République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique
Université Abderrahmane Mira
Faculté de la Technologie

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

ATE Department

Département d'Automatique, Télécommunication et d'Electronique

Projet de Fin d'Etudes

Pour l'obtention du diplôme de Master

Filière : Télécommunications.

Spécialité :Réseaux et Télécoms.

Thème

Descripteurs audio-visuels pour la reconnaissance des émotions

Préparé par :

- Fourar Ryma
- Ouari Samia

Dirigé par :

M. Tounsi Mohamed

M. Belabbaci El Ouanas

Examiné par :

M. Boualem Mohamed

M. Kasmi Reda

Année universitaire : **2023/2024**

Dédicace

Le devoir de reconnaissance m'oblige de dedier ce modeste memoire à tous ceux qui me sont chers,ce sont ceux à qui je dois mon succès
A mon chers père ,chaque mot semble bien fade pour exprimer l'amour profond et la gratitude infinie que j ai pour vous, pour les innombrables sacrifices que vous avez consentis pour mon éducation, vous avez été bien plus qu'un guide, vous avez été mon modèle d'honneteté, de sérieux et de responsabilité, votre présence incarne pour moi la quintessence de la perseverance et de la créativité
A ma merveilleuse mère aucun mot ne saurait capturer la profondeur de l'amour et de l'affection que je ressens pour vous, vous êtes bien plus qu'une mère, vous êtes mon phare de génerosité et mon exemple de dévouement, votre tendresse infinie est une source inépuisable de réconfort, et je suis infinement reconnaissante pour chaque instant ou vous avez été là pour moi, sans jamais faillir, merci pour votre présence rassurante et pour tous ces instants où ton amour inconditionnel a été ma plus grande force
A ma douce Lydia, ma lumière et ma douceur, ta présence est un précieux cadeau qui inonde ma vie de joie et de tendresse, tes sourires étincelants illuminent notre foyer et ton innocence m'apprend chaque jour la beauté de la simplicité et la puréte de l'amour. Je te souhaite d'être toujours entourée de bonheur et de sérénité, ma chere Lydia, car je t'aime au-dela des mots et je suis infinement reconnaissante de t'avoir comme soeur, ton amour est un trésor que je chérirai éternellement
A mes deux frères Fares et Hichem, vous êtes bien plus que des frères, vous êtes mes guides, votre presence dans ma vie est un cadeau précieux que je chéris chaque jour, merci pour votre soutien inconditionnel et votre amour sans faille, je suis infinement reconnaissante de vous avoir comme mes frères
A ma binome qui me supporte durant ces derniers mois
A tous mes amis avec lesquelles j'ai partagé de merveilleux moments, chaqu'un de vous a enrichi ma vie de votre amitié sincere et de votre présence précieuse
Je dédie ce travail en témoignage de mon amour et ma profonde reconnaissance

Fourar ryma

Dédicace

Je voudrais d'abord me prosterner remerciant "ALLAH Le Tout-Puissant" de m'avoir donnée le courage et la patience pour terminer ce travail.

Je dédie ce modeste travail :

À mes chers parents pour leur soutien et leurs encouragements durant toutes mes années d'étude.

À ma grand-mère ma source d'affection et d'amour.

À ma petite sœur la moitié de mon cœur.

À mes frères mes piliers dans cette vie.

À toute ma famille et à ma chère amie.

À ma binôme avec qui j'ai partagé les éxigences de ce travail.
À tous mes enseignants durant mes années d'études avec lesquels j'ai beaucoup appris.

Samia

Remerciements

Nous traversons actuellement une étape cruciale de notre parcours estudiantin. C'est pourquoi nous souhaitons adresser nos plus sincères remerciements à nos familles bien-aimées. Depuis le commencement de nos études, elles n'ont cessé de nous apporter un soutien indéfectible et des encouragements inestimables. Grâce à elles, nous avons pu bénéficier des meilleures conditions pour étudier dans la sérénité et nous épanouir pleinement. Nos proches ont été d'un appui essentiel durant cette période académique déterminante de nos vies.

Nous souhaitons témoigner notre profonde reconnaissance envers notre enseignant et encadrant **Tounsi Mohamed** Grâce à son encadrement bienveillant et ses conseils avisés, nous avons pu mener à bien ce projet dans les meilleures conditions. Ses enseignements et sa pédagogie nous ont permis de progresser et d'acquérir de nouvelles compétences précieuses pour notre parcours professionnel. Nous voudrons également remercier notre CO-Encadrant **Belabbaci El Ouanas** qui a été un pilier inébranlable tout au long de la réalisation de ce travail. Nous le remercions chaleureusement pour sa grande patience, ses précieux encouragements qui nous ont sans cesse motivées, ainsi que pour sa disponibilité et son implication remarquables. Son encadrement exemplaire a été un atout inestimable. Nous n'aurions pu rêver d'un meilleur guide et conseiller pour nous accompagner dans l'accomplissement de cette étude.

Nous remercions également les membres du jury qui ont bien voulu accepter de consacrer leur temps à examiner ce travail .

Nous tenons également à exprimer notre vive gratitude à l'égard de nos amis, de nos proches et de nos camarades de promotion. Ils ont été des compagnons de route indispensables durant ce parcours. Leurs mots d'encouragement, leur esprit de collaboration et le soutien mutuel dont nous avons pu bénéficier ont grandement contribué à rendre cette expérience académique d'autant plus enrichissante et mémorable. Partager ce voyage studieux ensemble a donné une dimension particulièrement significative et agréable à nos années d'apprentissage.

