, multiplié par le taux d'apprentissage $\boldsymbol{\eta}$, étant donné le nombre d'exemples d'apprentissage N :

$$\theta = \theta - \eta \frac{1}{N} \sum_{i=1}^{N} g_{\theta_i} \qquad (2.13)$$

Il existe plusieurs manières d'effectuer une descente de gradient (Maurice, s. d.) :

- De manière globale (batch gradient) : la totalité des données est envoyée au réseau en une seule fois, puis le calcul du gradient est effectué ainsi que l'ajustement des poids.

- Par lots (mini-batch gradient) : les données sont envoyées au réseau par petits groupes d'une taille définie par l'expérimentateur, dont les erreurs sont calculées, puis vient le calcul de l'erreur moyenne et enfin la mise à jour des poids.

- Par gradient unitaire, stochastique : une seule donnée est envoyée à la fois dans le réseau et ainsi, la mise à jour des poids se fait à chaque fois juste après.

2.5.2 Algorithme de rétropropagation

Bien que l'idée de la rétropropagation ait été introduite dans les années 1970(Schmidhuber, 2015), il a fallu attendre que Werbos l'applique pour la première fois en 1981 (Werbos, 1982) et que Rumelhart et al. la popularise en 1986 (Rumelhart et al., 1986) pour qu'elle soit utilisée pour les réseaux de neurones.

Le processus d'apprentissage se compose de deux processus principaux : la propagation en avant et la rétropropagation. D'abord, l'équation 3.3 est utilisée pour calculer les valeurs de chaque neurone $x_i^{(k)}$, c'est la propagation en avant. Ensuite, l'erreur ou la fonction de coût est évaluée sur la base des valeurs prédites et réelles, avant d'être rétro-propagée aux poids des couches précédentes, dans lesquelles le gradient de la fonction de coût est estimé de manière itérative. Ce processus est complété par la mise à jour des poids w_{ij} selon l'équation 2.11 en fonction des valeurs de gradient précédemment calculées(Dif, 2020).

Donc, pour calculer le gradient d'une fonction objective par rapport aux poids d'un réseau de neurones multicouches, la procédure de rétropropagation est une simple application pratique de la règle de la chaîne pour les dérivées(I. Goodfellow et al., 2016). L'équation de rétropropagation peut être appliquée de manière répétée pour propager les gradients à travers tous les modules, en partant de la sortie du haut (où le réseau produit sa prédiction) jusqu'au bas (où l'entrée externe est alimentée). Une fois que ces gradients ont été calculés, il est facile de calculer les gradients par rapport aux poids de chaque nœud.

Différentes variantes de la GD ont été proposées dans la littérature afin d'améliorer et d'accélérer l'apprentissage, comme la descente de gradient stochastique (SGD) et la descente de gradient en mini-batch (BGD). Par ailleurs, en vue d'optimiser la GD, plusieurs méthodes ont été développées, telles que : la descente de gradient inertielle, la descente de gradient accélérée de Nesterov (NAG), Adagrad, AdaDelta, Adam, et RmsProp(Dif, 2020).

2.6 Les techniques de régularisation

La descente de gradient et la rétropropagation sont utilisées pour entraîner un MLP sur un ensemble d'entraînement x_T avec les étiquettes correspondantes y, ceci lui permet de prédire les sorties de cet ensemble d'entraînement, mais cela ne signifie pas nécessairement que ce réseau est capable de prédire correctement les sorties pour des données inconnues. Pour optimiser les MLP, deux ensembles de données supplémentaires sont alors introduits : l'ensemble de validation x_{val} et l'ensemble de test x_{test}. Les trois ensembles (apprentissage, validation, test) doivent être disjoints. L'ensemble de validation est généralement utilisé pour ajuster les hyperparamètres du modèle FNN, tels que l'architecture du réseau ou le taux d'apprentissage. L'ensemble de

test n'est utilisé que pour l'évaluation finale, afin de vérifier les performances du FNN sur des données inédites.(I. Goodfellow et al., 2016).

Quand la capacité de généralisation d'un réseau de neurones est faible, c'est-à-dire quand la perte d'apprentissage est inférieure à la perte de test, on parle de *sur-apprentissage* (Overfitting), tandis que quand la perte de test est beaucoup plus faible que la perte d'apprentissage, on parle de *sous-apprentissage* (underfitting). En général, l'overfitting et le underfitting sont directement liés à la capacité de modélisation d'une méthode d'apprentissage automatique. Si cette capacité est trop faible, le réseau peut être incapable de s'adapter à l'ensemble d'apprentissage (sous-apprentissage), tandis qu'une capacité de modèle trop importante peut conduire à la mémorisation des échantillons d'apprentissage (sur-apprentissage). La Figure 2-8 illustre une représentation graphique de ses deux phénomènes dans un problème classique de l'apprentissage automatique ; le problème de classification(Serpe, 2021).

En général, les réseaux de neurones ne posent pas de problème de sous-apprentissage, car ils peuvent être résolus via l'utilisation d'une architecture de réseau plus puissante ou plus profonde avec plus de paramètres. En revanche, le sur-apprentissage est un problème classique important lorsqu'il s'agit d'utiliser les réseaux de neurones pour des données nouvelles et inédites.

La régularisation est le processus qui permet de réduire l'effet de l'overfitting ou de l'empêcher complètement(I. Goodfellow et al., 2016). Cette méthode consiste à ajouter un terme qui pénalise la fonction de coût. On distingue plusieurs méthodes de régularisation : La régularisation L_1 et L_2, la régularisation par abandon (Dropout), l'augmentation des données et l'arrêt prématuré (Early Stopping).

2.6.1 Les régularisations L1 et L2

Le sur-apprentissage concerne les réseaux qui obtiennent de très bonnes performances sur les données d'apprentissage et des performances médiocres sur les données de test, entraînant ainsi une importante variance entre les deux et une mauvaise généralisation. Une des solutions couramment utilisées est l'ajout d'un terme

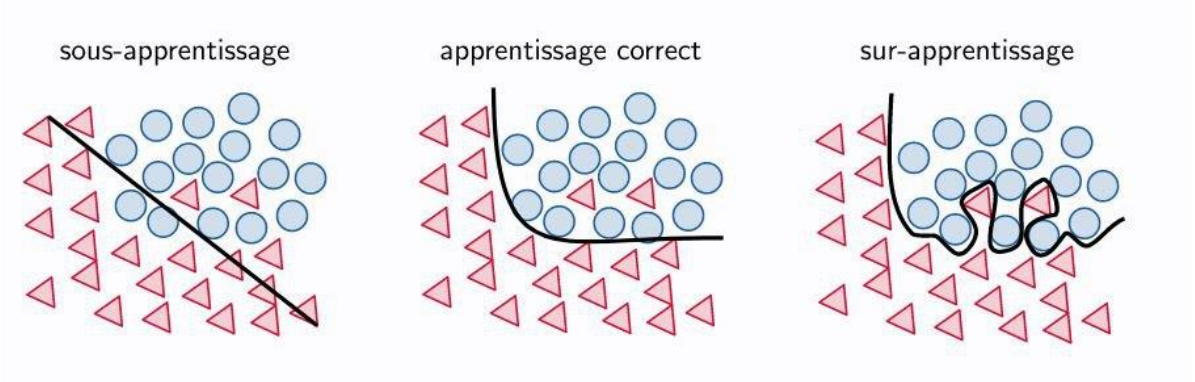


Figure 2-8 : Le sous-apprentissage et le sur-apprentissage pour un problème de classification.
Les triangles et les cercles représentent les données de deux classes différentes, la ligne noire vise à les séparer.

de régularisation à la fonction de coût J à minimiser. Cela permettra de pénaliser les poids de trop forte connexion(Hardy, 2019), en diminuant la valeur de ces poids par l'ajout du terme de régularisation. En pratique, elle est appliquée après le calcul de la fonction de perte, comme suit :

- La somme des poids absolus, multipliée par une constante λ, pour la Régularisation L1
- La somme des poids au carré multipliée par une constante λ, pour la Régularisation L2. Cette méthode est connue également sous le nom de méthode de dégradation du poids (weight decay).

2.6.2 La régularisation par abandon (Dropout)

Il s'agit d'une méthode proposée précisément pour les réseaux de neurones en 2014 par Srivastava et al. (Srivastava et al., 2014a) consistant à déconnecter aléatoirement des neurones de façon provisoire, à chaque itération, et cela en multipliant sa valeur de sortie par zéro, il est possible de supprimer efficacement un neurone artificiel d'un MLP. Ces neurones qui appartiennent aux couches d'entrée ou cachées sont sélectionnés aléatoirement en fonction d'un taux d'abandon, qui est généralement fixé à 0,5 pour les neurones cachés et à 0,8 pour les neurones d'entrée. Cela signifie que pour chaque échantillon de l'ensemble d'apprentissage, chaque neurone a une chance d'être actif, égale au taux d'abandon. Ainsi, le réseau de neurones ne dépend pas uniquement de neurones artificiels spécifiques, mais distribue le calcul à plusieurs neurones, ce qui réduit l'impact de l'adaptation excessive(I. Goodfellow et al., 2016).

2.6.3 L'augmentation des données

L'augmentation des données permet de réduire de manière simple mais efficace le sur-apprentissage, car elle permet d'augmenter le volume de données par la modification de celles qui sont déjà disponibles tout en maintenant la classe d'appartenance des données originales (Etienne, 2019).

Il est possible de procéder à l'augmentation des données soit avant l'étape d'apprentissage (hors ligne), soit pendant la formation en mini-lots (en ligne)(Dif, 2020). Plusieurs méthodes d'augmentation des données ont été proposées : la translation, la rotation, la division des patchs, l'écaillage, l'amélioration des couleurs, le bruit gaussien, la normalisation et la génération de nouvelles instances par des réseaux adversaires génératifs (GAN) sont quelques-unes des techniques les plus couramment utilisées.

2.6.4 L'arrêt prématuré (Early Stopping)

La procédure d'arrêt prématuré est une alternative à la régularisation comme moyen de contrôler la complexité effective d'un réseau (Bishop, 2016). Il s'agit d'une méthode de régularisation implicite (C. Zhang et al., 2021). Pour cette méthode, l'ensemble de données est subdivisé en deux bases de formation et de validation. Quand la capacité du modèle d'un réseau de neurones est suffisamment grande pour permettre le sur-apprentissage, la perte d'apprentissage diminue régulièrement jusqu'à la convergence, tandis que la perte de validation diminue au début et augmente à nouveau après que le réseau de neurone commence à sur-ajuster.

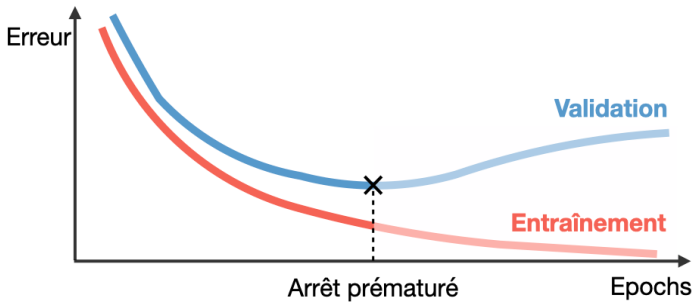


Figure 2-9 : Modélisation de l'arrêt prématuré et le sur-apprentissage (Amidi, 2018/2022).

Il est recommandé d'arrêter l'apprentissage lorsque la valeur de l'erreur sur la base de validation commence à augmenter (Figure 2-9) (Bishop, 2016). Au final, et pour la phase de prédiction, le modèle minimisant le taux d'erreur est enregistré. Ce modèle peut potentiellement mieux se généraliser aux données non vues. Par conséquent, l'arrêt précoce empêche essentiellement le surajustement avant qu'il n'ait un impact mesurable sur la perte de validation.

2.7 Les différents modèles de l'apprentissage profond

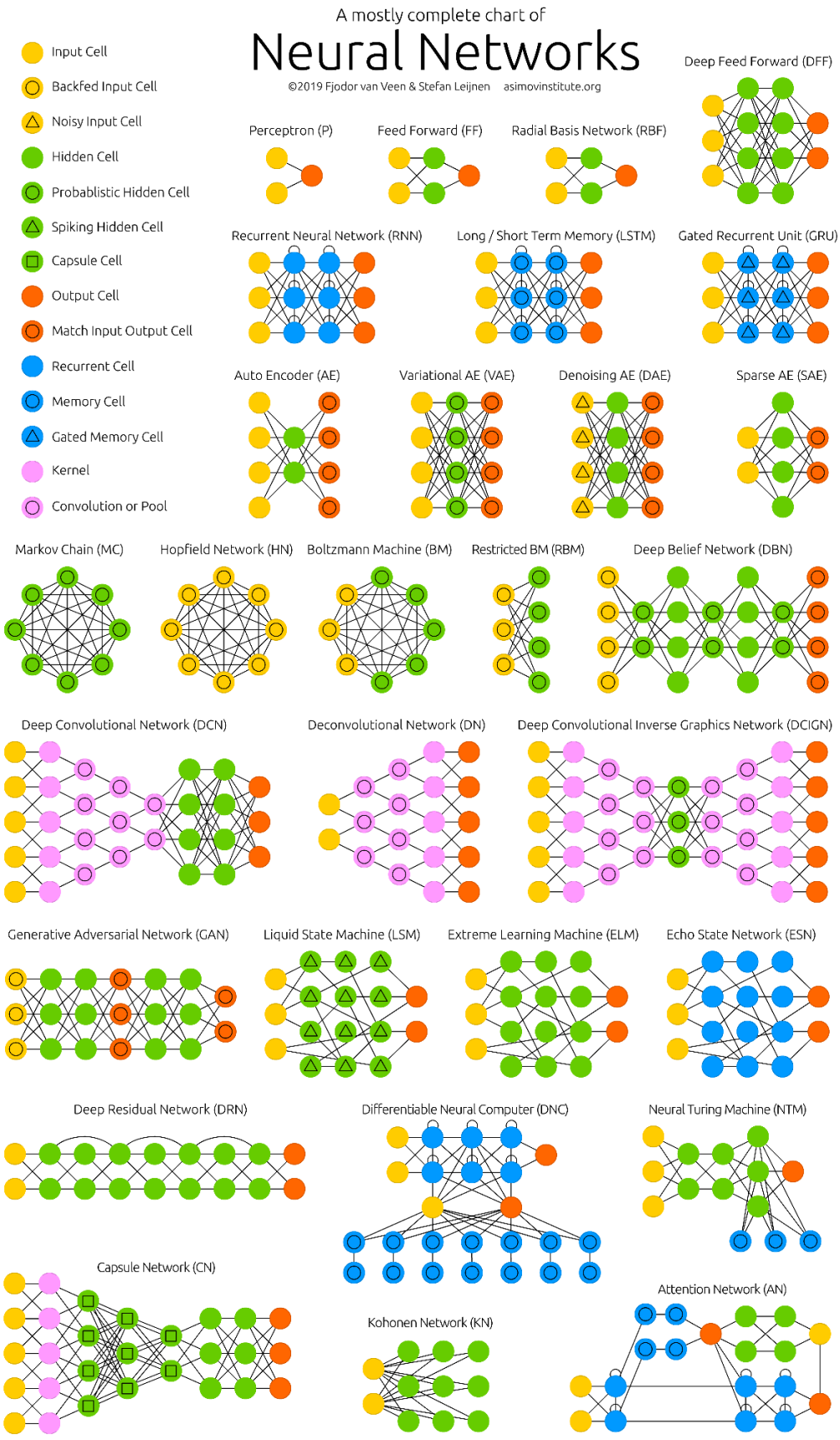


Figure 2-10 : Différents types de réseaux neuronaux développés de 2004 à 2019(Veen, 2016).

Pour effectuer la tâche d'apprentissage, la première étape des algorithmes d'apprentissage conventionnels consiste à extraire des caractéristiques à partir des données brutes. L'objectif visé est de parvenir à une représentation de plus haut niveau des données(Hardy, 2019). Une bonne connaissance des données et de la tâche d'apprentissage est nécessaire pour extraire des caractéristiques à partir de données brutes, ainsi qu'un travail d'ingénierie pour adapter les méthodes d'extraction. La mise en place de cette opération est relativement coûteuse, elle dépend du contexte, et ainsi une mauvaise extraction de caractéristiques conduit à de très faibles performances d'apprentissage.

Depuis leur émergence, les techniques d'apprentissage profond ont suscité un intérêt croissant de la part des chercheurs en raison de leur capacité à surmonter cet inconvénient. Les approches d'apprentissage profond se sont également révélées adaptées à l'analyse des données volumineuses, avec des applications réussies dans les domaines de la vision par ordinateur, de la reconnaissance des formes, de la reconnaissance vocale, du traitement du langage naturel et des systèmes de recommandation.

Dans les réseaux de neurones profonds, le terme "*profond*" fait référence au nombre de couches entre l'entrée et la sortie. Ainsi, un réseau ne comportant qu'une seule couche cachée est appelé réseau *peu profond*, et inversement, un réseau comportant plus de 2 couches cachées est dit profond. Aujourd'hui, des réseaux d'une centaine, voire d'un millier de couches sont déjà disponibles pour les réseaux les plus profonds(Szegedy et al., 2015).

Plusieurs types d'architectures optimisées ont été proposés pour l'apprentissage supervisé comme les réseaux de neurones à convolution (CNN), réseaux de neurones récurrents (RNN), mémoire à long terme (LSTM)) et pour l'apprentissage non supervisé comme les auto-encodeurs (AE), les réseaux adversaires génératifs (GAN). La Figure 2-10 est tirée de (Veen, 2016) illustre les principales architectures proposées dans la littérature. Il devient donc crucial d'automatiser leur conception. Dans cette partie du chapitre, nous présentons quelques architectures d'apprentissage profond largement utilisées et leurs applications pratiques.

2.7.1 Les réseaux de neurones à convolution

Le réseau de neurones convolutifs (Convolutional Neural Network, CNN ou ConvNet) est inspiré du cortex visuel des vertébrés et a été proposé pour la première fois par Fukushima en 1988 (Fukushima, 1988). Cependant, il n'a pas été largement utilisé en raison des limitations du matériel de calcul pour l'entraînement du réseau. Puis, en 1990, LeCun et al. (Lecun et al., 1998) ont appliqué avec succès un algorithme d'apprentissage basé sur le gradient aux CNN pour le problème de la classification des chiffres manuscrits. Par la suite, la communauté scientifique a continué à améliorer les

CNN et a obtenu des performances excellentes dans de nombreuses tâches (Alom et al., 2019) comme le traitement du langage naturel (Y. Kim, 2014), vision par ordinateur (Krizhevsky et al., 2012) et les systèmes de recommandation (Ying et al., 2018). En fait, dans les premières années 2000, à cause de la puissance insuffisante des processeurs et des capacités des mémoires internes limitées pour les besoins de tels algorithmes, les travaux de recherche sur les CNN ont stagné. Pendant cette période, les algorithmes classiques d'apprentissage automatique ont été largement exploités en raison de leurs exigences raisonnables en termes de complexité de calcul et d'espace de stockage. C'est en 2012 notamment que l'architecture AlexNet, de type CNN, a obtenu le meilleur taux d'erreur de l'état de l'art sur la base d'apprentissage ImageNet (Krizhevsky et al., 2012). Grâce à ces excellents résultats et à la capacité des GPU à optimiser la complexité temporelle, la communauté scientifique a commencé à proposer de nouvelles versions optimisées des CNN(Dif, 2020).

Les réseaux CNN se caractérisent par l'utilisation de l'opération de convolution dans les premières couches du réseau de neurones. Cette opération est à l'origine utilisée comme un filtre en imagerie ou dans le domaine du son pour mettre en évidence des motifs ou réduire un type de bruit (Hardy, 2019). Les réseaux CNN, plutôt que les méthodes d'apprentissage classiques, effectuent automatiquement le processus d'extraction des caractéristiques par le biais de ses couches convolutives, dont les premières représentent les caractéristiques simples, qui sont ensuite combinées pour former d'autres plus complexes dans les couches profondes(Dif, 2020).

2.7.1.1 L'architecture générale d'un réseau de neurones convolutif

La Figure 2-11 illustre l'architecture globale standard d'un CNN. Il est principalement composé de trois types de couches hétérogènes :

- La couche de convolution, qui a pour objectif à identifier la présence d'un ensemble de caractéristiques dans les images d'entrée.
- La couche de mise en commun qui consiste à réduire la taille des images, tout en préservant leurs caractéristiques importantes.
- La couche entièrement connectée a pour but de prédire la classe de l'image.

A. La couche de convolution

La couche de convolution est le composant clé des réseaux neuronaux convolutifs et constitue toujours leurs premières couches, pour effectuer une extraction implicite des

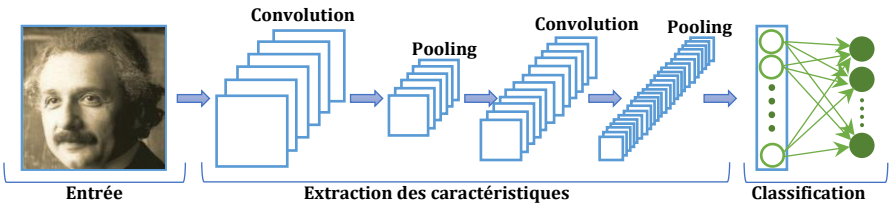


Figure 2-11 : Architecture globale d'un CNN.

caractéristiques (features). Ces couches prennent une image en entrée, et convoluent ces entrées avec un ensemble de banques de filtres de manière glissante pour produire des *cartes de caractéristiques* qui représentent une disposition spatiale de l'image d'entrée. Il existe plusieurs types de convolutions, la plus utilisée étant la convolution classique, mais d'autres types ont également été largement adoptés de nos jours, comme la convolution dilatée ou la convolution séparable.

La convolution classique utilise des filtres qui balaient l'image d'entrée I, selon ses dimensions, en effectuant des opérations de convolution. Elle peut être ajustée en adaptant la taille du filtre F et le pas S. La sortie O de cette opération est la carte de caractéristiques. Mathématiquement, pour une image bidimensionnelle I, la convolution est définie comme suit :

$$O(i,j) = (I * F)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \qquad (2.14)$$

où $K(m,n)$ est un noyau bidimensionnel(kernel), et la sortie $S(i,j)$ constitue la carte des caractéristiques(I. Goodfellow et al., 2016).

La couche de convolution est caractérisée par trois hyperparamètres : La profondeur de couche (c'est nombre de noyaux de convolution), le pas ou le « stride » (c'est le décalage du noyau entre chaque calcul) et le remplissage à zéro (il est courant de mettre des zéros à la limite de l'image d'entrée, cette marge permet de contrôler la dimension spatiale de l'image de sortie.

B. La couche de mise en commun

Cette couche de regroupement, dite aussi couche de sous échantillonnage, elle est plus connue sous son nom en anglais « *pooling* », Ce type de couche est généralement placé entre deux couches de convolution. Ces couches servent à réduire la résolution spatiale de la représentation en faisant par exemple la moyenne des cartes de caractéristiques d'entrée données, afin d'ignorer les variations des petits décalages et les distorsions géométriques(Lecun et al., 1998). Le pooling est une opération simple qui consiste à remplacer un carré de pixels (généralement 2×2 ou 3×3) par une seule valeur. Il existe plusieurs types de pooling :

- Le "max pooling", qui consiste à prendre la valeur maximale de la sélection. C'est le type le plus utilisé car il est rapide à calculer (immédiat), et permet de simplifier efficacement l'image.

- Le "mean pooling" (ou average pooling) est la moyenne des pixels de la sélection : la somme de toutes les valeurs est calculée et divisée par le nombre de valeurs. Ainsi, une valeur intermédiaire est obtenue pour représenter cet ensemble de pixels.

C. La couche entièrement connectée

Les couches entièrement connectées (Fully Connected FC) d'un réseau CNN ont la même structure qu'un perceptron multi couches MLP. Ces couches ont pour but d'apprendre les combinaisons non linéaires entre les caractéristiques extraites par les couches convolutionnelles. La sortie de la dernière couche de convolution $[N, N, N_c]$ est aplatie en un vecteur de taille $[N * N * N_c]$. Celui-ci constitue la couche d'entrée pour les couches entièrement connectées.

Lorsqu'il reçoit le vecteur aplati, il produit un nouveau vecteur de sortie, en appliquant une combinaison non linéaire et éventuellement une fonction d'activation (habituellement, la fonction Softmax) aux valeurs d'entrée. Le vecteur de sortie sera de taille N, où N est le nombre de classes dans le problème de classification. Chaque élément du vecteur indique la probabilité que l'image d'entrée appartienne à une classe.

2.7.1.2 Un aperçu des architectures de réseaux neuronaux convolutifs les plus connues

Depuis 2011, une compétition a été lancée, ImageNet's Large-Scale Visual Recognition Challenge (ILSVRC), qui évalue les algorithmes de détection d'objets et de classification d'images à grande échelle. La motivation est de permettre aux chercheurs de comparer les progrès de la détection sur une plus grande variété d'objets et de mesurer les progrès de la vision par ordinateur pour l'indexation d'images à grande échelle pour la recherche et l'annotation. Cette compétition a donné naissance à plusieurs architectures de réseaux de neurones convolutifs au cours de la dernière décennie.

LeNet est le premier CNN proposé par LeCun et al. (Lecun et al., 1995) pour identifier les numéros manuscrits sur les chèques dans la plupart des banques aux États-Unis. Ils ont utilisé deux couches convolutives pour l'extraction des caractéristiques, deux couches de max-pooling pour le sous-échantillonnage spatial et deux couches entièrement connectées pour la classification. Il s'agit d'environ 60 000 paramètres, la plupart dans les deux dernières couches. En 1998, LeCun et al. (Lecun et al., 1998) ont utilisé l'architecture LeNet-5 (un des multiples modèles proposés dans (Lecun et al., 1995)) pour la reconnaissance de caractères manuscrits, sur un ensemble de test de 10 000 exemples, il atteint un score de 0,7% de taux d'erreur brut. Ce réseau a défini les composants de base du CNN, mais il nécessitait une capacité de calcul élevée par rapport au matériel existant à cette époque, ce qui ne lui a pas permis d'être aussi populaire et utilisé que d'autres algorithmes, comme les SVM, qui pouvaient obtenir des résultats similaires, voire meilleurs.

En 2012, Krizhevsky et al. (Krizhevsky et al., 2012) a proposé le modèle AlexNet, où ils ont ajouté quelques couches supplémentaires au LeNet-5, et ils ont été les premiers à mettre en œuvre et à utiliser comme fonctions d'activation les unités linéaires rectifiées (ReLU). Ce réseau a été parmi les premiers à introduire la méthode du "dropout" pour

résoudre le problème de l'overfitting (suggéré par (Srivastava et al., 2014b)). AlexNet est composé de huit (8) couches : cinq convolutives, certaines de ces couches sont suivies de couches de max-pooling, et trois entièrement connectées. Ce réseau possède 60 millions de paramètres et 650 000 neurones. Il a été le vainqueur du concours ImageNet LSVRC-2012 ; il a obtenu des taux d'erreur de 15,3 % dans le top 5, alors que la deuxième meilleure entrée a atteint 26,2 %. L'idée de cette architecture est devenue le "modèle" de base pour les futurs réseaux, mais plus tard, de nouvelles architectures seront proposées, avec une taille beaucoup plus petite et la même précision.

C'était la première fois qu'un modèle réalisait des performances aussi brillantes, ce record de performance de ce réseau a illustré les avantages des CNNs. Depuis lors ; le défi pour les chercheurs a été de proposer des réseaux de plus en plus profonds, et d'étudier l'impact sur l'efficacité des CNNs.

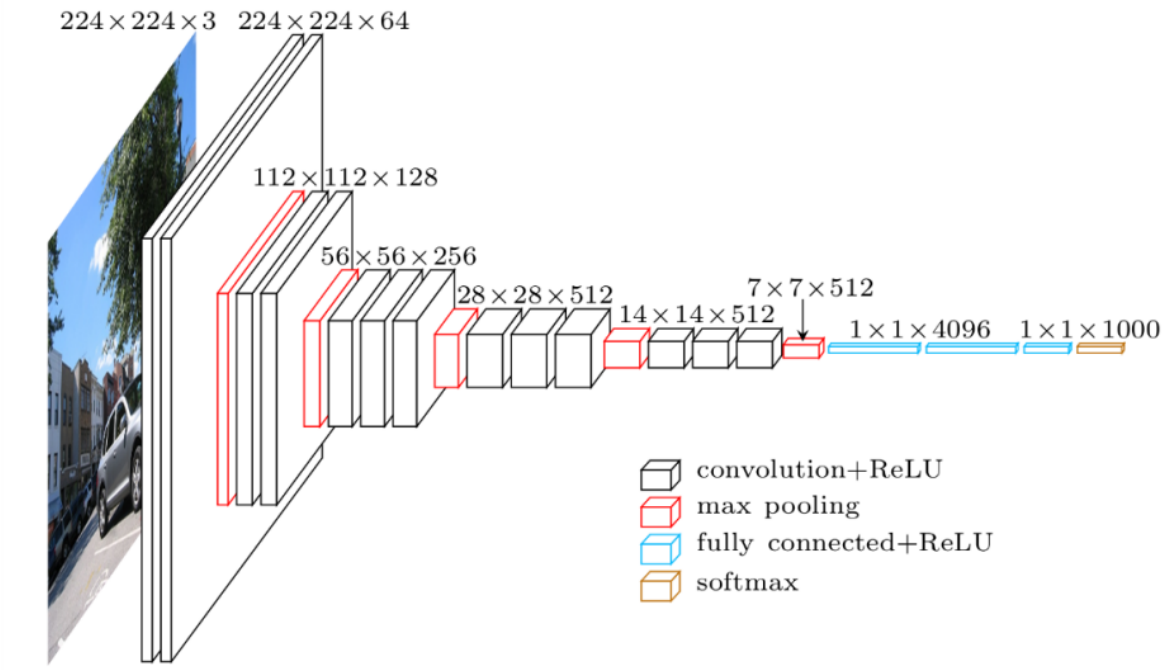


Figure 2-12 : Architecture du VGG16 (Hassan, 2018).

Ainsi, en 2014, Simonyan et Zisserman ont proposé un réseau très profond appelé VGG16 (Simonyan & Zisserman, 2014), dans leur travail, ils évaluent les réseaux, en augmentant la profondeur et avec des filtres de convolution très petits (3x3). Cette architecture a remporté la première et la deuxième place dans le domaine de la localisation et de la classification, respectivement, lors du défi ImageNet 2014. Cette architecture était simple et profonde, elle se compose de 16 couches : 13 couches de convolution et trois couches entièrement connectées, toutes les couches cachées sont équipées de la non-linéarité de rectification (ReLU). L'architecture est résumée dans la Figure 2-12. Ce réseau avait 138M de paramètres et utilisait environ 500MB d'espace de stockage, ce qui le ralentissait péniblement lors de l'entraînement.

Depuis, une nouvelle tendance a émergé dans les architectures CNN, à savoir la structure " network in network " (M. Lin et al., 2014). Par conséquent, le CNN n'est plus un simple empilement de couches, mais plutôt une architecture plus compliquée, basée sur l'utilisation d'un module d' « Inception » proposé par Google dans (Szegedy et al., 2015). Le module d'Inception est un bloc de filtres parallèles de différentes tailles (1×1, 3×3, 5×5) et un pooling max 3×3, dont les résultats sont ensuite concaténés.

L'implémentation de cette structure naïve du modèle d'inception pose un problème de coût de calcul des filtres (lorsqu'un grand nombre de filtres de taille importante sont utilisés). Pour l'améliorer, ils ont proposé d'utiliser un filtre de convolution 1x1. Sur la base de ce réseau et du module Inception, plusieurs autres améliorations ont été apportées à cette architecture, donnant naissance à d'autres réseaux tels que Inception v2 et Inception v3 (Szegedy et al., 2016), Inception v4 (Szegedy et al., 2017).

Après cela, He et al. (He et al., 2016) a proposé ResNet en 2015 ; ils ont introduit le concept d'apprentissage résiduel dans les CNN, ce qui a révolutionné le développement architectural des CNN. Resnet a été évalué sur le jeu de données ImageNet, avec une profondeur allant jusqu'à 152 couches, soit huit de plus que les réseaux VGG, tout en ne compromettant pas la capacité de généralisation du modèle et avec moins de complexité. Cette architecture a un taux d'erreur de 3,57% sur le jeu de test ImageNet, et remporte la première place dans la tâche de classification ILSVRC 2015. Elle vise à réduire l'erreur de formation. Dans ce travail, les auteurs ont popularisé l'utilisation des "skip connections" car ils ne sont pas les premiers à les utiliser. Leurs CNNs sont beaucoup plus profonds que leurs prédécesseurs, avec une forte utilisation de la normalisation des lots.

Diverses améliorations ont été proposées telles que Inception-ResNet V1 et V2 (Szegedy et al., 2017), Wide Residual Networks (Zagoruyko & Komodakis, 2017), ResNeXt (Xie et al., 2017), Deep Networks with Stochastic Depth (G. Huang et al., 2016). Dans (G. Huang et al., 2017), les auteurs présentent DenseNet comme un réseau convolutif dense qui connecte chaque couche à toutes les autres couches de manière feed-forward, ce qui le rend plus précis et plus facile à former. Dans leur expérimentation, ils ont utilisé la même architecture générale que le modèle ResNet, mais ont simplement remplacé le bloc dense par l'unité répétée. Par rapport aux modèles ResNet, les DenseNets sont considérés comme plus efficaces et moins complexes.

Chollet (Chollet, 2017a) a proposé une autre architecture CNN intéressante : Le modèle Xception, qui est une interprétation des modules Inception, en remplaçant ces derniers par des convolutions séparables en profondeur. Ce modèle possède un nombre de paramètres similaire à celui d'Inception V3. Cependant, par rapport à Inception V3, Xception présente de faibles avantages en termes de performance de classification sur le jeu de données ImageNet et de grands avantages sur le jeu de données JFT.

Récemment, une étude propose une nouvelle méthode de mise à l'échelle des modèles pour obtenir de meilleures performances. En proposant EfficientNet (Tan & Le, 2019), ils ont démontré une efficacité remarquable en augmentant conjointement la largeur, la profondeur et la résolution du réseau. En utilisant cette méthode de mise à l'échelle composée, ils démontrent qu'un modèle EfficientNet de taille mobile peut être mis à l'échelle très efficacement, dépassant la précision de l'état de l'art avec moins de paramètres et de FLOPS, à la fois sur ImageNet et sur cinq jeux de données d'apprentissage par transfert couramment utilisés.

Dans Tableau 2-1, nous résumons quelques-uns des CNN les plus populaires et les plus performants.

Tableau 2-1: Les réseaux CNN les plus connus et les plus performants.

CNN	**Référence**	**Description**	**Année**
LeNet	(Lecun et al., 1998)	Des empilements de convolutions pour l'extraction de caractéristiques et des opérations de regroupement maximal pour le sous-échantillonnage spatial.	1998
AlexNet	(Krizhevsky et al., 2012)	Architecture traditionnelle CNN en couches qui consiste en huit couches apprises : cinq couches de convolution suivies de couches de max-pooling et d'unités linéaires rectifiées (ReLU) et trois couches entièrement connectées.	2012
VGG	(Simonyan & Zisserman, 2014)	Une architecture d'une pile de couches de convolution avec de très petits (3×3) filtres de convolution. Cinq couches de max-pooling suivent certaines des couches de convolution. La mise en commun maximale est effectuée sur une fenêtre de 2×2 pixels, avec un pas de 2. Trois couches entièrement connectées (FC) suivent les couches convolutionnelles. La dernière couche est la couche soft-max. Toutes les couches cachées sont équipées de la rectification ReLU.	2014
GoogLeNet	(Szegedy et al., 2015)	L'architecture se compose de plusieurs couches "Inception", chacune agissant comme un micro-réseau dans un réseau plus large, ce qui permet à l'architecture de prendre des décisions plus complexes. Le réseau a une profondeur de 22 couches avec de petits filtres de convolution (1x1, 3x3 et 5x5). Toutes les convolutions utilisent une activation linéaire rectifiée.	2014
ResNet	(He et al., 2016)	Basée sur des blocs résiduels, cette architecture est composée de 152 couches et se caractérise par des connexions de saut spéciales et une utilisation importante de la normalisation par lots. L'architecture ne comporte pas non plus de	2015

		couches entièrement connectées à l'extrémité du réseau.	
Xception	(Chollet, 2017b)	Les couches de convolution à séparation profonde constituent la base de cette architecture de réseau CNN, qui compte 36 couches de convolution structurées en 14 modules, qui ont tous des connexions résiduelles linéaires autour d'eux, à l'exception du premier et du dernier modules.	2017
EfficientNet	(Tan & Le, 2019)	Une famille de modèles développée à partir d'une méthode qui met uniformément à l'échelle chaque dimension avec un ensemble fixe de coefficients d'échelle.	2019

2.7.1.3 Domaines d'application des CNN

Ces dernières années, l'apprentissage profond a été amplement utilisé dans divers domaines, et plus spécifiquement les réseaux de neurones convolutifs. Ceux-ci ont connu un véritable succès dans le domaine de la vision par ordinateur, où ils ont surpassé d'autres méthodes d'apprentissage, atteignant notamment une bonne précision dans la résolution de problèmes du monde réel.

Les domaines d'application des CNN vont de la classification d'images à la segmentation d'images, en passant par la reconnaissance faciale, l'estimation de pose, la détection et la localisation d'objets dans les images. Les CNN sont également exploités efficacement en imagerie médicale et dans de nombreux autres domaines.

Le but de la classification d'images est de classer une image dans une ou plusieurs classes. Ce problème, également connu sous le nom de classification d'objets ou de reconnaissance d'images, est un problème de base de la vision par ordinateur. D'autres tâches de vision par ordinateur, telles que la localisation, la détection ou la segmentation, en dépendent. La base de données ImageNet, composée de 15 millions d'images étiquetées, a attiré beaucoup d'attention, plusieurs CNN testés sur cette base ont atteint des performances encourageantes, ce qui attiré les chercheurs dans le domaine de vision à les exploiter dans divers applications, tel que l'analyse et classification des images médicales(Ayadi et al., 2021; Benzebouchi et al., 2019), identification des plantes médicinales (Nguyen Quoc & Truong Hoang, 2020) , Classification des images hyperspectrales (M. Zhang et al., 2018), reconnaissance de l'écriture (El-Sawy et al., 2017; Gan et al., 2019), reconnaissance faciale (Coşkun et al., 2017; Lawrence et al., 1997).

La détection d'objets, qui vise à cerner un ou plusieurs objets dans une image à l'aide de cadres de sélection, est une tâche plus difficile que la classification, du fait qu'elle réunit classification et localisation(Dif, 2020). Le réseau R-CNN proposé en 2014(Girshick et al., 2014), qui a obtenu des résultats satisfaisants tout en réduisant le temps de détection, fait partie des premiers travaux à avoir obtenu un grand succès en

matière de détection. Cependant, cette performance n'était pas à la hauteur des applications en temps réel, c'est pourquoi d'autres architectures ont été proposées comme le Fast R-CNN (Girshick, 2015) et le Faster R-CNN (Ren et al., 2015). Ensuite, c'est l'algorithme YOLO (Redmon et al., 2016), qui a connu avec ses différentes versions et améliorations beaucoup de succès, il a transformé le problème de détection en un problème de régression.

En reconnaissance faciale, qui est considéré comme le système le plus pratique en biométrie, le problème consiste à identifier les visages et les reconnaitre après leur détection sur les images ou les vidéos. Ce problème rencontre plusieurs défis, comme le changement de pose, les problèmes d'éclairage et la variation des expressions faciales. Plusieurs travaux basés sur les réseaux de neurones convolutif ont été proposés, parmi lesquels, nous citons : Face Descriptor(Parkhi et al., 2015), Facenet (Schroff et al., 2015), OpenFace (Amos et al., 2016) et light CNNs (X. Wu et al., 2018).

En plus de la vision par ordinateur, les CNN ont été exploité en traitement du langage naturel (Collobert & Weston, 2008; Kalchbrenner et al., 2014), en reconnaissance de la parole(Hou et al., 2018; Y. Xu et al., 2017), détection des intrusions dans les réseaux(Vinayakumar et al., 2017; H. Yang & Wang, 2019) et e, télédétection (J. Yang et al., 2019; W. Zhang et al., 2019).

2.7.2 Les auto-encodeurs

Un réseau neuronal auto-encodeur est un algorithme d'apprentissage non supervisé qui applique la rétropropagation, en fixant les valeurs cibles comme étant égales aux entrées(A. Ng, 2011). C'est la combinaison d'une fonction d'encodage qui convertit les données d'entrée en une représentation différente, et d'une fonction de décodage qui reconvertit la nouvelle représentation dans le format d'origine. Les auto-codeurs sont entraînés à préserver autant d'informations que possible lorsqu'une entrée passe par le codeur puis par le décodeur, mais ils sont également entraînés à faire en sorte que la nouvelle représentation ait diverses propriétés agréables. Différents types d'auto-codeurs visent à obtenir différents types de propriétés(I. Goodfellow et al., 2016).

Les auto-encodeurs sont des réseaux neuronaux constitués de deux parties : un encodeur et un décodeur. L'objectif principal d'un auto-encodeur est d'apprendre et de représenter (encodage) les données d'entrée, généralement pour la réduction de la dimensionnalité des données, la compression, la fusion et bien d'autres choses encore.

Dans la phase d'encodage, les données d'entrée ($x \in \mathbb{R}^n$) sont généralement mises en correspondance dans l'espace des caractéristiques de dimension inférieure($\mathbb{R}^m$) avec une représentation constructive des caractéristiques (c'est-à-dire $n > m$). Cette approche peut être répétée jusqu'à ce que l'espace dimensionnel souhaité soit atteint. Tandis que dans la phase de décodage, les caractéristiques réelles sont régénérées à partir des caractéristiques de dimension inférieure par un traitement inverse (c'est-à-dire les données x sont transformées par le décodeur de l'espace $\mathbb{R}^m$ vers l'espace $\mathbb{R}^n$)(Alom et al., 2019). L'auto-encodeur peut avoir autant de couches que nécessaire, généralement placées de façon symétrique dans l'encodeur et le décodeur.

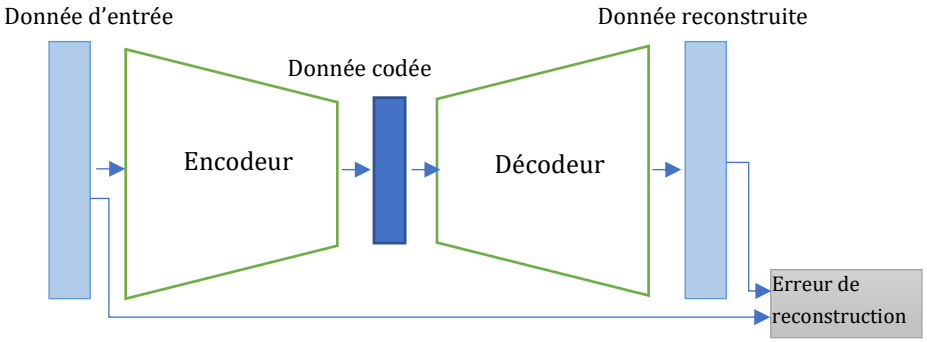


Figure 2-13: Structure d'un auto-encodeur.

La Figure 2-13 présente le schéma d'un auto-encodeur avec les phases d'encodage et de décodage. Cette architecture est généralement décomposée en plusieurs parties :

- La donnée d'entrée x
- La fonction de codage f
- La représentation interne « donnée codée » $h = f(x)$, Goulot d'étranglement (ou bottleneck).
- La fonction de décodage g
- La sortie « donnée reconstruite » $r = g(h) = g(f(x))$
- La fonction de perte L calculant un scalaire $L(r, x)$ qui mesure la qualité de la reconstruction r de l'entrée donnée x.

Ces deux réseaux neuronaux sont entrainés simultanément. L'objectif est de minimiser la valeur attendue de L sur l'ensemble d'exemples d'apprentissage $\{x\}$. Autrement dit, minimiser l'erreur de "reconstruction" du décodeur sur les données encodées par l'encodeur.

Le processus de formation d'un AE peut être divisé en deux étapes : la première étape consiste à apprendre des caractéristiques à l'aide d'un apprentissage non supervisé et la seconde consiste à affiner le réseau à l'aide d'un apprentissage supervisé. Pour être précis, dans la première étape, la propagation feed-forward est d'abord effectuée pour chaque entrée afin d'obtenir la valeur de sortie (la donnée reconstruite) r . Ensuite, les erreurs quadratiques sont utilisées pour mesurer la déviation de r par

rapport à la valeur d'entrée. Enfin, l'erreur sera rétro-propagée à travers le réseau pour mettre à jour les poids. Dans l'étape de réglage fin, le réseau ayant des caractéristiques appropriées à chaque couche, nous pouvons adopter la méthode d'apprentissage supervisé standard et l'algorithme de descente de gradient pour ajuster les paramètres à chaque couche(Bengio et al., 2006). Le problème de l'évanouissement du gradient est toujours un problème important avec le modèle plus profond de l'EA : le gradient devient trop petit lorsqu'il traverse de nombreuses couches d'un modèle d'EA (Alom et al., 2019).

L'un des principaux avantages de l'auto-encodeur est que ce modèle peut extraire des caractéristiques utiles en continu pendant la propagation et filtrer les informations inutiles. En outre, puisque le vecteur d'entrée est transformé en une représentation de dimension inférieure lors du processus de codage, l'efficacité du processus d'apprentissage peut être améliorée (W. Liu et al., 2017).