Résumé

Les émotions jouent un rôle crucial dans la communication et l'interaction humaines car elles permettent aux individus de s'exprimer au-delà du domaine verbal. La capacité de comprendre les émotions humaines est souhaitable pour les ordinateurs dans diverses applications. Les récentes avancées technologiques ont permis aux utilisateurs de communiquer avec les ordinateurs de manière auparavant inimaginable. Cette recherche présente une approche holistique de l'analyse des sentiments et des émotions, intégrant un ensemble diversifié d'algorithmes d'apprentissage automatique (machine learning) et d'apprentissage profond (deep learning) pour analyser de manière exhaustive les données faciales et vocales.

Les contributions de ce travail incluent l'utilisation d'une méthode de prétraitement connue sous le nom de Multiscale Retinex (MSR) pour améliorer la qualité des images et le contraste. De plus, des descripteurs discriminants handcrafted tels que LDP (Local Directional Pattern), BSIF (Binarized Statistical Image Features) et LBP (Local Binary Patterns), ainsi que des descripteurs de deep learning comme VGG19 et ResNet101, sont utilisés pour la reconnaissance des émotions basées sur les images faciales. Pour la reconnaissance des émotions dans la parole, nous utilisons le descripteur Handcrafted MFCC (Mel Frequency Cepstral Coefficient) et le modèle acoustique préentraîné VGGish basé sur un réseau CNN (Convolutional Neural Network). De plus, la méthode EDA (Exponential Discriminant Analysis) est utilisée pour la séparation maximale entre les classes et une fusion au niveau des scores utilisant la somme pondérée WSF (Weighted Sum Fusion) sert à améliorer le processus de correspondance. Des tests ont été effectués en utilisant trois bases de données et les résultats sont trés satisfaisants.

Mots clés : Reconnaissance des Emotions Audio-Visuels, EDA, CNN, MSR, MFCC, LDP, WSF.

Table des matières

Liste des Figures **v**

Liste des Tableaux **vi**

Introduction Générale **1**

I Concepts de Base Pour La Reconnaissance Des Emotions **5**
I.1 Introduction 6
I.2 La Reconnaissance des Emotions 6
I.3 Modèles Contemporains et aspects Psychologiques de l'émotion . . 6
I.4 Reconnaissance Multimodale des Emotions (RME) 7
I.4.1 Reconnaissance avec fusion de la parole et de l'image faciale 8
I.5 Motivations et applications 9
I.6 Défis et problème pour les systems de reconnaissance 10
I.6.1 Defis du système des émotions faciales 10
I.6.2 Defis des émotions pour le système de la parole 13
I.7 Structure du systeme de reconnaissance des emotions 14
I.7.1 Phases du système de reconnaissance des émotions 14
I.7.2 Étape du système de reconnaissance des émotions 14
I.8 Conclusion 16

II Méthode Pour La Reconnaissance Des Emotions **17**
II.1 Introduction 18
II.2 Travaux Connexes 18
II.3 Base de Données 20
II.4 Approches pour la reconnaissance d'émotion 22
II.4.1 Prétraitement 22
II.4.1.1 Prétraitement de la Parole 23
II.4.1.1.1 Spectrogramme : 24
II.4.1.2 Prétraitement de l'image facial 25
II.4.1.2.1 Détection de visage : 25

II.4.1.2.2 Multiscale Retinex (MSR) : 26
II.4.2 Extraction de Caractéristiques 27
II.4.2.1 Méthode basé sur les CNN 27
II.4.2.1.1 Réseau de Neurones Convolutif (CNN) . . 27
II.4.2.1.2 ResNet101 : 30
II.4.2.1.3 VGG19 : 30
II.4.2.1.4 VGGish : 31
II.4.2.2 Caractéristiques Handcrafted 32
II.4.2.2.1 Modèles binaires locaux (LBP) : 32
II.4.2.2.2 Caractéristiques des images statistiques binarisées (BSIF) : 32
II.4.2.2.3 Quantification de phase locale (LPQ) : . . 33
II.4.2.2.4 Motif Directionnel local (LDP) : 35
II.4.2.2.5 Coefficients cepstraux en fréquence Mel (MFCC) 36
II.4.3 Reduction de dimensionnalité 39
II.4.3.1 Analyse des Composants Principaux (PCA) 39
II.4.3.2 Analyse discriminante linéaire (LDA) 41
II.4.4 Classification 41
II.4.4.1 Machines à Vecteurs de Support (SVM) 41
II.4.4.2 K-plus proche voisin classifieur (KNN) 42
II.4.5 Décision 43
II.5 Métrique évaluation 43
II.6 Conclusion 45

III Présentation de la Solution **46**
III.1 Introduction 47
III.2 Méthodologie proposée 47
III.2.1 Système audio pour la reconnaissance des émotions 47
III.2.1.1 Audio vers représentation visuelle (spectrogramme) 47
III.2.1.1.1 STFT Spectrogramme : 48
III.2.1.1.2 Mel Spectrogramme : 49
III.2.1.2 Extraction de caractéristiques acoustique 49
III.2.2 Système visuel pour la reconnaissance des émotions 50
III.2.2.1 Pretraitement de visage 50
III.2.2.2 Extraction des caractéristiques visuels 50
III.2.3 Projection et Classification dans l'éspace par EDA 51
III.2.3.1 Comparaison avec CSS 52
III.2.4 Fusion Pondérée Audio-Visuelle 52
III.3 Conclusion 53

IV Résultats et Discussions **54**
IV.1 Introduction 55
IV.2 Environnement de travail 55
IV.3 Base de données 55
IV.3.1 Base de données de visages 55
IV.3.2 Base de données d'audio 56
IV.3.3 La base de données Audio-Visuels (IEMOCAP) 57
IV.4 Protocole de travail 57
IV.5 Expérimentations et Résultats 57
IV.6 Discussion 60
IV.6.1 L'avantage des caracteristiques handcrafted dans notre système 60
IV.6.2 Deep Vs. handcrafetd descripteur 60
IV.6.3 L'effet du nombre d'EDA 60
IV.6.4 Impact de la fusion WS 61
IV.6.5 Résultats sur IEMOCAP 61
IV.7 Conclusion 61

Conclusion Générale **63**

Perspectives **65**

A Interface graphique **75**
A.1 Fenêtre d'accueil 75
A.1.1 GUI pour le systeme de reconnaissance d'emotion a partir l'image faciale. 77
A.1.2 Test d'identification 79

Table des figures

I.1 Reconnaissance des émotions multimodales 8
I.2 Processus intégré de reconnaissance émotionnelle à partir d'images faciales et de la parole [Livingstone and Russo, 2018]. 9
I.3 Illustration d'une variation de pose [Spectroscopies et al., 2016]. . . 12
I.4 Illustration d'une Présence/absence d'éléments structurants/occultations [Spectroscopies et al., 2016]. 12
I.5 Le vieillissement du visage [Spectroscopies et al., 2016]. 12
I.6 Conditions d'éclairage variables [Spectroscopies et al., 2016]. 13
I.7 Faible résolution [Spectroscopies et al., 2016]. 13
I.8 Etapes de la Reconnaissance automatique d'emotion. 15

II.1 Étapes de prétraitement. 23
II.2 Spectrogramme en 2D (a) et 3D (b). 25
II.3 Détection de visage (Viola et Jones). 25
II.4 Exemples des caractéristiques de Haar [Viola, 2001]. 26
II.5 Réseau de Neurones Convolutif. 28
II.6 Exemple convolution avec un filtre (2*2) et pas égale à 1. 28
II.7 Exemple de la correction ReLU. 29
II.8 Exemple Pooling avec un filtre (2*2) et pas égale à 2. 29
II.9 Architecture resnet-101. 30
II.10 Architecture VGG19. 31
II.11 Architecture VGGish. 31
II.12 Un exemple de descripteur LBP. 33
II.13 Les étapes nécessaires pour extraire les caractéristiques LPQ. . . . 35
II.14 Un exemple de descripteur LDP de base. 36
II.15 Extraction des paramètres MFCC. 37
II.16 des examples des fenêtres. 37
II.17 Filtre Mel. 38
II.18 Machines à Vecteurs de Support (SVM). 42
II.19 k plus proches voisins (K-NN). 42

III.1 Schéma proposé de la reconnaissance des émotions basée sur l'audio. 48
III.2 Visualisation comparative des caractéristiques du signal audio à l'aide de différentes méthodes de spectrogramme et des résultats de techniques de descripteurs visuels de texture. 49
III.3 Schéma proposé de la reconnaissance des émotions basée sur les images faciales. 51
III.4 Schéma proposé de la reconnaissance des émotions Audio-Visuels. . 53

IV.1 Exemples d'images extraites des ensembles de données de visage et 8 classes d'émotions déférentes. 56

A.1 Fenêtre d'accueil. 76
A.2 Système de reconnaissance d'émotion a partir de l'image faciale. . . 77
A.3 Système de reconnaissance d'émotion a partir de l'image faciale. . . 78
A.4 Test d'identification d'émotion a partir de l'image faciale. 79

Liste des tableaux

IV.1 Le taux de reconnaissance (%) du système d'émotion basé sur l'image faciale. 58
IV.2 Le taux de reconnaissance (%) du système d'émotion basé sur la parole. 59
IV.3 Le taux de reconnaissance (%) avec la fusion des meilleurs résultats du système audio et visuel par WS Fusion 59
IV.4 Le taux de reconnaissance (%) du système d'émotion Audio-Visuels sur la base de donnée IEMOCAP. 62

Liste des acronymes

AdaBoost	*Adaptive Boosting*
ANN	*Artificial Neural Network*
AUC	*Area Under the Curve*
BSIF	*Binarized statistical image features*
CNN	*Convolutional neural network*
CK+	*cohn-kanade*
CSS	*Cosine similarity scoring*
DCT	*Discrete Cosine Transform*
EDA	*Exponential Discriminant Analysis*
EER	*Equal Error Rate*
ER	*Emotion Recognition*
FC	*Fully Connected*
FD	*Face Detection*
FAR	*False Acceptance Rate*
FRR	*False Rejection Rate*
GUI	*Graphical User Interface*
KNN	*K-nearest neighbor*
LBP	*Local Binary Patterns*
LDA	*Linear Discriminant Analysis*

LDP	*Local Derivative Patterns*
LPC	*Linear Predective Coding*
LPCC	*Linear Predective Cepectral Coefficient*
LPQ	*Local Phase Quantization*
MFCC	*Mel Frequency Cepstral Coefficient*
MSR	*Multiscale Retinex*
MSIDA	*Multilinear Side-Information based Discriminant Analysis*
MPCA	*Multilinear Principal Component Analysis*
Open Cv	*Open Source Computer Vision*
PCA	*Principal Component Analysis*
PSF	*Point Spread Function*
PLP	*Perceptual Linear Prediction*
RVE	*Reconnaissance Vocale des émotions*
RFE	*Reconnaissance Faciale des émotions*
ResNet	*Residual Network*
Relu	*Rectified linear unit*
RA	*Reconnaissances Automatique*
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machines*
SIDA	*Semantic Indexing based on Document Annotation*
STFT	*Short-Term Fourier Transform*
VGG	*Visual Geometry Group*
WSF	*Weighted Sum Fusion*
XQDA	*Cross-View Quadratic Discriminant Analysis*
JAFFE	*Expression Faciale Féminine Japonaise*
KDEF	*Karolinska Directed Emotional Faces*