Différents modèles avancés d'EA sont présentés dans la littérature, parmi lesquelles, nous citons :

- *Auto-encodeurs de débruitage* (Denoising Autoencoders)(Vincent et al., 2008), qui ajoutent intentionnellement des bruits dans les données d'apprentissage et entraîne les AE avec ces données déformées (dits aussi corrompues). Grâce au processus d'entraînement, le DAE peut récupérer la version sans bruit des données d'entraînement, ce qui implique une robustesse accrue.
- *Auto-encodeurs éparses* (sparse autoencoders) (H. Lee et al., 2006), qui sont une approche permettant d'apprendre automatiquement des caractéristiques à partir de données non étiquetées. Les représentations éparses sont utilisées pour produire une interprétation simple des données d'entrée en extrayant la structure cachée des données. L'algorithme d'apprentissage de la représentation éparse a été proposé pour la première fois par Ranzato en 2006(Ranzato et al., 2006). Il s'agit de construire une fonction objective en sanctionnant les activations d'une couche en fonction de l'observation en question. Ainsi, à partir de l'activation d'un certain nombre de neurones, le réseau de neurones effectue l'encodage et le décodage (Sow, 2020).
- *Auto-encodeurs variationnels* (Variational Autoencoders), ils se sont imposés comme l'une des approches les plus populaires de l'apprentissage non supervisé de distributions complexes, au cours des trois dernières années. Les VAE se sont déjà montrés prometteurs dans la génération de nombreux types de données complexes, notamment les chiffres manuscrits, les visages, les numéros de maison, les images CIFAR, les modèles physiques de scènes,

la segmentation et la prédiction de l'avenir à partir d'images statiques (Doersch, 2021). Dans ce modèle, il y a deux pertes, l'une est une erreur quadratique moyenne qui détermine la qualité de la reconstruction de l'image par le réseau, et la perte (la divergence de Kullback-Leibler (KL)) de la variable latente, qui détermine dans quelle mesure la correspondance de la variable latente est proche de la distribution gaussienne unitaire (Alom et al., 2019).

Les auto-encodeurs ont été appliqués en bio-informatique (T. Wang et al., 2021) et en cybersécurité (Alom & Taha, 2017). Ils ont également été utilisés comme technique de codage et de décodage avec ou pour d'autres approches d'apprentissage profond, notamment CNN, DNN, RNN et RL au cours de la dernière décennie.

2.7.3 Les réseaux génératifs antagonistes

Le concept de réseau génératif antagonistes (GAN) est devenu un outil puissant pour diverses tâches de synthèse d'images et de vidéos, permettant la synthèse de contenu visuel de manière inconditionnelle ou conditionnelle à l'entrée. Il a permis de générer des images et des vidéos photoréalistes à haute résolution, une tâche qui était difficile, voire impossible, avec les méthodes précédentes. Elle a également conduit au développement de nombreuses nouvelles applications dans la création de contenu(M.-Y. Liu et al., 2021).

2.7.3.1 L'architecture générale d'un réseau de génératif antagoniste

Ian Goodfellow a introduit les GAN en 2014 (I. Goodfellow et al., 2014), en se basant sur la théorie des jeux. Les Gan consistent en deux sous-modèles différentiables, généralement implémentés par des réseaux de neurones profonds. Ainsi, un GAN typique est constitué de deux composants, dont l'un est un discriminateur D (avec les paramètres θ_D) qui distingue les images réelles des images fausses, tandis que l'autre est un générateur G (avec les paramètres θ_G) qui crée des images fausses, mais plausibles, pour tromper le discriminateur. Réelles dans le sens où elles proviennent de données de la base d'apprentissage et fausses en tant que données synthétiques générées.

Tableau 2-2 : L'entrée et la sortie des composants du GAN.

	Générateur	***Discriminateur***
Entrée	Un vecteur de bruit aléatoire	Une image réelle provenant des données d'entraînement ou une fausse image provenant du générateur
Sortie	Une fausse image	Probabilité que l'entrée soit une image réelle image

Perte	Classer les images fausses dans la catégorie des "vraies" images (à partir du discriminateur)	Classer les images réelles dans la catégorie "réelles" et les images fausses dans la catégorie "fausses"

Le Tableau 2-2 énumère les entrées et sorties du discriminateur et du générateur et la Figure 2-14 illustre une configuration typique d'un GAN.

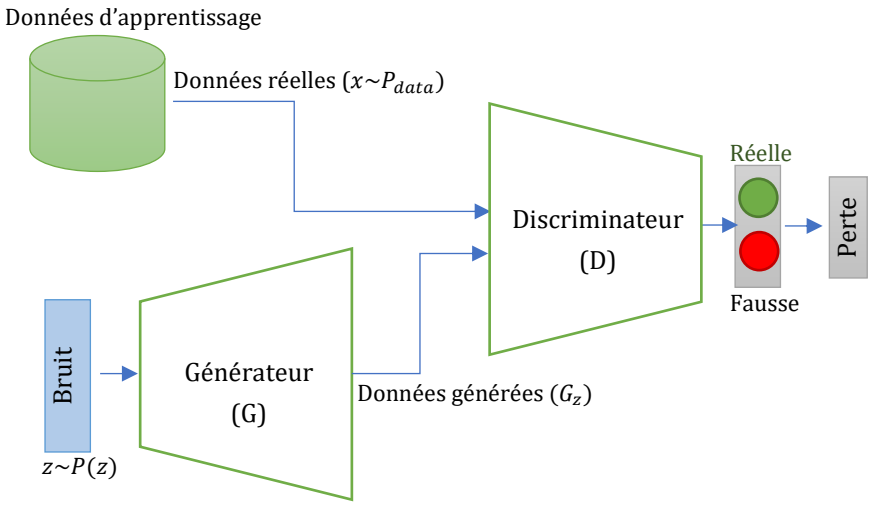


Figure 2-14 : Architecture typique d'un GAN

2.7.3.2 Fonctionnement et apprentissage d'un GAN

Goodfellow et ses collègues (I. Goodfellow et al., 2014) ont proposé un processus d'apprentissage contradictoire où ces deux composants sont simultanément formés. Le générateur et le discriminateur sont en concurrence, ils s'affrontent dans un jeu à somme nulle :

- Le générateur est entraîné à générer des images $G(z)$ qui ressemblent à la distribution des données d'entraînement P_{data}, en utilisant un vecteur de bruit latent z, échantillonné à partir de la distribution pz comme entrée,
- Quant au discriminateur, il reçoit les images générées $G(z)$ ainsi que les données d'entraînement réelles x comme entrée et est entraîné à distinguer les images générées des images réelles.

Ce processus se poursuit jusqu'à ce que les sorties du générateur soient proches des échantillons d'entrée réels. On peut considérer que le Discriminateur (D) et le Générateur (G) deux joueurs jouant le jeu min-max avec la fonction de $V\ (D, G)$ qui peut être exprimée de la manière suivante :

$$min_G max_D V(D,G) = E_{x \sim P_{data}(x)}[logD(x)] + E_{z \sim P(z)}[\log(1 - D(G(z)))] \qquad (2.15)$$

Le discriminateur idéal pour un générateur fixe, tel que présenté par Goodfellow et al. (I. Goodfellow et al., 2014), est donné par :

$$D_G^* = \frac{P_{data}}{P_{data} + P_G} \tag{2.16}$$

Où P_G est la distribution des données générées par le générateur. L'optimum du jeu minimax est atteint pour $P_{data} = P_G$, c'est-à-dire lorsque le générateur reproduit parfaitement la distribution des données d'apprentissage. Pour cet optimum, la sortie optimale du discriminateur est $D_G = \frac{1}{2}$, ce qui signifie que le discriminateur est incapable de différencier les distributions des données d'apprentissage et des données générées. Pour la mise en œuvre pratique des GAN, l'objectif minimax représenté dans l'équation 2.15 est optimisé en alternant l'optimisation de *G* et de *D*, en optimisant *G* une fois pour plusieurs étapes de *D*, afin de maintenir *D* dans sa région optimale. En pratique, cette équation peut ne pas fournir un gradient suffisant pour l'apprentissage de *G* (qui a commencé à partir d'un bruit gaussien aléatoire) aux premiers stades.

Dans les premiers stades, *D* peut rejeter des échantillons parce qu'ils sont clairement différents des échantillons d'apprentissage. Dans ce cas, $log(1 - D(G(z)))$ sera saturé. Au lieu d'entraîner G à minimiser $log(1 - D(G(z)))$, nous pouvons l'entraîner à maximiser la fonction objective $log(G(z))$ qui fournit de bien meilleurs gradients dans les premières étapes de l'apprentissage. Cependant, la première version présentait certaines limitations de convergence pendant l'apprentissage. En effet, les GAN ont au départ quelques limitations concernant les problèmes suivants (Alom et al., 2019):

- L'absence d'une fonction de coût heuristique (comme les erreurs carrées moyennes approximatives par pixel (MSE))
- Instabilité de leur apprentissage (qui peut parfois être due à la production de résultats absurdes).

La popularité et l'adoption généralisée des GAN dans la recherche s'expliquent en grande partie par le fait que les GAN peuvent être entraînés à l'aide de méthodes bien étudiées et comprises, également utilisées pour l'entraînement de modèles96++ discriminants, telles que la rétropropagation et la descente de gradient stochastique (SGD), sans nécessiter de calculs probabilistes complexes (I. Goodfellow et al., 2014). Les premiers exemples de GAN utilisaient des techniques d'apprentissage profond populaires, telles que Dropout (Srivastava et al., 2014b) et Rectied Linear Units (ReLUs) (Glorot et al., 2011), dans leur architecture, ce qui n'était possible que parce que le générateur et le discriminateur étaient des réseaux de neurones. L'optimisation entre le générateur et le discriminateur était très fragile, car l'un des réseaux pouvait parfois prendre le dessus sur l'autre, ce qui conduisait à une multitude d'états de défaillance possibles des GAN, par exemple l'effondrement de mode (mode collapse)(I. Goodfellow

et al., 2014). L'effondrement de mode décrit un problème courant pour les GANs, où le générateur apprend à faire correspondre plusieurs vecteurs d'entrée de bruit différents z à la même sortie G(z). Ceci est principalement dû au fait que les gradients du discriminateur sont calculés indépendamment les uns des autres, sans incorporer aucune sorte de mesure de similarité pour comparer les échantillons d'un minibatch donné (Salimans et al., 2016).

La recherche dans le domaine des GANs est en cours et de nombreuses versions améliorées ont été proposée (Salimans et al., 2016). Les GAN ont deux domaines de DL différents, semi-supervisé et non-supervisé. Certaines recherches dans ces domaines se concentrent sur la topologie de l'architecture du GAN afin d'améliorer la fonctionnalité et l'approche de formation.

2.7.3.3 Types de GAN

Diverses architectures dérivées de GAN ont été proposées afin d'améliorer les performances en termes de diversité et de qualité des données, ainsi que de stabilité de l'apprentissage. On distingue différents types de réseaux génératifs adversaires qui varient au gré de leur mise en œuvre. Les principaux types de GANs couramment utilisés sont les suivants :

A. GAN conditionnel (CGAN)

Comme l'entrée du générateur est le vecteur de bruit aléatoire z, ces entrées incontrôlées peuvent conduire à l'effondrement du mode de formation. Ainsi, Mirza et Osindero (Mirza & Osindero, 2014) ont proposé un réseau génératif adversarial conditionnel (CGAN), qui introduit la variable conditionnelle c (la variable c peut être des étiquettes, du texte ou d'autres données) dans le générateur et le discriminateur pour ajouter des conditions au modèle en utilisant des informations supplémentaires pour affecter le processus de génération de données(Pan et al., 2019). Dans la Figure 2-15, l'entrée du générateur est la variable conditionnelle c et le vecteur de bruit z, l'entrée du discriminateur est $G(z\backslash c)$ qui provient du générateur, et l'échantillon réel sous le contrôle de la même variable conditionnelle c. Par conséquent, la fonction objective peut être décrite comme suit :

$$min_G max_D V(D,G) = E_{x\sim P_{data}(x)}[logD(x\backslash c)] + E_{z\sim p(z)}\left[\log\left(1 - D\left(G(z\backslash c)\right)\right)\right] \quad (2.17)$$

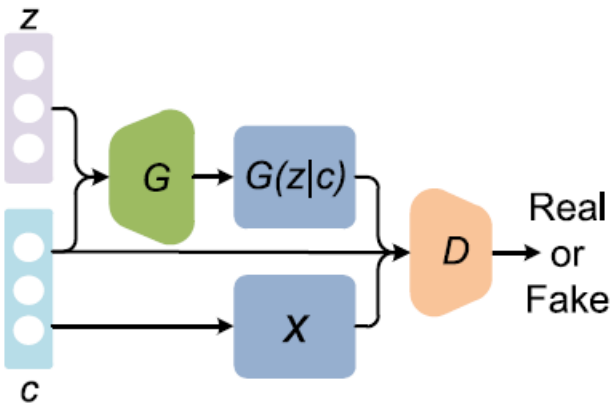


Figure 2-15 : Architecture du CGAN.

B. GAN à convolution profonde (DCGAN)

Suite au succès des GANs, Radford et al. ont (Radford et al., 2016) introduit le Deep Convolutional Generative Adversarial Network (DCGAN) en 2015, dans le but de mieux intégrer les méthodes récentes utilisées dans la formation des réseaux de neurones convolutifs (CNNs). Les DCGAN sont une famille d'architectures d'apprentissage non supervisé qui ont montré des résultats prometteurs par rapport à son homologue non supervisé, ils apprennent une hiérarchie de représentations dans les images, aussi bien les scènes elles-mêmes que les objets dans les scènes. L'architecture du DCGAN repose sur un ensemble de directives visant à améliorer la stabilité de la formation et la qualité des images (Radford et al., 2016):

- Remplacer toutes les couches de pooling par des convolutions stridées (discriminateur) et des convolutions stridées fractionnaires (générateur).
- Utiliser la normalisation par lots à la fois dans le générateur et le discriminateur.
- Supprimer les couches cachées entièrement connectées pour les architectures plus profondes.
- Utiliser l'activation ReLU dans le générateur pour toutes les couches sauf pour la sortie, qui utilise Tanh.
- Utiliser l'activation LeakyReLU dans le discriminateur pour toutes les couches.

C. Wasserstein GAN

Le modèle Wasserstein GAN (WGAN) présente deux avantages importants et majeurs par rapport au GAN traditionnel (Arjovsky et al., 2017). Premièrement, un WGAN établit une corrélation significative entre la métrique de perte et la convergence du générateur et la qualité de l'échantillon. Deuxièmement, les WGAN ont amélioré la stabilité du processus d'optimisation.

Gulrajani et al. (Gulrajani et al., 2017) ont constaté que le WGAN pouvait encore donner des résultats insatisfaisants ou ne pas converger en raison de l'utilisation de l'écrêtage des poids dans le discriminateur. Ils ont donc proposé une pénalité de

gradient appelée WGAN-GP pour faire respecter la contrainte de Lipschitz. La méthode présente également de meilleures performances que le WGAN original, et permet l'entraînement de diverses architectures de GANs plus stables qu'auparavant, presque sans réglage des hyperparamètres(Pan et al., 2019).

D. CycleGAN

CycleGAN (J.-Y. Zhu et al., 2017) a été proposé pour la traduction non supervisée d'image à image, afin d'apprendre la correspondance entre une image d'entrée et une image de sortie, lorsque des données d'apprentissage appariées ne sont pas disponibles. Il introduit une perte de cohérence de cycle afin d'appliquer l'exigence selon laquelle le mappage d'un domaine X à l'autre domaine Y est à peu près le même dans chaque direction(Z. Wang et al., 2021).

CycleGAN se compose de deux GAN distincts, l'un traduit une image d'un domaine à un autre (par exemple, cheval à zèbre) : $x_{trans} = G(x)$, l'autre effectue la traduction inverse (par exemple, zèbre à cheval) : $x = G_{inv}(x_{trans})$. Leur architecture de réseau suit celle de Johnson et al. (Johnson et al., 2016) qui s'est révélée efficace pour le transfert de style. Les deux GANs sont entraînés conjointement. La perte adverse utilise la fonction des moindres carrés au lieu d'une fonction logarithmique pour un entraînement plus stable. Outre les deux pertes adverses des deux GAN, la fonction objective comprend également une perte de cohérence de cycle L1, qui impose qu'une image se traduise par elle-même après le cycle de traduction $x \approx G_{inv}(G(x))$; $x_{trans} \approx G\big(G_{inv}(x_{trans})\big)$. Leur méthode a été appliquée avec succès à plusieurs tâches de traduction, notamment le transfert de style de collection, le transfert de saison, etc.

2.7.3.4 Les défis des GANs

Bien que les GAN aient connu un grand succès, ils doivent encore relever trois défis principaux (Mao & Li, 2021) :

1. Le premier défi est la qualité de l'image. De nombreuses études ont cherché à améliorer la qualité d'image des GANs (X. Huang et al., 2017; Jolicoeur-Martineau, 2018; Radford et al., 2016; H. Zhang et al., 2019). Radford et al. (2015) ont proposé les GANs convolutifs profonds (DCGANs), qui sont les premiers à introduire avec succès des couches convolutives dans les architectures de GANs. Denton et al. (2015) ont proposé la pyramide laplacienne des GANs (LAPGANs), dans laquelle une pyramide laplacienne est construite pour générer des images à haute résolution à partir d'images à basse résolution. Karras et al. (2017) ont proposé une nouvelle méthode de formation appelée formation progressive, qui génère d'abord des images réalistes à une résolution de 1024 × 1024.

   Brock et al. (2018) ont souligné que les GAN bénéficient considérablement de la mise à l'échelle de la taille du lot et du nombre de canaux dans chaque couche, et les

BigGAN proposés améliorent considérablement les performances sur des ensembles de données complexes tels qu'ImageNet (Russakovsky et al. 2015). Bien que les GAN puissent générer des images photoréalistes que même les humains ne peuvent pas distinguer des images réelles, ils sont limités à certains jeux de données simples dont les objets sont fortement " modélisés " et centrés avec de petites marges, comme les jeux de données de visages. Pour les jeux de données complexes, tels que les jeux de données de scènes, les performances des GAN restent limitées et les personnes peuvent facilement distinguer les images générées des images réelles.

2. Le deuxième défi est la stabilité de leur apprentissage. D'une manière générale, il est difficile d'entraîner les GAN dans la pratique en raison du problème de l'effondrement des modes(Radford et al., 2016). Certains travaux (Arjovsky et al., 2017; Nowozin et al., 2016) ont cherché à résoudre ce problème en analysant les fonctions objectives des GANs. Les techniques de régularisation sont également efficaces pour améliorer la stabilité de formation des GANs, comme la pénalité de gradient (Gulrajani et al., 2017) et la normalisation spectrale (Miyato et al., 2018). Notez que l'amélioration de la stabilité de l'apprentissage peut généralement conduire à des images générées de meilleure qualité.
3. Le troisième défi est l'évaluation des GANs. L'Inception Score (IS) (Salimans et al., 2016) et la Fréchet Inception Distance (FID) (Heusel et al., 2017) sont deux mesures d'évaluation largement utilisées pour les GAN. IS corrèle la qualité de l'image avec le degré auquel les images sont hautement classables en utilisant un classificateur pré-entraîné. FID modélise les caractéristiques des données générées et réelles comme des distributions gaussiennes multivariées continues et utilise la distance de Fréchet pour mesurer la distance entre les données générées et réelles. Bien que l'IS et le FID soient largement utilisés, des questions subsistent, comme l'utilisation de réseaux pré-entraînés et les approximations des distributions gaussiennes (Borji, 2019).

2.7.3.5 Domaines d'applications de GAN

Les GAN ont été appliqués à divers domaines tels que

- la vision par ordinateur (L. Ma et al., 2017; H. Wu et al., 2019),
- le traitement du langage naturel (Fedus et al., 2018; Nie et al., 2022),
- la synthèse de séries temporelles (Brophy et al., 2019; D. Li et al., 2019),
- la super-résolution (Ledig et al., 2017; You et al., 2020),
- la segmentation sémantique (Dong et al., 2017; Vo & Sugimoto, 2018),
- le traitement de la parole et de l'audio (Pascual et al., 2017; Serban et al., 2016; L.-C. Yang et al., 2017),
- Le traitement de l'information médicale
- Les jeux vidéo
- etc.

2.8 Conclusion

Les techniques d'apprentissage profond ont suscité un intérêt croissant de la part des chercheurs en raison de leur capacité inhérente à surmonter l'inconvénient des algorithmes traditionnels de machine learning, dépendant de caractéristiques conçues à la main. Les approches d'apprentissage profond se sont également avérées adaptées à l'analyse des données volumineuses, avec des applications réussies en vision par ordinateur, en reconnaissance des formes, en reconnaissance vocale, en traitement du langage naturel et en systèmes de recommandation.

Dans ce chapitre, nous avons abordé certaines architectures d'apprentissage profond largement utilisées et leurs applications pratiques. Une vue d'ensemble actualisée est fournie sur trois architectures d'apprentissage profond, à savoir le réseau neuronal convolutif, l'auto-encodeur et les réseaux adversaires génératifs. Les différents types de réseaux neuronaux profonds sont passés en revue et les avancées récentes sont résumées.

Chapitre 3 : Les techniques de l'animation faciale

3.1 Introduction

Les animations faciales par ordinateur sont principalement un domaine de l'infographie qui couvre les techniques de création et d'animation du visage d'un personnage. La génération d'animations faciales a toujours constitué un important challenge pour les chercheurs travaillant dans le domaine de la visualisation graphique. De nombreuses recherches ont été menées pour parvenir à un haut niveau de réalisme dans l'animation des visages. Malheureusement, la complexité de l'anatomie faciale humaine et notre sensibilité à la façon dont le visage apparaît naturellement ne permettent pas à un système de capturer avec réalisme les expressions et les émotions subtiles d'un avatar en temps réel.

Dans ce qui suit, nous dressons une synthèse des différentes techniques utilisées dans l'animation du visage permettant d'obtenir une animation faciale réaliste. Nous examinons les techniques de modélisation du visage de différents points de vue ; les manipulations géométriques connexes (qui peuvent être classées en interpolations, paramétrisation, modèle basé sur les muscles et pseudo-muscules) et les techniques d'animation faciale basées sur la parole, l'image et la capture de données.

3.2 Quelques concepts et notions de base

3.2.1 Animation

Le terme d'animation est très vaste et recouvre un grand nombre de techniques appliquées à différents niveaux. C'est la production d'images consécutives qui, lorsqu'elles sont visualisées, donnent une impression de mouvement.

Traditionnellement, l'animation était créée en dessinant des images des personnages pour chaque image de l'action. Au début de la production, l'animateur reçoit des storyboards, qui sont des croquis décrivant la séquence des principales actions et illustrant les expressions des personnages. L'animateur travaille également à partir d'une bande sonore finalisée, qui détermine le timing de la pièce. La plupart des animations manuelles ont été créées à l'aide d'images clés : un animateur principal crée les images clés, ou les plus importantes, et un second animateur crée les images intermédiaires. Le défi pour l'animateur, indépendamment des moyens utilisés, est de

créer des images qui donnent de l'expressivité et de la vie aux personnages (Hodgins & O'Brien, 2003).

Avec les progrès de l'infographie et de l'intelligence artificielle, il est désormais possible de créer des personnages virtuels de plus en plus réalistes et de reproduire la communication homme-machine de manière encore plus fine, cette évolution ouvre de nouvelles possibilités aux animateurs.

3.2.2 Visage

Le visage est la surface externe de la partie antérieure de la tête humaine, également appelée face ou figure. Différent d'une population à l'autre, le visage humain est le trait le plus distinctif permettant d'identifier et de reconnaître les autres. La peau et les muscles faciaux cachent un complexe squelettique de 14 os distincts qui abritent des parties des systèmes digestif, respiratoire, visuel et olfactif. Ce trait joue un rôle primordial dans les interactions sociales et la communication(Lacruz et al., 2019), à deux niveaux :

- Au niveau de la communication verbale, grâce à la bouche et à la mâchoire qui jouent un rôle important ;
- Au niveau de la communication non verbale, réalisée grâce notamment aux multiples déformations du visage.

Un système biomécanique complexe, composé de trois structures : la peau, les os et les muscles ainsi que d'autres organes tels que les yeux, la langue, les oreilles, etc. rendent possible ces deux formes de communication(Dutreve, 2011).

3.2.3 Animation du visage

L'animation de visage vise à synthétiser automatiquement des images de visage continues à partir d'une seule image source, pilotée par un ensemble de mouvements de visage. Elle désigne donc les méthodes permettant d'animer le visage en fonction des mouvements d'un être humain réel(Shakir & Al-Azza, 2022). En raison de divers facteurs, ce processus est considéré comme l'une des tâches les plus difficiles dans le domaine de l'animation. En effet, la plupart des gens étant confrontés quotidiennement à de nombreuses interactions humaines naturelles, les humains sont capables de reconnaître les activités faciales non naturelles. Ainsi, le moindre changement dans un visage animé attire directement l'attention de l'observateur et l'animation perd son réalisme. Il convient de noter que le visage humain est un système complexe composé d'un grand nombre de muscles qui doivent être habilement coordonnés pour être considérés comme réels. Un autre facteur qui contribue à la difficulté de modéliser et d'animer le visage humain est sa variété ; en effet, les caractéristiques du visage varient d'une personne à l'autre, en raison de structures osseuses et de proportions musculaires différentes.

Les applications de l'animation faciale sont largement répandues dans l'industrie du divertissement : les films, les jeux vidéo et les avatars interactifs. L'animation faciale trouve d'autres applications dans divers domaines tels que les sciences médicales et la robotique.

3.2.4 Expression faciale

Les expressions faciales sont considérées comme l'une des formes de communication humaine les plus marquantes et les plus influentes, qu'il s'agisse de l'écrit, de la parole, du langage corporel, etc.

Le terme "expression faciale" est réservé par les chercheurs aux configurations récurrentes des mouvements des muscles faciaux qui communiquent une pensée, une émotion ou un comportement(Frank, 2001).

Sur le plan anatomique, une expression faciale est issue de la déformation des traits du visage engendrée par une émotion. La déformation des principaux traits permanents du visage, à savoir les yeux, le nez, les sourcils et la bouche, contient l'information essentielle d'une expression(Khalfi, 2010).

Les expressions faciales peuvent également être définies comme un système universel de signaux qui reflètent les fluctuations momentanées de l'état émotionnel d'une personne. Une expression faciale est issue de la contraction de certains muscles et du relâchement d'autres.

3.2.5 Emotion

Philippe Claudon et Margot Weber donnent la définition suivante de l'émotion, dans leur article (Claudon & Weber, 2009):

« L'émotion est un état de conscience complexe, généralement brusque et momentané, accompagné de signes physiologiques (par exemple : rougissement, sudation) »

En psychologie, l'émotion est souvent définie comme un état complexe de sentiment qui entraîne des changements physiques et psychologiques qui influencent la pensée et le comportement. Selon l'auteur David G. Myers(Myers, 2004), l'émotion humaine implique "...une excitation physiologique, des comportements expressifs et une expérience consciente."

Les travaux de recherche se concentrent sur la reconnaissance des émotions à l'aide de différentes modalités, le plus souvent à partir de l'analyse des expressions faciales.

Sur la base d'une étude interculturelle, Ekman et al.(Ekman & Friesen, 1971) ont défini six expressions émotionnelles de base : le dégoût, la colère, la peur, le bonheur, la tristesse et la surprise. Ces expressions étant universelles chez les êtres humains, ils démontrent que certaines émotions de base sont perçues de la même manière chez les

êtres humains, indépendamment de leur culture. Bien que de récentes recherches avancées en neuroscience et en psychologie aient indiqué que le modèle des six émotions de base n'est pas universel(Jack et al., 2013), mais spécifique à une culture, la plupart des études dans le domaine de la reconnaissance des expressions faciales se concentrent sur ce modèle.

3.2.6 Reciblage du visage (*facial retargeting*)

Le reciblage facial est un processus permettant de transférer l'animation faciale d'un modèle source (un visage réel) à un modèle cible (un visage ou un objet virtuel) tout en préservant la signification sémantique des expressions faciales(Ribera et al., 2017).

L'objectif est de réduire le temps et les ressources nécessaires par la réutilisation de l'animation déjà existante d'un visage donné sur un autre visage. Ce transfert d'animation n'est pas évident et nécessite la mise en œuvre de techniques particulières, compte tenu des différences de morphologie entre les différents visages.

3.2.7 Reconstitution de visage (*reenactment*)

La reconstitution de visage est une tâche de synthèse de visage dans laquelle les expressions faciales et la pose d'un visage source sont transférées sur un visage cible tout en préservant l'apparence et les détails du visage cible. Les systèmes de reconstitution de visages ont des applications pratiques dans les jeux multi-joueurs (AR/VR) ainsi que dans l'industrie du divertissement.

Le développement d'un système de reconstitution de visage réaliste requiert un grand ensemble de données pour une identité donnée (par exemple, un grand nombre de vidéos d'une seule personne). Ces exigences en matière de données limitent les applications du système.

3.2.8 Clonage d'expression

Le problème du clonage d'expression a été posé dans Noh et Neumann (2001), où la solution a été donnée comme une mise en correspondance en trouvant des paires de points correspondants sur les visages source et cible en utilisant des heuristiques spécifiques aux visages(Dhere et al., 2020). Les premiers algorithmes de clonage d'expression ne tiennent pas compte de l'adaptation de la dynamique temporelle du mouvement à la cible, ce qui signifie que leur efficacité n'est prouvée que si la source et la cible sont de proportions similaires.

L'approche proposée par Noh et al. (Noh & Neumann, 2001) consiste à produire des animations faciales par réutilisation des données. Le clonage d'expression (Expression Cloning EC) réutilise les vecteurs de mouvement des sommets du modèle source pour créer des animations similaires sur un nouveau modèle cible.

3.3 Les Techniques conventionnelles d'animation faciale

L'animation d'une image fixe d'un objet ou d'une scène de manière contrôlable et efficace permet de nombreuses applications intéressantes dans le domaine de l'édition/amélioration d'images, de la post-production de films et de l'interaction homme-machine. Dans ce domaine, une approche fréquente consiste à modéliser des visages paramétriques en 3D, puis à les animer en 3D. Ces approches nécessitent généralement un post-traitement par des techniques d'infographie, un équipement coûteux et une importante quantité de travail pour produire des résultats réalistes. Les chercheurs s'intéressent actuellement à des méthodes entièrement automatiques basées sur la 2D et utilisant des techniques d'apprentissage automatique afin de réduire le coût et le temps nécessaires à la production d'animations de haute qualité. Ces dernières années, les méthodes basées sur la 2D ont accompli des avancées considérables dans l'animation de visages/têtes, d'objets humains, de personnages de dessins animés, d'autres objets et de scènes(Yi et al., 2020).

Une classification rigoureuse de tous les travaux publiés est une tâche délicate(Dutreve, 2011). En effet, la frontière entre les différentes techniques est parfois floue, tandis que certaines méthodes intègrent souvent plusieurs approches pour produire de meilleurs résultats(Shakir & Al-Azza, 2022), ou encore que plusieurs recherches se focalisent sur une question bien spécifique de l'animation faciale, alors que d'autres tentent de mettre en place un système complet, etc.

Parmi ces approches classiques, nous citons l'interpolation ou du mélange de formes, de l'animation guidée par les performances, des modèles paramétriques et de l'animation guidée par les pseudo-muscles, les muscles et le langage.

De notre point de vue, nous classons les approches traitants l'animation faciale en deux catégories, les approches conventionnelles, qui sont les méthodes classiques basées sur les transformations géométriques, et les approches à base des techniques de l'apprentissage profond, qui sont des méthodes actuelles, plus utilisées, et se basées sur les manipulations de visage.

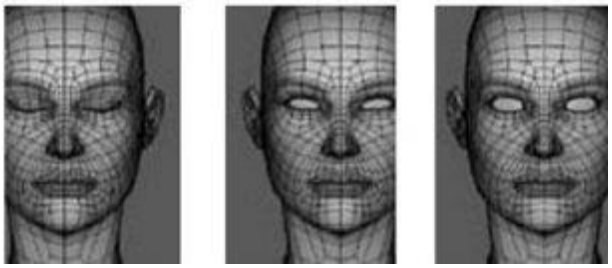
Dans ce qui suit, nous détaillons les principales approches conventionnelles, qui sont l'interpolation, la paramétrisation, la modélisation physique des muscles, les muscles simulés, ainsi que l'animation faciale basée sur les données, tandis que les approches basées sur l'apprentissage profond seront détaillées dans la section suivante.

3.3.1 Interpolation de formes (*blendshape*)

Les techniques de mélange de formes sont une approche standard pour réaliser des animations faciales expressives dans le domaine de la production numérique. Cette approche consiste à créer des images clés (appelés frames), puis à effectuer une interpolation entre elles pour obtenir une déformation continue du visage dans le temps.

Ce concept a été introduit par F. Parke en 1972 (Parke, 1972), où la surface d'un visage a été approximée par une peau polygonale contenant environ 250 polygones définis par environ 400 sommets. Afin de produire un mouvement facial réaliste, l'animation a été réalisée en utilisant l'interpolation en cosinus pour remplir les images intermédiaires et obtenir des transitions douces entre les différentes expressions faciales. En effet, le visage humain étant régi par la physique, le mouvement ne pouvait pas être linéaire et devait subir des accélérations et des décélérations.

D'autre fonctions d'interpolation ont été utilisé dans la littérature, tel que l'interpolation bilinéaire (Arai et al., 1996), L'interpolation B-Spline a été utilisée pour créer un effet d'animation plus rapide au début et plus lent à la fin de l'animation (Reeves, 1981), l'interpolation de Bézier(Lu et al., 2007).



yeux fermés image interpolée yeux ouverts

Figure 3-1: Interpolation linéaire effectuée sur les yeux, de fermés à ouverts (Parke, 1972).

Le principe des Blendshape est de générer un espace linéaire d'expressions faciales en interpolant les formes de base pour en produire une nouvelle, comme le montre la Figure 3-1. La définition d'une expression est alors une combinaison linéaire de n vecteurs, qui définissent chacun une forme. Pour un modèle à maillage, il s'agit des coordonnées de chacun de ses sommets sachant que la topologie entre chaque forme doit être identique. Pour un modèle spline, il s'agit de la position des points de contrôle(Dutreve, 2011).

On considère le maillage de visage comme un vecteur colonne *f* contenant toutes les coordonnées des sommets $\overrightarrow{v_j}$ (la position du $j^{ème}$ sommet du maillage). Si nous considérons le modèle de visage composé de *n* blendshapes, la position du sommet $\overrightarrow{v_j}$ est définie comme suit :

$$\overrightarrow{v_j} = \sum_{i=1}^{n} w_i \overrightarrow{v_{ji}} \qquad (3.1)$$

Avec *w* étant le poids appliqué à la forme *i*, et $\overrightarrow{v_{ji}}$ représente la position du même sommet pour la forme *i*.

Cette technique est adaptée à la création de séquences de courte durée. Cependant, elle n'est pas nécessairement appropriée pour créer des séquences en temps réel (en

fonction de la complexité des modèles utilisés) permettant l'interactivité avec un modèle virtuel.

3.3.2 Paramétrisation

Pour surmonter certaines des limitations et restrictions des interpolations simples, la paramétrisation est proposée(M. M. Cohen & Massaro, 1993a; Parke, 1982). Une paramétrisation idéale définit le visage ou l'expression par une combinaison de valeurs de paramètres indépendants (ensembles de paramètres ou points de contrôle). Les combinaisons de paramètres fournissent une large gamme d'expressions faciales avec des coûts de calcul relativement faibles.

Afin de faciliter le processus de paramétrage, des systèmes de codages sont établis pour représenter de manière paramétrique les expressions ou les animations faciales(Ping et al., 2013). L'un des plus utilisés de ces systèmes provient de travaux dans le domaine de la psychologie(Ekman et al., 2002; Ekman & Friesen, 1978) et n'était pas initialement conçu pour être utilisé en infographie. En revanche, d'autres ont été proposés spécifiquement pour cet usage, notamment la partie dédiée à l'animation faciale de la norme MPEG-4(ISO, 2001), ou encore les systèmes MPA(Kalra et al., 1992) ou AMA (Magnenat-Thalmann et al., 1988).

Contrairement aux techniques d'interpolation, les paramétrisations permettent le contrôle explicite de configurations faciales spécifiques. Toutefois, elle présente également des limites, malgré ses avantages ; comme il n'existe aucun moyen systématique d'arbitrer entre deux paramètres conflictuels pour mélanger des expressions qui affectent les mêmes sommets, la paramétrisation produit rarement des expressions ou des configurations humaines naturelles lorsqu'un conflit entre paramètres se produit(Waters & Frisbie, 1995). Par conséquent, les paramétrisations sont conçues pour n'affecter que des régions spécifiques du visage, mais cela introduit souvent des limitations de mouvement importantes(Noh & Neumann, 1998b). Une autre limite de la paramétrisation est que le choix des paramètres dépend de la topologie du maillage facial et, par conséquent, une paramétrisation générique complète n'est pas possible. L'animation faciale peut être paramétrée de manière relativement intuitive pour les modèles faciaux simples car ils comportent à peine une centaine de sommets(Ping et al., 2013). Cependant, pour un modèle facial complexe qui produit un très grand nombre de sommets, la paramétrisation n'est pas pratique. Par ailleurs, il faut un processus de réglage manuel fastidieux et complexe pour adapter une paramétrisation existant à un nouveau modèle facial.

Les limites de la paramétrisation ont conduit au développement de diverses techniques telles que la modélisation basée sur les muscles, ainsi que les techniques d'animation faciale basées sur les données.

3.3.3 La Modélisation physique des muscles

D'autre chercheurs se sont penchés sur l'approche de modélisation physique qui permet de représenter le visage humain, et de simuler ses déformations réelles. Les modèles musculaires basés sur la physique se répartissent en trois catégories : les systèmes masse-ressort, les représentations vectorielles et les maillages de ressorts en couches(Noh & Neumann, 1998b).

3.3.3.1 Les systèmes masse-ressort

Depuis les années 1980, Platt et Badler (Platt & Badler, 1981) ont été les pionniers de ces modèles. Leur modèle masse-ressort proposé relie les nœuds de la peau, du muscle et de l'os, et propage ainsi les forces musculaires dans le maillage élastique du ressort qui modélise la déformation de la peau. Ces forces appliquées à travers les arcs musculaires génèrent diverses expressions faciales. Quelques exemples d'expressions faciales générées par ce modèle sont représentés dans la Figure 3-2.

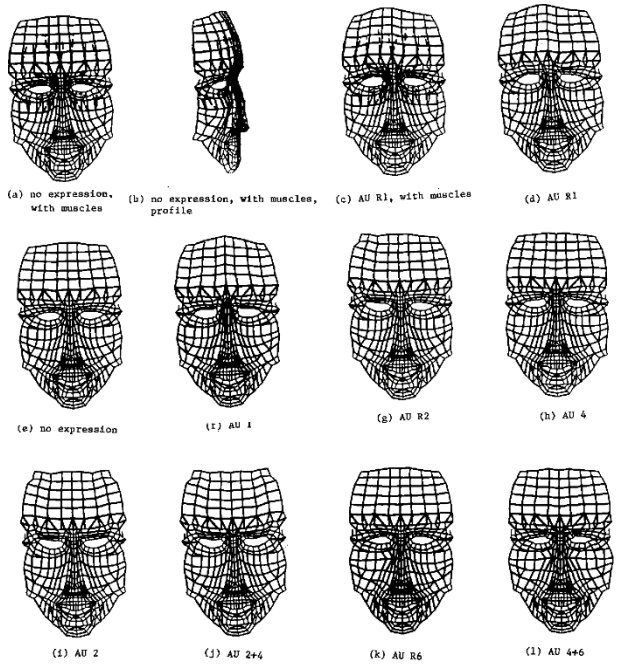


Figure 3-2: Modèle masse-ressort de Platt (Platt, 1985), exemples d'expressions faciales.

Les travaux ultérieurs de Platt (Platt, 1985) ont présenté un modèle facial avec des muscles représentés comme des collections de blocs fonctionnels dans des régions définies de la structure faciale. Le modèle de Platt est constitué de 38 blocs musculaires régionaux interconnectés par un réseau de ressorts. Les unités d'action sont créées en appliquant des forces musculaires pour déformer le réseau de ressorts.

3.3.3.2 Les représentations vectorielles

Un modèle musculaire vectoriel très réussi et plus développé a été proposé par Waters (Waters, 1987) à la fin des années 1980 et a constitué les principes de base du modèle basé sur la physique. L'approche vectorielle déforme un maillage facial à l'aide

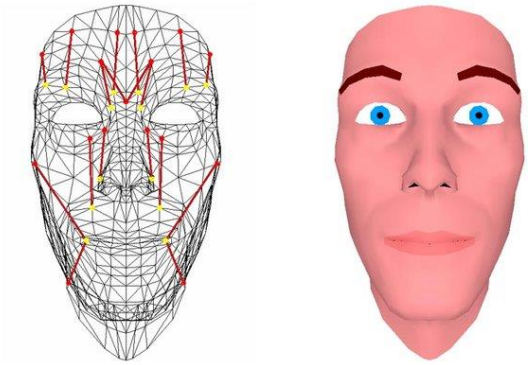


Figure 3-3: Modèle de muscle de Waters (Waters, 1987).

L'image de gauche montre le maillage du modèle et les vecteurs du muscle. Les points rouges indiquent les points d'origine et les points jaunes les points d'insertion. L'image de droite montre le modèle entièrement ombré.

de champs de mouvement (également appelés champs de déformation) dans des régions d'influence délimitées. Waters avait modélisé les muscles par des vecteurs, où la direction du champ vectoriel, une origine et un point d'insertion définissent chaque vecteur. La Figure 3-3 illustre le modèle de muscles proposé par Waters.

Il a également utilisé FACS (Facial action coding system, désigne un ensemble de mouvements des muscles du visage qui correspondent à une émotion affichée, il a été proposé en 1978 (Ekman & Friesen, 1978), mais a été considérablement mis à jour en 2002 (Ekman et al., 2002)) pour relier les expressions faciales à l'activité musculaire, indépendamment de la morphologie du visage. Et cela lui a permis d'animer des émotions humaines telles que la colère, la peur, la surprise, le dégoût, la joie et le bonheur. Dans le court-métrage Tin Toy (Pixar), ils ont utilisé 47 muscles inspirés de la méthode de Waters pour animer le visage de bébé Billy(Noh & Neumann, 1998b).

3.3.3.3 Les maillages de ressorts en couches

Enfin, dans la troisième catégorie, le maillage élastique en couches est une structure masse-ressort en trois couches de maillage connectées pour modéliser plus fidèlement le comportement anatomique du visage.

Dans les travaux de Terzopoulos et Waters (Terzopoulos & Waters, 1990), les trois couches déformables du maillage correspondent à la peau, au tissu graisseux et aux muscles liés aux os. Des éléments élastiques à ressort relient chaque nœud de maillage et chaque couche, et les forces musculaires se propagent dans les systèmes de maillage pour créer une animation (Figure 3-4). Ces travaux ont été amélioré par d'autres chercheurs, en introduisant un modèle musculaire basé sur la physique(Y. Lee et al., 1993; W. Wang et al., 2009) et un système basé sur les muscles NURBS(Tang et al., 2003, 2004). Par la suite, d'autres travaux ont été réalisés sur la base d'une modélisation similaire de la peau (Kähler et al., 2001; Y. Lee et al., 1995; Yu Zhang et al., 2001).

Bien que ce modèle soit fidèle à l'anatomie du visage et atteigne un grand réalisme, des calculs importants sont nécessaires pour simuler les déformations volumétriques à l'aide de maillages tridimensionnels. La génération précise d'un modèle musculaire pendant différentes activités en temps réel reste également un des problèmes rencontrés dans cette approche.

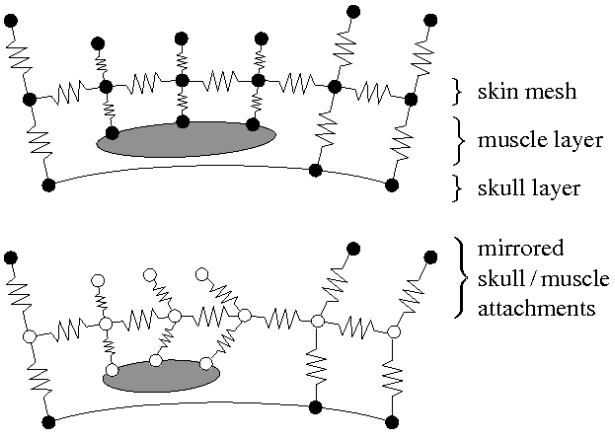


Figure 3-4: Système masse-ressort proposé par Terzopoulos et Waters (Terzopoulos & Waters, 1990).

En haut : muscle détendu, ressorts extérieurs reflétant les attaches du crâne et des muscles.
En bas : Muscle contracté, les points de masse se déplaçant en raison de la contraction marquée par

3.3.4 Animation par Pseudo-muscle ou muscle simulé

Les pseudo-muscles ou muscles simulés produisent souvent des animations de qualité en déformant le maillage du visage comme un muscle, tout en ignorant l'anatomie sous-jacente complexe et les structures musculaires réelles. En revanche, la création de diverses expressions faciales par la manipulation du maillage fin du visage est privilégiée. Son principe consiste, donc, à reproduire les effets visuels du visage, et non sa mécanique. Cette catégorie englobe le morphing entre différents modèles et les pseudo-muscles simulés en forme de Splines, ou encore les déformations de forme libre.

3.3.4.1 Le Morphing

Le morphing est l'une des techniques d'animation appliquées pour la 2D et la 3D dans cette approche. Il s'agit du processus de génération d'une transition animée fluide d'une image à une autre, ou d'un modèle à un autre. Il consiste en une déformation entre des points correspondants dans les images cibles et une dissolution croisée simultanée. En général, les correspondances sont sélectionnées manuellement pour répondre aux besoins de l'application.

Le morphing comporte généralement deux étapes : tout d'abord, un animateur humain fait correspondre des caractéristiques (un ensemble de points ou de segments de ligne correspondants) dans une paire ou un ensemble d'images, Ensuite, Beier et al. utilisent un algorithme pour déterminer la correspondance (mapping) pour le reste des images (Beier & Neely, 1992). La deuxième étape du processus consiste à utiliser le mappage pour interpoler la forme de chaque image à l'autre, en fonction de l'image

intermédiaire particulière à synthétiser, et à mélanger les valeurs des pixels des deux images déformées par les mêmes coefficients respectifs, achevant ainsi le morphing(S. E. Chen & Williams, 1993).

Noh et al.(Noh & Neumann, 1998c) ont constaté que le réalisme, avec cette approche, nécessite une interaction manuelle importante pour l'équilibrage des couleurs, la sélection des correspondances et le réglage des paramètres de morphing. Les variations des points de vue ou des caractéristiques de l'image cible compliquent la sélection des correspondances. Il est difficile de synthétiser des mouvements de tête réalistes car les caractéristiques de la cible sont masquées ou révélées pendant l'animation.

Pighin et al. (Pighin et al., 1997) ont tenté de surmonter les limites du morphing 2D en le combinant avec des transformations 3D d'un modèle géométrique. Ils ont animé des expressions faciales clés par interpolation géométrique 3D, tandis que le morphing d'image est effectué entre les cartes de texture correspondantes. Bien que cette approche atteigne un réalisme indépendant du point de vue, l'interpolation entre les expressions clés prédéfinies reste la limite de l'animation.

Si les méthodes de morphing 2D et 3D parviennent à produire des expressions faciales de qualité, elles présentent les mêmes limites que les approches d'interpolation. En effet, la sélection des points correspondants dans les images cibles nécessite un travail manuel conséquent, elle dépend du point de vue et elle ne peut être généralisée à différents visages. De plus, le point de vue de l'animation est contraint de correspondre approximativement à celui des images cibles.

Il est à noter que bien que dans certaines recherches il y ait une confusion entre la notion de morphing et la notion d'interpolation de formes (Blenshapes), nous retenons que le morphing permet de transformer de manière fluide un visage en un autre complètement différent, alors que l'interpolation de formes permet d'effectuer des changements sur un visage animé, sa taille, sa couleur et son positionnement.

3.3.4.2 Déformation de forme libre

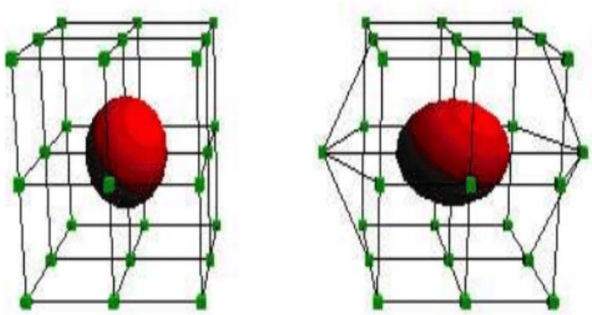


Figure 3-5: Déformation de forme libre tridimensionnel (Sederberg & Parry, 1986).

La déformation de forme libre (Free Form Deformation FFD) permet de générer des déformations faciales en déformant des objets volumétriques par la manipulation de points de contrôle disposés dans un réseau cubique tridimensionnel (Sederberg & Parry, 1986). Conceptuellement, l'objet ou la surface à animer est enveloppé dans une boîte de contrôle imaginaire, transparente et flexible, contenant une grille 3D de points de contrôle, et pour déformer cette boîte ; toute déformation faite à la boîte de contrôle (écrasement, pliage ou torsion en formes arbitraires), est ensuite appliquée à l'objet enveloppé en conséquence, comme le montre la Figure 3-5. La déformation ne se fait pas directement sur l'objet, mais en déplaçant les points de contrôle(Dutreve, 2011).

Les FFD peuvent déformer de nombreux types de primitives de surface, notamment des polygones, des surfaces quadriques, paramétriques et implicites, ainsi que des modèles solides. Des extensions de cette méthode ont été proposé pour la déformation de surfaces géométriques, tel que EFFD (Extented FFD)(Coquillart, 1990), RFFD (Rational FFD)(Kalra et al., 1992), et elles sont disponibles dans différents logiciels tel que Blender, Maya et 3ds Max. Lorsque l'on utilise le FFD (EFFD, RFFD) pour abstraire le contrôle de la déformation de celui de la description de la surface réelle, l'un des principaux avantages réside dans le fait que la transition de la forme ne dépend plus des spécificités de la surface elle-même.

Par rapport au modèle physique de Waters(Waters, 1987), la manipulation des positions ou des poids des points de contrôle est plus intuitive et plus simple que la manipulation des vecteurs musculaires avec une zone d'influence délimitée. Cependant, la FFD (EFFD, RFFD) ne fournit pas une simulation précise du comportement réel des muscles et de la peau. De plus, comme la FFD (EFFD, RFFD) est basée sur la déformation de surface, les changements volumétriques qui se produisent dans le muscle physique ne sont pas pris en compte (Noh & Neumann, 1998b) ; ces approches sont incapables de modéliser les bosses et les rides de la peau, car aucune simulation particulière n'est effectuée sur les muscles et la peau réels.

3.3.4.3 Pseudo-muscles à base de Splines

Les modèles polygonaux du visage sont largement utilisés, mais ils ne parviennent souvent pas à reproduire de manière adéquate la douceur ou la flexibilité du visage humain. Idéalement, la représentation de la surface d'un modèle facial doit permettre des déformations lisses et flexibles. Une solution plausible est proposée par les modèles musculaires splines. En effet, les splines sont généralement continues jusqu'à C2, garantissant ainsi un patch de surface lisse, et permettant une déformation localisée sur la surface. En outre, les transformations affines sont définies par la transformation d'un petit ensemble de points de contrôle au lieu de tous les sommets du maillage, réduisant ainsi la complexité du calcul. Des exemples d'animations basées sur les splines sont présentés dans (Hoch et al., 1994; Nahas et al., 1990; Viaud & Yahia, 1992). Cette

technique est principalement adaptée à la modélisation des plis aigus sur une surface ou des discontinuités entre les surfaces.

D'autres variantes ont été proposées dans la littérature : par exemple, pour surmonter l'inconvénient des B-splines classiques, Eisert et Girod (Eisert & Girod, 1998) ont utilisé des B-splines triangulaires qui n'affinent pas localement les zones courbées puisqu'elles sont définies sur une topologie rectangulaire. Cependant, le recours à un modèle spline hiérarchique permet de réduire le nombre de points de contrôle inutiles. Wang et al [125] ont présenté un système qui intègre des modèles splines hiérarchiques avec des muscles simulés basés sur les déformations locales de la surface.

Néanmoins, l'inconvénient de l'utilisation de B-splines naïves pour les surfaces complexes devient évident lorsqu'une déformation doit être plus fine que la résolution du patch(Noh & Neumann, 1998b). Pour produire une résolution de patch plus fine, une ligne ou une colonne entière de la surface est subdivisée. Ainsi, plus de détails (et de points de contrôle) sont ajoutés là où ils ne sont pas nécessaires.

3.3.5 Animation faciale basée sur les données

Les techniques basées sur les données sont largement utilisées au cours de ces dernières années, car elles présentent une animation plus riche, plus détaillée et plus précise à partir des données empiriques collectées. Il s'agit d'animer des visages virtuels à partir de données réelles capturées. Ces techniques sont souvent utilisées dans les interfaces homme-machine (agents virtuels), et dans la communication anonymisée entre humains en temps réel par le biais de ces avatars. Cette approche comporte trois catégories : les techniques basées sur l'image, les techniques basées sur la parole et l'animation basée sur la performance.

3.3.5.1 Les techniques basées sur l'image

Afin de produire un modèle de visage humain photoréaliste, des techniques basées sur l'image sont développées, ce qui aurait été irréalisable en adaptant uniquement des interpolations de formes et des déformateurs de muscles(Ping et al., 2013). Cette technique consiste à capturer les données de surface et de position du visage à partir d'images provenant de différentes vues afin de reconstruire le modèle du visage. Le calcul de la profondeur du modèle peut être effectué en appliquant la triangulation sur deux images. Pour déterminer l'efficacité du système, il faut tenir compte du nombre et de la vue des images d'entrée et de la connaissance de la géométrie de la scène.

Les techniques basées sur l'image ont été utilisées comme effets spéciaux dans certains films à grand succès comme « The Matrix Reloaded » (Borshukov et al., 2005)et « The Lord of the Rings »(Lewis et al., 2005). Certains chercheurs (Deng et al., 2006; Lewis et al., 2005) proposent de combiner cette technique avec d'autres techniques géométriques pour développer une meilleure déformation du modèle facial. Les

animateurs ne cessent d'utiliser la technique de l'image plutôt que le modèle facial par balayage laser 3D, car elle est peu coûteuse et produit un modèle facial réaliste. La technique basée sur l'image doit être améliorée pour maintenir son utilisation à l'avenir, ses contraintes doivent être traitées, par exemple comment automatiser la représentation du modèle facial 3D en utilisant certains paramètres pendant l'animation(Ping et al., 2013).

Une des premières recherches utilisant cette technique est le travail de Pighin et al. (Pighin et al., 1998) qui ont adapté des maillages de modèles tridimensionnels à une certaine pose du visage et ont mélangé les différentes poses afin d'obtenir une animation, en utilisant seulement quelques photographies. Un an plus tard, en 1999, Pighin a utilisé cette approche afin de mettre en évidence les problèmes et les enjeux de la modélisation et de l'animation de visages très réalistes à partir d'images(Pighin, 1999).

3.3.5.2 Les techniques basées sur la parole

La génération d'une animation réaliste et visuelle de la parole correspondant aux paroles d'entrée est un autre enjeu de l'animation faciale. Cette technique guidée par la parole consiste à synchroniser le discours d'entrée avec l'animation faciale. En parlant, l'intonation et l'état émotionnel utilisés affectent l'expression.

Synthétiser des animations faciales réalistes à partir d'un nouveau texte ou d'une entrée vocale acoustique préenregistrée est une tâche difficile depuis des décennies, car les langages humains, comme l'anglais, ont non seulement un grand vocabulaire et un grand nombre de phonèmes (le phonème est l'unité de parole standard), mais aussi le phénomène de co-articulation de la parole qui complique les correspondances entre les signaux vocaux acoustiques et les mouvements vocaux visuels(Deng & Noh, 2008). L'expression "co-articulation de la parole" est définie dans la littérature linguistique de la manière suivante : les phonèmes ne sont pas prononcés comme une séquence indépendante de sons, mais le son d'un phonème particulier est affecté par les phonèmes adjacents. La co-articulation visuelle de la parole est analogue.

La méthode de synthèse des mouvements des lèvres est la méthode de base utilisée pour l'animation faciale, et afin de produire les formes des lèvres souhaitées, les recherches antérieures sur l'animation de la parole ont commencé par établir une correspondance entre l'annotation des phonèmes et l'animation, manuellement (Parke, 1975) ou automatiquement(Lewis & Parke, 1986). Dans cette dernière recherche, Lewis & Parke (Lewis & Parke, 1986) ont présenté un système de synchronisation des lèvres dans lequel la parole enregistrée est ensuite décomposée en phonèmes et un modèle paramétrique modifie la forme de la bouche.

Par la suite, en 1990, Cohen et Massaro ont utilisé un modèle paramétrique en termes de perception de la parole pour produire un système de synchronisation labiale.

Ils ont également élargi le modèle pour la langue et pour modéliser les effets de co-articulation (M. M. Cohen & Massaro, 1993b; Massaro & Cohen, 1990). Bregler et al. (Bregler et al., 1997) ont conçu *Video Rewrite*, un exemple typique de techniques basées sur la parole qui permet de créer une nouvelle séquence audiovisuelle en concaténant des triphones à partir de bases de données construites. Cao et al. (Y. Cao et al., 2005) ont proposé une animation faciale expressive guidée par la parole. Dans cette approche, on considère que la précision de la synchronisation des lèvres est acceptable pour capturer les émotions par rapport aux modèles existants. Et afin de développer une animation faciale de synchronisation labiale de haute performance en temps réel, Y.M. Chen et al. (Y.-M. Chen et al., 2012) ont présenté un modèle d'animation dominée (DAM) qui intègre la fonction de dominance latente et la fonction d'animation intrinsèque pour affiner la coarticulation. Deena et al. (Deena & Galata, 2009)ont pu synthétiser une animation faciale en modélisant la correspondance entre le mouvement facial et la parole à l'aide du modèle de variable latente à processus gaussien partagé. Les deux données sont traitées séparément puis couplées ensemble pour produire un espace latent partagé. Ils ont donc pu modéliser la co-articulation en ayant un modèle dynamique sur l'espace latent. Les vidéos générées sont correctement synchronisées avec l'audio et présentent une dynamique faciale correcte. Ensuite, ce sont les méthodes d'apprentissage automatique qui ont été largement utilisées, comme les réseaux de neurones artificiels (Hofer & Richmond, 2010)les HMM (les modèles de Markov cachés)(L. Wang & Soong, 2015). Des améliorations supplémentaires peuvent être apportées pour assurer une synchronisation précise de la synthèse labiale en temps réel et l'intégration des émotions dans l'animation faciale guidée par la parole. Comme le travail récemment proposé par Viktor Igeland dans (Igeland, 2019), où une animation faciale a été généré à partir de la parole avec émotion en combinant différents modèle d'apprentissage profond.

3.3.5.3 L'animation guidée par la performance

Comme les techniques d'infographie relèvent le défi de rendre des acteurs plus vrais que nature, des performances plus vraies que nature sont nécessaires. Les nouvelles technologies de capture de données 3D se sont considérablement développées, notamment la capture de mouvement (MoCap), qui consiste à suivre par triangulation les positions 3D d'un nombre limité de marqueurs placés sur le corps d'un acteur, ainsi que les caméras de type RVB-D (ou caméras avec capteurs de profondeur) basées sur la vision binoculaire ou l'émission de lumière structurée, même des domaines différents tels que l'audiovisuel et les jeux vidéo sont de plus en plus demandeurs de ce type de technologie. Cela a donné naissance à une nouvelle approche qui est l'animation guidée par la performance, qui consiste à utiliser les données de mouvement capturées par certains dispositifs d'entrée et des caméras de capture de mouvement pour piloter le personnage synthétique. Les données de mouvement peuvent être utilisées pour

générer directement une animation faciale (Essa et al., 1996) ou pour déduire les UA des FACS dans la génération des expressions faciales.

Williams (Williams, 1990)a été le premier à introduire le concept de capture de mouvement pour le visage, et à se servir de l'animation d'un visage réel pour l'appliquer à un visage virtuel, afin de remédier aux difficultés liées à la gestion manuelle des paramètres de contrôle, il a utilisé quelques marqueurs sur le visage et une caméra vidéo pour scanner le visage et suivre les marqueurs. Kouadio et al. (Kouadio et al., 1998) ont proposé un système d'animation qui capture les expressions faciales d'un acteur et les ont utilisées pour animer en temps réel un personnage synthétique. Les points critiques du modèle de visage sont mis en correspondance avec des marqueurs en direct. Une combinaison linéaire des expressions de base obtenues en minimisant la distance euclidienne entre les points et les marqueurs correspondants est ensuite utilisée pour construire l'expression faciale intermédiaire. Différentes techniques de capture de mouvement ont été utilisé : par suivi de marqueurs, qui sont des billes réfléchissantes collées sur le visage ou des points dessinés sur la peau(Guenter et al., 1998; Williams, 1990), par suivi des point caractéristiques, ou par modèles déformables.

Depuis l'apparition des caméra RGB-D tels que le Kinect, la capture en temps réel de la performance faciale sans marqueur basée sur des capteurs ordinaires a été démontrée. De nombreux défis techniques se sont posés dans le développement de ce genre d'animation. Il s'agit notamment de savoir comment suivre avec précision les mouvements rigides et non rigides du visage de l'utilisateur, et comment faire correspondre les paramètres de suivi extraits qui pilotent l'animation faciale. Les pionniers de cette technique sont Weise et al.(Weise et al., 2011). Ils ont proposé un système d'acquisition sans marqueur qui utilise un système de suivi du visage, produit de l'enregistrement de la géométrie et de la texture, afin de surmonter les défis suscités.

L'animation faciale basée sur la performance peut intégrer d'autres techniques discutées ci-dessus ou d'autres effets de détail comme les cheveux et le mouvement du cou avec moins de marqueurs impliqués, afin d'augmenter le réalisme du modèle facial généré. En outre, une animation faciale intelligente axée sur la performance, capable d'adapter les données de mouvement d'entrée en fonction des exigences actuelles de l'animation faciale, sera certainement bénéfique pour les animateurs et les chercheurs(Ping et al., 2013).

3.4 Les Techniques d'animation faciale basées sur l'apprentissage profond

Bien que les méthodes conventionnelles, basées sur l'infographie, permettent de capturer une grande variabilité dans un petit vecteur de paramètres, elles ne répondent généralement pas à la qualité visuelle nécessaire. En effet, l'animation basée sur la

géométrie ne permet souvent pas de réaliser des déformations fines et ne parvient pas à produire des rendus réalistes dans les zones difficiles (bouche, yeux). Pour éviter ces problèmes, les techniques d'animation basées sur l'image utilisent des textures dynamiques qui capturent les détails et les petits mouvements qui ne sont pas expliqués par la géométrie. Cependant, cette technique a pour inconvénient d'exiger une mémoire importante et de limiter la flexibilité de l'animation, car les séquences de textures dynamiques doivent être concaténées de manière transparente, ce qui n'est pas toujours possible et est sujet à des artefacts visuels(Paier et al., 2020).

Le problème de l'animation faciale a suscité une multitude de recherches, car les différentes utilisations de l'animation faciale peuvent avoir des objectifs et des exigences radicalement différents. Alors que les algorithmes qui génèrent les expressions des personnages pour un jeu vidéo privilégient la vitesse par rapport à la qualité afin de maintenir une fréquence d'images élevée, les algorithmes utilisés pour générer les expressions des personnages pour les films privilégient la précision par rapport à la vitesse car chaque image est préenregistrée(Johnson, 2021).

Au cours de la dernière décennie de recherche, de nombreux domaines ont utilisé avec succès l'apprentissage profond en raison de ses perspectives très prometteuses. Il a obtenu de bons résultats d'application dans divers domaines, comme cela a été évoqué dans le chapitre précédent.

Le domaine qui a connu le plus de progrès grâce à l'apprentissage profond est celui de la vision par ordinateur et du traitement des images. Dans ce domaine, la génération de visages de personnages d'animation est une recherche à la fois intéressante et de grande ampleur. Ceci est dû à plusieurs raisons, parmi lesquelles la variété des expressions qu'un visage peut exprimer, l'utilisation d'images animées sur les plateformes sociales comme avatars sur les plateformes sociales. Par ailleurs, de nombreux travaux d'animation ont vu le jour dans le domaine de l'informatique affective. Des modèles génératifs puissants ont été développés, avec cet avènement des réseaux neuronaux profonds. En particulier, les VAE (Kingma & Welling, 2014) et les réseaux contradictoires génératifs (GAN) (I. Goodfellow et al., 2014) gagnent en popularité, car ils sont suffisamment puissants pour représenter de grandes distributions de données hautement dimensionnelles de haute qualité (par exemple, des images de visages ou des textures) (Paier et al., 2020).

Dans ce qui suit, une synthèse des travaux et approches proposés en matière d'animation faciale par apprentissage profond est présentée. Il est important de mentionner que la plupart des travaux sur l'animation faciale par apprentissage profond sont des méthodes guidées par les données, ou dans certains cas, ils sont fondés sur des approches hybrides.

3.4.1 Génération d'animation faciale guidée par la parole

L'animation faciale basée sur la parole est un processus qui permet de synthétiser automatiquement des personnages parlants à partir de signaux vocaux. Ces systèmes d'animation sont en mesure de simplifier le processus d'animation des films grâce à la génération automatique à partir du jeu de voix (Vougioukas et al., 2020). Ils sont également utilisés en post-production pour obtenir une meilleure synchronisation des lèvres lors du doublage de films. Par ailleurs, ces systèmes sont utilisés pour générer les parties du visage qui sont occultées ou manquantes dans une scène. Cette technologie permet aussi d'améliorer les télécommunications visuelles à bande limitée, soit en générant l'intégralité du contenu visuel à partir de l'audio, soit en remplissant les images manquantes.

Parmi les premiers travaux qui ont réussi à synthétiser une vidéo de haute qualité avec une synchronisation labiale précise, le travail proposé dans (Suwajanakorn et al., 2017). Les auteurs ont montré qu'en entraînant un réseau neuronal récurrent sur de nombreuses heures de séquences du discours hebdomadaire d'Obama, afin que le réseau apprenne la correspondance entre les caractéristiques audio brutes et les formes de la bouche, ils peuvent créer une vidéo crédible à partir d'enregistrements audio avec une synchronisation labiale convaincante. Cette méthode comprend une étape manuelle pour chaque vidéo cible, l'utilisateur devant sélectionner et masquer un substitut de dent.

Toutefois, leur réseau ne peut pas être entraîné sur une autre personne, en raison de la difficulté à obtenir des vidéos d'entraînement de longue durée. Cependant, le réseau formé sur Obama peut être réentraîné pour une autre personne avec bien moins de données de formation supplémentaires. La méthode présente certaines limites, telles que des erreurs de géométrie 3D ; par exemple, lorsque le menton occulte une partie de la chemise, cela produit un artefact de double menton. De plus, la méthode ne modélise pas explicitement les émotions et ne prédit pas le sentiment du discours d'entrée. Une autre limite concerne la modélisation de la langue, car dans ce système, les auteurs supposent que la texture de la bouche peut être entièrement déterminée par les positions des repères labiaux, ce qui n'est pas entièrement vrai pour certains sons, tels que "th".

Karras et al. (Karras et al., 2017) ont proposé un modèle de bout en bout qui pilote l'animation faciale 3D en temps réel avec une faible latence par le biais de l'entrée audio. Ils utilisent des CNN pour transformer des caractéristiques audios en maillages 3D d'une personne spécifique, et découvre simultanément un code latent compact qui désambiguïse les variations de l'expression faciale qui ne peuvent pas être expliquées par l'audio seul. Ce système est conceptuellement décomposé en sous-réseaux chargés de capturer la dynamique des articulations et d'estimer les points 3D du maillage.

La tâche du réseau est d'inférer l'expression faciale au centre de la fenêtre, à partir d'une courte fenêtre audio. L'expression est représentée directement sous forme de vecteurs de différence par sommet à partir d'une pose neutre dans un maillage de visage à topologie fixe. Une fois le réseau entraîné, le maillage est animé en glissant une fenêtre sur une piste audio vocale et le réseau est évalué indépendamment à chaque pas de temps. Bien que le réseau lui-même n'ait aucune mémoire des images d'animation passées, il produit des résultats temporellement stables dans la pratique.

Tandis que les auteurs de « *You Said That?: Synthesising Talking Faces from Audio* » ont proposé un modèle, *Speech2Vid* (Jamaludin et al., 2019), qui permet de générer des vidéos d'un visage parlant en utilisant uniquement un segment de parole audio et des images de visage de l'identité cible. Ce segment de parole peut ne pas être prononcé à l'origine par la personne cible. Cette méthode se distingue des approches précédentes car elle apprend directement les correspondances entre les caractéristiques audio et les données vidéo, au lieu d'apprendre les correspondances entre les phonèmes et les visèmes. Le modèle Speech2Vid est capable de produire des vidéos d'un visage parlant au moment du test, même en utilisant des images et de l'audio en dehors de l'ensemble de données d'entraînement, car il se concentre sur la partie vocale de l'audio et sur les régions faciales étroites des locuteurs dans les images.

Le modèle *Speech2Vid* se compose de trois éléments principaux : un encodeur audio, un encodeur d'image d'identité et un décodeur d'image de visage parlant. Pour un échantillon d'entrée donné, le modèle génère une image de sortie qui représente au mieux l'échantillon audio à un pas de temps spécifique. Au moment du test, la vidéo est générée image par image en glissant une fenêtre temporelle sur l'ensemble du segment audio tout en utilisant les mêmes images d'identité. La particularité de *Speech2Vid* est sa capacité à générer une vidéo de n'importe quelle identité parlante à partir de n'importe quelle source audio d'entrée. Cependant, même si *Speech2Vid* génère l'ensemble du visage, il est principalement axé sur l'obtention de mouvements de lèvres réalistes, et négligent généralement l'importance de la génération des expressions faciales (Vougioukas et al., 2020). Il souffre donc de la difficulté de faire bouger les mentons avec la bouche.

Duarte et al. (Duarte et al., 2019) ont proposé une nouvelle variante de réseau contradictoire génératif qui est conditionnée par le signal de parole brut, pour synthétiser des images faciales contenant des expressions et des poses. Pour l'apprentissage de ce GAN, les auteurs ont collecté et conservé un nouveau jeu de données, le jeu de données Youtubers, qui contient des signaux visuels et vocaux de haute qualité. Les tests de validation montrent que l'approche proposée parvient à synthétiser des images faciales crédibles avec une précision de 90,25 %, tout en préservant l'identité du locuteur dans environ 50 % des cas. Cependant, les résultats

restent ambigus (X. Li et al., 2022), bien qu'ils ont amélioré la qualité de génération de leur modèle en recherchant la longueur optimale de l'audio d'entrée.

Dans le but de mieux apprendre les visages normalisés à partir de la parole, Oh et al. (Oh et al., 2019) ont exploré un modèle reconstructif, *Speech2Face*, qui est entraîné sur un grand nombre de vidéos. Les chercheurs ont entraîné un encodeur audio en apprenant à aligner l'espace des caractéristiques de la parole avec celui d'un décodeur de visage pré-entraîné en utilisant des millions de vidéos naturelles de personnes parlant. Au cours de cet entraînement, qui se fait de manière auto-supervisée, le modèle apprend les corrélations voix-visage qui lui permettent de produire des images capturant divers attributs physiques des locuteurs tels que l'âge, le sexe et l'origine ethnique.

Le modèle *Speech2Face* s'est avéré stable et a produit des images de visage cohérentes déduites de différents segments de parole de la même personne, provenant de différentes parties de la même vidéo et d'une autre vidéo. Cependant, le modèle a inféré des visages différents en fonction de la langue parlée, lorsqu'il a été testé avec un exemple d'homme asiatique, disant la même phrase en anglais et en chinois. Alors que l'idéal aurait été d'avoir le même visage reconstruit dans les deux cas, le modèle a inféré des visages différents en fonction de la langue parlée.

Vougioukas et al. (Vougioukas et al., 2020) ont aussi présenté un système de bout en bout qui génère des vidéos d'une tête parlante, en utilisant uniquement une image fixe d'une personne et un clip audio contenant de la parole, et ce sans s'appuyer sur des caractéristiques intermédiaires conçues à la main. Les auteurs ont proposé un GAN temporel qui utilise trois discriminateurs permettant d'obtenir des images détaillées, une synchronisation audio-visuelle et des expressions réalistes. Le modèle a été entraîné sur les deux ensembles de données TCD-TIMIT, GRID, CREMA-D et LRW séparément. Les auteurs démontrent que leur méthode produit des séquences plus cohérentes et des mouvements de bouche plus précis par rapport à la méthode de base statique basée sur le GAN et à la méthode Speech2Vid. Bien que la méthode produit un mouvement labial plausible mais rend une identité incorrecte, et elle ne parvient pas à capturer avec précision les formes de la bouche et la texture détaillée du visage d'un sujet de test inconnu dont les caractéristiques faciales diffèrent des données d'entraînement.

Plusieurs autres recherches ont été menées en se basant sur les GANs, tel que celle proposée par Wen et al. (Wen et al., 2019), qui ont exploré la relation entre deux modalités, en proposant un GAN capable de générer des visages à partir de voix en faisant correspondre les identités des visages générés à celles des locuteurs, sur un ensemble d'apprentissage. Ce GAN reconstruit un visage via un vecteur audio capturé par un réseau d'intégration de la voix puis le visage et l'identité générés sont distingués par un discriminateur et un classificateur, respectivement. Les visages générés

présentent des associations d'identité avec le véritable locuteur. Mais, le résultat pose des problèmes évidents : des caractéristiques qui sont présents tel que les cheveux qui ne sont pas prédites par la voix), mais simplement obtenues à partir de leur cooccurrence avec d'autres caractéristiques.

En raison de la disponibilité d'importants ensembles de données vidéo 2D, la majorité des travaux existants visent à produire des vidéos 2D de têtes parlantes (Fan et al., 2022). Néanmoins, ces vidéos 2D ne sont pas adaptées à des applications telles que les jeux 3D et la réalité virtuelle, qui requièrent l'animation de modèles 3D dans un environnement 3D. Fan et al. (Fan et al., 2022)ont proposé un modèle de génération visuelle de parole autorégressif basé sur un transformateur, appelé *FaceFormer*, pour encoder les informations de contexte audio à long terme et prédire une séquence de maillages de visages 3D animés. Dans ce récent travail, Le problème de l'animation faciale 3D pilotée par la parole est formulé comme un problème d'apprentissage de séquence à séquence (seq2seq) et une nouvelle architecture *seq2seq* est proposée pour prédire de manière autorégressive les mouvements faciaux conditionnés à la fois par le contexte audio et la séquence de mouvements faciaux passés.

FaceFormer permet d'obtenir une meilleure qualité de synchronisation labiale et d'animation faciale réaliste par rapport à l'état de l'art. Cependant, il nécessite l'accès à la séquence audio complète, ce qui le rend inadapté aux applications de diffusion en ligne. La complexité quadratique en mémoire et en temps de l'auto-attention constitue un autre problème.

La génération d'animations faciales guidées par la parole est une tâche importante et difficile. Ce problème provient essentiellement de l'écart important entre les modalités audio et visuelles(H. Zhu et al., 2021). Afin de réduire cet écart, certains chercheurs ont introduit des informations supplémentaires dans leur modèle, notamment des marqueurs, des points clés, des flux optiques, etc. Les approches les plus courantes consistent à modifier la structure du réseau en se basant sur le GAN ou d'autres modèles génératifs.

3.4.2 Génération d'animation faciale guidée par l'image

L'animation automatique des expressions faciales à partir d'une seule image est à l'origine de nombreuses applications dans différents domaines, notamment l'industrie cinématographique, les technologies de la photographie, la mode et le commerce électronique et l'interaction homme-machine entre autres (Pumarola et al., 2018).

Ces dernières années, l'apprentissage profond a été largement exploité et a obtenu des résultats très performants dans plusieurs domaines tels que la synthèse des expressions faciales et les manipulations du visage, ces dernières étant généralement classées en quatre grands groupes différents en fonction du niveau de manipulation

(Tolosana et al., 2020), à savoir : Synthèse de visage entier, échange d'identité, manipulation d'attributs, échange d'expression.

Dans le but de transformer les images faciales, un certain nombre de méthodes d'apprentissage profond ont été proposées. Cette tâche a connu des avancées significatives, avec des architectures telles que StarGAN (Choi et al., 2018), qui est capable non seulement de synthétiser de nouvelles expressions, mais aussi de modifier d'autres attributs du visage, comme l'âge, la couleur des cheveux ou le sexe. Malgré sa généralité, StarGAN ne peut modifier qu'un aspect particulier d'un visage parmi un nombre discret d'attributs définis par la granularité d'annotation de l'ensemble de données. Par exemple, pour la tâche de synthèse d'expressions faciales, StarGAN est entraîné sur le jeu de données RaFD qui ne comporte que huit étiquettes binaires pour les expressions faciales, à savoir triste, neutre, colère, méprisant, dégoût, surpris, peur et heureux. Il peut ainsi transférer une expression à une image faciale donnée en fournissant l'étiquette discrète de l'expression cible. Par conséquent, ses capacités en matière d'édition d'expressions et de transfert d'expressions arbitraires sont assez limitées.

Le modèle GANimation (Pumarola et al., 2018), qui est proposé pour générer des animations faciales tenant compte de l'anatomie, est une autre architecture qui a connu un succès intéressant dans le domaine de l'animation faciale. Il synthétise continuellement les mouvements anatomiques du visage en contrôlant la magnitude d'activation de chaque unité d'action (UA). Bien que le codage des UA soit un modèle suffisamment complet pour décrire le mouvement du visage, la détection des UA constitue un problème ouvert, qui affecte la précision de la génération de la GANimation. La faible précision de l'annotation automatique des UA s'explique notamment par le manque de données annotées et le coût élevé de l'annotation, qui doit être effectuée par des experts hautement qualifiés.

StarGAN (Choi et al., 2018) et GANimation (Pumarola et al., 2018) ont été conçus pour le transfert d'expressions faciales en effectuant des traductions d'image à image, où la génération d'images faciales est conditionnée par des états émotionnels discrets. Cependant, les émotions humaines sont exprimées de manière continue, et ces états discrets ne suffisent pas à décrire les caractéristiques détaillées des expressions faciales. D'autres travaux ont été effectués pour réaliser la même tâche, en proposant de guider la génération par la géométrie. Comme Qiao et al. (Qiao et al., 2018)qui ont dérivé un modèle basé sur VAEGANs pour synthétiser des expressions faciales à partir d'une seule image et de plusieurs points de repère à travers quelques étapes de transfert. Leur modèle ne requiert ni l'étiquette de classe cible de l'image générée ni l'expression neutre d'un sujet spécifique comme niveau intermédiaire dans la procédure de transfert des expressions faciales.

Alors que *G2GAN* (Song et al., 2018), qui se concentre également sur les caractéristiques géométriques pour guider la procédure de synthèse des expressions, nécessite une expression de visage neutre et des paramètres de forme pour l'expression cible. Song et al. (Song et al., 2018) utilisent des générateurs doubles pour effectuer simultanément la synthèse et la suppression des expressions faciales. Les images faciales neutres sont générées par le réseau de suppression et utilisées pour le transfert ultérieur des expressions faciales. Cette procédure entraîne des artefacts supplémentaires et dégrade les performances, en particulier lorsque les visages d'entraînement sont collectés auprès d'autres sujets dans des émotions différentes.

ExprGAN a été proposé par Ding et al. (Ding et al., 2018) , lequel est capable de synthétiser des expressions faciales avec une intensité contrôlable, et un réseau de contrôle d'expression est proposé pour apprendre le code d'expression. Toutefois, ExprGAN génère des images conditionnées par des étiquettes d'expression et des valeurs d'intensité, contrairement à *G2GAN* qui utilise la géométrie du visage comme condition de contrôle qui n'est pas limitée à certains styles d'expression. Les méthodes génératives basées sur des modèles peuvent difficilement générer des images de visage photoréalistes et préservant l'identité tout en permettant un ajustement continu de l'expression cible (Song et al., 2018).

Tandis que Geng et al. (Geng et al., 2018) ont proposé d'utiliser un GAN guidé par la déformation *wg-GAN* (Warp-Guided GAN) qui génère des animations faciales en temps réel à l'aide d'une seule photo. Ils ont besoin d'une seule photo de portrait cible avec le visage dans la pose neutre-frontale et génère des animations photo-réalistes imitant une source de conduite. La méthode fusionne instantanément les détails du visage tels que les rides et les plis pour obtenir une expression faciale haute-fidélité. Pour y parvenir, l'image est affinée par une série de GAN : un pour affiner le visage déformé et un autre pour peindre les occlusions (yeux et bouche). Cette méthode a obtenu des résultats nettement meilleurs que certaines méthodes de l'état de l'art en principalement en terme synthèse de détails à petite échelle, tels que les rides, les plis, les ombres propres, les dents, etc. Néanmoins, des limites demeurent ; le processus de déformation est sensible au mouvement de la tête (changement de pose) en raison de la déformation 2D. Une autre limitation, la photo du portrait d'entrée doit être prise dans une pose frontale. En outre, la procédure de synthèse de réseau n'est pas en mesure de récupérer exactement les mêmes détails que le visage original et peut entraîner une certaine déviation de l'expression source.

Ververas et al. (Ververas & Zafeiriou, 2020) ont transformé une image de visage d'entrée en une nouvelle image en fonction des valeurs continues d'un modèle statistique de mélange de formes du mouvement du visage. Ils ont utilisé les paramètres du 3DMM comme conditions pour le cadre de synthèse conditionnelle. Contrairement à *StarGAN* (Choi et al., 2018), qui utilise des émotions discrètes, et à *GANimation*

(Pumarola et al., 2018), qui est basé sur les unités d'action du visage, tous deux incapables de décrire le mouvement produit par la parole ou la combinaison du mouvement de la parole et de l'expression, *SliderGAN* (Ververas & Zafeiriou, 2020), en revanche, est capable de synthétiser des déformations lisses de l'expression et de la parole dans les images en utilisant des modèles de mélange de formes 3D de l'expression et de la parole respectivement, car ils peuvent décrire à la fois le mouvement produit par l'expression et/ou le mouvement produit par la parole. À cette fin, un nouveau générateur basé sur des réseaux neuronaux convolutifs profonds (DCNN) est proposé, ainsi qu'une stratégie d'apprentissage qui fait appel à l'apprentissage contradictoire. Un des avantages du processus d'apprentissage est un réseau de régression très puissant qui convertit l'image en un certain nombre de paramètres de mélanges de formes, qui peuvent ensuite être utilisés pour conditionner les entrées du générateur.

Le modèle SliderGAN a également certaines limites - il s'est avéré incapable de conserver l'identité des images d'entrée complètement inchangée, ou bien il s'adapte trop à des images spécifiques. La présence d'expressions extrêmes constitue une autre limite, que ce soit lors de la génération ou de la manipulation. Dans les deux cas, les images présentent souvent beaucoup d'artefacts.

D'autre chercheurs ont concentré leur travail sur l'animation du visage par reconstitution de visage (face reenactment), où l'identité de la personne est tirée de l'image source et le mouvement du visage de l'image de conduite. Bien que ces méthodes aient montré des résultats de haute qualité, lorsque l'image source et l'image de conduite représentent la même personne ou des structures faciales très similaires, elles présentent une précision limitée et les structures faciales du visage conducteur s'échappent vers la sortie, déformant ainsi le résultat de la reconstitution. Les recherches actuelles tentent d'améliorer les modèles soit en améliorant les entrées (unités d'actions), soit en améliorant les architectures (blocs supplémentaires), comme *FACEGAN proposé par* Tripathy et al. (Tripathy et al., 2021). Ces auteurs présentent un cadre de reconstitution de visage interprétable et contrôlable pour contrôler la pose et les expressions de différentes sources et identités de conduite. Ce modèle ne pose aucune restriction sur la correspondance entre les paires source et conduite, car il combine les meilleures propriétés de l'unité d'action et des représentations du mouvement des points de repère du visage pour réduire le problème de fuite d'identité. En plus, *FACEGAN* traite séparément les régions de l'arrière-plan et du visage pour optimiser la qualité du résultat.

L'animation faciale guidée par l'image est un problème considérable qui vise à synthétiser automatiquement des images (généralement de façon continue) de visages à partir d'une seule image source. Il est généralement formulé comme un problème de génération conditionnelle (X. Wu et al., 2021) : étant donné un ensemble de variables conditionnelles décrivant le mouvement du visage (comme l'expression ou la pose), le

système de synthèse doit être capable de transformer une image source d'un visage en images cibles correspondantes. Le mouvement du visage est modélisé soit par sa forme soit par son expression en utilisant différentes représentations telles que des points de repère 2D, des unités d'action ou des codes d'émotion. Malgré leur capacité à produire de bons résultats, les images synthétisées présentent toujours des artefacts ou même des formes irrégulières. Les réseaux adversariaux génératifs (GAN) ont récemment fait preuve de performances remarquables dans plusieurs tâches, en particulier la génération et la manipulation de visages La plupart des GANs pour la génération de poses ont utilisé des GANs conditionnels grâce à la disponibilité de jeux de données appariés. En revanche, pour la génération d'expressions faciales, les GAN ont utilisé des modèles de traduction image à image altérés en raison du nombre limité de données appariées. Par conséquent, les scores de précision sont encore faibles par rapport aux ensembles de données appariées (Kammoun et al., 2022).

3.4.3 Génération d'animation faciale guidée par la vidéo

L'une des approches les plus répandues et les plus performantes pour créer des visages virtuels consiste souvent à capturer les performances faciales de personnes réelles en vidéo. L'une de ces techniques les plus remarquables est la méthode de Thies et al. (Thies et al., 2016) connue sous le nom de *Face2Face*. Elle permet de transférer les expressions faciales d'une personne à une autre en temps réel en utilisant uniquement du matériel de base. Cette approche est considérée comme étant le premier système de reconstruction faciale en temps réel qui ne nécessite qu'une entrée RVB monoculaire. Son objectif est d'animer les expressions faciales de la vidéo cible par un acteur source et de rendre la vidéo de sortie manipulée de manière photo-réaliste, autrement dit, de remplacer l'expression faciale d'une personne dans une vidéo par l'expression faciale d'une autre personne. La séquence cible peut être n'importe quelle vidéo monoculaire, comme une ancienne vidéo YouTube avec une performance faciale.

Les auteurs commencent par reconstruire l'identité de la forme de l'acteur cible par un regroupement non rigide basé sur un modèle sur une séquence d'entraînement préenregistrée. Au moment de l'exécution, les expressions faciales des vidéos source et cible sont suivies à l'aide d'une mesure de cohérence photométrique dense par une approche d'analyse de synthèse dense basée sur des antécédents faciaux statistiques. Une nouvelle fonction de transfert est proposée pour transférer les expressions de l'acteur source à l'acteur cible en temps réel. Pour la synthèse finale de l'image, le visage cible est rendu à nouveau avec les coefficients d'expression transférés et composé avec le fond de la vidéo cible en tenant compte de l'éclairage estimé de l'environnement. Enfin, une nouvelle approche de synthèse de bouche basée sur l'image est introduite afin de générer un intérieur de bouche réaliste en récupérant et en déformant les formes de bouche les mieux adaptées à partir de la séquence échantillon hors ligne.

Bien que les résultats obtenus surpassent les méthodes de l'état de l'art à la fois en termes de qualité de la vidéo obtenue et de temps d'exécution, cette approche présente certaines limites, telles que certaines difficultés rencontrées avec les occlusions du visage par les cheveux longs et la barbe, le modèle de mélange utilisé pour la reconstruction est de faible dimension (76 coefficients d'expression), la synthèse de la bouche basée sur la récupération suppose une variation suffisante des expressions visibles dans la séquence cible.

Parmi les approches qui ont été proposées pour contourner l'étape d'optimisation coûteuse, l'approche proposée par Laine et al. (Laine et al., 2017) présente un cadre d'apprentissage profond en temps réel pour la capture de performance faciale basée sur la vidéo et le suivi 3D dense du visage d'un acteur à partir d'une vidéo monoculaire. Cette approche entraîne un réseau neuronal convolutif profond (CNN) à produire une sortie de haute qualité, y compris des régions auto-occultantes, à partir d'une séquence vidéo monoculaire de ce sujet. Selon les auteurs, le système proposé réduit considérablement la quantité de travail nécessaire au développement de jeux vidéo ou de films par rapport aux autres pipelines de capture des performances faciales actuellement utilisés. Cependant, un inconvénient majeur de ce système est la nécessité d'un calibrage par utilisateur. Pour obtenir un résultat de haute qualité, chaque nouvelle identité nécessite un ensemble de données de 5 à 10 minutes, ce qui implique que l'acteur doit avoir un rôle suffisamment important dans le jeu pour justifier le coût.

Depuis les premières années de l'émergence de la technologie, des modèles GAN permettant de générer une séquence d'images ou une vidéo/animation ont été proposés, tels que le GAN vidéo, VGAN (Vondrick et al., 2016) , et le GAN temporel, TGAN (Saito et al., 2017) . Bien que ces modèles puissent apprendre une représentation sémantique de vidéos non étiquetées, ils produisaient un clip vidéo de longueur fixe. Pour surmonter ce problème, Tulyakov et al. (Tulyakov et al., 2018) ont proposé MoCoGAN. Ce modèle fournit une décomposition du mouvement et du contenu pour la génération de vidéos. Par exemple, pour les vidéos de personnes exécutant différentes expressions faciales, MoCoGAN apprend à séparer l'identité d'une personne de son expression, ce qui lui permet de synthétiser une nouvelle vidéo d'une personne exécutant différentes expressions, ou de fixer l'expression et de générer différentes identités. En plus de son efficacité prouvée, MoCoGAN peut générer des vidéos avec un contenu identique mais un mouvement différent, ainsi que des vidéos avec un contenu différent et un mouvement identique. En revanche, il présente le problème de l'apparence non naturelle des objets en mouvement et de l'assimilation des objets en arrière-plan.

"*Deep Video Portraits*" est un système de reconstruction intégrale de vidéos de portraits, c'est un autre travail basé sur le principe de la reconstruction de visage qui a été proposé en 2018, par Kim et al. (H. Kim et al., 2018). Il permet d'animer des expressions faciales par transfert d'émotion de vidéo à vidéo, nécessitant ainsi une vidéo

source en entrée et une séquence de conduite pour piloter l'animation. La pose de la tête, les expressions faciales et les mouvements des yeux de la personne dans une vidéo sont transférés d'un acteur source à un acteur cible. Les auteurs proposent un réseau de rendu-à-vidéo qui convertit une séquence de rendus graphiques simples en vidéo photoréaliste et cohérente dans le temps. Le réalisme de ce transfert de rendu à vidéo est obtenu par un entraînement contradictoire minutieux, produisant des vidéos cibles modifiées qui imitent le comportement de l'entrée créée synthétiquement. Ce système ouvre ainsi un nouveau niveau de capacité dans de nombreuses applications, telles que la reconstruction vidéo pour la réalité virtuelle et la téléprésence, le montage vidéo interactif et le doublage visuel.

Cependant, ce système souffre encore de certaines limitations. Telles que : les positions extrêmes de la tête de la cible, comme les grandes rotations, ou les expressions bien en dehors de cette plage peuvent entraîner une dégradation de la qualité visuelle du portrait vidéo généré. Le mouvement du torse, des cheveux ou de l'arrière-plan ne peut pas être contrôlé activement. De plus, en raison des limites de mémoire et de temps d'apprentissage, le système ne produit que des images de résolution moyenne, ce qui rend particulièrement difficile la reproduction de détails à petite échelle, comme les dents individuelles, de manière cohérente dans le temps.

Ainsi, pour permettre le contrôle de la pose de la tête en plus des expressions faciales, le modèle X2Face proposé par Wiles et al. (Wiles et al., 2018) utilise une photo du visage ou un échantillon d'une autre modalité (par exemple, audio) pour modifier la pose et l'expression d'un visage donné pour le montage vidéo/photo. Ils entraînent le modèle de manière auto-supervisée en recevant deux échantillons : un échantillon source (vidéo) et un échantillon pilote (vidéo, audio ou, une combinaison). L'échantillon généré hérite de la même identité et du même style (par exemple, la coiffure) que l'échantillon source et obtient la pose et l'expression de l'échantillon conducteur. Les auteurs ont utilisé un réseau d'intégration qui factorise la représentation du visage de l'échantillon source et applique la frontalisation du visage. Malheureusement, les auteurs n'ont rapporté que les échantillons visuels générés et aucune autre métrique n'est utilisée pour la comparaison.