Introduction Générale

La communication est le socle fondamental de nos échanges quotidiens, qu'ils soient personnels, professionnels ou sociaux. C'est le moyen par lequel nous partageons nos idées, nos expériences et tissons des liens avec autrui. Toutefois, au-delà des mots prononcés, les émotions jouent un rôle crucial dans notre capacité à transmettre et recevoir des messages avec précision et profondeur. Elles sont une part intégrante de la communication humaine, ajoutant une dimension unique et subtile à nos interactions. Ces émotions se manifestent à travers divers canaux, comme les expressions faciales, les intonations vocales, les gestes et les postures corporelles, fournissant des informations essentielles sur notre état émotionnel, nos intentions et nos attitudes, enrichissant ainsi le sens littéral de nos paroles [Ekman, 1992, Koolagudi and Rao, 2012].

La capacité à comprendre et interpréter ces émotions est une compétence essentielle tant pour les individus que pour les machines interactives et intelligentes, garantissant une communication efficace et harmonieuse. Dans ce contexte, la reconnaissance automatique des émotions est un domaine de recherche en plein essor visant à analyser et interpréter les différents états émotionnels grâce aux avancées technologiques [Gračanin et al., 2021]. Les progrès de l'intelligence artificielle, du traitement du signal et de l'apprentissage automatique ont permis le développement de systèmes sophistiqués capables de détecter et analyser les émotions à partir de différents types de données, comme la parole, les images et les vidéos. Ces systèmes offrent de vastes possibilités d'application, de l'amélioration des interactions homme-machine à l'évaluation des réponses émotionnelles dans des domaines tels que la psychologie et les sciences sociales.

La reconnaissance automatique des émotions suit un processus de la reconnaissance qui est défini par les étapes suivantes : prétraitement, extraction des caractéristiques, classification et décision.

Plusieurs travaux ont été développés dans le domaine de l'apprentissage automatique sur la reconnaissance des émotions en utilisant plusieurs algorithmes.

C'est dans ce contexte qu'on a abordé notre travail. L'objectif de notre travail est de concevoir un système qui détecte les émotions en utilisant la voix et les expressions faciales et émotionnelles.

Contexte Générale

Le traitement de la parole et de l'image faciale sont deux branches complémentaires et cruciales de la recherche en traitement du signal pour la reconnaissance automatique des émotions. Les chercheurs ont toujours accordé une attention particulière aux signaux vocaux et aux expressions faciales, car ce sont les moyens de communication les plus naturels et les plus riches en informations pour les êtres humains. Avec les progrès des technologies de l'information et de la communication, le rêve d'une interaction naturelle multimodale avec les machines, combinant voix et visages expressifs, devient réalisable. Les recherches actuelles proposent de nombreux systèmes de reconnaissance automatique des émotions à partir des signaux vocaux et faciaux. Des progrès considérables ont été accomplis, exploitant à la fois les caractéristiques acoustiques du signal vocal comme la prosodie, le ton, le rythme, et les caractéristiques visuelles des expressions faciales comme les mouvements des sourcils, de la bouche, etc. Dans ce projet, nous combinons ces deux modalités pour identifier avec précision l'état émotionnel du locuteur comme la joie, la tristesse, la colère, etc. à partir de leurs manifestations vocales et faciales. Cette reconnaissance multimodale des émotions ouvre la voie à des interfaces homme-machine plus naturelles, expressives et contextuelles.

Problématique

La reconnaissance automatique des émotions à partir de la parole et des expressions faciales trouve son intérêt dans de nombreux domaines d'application. Dans les interfaces homme-machine multimodales, elle permettrait une interaction plus naturelle et contextuelle en détectant les états émotionnels de l'utilisateur à travers sa voix et son visage. Dans le domaine du multimédia, elle ouvrirait la voie à un nouveau niveau de métadonnées émotionnelles audio et visuelles pour l'indexation et la recherche de contenus. En surveillance, elle pourrait détecter des situations de stress, de colère ou d'autres émotions potentiellement à risque grâce à l'analyse combinée des signaux vocaux et faciaux. De nos jours, sous des conditions contrôlées en laboratoire, les systèmes de reconnaissance émotionnelle automatique donnent de bonnes performances pour distinguer les différents états à partir des caractéristiques acoustiques de la voix et des mouvements faciaux. Cependant, dans des environnements bruités ou avec des enregistrements de courte durée, une dégradation des performances est constatée avec les signaux vocaux

et visuels. La variabilité des expressions émotionnelles vocales et faciales entre les individus et la subtilité de certains états affectifs restent des défis majeurs à surmonter pour une reconnaissance multimodale fiable .

Contribution

Cette étude réside dans le développement et l'exploration de descripteurs audio-visuels innovants pour la reconnaissance d'émotions. En combinant les caractéristiques de l'audio et de la vidéo, cette étude vise à capturer de manière holistique et multidimensionnelle les nuances des expressions émotionnelles humaines. Les descripteurs sont conçus pour capturer non seulement les aspects acoustiques de la parole, mais aussi les informations visuelles telles que les expressions faciales, les gestes et les mouvements corporels, qui sont tous des indicateurs importants de l'état émotionnel. En intégrant ces différentes modalités, notre approche cherche à améliorer la robustesse et la précision des systèmes de reconnaissance d'émotions, en tenant compte de la complexité et de la variabilité des signaux émotionnels dans des contextes réels. Les résultats de notre travail sont susceptibles d'avoir un impact significatif dans divers domaines d'application, tels que la psychologie, la santé mentale, les interactions homme-machine et l'analyse du contenu multimédia, en offrant des outils avancés pour comprendre et interpréter les émotions humaines de manière plus complète et contextuelle.