Récemment, le succès et la simplicité des techniques basées sur les images 2D et leurs parallèles avec les flux de travail 3D ont incité Moser et al. (Moser et al., 2021) à proposer un algorithme simple pour le transfert automatique des expressions faciales des vidéos vers un personnage 3D, ainsi qu'entre des personnages 3D distincts à travers leurs animations rendues. Leur méthode commence par l'apprentissage d'une représentation latente commune et sémantiquement cohérente pour les différents domaines d'images d'entrée, à l'aide d'un modèle de traduction d'images non supervisé. Elle apprend ensuite, de manière supervisée, une correspondance linéaire entre la représentation codée des images des personnages et les coefficients d'animation. Au

moment de l'inférence, étant donné le domaine source, il régresse les coefficients d'animation correspondants pour le personnage cible.

La méthode s'est avérée efficace pour les personnages spécifiques à un acteur ou non, et nous l'avons testée dans des conditions strictes, avec des résultats prometteurs pour la généralisation aux modèles multi-identités. Cette méthode peut également être utilisée pour transférer des animations faciales entre des personnages distincts sans paramètres de maillage cohérents ni prieurs géométriques élaborés. Cependant, la qualité des résultats est limitée par la qualité et la complexité des déformations faciales du personnage. Une autre limitation est que la recherche a été restreinte à des conditions d'enregistrement contrôlées.

La production d'une animation faciale guidée par vidéo englobe plusieurs tâches telles que la capture des mouvements du visage, le reciblage des mouvements du visage, la reconstruction du visage ou le transfert des expressions faciales. Le fait de disposer d'une vidéo du visage cible et d'une vidéo de conduite de la source peut atténuer certains des problèmes de réalisme, par exemple, car le contenu peut être soit hérité de la séquence cible, soit emprunté à la source. Certains travaux ont utilisé un modèle déformable pour piloter l'expression faciale dans une vidéo en modifiant les paramètres d'expression, tandis que d'autres l'ont utilisé pour la reconstruction du visage ; d'autres encore ont opté pour le remplacement du visage dans une vidéo.

3.5 Conclusion

L'animation faciale par ordinateur a connu une croissance continue et rapide depuis les travaux pionniers de Frederick I. Parke en 1972. Cette évolution est notamment due à une forte hausse de la demande de personnages virtuels ou d'avatars dans le domaine de la programmation des jeux, de la réalisation de films, de l'interaction homme-machine et de la communication homme-machine. L'objectif de la recherche liée à l'animation du visage, qui consiste à atteindre le réalisme en temps réel de manière automatisée, n'a pas encore été atteint. Cependant, des succès dans chaque domaine ont été récemment rapportés.

Dans ce chapitre, nous avons défini plusieurs concepts de base du domaine de l'animation faciale. Ensuite, nous avons discuté et passé en revue plusieurs techniques classiques et conventionnelles d'animation faciale par ordinateur, telles que le mélange et l'interpolation de formes, les méthodes basées sur la paramétrisation, les approches basées sur la déformation, la modélisation musculaire basée sur la physique et les approches d'animation basées sur les données. Pour chaque technique, nous avons décrit les idées principales et comparé leurs forces et faiblesses. La plupart des techniques sont en fait synthétisées par plusieurs méthodes afin d'obtenir de meilleures performances. Ensuite, nous avons examiné plusieurs études de recherche récentes qui

ont abordé l'animation faciale via des méthodes d'apprentissage profond. Ces nouvelles méthodes sont axées sur les données et nous les avons classées en trois catégories, en fonction de leurs données d'entrée : les méthodes axées sur la parole, les méthodes axées sur l'image et les méthodes axées sur la vidéo. La maîtrise de l'animation faciale est depuis longtemps un défi dans le domaine de l'animation par ordinateur et de l'infographie, et ce défi continue de prendre de l'importance avec l'émergence de nouvelles techniques d'apprentissage profond.

Deuxième partie

Présentation et évaluation des contributions

Chapitre 4 : Animation faciale basée sur les points d'intérêt

4.1 Introduction

Ces dernières années, l'animation faciale est devenue un domaine d'intérêt croissant, malgré sa complexité. Les difficultés de ce domaine se reflètent dans les approches actuelles, car il est encore très difficile de créer une animation de qualité sans un travail long et fastidieux. Ainsi, de nombreux travaux ont été réalisés pour tenter de reproduire fidèlement le visage humain, contribuant notamment à l'évolution de l'animation faciale, et à l'augmentation considérable de son utilisation.

On peut distinguer plusieurs domaines d'application (Dutreve, 2011), à savoir : les applications en temps réel, pour lesquelles le temps interactif est nécessaire et les ressources sont partagées entre différentes tâches tels que les jeux vidéo, les mondes virtuels ou les systèmes d'interaction homme-machine, etc. D'autre part, des applications dites "off-line", ne nécessitant pas de temps réel, mais la qualité du résultat est prioritaire par rapport à la vitesse de calcul, telles que dans le cinéma avec la généralisation des films d'animation 3D, la création d'effets spéciaux, la réalisation de spots publicitaires, etc.

Dans ce chapitre, nous nous intéressons à l'animation faciale basée sur les points caractéristiques, nous présentons notre contribution, qui peut être résumée en deux axes : premièrement, la synthèse d'animation par déformation basée sur la paramétrisation MPEG-4 (Bordjiba & Merouani, 2012), et deuxièmement, le transfert d'animation faciale basé sur les points caractéristiques du visage (Bordjiba & Merouani, 2013).

Dans le premier travail, notre objectif était d'animer un modèle de visage tridimensionnel, par interpolation de forme, et par déformation des points de contrôle. Un des problèmes de ce type de méthode réside dans la définition des points de contrôle et de leurs zones d'influence. C'est pourquoi nous avons présenté une méthode permettant de calculer automatiquement les zones d'influence de chaque point de contrôle afin de réduire toute intervention manuelle.

Dans la seconde contribution, nous avions exploré le reciblage d'animation faciale, qui consiste à animer un visage cible (virtuel) à partir de données extraites d'un visage source. Grâce au transfert automatique des mouvements du visage d'un modèle existant (source) vers un nouveau modèle (cible), il est possible d'économiser considérablement la spécification et l'animation minutieuse d'un modèle en un nouveau modèle de visage.

La source des mouvements de visages peut avoir plusieurs formats, notamment des vidéos de visages en 2D, des données de mouvements de visages capturés en 3D et des maillages de visages animés, et les modèles cibles se présentent principalement comme des maillages de visages statiques en 2D ou 3D.

4.2 Synthèse d'animation faciale par déformation

Au début, la recherche était principalement axée sur la synthèse d'animation, de sorte que la création était entièrement réalisée par ordinateur. En effet, produire des expressions faciales humaines réalistes à partir de modèles faciaux en 3D est l'un des problèmes les plus difficiles à résoudre dans le domaine de l'infographie. Bien que des progrès considérables soient réalisés dans les techniques de modélisation et d'animation des visages, les manipulations sophistiquées suscitent toujours beaucoup de méfiance de la part des novices et coûtent même beaucoup de temps aux animateurs professionnels pour en saisir l'essentiel(B. Li et al., 2013). Dès lors, un système intuitif, facile et efficace pour synthétiser les expressions faciales serait utile dans une variété d'applications telles que l'industrie cinématographique, les jeux vidéo et la téléconférence.

La réutilisation des données dans le domaine de l'animation est tout aussi importante et intéressante(Garchery, 2004). Le recours à une technique à base de paramétrisation devrait ainsi permettre de dépasser cette contrainte ; plutôt qu'un système basé sur l'interpolation des modèles dans leur ensemble, notre objectif est de pouvoir réutiliser les données d'animation sur différents modèles. Plusieurs systèmes de paramétrisation ont été proposé, tel que MPA (*An Interactive Multimodal Facial Animation System*, 1993), FACS (Ekman & Friesen, 1978) et MPEG 4(Koenen et al., 1997; *Video | MPEG*, s. d.).

À cette fin, nous proposons un système simple et rapide de déformation des maillages géométriques. Le maillage géométrique peut être caractérisé par la localisation de points caractéristiques, et l'animation du maillage peut être définie par les déplacements de ces points caractéristiques. L'algorithme décrit ici peut être appliqué à l'animation de tels maillages. Pour illustrer cette approche, on a choisi d'utiliser l'algorithme pour le maillage facial MPEG-4, qui est caractérisé par les paramètres de définition du visage (FDP). Nous examinons les résultats de la déformation du maillage appliquée à l'animation faciale.

4.2.1 La paramétrisation MPEG-4

Parke (Parke, 1982) fut le premier à proposer un visage virtuel animé. Pour chaque expression clé, le maillage du visage est modifié sommet par sommet. Quant aux positions intermédiaires, elles sont obtenues par interpolation entre ces expressions clés. Une telle manipulation permet de bénéficier d'une importante liberté artistique, car

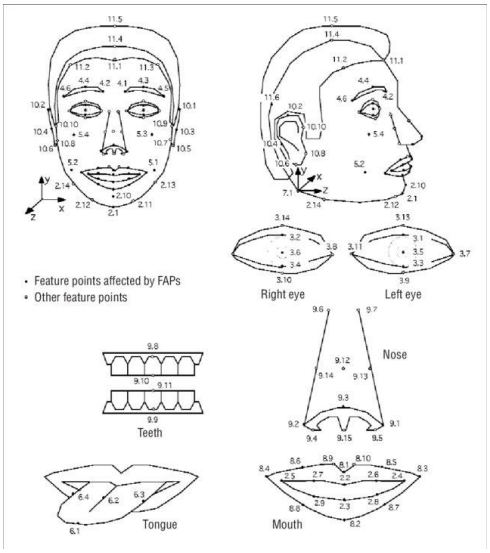


Figure 4-1 : L'ensemble des points de caractéristiques des paramètres de définition des visages (FDP) du MPEG-4 (Monjaux, 2007).

elle ne repose sur aucun paramètre prédéfini. Toutefois, pour obtenir des résultats, il faut faire appel aux connaissances, à l'expérience et à l'intuition du concepteur. Ce travail fastidieux sur un grand maillage empêche de produire des résultats vraiment réalistes, et la complexité et la précision des modèles 3D actuels rendent une telle méthode inenvisageable. En plus, Les informations relatives à la déformation doivent être définies pour chaque modèle, ce qui n'est pas trivialement réutilisable pour d'autres modèles.

Ce modèle de Parke a été le modèle de départ dans les décodeurs d'animation faciale MPEG-4 (Abrantes & Pereira, 1999). Le Moving Picture Experts Group (MPEG) a développé cette norme MPEG-4 pour répondre aux besoins de compression de fichiers et de communication multimédia en utilisant des points de contrôle pour déformer les images. Cette norme est adaptée aux scènes en 3D et comprend une méthode de déformation des visages.

Ce système de paramétrage comprend 84 paramètres décrivant le modèle du visage en termes de forme et de texture, *Face Definition Parameters* (FDP), et 68 paramètres d'animation du modèle, *Face Animation Parameters* (FAP), qui spécifient le mouvement global et/ou les déformations locales du visage, ainsi qu'une normalisation des visages utilisant les distances entre les points clés, *Facial Animation Parameter Units* (FAPU) et une manière de définir les déformations sur l'ensemble du visage à l'aide de tables d'animation, *Facial Animation Tables* (FAT). Ainsi, le descripteur MPEG-4 permet de reproduire une large gamme d'expressions faciales, d'émotions ou de mouvements de la parole (Monjaux, 2007).

La Figure 4-1 montre l'emplacement des FDPs sur le visage, ce sont les points pleins qui subissent l'influence des paramètres de l'animation FAP. Quant aux points creux, ils sont fixes et servent à la calibration du visage. Les FDP sont regroupés par zone (9 pour le nez et 3 pour les yeux par exemple).

4.2.2 Méthode proposée de la synthèse par déformation

L'objectif visé était de créer une animation faciale en 3D à l'aide de la technique d'interpolation des formes et de déformation des points de contrôle. Cependant, une des difficultés de cette méthode est la nécessité de définir les points de contrôle et leurs zones d'influence, ce qui peut être une tâche fastidieuse et requérir une intervention manuelle importante. Afin de résoudre ce problème, nous avons développé une méthode permettant de calculer automatiquement les zones d'influence de chaque point de contrôle, réduisant ainsi le besoin d'intervention manuelle.

La technique de déformation basée sur l'interpolation des images-clés consiste en la définition des différentes déformations envisageables et leur interpolation dans le but de créer des animations.

Notre algorithme comprend plusieurs étapes, la première étant le chargement du modèle 3D représentant le visage, et ayant une structure polygonale. Cette étape préliminaire prépare les données sur lesquelles l'animation sera appliquée. Le modèle de visage utilisé est représenté par un maillage 3D et est associé à des textures pour rendre le visage plus réaliste. Ce maillage de visage correspond à un ensemble de points, appelés « sommets du maillage », et qui peuvent être des points de contrôle ou des points ordinaires. En utilisant la paramétrisation MPEG-4, les points de contrôle sont les FDP.

Le moteur de déformation se compose de deux étapes principales :

- *L'initialisation*, qui utilise les données des points de contrôle et du maillage. Cette étape est réalisée une seule fois pour calculer les poids qui reflètent l'influence de chaque point de contrôle sur les points ordinaires du maillage.
- *La déformation*, quant à elle, consiste à calculer le déplacement de chaque sommet du maillage après avoir calculé les zones d'influence et les poids de chaque point de contrôle sur les sommets du maillage.

Avant d'animer un modèle de visage 3D basé sur cet algorithme de déformation basé sur les points de contrôle, il est nécessaire de sélectionner les points de contrôle (FDP) qui seront utilisés pour la déformation. Cette étape est réalisée à l'aide de l'outil XFACE, une boîte à outils open source conçue pour la génération d'agents parlants en 3D. XFACE s'adresse à la fois aux chercheurs travaillant sur des sujets similaires et aux développeurs de l'industrie du logiciel, et est basé sur les spécifications de la norme MPEG-4 (Balci, 2004). Une fois les points de contrôle sélectionnés, ils sont exportés au format XML. Dans ce qui suit, nous examinerons en détail chacune des deux étapes de l'algorithme de déformation utilisé.

4.2.2.1 Initialisation

La méthode d'initialisation repose sur une méthode proposée par Stéphane Garchery (Garchery, 2004). Cet algorithme calcule automatiquement les zones d'influence et les poids uniquement à partir des FDP. Ce calcul est effectué en trois phases : le calcul de la répartition des points de contrôle sur le maillage, la sélection des points de contrôle influents pour chaque point ordinaire, et enfin le calcul des influences de chaque point de contrôle sur les points ordinaires.

a. Calcul de la zone d'influence :

Avant de procéder au calcul de la zone d'influence de chaque point de contrôle, le modèle 3D est subdivisé en zones spécifiques correspondant aux différentes caractéristiques du visage, telles que les yeux, la bouche et le nez. Cette subdivision permet de réduire les calculs de distance de surface. Par exemple, pour la région des lèvres, seuls les points de contrôle qui définissent le contour des lèvres ont une influence sur les points ordinaires de cette région, et non les points de contrôle situés en dehors de cette zone. Nous avons utilisé une méthode basée sur l'extraction de caractéristiques à partir des relations anthropométriques pour parvenir à ce résultat, comme le montre la Figure 4-2. Bien que plusieurs formules existent pour cette méthode, celles qui ont été décrites par Farkas (Farkas, 1994) ont montré leur efficacité :

$$h_{face} \cong 1{,}8 \times d_{eye} \qquad (4.1)$$

$$h_{eye} \cong h_{face}/5 \qquad (44.2)$$

$$w_{eye} \cong 0{,}255 \times h_{face} \qquad (44.3)$$

Où :

h_{face} : est la distance entre le dessous des lèvres et le dessus des sourcils.

w_{eye} : est la distance entre la coordonnée du centre de l'œil et le côté du visage.

d_{eye} : est la distance entre les yeux.

h_{eye} : est la distance entre le dessus du sourcil et le dessous de l'œil.

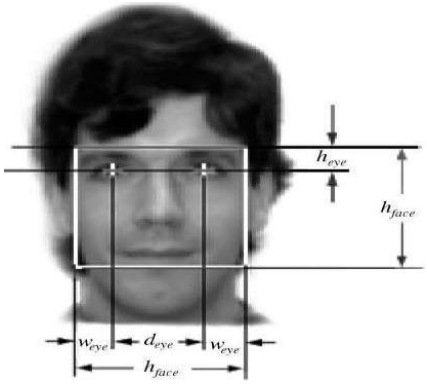


Figure 4-2 : Extraction des caractéristiques par les relations anthropométriques.

Par la suite, pour chaque région ainsi définie, la répartition des points de contrôle sur le maillage est définie, ce qui revient à définir des zones d'influence pour chaque point de contrôle, et il est donc nécessaire d'utiliser une métrique. La métrique consiste à calculer une approximation du diagramme de Voronoï sur la surface, avec la totalité des points de contrôle, utilisant la distance surfacique.

Grâce à l'approximation de Voronoï, il est possible de déterminer, pour chaque sommet du maillage, le point de contrôle le plus proche en termes de distance surfacique. Cela permet de définir les régions PDF dites "strictes". La distance surfacique entre deux points est calculée comme la somme des distances des arcs parcourus entre deux points. Dans ce cas, le maillage est assimilé à un graphe, les sommets étant les points, les arêtes étant les droites des triangles. On peut alors utiliser l'algorithme de Dijkstra pour calculer le chemin le plus court entre deux points donnés (Garchery, 2004), et la distance surfacique correspond alors à la somme des longueurs des arêtes de ce chemin entre les deux points. L'algorithme de Dijkstra n'est applicable qu'aux graphes ayant des poids positifs ou nuls, ce qui est le cas de nos méthodes du fait que le poids d'une arête correspond à la distance entre deux points. L'algorithme utilisé peut-être décrit comme suit :

Algorithme 2 : Algorithme de Dijkstra

Entrée : un maillage contenant *n* vertices, le point source *s*, le point destination *d*.

Sortie : le plus court chemin entre *s* et *d*.

Début

```
Pour i allant de 0 à n
        Longueur (vi) = Infinie
        L(s) = 0
        S = vide
        Tant que d n'existe pas dans S faire
        u = a // vertice n'est pas dans S qui possède le poids minimal.
        Ajouter u à S
        Pour toute vertice v qui n'existe pas dans S
                Si L(u) + distance (u, v) < L(v) alors L(v)= L(u) + distance (u, v)
```

Fin

Pour calculer la zone d'influence, Stéphane Garchery (Garchery, 2004) a présenté une méthode qui peut être résumée comme suit : à partir de tous les points de contrôle, examiner tous les voisins le long des arcs. À chaque étape, plusieurs situations peuvent se présenter :

- Si le point n'a jamais été visité, il est considéré comme un chemin potentiel et est associé à son point de départ. Il est ajouté à la liste des prochains points à traiter.
- Si le point a déjà été visité à partir du même point de contrôle, on conserve la distance de surface minimale.

- Si le point est déjà sous l'influence d'un autre point, il est placé sous l'influence du point ayant la distance surfacique minimale.
- Si le point est un point de contrôle, il n'est pas sous l'influence de son point de départ et est retiré de la liste des points à traiter.
- Et ainsi de suite, en prenant en compte les points déjà parcourus.

Le chemin le plus court entre les points de contrôle correspond à la somme minimale des distances des chemins allant des points de contrôle jusqu'à la frontière entre les régions. Cette information relative aux points de contrôle nous permet de déterminer, pour tout point ordinaire du maillage, le point de contrôle le plus proche ainsi que la distance entre les points de contrôle.

Après avoir testé cet algorithme, nous avons constaté que les résultats n'étaient pas satisfaisants, en particulier dans la région de la bouche où la structure du visage est irrégulière. Les déformations générées ne sont pas réalistes car l'algorithme ne passe pas par tous les sommets du maillage, ce qui signifie que certains points ne sont sous l'influence d'aucun point de contrôle. Nous proposons donc la solution suivante :

- Pour chaque point de contrôle, calculer la distance surfacique avec chaque sommet du maillage en parcourant tous les sommets.
- Le point est placé sous l'influence du point de contrôle ayant la distance surfacique minimale.
- Si le point est un point de contrôle, il n'est sous l'influence d'aucun point de contrôle.

b. Choix des points influents

Cette étape vise à identifier les points de contrôle qui influencent les points ordinaires. Dans le cadre de notre approche pour l'animation faciale, nous souhaitons définir des régions à l'aide d'un ensemble de points de contrôle qui entoure des points ordinaires. Pour ce faire, nous utilisons le critère de la distance d'angle θ_i (Garchery, 2004), entre les points de contrôle FDP_i et le point ordinaire FP , comme le montre la Figure 4-3. L'angle θ_i représente l'angle entre les vecteurs $\overrightarrow{FP\ P}$ et $\overrightarrow{FP\ FDP_i}$ où FDP_i sont les points de contrôle voisins de FP :

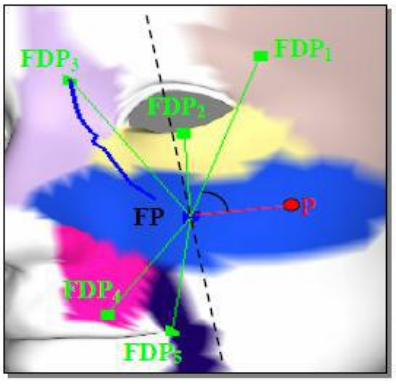


Figure 4-3: Choix des points influents.

$$\theta_i = \left(\overrightarrow{FP\ P}, \overrightarrow{FP\ FDP_i}\right) = P\ \widehat{FP\ FDP_i} \quad (4.4)$$

Chaque sommet P peut être influencé par un ou plusieurs points de contrôle. Il convient de parcourir les points de contrôle voisins calculés précédemment, tout en calculant les θ_i, et en sélectionnant les *FDP* voisins de *FP* qui peuvent influencer P de la manière suivante : l'angle θ_i doit être inférieur à $\pi/2$ ($\theta_i < \pi/2$). Dans l'exemple illustré dans la Figure 4-3, seuls les points FDP_1 et FDP_2 satisfont à ce critère.

c. Calcul des influences de chaque point de contrôle sur un point ordinaire

L'influence des points de contrôle sur les points ordinaires est mesurée en termes de distances surfaciques et d'angles. Le calcul de la somme pondérée, d, entre les points de contrôle influents et le point ordinaire P s'effectue comme suit :

$$d = \frac{\sum_{i=1}^{i=n} d_i + \cos\theta_i}{\sum_{i=1}^{i=n} \cos\theta_i} \quad (4.5)$$

Où :

n est le nombre de points de contrôle influençant le point ordinaire P.

θ_i sont les angles décrits précédemment.

La valeur de d est utilisée pour la normalisation des poids relatifs des points de contrôle pour le point ordinaire P.

Le point ordinaire P a un poids associé au déplacement d'un point de contrôle FP_i qui est inversement proportionnel à sa distance par rapport à ce point de contrôle FP_i.

Le calcul de ce poids $W_{i,P}$, qui est associé au point ordinaire P par le point de contrôle FP_i, est défini comme suit :

$$W_{i,P} = \sin\left(\frac{\pi}{2}\left(1 - \frac{d_i}{d}\right)\right) \quad (4.6)$$

Ainsi, l'influence d'un point de contrôle sur un point ordinaire peut être déterminée en tenant compte de l'angle entre les deux. Cette introduction des angles permet de réduire l'influence de la topologie du maillage, en agissant comme des normalisateurs. Dans le cas contraire, l'influence serait fortement dépendante de la topologie.

4.2.2.2 Déformation

Le modèle est désormais prêt pour l'animation, après que les poids de chaque point de maillage ont été calculés. Si un point de contrôle est déplacé au cours de l'animation, le déplacement de tous les points de maillage influencés par ce point doit être recalculé. A noter que dans un tel cas, il faut tenir compte des effets du déplacement simultané de deux ou plusieurs points de contrôle, qui peuvent influencer le même point ordinaire. La somme de tous les déplacements du point P causés par les points de contrôle environnants est alors calculée.

Le déplacement D_p d'un point P est calculé comme suit :

$$D_p = \frac{\sum_{i=1}^{i=n} \frac{W_{i,P} * D_i}{d_{i,P}^2}}{\sum_{i=1}^{i=n} \frac{W_{i,P}}{d_{i,P}^2}} \qquad (4.7)$$

Avec :

n représente le nombre de points de contrôle FP_i influençant le point ordinaire P.

D_i est le déplacement spécifique du point de contrôle FP_i.

$W_{i,P}$ est le poids du point P associé au point de contrôle FP_i, calculé dans l'étape précédente.

$d_{i,P}$ représente la distance surfacique entre P et FP_i.

Dans le cadre des calculs d'animation du modèle, cette opération est effectuée pour chaque image.

4.2.3 Discussion

L'approche par points de contrôle est couramment utilisée dans l'animation faciale pour résoudre trois problèmes majeurs :

- Comme la paramétrisation d'un visage est une étape longue et ardue, de quelle manière peut-on aider le graphiste dans sa réalisation ; voir comment rendre ce travail accessible à un utilisateur novice ?
- Comment peut-on éviter les effets indésirables de déformation qui résultent du fait que certains sommets du maillage appartiennent à plusieurs zones d'influence ?
- Comment peut-on minimiser l'intervention humaine ?

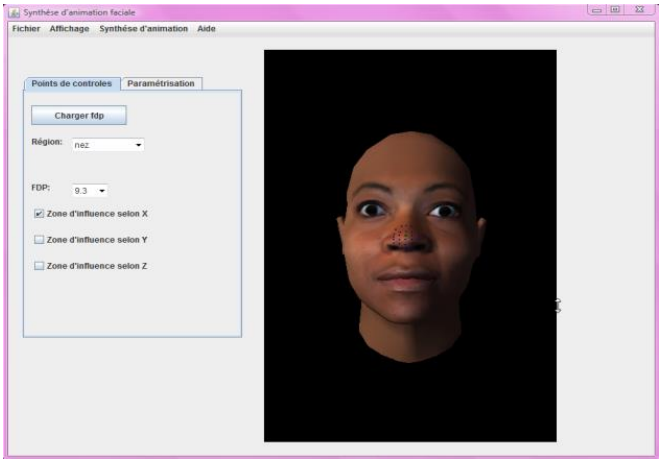


Figure 4-4 : L'interface graphique développée.

Nous avons dû faire face à plusieurs défis majeurs lors de la mise en œuvre de cette méthode, la Figure 4-4 présente l'interface graphique développée. Tout d'abord, la sélection et la correspondance des points de contrôle avec le maillage du modèle ont été résolues grâce à l'utilisation de l'outil XFACE. Ensuite, le calcul des zones d'influence des

points de contrôle dans la région de la bouche a posé un problème important en raison de la topologie irrégulière de cette partie du modèle. Malgré les solutions adoptées pour surmonter ces limitations, les principales contraintes de cette méthode restent la modélisation du visage et la détection des points de contrôle.

4.3 Transfert d'animation faciale par points caractéristiques

Une autre approche active dans le domaine de l'animation faciale est la méthode prometteuse du transfert d'animation, qui consiste à transférer les expressions faciales d'un visage existant vers un nouveau visage, plutôt que de créer une animation entièrement nouvelle à partir de zéro. Ce processus comprend plusieurs étapes qui sont la détection du visage, la localisation et le suivi des points de contrôle dans le visage source, la localisation des points de contrôle dans le visage cible et le transfert des mouvements des points de contrôle du visage source vers le visage cible. Dans ce travail (Bordjiba & Merouani, 2013), nous utilisons les fonctions à base radiale (RBF) pour transférer les expressions faciales de visages réels (2D) vers des visages virtuels. Dans notre cas, les entrées sont le visage source à l'état neutre et avec diverses expressions, capturées avec une webcam, et le visage cible à l'état neutre, la sortie est le visage cible (l'objet virtuel) avec les expressions du visage source transférées. La Figure 4-5 illustre la séquence des étapes de la méthode proposée.

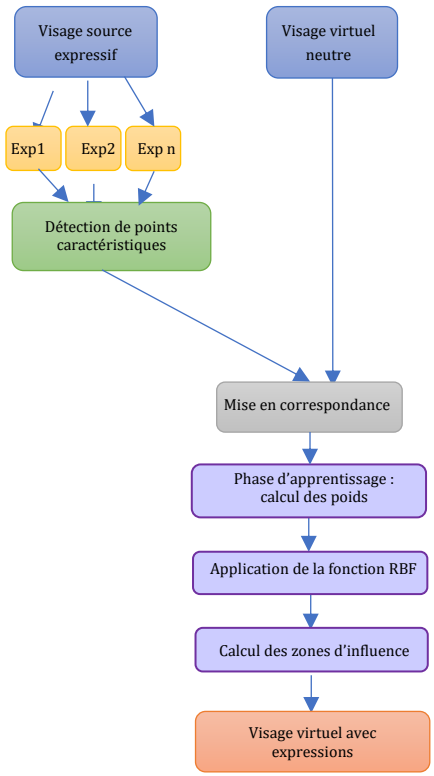


Figure 4-5: Architecture de la méthode de transfert proposée.

Dans notre étude, nous avons abordé deux problématiques principales. Tout d'abord, nous avons étudié la localisation et le suivi des points de contrôle d'un visage expressif à partir d'une vidéo réelle(Bordjiba & Merouani, 2015). Ensuite, nous nous

sommes penchés sur la localisation des points de contrôle sur un visage cible en les associant aux points caractéristiques du visage expressif (Bordjiba et al., 2016, 2018).

4.3.1 Détection et suivi des points caractéristiques du visage

La détection de points de repère faciaux est une technique d'analyse d'images qui permet d'identifier et de localiser des points spécifiques sur le visage d'une personne, tels que les coins des yeux, le bout du nez, les coins de la bouche, etc. Ces points, également appelés points de repère ou points d'intérêt, sont des indicateurs clés pour l'analyse des expressions faciales, la reconnaissance faciale, la synthèse d'animation faciale et d'autres applications liées au visage. Ces points, également appelés "points de repère" ou "points d'intérêt", sont des indicateurs clés pour l'analyse des expressions faciales, la reconnaissance faciale, la synthèse d'animation faciale et d'autres applications liées au visage.

La détection des points caractéristiques du visage est une étape importante de l'animation faciale. Cette technique permet de localiser des points d'intérêt sur le visage, afin de décrire la position, l'orientation et la forme du visage. Ces informations sont utilisées pour créer une représentation numérique du visage, qui peut être utilisée pour l'animation faciale en temps réel ou pour transférer l'animation faciale d'un visage source à un visage cible. La détection précise des points caractéristiques du visage est essentielle pour une animation faciale réaliste et naturelle. Dans l'animation faciale, les points de caractéristiques faciales sont généralement définis comme des points clés du visage utilisés pour contrôler les mouvements et les expressions faciales d'un personnage animé. Ces points sont également appelés "points de contrôle" ou "points d'animation".

Notre étude présente une méthode pour la détection et le suivi des points caractéristiques du visage dans une vidéo réelle ou une séquence d'images (Bordjiba et al., 2018). La détection est réalisée sur la première image de la vidéo en utilisant l'algorithme de l'histogramme cumulatif, qui sert également à effectuer le suivi des points détectés. Nous avons également testé l'efficacité de l'implémentation pyramidale de l'algorithme de Lucas et Kanade pour suivre les points détectés, afin de comparer nos résultats.

4.3.1.1 Méthode proposée pour la détection et le suivi

L'objectif de ce travail est de repérer et de suivre les points clés du visage dans une vidéo, afin de les utiliser comme une étape initiale dans diverses applications telles que : la reconnaissance faciale, la reconnaissance des émotions, l'animation de personnages virtuels et la simulation de leurs expressions, entre autres.

La méthode proposée permet de traiter soit une séquence d'images issues d'une base de données, soit une vidéo réelle en entrée. La première image est utilisée pour

détecter le visage, ses composants et ses points clés, tandis que le reste des images sont utilisées pour suivre ces points tout au long de la vidéo. Il est important de noter que trois conditions doivent être respectées pour un fonctionnement optimal :

- Le visage de la personne doit être en face de la caméra.
- Le sujet doit débuter avec une expression faciale neutre pendant la phase d'initialisation.
- Les conditions d'éclairage doivent être stables.

Au final, le résultat consiste en une vidéo avec les points caractéristiques d'un visage détectés dans chaque image. La Figure 4-6 illustre les principales étapes de l'architecture proposée.

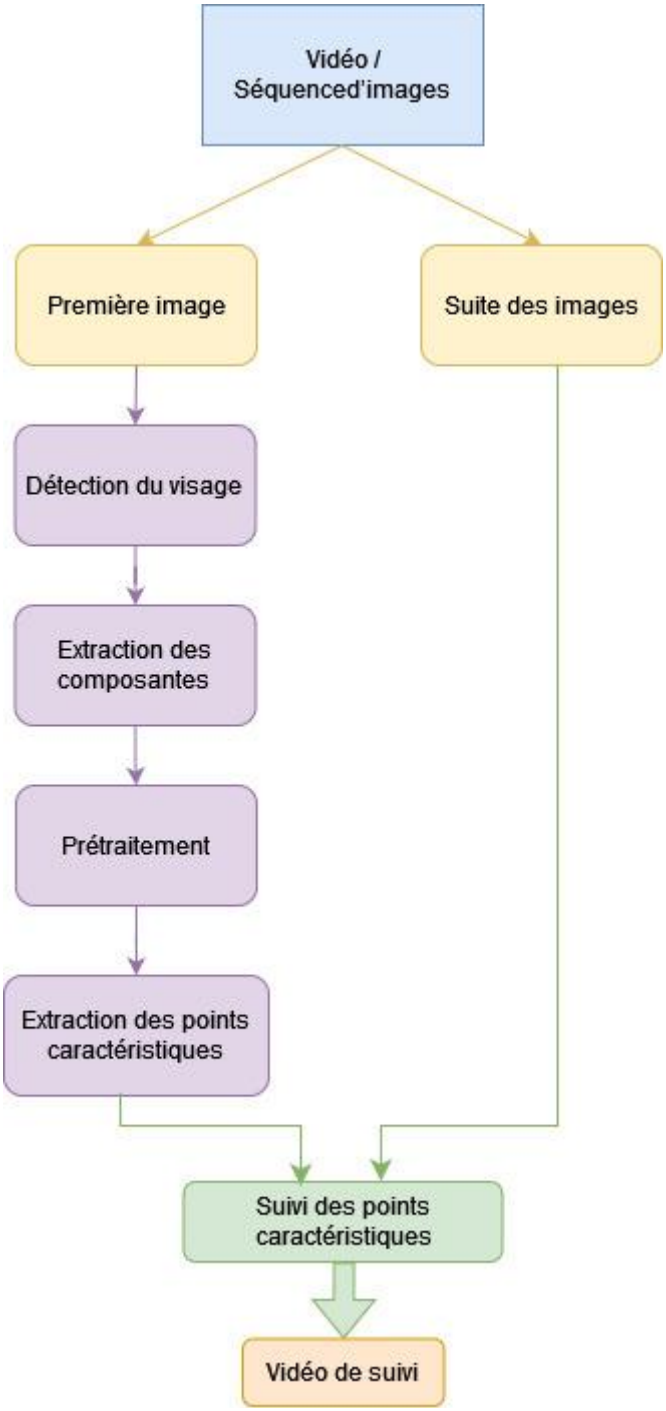


Figure 4-6 : Architecture de la méthode de détection et suivi des points caractéristiques proposée.

Nous avons opté pour la méthode Viola & Jones pour la détection des visages en raison de ses performances éprouvées. En ce qui concerne l'extraction des points caractéristiques, une technique basée sur les histogrammes cumulés a été mise en œuvre. Enfin, pour le suivi des points, nous avons mis en œuvre deux méthodes distinctes : les histogrammes cumulatifs et le flux optique. Plus de détail sur chacune de ces étapes est donné dans ce qui suit.

a. Extraction des points caractéristiques par histogrammes cumulées

Après la détection du visage, En appliquant l'algorithme de détection de visage Viola-Jones (Viola & Jones, 2004), il est nécessaire d'extraire les parties importantes qui sont les yeux, les sourcils, le nez et la bouche, à l'aide de boîtes englobantes. Pour cela, la région du visage détectée est d'abord recadrée, puis divisée verticalement en parties supérieure, moyenne et inférieure.

Conformément au schéma de la structure frontale du visage humain, les sourcils et les yeux, le nez et la bouche sont situés respectivement dans les parties supérieure, moyenne et inférieure de l'image du visage. Une fois encore, la partie supérieure est divisée horizontalement en segments gauche et droit pour isoler respectivement le sourcil droit et l'œil droit, ainsi que le sourcil gauche et l'œil gauche. Comme le montre la Figure 4-7, les dimensions ($H_{œil}$, $W_{œil}$, $H_{sourcil}$, $W_{sourcil}$, H_{nez}, W_{nez} , H_{bouche} , W_{bouche}) des rectangles englobant les yeux , les sourcils, le nez et la bouche sont définies par les relations suivantes (Kumar et al., 2012), à partir du visage et de ses dimensions (*H :* largeur du cadre du visage, *W :* hauteur du cadre du visage) :

- (a) sourcils droite (taille : 0.375W×0.12H)
- (b) sourcils gauche (taille : 0.375W×0.12H)
- (c) œil droite (taille : 0.375W×0.25H)
- (d) œil gauche (taille : 0.375W×0.25H)
- (e) Nez (taille : 0.50W×0.19H)
- (f) Bouche (taille : 0.50W×0.16H)

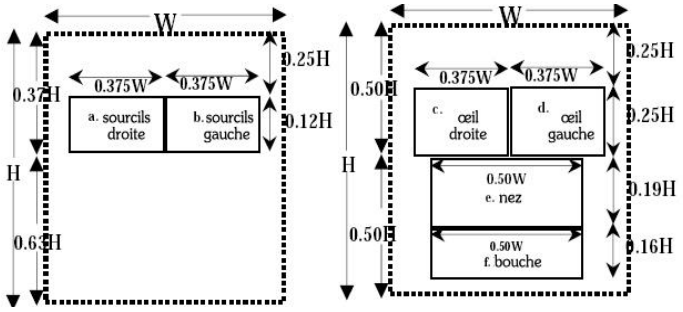


Figure 4-7 : Localisation et taille des six boîtes englobantes d'une image de visage.

Les différentes parties du visage sont soumises à un prétraitement spécifique. Pour la région de la bouche, un filtre gaussien est appliqué, tandis que pour la région du nez, une égalisation de l'histogramme est utilisée. Aucun prétraitement n'est appliqué aux autres parties du visage.

Pour détecter les points caractéristiques du visage, l'algorithme de l'histogramme cumulatif est appliqué à chaque partie du visage (yeux, bouche et nez). Cette méthode permet d'extraire la localisation de douze points caractéristiques, soit huit points pour les sourcils droit et gauche, les yeux droit et gauche, deux points pour le nez et deux points pour la bouche. Pour ce faire, plusieurs étapes sont suivies pour chaque boîte de délimitation :

- Convertir l'image en niveaux de gris.
- Calculer la probabilité d'occurrence de chaque niveau de gris.
- Calculer l'histogramme cumulatif pour chaque pixel de niveau de gris.
- Binariser l'image.
- Effectuer une recherche linéaire sur l'image binarisée pour détecter les premiers pixels blancs dans les boîtes englobantes de chaque partie du visage (sourcil droit, œil droit, sourcil gauche, œil gauche, nez et bouche).

b. Suivi des points caractéristiques du visage

Deux approches ont été expérimentées pour le suivi des points caractéristiques du visage :

- Suivi par histogramme cumulatif : cette méthode a fait preuve d'une efficacité considérable dans la détection des points caractéristiques, raison pour laquelle nous avons décidé de l'utiliser pour le suivi. Elle ne s'applique qu'à une fenêtre entourant le point caractéristique.
- Suivi par flux optique : nous avons opté pour l'implémentation pyramidale de l'algorithme de Lucas-Kanade.

4.3.1.2 Résultats et discussion

Nous utilisons la base de données d'images étendue de Cohn-Kanade, composée de 123 visages d'adultes, qui a été prise par deux caméras synchronisées Panasonic AG-7500. Les participants étaient âgés de 18 à 50 ans, 69% de femmes, 81% d'euro-américains, 13% d'afro-américains et 6% d'autres groupes. Chaque vidéo commence et se termine par un visage neutre. Les séquences d'images pour une vue de face et une vue de 30 degrés ont été numérisées dans des matrices de 640x490 ou 640x480 pixels avec un niveau de gris de 8 bits ou une couleur de 24 bits. Tous les détails de cette base de données sont donnés dans (Lucey et al., 2010).

Dans les figures suivantes, nous présentons quelques résultats obtenus des étapes de détection et de suivi :

Figure 4-9 : Les points caractéristiques d'un visage mal détectés.

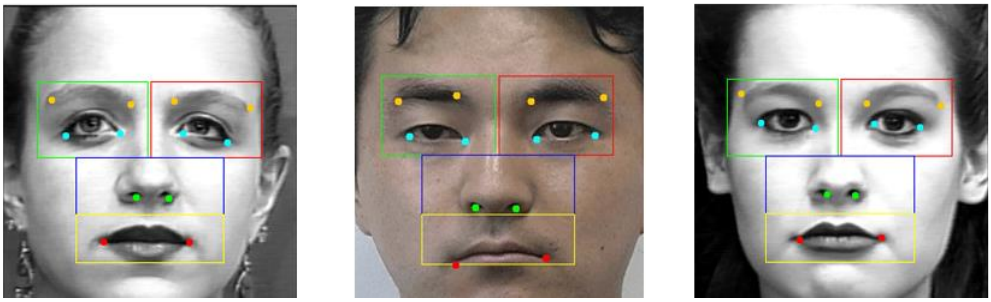


Figure 4-8 : Les points caractéristiques d'un visage correctement détectés.

Les figure 4-8 et 4-9 illustre des résultats de bonne et de mauvaise détection de points caractéristiques. Le Tableau 4-1 présente les de taux de bonne et mauvaise détection des différents points du visages.

Tableau 4-1 : taux de détection des points de visage par histogrammes cumulatifs.

Les points de visage	**Taux de bonne détection**		**Taux de mauvaise détection**	
	Points droit	Points gauche	Points droit	Points gauche
Les points du sourcil droit	58,37 %	78,03 %	42,63 %	21,97 %
Les points du sourcil gauche	75,47 %	67,99 %	24,53 %	32,01 %
Les points du l'œil droite	69,64 %	75,98 %	25 ,36 %	37,02 %
Les points de l'œil gauche	64,33 %	72,02 %	35,67 %	27,98 %
Les points du nez	96,63 %	79,07 %	3,36 %	10,92 %
Les points de la bouche	93,27 %	78,15 %	6,72 %	21,84 %

En ce qui concerne le suivi, Les deux méthodes de suivi ont donné des résultats acceptables. Lors de notre analyse, nous avons constaté que les résultats peuvent être divisés en trois catégories distinctes pour chaque méthode de suivi :

1. Bon suivi associé à une bonne détection : dans ce cas, le suivi est précis et la détection est correcte pour les séquences d'images et les vidéos.

2. Mauvais suivi associé à une bonne détection : certains cas présentent des difficultés liées à la luminosité, à la taille du voisinage, à la nature du mouvement et à la taille du visage, ce qui entraîne des résultats insatisfaisants pour le suivi dans les deux méthodes.

3. Mauvais suivi associé à une mauvaise détection : cela peut se produire lorsque les coordonnées des points détectés dans la première image ne sont pas correctes ni précises, ce qui conduit à de mauvais résultats.

Les figures suivantes, de Figure 4-10 à Figure 4-12, présentent des exemples de chaque catégorie par la méthode de suivi par histogrammes cumulatifs.

Figure 4-12 : Bon suivi associé à une bonne détection par l'histogramme cumulé.

Figure 4-11 : Mauvais suivi associé à une bonne détection par l'histogramme cumulé.

Figure 4-10 : Mauvais suivi associé à mauvaise détection par l'histogramme cumulatif.

Ainsi, les figures, de Figure 4-13 à Figure 4-15, illustrent ceux de la méthode de suivi par flux optique.

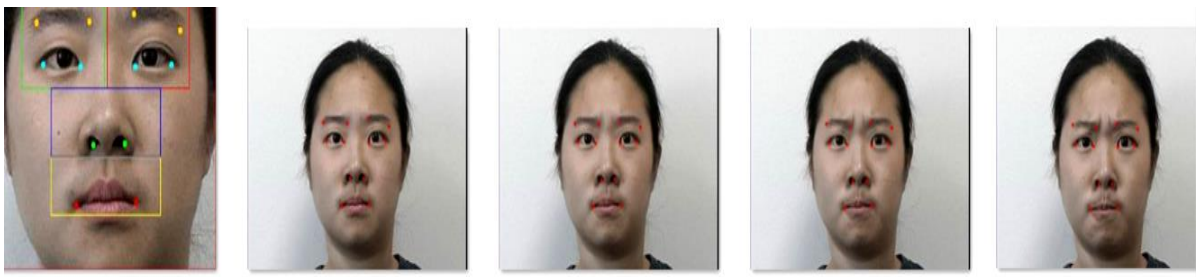


Figure 4-13 : Bon suivi associé à une bonne détection par flux optique.

Figure 4-15 : Mauvais suivi associé à une bonne détection par flux optique.

Figure 4-14 : Mauvais suivi associé à mauvaise détection par flux optique.

En général, le suivi par flux optique a donné de meilleurs résultats que l'histogramme cumulé. Cependant, il convient de noter que le suivi par flux optique est plus coûteux en termes de temps de calcul que le suivi par histogramme cumulé.

4.3.2 Correspondance des points caractéristiques de deux visages différents

Afin de transférer les expressions faciales d'un maillage source à la géométrie d'un maillage cible, il est nécessaire de trouver un certain nombre de points communs sur chaque maillage, appelés points caractéristiques. Chaque point caractéristique d'un maillage doit avoir son équivalent sur l'autre maillage. Par exemple, le point situé au coin de l'œil droit sur le maillage source doit être placé au même endroit sur le maillage cible. La mise en correspondance est une étape cruciale dans diverses applications, notamment la reconstruction tridimensionnelle, où elle consiste à localiser les projections des mêmes points de la scène observée dans deux images différentes. Elle est également importante dans le transfert d'animations faciales, où il est nécessaire d'établir une correspondance entre les points caractéristiques de deux visages différents. Nous nous concentrons sur la localisation des points de contrôle sur un visage cible en faisant correspondre les points caractéristiques d'un visage expressif. Pour résoudre ce problème, nous proposons d'utiliser un algorithme évolutionnaire, à savoir l'algorithme génétique.

Plusieurs recherches ont été menées sur ce sujet et les méthodes proposées peuvent être classées en deux catégories : les méthodes locales et les méthodes globales. Si l'on considère la mise en correspondance comme un problème d'optimisation, les algorithmes génétiques peuvent constituer une solution efficace. Ces algorithmes explorent l'espace des solutions en utilisant une population initiale de solutions et en appliquant des opérations de sélection, de croisement et de mutation.

L'algorithme génétique (AG) (Holland, 1992) est un algorithme de recherche basé sur les mécanismes de la sélection naturelle et de la génétique. Il combine une stratégie de "survie du plus apte" avec un échange d'informations aléatoire mais structuré. Pour un problème dont la solution est inconnue, un ensemble de solutions possibles est créé aléatoirement ; cet ensemble est appelé population. Les caractéristiques (ou variables à déterminer) sont ensuite utilisées dans les séquences de gènes qui seront combinées avec d'autres gènes pour former des chromosomes puis des individus (Reynès, 2007).

4.3.2.1 Méthode proposée pour la mise en correspondance

L'objectif de ce travail est de détecter les points caractéristiques d'un visage et de les associer aux points caractéristiques correspondants d'un autre visage à l'aide d'un algorithme génétique. Pour ce faire, nous allons d'abord détecter les points caractéristiques du premier visage à l'aide de la méthode de détection décrite dans la section précédente, puis utiliser l'algorithme génétique pour rechercher les points caractéristiques correspondants sur le second visage sur la base de critères de similarité. Cette approche devrait permettre d'effectuer des tâches telles que la reconstruction tridimensionnelle de visages et la création de caricatures ou d'animations faciales.

La méthode proposée utilise deux images comme données d'entrée. Ces images, issues d'une base de données ou représentant deux visages réels, peuvent représenter le même visage avec deux expressions différentes ou deux visages distincts.

Tout d'abord, la première image est utilisée pour détecter le visage, extraire les composantes faciales et enfin extraire les points caractéristiques. Ensuite, la deuxième image est utilisée pour détecter le visage et faire correspondre les points caractéristiques avec le premier visage. Les principales étapes de l'architecture proposée sont illustrées à la Figure 4-16.

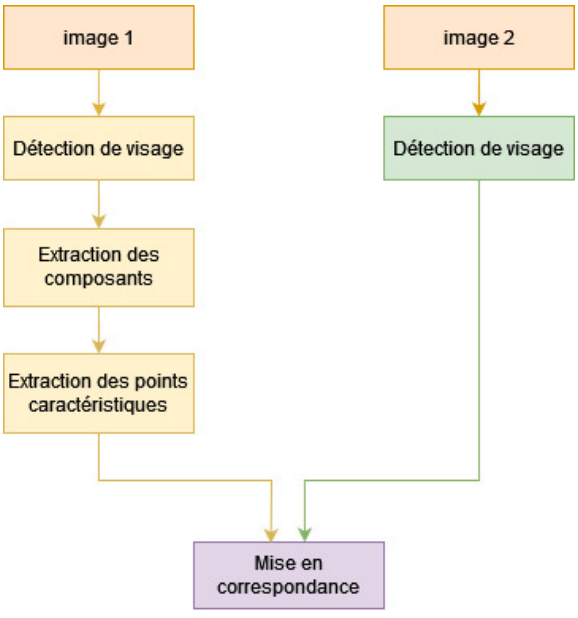


Figure 4-16 : Architecture de la méthode proposée de mise en correspondance.

a. Extraction des points caractéristiques du visage source

La première étape est l'extraction des points caractéristiques du visage source. Tout d'abord, pour la détection des visages, nous avons utilisé la méthode de Viola et Jones (Viola & Jones, 2004), qui utilise les descripteurs de Haar. Ensuite, les composantes du visage (sourcils, yeux, nez et bouche) sont extraites. Enfin, nous utilisons la méthode des histogrammes accumulés pour extraire les points caractéristiques de chaque boîte englobante. La méthode est détaillée dans la section 4.3.1.

b. Mise en correspondance par algorithme génétique

La deuxième étape consiste à rechercher sur le visage cible les points correspondant aux points caractéristiques extraits du premier visage à l'aide d'un algorithme génétique, dont le schéma général est illustré dans la Figure 4-17. Celui-ci est utilisé pour explorer l'espace des solutions, en partant d'une population initiale de solutions et en appliquant les opérateurs de sélection, de croisement et de mutation.

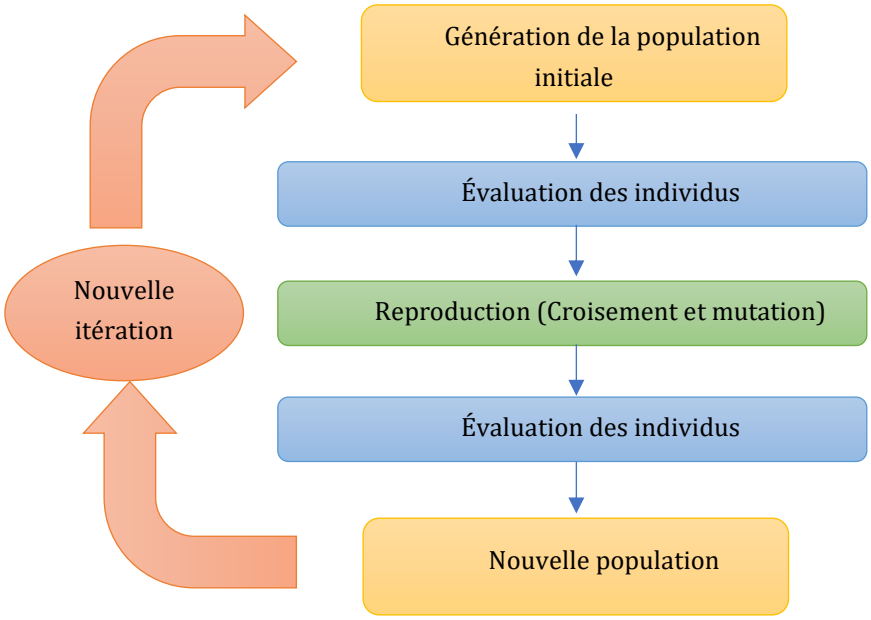


Figure 4-17: Schéma général de l'algorithme génétique.

La population initiale est générée aléatoirement dans l'espace de recherche. Pour ce dernier, on distingue deux cas : le visage et les boîtes englobantes.

La fonction objective est cruciale pour l'algorithme génétique. Dans notre cas, nous avons opté pour la corrélation afin de mesurer la similarité entre deux fenêtres centrées sur les points caractéristiques.

La corrélation normalisée est utilisée dans les opérateurs de sélection et de croisement de l'algorithme génétique. Pour la calculer, nous avons utilisé la formule suivante :

Le meilleur individu, correspondant au point dont la corrélation est comprise entre 0,5 et 1, est sélectionné comme point final.

$$Corr = \frac{\sum i \sum j \left[I(x+i, y+j) - \hat{I}(x,y) - I'^{(x'+i,y'+j)} - \hat{I}'^{(x',y')} \right]^2}{\sqrt{\sum i \sum j \left[I(x+i, y+j) - \hat{I}(x,y) \right]^2 * \sum i \sum j \left[I'^{(x'+i,y'+j)} - \hat{I}'^{(x',y')} \right]^2}} \quad (4.8)$$

Où : x et y sont les coordonnées du point caractéristique du premier visage.

x' et y' sont les coordonnées du point candidat du deuxième visage.

Les opérateurs génétiques utilisés sont la sélection, le croisement et la mutation, chacun de ces opérateurs est décrit dans ce qui suit :

a. Sélection

Nous avons appliqué l'opérateur à la moitié de la population initiale, en utilisant la méthode de sélection par tournoi.

Cette méthode consiste à sélectionner en plusieurs tours, à partir d'une paire d'individus, celui qui a le meilleur score de corrélation est choisi pour passer à l'étape suivante. Cette méthode assure une meilleure diversité génétique et une sélection plus robuste des individus les plus performants.

b. Croisement :

Nous avons appliqué l'opérateur à la moitié de la population initiale en utilisant le croisement linéaire de deux points sélectionnés, $P_1(x_1, y_1)$ et $P_2(x_2, y_2)$. Ainsi, ce processus permet de générer trois fils selon la formule suivante (Davis, 1991) :

$$C_1 = 0{,}5 * P_1 + 0{,}5 * P_2 \quad (4.8)$$
$$C_2 = 1{,}5 * P_1 - 0{,}5 * P_2 \quad (4.9)$$
$$C_3 = -0{,}5 * P_1 + 1{,}5 * P_2 \quad (4.10)$$

Nous choisissons ensuite les deux meilleurs points en fonction de leurs résultats de corrélation. Le point croisé doit être inclus dans l'espace de recherche (cadre du visage, boîte englobante).

c. Mutation :

Cet opérateur est rarement utilisé, nous l'avons appliqué une fois toutes les dix itérations, donc avec une probabilité de 0,1.

Le point muté doit être inclus dans l'espace de recherche (cadre de visage, boîte englobante). Ainsi, un point P(x, y) est choisi aléatoirement pour la mutation, puis x ou y est multiplié par 1,5 à tour de rôle.

Le Tableau 4-2 résume les différents paramètres utilisés pour l'algorithme génétique.

Tableau 4-2 : Paramètres de l'algorithme génétique.

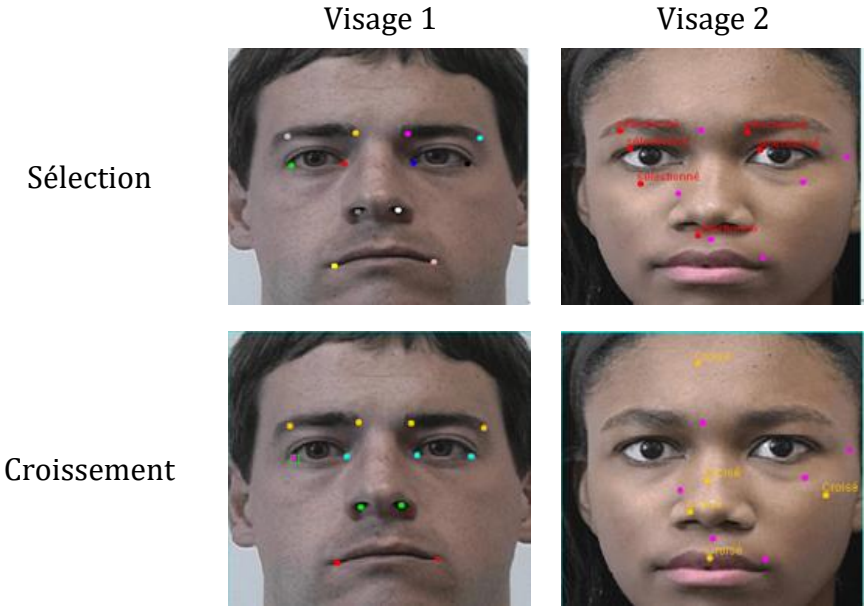
Paramètre	Valeur du paramètre

Paramètres dimensionnels	• Taille de la population : 12 points/22 points • Probabilité de croissement : 0,5 • Probabilité de mutation : 0,1
Fonction objective	• Corrélation normalisée
Codage	• Réel
Opérateurs	• Sélection par tournois • Croisement linéaire • Mutation
Condition d'arrêt	• Corrélation entre 0.5 et 1 • Nombre d'itérations = 1000

4.3.2.2 Résultats et discussion

La base de données d'images Cohn-Kanade étendue (Lucey et al., 2010), déjà décrite dans la section 4.3.1.2, est utilisée pour évaluer la méthode de mise en correspondance proposée.

Les résultats de la recherche de points correspondants dans la deuxième image se sont avérés insatisfaisants, les tests ayant été effectués avec une population initiale de 12 points (la Figure 4-18 présente les résultats de l'application des différents opérateurs de l'algorithme génétique), et même après avoir augmenté la taille de la population initiale à 22 points (le résultat est présenté dans la Figure 4-19). Nous avons constaté qu'environ 50 % seulement des points correspondants étaient trouvés.



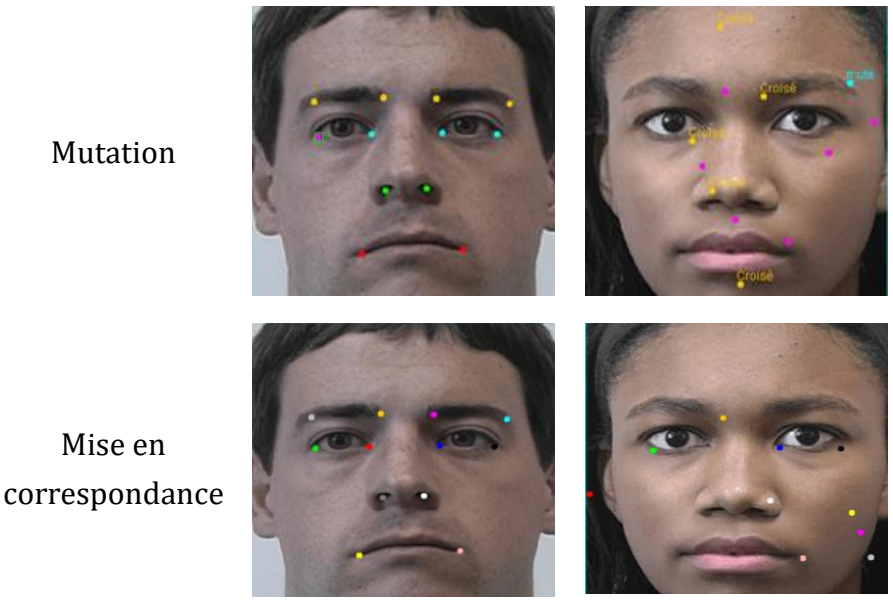


Figure 4-18 : Résultat d'application des opérateurs de l'algorithme génétique pour une population initiale de 12 points dans le cadre de visage.

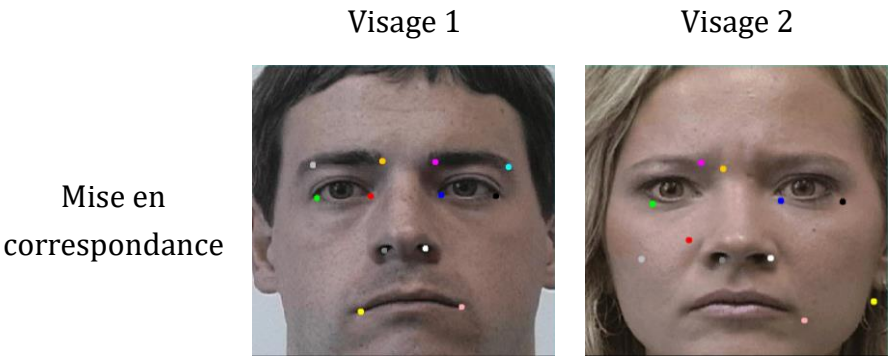
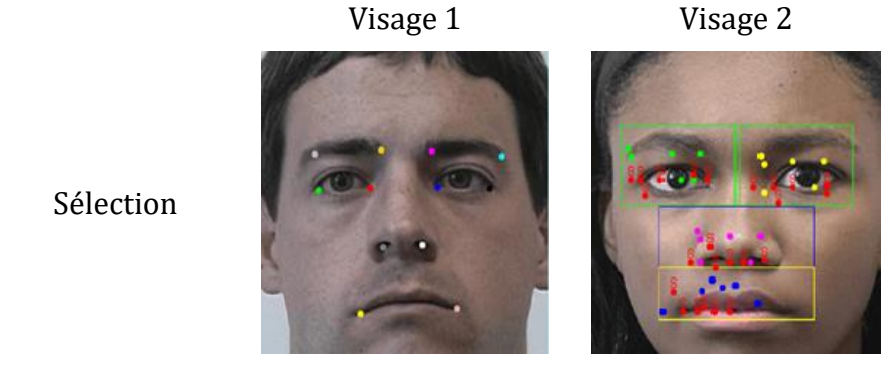


Figure 4-19 : Résultat de mise en correspondance pour une population initiale 22 points dans le cadre de visage.



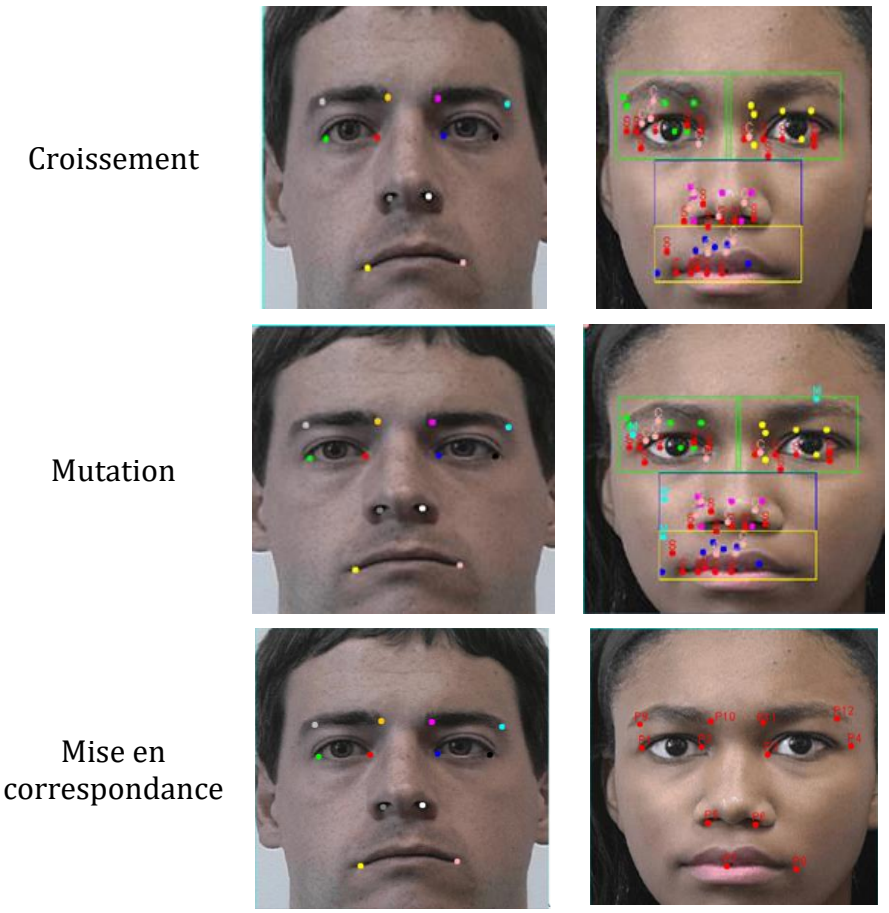


Figure 4-20 : Résultat d'application des opérateurs de l'algorithme génétique pour une population initiale de 12 points dans les boites englobantes.

Pour améliorer les résultats, nous avons proposé de modifier l'espace de recherche. Au lieu de rechercher des points correspondants dans l'ensemble du cadre du visage, nous nous sommes limités aux boîtes englobantes. Après cette modification, les résultats obtenus étaient nettement meilleurs, comme le montrent la Figure 4-20.

4.4 Conclusion

Nous avons examiné deux approches très intéressantes parmi les méthodes classiques d'animation faciale présentées dans le chapitre 3. La première approche, qui a connu un grand succès dans les premières années de l'animation faciale grâce aux travaux de F. Parke, concerne la synthèse d'animation faciale. La seconde approche, quant à elle, vise à transférer les expressions faciales d'un visage source à un visage cible, sans utiliser de marqueurs, en se basant sur une méthode automatique de détection des points caractéristiques, ainsi que sur une mise en correspondance par un algorithme évolutionnaire, l'algorithme génétique.

Malgré ses avantages, le transfert d'animation faciale est confronté à une difficulté majeure : la similarité entre le modèle cible et la personne originale est essentielle, tout comme le maintien de l'harmonie de l'expression source. Par conséquent, les techniques

de transfert doivent être capables d'adapter les déformations du visage source à la morphologie du visage cible.

Cela nous a incités à envisager de nouvelles approches qui nous permettront de concevoir une méthode d'animation faciale basée sur la reconnaissance des expressions faciales, qui servira à générer de nouveaux visages avec cette expression transférée.

Chapitre 5 : Reconnaissance des expressions faciales

5.1 Introduction

La communication entre humains face à face est incontestablement le mode de communication le plus naturel qui soit pour nous, et pourtant elle est immensément riche et nuancée, et difficile à comprendre entièrement. Lors de chaque rencontre avec d'autres personnes, nous établissons automatiquement plusieurs conclusions sur la base de signaux non verbaux, tels que leur posture, leurs mouvements et leurs expressions faciales.

De la même manière, les personnages animés dans les films, les jeux ou les environnements virtuels sont interprétés de manière complexe dans leur langage corporel et leurs expressions faciales. Ces expressions faciales, en particulier, sont censées influencer les émotions que nous attribuons à une personne ou à un personnage virtuel. Un visage d'un personnage humain animé enverra toujours un message sur les émotions du personnage, il est donc indispensable de comprendre comment les gens perçoivent les expressions faciales afin de contrôler ce message.

5.2 La reconnaissance des émotions par les expressions faciales

Selon l'Organisation Mondiale de la Santé (OMS) l'émotion peut être définie comme (D. Cohen, 2019) : « Une expérience de plaisir ou de douleur indissociable du caractère attractif ou répulsif de certains événements vécus ou appréhendés. L'épisode émotionnel qui en découle (…) se caractérise par des réactions physiologiques, motrices et subjectives se développant de manière synchrone ».

Selon Darwin (Darwin & Prodger, 1998), l'expérience émotionnelle se manifeste par des modifications corporelles spécifiques dans la posture, les gestes, le ton de la voix et surtout dans le visage ; on parle alors d'expressions faciales émotionnelles.

La reconnaissance des émotions est activement explorée dans le cadre de la recherche sur la vision par ordinateur, l'apprentissage automatique et l'apprentissage profond. Il est désormais possible de construire des systèmes intelligents capables de reconnaître les émotions d'un humain. Plusieurs termes tels que les signaux d'électroencéphalographie (EEG), les gestes, le ton de la voix, les expressions faciales ont

un impact puissant sur l'identification des émotions chez un être humain (Canedo & Neves, 2019).

Les expressions faciales sont des formes de communication non verbale qui donnent des indices sur les émotions humaines et la reconnaissance des expressions faciales consiste à classer les expressions de visages dans diverses catégories (les expressions de base) telles que la colère, la peur, la surprise, la tristesse, la joie, etc.

5.3 Processus de reconnaissance des expressions faciales

Au cours des deux dernières décennies, la reconnaissance des expressions faciales est devenue un sujet important et stimulant dans plusieurs domaines tels que la vision par ordinateur, l'intelligence artificielle et l'interaction homme-machine.

En général, les travaux traditionnels sur la reconnaissance des expressions faciales sont réalisés après l'acquisition du visage en deux étapes principales après la détection du visage et de ses composants, comme le montre la Figure 5-1. Les caractéristiques des expressions sont extraites pour représenter l'image ou la vidéo donnée, puis une étape de classification est réalisée pour reconnaître les différentes expressions à partir des caractéristiques extraites.

La plupart des méthodes conventionnelles s'appuient sur des caractéristiques fabriquées à la main ou sur un apprentissage superficiel, comme les réseaux de neurones [3], les réseaux bayésiens (BN) [4], les machines à vecteurs de support (SVM) [5], Adaboost [6] et Random Forest [7].

Figure 5-1 : Schéma d'un système FER conventionnel.

5.4 Techniques de reconnaissance des expressions faciales basée sur l'apprentissage profond

Pour les systèmes conventionnels de reconnaissance des expressions faciales, l'extraction des caractéristiques faciales est une étape très cruciale, qui affecte la décision de classification finale.

En général, il convient de noter que les méthodes employées pour extraire ces caractéristiques élaborées à la main utilisent des données étiquetées dans le contexte de l'apprentissage supervisé. En outre, ces caractéristiques artisanales, telles que la représentation du LBP et des ondelettes de Gabor, capturent des informations de bas niveau sur les images faciales, à l'exception de la représentation de haut niveau des images faciales (X. Zhao et al., 2015a). En plus de cela, les approches conventionnelles nécessitent relativement moins de puissance de calcul et de mémoire que les approches

basées sur l'apprentissage profond. Pour ces raisons, ces approches sont encore à l'étude pour une utilisation dans les systèmes embarqués en temps réel en raison de leur faible complexité de calcul et de leur grande précision (Suk & Prabhakaran, 2014).

Les méthodes d'apprentissage profond qui ont été introduites pour la reconnaissance automatique des expressions faciales dans des images statiques ont tenté de résoudre les principaux problèmes de la FER, qui varient des problèmes liés aux performances en temps réel, à l'amélioration de la précision de la reconnaissance, au problème réel de l'occlusion partielle du visage et à la mauvaise généralisation lors de l'entraînement des CNN sur de petits échantillons de données de FER (Saurav et al., 2022). Avec ces approches, la dépendance vis-à-vis des modèles basés sur la physique du visage et d'autres techniques de prétraitement est considérablement réduite en permettant un apprentissage "de bout en bout" dans le pipeline directement à partir des images d'entrée (Walecki et al., 2017). Les approches d'apprentissage profond de la FER se concentrent fondamentalement sur deux méthodes : la méthode basée sur le DBN (deep belief network) et la méthode basée sur le CNN (convolution neural network). Cependant, au cours de ces dernières années, les réseaux neuronaux à convolution se sont imposés comme l'approche la plus populaire parmi les chercheurs dans ce domaine (Ko, 2018a), (X. Zhao et al., 2015b).

Les auteurs de (X. Zhao et al., 2015b) ont proposé d'intégrer l'avantage des DBNs (Deep belief networks) pour l'apprentissage non supervisé des caractéristiques avec l'avantage de la classification des MLPs (multi-layer perceptron). Tout d'abord, ils ont utilisé les DBN pour apprendre les pixels bruts des images d'expressions faciales et pour obtenir un niveau plus élevé de caractéristiques abstraites. Ensuite, en utilisant les résultats de l'apprentissage des DBN, ils ont initialisé un modèle MLP pour effectuer la classification de l'expression faciale.

En utilisant un réseau profond est basé sur deux modèles différents, les auteurs de (Jung et al., 2015) ont extrait les caractéristiques d'apparence temporelles des séquences d'images et les caractéristiques géométriques des points de repère temporels du visage. Ces deux modèles sont combinés pour améliorer la performance de la reconnaissance des expressions faciales, en utilisant une nouvelle méthode d'intégration. Le réseau proposé a été validé sur les bases de données CK+ et Oulu-CASIA, et ils ont montré que les performances obtenues sont supérieures aux autres méthodes de pointe.

Ebrahimi Kahou et al. (Ebrahimi Kahou et al., 2015) ont proposé une architecture hybride CNN-RNN pour modéliser l'évolution spatio-temporelle de l'expression faciale afin d'effectuer la reconnaissance des émotions dans la vidéo. Tout d'abord, un CNN est entraîné pour classer les images statiques, puis un RNN est utilisé pour agréger les caractéristiques des images. La représentation de la couche supérieure du CNN fournit

des informations structurelles d'une image donnée et le RNN modélise l'évolution spatio-temporelle de la structure dans le temps. Ils explorent également deux méthodes de fusion, opérant au niveau des caractéristiques et au niveau de la décision.

Dans (H.-W. Ng et al., 2015), les auteurs ont testé deux modèles pré-entraînés (AlexNet et VGGNet) basés sur le jeu de données ImageNet, suivis d'un réglage fin supervisé en deux étapes, après un premier réglage fin sur Fer-2013, le modèle est affiné par une deuxième étape de réglage fin sur le jeu de données Static Facial Expression Sub-Challenge (SFEW) du défi EmotiW 2015. Les résultats expérimentaux démontrent que cette approche de réglage fin en cascade donne de meilleurs résultats que le réglage fin en une seule étape avec des ensembles de données combinés. Leur meilleure précision est de 55,6%, obtenue à partir du modèle pré-entraîné AlexNet qui est suivi de deux étapes de réglage fin, la première étape avec le jeu de données Fer28, et la seconde étape avec le jeu de données EmotiW.

Le travail présenté dans (K. Zhao et al., 2016) a proposé un réseau profond unifié appelé " deep region and multi-label learning " (DRML). Il s'agit d'une nouvelle couche de régions qui utilise des fonctions de feed-forward pour induire des régions faciales importantes afin de capturer les informations structurelles du visage. Cette couche de régions sert de conception alternative entre les couches connectées localement et les couches convolutionnelles classiques. Le réseau complet peut être entraîné de bout en bout, et apprend automatiquement des représentations robustes des variations inhérentes à une région locale.

La référence (Arriaga et al., 2017) a proposé un système de vision en temps réel qui permet la détection des visages, la classification des sexes et des émotions en une seule étape combinée, en utilisant une architecture CNN moderne. Le système a été intégré avec succès dans un robot Care-O-bot 3, et a été étendu aux plateformes robotiques générales et aux défis de la compétition RoboCup@Home.

Les auteurs de (Miao et al., 2019) proposent un système basé sur CNN utilisant l'apprentissage par transfert pour l'estimation en temps réel des expressions faciales à partir d'une webcam. Le modèle proposé, MobileNet, est mis en œuvre dans un cadre à la fois hors ligne et en temps réel qui permet une sortie rapide et précise en temps réel. Le processus de formation du CNN est développé par une stratégie de réglage fin en deux étapes. Pour évaluer le modèle proposé, des expériences sont réalisées sur deux ensembles de données, JAFFE et CK+. Une précision de 95,24 % est obtenue sur l'ensemble de données JAFFE, tandis qu'une précision de 96,92 % est atteinte sur l'ensemble de données CK+ à 6 classes.

Dans (Jain et al., 2019), les auteurs proposent d'utiliser un seul réseau neuronal convolutif profond qui se compose de couches de convolution et de blocs résiduels profonds. Ce modèle est évalué sur deux jeux de données publics pour évaluer ses

performances, il atteint une précision de 95,23% pour le jeu de données JAFEE, et 93,24 pour le jeu de données CK+.

Récemment, dans (K. Li et al., 2020), les auteurs proposent d'utiliser un CNN simple pour la reconnaissance d'expressions faciales, sans couches entièrement connectées, et avec de nouvelles stratégies de recadrage et de rotation des visages. Le but est d'extraire uniquement les caractéristiques faciales utiles. Les jeux de données utilisés dans ce travail pour l'évaluation sont CK+ et JAFFE, les expériences ont atteint une précision de 97,38% pour la base de données CK+, et 97,18% pour la base de données JAFFE.

D'autres travaux existent qui traitent de la reconnaissance d'expressions faciales à partir de séquences d'images, ainsi que d'autres travaux qui font la reconnaissance à partir d'autres types d'images comme les images infrarouges. La majorité des études de recherche liées à la reconnaissance des expressions faciales ont utilisé un seul CNN, soit en apprenant à partir de zéro, soit par apprentissage de transfert, tandis que certaines d'entre elles ont modifié la structure classique du CNN, en supprimant certaines couches ou en les remplaçant par d'autres. D'autres études ont été basées sur la combinaison et la fusion de différents CNN, ou l'hybridation de divers modèles d'apprentissage profond, ou même une hybridation entre des méthodes conventionnelles et des méthodes basées sur l'apprentissage profond.

Parmi ces travaux, les auteurs de (J. Chen et al., 2019) ont proposé un cadre en deux étapes, basé sur un réseau de neurones à convolution de différence (DCNN) pour éliminer les différences individuelles. Tout d'abord, les cadres neutres et les cadres à pleine expression sont automatiquement détectés par un réseau neuronal convolutif binaire. Ensuite, un réseau DCNN de bout en bout est utilisé pour contenir les informations de différence entre ces deux trames, qui se compose de deux branches identiques de couches convolutionnelles et de couches entièrement connectées avec une couche de différence ajoutée. Les résultats obtenus sur les jeux de données CK+ et BU-4DFE sont considérés comme prometteurs (95,4 % pour CK+, et 77,4 % pour BU-4DFE) et cette méthode a été utilisée avec succès pour l'analyse de l'état affectif des étudiants dans un environnement d'apprentissage en ligne.

Dans (Aghamaleki & Ashkani Chenarlogh, 2019), les auteurs ont traité le problème des données limitées dans les CNN en utilisant une structure multi-flux et trois caractéristiques créées à la main (l'extracteur de code de motif binaire local LBP et l'opérateur de détection de bord de Sobel dans les directions horizontale et verticale des images). Le système a été évalué sur les jeux de données CK+ et MUG, avec des données limitées et étendues, et les auteurs ont remarqué que la précision de la reconnaissance est améliorée en utilisant les données limitées.

Le travail présenté dans (H. Zhang et al., 2021) tente d'améliorer l'influence des informations d'identité dans la reconnaissance des expressions faciales en utilisant un

réseau à double branche identité-expression (Identity-Expression Dual Branch Network, IE-DBN). Ce réseau extrait d'abord les caractéristiques d'identité et d'expression par deux branches, suivies d'un module bilinéaire pour l'agrégation des deux caractéristiques, afin de souligner l'impact de l'identité, ainsi que d'améliorer les variations interclasses et les similarités intra classes. La méthode a été évaluée sur trois ensembles de données CK+, Oulu-CASIA et RAF-DB, et a obtenu une précision de reconnaissance de 96,02%, 85,21% et 84,75% respectivement.

Les auteurs de (D. Zhu et al., 2021) ont proposé de baser l'apprentissage de l'architecture qu'ils proposent sur les informations clés du visage en introduisant un modèle LKRNet à double branche, et d'utiliser la perte triplet pour réduire la distance intra-classe et augmenter l'inter-classe des caractéristiques. Les résultats expérimentaux sur les ensembles de données CK+ et FER2013 ont montré son avantage par rapport au CNN général.

Dans (Zou et al., 2022), les auteurs proposent un réseau de neurones convolutif basé sur la fusion multi-fonctions (Multi-Feature Fusion based Convolutional Neural Network, MFF-CNN), qui est également un réseau à double branche, mais qui présente l'avantage de la légèreté (son nombre de paramètres est de 1,21M). Les deux branches parallèles (la branche de l'image entière et la branche du patch) permettent l'extraction de caractéristiques faciales globales et locales. Le modèle a été testé sur quatre jeux de données : CK+, JAFFE, Oulu-CASIA et SFEW2.0.

Afin d'éviter le problème de l'extraction insuffisante de caractéristiques et d'améliorer les performances de la reconnaissance des expressions faciales, les auteurs de (Shi et al., 2021) ont proposé d'utiliser un réseau de neurones convolutif multi-branches à connexions croisées (Multi-Branch Cross-Connection Convolutional Neural Network , MBCC-CNN), qui est basé sur les approches de connexion résiduelle, de réseau dans le réseau et de structure arborescente, puis ont ajouté une connexion croisée raccourcie pour la sommation de la couche de sortie convolutive, puis la fusion de caractéristiques est effectuée comme un ajout. Enfin, le classificateur SoftMax est utilisé pour la reconnaissance des expressions faciales. Cette méthode permet d'obtenir de meilleurs résultats au prix d'un grand nombre de données d'entraînement et d'une augmentation des coûts de calcul.

Enfin, plusieurs architectures de réseaux neuronaux intéressantes et prometteuses, comme les réseaux de capsules (S. Cao et al., 2020), les réseaux adversaires génératifs (Xie et al., 2021) et les transformateurs (F. Ma et al., 2021), ont également été étudiées pour résoudre les limitations des systèmes de FER et ont atteint des performances intéressantes.

5.5 Les bases de données de reconnaissance des expressions faciales

De nombreuses bases de données ont été utilisées dans le domaine de la recherche sur les expressions faciales pour mener des expériences comparatives et approfondies. Traditionnellement, les expressions faciales humaines ont été étudiées à l'aide d'images statiques 2D ou de séquences vidéo 2D (Ko, 2018b). Dans ce qui suit, nous présentons brièvement certaines bases de données populaires liées à la FER, constituées de séquences vidéo 2D et d'images fixes. Ces ensembles de données sont généralement divisés en deux catégories : des petites bases de données contrôlées en laboratoire, et les bases de données naturelles à grande échelle.

5.5.1 Ensembles de données FER contrôlés en laboratoire

5.5.1.1 Extended Cohn Kanade (CK+)

La base de données Extended Cohn Kanade (CK+) (Lucey et al., 2010) est la base de données contrôlée en laboratoire la plus communément utilisée pour l'évaluation des systèmes de FER. Elle se compose de 593 séquences vidéo obtenues de 123 sujets. Parmi celles-ci, 327 séquences provenant de 118 sujets sont étiquetées comme l'une des sept expressions. Pour chaque séquence, seule la dernière image est étiquetée. Les trois dernières images sont extraites de chaque séquence du jeu de données CK+, qui contient 981 expressions faciales.

5.5.1.2 Japanese Female Facial Expressions (JAFFE)

La base de données JAFFE (Lyons, Michael et al., 1998) contient 213 images de sept émotions faciales (six émotions faciales de base et une neutre) posées par dix modèles féminins japonais différents. Chaque image a été évaluée sur la base de six adjectifs émotionnels par 60 sujets japonais. La taille originale de chaque image faciale est de 256 x 256 pixels.

5.5.1.3 Oulu-CASIA NIR-VIS Database (Oulu-CASIA)

La base de données NIR-VIS de l'Oulu-CASIA (G. Zhao et al., 2011) est constituée de 2880 séquences d'images avec six expressions de base provenant de 80 personnes âgées de 23 à 58 ans. Toutes les expressions sont capturées dans la direction frontale avec trois conditions d'illumination différentes : normale, faible et sombre. On a demandé aux sujets de faire une expression faciale selon un exemple d'expression montré dans les séquences d'images. Le matériel d'imagerie fonctionne à une fréquence de 25 images par seconde et la résolution des images est de 320 x 240 pixels.

5.5.2 Ensembles de données FER naturelles à grande échelle.

5.5.2.1 Facial Emotion Recognition 2013 (FER-2013)

La base de données Facial Emotion Recognition 2013 (FER-2013) a été créée par Pierre Luc Carrier et Aaron Courville et a été présentée dans le cadre du défi de reconnaissance d'expressions faciales de l'atelier ICML 2013 (I. J. Goodfellow et al., 2013). Le jeu de données est composé de 35887 images faciales, la plupart dans des environnements naturels. Il se compose de trois parties : les données d'entraînement originales (OTD), qui comprennent 28709 images, les données de test publiques (PTD), qui comprennent 3589 images, et les données de test finales (FTD), qui comprennent 3589 images utilisées pour noter les modèles finaux.

5.5.2.2 Denver Intensity of Spontaneous Facial Action Database (DISFA)

DISFA (Mavadati et al., 2013) est constituée de 130 000 images vidéo stéréo à haute résolution (1024 768) de 27 sujets adultes (12 femmes et 15 hommes) de différentes ethnies. Les intensités des UA (échelle de 0 à 5) pour toutes les images vidéo ont été notées manuellement par deux experts humains en FACS. La base de données comprend également 66 points de repère faciaux pour chaque image de la base. La taille originale de chaque image faciale est de 1024 x 768 pixels.

5.5.2.3 Acted Facial Expressions in the Wild dataset (AFEW)

AFEW (Dhall et al., 2012) est un corpus de données d'expressions faciales dynamiques et temporelles avec des expressions spontanées, différentes poses de la tête, des occlusions et des illuminations, ce qui est proche de l'environnement du monde réel. Les échantillons sont étiquetés avec sept catégories d'émotions (six émotions de base et une neutre). AFEW est divisé en trois partitions de données de manière indépendante en termes de sujet et de source de film/télévision : Train (773 échantillons), Val (383 échantillons) et Test (653 échantillons), ce qui garantit que les données des trois ensembles appartiennent à des films et des acteurs mutuellement exclusifs. L'ensemble de données sur les expressions faciales statiques dans la nature (SFEW) (Dhall et al., 2011) a été développé en sélectionnant des images de l'AFEW et a été divisé en trois ensembles : Train (958 échantillons), Val (436 échantillons) et Test (372 échantillons). Il est étiqueté avec sept expressions faciales (six émotions de base et une neutre).

5.5.2.4 AffectNet

AffectNet (Mollahosseini et al., 2017) est une base de données contenant plus d'un million d'images provenant d'Internet et obtenues en interrogeant différents moteurs de recherche à l'aide d'étiquettes liées aux émotions. Cette base de données est sans conteste la plus importante à fournir des expressions faciales dans deux modèles d'émotion différents (modèle catégorique et modèle dimensionnel), dont 450 000 images ont des étiquettes annotées manuellement pour huits expressions de base.

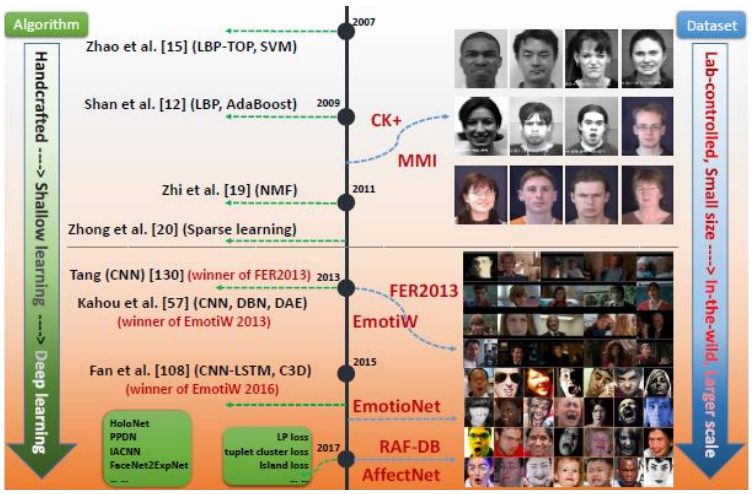


Figure 5-2 : L'évolution de la reconnaissance des expressions faciales en termes d'ensembles de données et de méthodes (S. Li & Deng, 2020).

Alors que ces bases de données deviennent de plus en plus efficaces sur les expressions faciales, les techniques d'apprentissage profond sont de plus en plus mises en œuvre pour relever les défis de la reconnaissance des émotions dans la nature (S. Li & Deng, 2020).

La **Erreur ! Source du renvoi introuvable.** illustre cette évolution de la RIE du point de vue des algorithmes et des ensembles de données.

5.6 La méthode proposée pour la reconnaissance des expressions faciales

En investiguant plusieurs approches pour la reconnaissance des expressions faciales, les réseaux de neurones profonds offrent généralement de meilleures performances de classification et obtiennent de très bons résultats en termes de précision dans la reconnaissance des expressions faciales par rapport aux approches conventionnelles en raison de leur extraction automatique et intelligente des caractéristiques. Les recherches montrent que la reconnaissance des expressions faciales est influencée de manière significative et efficace par les caractéristiques faciales extraites.

Le défi dans le processus d'entraînement de ces réseaux est la limite des échantillons disponibles dans les ensembles de données de reconnaissance des expressions faciales. En se concentrant sur l'amélioration des performances des systèmes de reconnaissance des expressions faciales (FER), nous proposons un modèle hybride combinant des caractéristiques extraites basées sur les CNN pour assurer la complémentarité et la diversité, et des avantages d'apprentissage de transfert dans la classification pour les applications FER. Nos contributions dans ce travail sont les suivantes :

1. Un modèle à double branche basé sur un réseau CNN simple et performant inspiré de l'architecture VGGnet et un réseau CNN pré-entraîné, qui est un

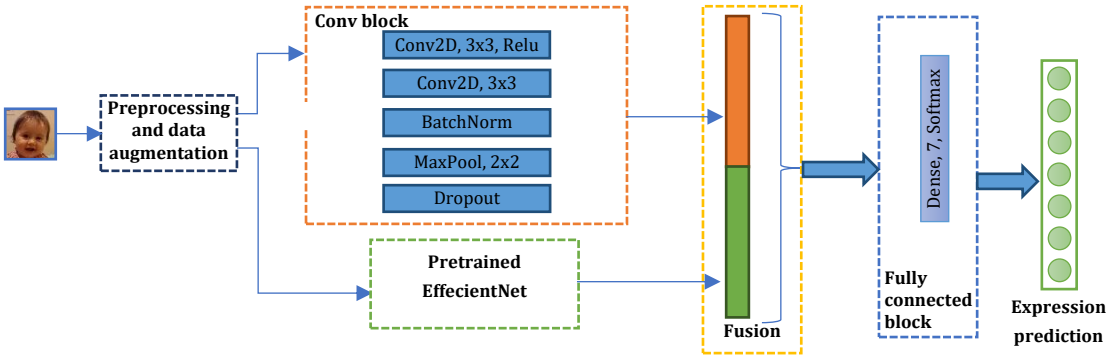


Figure 5-3 : Aperçu de l'approche FER à double branche proposée.

réseau puissant, est proposé pour compenser le manque d'échantillons d'entraînement, en fusionnant leurs caractéristiques extraites et pour améliorer la précision de la reconnaissance.

2. Une stratégie de formation conjointe est conçue pour le modèle à double branche proposé.
3. Deux jeux de données, Fer-2013 et CK+, sont utilisés pour valider l'efficacité de nos architectures. CK+ est un jeu de données classique sur les expressions faciales et FER2013 fournit des échantillons de visages capturés dans le monde réel.