Organisation du mémoire

Notre mémoire est strucuré en quatre chapitres comme suit :

— **Chapitre 1 :** Ce chapitre est consacré à la présentation du système de reconnaissance automatique des émotions à partir des visages et de la voix en présentant quelques définitions et terminologies essentielles pour comprendre la thématique.

— **Chapitre 2 :** Dans ce chapitre, nous présentons les bases de données publiques disponibles ainsi que quelques travaux connexes. Nous abordons également les approches de la reconnaissance d'emotion telles que le prétraitement des visages et les méthodes d'extraction des caractéristiques (handcrafted et méthode de deep learning) et la classification.

— **Chapitre 3 :** Ce chapitre est consacré à la conception de la solution proposée, où nous expliquons les différentes étapes utilisées pour la mise en œuvre de notre système : le prétraitement des visages (avec viola-jones +retinex), l'extraction des caractéristiques à l'aide de caractéristiques profondes

(VGG19, ResNet101 et VGGish) et des caractéristiques handcrafted (LBP, LDP,BSIF et MFCC).

— **Chapitre 4 :** Dans ce dernier chapitre, nous présentons et commentons nos résultats expérimentaux. L'évaluation des performances de notre solution de reconnaissance automatique des émotions est réalisée sur plusieurs investigations. Notre interface graphique créée à l'aide du Graphical User Interface (GUI) sous Matlab est donnée en Annexe A.

— **Conclusion Générale et l'Annexe :** Le mémoire se termine par une conclusion générale suivie des références bibliographiques utilisées. Dans notre conclusion, nous présentons un bilan du travail réalisé dans ce mémoire et nous exposons les perspectives pour des travaux futurs afin d'améliorer et de compléter nos investigations.

Chapitre I

Concepts de Base Pour La Reconnaissance Des Emotions

I.1 Introduction

La reconnaissance des émotions est un domaine émergent de l'intelligence artificielle qui vise à permettre aux machines de détecter et comprendre les émotions humaines . Cette procédure identifie les émotions d'une personne en utilisant les expressions faciales et la parole.

Ce chapitre aborde les connaissances et les concepts de base nécessaires pour comprendre le sujet de la reconnaissance des émotions. Il explique clairement en quoi un système de reconnaissance des émotions diffère d'un système de reconnaissance faciale et de la parole. Nous fournissons également un résumé des défis liés à la reconnaissance des émotions à partir des images faciales et de la parole.

I.2 La Reconnaissance des Emotions

Les émotions jouent un rôle puissant dans le façonnement du comportement humain, étant souvent associées à une gamme de sentiments tels que le bonheur, la tristesse, la colère, la peur, la joie, la haine et la surprise. Les variations dans les expressions faciales, le langage corporel et la voix sont couramment utilisées pour identifier ces différentes émotions [Cowie et al., 2001b, Gross, 2015]. Ce processus, connu sous le nom de reconnaissance ou d'analyse des sentiments, est devenu une caractéristique essentielle des systèmes d'intelligence artificielle.

L'observation des expressions faciales et de la voix est l'un des moyens principaux de déduire les émotions des individus, et cela peut être quantifié en analysant des images faciales et la voix. Les approches algorithmiques pour l'analyse des sentiments peuvent varier selon les applications, mais elles reposent généralement sur la reconnaissance des expressions faciales et sur l'analyse de la parole [Chanel et al., 2011]. Traditionnellement, les systèmes de reconnaissance des émotions faciales (REF) utilisent des données annotées pour l'apprentissage automatique, où les algorithmes sont formés à partir de ces données étiquetées. Au fil des années, deux facteurs clés ont émergé comme étant cruciaux pour les systèmes de reconnaissance des émotions : les bases de données et les algorithmes. Les premières stockent les informations nécessaires, tandis que les seconds visent à modéliser ces données comme un ensemble de caractéristiques logiques projetées dans des espaces multidimensionnels [Teixeira, 2020].

I.3 Modèles Contemporains et aspects Psychologiques de l'émotion

Plusieurs modèles contemporains ont tenté de classer les émotions en fonction de diverses variables. Un modèle décompose les émotions en huit schémas com-

portementaux prototypiques (protection, destruction, reproduction, réintégration, incorporation, rejet, exploration, orientation) associés à huit émotions primaires (joie, attirance, peur, surprise, tristesse, dégoût, colère, anticipation) [Plutchik, 1984, Oatley and Johnson-Laird, 1987]. Selon ce modèle, les émotions secondaires découlent de la combinaison de ces émotions primaires. De manière similaire, certains chercheurs proposent un modèle hiérarchique des émotions. Au niveau le plus élevé se trouve la valence (positive ou négative) des émotions. Viennent ensuite les différentes émotions positives (joie et amour) et négatives (colère et tristesse), qui sont à nouveau subdivisées en sous-catégories (par exemple, angoisse, chagrin et culpabilité pour la tristesse).