Nous proposons dans ce travail (Bordjiba et al., 2022) une nouvelle architecture de système à double branche pour exploiter les deux types d'apprentissage de CNN, à savoir l'apprentissage à partir de zéro et l'apprentissage par transfert, afin de diversifier les cartes de caractéristiques fournies au classificateur. Le système FER proposé, tel qu'illustré dans la Figure 5-3, comprend les étapes suivantes : deux branches convolutives, générant chacune ses propres cartes de caractéristiques. Ces sorties seront combinées pour représenter le vecteur de caractéristiques responsable de l'étape de classification. La première branche extrait les caractéristiques de l'image d'entrée à travers les couches convolutives d'un CNN inspiré de VGG, tandis que la seconde branche extrait les caractéristiques de la même image à travers un réseau pré-entraîné. Enfin, les cartes de caractéristiques sont fusionnées en utilisant la concaténation, et la classification est effectuée en utilisant une couche entièrement connectée avec une fonction d'activation Softmax pour reconnaître les expressions. Ces étapes seront présentées en détail dans les sections suivantes.

5.6.1 Branche inspirée du VGG

En nous appuyant sur l'architecture VGG16, et après avoir testé et expérimenté plusieurs configurations (Bordjiba et al., 2019), nous proposons deux modèles simples et profonds, VGGinspiredCNN1 et VGGinspiredCNN2, enrichis par la normalisation des lots pour améliorer la généralisation et les couches d'optimisation et d'abandon (Dropout).

Leur architecture est composée de cinq blocs convolutifs et d'un bloc entièrement connecté. Chaque bloc convolutif est composé de deux couches de convolution (tous les filtres utilisés sont de taille 3x3 comme les modèles VGGNet) suivies d'une normalisation par lot et d'une couche de max-pooling (avec un noyau de taille 2x2) et d'une couche de dropout. Chaque couche de convolution est équipée d'une rectification non linéaire (Relu). Le bloc entièrement connecté du premier modèle est composé de deux couches entièrement connectées avec 512 et 7 sorties respectivement. La couche entièrement connectée du second modèle est une seule couche avec sept sorties. Les architectures CNN inspirées du VGG sont décrites dans la Figure 5-4 et Tableau 5-1.

Tableau 5-1: La structure de la couche convolutive des deux modèles proposés inspirés de VGGnet.

	Entrée	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling
Taille du noyau	48 x 48	3	2	3	2	3	2	3	2	3	2
Stride		1	2	1	2	1	2	1	2	1	2
remplissage		0	0	0	0	0	0	0	0	0	0
# filtres		64		128		256		512		512	
#Réplications		2	1	2	1	2	1	2	1	2	1

5.6.2 Branche basée sur l'apprentissage par transfert

Pour permettre aux CNN d'apprendre et d'extraire des caractéristiques et d'atteindre une grande précision, des millions d'échantillons doivent être utilisés dans leur base d'apprentissage. Cependant, les ensembles de données d'expression faciale existants ne contiennent que quelques centaines ou milliers d'échantillons. Cette taille insuffisante est l'un des principaux problèmes de la FER basée sur les CNN. Pour surmonter cette limite, l'utilisation de l'apprentissage par transfert sera une solution

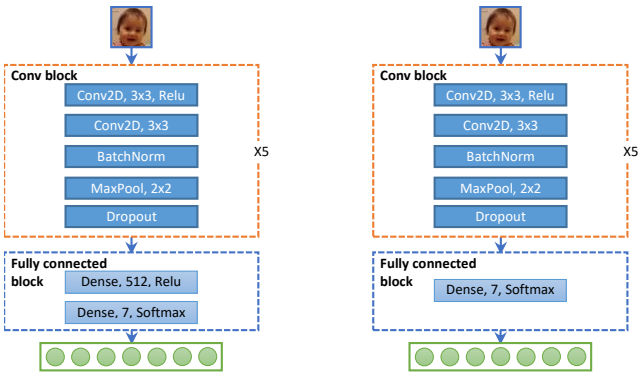


Figure 5-4 : L'architecture VGGinspiredCNN proposé.

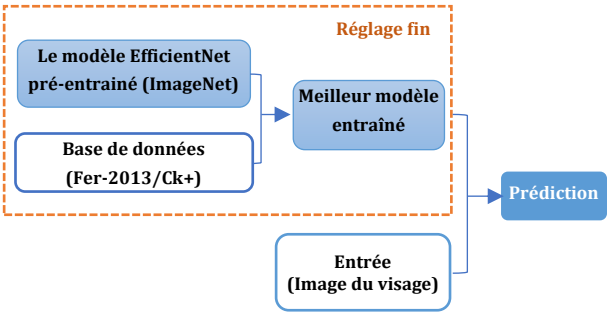


Figure 5-5 : L'architecture du réseau CNN pré-entraîné.

possible ; il s'agit d'une pratique courante où le réseau est d'abord initialisé avec un ensemble de poids (et de biais) préformés basés sur un ensemble de données à grande échelle d'une tâche et ces paramètres sont ensuite recyclés à une autre nouvelle tâche cible.

Afin d'obtenir une meilleure précision que les CNN traditionnels, les auteurs de (Tan & Le, 2019) ont proposé une famille de modèles, les EfficientNets, qui peuvent être systématiquement dimensionnés en fonction des ressources disponibles. Un équilibre entre les dimensions du réseau est obtenu en les mettant simplement à l'échelle avec un rapport constant. Les modèles EfficientNets se transfèrent bien aux ensembles de données tels que CIFAR-100 (Tan & Le, 2019), fruits (Duong et al., 2020), etc. avec moins de paramètres. Huit modèles d'EfficientNetB0-EfficientNetB7 ont été examinés afin de déterminer leur efficacité et leurs performances. Dans cette branche par transfert, les poids pré-entraînés du jeu de données ImageNet sont utilisés étant donné qu'il contient un grand nombre d'images de personnes (*ImageNet*, 2020), environ 952K images, ce qui est très pertinent pour classer les jeux de données d'expressions faciales Fer-2013 et Ck+ utilisés dans l'évaluation. Ainsi, ces paramètres de réseau pré-entraînés sont utilisés pour l'initialisation. Ensuite, le modèle est entraîné, et un réglage fin sera effectué pour extraire des caractéristiques plus spécifiques. La Figure 5-5 montre l'architecture du réseau CNN pré-entraîné.

Tableau 5-2 : Le réseau de base EFFICIENTNET-B0 (Tan & Le, 2019).

Stage i	Operator F_i	Input Resolution H_i x W_i	Output Channels C_i	Layers Li
1	Conv3x3	224 x 224	32	1
2	MBConv1, k3x3	112 x 112	16	1
3	MBConv6, k3x3	112 x 112	24	2
4	MBConv6, k5x5	56 x 56	40	2
5	MBConv6, k3x3	28 x 28	80	3
6	MBConv6, k5x5	14 x 14	112	3
7	MBConv6, k5x5	14 x 14	192	4
8	MBConv6, k3x3	7 x 7	320	1
9	Conv1x1 & Pooling & FC	7 x 7	1280	1

Le modèle de base EfficientNet-B0 est constitué de 18 couches de convolution (avec une taille de noyau de 3x3 ou 5x5). Ensuite, une couche d'aplatissement suit le max pooling tel que décrit dans le **Erreur ! Source du renvoi introuvable.**, son principal bloc de construction est le goulot d'étranglement inversé mobile MBConv, auquel ils ajoutent également l'optimisation squeeze-and-excitation. Les autres configurations EfficientNet, c'est-à-dire B1 à B7, sont mises à l'échelle à partir de la configuration de base EfficientNet-B0 avec différents coefficients composés. Une nouvelle couche de classification remplace les dernières couches entièrement connectées avec sept classes (correspondant à sept expressions).

5.6.3 Module de fusion

La concaténation des vecteurs de caractéristiques est couramment utilisée pour fusionner et intégrer plusieurs canaux ou branches dans plusieurs architectures (Aghamaleki & Ashkani Chenarlogh, 2019; Zou et al., 2022). L'opération qui combine les caractéristiques extraites de la branche inspirée par le VGG et les caractéristiques extraites du modèle pré-entraîné est définie par la formule suivante :

$$(x_1^V, x_2^V, \dots)^T \oplus (x_1^P, x_2^P, \dots)^T = (x_1^V, x_2^V, \dots, x_1^P, x_2^P, \dots)^T \quad (5.1)$$

Où : $\oplus$ désigne l'opérateur de concaténation de vecteurs,

$(x_1^V, x_2^V, \dots)^T$ désigne les caractéristiques extraites de la branche inspirée par VGG,

$(x_1^P, x_2^P, \dots)^T$ désigne les caractéristiques extraites du modèle pré-entraîné.

5.7 Résultats et discussions

5.7.1 Ensembles de données utilisés et prétraitement

Les performances des modèles proposés en matière de reconnaissance des expressions faciales sont démontrées à l'aide de deux bases de données largement utilisées, à savoir la base de données Fer-2013 (I. J. Goodfellow et al., 2013) et la base de données Cohn Kanade (Lucey et al., 2010).

Toutes les images de visage sont redimensionnées à 48 × 48 pixels, puis normalisées pour avoir une moyenne nulle et une variance unitaire. Pour rendre le modèle proposé plus robuste aux transformations légères et au bruit, l'augmentation des données est appliquée en utilisant différentes transformations linéaires : rotation, retournement horizontal, zoom et inclinaison de la zone centrale.

5.7.2 Paramètres expérimentaux

Afin de démontrer la performance du modèle à double branche proposé, trois expériences d'apprentissage différentes sont menées : le modèle CNN classique basé sur l'architecture VGG, l'apprentissage par transfert de tous les modèles EfficientNet et l'apprentissage joint du modèle à double branche.

Pour la formation, les images des ensembles de données CK+ sont mélangées de manière aléatoire et sont réparties comme suit : 85% pour l'entraînement, 15% pour le test. Pour le jeu de données FER-2013, l'ensemble d'entraînement complet (28 709) et l'ensemble de test public (3 589) sont utilisés pour l'entraînement et la validation, respectivement.

La fonction de perte totale est optimisée pendant la rétropropagation à l'aide de l'optimiseur Adam. Il convient de noter que différents optimiseurs ont été testés, y compris la descente de gradient stochastique, et qu'Adam s'est avéré plus performant.

L'implémentation est basée sur la bibliothèque Keras (*Keras*, 2015/2022) avec un backend TensorFlow (Abadi et al., 2016). OpenCV(« OpenCV Face Recognition », s. d.) est utilisé pour toutes les opérations sur les images. Toutes les expériences ont été exécutées avec PyTorch et entraînées à l'aide de Google Colaboratory (*Google Colaboratory*, s. d.). Compte tenu des limitations d'un compte Google Colab gratuit, telles qu'un maximum de 12 heures par session d'entraînement, le nombre et le type de GPU ou la capacité VRAM, la phase d'entraînement a été réalisée en utilisant plusieurs paramètres comme indiqué dans le Tableau 5-3 pour les deux ensembles de données Fer-2013 et Ck+.

Tableau 5-3 : Configurations expérimentales pour FER-2013 et CK+.

Modèle		Fer-2013			CK+		
		Époques	Taille du lot	Taux d'apprentissage	Époques	Taille du lot	Taux d'apprentissage
Learning from scratch	VGGinspiredCNN1	60	64	0,001	60	8	0,001
	VGGinspiredCNN2	60	64	0,001	100	16	0,001
Apprentissage par transfert	EfficientNet-B0	80	32	0.00001	80	16	0.0001
	EfficientNet-B1	100	32	0.00001	80	16	0,0001
	EfficientNet-B2	100	32	0.00001	80	8	0.0001
	EfficientNet-B3	100	32	0.00001	80	8	0.0001
	EfficientNet-B4	80	32	0.00001	80	8	0.0001
	EfficientNet-B5	80	32	0.00001	80	16	0.0001
	EfficientNet-B6	80	32	0.00001	80	16	0.0001
	EfficientNet-B7	80	32	0.00001	80	16	0.0001
Double branche	CNN+EfficientNet-B0	60	64	0.001	120	8	0.001
	CNN+EfficientNet-B1	60	64	0.001	100	8	0.0001
	CNN+EfficientNet-B2	50	64	0.001	80	8	0.0001
	CNN+EfficientNet-B3	40	64	0.0002	100	8	0.0001
	CNN+EfficientNet-B4	40	128	0.0001	120	8	0.0001
	CNN+EfficientNet-B5	40	64	0.0001	120	16	0.0001
	CNN+EfficientNet-B6	30	64	0.0001	120	8	0.0001
	CNN+EfficientNet-B7	40	64	0.0002	120	16	0.0001

5.7.3 Résultats et discussions

Comme mentionné précédemment, des expériences ont été menées pour déterminer l'efficacité du modèle CNN à double branche proposé par rapport aux modèles CNN par apprentissage à partir de zéro et EfficientNet par apprentissage par transfert. Ainsi, les deux modèles CNN inspirés de VggNet, l'apprentissage par transfert de l'EfficientNet avec ses huit configurations, et le modèle à double branche proposé sont testés sur deux jeux de données FER largement utilisés : CK + et FER-2013. Le jeu de données FER-2013 comprend les expressions de sept étiquettes : colère, dégoût, peur,

joie, tristesse, surprise et neutre, tandis que le jeu de données CK+ comprend les mêmes expressions à l'exception de l'expression neutre, et comprend également l'expression mépris.

5.7.3.1 Évaluation du jeu de données FER-2013 :

Les résultats expérimentaux des modèles CNN entraînés à partir de zéro, les modèles EfficientNet pré-entraînés ainsi que le modèle à double branche proposé sur le jeu de données FER-2013 sont donnés dans les Figure 5-6 à Figure 5-8 et le **Erreur ! Source du renvoi introuvable.**.

Tableau 5-4 : Performances des modèles proposés par rapport aux autres modèles sur l'ensemble de données Fer-2013.

Modèle		Taux de précision
Learning from Scratch	VGGinspiredCNN1	66.71
	VGGinspiredCNN2	67.70
Apprentissage par transfert	EfficientNet-B0	56.23
	EfficientNet-B1	57.48
	EfficientNet-B2	57.43
	EfficientNet-B3	58.32
	EfficientNet-B4	57.13
	EfficientNet-B5	57.70
	EfficientNet-B6	57.17
	EfficientNet-B7	60.10
Double branche	CNN+EfficientNet-B0	63.36
	CNN+EfficientNet-B1	62.88
	CNN+EfficientNet-B2	62.77
	CNN+EfficientNet-B3	63.80
	CNN+EfficientNet-B4	62.25
	CNN+EfficientNet-B5	62.09
	CNN+EfficientNet-B6	62.31
	CNN+EfficientNet-B7	62.22
Modèles de l'état de l'art	(Mollahosseini et al., 2016)	66.4 (Top-1)
	(Arriaga et al., 2017)	66
	(Giannopoulos et al., 2018)	65.2
	(Shao & Qian, 2019)	71.14

La précision de tous les modèles proposés par rapport aux autres modèles de pointe est indiquée dans le Tableau 5-4, et les figures de 5.6 à 5.8 montrent les matrices de confusion normalisées correspondantes. Pour la base de données Fer-2013, les modèles CNN inspirés de VGG offrent des performances compétitives par rapport aux résultats des modèles de l'état de l'art et dépassent la précision de niveau humain, sur la base des taux de précision obtenus par l'expérimentation. Comme l'indique l'étude menée par (S. Li & Deng, 2020) sur la reconnaissance des expressions faciales basée sur l'apprentissage profond, les tests les plus précis sur le jeu de données FER-2013 à l'aide d'un seul réseau CNN se situent dans une fourchette de 67 à 71 % et les modèles qui y parviennent sont très performants. Il peut être observé à partir des matrices de confusion que les expressions "heureux" et "surprise" sont plus faciles à reconnaître,

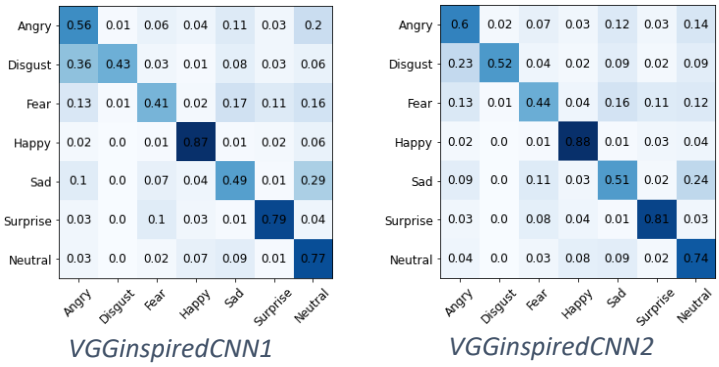


Figure 5-8 : Matrices de confusion pour les modèles VGGinspiredCNN sur la base de données Fer-2013.

avec une précision de plus de 80%, tandis que l'expression "peur" et les expressions "triste" et "dégoût" sont les plus difficiles à reconnaître avec notre meilleur modèle VGGinspiredCNN2, avec une précision de 44%, 51% et 52% respectivement. Il est important de mentionner que parfois, en tant qu'être humain, il est difficile de reconnaître une expression de tristesse ou de peur, ceci est dû au fait que les gens n'expriment pas tous leurs émotions de la même manière, et le faible taux de précision de l'expression " dégoût " est dû au petit nombre d'échantillons de cette expression dans le jeu de données Fer-2013.

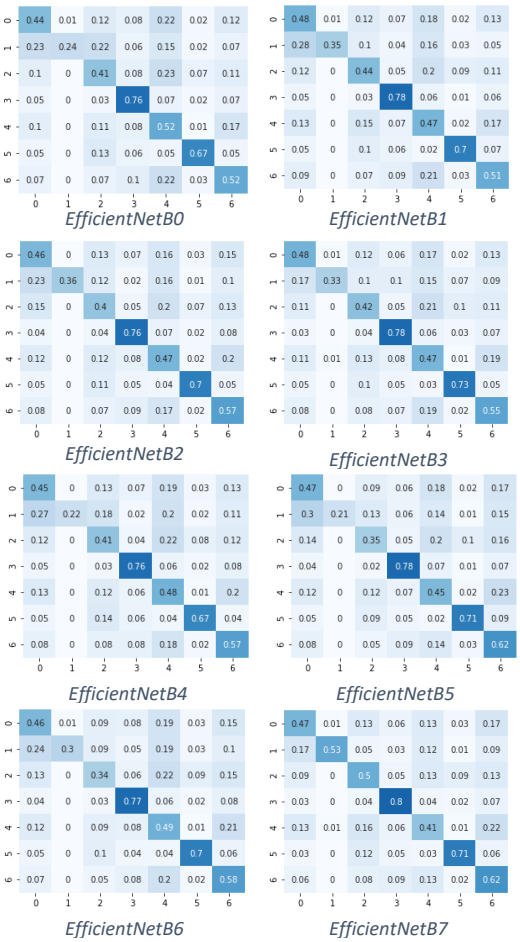


Figure 5-7 : Matrices de confusion pour les modèles EfficientNet sur la base de données Fer-2013.

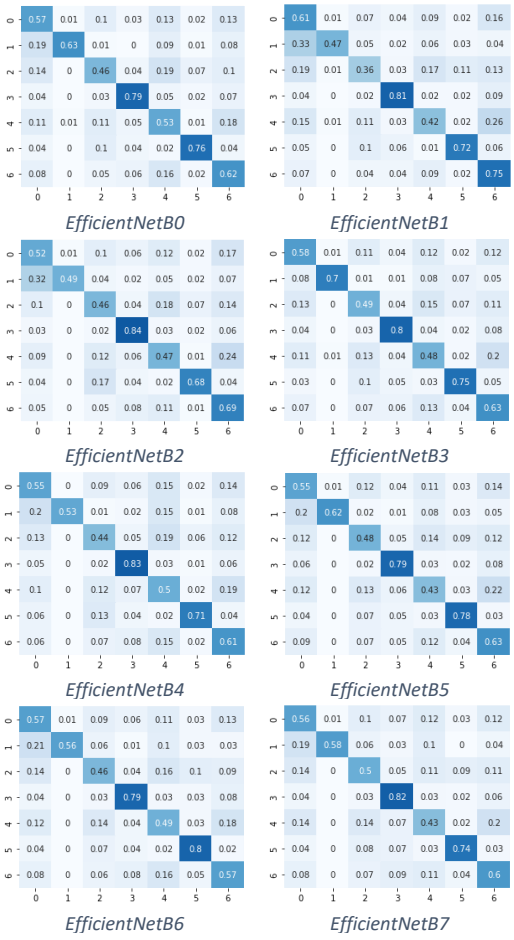


Figure 5-7 : Matrices de confusion pour le modèle à double branche proposé sur la base de données Fer-2013.

D'autre part, les modèles EfficientNet pré-entraînés ont atteint des taux de précision acceptables pour ce jeu de données, avec le modèle EfficientNet -B7 atteignant sa plus haute précision de 60,10%. Notez que dans tous les modèles EfficientNet, l'expression "heureux" est toujours la plus reconnue avec la plus grande précision (plus de 76%), tandis que l'expression "dégoût" est toujours la moins reconnue par les différents modèles EfficientNet, avec une précision entre 21% et 35%, sauf pour le modèle EfficientNet-B7, où cette classe atteint une précision de 53%.

Tandis que le modèle à double branche proposé fournit une amélioration significative pour toutes les configurations d'EfficientNet, en particulier pour le modèle EfficientNet B0, qui gagne 7,13% en précision, mais la meilleure performance est obtenue par le modèle EfficientNet-B3, avec une précision de 63,80%. Les matrices de confusion des modèles à double branche montrent une amélioration du taux de reconnaissance de l'expression "dégoût", qui varie de 47% à 70%, obtenue par le meilleur modèle à double branche avec EfficientNet-B3 pour le jeu de données Fer-2013, alors qu'aucune amélioration n'est enregistrée sur le taux de reconnaissance de l'expression "heureux". Il est certain que le modèle à double branche proposé a amélioré les résultats pour ce jeu de données, cependant, le Tableau 5-4 révèle également que la précision de reconnaissance obtenue n'est pas aussi performante que les modèles de pointe (S. Li & Deng, 2020) qui ont été conçus spécifiquement pour la reconnaissance des expressions faciales sans contrainte. Dans le modèle à double branche proposé, une seule couche entièrement connectée est utilisée afin d'obtenir un réseau efficace, ce qui limite ses performances lors du traitement de la tâche de FER sans contrainte.

5.7.3.2 Évaluation du jeu de données Ck+ :

De la même manière que pour l'évaluation du jeu de données Fer-2013, les matrices de confusion pour chaque expression et chacun des modèles proposés sur le jeu de données CK+ sont présentées dans les Figures de Figure 5-9 à Figure 5-11, et le résultat de la comparaison avec les autres modèles concurrents est donné dans le Tableau 5-5.

Tableau 5-5 : Performances des modèles proposés par rapport aux autres modèles sur l'ensemble de données CK+.

Network		Accuracy rate
Learning from scratch	VGGinspiredCNN1	93.40
	VGGinspiredCNN2	98.48
Transfer learning	EfficientNet-B0	99.32
	EfficientNet-B1	97.30
	EfficientNet-B2	97.97
	EfficientNet-B3	97.97
	EfficientNet-B4	97.30
	EfficientNet-B5	96.62
	EfficientNet-B6	96.62
	EfficientNet-B7	98.65
Dual Branch	CNN+EfficientNet-B0	99.32
	CNN+EfficientNet-B1	93.92
	CNN+EfficientNet-B2	94.59
	CNN+EfficientNet-B3	95.94

	CNN+EfficientNet-B4	95.94
	CNN+EfficientNet-B5	97.30
	CNN+EfficientNet-B6	98.65
	CNN+EfficientNet-B7	95.94
State of art models	(Mollahosseini et al., 2016)	93.2(Top-1)
	(Breuer & Kimmel, 2017)	72.1
	(Ding et al., 2017)	96.8
	(Shao & Qian, 2019)	95.29
	(Jain et al., 2019)	93.24

Les modèles CNN inspirés par VGG obtiennent des résultats très intéressants et compétitifs pour le jeu de données CK+, par rapport aux modèles de l'état de l'art, en particulier, le modèle VGGinspiredCNN2 atteint un taux de précision de 98,48% qui est meilleur que tous les travaux de référence cités ci-dessus.

Selon la matrice de confusion normalisée, 4 des 7 expressions ('colère', 'peur', 'joie', 'tristesse') sont reconnues à 100%, deux autres expressions ('dégoût' et 'surprise') sont reconnues à 98%, et seule l'expression 'mépris' est reconnue à 89%, qui est confondue avec l'expression 'triste'.

Selon les tableaux, Tableau 5-4 et Tableau 5-5 en comparant les résultats obtenus par les modèles proposés et certains modèles de référence, l'utilisation de l'approche à double branche améliore les résultats pour le jeu de données CK+, mais pas autant que pour le jeu de données Fer-2013. Alors que la précision pour les sept expressions du jeu de données CK+ est élevée, le jeu de données Fer2013 présente une faible précision de classification en raison d'un mauvais étiquetage dans le jeu de test, sauf pour la catégorie " heureux ". Selon le tableau 5.5, on peut voir que tous les modèles EfficientNet atteignent une précision de plus de 96%. EfficientNet-B5 et EfficientNet-B6 obtiennent la plus faible précision de 96,62%, tandis que EfficientNet-B0 réalise la meilleure performance avec une précision de 99,32%. Le même taux de précision est obtenu dans le modèle à double branche proposé.

Néanmoins, le jeu de données Fer-2013 est le jeu de données le plus couramment utilisé pour la reconnaissance des expressions faciales. Il convient de noter que l'œil humain peut difficilement distinguer l'émotion appropriée pour certaines d'entre elles.

Figure 5-9 : Matrices de confusion pour les modèles VGGinspiredCNN sur la base de données Ck+.

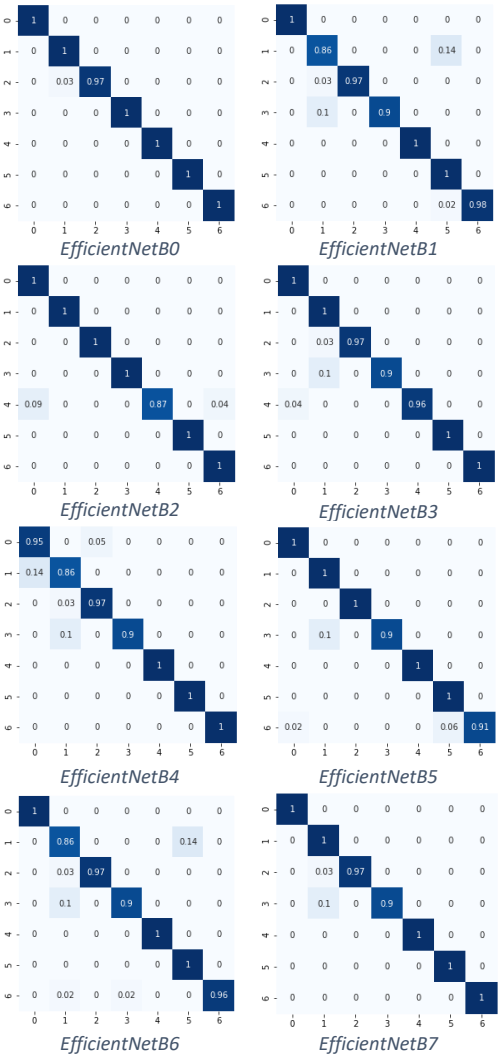


Figure 5-11 : Matrices de confusion pour les modèles EfficientNet sur la base de données CK+.

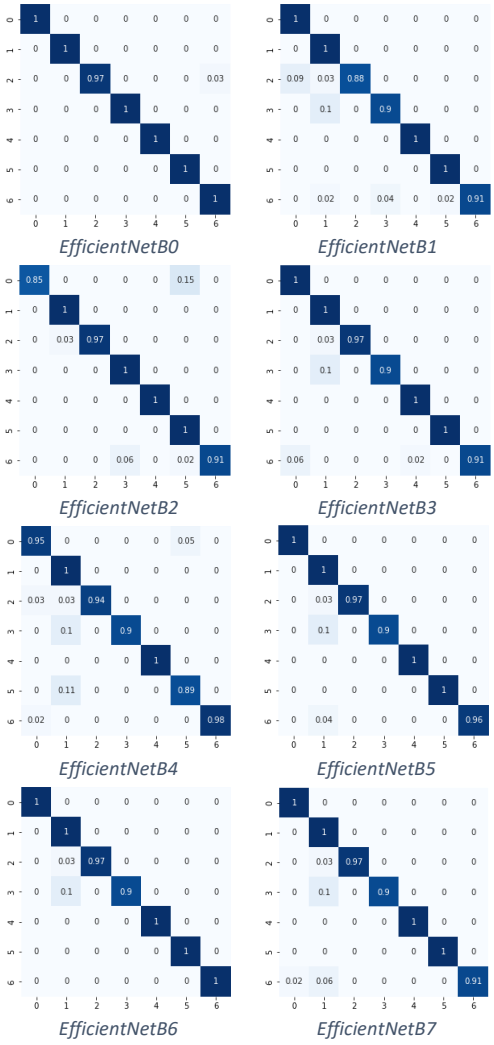


Figure 5-10 : Matrices de confusion pour le modèle à double branche proposé sur la base de données CK+.

5.8 Conclusion :

Dans cette étude, nous avons proposé un modèle à double branche qui exploite les méthodes d'apprentissage courantes, telles que l'apprentissage à partir de zéro et l'apprentissage par transfert, pour reconnaître les expressions faciales humaines dans différents contextes, qu'ils soient naturels ou contrôlés en laboratoire. Les expériences menées sur deux ensembles de données de référence, Fer-2013 et CK+, ont abouti à des résultats très prometteurs et compétitifs, surpassant les travaux existants. Cependant, les jeux de données FER présentent une limitation majeure en termes de taille et de déséquilibre des images entre les différentes classes, ce qui nuit à l'apprentissage profond. Pour relever ce défi, notre modèle combine deux types d'apprentissage, en affinant l'entraînement des modèles EfficientNets préalablement

entraînés sur Imagenet, et en concaténant les vecteurs de caractéristiques obtenus avec un CNN classique basé sur l'architecture robuste et bien connue de VGG.

Chapitre 6 : Transfert d'expression faciale à un visage synthétique par StyleGAN2

6.1 Introduction

L'animation faciale est un défi sérieux et permanent pour l'industrie infographique. Comme diverses émotions complexes doivent être exprimées par différentes déformations et animations faciales, la copie des déformations faciales d'un personnage existant à un autre est largement nécessaire dans l'industrie et le monde universitaire, afin de réduire le travail manuel fastidieux et répétitif de modélisation pour chaque nouveau personnage(Bian et al., 2020). Mais le transfert d'animations faciales réalistes entre deux modèles 3D est limité et peu pratique pour une utilisation générale. La plupart de ces méthodes de transfert de déformations nécessitent un mappage des correspondances, qui est une tâche fastidieuse dans la plupart des cas.

Comme susmentionné, ces dernières années, l'intérêt des chercheurs en vision par ordinateur s'est accru pour les réseaux adversaires génératifs (GAN), dont les applications comprennent la génération, la traduction, l'imputation et la super-résolution d'images.

Les films générés par ordinateur et les jeux numériques modernes reposent sur l'animation faciale expressive. Actuellement, la capture de performance basée sur la vision, c'est-à-dire le pilotage du visage animé avec le mouvement observé d'un acteur humain, fait partie intégrante de la plupart des pipelines de production. Si la qualité des systèmes de capture s'améliore régulièrement, le coût de production d'une animation faciale de haute qualité reste élevé (Karras et al., 2017). Tout d'abord, les systèmes de vision par ordinateur nécessitent des installations complexes et, souvent, un traitement et un entretien intensifs. Par ailleurs, lorsque de nouvelles prises de vue sont enregistrées, les acteurs doivent être sur place et, idéalement, conserver leur apparence. Cela peut s'avérer difficile si, par exemple, un autre rôle leur impose de se laisser pousser la barbe.

Dans ce chapitre, nous nous focalisons sur le problème du transfert d'expressions faciales. L'objectif est en effet le suivant : générer un visage synthétique ayant la même expression qu'un visage source, en expérimentant le StyleGAN2.

6.2 Travaux connexes

La maîtrise de l'animation faciale est un véritable défi pour l'animation par ordinateur et l'infographie. Pour exprimer les émotions d'un personnage, son état d'esprit et indiquer ses actions futures, le visage est un élément très important. Naturellement, le public est particulièrement entraîné à remarquer les traits subtils des visages et à identifier les différentes allusions dans les visages.

La recherche sur la modélisation et l'animation du visage comprend principalement des manipulations géométriques et d'images. Les manipulations géométriques comprennent les interpolations géométriques, les paramétrages, les méthodes d'éléments finis, la modélisation basée sur les muscles, la simulation visuelle utilisant des pseudo-muscles, les modèles splines et les déformations de forme libre. Les manipulations d'images comprennent le morphing d'images entre des images photographiques, le mélange d'images et la génération des expressions.

Dans cette section, nous fournissons une vue d'ensemble sur la génération d'image par GANs avec un accent particulier sur les algorithmes et les applications pour la synthèse de visage. Nous abordons plusieurs techniques importantes et différents modèles de GAN proposés dans les recherches récentes. Les études sur l'animation et la génération d'expressions faciales ont été traitées de différentes manières, et pour différentes tâches.

Certains travaux ont tenté de générer une image donnée avec une catégorie d'expression arbitraire dans un espace continu, de sorte que les images générées aient l'expression cible, tout en préservant l'identité de l'image source. Cette tâche a un large éventail d'applications dans l'interaction homme-machine, la réalité virtuelle (Wei et al., 2019), la planification de la chirurgie faciale (Keeve et al., 1998), etc., et a attiré une attention considérable de la recherche ces dernières années (Choi et al., 2018; Otberdout et al., 2022; Pumarola et al., 2018).

De leur côté, Li et al. (X. Li & Yu, 2019) ont proposé un cadre pour la génération d'expressions faciales et le transfert de photos vers des dessins animés. Premièrement, Star-GAN a été entraîné séparément pour la génération d'expressions faciales et Cartoon-GAN pour le transfert de style de dessin animé. Ensuite, deux modèles empilés ont été présentés. Le modèle Stack-GAN-A, dans lequel Star-GAN est considéré comme le premier GAN et Cartoon-GAN le second, et le modèle Stack-GAN-B dans lequel l'ordre des GAN est inversé.

ExprGAN a été proposé par Ding et al. (Ding et al., 2018) , lequel est capable de synthétiser des expressions faciales avec une intensité contrôlable, et un réseau de contrôle d'expression est proposé pour apprendre le code d'expression. Toutefois, ExprGAN génère des images conditionnées par des étiquettes d'expression et des

valeurs d'intensité, contrairement à *G2GAN* qui utilise la géométrie du visage comme condition de contrôle qui n'est pas limitée à certains styles d'expression. Les méthodes génératives basées sur des modèles peuvent difficilement générer des images de visage photoréalistes et préservant l'identité tout en permettant un ajustement continu de l'expression cible (Song et al., 2018).

6.3 Le modèle StyleGAN et ses versions

En mars 2019, les chercheurs Karras et al., de chez Nvidia,ont lancé *StyleGAN* (Karras et al., 2019), une architecture alternative de générateur, qui a constitué un grand pas en avant dans le domaine de la modélisation générative inconditionnelle d'images basée sur les données. En effet, cette nouvelle architecture conduit à une séparation automatiquement apprise entre les caractéristiques de haut niveau telles que la pose ou l'identité lorsqu'elles sont entraînées sur des visages humains et les variations stochastiques telles que la structure des cheveux ou les taches de rousseur, ce qui permet des opérations d'interpolation et un mélange spécifique à l'échelle dans l'espace image. Il permet ainsi d'effectuer plusieurs tâches comme le mélange de styles, la variation stochastique, la séparation des effets globaux et de la stochasticité.

Les auteurs proposent, en s'appuyant sur la littérature sur le transfert de style, de modifier l'architecture du générateur par l'utilisation d'un réseau feed-forward (mapping network *f*) pour projeter et démêler l'entrée dans un espace latent intermédiaire (*W*), au lieu de fournir le code latent (*Z*) directement au générateur. Par la suite, une transformation affine peut être calculée à partir de W, pour contrôler directement la normalisation adaptative d'instance (AdaIN) après chaque convolution. En utilisant cette normalisation, l'information sur le style est injectée dans le réseau G d'une manière bien plus efficace qu'en utilisant un simple vecteur latent d'entrée. Ainsi, *W* sera encouragé à se spécialiser dans différents styles, par les paramètres de la transformation affine. Enfin, un bruit gaussien supplémentaire est ajouté à chaque carte de caractéristiques pour faciliter la génération de détails stochastiques.

Plusieurs versions de StyleGAN ont été lancées, qui sont des améliorations successives du modèle original. Ces améliorations concernent la qualité de l'image et les performances par rapport à la version précédente. La Figure 6-1 montre la différence entre le réseau traditionnel et le réseau générateur basé sur le style (StyleGAN), et sa version améliorée StyleGAN2. Afin de résoudre les problèmes d'artefacts en forme de bulles dans certaines images, StyleGAN2 (Karras, Laine, et al., 2020) a été proposé en décembre 2019, il a amélioré la modélisation inconditionnelle des images en termes de métriques de qualité de distribution et de qualité d'image. Les performances de formation de StyleGAN2 se sont également améliorées, la plupart de cette accélération

en apprentissage provient de la simplification du flux de données due à la démodulation de poids, à la régularisation paresseuse et aux optimisations du code.

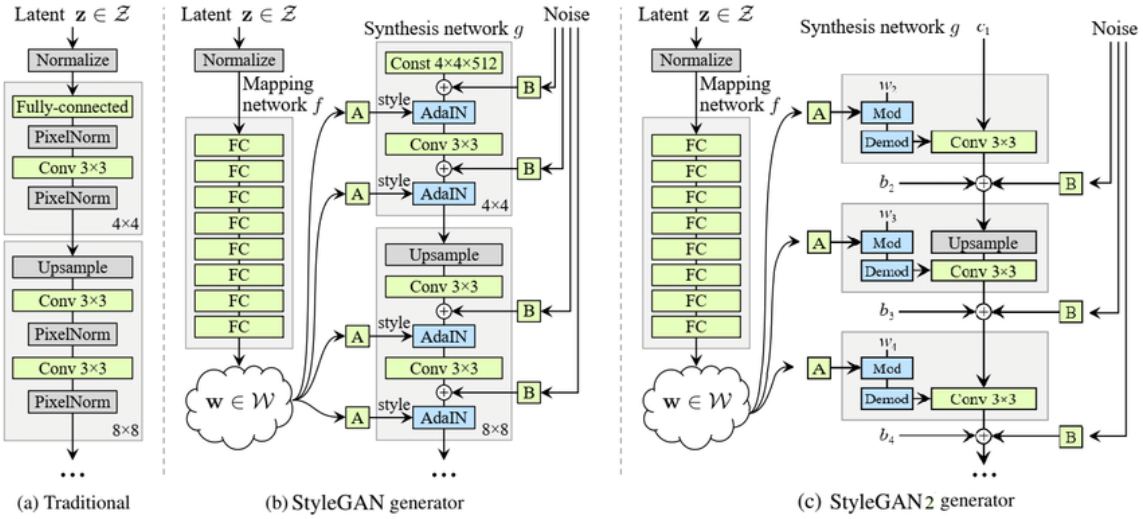


Figure 6-1 : Comparaison entre le générateur d'un GAN traditionnel, StyleGAN et StyleGAN2 (Karras, Laine, et al., 2020).

Enfin, la projection des images dans l'espace latent W fonctionne beaucoup mieux avec le nouveau générateur StyleGAN2 régularisé par la longueur du chemin qu'avec le StyleGAN original. Il est ainsi plus facile d'attribuer une image générée à sa source. Pour les humains, les images générées sont difficiles à distinguer des images réelles. Des exemples d'images synthétisées par StyleGAN2 sont disponibles sur https://thispersondoesnotexist.com/.

Afin d'élargir le champ d'application des GAN, les auteurs (Karras, Aittala, et al., 2020) se sont intéressés aux cas des données limités, où l'apprentissage des réseaux adversaires génératifs (GAN) provoque un surajustement du discriminateur, qui à son tour provoque une divergence de l'entraînement. Karras et al. (Karras, Aittala, et al., 2020) ont proposé d'utiliser un mécanisme d'augmentation adaptative du discriminateur pour stabiliser de manière significative la formation dans ces cas. Cette approche ne nécessite pas de modification des fonctions de perte ou des architectures de réseau, et elle est applicable à la fois lors de la formation à partir de zéro et lors du réglage fin d'un GAN existant sur un autre jeu de données.

Une dernière version a été lancée en Juin 2021, le StyleGAN3 (Karras et al., 2021). Son objectif est de s'attaquer au problème du " collage de texture " (aliasing) observé dans les GAN et de proposer une nouvelle architecture en considérant l'effet d'aliasing dans le domaine continu et en filtrant les résultats de façon passe-bas appropriée, ce qui est mieux adapté à la vidéo et à l'animation.

Pour StyleGAN et StyleGAN2, le nombre de couches L est déterminé par la taille de l'image de sortie R : $L = 2\,log_2\,R - 2$; il a également une résolution maximale de 1024x1024 avec 18 couches. Pour StyleGAN3, le nombre de couches est un paramètre libre et n'a pas de relation directe avec la résolution de sortie(Xia et al., 2023).

6.4 La démarche suivie

Dans le domaine de la conception et de la génération d'animations, le processus de génération d'expressions d'animation n'est pas évident en raison du manque de détails de l'image, ce qui entraîne un manque de réalisme des expressions d'animation générées. En vue de traiter ce problème, une méthode de génération d'expressions faciales de personnes synthétiques, basée sur l'apprentissage profond, est proposée. La méthode, basée sur les images d'expressions faciales réelles, utilise l'apprentissage profond, à savoir les GANs.

Tous les travaux susmentionnés sont des évaluations axées sur la performance. Ils animent l'image d'un seul visage à partir d'une vidéo entraînée qui pourrait fournir beaucoup d'informations(Y. Zhao et al., 2019), tandis que notre travail diffère en générant l'image d'un seul visage à partir d'étiquettes sémantiques abstraites, c'est-à-dire des étiquettes d'expression faciale, ce qui est une tâche plus difficile. Ces étiquettes sémantiques peuvent être considérées comme des paramètres d'animation plus intuitifs.

Une orientation populaire dans ce domaine consiste à modéliser des visages paramétriques en 3D, puis à les animer en 3D comme dans (Thies et al., 2016). Les approches dans ce domaine nécessitent souvent un post-traitement utilisant des techniques d'infographie, un équipement coûteux et une quantité importante de travail pour produire des résultats réalistes(Yi et al., 2020).

Afin de réduire le coût et le temps nécessaires à la production d'animations de haute qualité, les chercheurs se penchent sur des méthodes entièrement automatiques basées sur la 2D et utilisant des techniques d'apprentissage automatique. Ces dernières années, les méthodes basées sur la 2D ont fait des progrès significatifs dans l'animation de visages/têtes (Pumarola et al., 2018; Wiles et al., 2018), d'objets humains (C.-Y. Lin et al., 2019; Y. Zhao et al., 2019), de personnages de dessins animés (Sin et al., 2019) et d'autres objets (Siarohin et al., 2019) (Yi et al., 2020).

Des progrès considérables ont été réalisés dans le domaine de la génération d'images et de vidéos réalistes depuis l'introduction des réseaux adversaires génératifs (GAN) par Goodfellow et al. (I. Goodfellow et al., 2014). Or, pour contrôler le contenu de cette génération, il est souvent nécessaire de conditionner le processus en utilisant des GANs conditionnels (Mirza & Osindero, 2014). Concrètement, les GANs conditionnels

présentent un vif intérêt, dans la mesure où ils permettent la génération et le contrôle de nombreuses sorties au moyen d'un seul modèle. Quelques exemples d'applications des GANs conditionnels incluent la traduction d'image à image (J.-Y. Zhu et al., 2017), la génération conditionnée par classe (Brock et al., 2023), la manipulation d'images (Yu et al., 2018) et la génération de texte à image (T. Xu et al., 2018). Cependant, pour une génération réaliste et un entraînement stable, l'entraînement des GAN conditionnels nécessite de grandes données d'entraînement, incluant des étiquettes de conditionnement. La collecte de ces données rencontre souvent beaucoup de difficultés, principalement en raison de la confidentialité, de la qualité et de la diversité requises et de la collecte des étiquettes pour l'apprentissage conditionnel (Shahbazi et al., 2022).

L'animation faciale ne concerne pas seulement la transition fluide dans la séquence générée, elle doit également considérer les détails de l'expression et de l'identité du visage. Dans la plupart des méthodes d'animation d'expression existantes, des étiquettes d'expression continues sont utilisées, notamment des unités d'action (UA) ou des séquences de repères. Comparativement aux étiquettes d'expression discrètes, les annotations de nombre d'entre elles sont ambiguës et susceptibles de générer des erreurs. Cependant, l'animation d'une expression faciale conditionnée par des étiquettes d'expression discrètes est peu étudiée et les méthodes existantes ne peuvent pas générer des détails faciaux satisfaisants.

Dans ce contexte, nous nous intéressons particulièrement au problème de la génération d'expressions faciales. A cette fin, nous proposons d'utiliser un modèle de réseau adversarial génératif pré-entraîné StyleGAN2-ADA guidé par des expressions de visage d'entrée réel. De plus, nous proposons d'intégrer ces expressions dans StyleGAN2-ADA, ce qui en fait un GAN conditionnel. Ces expressions sont obtenues à partir d'un réseau CNN que nous proposons. Nous avons mené de nombreuses expérimentations, qui ont montré que cette méthode peut générer un visage synthétique réalisant l'expression du visage source, uniquement conditionné par des étiquettes discrètes générées par le CNN proposé.

6.4.1 Formulation du problème

Définissons une image d'entrée comme $I \in R^{H\times W\times 3}$, capturée sous une expression faciale arbitraire. Chaque expression est codée au moyen d'un ensemble de N (N=7) expressions de base $y_r = (y_1, \dots, y_N)^T$, où chaque y_i désigne une valeur normalisée entre 0 et 1 pour moduler la magnitude de la $i^{ième}$ expression de base.