Ces deux modèles ont été élaborés en mesurant la similarité sémantique entre les mots de différents corpus. Dans la même veine, d'autres chercheurs se basent sur le contenu propositionnel des mots pour aboutir à une classification des émotions. Selon eux, les émotions de base sont constituées de primitives sémantiques abstraites difficilement accessibles à ce type de contenu propositionnel (bonheur, peur, colère et dégoût, par exemple, sont difficiles à définir par des mots), tandis que les émotions secondaires sont plus facilement objectivables [Shaver et al., 1987]. Parallèlement à ces modèles théoriques reposant sur la proximité sémantique des mots, d'autres modèles psychologiques ont émergé à la fin du 20 siècle et peuvent être classés en trois catégories : les modèles dimensionnels, les modèles cognitifs et les théories de l'incorporation.

I.4 Reconnaissance Multimodale des Emotions (RME)

La reconnaissance multimodale des émotions consiste à identifier et comprendre les états émotionnels humains en exploitant de multiples canaux d'information, comme l'illustre la figure I.1. Ces différentes modalités incluent les expressions faciales, le langage corporel, les patterns vocaux et bien d'autres signaux comportementaux et physiologiques. Chacune de ces sources fournit des indices pertinents sur l'émotion ressentie par une personne, permettant ainsi une analyse plus complète et approfondie [Sebe et al., 2005]. Pour combiner efficacement ces différents flux d'informations, des techniques d'intégration multimodale sont mises en œuvre. Celles-ci peuvent opérer au niveau des caractéristiques extraites de chaque modalité, en fusionnant ces représentations dans un vecteur unique avant la phase de classification. D'autres approches réalisent l'intégration au niveau décisionnel, en combinant les prédictions issues séparément de chaque modalité pour obtenir un résultat consolidé.

Ainsi, par cette analyse conjointe de plusieurs modalités expressives, la reconnaissance multimodale vise à saisir de manière plus riche et nuancée la complexité des processus émotionnels humains [Chen et al., 2020].

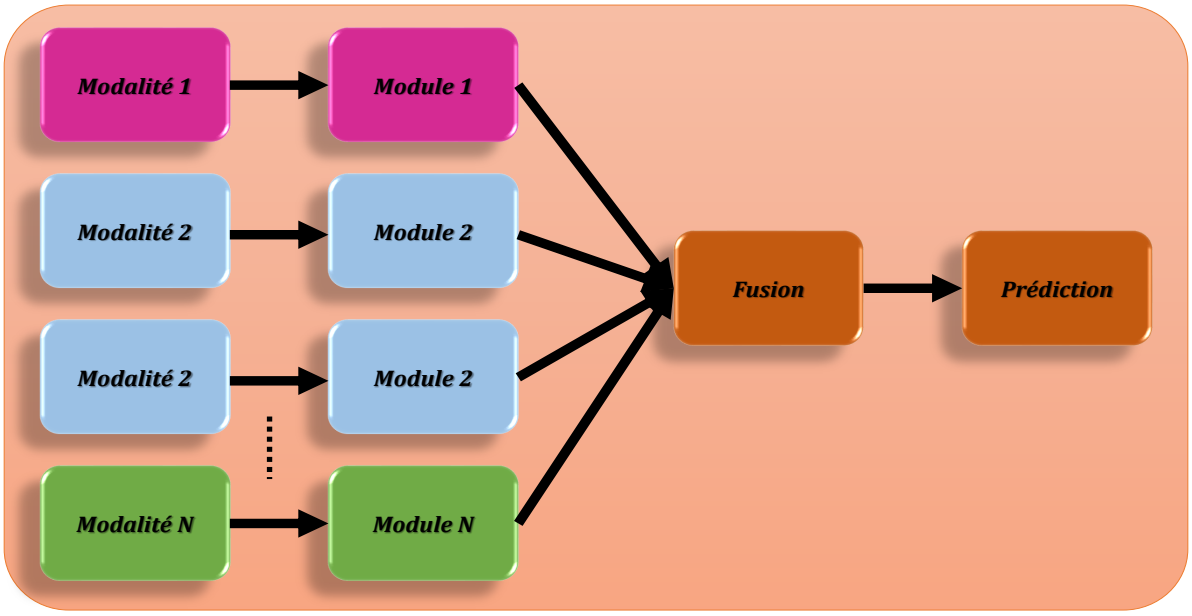

FIGURE I.1 – Reconnaissance des émotions multimodales

I.4.1 Reconnaissance avec fusion de la parole et de l'image faciale

La fusion des signaux vocaux et des expressions faciales représente une avancée significative dans le domaine de la reconnaissance automatique des émotions. En intégrant ces deux sources d'informations complémentaires, nous pouvons obtenir une compréhension plus complète et précise des états émotionnels d'un individu [Haq and Jackson, 2011]. Cette section est dédiée à la fusion des deux modèles finaux sélectionnés pour la reconnaissance vocale des émotions (RVE) et la reconnaissance faciale des émotions (RFE).

L'association de plusieurs modèles ou prédictions individuelles en un seul modèle ou prédiction unifiée est un processus connu sous le nom de fusion de modèles. Cette approche vise à tirer parti des points forts des différents modèles dans le but d'améliorer les performances globales et d'obtenir de meilleurs résultats. Selon Sanderson et Paliwal [Sanderson and Paliwal, 2002], on distingue deux grandes catégories de fusion : la fusion pré-classification et la fusion post-classification, comme illustré dans la figure I.2.

- **Fusion pré-classification [Sahoo et al., 2012] :** Ce type de fusion consiste à combiner les informations avant la classification, et peut se réaliser à deux niveaux :
  - **Capteurs :** Cette approche de fusion est envisageable lorsque les données sont de type identique mais proviennent de sources différentes. Par exemple, il est possible de fusionner les images faciales acquises par plusieurs caméras pour générer une image unique du visage en combinant ces diverses entrées.