Notre objectif est d'apprendre un appariement (mapping) A pour traduire l'image I en une image de sortie I_s conditionnée par l'expression y_r, c'est-à-dire que nous cherchons à estimer l'appariement A : $(I, y_r) \rightarrow I_s$. À cette fin, nous proposons d'entraîner A de manière non supervisée, en utilisant M paires d'entraînement

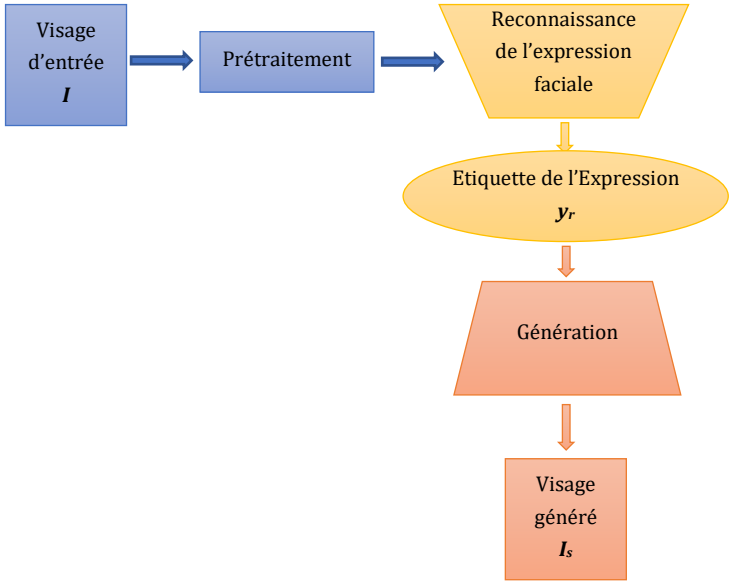


Figure 6-2 : Le schéma fonctionnel du système de transfert d'expression faciale proposé.

$\{I^m, y_r^m\}_{m=1}^{m=M}$. Il est important de noter que nous n'avons pas besoin de paires d'images de la même personne sous différentes expressions, ni de l'image cible attendue I_{yr}.

6.4.2 Architecture du système de génération de visage expressif

Pour générer un visage synthétique qui imite l'expression faciale réalisée par un visage source, nous entraînons deux réseaux indépendants (comme le montre la Figure 6-2) : un réseau de reconnaissance d'expression faciale et un réseau de génération de visage expressif. Le premier réseau est entraîné à reconnaître et à générer une étiquette de l'expression exprimée par le visage d'entrée, tandis que le second est entraîné à générer un visage expressif synthétique à partir de l'étiquette générée par le premier réseau. Après l'entraînement des deux réseaux, leur combinaison permet de générer le visage expressif synthétique à partir du visage source réel.

Après l'étape d'apprentissage, le système de transfert est prêt à générer les images de visages synthétiques, en utilisant ces deux modules : le module de reconnaissance des expressions faciales et le module de génération de visages synthétiques. Tout d'abord, le module de reconnaissance des expressions faciales génère une étiquette d'expression faciale à partir d'une image faciale frontale source, par un réseau CNN entrainé sur la base de données FER-2013. Ensuite, l'étiquette obtenue est utilisée par le module de génération de visage pour construire un visage synthétique réalisant la même expression que le visage source, par un générateur du modèle StyleGAN2. Nous utilisons StyleGAN2 dans ce travail, car il atteint une qualité visuelle de pointe sur des images à haute résolution avec un minimum d'artefacts. La Figure 6-2 montre la vue d'ensemble

de notre pipeline de ce système de transfert. Dans ce qui suit, une description détaillée de chacun de ces modules est fournie.

6.4.2.1 Le module de reconnaissance d'expression faciale

L'objectif de ce module est d'analyser et de reconnaître l'expression réalisée par l'image d'entrée, qui est une image frontale d'un visage réel. Les expressions reconnues sont les six expressions de base définies par Ekman et al. (Ekman & Friesen, 1971) : joie, colère, dégoût, peur, surprise, tristesse et l'expression neutre. Dans la littérature, différentes méthodes ont été proposées, des méthodes classiques et des méthodes basées sur l'apprentissage profond. Nous nous intéressons à ces dernières, dont un état de l'art détaillé est proposé dans le chapitre précédent, section 5.4.

Le modèle d'apprentissage profond utilisé est un CNN profond constitué d'un ensemble de blocs convolutionnels (quatre blocs) et un bloc entièrement connecté. Pour les blocs

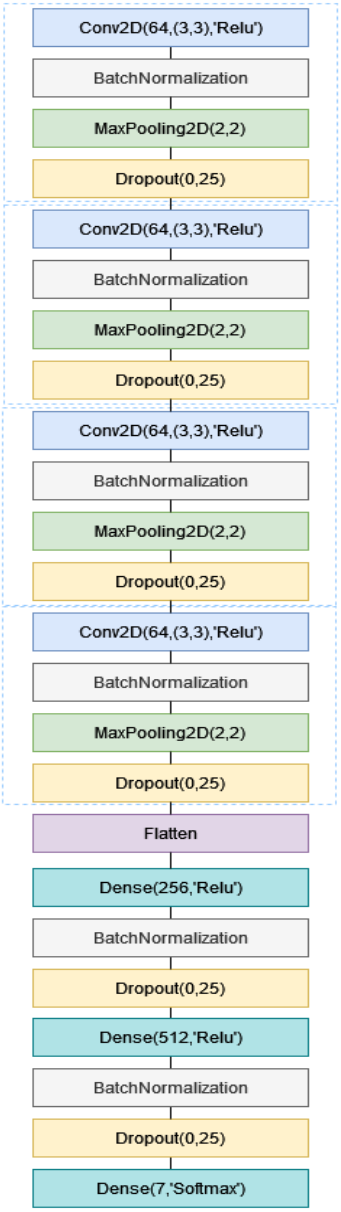


Figure 6-3 : Architecture du module de reconnaissance d'expressions faciales.

convolutionnels, chacun comprend une couche convolutive qui est suivie d'une normalisation par lot puis d'une couche de mise en commun maximale, chacune de ces couches est dotée d'une rectification non linéaire ('Relu'), tandis que le bloc entièrement connecté est constitué de trois couches entièrement connectées avec 256, 512 et 7 sorties, alors que la dernière couche adopte la fonction d'activation Softmax. Comme la montre la Figure 6-3.

6.4.2.2 Le module de génération de visage synthétique

L'objectif de ce module est de générer une image synthétique de visage à partir d'une étiquette représentant une classe d'expression faciale (joie, colère, dégoût, peur, surprise, tristesse et expression neutre). Comme mentionné précédemment, plusieurs travaux ont été réalisés, et différents générateurs ont été proposés et testés. Dans ce travail, nous avons opté pour le générateur pré-entrainé de styleGAN2, en raison de la qualité supérieure de ses images générées.

L'idée initiale est de contrôler le processus de transfert de l'expression faciale du visage d'entrée vers un autre visage en agissant sur StyleGan, à travers son espace latent, afin de favoriser son succès sans exiger de conditions. Pour ce faire, nous nous servons de l'expression faciale prédite par le système FER comme espace latent initial (vecteur de 7 valeurs réelles), pour ensuite constituer l'espace latent du générateur pré-entraîné StyleGan2. Ce dernier sera utilisé comme entrée pour le générateur StyleGan2 pré-

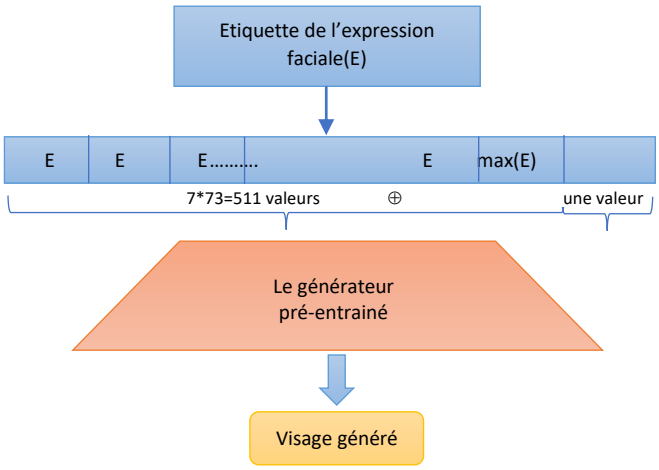


Figure 6-4 : Approche de génération de visage par StyleGAN2 pré-entrainé.

entraîné.

Sachant que chaque expression faciale reconnue par le module FER est un vecteur *E* de 7 valeurs réelles, et que l'entrée du générateur de StyleGAN2 pré-entrainé est un vecteur de 512 valeurs réelles (ce vecteur est habituellement crées aléatoirement), nous passons par une étape de construction d'un nouvel espace latent comme suit : chaque vecteur E représentant une expression est dupliqué, puis concaténé 73 fois pour obtenir

un vecteur de 511 valeurs, auquel est concaténée la valeur maximale contenue dans le vecteur E. Cette approche est illustrée à la Figure 6-4.

Malheureusement, les résultats ne correspondaient pas nécessairement à nos attentes, bien que le générateur ait réussi à produire des visages expressifs de haute qualité, mais avec des expressions différentes de celles exprimées par les visages d'entrée.

Après les résultats inespérés du générateur pré-entraîné, nous avons tenté d'intégrer l'expression prédite du visage d'entrée en tant que condition pour le générateur. Or, dans sa version actuelle, le modèle StyleGan2 n'est pas un modèle conditionnel, raison pour laquelle nous avons dû envisager un autre modèle conditionnel. En effet, il s'agit de parvenir à exploiter l'information de classe en guise de contrôle supplémentaire pour augmenter ainsi la qualité de la génération du visage. Dans cette optique, nous avons opté pour le modèle StyleGAN2 avec augmentation adaptative des données (ADA), qui est une méthode de pointe permettant de générer des images inconditionnelles et conditionnelles par classe dans un contexte de limitation des données.

Hélas, dans des contextes où les données sont limitées, le conditionnement par classe entraîne un effondrement des modes. Pour cela, Shahbazi et al. (Shahbazi et al., 2022) suggèrent une nouvelle approche d'entraînement des GANs avec conditionnement par classe (cGANs) permettant d'éviter efficacement le phénomène observé d'effondrement des modes. Dans cette approche, les GANs sont inconditionnels au départ et le conditionnement par classe est progressivement injecté au générateur ainsi qu'à la fonction objective. Grâce à cette proposition, non seulement la méthode d'entraînement avec des données limitées a été stable, mais elle a aussi permis de générer des images de haute qualité, grâce à l'exploitation précoce des informations partagées entre les classes.

6.5 Détail de l'implémentation

La mise en œuvre de l'architecture proposée a été réalisée dans le cadre d'un projet de fin d'études universitaires à travers une application expérimentale. La mise en œuvre est réalisée à l'aide de la bibliothèque Keras (*Keras*, 2015/2022) avec un backend TensorFlow (Abadi et al., 2016). La bibliothèque OpenCV (« OpenCV Face Recognition », s. d.) a servi à effectuer les différentes opérations sur les images. Les expérimentations ont été réalisées avec PyTorch et entraînées avec Google Colaboratory (*Google Colaboratory*, s. d.). En tenant compte des restrictions imposées par la gratuité du compte Google Colab, notamment un maximum de 12 heures par session d'entraînement, le nombre et le type de GPU ou la capacité VRAM, le processus de

formation a été réalisé en utilisant plusieurs paramètres, qui seront détaillés pour chaque module dans ce qui suit.

L'entraînement du système proposé a été effectué en deux étapes : dans un premier temps, le module de reconnaissance des expressions faciales a été entraîné et évalué, puis le second module, à savoir le module de génération de visages, a été entraîné.

6.5.1 Implémentation et apprentissage du module FER

L'apprentissage de ce module est effectué sur la base de données FER-2013, précédemment présenté dans le chapitre 5. Cet ensemble est composé de (35887 images en niveaux de gris de 48x48 pixels, ces images sont étiquetées en sept classes, comme suit : 0=colère, 1=dégoût, 2=peur, 3=joie, 4=triste, 5=surprise, 6=neutre. Nous avons testé plusieurs configurations, présentées dans le Tableau 6-1 ci-dessous :

Tableau 6-1 : Configurations expérimentales pour le module FER.

	Nombre d'époch	Batch size	Taux d'apprentissage
Test 1	45	128	0,6
Test 2	40	128	0,3
Test 3	35	32	0,7
Test 4	45	128	0,1
Test 5	45	64	0,1
Test 6	45	64	0,8

6.5.2 Implémentation et apprentissage du module de génération de visages

L'entraînement du second module de génération de visages est également réalisé sur la base de données FER-2013, mais il est nécessaire de procéder à quelques ajustements et pré-traitements sur cette base. Les étapes à suivre sont :

- Réduire la taille de la base de données FER-2013 à 200 images, sélectionnées aléatoirement, par classe.
- Redimensionner les images de la base de données à 64*64, étant donné que styleGAN2 requiert que la dimension des images soit une puissance de 2 (les images de la base FER-2013 sont 48*48).
- Convertir les images au format RGB.
- Classer les images par chaque catégorie (classe) dans un fichier json, à placer dans le répertoire de la base de données.
- Compresser les images ajustées de la base de données et le fichier json au format zip.

Après la préparation des données d'apprentissage, l'entrainement progressif est lancé avec les paramètres suivants : cond=1, kimg=1500 --t_start_kimg=500 --t_end_kimg=1000 \ --gpus=1 \

--cond doit être défini à 1, afin que l'entraînement soit effectué de manière conditionnelle, t_start_kimg est le début de la transition de la formation inconditionnelle à la formation conditionnelle. Enfin, t_end_kimg est la fin de cette transition.

6.6 Evaluation expérimentale

Comme mentionné ci-dessus, des expérimentations ont été menées pour déterminer l'efficacité du système proposé. Ainsi, les deux modules sont testés sur le jeu de données FER-2013 largement utilisé.

6.6.1 Évaluation du module FER proposé

Les résultats expérimentaux du module FER entraîné sur le jeu de données FER-2013 sont donnés dans le Tableau 6-2.

Tableau 6-2 : Performance du module FER proposé.

	Précision	Perte
Test 1	0,67	0,92
Test 2	0,67	0,91
Test 3	0,68	0,94
Test 4	0,66	0,92
Test 5	0,65	0,94
Test 6	0,68	1,22

Suite à la réalisation de plusieurs tests, l'expérience 5 a obtenu une précision de 65%. Dans le cas de l'expérience 3 et de l'expérience 6, nous avons obtenu une meilleure précision de 68% respectivement, mais selon les courbes de précision et de perte (Figure 6-6, Figure 6-7), ces tests présentent un sur-apprentissage. Pour cette raison, nous avons opté pour le CNN du test N°5 (Figure 6-5).

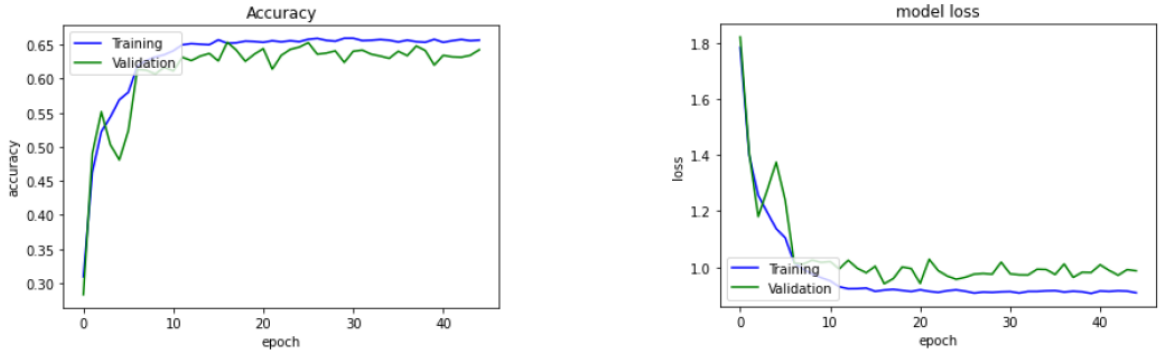


Figure 6-5 : Les courbes de précision et de perte de l'expérimentation 5.

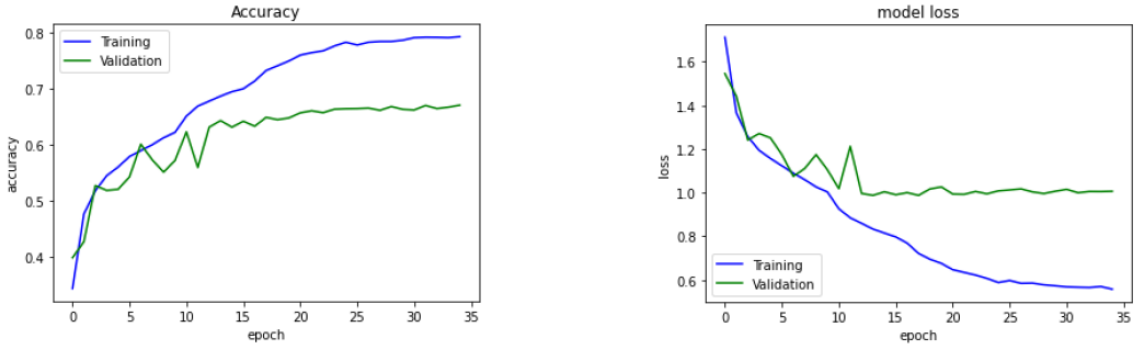


Figure 6-6 : Les courbes de précision et de perte de l'expérimentation 3.

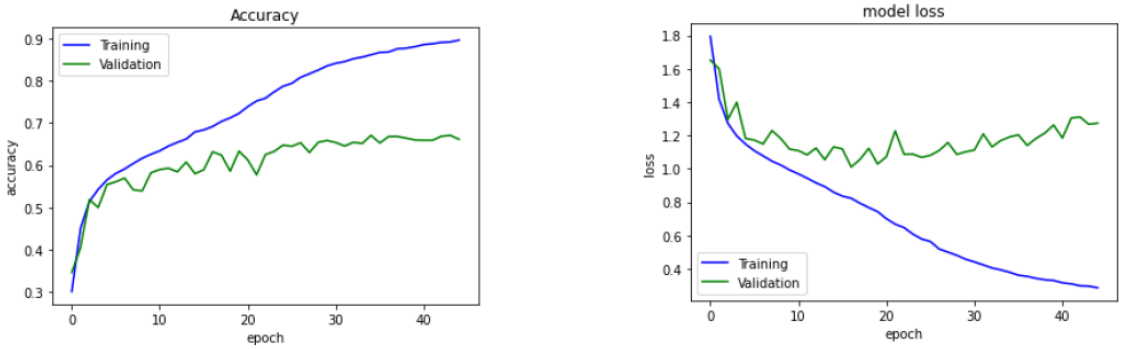


Figure 6-7 : Les courbes de précision et de perte de l'expérimentation 6.

Pour tester ce module, quelques images de visage de l'ensemble de donnée CK+ ont été utilisé, les résultats obtenus sont illustrés dans la Figure 6-8.

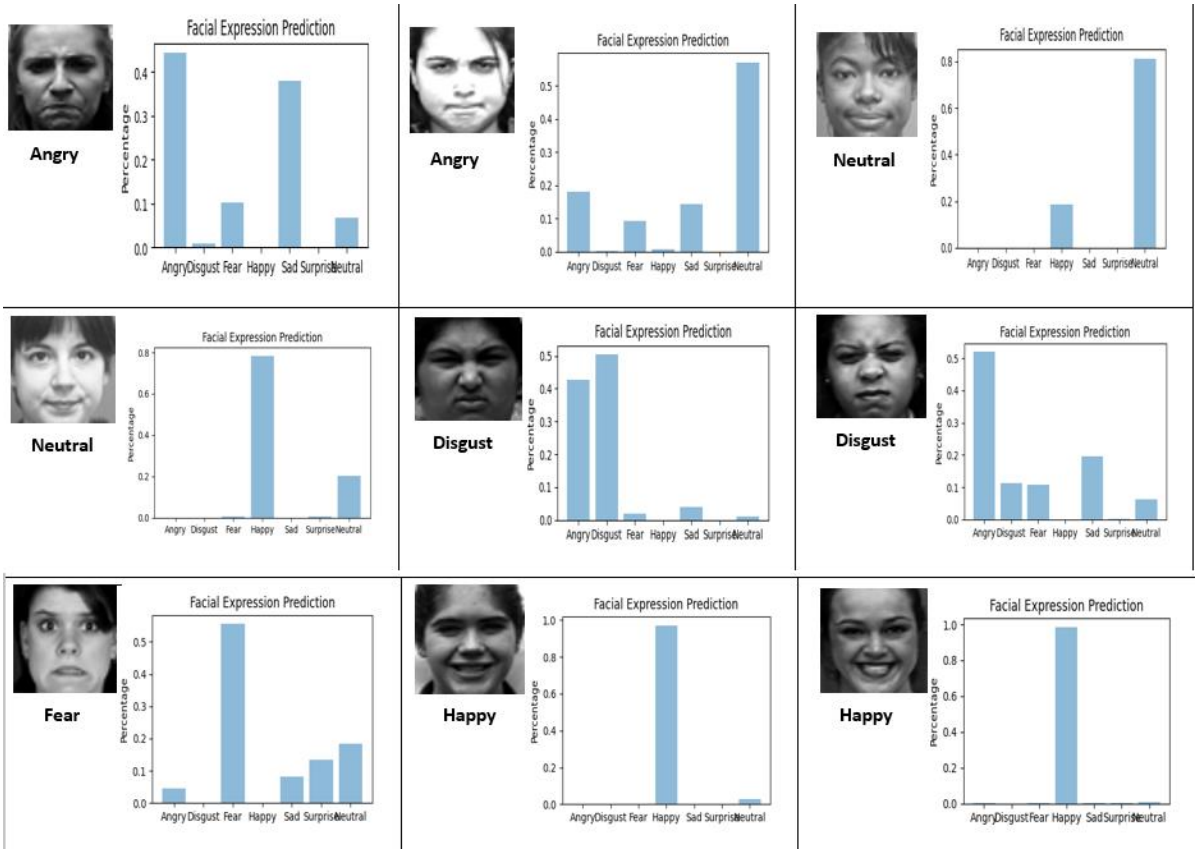


Figure 6-8: Quelques tests du module FER sur des images de CK+.

La matrice de confusion, présentée à la Figure 6-9, montre que les expressions "joie" et "surprise" sont de loin les plus aisément reconnaissables, avec une précision de plus de 80%, tandis que l'expression "peur" est la plus difficile à reconnaître, avec une précision de 29%. Il faut noter qu'il est parfois difficile de distinguer une expression de tristesse d'une expression de peur, même pour les humains, étant donné que tous les individus ne manifestent pas leurs émotions de la même manière.

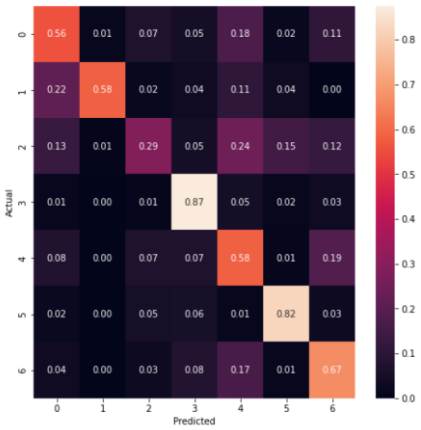


Figure 6-9 : Matrice de confusion de l'expérimentation 5.

6.6.2 Évaluation du module de génération de visage proposé

Les deux idées proposées ont été testées. L'utilisation du générateur StyleGAN2 pré-entraîné (Figure 6-10) a été testée à plusieurs reprises et nous avons constaté que les

images générées étaient de très haute qualité mais ne permettaient pas d'atteindre notre objectif, à savoir le transfert de l'expression faciale. En effet, dans la plupart des tests effectués, le résultat de cette méthode correspond à des visages aux expressions neutres, ce qui nous a poussé à faire appel à un générateur conditionnel.

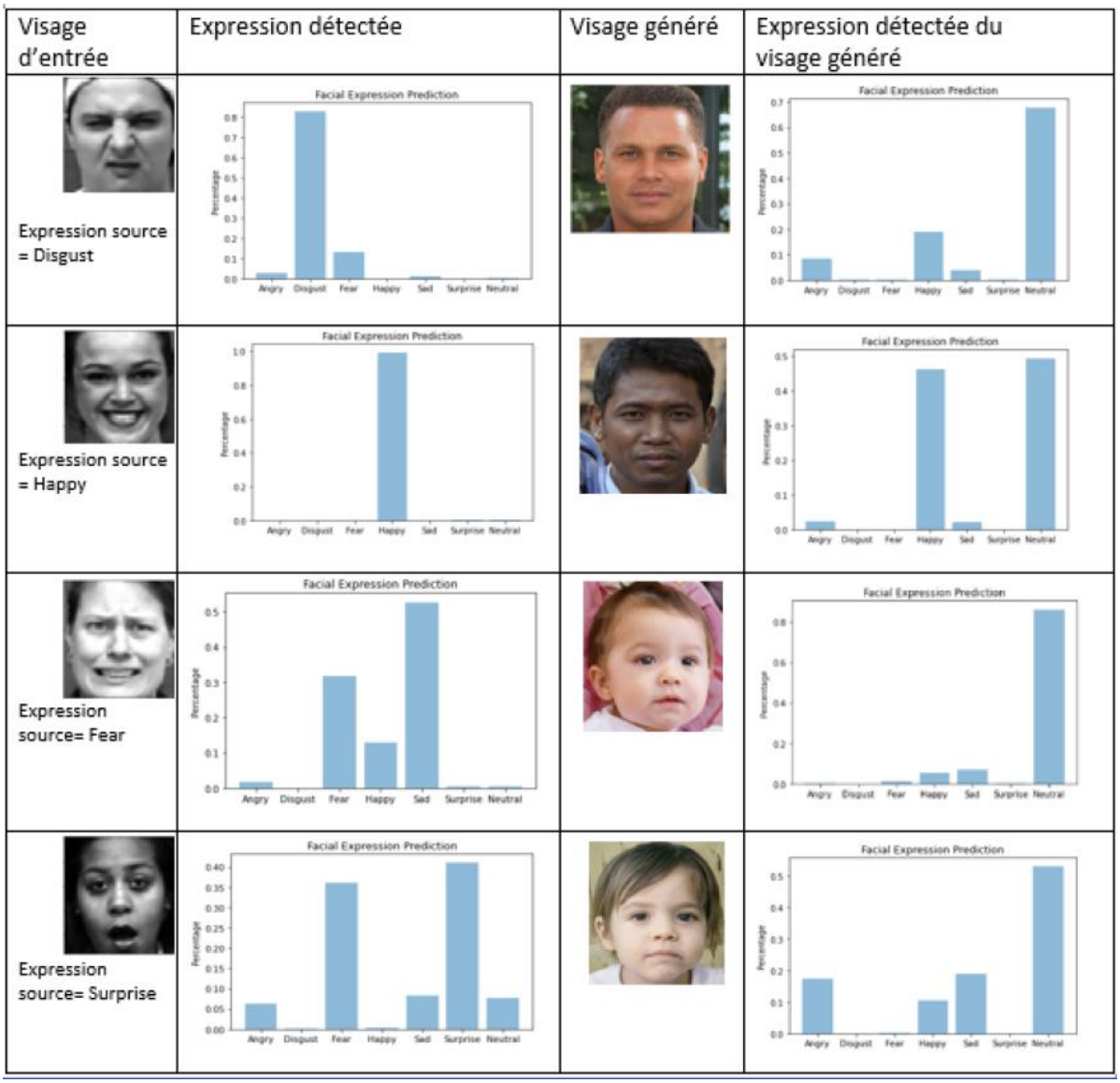


Figure 6-10 : Résultats de génération de visage avec StyleGAN2 pré-entrainé.

Après la phase d'entraînement, les résultats obtenus sont présentés dans la Figure 6-11.

Il est évident que les images produites ne sont pas aussi précises que celles générées par le modèle pré-entraîné StyleGAN2. Cela est dû à la nécessité d'utiliser au moins deux cartes graphiques (GPU) très puissantes et de s'entraîner pendant plusieurs jours pour obtenir un résultat de qualité supérieure.

Toutefois, les expressions des visages générés ressemblent en apparence aux expressions des visages d'entrée, et ce, malgré cette mauvaise qualité due au manque d'entraînement du modèle.

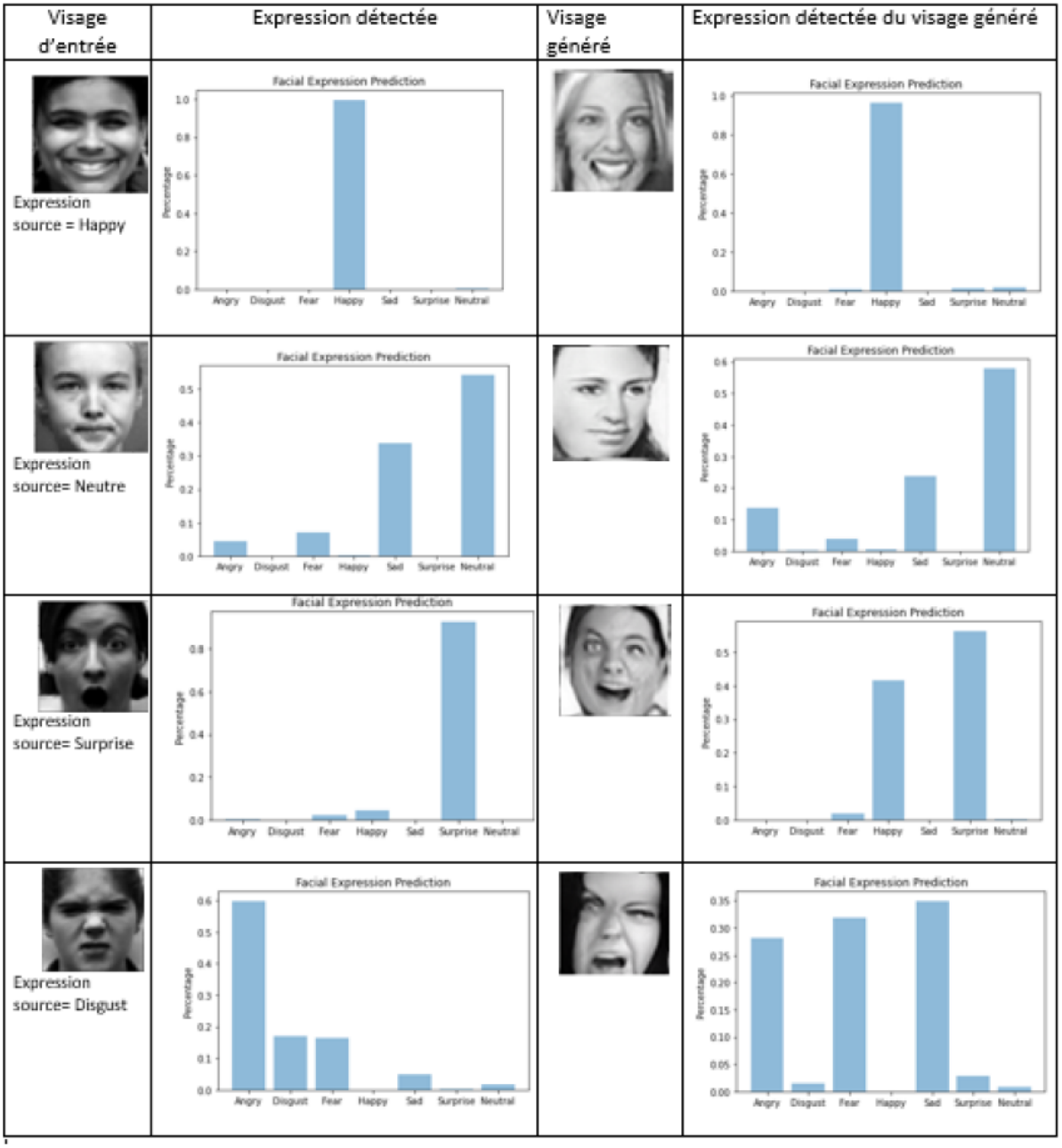


Figure 6-11 : Résultats de génération de visage avec StyleGAN2-ADA conditionnel.

Les résultats ainsi obtenus sont prometteurs et les performances de ce système pourraient être renforcées par un jeu de données plus important pour l'entraînement du générateur StyleGAN2-ADA. Par ailleurs, il faudrait disposer de machines et de dispositifs (comme Colab pro, non encore disponible en Algérie) plus puissants et dotés de capacités notables pour générer des images de haute qualité.

6.7 Conclusion

L'animation faciale est un véritable enjeu pour l'industrie de l'infographie. Comme les diverses émotions complexes doivent être exprimées par des déformations et des animations faciales différentes, le transfert des déformations faciales d'un personnage existant à un autre est largement requis, aussi bien dans l'industrie que dans la recherche, afin de réduire le travail manuel répétitif et fastidieux.

Dans ce travail, nous nous sommes concentrés sur le problème spécifique du transfert d'expression faciale d'un visage réel à un visage synthétique, nous utilisons module de reconnaissance d'expression faciale pour détecter l'émotion exprimée par la personne source, et sans avoir à détecter les repères faciaux, ni à effectuer l'appariement entre le visage source et le visage cible, nous générons un visage synthétique réalisant l'expression détectée.

Les expérimentations réalisées montrent l'intérêt de la méthode que nous proposons comme moyen simple et efficace d'obtenir des résultats de bonne qualité, crédibles et compétitifs de génération de visages expressifs synthétiques sans le moindre travail manuel.

Conclusion générale et perspectives

Les personnages virtuels peuvent jouer un rôle important dans l'interaction entre les humains et les ordinateurs. Le visage humain est une partie importante du corps qui exprime des émotions ainsi que l'identification pendant les conversations. Par exemple, ces dernières années, la communication a suscité beaucoup d'attention dans les services de réseaux sociaux en utilisant des avatars, qui sont des personnages virtuels exprimant des animations synchronisées par les lèvres avec des expressions émotionnelles dans un environnement de réalité virtuelle.

Il est difficile de créer des expressions faciales animées qui communiquent naturellement comme des personnes réelles. Il faut être capable d'exprimer des émotions très similaires à celles des vrais visages. De nombreuses études ont présenté différents moyens de créer des animations faciales à partir de différentes sources, telles que la parole, l'image ou la vidéo. L'objectif est toujours d'obtenir une animation faciale plus naturelle et plus réaliste.

Les techniques classiques d'animation faciale, depuis les premiers travaux de F. Parke, ont tenté d'atteindre cet objectif, mais il s'est avéré que ce sont des méthodes lourdes qui souffrent de plusieurs lacunes, la plus importante étant le manque de réalisme. En outre, les approches les plus efficaces devaient néanmoins extraire des caractéristiques des images sources, passer par des étapes de mise en correspondance ou utiliser des équipements spécialisés coûteux. Au cours des deux dernières décennies, et grâce aux progrès fulgurants de l'apprentissage profond, la génération de visages a connu un vif intérêt et beaucoup de succès.

Le transfert d'expressions faciales à partir d'une seule image est une tâche difficile qui a suscité beaucoup d'attention dans les domaines de la vision par ordinateur et de l'infographie. C'est un domaine de recherche pour reproduire et générer des images souhaitées d'un personnage et d'une expression faciale spécifiques. Récemment, les réseaux adversaires génératifs (GAN) ont été utilisés comme une nouvelle approche du transfert d'expressions faciales et ont obtenu des résultats significatifs pour les images à haute résolution, en s'appuyant sur des repères faciaux ou des unités d'action pour représenter l'expression à transférer. Cependant, il est encore difficile de générer des images de visages expressifs qui soient fidèles aux expressions du visage source.

Dans cette thèse, nous présentons une nouvelle méthode basée sur les GAN pour générer des expressions faciales émotionnelles à partir d'une seule image et sans repères faciaux, mais basée sur la détection de l'expression faciale source, qui est

incorporée dans un modèle GAN pour contrôler l'expression faciale générée à partir d'un espace latent.

Nous nous sommes particulièrement intéressés à cette première étape, qui est la reconnaissance des expressions faciales. Dans le but d'améliorer les performances des systèmes de reconnaissance des expressions faciales (FER), nous avons proposé un modèle hybride combinant des caractéristiques extraites à partir de CNNs afin d'assurer la complémentarité et la diversité, ainsi que de profiter des avantages de l'apprentissage par transfert dans la classification pour les applications FER. Le modèle à double branche proposé a été implémenté, testé et comparé à deux autres modèles, un CNN classique entraîné et un modèle récent avec transfert d'apprentissage, le modèle EffecientNet avec toutes ses variantes. Les expérimentations et l'évaluation des modèles sur deux jeux de données (Fer-2013 et CK+) ont abouti à des résultats motivants et prometteurs pour les deux jeux de données. Ils sont compétitifs et offrent de meilleures performances que des travaux existants.

La limite de la taille des jeux de données FER et le déséquilibre des images des différentes classes, qui ne favorisent pas l'apprentissage profond, constituent le principal obstacle. Pour pallier cette limitation, deux types d'apprentissage dans le même modèle sont simultanément mis en œuvre. L'apprentissage des modèles EfficientNets est raffiné ; en effet, ces modèles sont déjà entraînés pour le jeu de données Imagenet, sur les jeux de données FER-2013 et CK+, puis le vecteur de caractéristiques résultant est concaténé à celui obtenu à partir d'un CNN classique basé sur une architecture VGG très connue et robuste.

La seconde contribution de cette thèse concerne la génération de visage expressif synthétique, en utilisant uniquement un visage réel comme entrée. Pour cela, nous avons exploré plusieurs travaux traitant du sujet, puis proposé et testé un système qui atteint cet objectif en passant par deux étapes, la reconnaissance de l'expression faciale faite par les entrées du visage réel, puis l'utilisation de l'étiquette détectée comme condition pour le modèle StyleGan2-ADA. Ce dernier étant connu pour être un modèle inconditionnel, nous avons proposé dans son apprentissage d'intégrer progressivement une condition.

Certes, le résultat obtenu est encourageant, mais une plus longue formation permettra d'améliorer significativement la qualité des visages générés. Dans les travaux futurs, une étude plus approfondie sur le transfert séquentiel d'expressions est nécessaire.

Comme directions de recherche futures, nous envisageons de :

1. Exploration de la dernière version de StyleGAN3, pour la génération de séquence d’image de visage expressif.

2. Intégration de l'identité dans l'espace latent du générateur, en plus de l'étiquette de l'expression, ce qui permettra de générer une expression bien précise avec un visage bien précis.
3. Certaines améliorations peuvent être apportées à l'avenir, comme étendre la méthode à des modèles génériques, par exemple, le corps humain, la main ou le modèle animal.
4. L'utilisation des paramètres de mélange de formes comme entrées du générateur, et par conséquent l'hybridation de l'une des méthodes les plus classiques et les plus utilisées dans l'animation faciale avec un modèle récent et puissant, afin de bénéficier des avantages des deux.

Liste des productions scientifiques

Publication internationale

- Yamina Bordjiba, Hayet Farida Merouani and Nabiha Azizi, “Facial expression recognition via a jointly-learned dual-branch network”. International Journal of Electrical and Computer Engineering Systems, Volume 13, Number 6, 2022.

Communications internationales

- Yamina Bordjiba, Hayet Farida Merouani, "MPEG4 parameterization for facial deformation”, the third International Conference on Multimedia Computing and Systems (ICMCS), May 10-12, tangiers, Morocco, 2012.
- Yamina Bordjiba, Hayet Farida Merouani, “Facial feature points detection and tracking for facial animation retargeting” International Conference of Computing for Engineering and Sciences, Istanbul, Turkey, 29 July- 2 August 2015.
- Yamina Bordjiba, Hayat Merouani, Ali seridi, “A genetic algorithm for characteristic points matching of two different faces”, icraes'18, dec 23-25 hammamet, tunisia.
- Yamina Bordjiba, Hayet Farida Merouani and Nabiha Azizi, “A New Convolutional Neural Network for Facial Expression Recognition”, International Conference on Operating Systems, Cyber Security, Engineering Technology & Applied Sciences, 14-15 Dec 2019, Istanbul, Turkey.

Communications nationales

- Yamina Bordjiba, Merouani Hayet Farida, (2013), "Proposition d'une méthode de transfert d'une animation faciale à partir d'une vidéo réelle” ; journée nationale JSTIC, le 1 juillet 2013 Guelma, Algérie, 2013
- Yamina Bordjiba, Hayet Farida Merouani, F. Z. Adjabi, A. Guerib, « La mise correspondance entre les points caractéristiques de deux visages différents », SNSA 2016, Guelma.

Références

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). *TensorFlow : A System for Large-Scale Machine Learning*. 265-283.

Abrantes, G. A., & Pereira, F. (1999). MPEG-4 facial animation technology : Survey, implementation, and results. *IEEE Transactions on Circuits and Systems for Video Technology*, *9*(2), 290-305. https://doi.org/10.1109/76.752096

Aghamaleki, J. A., & Ashkani Chenarlogh, V. (2019). Multi-stream CNN for facial expression recognition in limited training data. *Multimedia Tools and Applications*, *78*(16), 22861-22882.
https://doi.org/10.1007/s11042-019-7530-7

Alom, M. Z., & Taha, T. M. (2017). Network intrusion detection for cyber security using unsupervised deep learning approaches. *2017 IEEE National Aerospace and Electronics Conference (NAECON)*, 63-69.
https://doi.org/10.1109/NAECON.2017.8268746

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, *8*(3), Article 3. https://doi.org/10.3390/electronics8030292

Amidi, A. (2022). *Deep Learning cheatsheets for Stanford's CS 230* [Logiciel]. https://github.com/afshinea/stanford-cs-230-deep-learning/blob/4653bc01297b269edb19e844b01127ba13de59df/fr/pense-bete-petites-astuces-apprentissage-profond.pdf (Édition originale 2018)

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface : A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, *6*(2), 20.

An interactive multimodal facial animation system. (1993). [EPFL]. https://doi.org/10.5075/epfl-thesis-1183

Arai, K., Kurihara, T., & Anjyo, K. (1996). Bilinear interpolation for facial expression and metamorphosis in real-time animation. *The Visual Computer*, *12*(3), 105-116. https://doi.org/10.1007/BF01725099

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 214-223.

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time Convolutional Neural Networks for Emotion and Gender Classification. *arXiv:1710.07557 [cs]*. http://arxiv.org/abs/1710.07557

Ayadi, W., Elhamzi, W., Charfi, I., & Atri, M. (2021). Deep CNN for Brain Tumor Classification. *Neural Processing Letters*, *53*(1), 671-700. https://doi.org/10.1007/s11063-020-10398-2

Balci, K. (2004). Xface : MPEG-4 based open source toolkit for 3D Facial Animation. *Proceedings of the working conference on Advanced visual interfaces*, 399-402. https://doi.org/10.1145/989863.989935

Beier, T., & Neely, S. (1992). Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics*, *26*(2), 35-42.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy Layer-Wise Training of Deep Networks. *Advances in Neural Information Processing Systems*, *19*.

Benzebouchi, N. E., Azizi, N., & Ayadi, K. (2019). A Computer-Aided Diagnosis System for Breast Cancer Using Deep Convolutional Neural Networks. In H. S. Behera, J. Nayak, B. Naik, & A. Abraham (Éds.), *Computational Intelligence in Data Mining* (p. 583-593). Springer. https://doi.org/10.1007/978-981-10-8055-5_52

Bian, S., Zheng, A., Gao, L., Maguire, G., Kokke, W., Macey, J., You, L., & Zhang, J. J. (2020). Fully Automatic Facial Deformation Transfer. *Symmetry*, *12*(1), Article 1. https://doi.org/10.3390/sym12010027

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning* (Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009)). Springer New York.

Bordjiba, Y., & Merouani, H. F. (2012). MPEG4 parameterization for facial deformation. *2012 International Conference on Multimedia Computing and Systems*, 414-419. https://doi.org/10.1109/ICMCS.2012.6320283

Bordjiba, Y., & Merouani, H. F. (2013, juillet 1). Proposition d'une méthode de transfert d'une animation faciale à partir d'une vidéo réelle.. *JSTIC 2013*.

Bordjiba, Y., & Merouani, H. F. (2015, juillet 29). Facial feature points detection and tracking for facial animation retargeting.. *International Conference of Computing for Engineering and Sciences*. International Conference of Computing for Engineering and Sciences, Istanbul, Turkey.

Bordjiba, Y., Merouani, H. F., & Azizi, N. (2022). Facial expression recognition via a jointly-learned dual-branch network. *International Journal of Electrical and Computer Engineering Systems*, *13*(6), Article 6. https://doi.org/10.32985/ijeces.13.6.4

Bordjiba, Y., Merouani, H. F., & Azizi, N. (2019, décembre 14). *A New Convolutional Neural Network for Facial Expression Recognition*. International Conference on Operating Systems, Cyber Security, Engineering Technology & Applied Sciences, Istanbul Turkey.

Bordjiba, Y., Merouani, H. F., F.Z, A., & A, G. (2016, décembre 15). *La mise en correspondance entre les points caractéristiques de deux visages différents*. Seminaire National sur la simulation Numérique dans les Sciences Appliquées, Guelma, Algérie.

Bordjiba, Y., Merouani, H. F., & Seridi, A. (2018, décembre 23). *A genetic algorithm for characteristic points matching of two different faces.* The 3rd International Conference on Recent Advances in Electrical Systems (ICRAES'18), Hammamet, Tunisia.

Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, *179*, 41-65. https://doi.org/10.1016/j.cviu.2018.10.009

Borshukov, G., Piponi, D., Larsen, O., Lewis, J. P., & Tempelaar-Lietz, C. (2005). Universal capture—Image-based facial animation for « The Matrix Reloaded ». *ACM SIGGRAPH 2005 Courses*, 16-es. https://doi.org/10.1145/1198555.1198596

Bregler, C., Covell, M., & Slaney, M. (1997). Video Rewrite : Driving visual speech with audio. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 353-360. https://doi.org/10.1145/258734.258880

Breuer, R., & Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*.

Brock, A., Donahue, J., & Simonyan, K. (2023, janvier 29). *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. International Conference on Learning Representations.