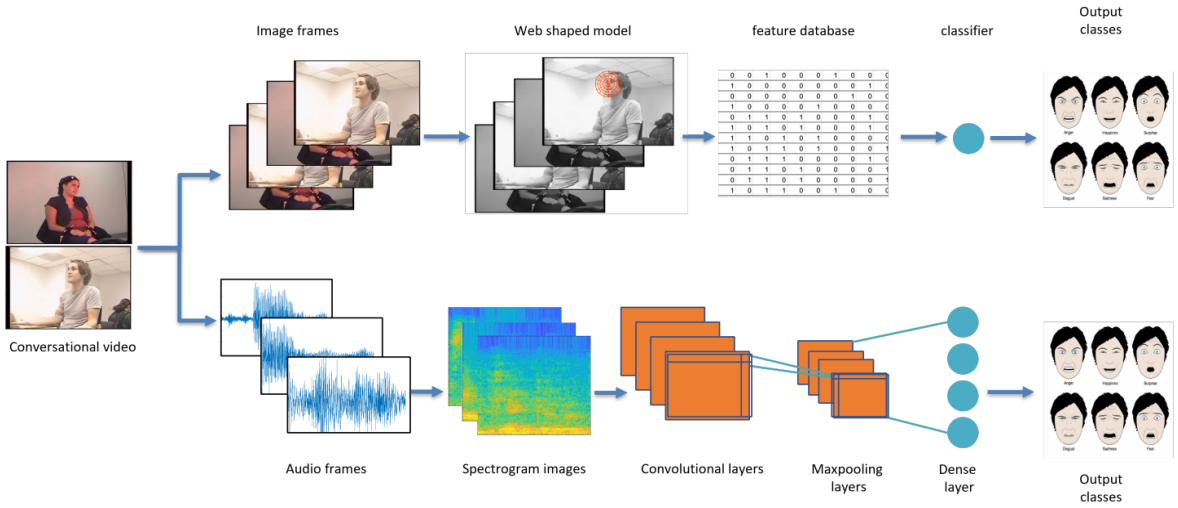

FIGURE I.2 – Processus intégré de reconnaissance émotionnelle à partir d'images faciales et de la parole [Livingstone and Russo, 2018].

- **Caractéristiques :** Une autre méthode consiste à fusionner les différentes caractéristiques extraites en les rassemblant au sein d'un seul vecteur par concaténation. Cependant, cette approche nécessite une attention particulière concernant les caractéristiques étroitement liées entre elles, ainsi que le risque de tomber dans ce qui est communément appelé "la haute dimensionnalité" lorsque le nombre de caractéristiques concaténées devient trop important.

— **Fusion post-classification [Sahoo et al., 2012] :** Cette méthode de fusion intervient après l'étape de classification. Elle consiste à combiner les prédictions (résultats) obtenues par différents classificateurs.

I.5 Motivations et applications

La reconnaissance automatique des émotions est un champ de recherche dynamique en constante expansion, offrant de multiples applications concrètes [Cowie et al., 2001b] :

Applications

— **Interaction homme-machine :** Elle permet aux systèmes informatiques de déceler les émotions des utilisateurs et d'adapter leur comportement en conséquence, améliorant ainsi l'expérience utilisateur.

— **Santé mentale :** Elle permet d'identifier précocement des troubles tels que la dépression, l'anxiété, le stress ou l'autisme en détectant les signaux émotionnels, facilitant ainsi une intervention rapide auprès des patients.

— **Marketing :** L'analyse des réactions émotionnelles face aux produits et services aide les entreprises à affiner leurs stratégies marketing pour mieux répondre aux besoins et désirs des consommateurs.
— **Détection de fraude :** Elle permet de repérer les émotions suspectes ou trompeuses lors d'entretiens, d'interrogatoires ou de transactions financières, renforçant ainsi la sécurité de ces processus.
— **Éducation :** Elle évalue l'engagement, la concentration et les réactions émotionnelles des étudiants, permettant aux enseignants d'adapter leur approche pédagogique pour optimiser le processus d'apprentissage.

Motivations

La reconnaissance automatique des émotions est une motivation de recherche en plein essor motivée par le désir fondamental de permettre aux systèmes informatiques d'interagir de manière plus naturelle et intuitive avec les humains[Gračanin et al., 2021].

— **Surveillance et sécurité :** La détection des émotions négatives comme la colère, l'agressivité ou la peur dans les lieux publics permet d'anticiper les comportements à risque et d'améliorer la sécurité publique.
— **Véhicules autonomes :** L'analyse des expressions faciales et des émotions des piétons permet aux véhicules autonomes de mieux anticiper leurs actions et de réduire les risques d'accident, améliorant ainsi la sécurité routière.
— **Jeux vidéo :** L'adaptation de l'expérience de jeu en fonction des émotions du joueur rend les jeux plus immersifs et captivants, offrant ainsi une expérience de divertissement plus enrichissante.
— **Robotique d'assistance :** La reconnaissance des émotions dans le contexte de la robotique d'assistance aux personnes âgées ou handicapées permet d'instaurer des interactions plus naturelles et empathiques, améliorant ainsi le bien-être des utilisateurs.
— **Cinéma et publicité :** L'analyse des réactions émotionnelles du public permet aux créateurs de peaufiner leurs œuvres cinématographiques ou publicitaires pour susciter les émotions souhaitées, renforçant ainsi leur impact et leur efficacité.