Brophy, E., Wang, Z., & Ward, T. E. (2019). *Quick and Easy Time Series Generation with Established Image-based GANs* (arXiv:1902.05624). arXiv. https://doi.org/10.48550/arXiv.1902.05624

Canedo, D., & Neves, A. J. R. (2019). Facial Expression Recognition Using Computer Vision : A Systematic Review. *Applied Sciences*, *9*(21), Article 21. https://doi.org/10.3390/app9214678

Cao, S., Yao, Y., & An, G. (2020). E2-capsule neural networks for facial expression recognition using AU-aware attention. *IET Image Processing*, *14*(11), 2417-2424. https://doi.org/10.1049/iet-ipr.2020.0063

Cao, Y., Tien, W. C., Faloutsos, P., & Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Transactions on Graphics*, *24*(4), 1283-1302. https://doi.org/10.1145/1095878.1095881

Chen, J., Lv, Y., Xu, R., & Xu, C. (2019). Automatic social signal analysis : Facial expression recognition using difference convolution neural network. *Journal of Parallel and Distributed Computing*, *131*, 97-102. https://doi.org/10.1016/j.jpdc.2019.04.017

Chen, S. E., & Williams, L. (1993). View interpolation for image synthesis. *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 279-288.

Chen, Y.-M., Huang, F.-C., Guan, S.-H., & Chen, B.-Y. (2012). Animating Lip-Sync Characters With Dominated Animeme Models. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(9), 1344-1353. https://doi.org/10.1109/TCSVT.2012.2201672

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). *StarGAN : Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. 8789-8797. https://openaccess.thecvf.com/content_cvpr_2018/html/Choi_StarGAN_Unified_Generative_CVPR_2018_paper.html

Chollet, F. (2017a). Xception : Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800-1807. https://doi.org/10.1109/CVPR.2017.195

Chollet, F. (2017b). Xception : Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251-1258.

Claudon, P., & Weber, M. (2009). L'émotion. Contribution à l'étude psychodynamique du développement de la pensée de l'enfant sans langage en interaction. *Devenir*, *21*(1), 61-99. https://doi.org/10.3917/dev.091.0061

Cohen, D. (2019). *Étude des émotions à travers les expressions faciales, intérêt dans les soins en médecine bucco-dentaire : Données actuelles de la littérature*. 95.

Cohen, M. M., & Massaro, D. W. (1993a). Modeling Coarticulation in Synthetic Visual Speech. In N. M. Thalmann & D. Thalmann (Éds.), *Models and Techniques in Computer Animation* (p. 139-156). Springer Japan. https://doi.org/10.1007/978-4-431-66911-1_13

Cohen, M. M., & Massaro, D. W. (1993b). Modeling Coarticulation in Synthetic Visual Speech. In N. M. Thalmann & D. Thalmann (Éds.), *Models and Techniques in Computer Animation* (p. 139-156). Springer Japan. https://doi.org/10.1007/978-4-431-66911-1_13

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160-167. https://doi.org/10.1145/1390156.1390177

Coquillart, S. (1990). Extended free-form deformation : A sculpturing tool for 3D geometric modeling. *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, 187-196. https://doi.org/10.1145/97879.97900

Coşkun, M., Uçar, A., Yildirim, Ö., & Demir, Y. (2017). Face recognition based on convolutional neural network. *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, 376-379. https://doi.org/10.1109/MEES.2017.8248937

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Davis, L. (ed). (1991). *Handbook of genetic algorithms*. CUMINCAD. http://cumincad.scix.net/cgi-bin/works/Show?eaca

Deena, S., & Galata, A. (2009). Speech-Driven Facial Animation Using a Shared Gaussian Process Latent Variable Model. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, Y. Kuno, J. Wang, J.-X. Wang, J. Wang, R. Pajarola, P. Lindstrom, A. Hinkenjann, M. L. Encarnação, C. T. Silva, & D. Coming (Éds.), *Advances in Visual Computing* (p. 89-100). Springer. https://doi.org/10.1007/978-3-642-10331-5_9

Deng, Z., Chiang, P.-Y., Fox, P., & Neumann, U. (2006). Animating blendshape faces by cross-mapping motion capture data. *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, 43-48. https://doi.org/10.1145/1111411.1111419

Deng, Z., & Noh, J. (2008). Computer Facial Animation : A Survey. In Z. Deng & U. Neumann (Éds.), *Data-Driven 3D Facial Animation* (p. 1-28). Springer. https://doi.org/10.1007/978-1-84628-907-1_1

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2106-2112. https://doi.org/10.1109/ICCVW.2011.6130508

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia*, *19*(3), 34-41. https://doi.org/10.1109/MMUL.2012.26

Dhere, S., Rathod, S. B., Aarankalle, S., Lad, Y., & Gandhi, M. (2020). A Review on Face Reenactment Techniques. *2020 International Conference on Industry 4.0 Technology (I4Tech)*, 191-194. https://doi.org/10.1109/I4Tech48345.2020.9102668

Dif, N. (2020). *L'apprentissage profond pour le traitement des images* [Theses, Djillali Liabes University]. https://hal.archives-ouvertes.fr/tel-03107095

Ding, H., Sricharan, K., & Chellappa, R. (2018). ExprGAN : Facial Expression Editing With Controllable Expression Intensity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.12277

Ding, H., Zhou, S. K., & Chellappa, R. (2017). FaceNet2ExpNet : Regularizing a Deep Face Recognition Net for Expression Recognition. *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 118-126. https://doi.org/10.1109/FG.2017.23

Doersch, C. (2021). *Tutorial on Variational Autoencoders* (arXiv:1606.05908). arXiv. https://doi.org/10.48550/arXiv.1606.05908

Dong, H., Yu, S., Wu, C., & Guo, Y. (2017). *Semantic Image Synthesis via Adversarial Learning*. 5706-5714. https://openaccess.thecvf.com/content_iccv_2017/html/Dong_Semantic_Image_Synthesis_ICCV_2017_paper.html

Duarte, A., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., Mohedano, E., McGuinness, K., Torres, J., & Giro-i-Nieto, X. (2019). Wav2Pix : Speech-conditioned Face Generation Using Generative Adversarial Networks. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8633-8637. https://doi.org/10.1109/ICASSP.2019.8682970

Duong, L. T., Nguyen, P. T., Di Sipio, C., & Di Ruscio, D. (2020). Automated fruit recognition using EfficientNet and MixNet. *Computers and Electronics in Agriculture*, *171*, 105326. https://doi.org/10.1016/j.compag.2020.105326

Dutreve, L. (2011). *Paramétrisation et transfert d'animations faciales 3D à partir de séquences vidéo : Vers des applications en temps réel* [Thèse de doctorat, Université Claude Bernard - Lyon I]. https://tel.archives-ouvertes.fr/tel-00863883

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., & Pal, C. (2015). Recurrent Neural Networks for Emotion Recognition in Video. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 467-474. https://doi.org/10.1145/2818346.2830596

Eisert, P., & Girod, B. (1998). Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics and Applications*, *18*(5), 70-78. https://doi.org/10.1109/38.708562

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system : Manual*. Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system*. A Human Face.

Ekmen, B., & Ekenel, H. K. (2019). From 2D to 3D real-time expression transfer for facial animation. *Multimedia Tools and Applications*, *78*(9), 12519-12535. https://doi.org/10.1007/s11042-018-6785-8

El-Sawy, A., EL-Bakry, H., & Loey, M. (2017). CNN for Handwritten Arabic Digits Recognition Based on LeNet-5. In A. E. Hassanien, K. Shaalan, T. Gaber, A. T. Azar, & M. F. Tolba (Éds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016* (p. 566-575). Springer International Publishing. https://doi.org/10.1007/978-3-319-48308-5_54

Essa, I., Basu, S., Darrell, T., & Pentland, A. (1996). Modeling, tracking and interactive animation of faces and heads//using input from video. *Proceedings Computer Animation '96*, 68-79. https://doi.org/10.1109/CA.1996.540489

Etienne, C. (2019). *Apprentissage profond appliqué à la reconnaissance des émotions dans la voix* [Phdthesis, Université Paris Saclay (COmUE)]. https://tel.archives-ouvertes.fr/tel-02479126

Fan, Y., Lin, Z., Saito, J., Wang, W., & Komura, T. (2022). *FaceFormer : Speech-Driven 3D Facial Animation With Transformers*. 18770-18780. https://openaccess.thecvf.com/content/CVPR2022/html/Fan_FaceFormer_Speech-Driven_3D_Facial_Animation_With_Transformers_CVPR_2022_paper.html

Farkas, L. G. (1994). *Anthropometry of the head and face* (2nd ed). Raven Press.

Fedus, W., Goodfellow, I., & Dai, A. (2018). *MaskGAN : Better Text Generation via Filling in the ____*. https://openreview.net/pdf?id=ByOExmWAb

Frank, M. G. (2001). Facial Expressions. In N. J. Smelser & P. B. Baltes (Éds.), *International Encyclopedia of the Social & Behavioral Sciences* (p. 5230-5234). Pergamon. https://doi.org/10.1016/B0-08-043076-7/01713-7

Fukushima, K. (1988). Neocognitron : A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119-130. https://doi.org/10.1016/0893-6080(88)90014-7

Gan, J., Wang, W., & Lu, K. (2019). A new perspective : Recognizing online handwritten Chinese characters via 1-dimensional CNN. *Information Sciences*, *478*, 375-390. https://doi.org/10.1016/j.ins.2018.11.035

Garchery, S. (2004). *Animation faciale temps réel multi plates-formes*. Université-Faculté des SES Département de systèmes d'information.

Geng, J., Shao, T., Zheng, Y., Weng, Y., & Zhou, K. (2018). Warp-guided GANs for single-photo facial animation. *ACM Transactions on Graphics*, *37*(6), 231:1-231:12. https://doi.org/10.1145/3272127.3275043

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep Learning Approaches for Facial Emotion Recognition : A Case Study on FER-2013. In I. Hatzilygeroudis & V. Palade (Éds.), *Advances in Hybridization of Intelligent Methods : Models, Systems and Applications* (p. 1-16). Springer International Publishing. https://doi.org/10.1007/978-3-319-66790-4_1

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580-587. https://doi.org/10.1109/CVPR.2014.81

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315-323. https://proceedings.mlr.press/v15/glorot11a.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., & Lee, D.-H. (2013). Challenges in representation learning : A report on three machine learning contests. *International conference on neural information processing*, 117-124.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 2672-2680.

Google Colaboratory. (s. d.). Consulté 20 novembre 2022, à l'adresse https://colab.research.google.com/notebooks/intro.ipynb

Guenter, B., Grimm, C., Wood, D., Malvar, H., & Pighin, F. (1998). Making faces. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 55-66. https://doi.org/10.1145/280814.280822

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved Training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

Hardy, C. (2019). *Contribution au développement de l'apprentissage profond dans les systèmes distribués* [Phdthesis, Université Rennes 1]. https://tel.archives-ouvertes.fr/tel-02284916

Hassan, M. ul. (2018, novembre 20). *VGG16—Convolutional Network for Classification and Detection*. https://neurohive.io/en/popular-networks/vgg16/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026-1034. https://doi.org/10.1109/ICCV.2015.123

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

Hoch, M., Fleischmann, G., & Girod, B. (1994). Modeling and animation of facial expressions based on B-Splines. *The Visual Computer*, *11*(2), 87-95. https://doi.org/10.1007/BF01889979

Hodgins, J. K., & O'Brien, J. F. (2003). Computer animation. In *Encyclopedia of Computer Science* (p. 301-304). John Wiley and Sons Ltd.

Hofer, G., & Richmond, K. (2010). *Comparison of HMM and TMD Methods for Lip Synchronisation*. https://era.ed.ac.uk/handle/1842/4558

Holland, J. H. (1992). *Adaptation in natural and artificial systems : An introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., & Wang, H.-M. (2018). Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117-128. https://doi.org/10.1109/TETCI.2017.2784878

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700-4708.

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. (2016). Deep Networks with Stochastic Depth. *arXiv:1603.09382 [cs]*. http://arxiv.org/abs/1603.09382

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., & Belongie, S. (2017). Stacked Generative Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1866-1875. https://doi.org/10.1109/CVPR.2017.202

Igeland, V. (2019). *Generating Facial Animation With Emotions In A Neural Text-To-Speech Pipeline*. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-160535

ImageNet. (2020, septembre 25). http://www.image-net.org/about-stats

ISO, I. (2001). *IEC 14496-2,"Coding of audio-visual objects-part2 : Visual," ISO*. IEC, Tech. Rep.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2013). Reply to Sauter and Eisner : Differences outweigh commonalities in the communication of emotions across human cultures. *Proceedings of the National Academy of Sciences*, *110*(3). https://doi.org/10.1073/pnas.1211865110

Jain, D. K., Shamsolmoali, P., & Sehdev, P. (2019). Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, *120*, 69-74.

Jamaludin, A., Chung, J. S., & Zisserman, A. (2019). You Said That? : Synthesising Talking Faces from Audio. *International Journal of Computer Vision*, *127*(11), 1767-1779. https://doi.org/10.1007/s11263-019-01150-y

Johnson, J. (2021). A Survey of Computer Graphics Facial Animation Methods : Comparing Traditional Approaches to Machine Learning Methods. *Master's Theses*. https://digitalcommons.calpoly.edu/theses/2315

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Éds.), *Computer Vision – ECCV 2016* (p. 694-711). Springer International Publishing. https://doi.org/10.1007/978-3-319-46475-6_43

Jolicoeur-Martineau, A. (2018). *The relativistic discriminator : A key element missing from standard GAN* (arXiv:1807.00734). arXiv. https://doi.org/10.48550/arXiv.1807.00734

Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2983-2991. https://doi.org/10.1109/ICCV.2015.341

Kähler, K., Haber, J., & Seidel, H.-P. (2001). Geometry-based muscle modeling for facial animation. *Proceedings of Graphics Interface 2001*, 37-46.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 655-665. https://doi.org/10.3115/v1/P14-1062

Kalra, P., Mangili, A., Thalmann, N. M., & Thalmann, D. (1992). Simulation of Facial Muscle Actions Based on Rational Free Form Deformations. *Computer Graphics Forum*, *11*(3), 59-69. https://doi.org/10.1111/1467-8659.1130059

Kammoun, A., Slama, R., Tabia, H., Ouni, T., & Abid, M. (2022). Generative Adversarial Networks for Face Generation : A Survey. *ACM Computing Surveys*, *55*(5), 94:1-94:37. https://doi.org/10.1145/3527850

Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics*, *36*(4), 94:1-94:12. https://doi.org/10.1145/3072959.3073658

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training Generative Adversarial Networks with Limited Data. *Advances in Neural Information Processing Systems*, *33*, 12104-12114. https://proceedings.neurips.cc/paper/2020/hash/8d30aa96e72440759f74bd2306c1fa3d-Abstract.html

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-Free Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, *34*, 852-863. https://proceedings.neurips.cc/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html

Karras, T., Laine, S., & Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks*. 4401-4410. https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and Improving the Image Quality of StyleGAN*. 8110-8119. https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html

Keeve, E., Girod, S., Kikinis, R., & Girod, B. (1998). Deformable modeling of facial tissue for craniofacial surgery simulation. *Computer Aided Surgery*, *3*(5), 228-238. https://doi.org/10.1002/(SICI)1097-0150(1998)3:5<228::AID-IGS2>3.0.CO;2-I

Keras : Deep Learning for humans. (2022). [Python]. Keras. https://github.com/keras-team/keras (Édition originale 2015)

Khalfi, F. A. (2010). *Reconnaissance automatique des émotions par données multimodales : Expressions faciales et des signaux physiologiques* [Thèse de doctorat, Université Paul Verlaine - Metz]. https://hal.univ-lorraine.fr/tel-01748925

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics*, *37*(4), 163:1-163:14. https://doi.org/10.1145/3197517.3201283

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751. https://doi.org/10.3115/v1/D14-1181

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., & Welling, M. (2014). Semi-supervised Learning with Deep Generative Models. *Advances in Neural Information Processing Systems*, *27*. https://papers.nips.cc/paper/2014/hash/d523773c6b194f37b938d340d5d02232-Abstract.html

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In Y. Bengio & Y. LeCun (Éds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. http://arxiv.org/abs/1312.6114

Ko, B. C. (2018a). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, *18*(2), Article 2. https://doi.org/10.3390/s18020401

Ko, B. C. (2018b). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, *18*(2), Article 2. https://doi.org/10.3390/s18020401

Koenen, R., Pereira, F., & Chiariglione, L. (1997). MPEG-4 : Context and objectives. *Signal Processing: Image Communication*, *9*(4), 295-304. https://doi.org/10.1016/S0923-5965(97)00003-9

Kouadio, C., Poulin, P., & Lachapelle, P. (1998). Real-time facial animation based upon a bank of 3D facial expressions. *Proceedings Computer Animation '98 (Cat. No.98EX169)*, 128-136. https://doi.org/10.1109/CA.1998.681917

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

Kumar, P. S., Uddin, M. S., & Bouakaz, S. (2012). Extraction of Facial Feature Points Using Cumulative Histogram. *IJCSI International Journal of Computer Science Issues*, *Vol. 9, Is*, 44.

Lacruz, R. S., Stringer, C. B., Kimbel, W. H., Wood, B., Harvati, K., O'Higgins, P., Bromage, T. G., & Arsuaga, J.-L. (2019). The evolutionary history of the human face. *Nature Ecology & Evolution*, *3*(5), Article 5. https://doi.org/10.1038/s41559-019-0865-7

Laine, S., Karras, T., Aila, T., Herva, A., Saito, S., Yu, R., Li, H., & Lehtinen, J. (2017). Production-level facial performance capture using deep convolutional neural networks. *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 1-10. https://doi.org/10.1145/3099564.3099581

Larochelle, H. (2008). *Étude de techniques d'apprentissage non-supervisé pour l'amélioration de l'entraînement supervisé de modèles connexionnistes* [Thèse de doctorat, Université de Montréal]. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/6435

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition : A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98-113. https://doi.org/10.1109/72.554195

Lecun, Y. (2016). Les enjeux de la recherche en intelligence artificielle. *Interstices*. https://hal.inria.fr/hal-01350469

LeCun, Y. (2016, février 29). *Les enjeux de la recherche en intelligence artificielle*. interstices. https://interstices.info/les-enjeux-de-la-recherche-en-intelligence-artificielle/

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Lecun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P., & Vapnik, V. (1995). Learning algorithms for classification : A comparison on handwritten digit recognition. In J. H. Oh, C. Kwon, & S. Cho (Éds.), *Neural networks* (p. 261-276). World Scientific.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. 4681-4690. https://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html

Lee, H., Battle, A., Raina, R., & Ng, A. (2006). Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems*, *19*. https://proceedings.neurips.cc/paper/2006/hash/2d71b2ae158c7c5912cc0bbde2bb9d95-Abstract.html

Lee, Y., Terzopoulos, D., & Waters, K. (1993). Constructing physics-based facial models of individuals [Application/pdf]. *Proceedings of Graphics Interface '93*, *Toronto*, 8 pages, 8.77 MB. https://doi.org/10.20380/GI1993.01

Lee, Y., Terzopoulos, D., & Waters, K. (1995). Realistic modeling for facial animation. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 55-62. https://doi.org/10.1145/218380.218407

Lewis, J. P., Mooser, J., Deng, Z., & Neumann, U. (2005). Reducing blendshape interference by selected motion attenuation. *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games - SI3D '05*, 25. https://doi.org/10.1145/1053427.1053431

Lewis, J. P., & Parke, F. I. (1986). Automated lip-synch and speech synthesis for character animation. *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*, 143-147. https://doi.org/10.1145/29933.30874

Li, B., Zhang, Q., Zhou, D., & Wei, X. (2013). Facial Animation Based on Feature Points. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *11*(3), Article 3.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., & Ng, S.-K. (2019). MAD-GAN : Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV*, 703-716. https://doi.org/10.1007/978-3-030-30490-4_56

Li, K., Jin, Y., Akram, M. W., Han, R., & Chen, J. (2020). Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. *The Visual Computer*, *36*(2), 391-404. https://doi.org/10.1007/s00371-019-01627-4

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition : A Survey. *IEEE Transactions on Affective Computing*, 1-1. https://doi.org/10.1109/TAFFC.2020.2981446

Li, X., & Yu, X. (2019). *Generating Cartoon Style Facial Expressions with StackGAN* (CS230 Group Project). Stanford University. https://cs230.stanford.edu/projects_fall_2019/reports/26242839.pdf

Li, X., Zhang, J., & Liu, Y. (2022). Speech driven facial animation generation based on GAN. *Displays*, *74*, 102260. https://doi.org/10.1016/j.displa.2022.102260

Lin, C.-Y., Huang, Y.-W., & Shih, T. K. (2019). Creating waterfall animation on a single image. *Multimedia Tools and Applications*, *78*(6), 6637-6653. https://doi.org/10.1007/s11042-018-6332-7

Lin, M., Chen, Q., & Yan, S. (2014). *Network In Network* (arXiv:1312.4400). arXiv. https://doi.org/10.48550/arXiv.1312.4400

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., & Mallya, A. (2021). Generative Adversarial Networks for Image and Video Synthesis : Algorithms and Applications. *Proceedings of the IEEE*, *109*(5), 839-862. https://doi.org/10.1109/JPROC.2021.3049196

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11-26. https://doi.org/10.1016/j.neucom.2016.12.038

Liu, Z. (2021). *Automated Deep Learning : Principles and Practice* [Phdthesis, Université Paris-Saclay]. https://tel.archives-ouvertes.fr/tel-03464519

Lu, T., Mu, D., & Tsai, M. (2007). Facial Expressions for 3D Game Applications. *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, *1*, 341-344. https://doi.org/10.1109/IIHMSP.2007.4457559

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

Lyons, Michael, Kamachi, Miyuki, & Gyoba, Jiro. (1998). *The Japanese Female Facial Expression (JAFFE) Dataset* [jeu de données]. Zenodo. https://doi.org/10.5281/ZENODO.3451524

Ma, F., Sun, B., & Li, S. (2021). Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing*, 1-1. https://doi.org/10.1109/TAFFC.2021.3122146

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose Guided Person Image Generation. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/34ed066df378efacc9b924ec161e7639-Abstract.html

Magnenat-Thalmann, N., Primeau, E., & Thalmann, D. (1988). Abstract muscle action procedures for human face animation. *The Visual Computer*, *3*(5), 290-297. https://doi.org/10.1007/BF01914864

Mao, X., & Li, Q. (2021). *Generative Adversarial Networks for Image Generation*. Springer. https://doi.org/10.1007/978-981-33-6048-8

Massaro, D. W., & Cohen, M. M. (1990). Perception of Synthesized Audible and Visible Speech. *Psychological Science*, *1*(1), 55-63. https://doi.org/10.1111/j.1467-9280.1990.tb00068.x

Maurice, B. (s. d.). Cours théoriques—Deep learning Archives. *Deeply Learning*. Consulté 1 novembre 2022, https://deeplylearning.fr/category/cours-theoriques-deep-learning/

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). Disfa : A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, *4*(2), Article 2.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), Article 4. https://doi.org/10.1609/aimag.v27i4.1904

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259

Miao, Y., Dong, H., Jaam, J. M. A., & Saddik, A. E. (2019). A deep learning system for recognizing facial expression in real-time. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(2), Article 2.

Mirza, M., & Osindero, S. (2014). *Conditional Generative Adversarial Nets* (arXiv:1411.1784). arXiv. https://doi.org/10.48550/arXiv.1411.1784

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). *Spectral Normalization for Generative Adversarial Networks* (arXiv:1802.05957). arXiv. https://doi.org/10.48550/arXiv.1802.05957

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1-10. https://doi.org/10.1109/WACV.2016.7477450

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet : A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.

Monjaux, P. (2007). *Modélisation et animation interactive de visages virtuels de dessins animés* [Phdthesis, Université René Descartes - Paris V]. https://theses.hal.science/tel-00274869

Moschoglou, S., Ploumpis, S., Nicolaou, M. A., Papaioannou, A., & Zafeiriou, S. (2020). 3DFaceGAN : Adversarial Nets for 3D Face Representation, Generation, and Translation. *International Journal of Computer Vision*, *128*(10), 2534-2551. https://doi.org/10.1007/s11263-020-01329-8

Moser, L., Chien, C., Williams, M., Serra, J., Hendler, D., & Roble, D. (2021). Semi-supervised video-driven facial animation transfer for production. *ACM Transactions on Graphics*, *40*(6), 222:1-222:18. https://doi.org/10.1145/3478513.3480515

Myers, D. G. (2004). Theories of emotion. *Psychology: Seventh Edition, New York, NY: Worth Publishers*, *500*.

Nahas, M., Huitric, H., Rioux, M., & Domey, J. (1990). Facial image synthesis using skin texture recording. *The Visual Computer*, *6*(6), 337-343. https://doi.org/10.1007/BF01901020

Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, *72*(2011), 1-19.

Ng, H.-W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 443-449.

https://doi.org/10.1145/2818346.2830593

Nguyen Quoc, T., & Truong Hoang, V. (2020). Medicinal Plant identification in the wild by using CNN. *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 25-29. https://doi.org/10.1109/ICTC49870.2020.9289480

Nie, W., Narodytska, N., & Patel, A. (2022, février 10). *RelGAN : Relational Generative Adversarial Networks for Text Generation*. International Conference on Learning Representations. https://openreview.net/forum?id=rJedV3R5tm

Noh, J., & Neumann, U. (1998a). *A survey of facial modeling and animation techniques*. USC Technical Report, 99–705.

Noh, J., & Neumann, U. (1998b). *A survey of facial modeling and animation techniques*. USC Technical Report, 99–705.

Noh, J., & Neumann, U. (1998c). *A survey of facial modeling and animation techniques*. USC Technical Report, 99–705.

Noh, J., & Neumann, U. (2001). Expression cloning. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 277-288. https://doi.org/10.1145/383259.383290

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-GAN : Training Generative Neural Samplers using Variational Divergence Minimization. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html

Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., & Matusik, W. (2019). *Speech2Face : Learning the Face Behind a Voice*. 7539-7548. https://openaccess.thecvf.com/content_CVPR_2019/html/Oh_Speech2Face_Learning_the_Face_Behind_a_Voice_CVPR_2019_paper.html

Olgac, A., & Karlik, B. (2011). Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence And Expert Systems*, *1*, 111-122.

OpenCV Face Recognition. (s. d.). *OpenCV*. Consulté 20 novembre 2022, à l'adresse https://opencv.org/opencv-face-recognition/

Otberdout, N., Daoudi, M., Kacem, A., Ballihi, L., & Berretti, S. (2022). Dynamic Facial Expression Generation on Hilbert Hypersphere With Conditional Wasserstein Generative Adversarial Nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(2), 848-863. https://doi.org/10.1109/TPAMI.2020.3002500

Paier, W., Hilsmann, A., & Eisert, P. (2020). Interactive facial animation with deep neural networks. *IET Computer Vision*, *14*(6), 359-369. https://doi.org/10.1049/iet-cvi.2019.0790

Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). Recent Progress on Generative Adversarial Networks (GANs) : A Survey. *IEEE Access*, *7*, 36322-36333. https://doi.org/10.1109/ACCESS.2019.2905015

Parke. (1982). Parameterized Models for Facial Animation. *IEEE Computer Graphics and Applications*, *2*(9), 61-68. https://doi.org/10.1109/MCG.1982.1674492

Parke, F. I. (1972). Computer generated animation of faces. *Proceedings of the ACM annual conference - Volume 1*, 451-457. https://doi.org/10.1145/800193.569955

Parke, F. I. (1975). A model for human faces that allows speech synchronized animation. *Computers & Graphics*, *1*(1), 3-4. https://doi.org/10.1016/0097-8493(75)90024-2

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*. https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1

Pascual, S., Bonafonte, A., & Serrà, J. (2017). *SEGAN : Speech Enhancement Generative Adversarial Network* (arXiv:1703.09452). arXiv. https://doi.org/10.48550/arXiv.1703.09452

Pighin, F. (1999). *Modeling and Animating Realistic Faces from Images* [Thèse de doctorat]. University of Washington.

Pighin, F., Auslander, J., Lischinski, D., Salesin, D. H., & Szeliski, R. (1997). Realistic facial animation using image-based 3D morphing. *Microsoft Research*.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., & Salesin, D. H. (1998). Synthesizing realistic facial expressions from photographs. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 75-84. https://doi.org/10.1145/280814.280825

Ping, H. Y., Abdullah, L. N., Sulaiman, P. S., & Halin, A. A. (2013). Computer Facial Animation : A Review. *International Journal of Computer Theory and Engineering*, 658-662. https://doi.org/10.7763/IJCTE.2013.V5.770

Platt, S. M. (1985). *A structural model of the human face (graphics, animation, object representation)* [Phd]. University of Pennsylvania.

Platt, S. M., & Badler, N. I. (1981). Animating facial expressions. *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, 245-252. https://doi.org/10.1145/800224.806812

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). GANimation : Anatomically-Aware Facial Animation from a Single Image. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Éds.), *Computer Vision – ECCV 2018* (p. 835-851). Springer International Publishing. https://doi.org/10.1007/978-3-030-01249-6_50

Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., & Wang, H. (2018). Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, *29*(3-4), e1819. https://doi.org/10.1002/cav.1819

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks* (arXiv:1511.06434). arXiv. https://doi.org/10.48550/arXiv.1511.06434

Ranzato, M. aurelio, Poultney, C., Chopra, S., & Cun, Y. (2006). Efficient Learning of Sparse Representations with an Energy-Based Model. *Advances in Neural Information Processing Systems*, *19*. https://proceedings.neurips.cc/paper/2006/hash/87f4d79e36d68c3031ccf6c55e9bbd39-Abstract.html

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once : Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. https://doi.org/10.1109/CVPR.2016.91

Reeves, W. T. (1981). Inbetweening for computer animation utilizing moving point constraints. *ACM SIGGRAPH Computer Graphics*, *15*(3), 263-269. https://doi.org/10.1145/965161.806814

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, *28*. https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

Reynès, C. (2007). *Etude des Algorithmes génétiques et application aux données de protéomique* [Phdthesis, Université Montpellier I]. https://theses.hal.science/tel-00268927

Ribera, R. B. i, Zell, E., Lewis, J. P., Noh, J., & Botsch, M. (2017). Facial retargeting with automatic range of motion alignment. *ACM Transactions on Graphics*, *36*(4), 154:1-154:12. https://doi.org/10.1145/3072959.3073674

Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-Supervised Self-Training of Object Detection Models. *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, *1*, 29-36. https://doi.org/10.1109/ACVMOT.2005.107

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386-408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), Article 6088. https://doi.org/10.1038/323533a0

Saito, M., Matsumoto, E., & Saito, S. (2017). *Temporal Generative Adversarial Nets With Singular Value Clipping*. 2830-2839. https://openaccess.thecvf.com/content_iccv_2017/html/Saito_Temporal_Generative_Adversarial_ICCV_2017_paper.html

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., & Chen, X. (2016). Improved Techniques for Training GANs. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Saurav, S., Gidde, P., Saini, R., & Singh, S. (2022). Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *The Visual Computer*, *38*(3), 1083-1096. https://doi.org/10.1007/s00371-021-02069-7

Schmidhuber, J. (2015). Deep learning in neural networks : An overview. *Neural Networks*, *61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet : A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815-823. https://doi.org/10.1109/CVPR.2015.7298682

Sederberg, T. W., & Parry, S. R. (1986). Free-form deformation of solid geometric models. *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 151-160. https://doi.org/10.1145/15922.15903

Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), Article 1. https://doi.org/10.1609/aaai.v30i1.9883

Serpe, O. (2021, février 9). *Le Coeur de l'intelligence articielle : L'apprentissage automatique (part 1)*. https://www.echosciences-auvergne.fr/articles/le-coeur-de-l-intelligence-articielle-l-apprentissage-automatique-part1

Shahbazi, M., Danelljan, M., Paudel, D. P., & Gool, L. V. (2022, mars 16). *Collapse by Conditioning : Training Class-conditional GANs with Limited Data*. International Conference on Learning Representations. https://openreview.net/forum?id=7TZeCsNOUB_

Shakir, S., & Al-Azza, A. (2022). Facial Modelling and Animation : An Overview of The State-of-The Art. *Iraqi Journal for Electrical and Electronic Engineering*, *18*(1), 28-37. https://doi.org/10.37917/ijeee.18.1.4

Shao, J., & Qian, Y. (2019). Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing*, *355*, 82-92. https://doi.org/10.1016/j.neucom.2019.05.005

Shi, C., Tan, C., & Wang, L. (2021). A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network. *IEEE Access*, *9*, 39255-39274. https://doi.org/10.1109/ACCESS.2021.3063493

Siarohin, A., Lathuiliere, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). *Animating Arbitrary Objects via Deep Motion Transfer*. 2377-2386. https://openaccess.thecvf.com/content_CVPR_2019/html/Siarohin_Animating_Arbitrary_Objects_via_Deep_Motion_Transfer_CVPR_2019_paper.html

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *arXiv e-prints*. https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S

Sin, Z. P. T., Ng, P. H. F., Shiu, S. C. K., Chung, F., & Leong, H. V. (2019). 2D character animating networks : Bringing static characters to move via motion transfer. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 196-203. https://doi.org/10.1145/3297280.3297301

Song, L., Lu, Z., He, R., Sun, Z., & Tan, T. (2018). Geometry Guided Adversarial Facial Expression Synthesis. *Proceedings of the 26th ACM international conference on Multimedia*, 627-635. https://doi.org/10.1145/3240508.3240612

Sow, A. M. (2020). *Classification, réduction de dimensionnalité et réseaux de neurones : Données massives et science des données* [PhD Thesis]. Université du Québec à Trois-Rivières.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014a). Dropout : A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), Article 1.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014b). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929-1958.

Suk, M., & Prabhakaran, B. (2014). Real-time mobile facial expression recognition system-a case study. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 132-137.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama : Learning lip sync from audio. *ACM Transactions on Graphics*, *36*(4), 95:1-95:13. https://doi.org/10.1145/3072959.3073640

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4278-4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

Tan, M., & Le, Q. (2019). Efficientnet : Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, 6105-6114.

Tang, S., Liew, A. W. C., & Yan, H. (2004). Lip-sync in human face animation based on video analysis and spline models. *10th International Multimedia Modelling Conference, 2004. Proceedings.*, 102-108. https://doi.org/10.1109/MULMM.2004.1264973

Tang, S., Yan, H., & Liew, A. W.-C. (2003). A NURBS-based vector muscle model for generating human facial expressions. *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, *2*, 758-762 vol.2. https://doi.org/10.1109/ICICS.2003.1292558

Terzopoulos, D., & Waters, K. (1990). Physically-based facial modelling, analysis, and animation. *The Journal of Visualization and Computer Animation*, *1*(2), 73-80. https://doi.org/10.1002/vis.4340010208

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face : Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387-2395.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond : A Survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

Tripathy, S., Kannala, J., & Rahtu, E. (2021). *FACEGAN : Facial Attribute Controllable rEenactment GAN*. 1329-1338. https://openaccess.thecvf.com/content/WACV2021/html/Tripathy_FACEGAN_Facial_Attribute_Controllable_rEenactment_GAN_WACV_2021_paper.html

Tulyakov, S., Liu, M.-Y., Yang, X., & Kautz, J. (2018). *MoCoGAN : Decomposing Motion and Content for Video Generation*. 1526-1535. https://openaccess.thecvf.com/content_cvpr_2018/html/Tulyakov_MoCoGAN_Decomposing_Motion_CVPR_2018_paper.html

Veen, F. V. (2016, septembre 14). *The Neural Network Zoo*. The Asimov Institute. https://www.asimovinstitute.org/neural-network-zoo/

Ververas, E., & Zafeiriou, S. (2020). SliderGAN : Synthesizing Expressive Face Images by Sliding 3D Blendshape Parameters. *International Journal of Computer Vision*, *128*(10), 2629-2650. https://doi.org/10.1007/s11263-020-01338-7

Viaud, M. L., & Yahia, H. M. (1992). Facial Animation with Wrinkles. In E. T. R. Series (Éd.), *Third Eurographics Workshop on Animation and Simulation*. Eurographics 1992. https://hal.inria.fr/inria-00423779

Video | MPEG. (s. d.). Consulté 30 août 2022, à l'adresse https://mpeg.chiariglione.org/standards/mpeg-4/video

Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017). Applying convolutional neural network for network intrusion detection. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1222-1228. https://doi.org/10.1109/ICACCI.2017.8126009

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning*, 1096-1103. https://doi.org/10.1145/1390156.1390294

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137-154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Vo, D. M., & Sugimoto, A. (2018). Paired-D GAN for Semantic Image Synthesis. *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV*, 468-484. https://doi.org/10.1007/978-3-030-20870-7_29

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating Videos with Scene Dynamics. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/04025959b191f8f9de3f924f0940515f-Abstract.html

Vougioukas, K., Petridis, S., & Pantic, M. (2020). Realistic Speech-Driven Facial Animation with GANs. *International Journal of Computer Vision*, *128*(5), 1398-1413. https://doi.org/10.1007/s11263-019-01251-8

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., & Pantic, M. (2017). Deep structured learning for facial expression intensity estimation. *Image Vis. Comput*, *259*, 143-154.

Wang, L., & Soong, F. K. (2015). HMM trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, *74*(22), 9849-9869. https://doi.org/10.1007/s11042-014-2118-8

Wang, T., Shao, Z., Xiao, Y., Zhang, X., Chen, Y., Shi, B., Chen, S., Wang, Y., Peng, J., & Shang, X. (2021). Predicting Hepatoma-Related Genes Based on Representation Learning of PPI network and Gene Ontology Annotations. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1892-1898. https://doi.org/10.1109/BIBM52615.2021.9669479

Wang, W., Yan, X., Xie, Y., Qin, J., Pang, W.-M., & Heng, P.-A. (2009). A Physically-Based Modeling and Simulation Framework for Facial Animation. *2009 Fifth International Conference on Image and Graphics*, 521-526. https://doi.org/10.1109/ICIG.2009.26

Wang, Z., She, Q., & Ward, T. E. (2021). Generative Adversarial Networks in Computer Vision : A Survey and Taxonomy. *ACM Computing Surveys*, *54*(2), 37:1-37:38. https://doi.org/10.1145/3439723

Waters, K. (1987). A muscle model for animation three-dimensional facial expression. *ACM SIGGRAPH Computer Graphics*, *21*(4), 17-24. https://doi.org/10.1145/37402.37405

Waters, K., & Frisbie, J. (1995). A coordinated muscle model for speech animation [Application/pdf]. *Proceedings of Graphics Interface '95*, *Québec*, 8 pages, 7.01 MB. https://doi.org/10.20380/GI1995.19

Wei, S.-E., Saragih, J., Simon, T., Harley, A. W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., & Sheikh, Y. (2019). VR facial animation via multiview image translation. *ACM Transactions on Graphics*, *38*(4), 67:1-67:16. https://doi.org/10.1145/3306346.3323030

Weise, T., Bouaziz, S., Li, H., & Pauly, M. (2011). Realtime performance-based facial animation. *ACM Transactions on Graphics*, *30*(4), 77:1-77:10. https://doi.org/10.1145/2010324.1964972

Wen, Y., Raj, B., & Singh, R. (2019). Face Reconstruction from Voice using Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/eb9fc349601c69352c859c1faa287874-Abstract.html

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick & F. Kozin (Éds.), *System Modeling and Optimization* (p. 762-770). Springer. https://doi.org/10.1007/BFb0006203

Wiles, O., Koepke, A. S., & Zisserman, A. (2018). *X2Face : A network for controlling face generation using images, audio, and pose codes*. 670-686. https://openaccess.thecvf.com/content_ECCV_2018/html/Olivia_Wiles_X2Face_A_network_ECCV_2018_paper.html

Williams, L. (1990). Performance-driven facial animation. *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, 235-242. https://doi.org/10.1145/97879.97906

Wu, H., Zheng, S., Zhang, J., & Huang, K. (2019). GP-GAN : Towards Realistic High-Resolution Image Blending. *Proceedings of the 27th ACM International Conference on Multimedia*, 2487-2495. https://doi.org/10.1145/3343031.3350944

Wu, X., He, R., Sun, Z., & Tan, T. (2018). A Light CNN for Deep Face Representation With Noisy Labels. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2884-2896. https://doi.org/10.1109/TIFS.2018.2833032

Wu, X., Zhang, Q., Wu, Y., Wang, H., Li, S., Sun, L., & Li, X. (2021). F^{3}A-GAN : Facial Flow for Face Animation With Generative Adversarial Networks. *IEEE Transactions on Image Processing*, *30*, 8658-8670. https://doi.org/10.1109/TIP.2021.3112059

Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., & Yang, M.-H. (2023). GAN Inversion : A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3121-3138. https://doi.org/10.1109/TPAMI.2022.3181070

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492-1500.

Xie, S., Hu, H., & Chen, Y. (2021). Facial Expression Recognition With Two-Branch Disentangled Generative Adversarial Network. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(6), 2359-2371. https://doi.org/10.1109/TCSVT.2020.3024201

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). *AttnGAN : Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks*. 1316-1324. https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html

Xu, Y., Kong, Q., Huang, Q., Wang, W., & Plumbley, M. D. (2017). Convolutional gated recurrent neural network incorporating spatial features for audio tagging. *2017 International Joint Conference on Neural Networks (IJCNN)*, 3461-3466. https://doi.org/10.1109/IJCNN.2017.7966291

Yang, H., & Wang, F. (2019). Wireless Network Intrusion Detection Based on Improved Convolutional Neural Network. *IEEE Access*, *7*, 64366-64374. https://doi.org/10.1109/ACCESS.2019.2917299

Yang, J., Guo, J., Yue, H., Liu, Z., Hu, H., & Li, K. (2019). CDnet : CNN-Based Cloud Detection for Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(8), 6195-6211. https://doi.org/10.1109/TGRS.2019.2904868

Yang, L.-C., Chou, S.-Y., & Yang, Y.-H. (2017, octobre 23). MidiNet : A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*. ISMIR. https://doi.org/10.5281/zenodo.1415990

Yi, Z., Tang, Q., Ramiya Srinivasan, V. S., & Xu, Z. (2020). Animating Through Warping : An Efficient Method for High-Quality Facial Expression Animation. *Proceedings of the 28th ACM International Conference on Multimedia*, 1459-1468. https://doi.org/10.1145/3394171.3413926

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018). Graph Convolutional Neural Networks for Web-Scale Recommender Systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 974-983. https://doi.org/10.1145/3219819.3219890

You, C., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z., Cong, W., Vannier, M. W., Saha, P. K., Hoffman, E. A., & Wang, G. (2020). CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). *IEEE Transactions on Medical Imaging*, *39*(1), 188-203. https://doi.org/10.1109/TMI.2019.2922960

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). *Generative Image Inpainting With Contextual Attention*. 5505-5514. https://openaccess.thecvf.com/content_cvpr_2018/html/Yu_Generative_Image_Inpainting_CVPR_2018_paper.html

Yu Zhang, Prakash, E. C., & Sung, E. (2001). Real-time physically-based facial expression animation using mass-spring system. *Proceedings. Computer Graphics International 2001*, 347-350. https://doi.org/10.1109/CGI.2001.934696

Zagoruyko, S., & Komodakis, N. (2017). *Wide Residual Networks* (arXiv:1605.07146). arXiv. https://doi.org/10.48550/arXiv.1605.07146

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115. https://doi.org/10.1145/3446776

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-Attention Generative Adversarial Networks. *Proceedings of the 36th International Conference on Machine Learning*, 7354-7363. https://proceedings.mlr.press/v97/zhang19d.html

Zhang, H., Su, W., Yu, J., & Wang, Z. (2021). Identity–Expression Dual Branch Network for Facial Expression Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(4), 898-911. https://doi.org/10.1109/TCDS.2020.3034807

Zhang, J., Chen, K., & Zheng, J. (2020). Facial Expression Retargeting from Human to Avatar Made Easy. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, M., Li, W., & Du, Q. (2018). Diverse Region-Based CNN for Hyperspectral Image Classification. *IEEE Transactions on Image Processing*, *27*(6), 2623-2634. https://doi.org/10.1109/TIP.2018.2809606

Zhang, W., Tang, P., & Zhao, L. (2019). Remote Sensing Image Scene Classification Using CNN-CapsNet. *Remote Sensing*, *11*(5), Article 5. https://doi.org/10.3390/rs11050494

Zhao, G., Huang, X., Taini, M., Li, S. Z., & PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, *29*(9), Article 9.

Zhao, K., Chu, W.-S., & Zhang, H. (2016). *Deep Region and Multi-Label Learning for Facial Action Unit Detection*. 3391-3399. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Zhao_Deep_Region_and_CVPR_2016_paper.html

Zhao, X., Shi, X., & Zhang, S. (2015a). Facial Expression Recognition via Deep Learning. *IETE Technical Review*, *32*(5), Article 5. https://doi.org/10.1080/02564602.2015.1017542

Zhao, X., Shi, X., & Zhang, S. (2015b). Facial Expression Recognition via Deep Learning. *IETE Technical Review*, *32*(5), 347-355. https://doi.org/10.1080/02564602.2015.1017542

Zhao, Y., Oveneke, M. C., Jiang, D., & Sahli, H. (2019). A video prediction approach for animating single face image. *Multimedia Tools and Applications*, *78*(12), 16389-16410. https://doi.org/10.1007/s11042-018-6952-y

Zhu, D., Tian, G., Zhu, L., Wang, W., Wang, B., & Li, C. (2021). LKRNet : A dual-branch network based on local key regions for facial expression recognition. *Signal, Image and Video Processing*, *15*(2), 263-270. https://doi.org/10.1007/s11760-020-01753-w

Zhu, H., Luo, M.-D., Wang, R., Zheng, A.-H., & He, R. (2021). Deep Audio-visual Learning : A Survey. *International Journal of Automation and Computing*, *18*(3), 351-376. https://doi.org/10.1007/s11633-021-1293-0

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242-2251. https://doi.org/10.1109/ICCV.2017.244

Zou, W., Zhang, D., & Lee, D.-J. (2022). A new multi-feature fusion based convolutional neural network for facial expression recognition. *Applied Intelligence*, *52*(3), 2918-2929. https://doi.org/10.1007/s10489-021-02575-0