I.6 Défis et problème pour les systems de reconnaissance

I.6.1 Defis du système des émotions faciales

Les systèmes de reconnaissance des émotions partagent également des défis communs avec la reconnaissance faciale générale : Variations d'éclairage affectant

la perception des traits du visage. Occultations partielles du visage (lunettes, cheveux, etc.) Poses et angles de vue non optimaux pour l'analyse. Expressions faciales volontairement masquées ou simulées [Kohli, 2019].

Les expressions faciales des émotions varient grandement d'un individu à l'autre, ce qui rend la reconnaissance des émotions différente de la reconnaissance faciale générale. Pour la reconnaissance des émotions, le défi consiste à extraire les caractéristiques faciales similaires entre différentes expressions d'une même émotion chez des personnes différentes. Trois défis majeurs uniques à la reconnaissance des émotions ont été identifiés, ce qui aide à comprendre le problème et à guider le développement de systèmes informatiques dédiés :

— **Variabilité interindividuelle** Chaque personne exprime les émotions de manière unique, avec des intensités, des timings et des combinaisons d'expressions faciales différentes pour une même émotion.

— **Subtilité des micro-expressions** Certaines émotions se manifestent par des changements faciaux très subtils et brefs, difficiles à détecter et à interpréter.

— **Expressions volontairement masquées** : Dans certains contextes, les gens peuvent volontairement masquer ou simuler leurs expressions faciales réelles, ajoutant de la complexité.

Comprendre ces défis spécifiques liés aux variations interindividuelles et à la nature subtile et potentiellement trompeuse des expressions émotionnelles permet de mieux guider la conception de systèmes informatiques capables d'analyser et reconnaître avec précision les émotions à partir d'images faciales [Chanel et al., 2011].

La reconnaissance des émotions est un sous-ensemble de l'analyse d'expressions faciales, et l'apparence des expressions est sensible aux changements d'expressions faciales, aux occultations et aux poses. De plus, l'éclairage, le flou et la basse résolution peuvent également affecter l'apparence des expressions émotionnelles. Ces facteurs peuvent influencer différemment l'apparence des expressions :

— **Variations de pose :** Les mouvements de tête, tels que l'inclinaison, le roulis et le lacet, ou les changements de point de vue de la caméra, peuvent entraîner des changements significatifs dans l'apparence et/ou la forme du visage, rendant difficile la reconnaissance automatique des expressions selon la pose (voir la figure I.3 example faciale pour variations de pose) [Nash et al., 2016, Spectroscopies et al., 2016]. La correction de la pose est essentielle et peut être réalisée en utilisant des techniques efficaces pour faire pivoter le visage et/ou l'aligner sur l'axe de l'image.

— **Présence/absence d'éléments structurants/occultations :** Les images faciales prises dans un environnement non contrôlé nécessitent souvent une

FIGURE I.3 – Illustration d'une variation de pose [Spectroscopies et al., 2016].

reconnaissance efficace des expressions émotionnelles en présence de déguisements, d'accessoires ou d'occultations. Ceci est illustré dans la figure I.4, où des éléments tels que des chapeaux, des lunettes ou une barbe peuvent représenter un facteur d'occultation [Spectroscopies et al., 2016].

FIGURE I.4 – Illustration d'une Présence/absence d'éléments structurants/occultations [Spectroscopies et al., 2016].

— **Le vieillissement du visage :** Les changements dans l'apparence du visage peuvent être causés par le vieillissement, ce qui peut avoir un impact significatif sur le processus de reconnaissance des émotions [Liu et al., 2015].

FIGURE I.5 – Le vieillissement du visage [Spectroscopies et al., 2016].

— **Conditions d'éclairage variables :** De grandes variations d'éclairage peuvent dégrader les performances des systèmes de reconnaissance des émotions, avec de faibles niveaux de lumière rendant difficiles la détection et la reconnaissance des expressions faciales [Spectroscopies et al., 2016]. Des niveaux d'éclairage trop élevés peuvent entraîner une surexposition et des motifs faciaux indiscernables. Des techniques de traitement d'image telles que la normalisation de l'éclairage et l'apprentissage automatique sont utilisées pour

gérer ces variations (voir la figure I.6 example pour les Conditions d'éclairage different).

FIGURE I.6 – Conditions d'éclairage variables [Spectroscopies et al., 2016].

— **Résolution et modalité d'image :** Les performances de la reconnaissance des émotions sont influencées par la qualité et la résolution de l'image du visage, la configuration et les modalités de l'équipement numérique, et l'utilisation de différents matériels photographiques (voir la figure I.7 example pour des image avec faible résolution). Les visages acquis dans des conditions réelles peuvent mener à d'autres défis en raison des multiples modalités [Spectroscopies et al., 2016].

FIGURE I.7 – Faible résolution [Spectroscopies et al., 2016].

I.6.2 Defis des émotions pour le système de la parole

Les principaux défis de la reconnaissance des émotions basée sur la parole sont les suivants [Cowie et al., 2001a] :

— **Variabilité individuelle :** Les expressions émotionnelles varient considérablement d'une personne à l'autre en raison de différences physiologiques, culturelles et de personnalité.

— **Expressions émotionnelles subtiles et mélangées :** Les émotions exprimées dans la parole ne sont pas toujours claires et peuvent être mélangées ou subtiles, ce qui rend leur reconnaissance plus difficile.

— **Influence du contexte :** Le contexte, y compris l'environnement, la situation et les relations interpersonnelles, peut influencer l'expression et la perception des émotions.

— **Manque de données d'entraînement :** La plupart des ensembles de données disponibles pour l'entraînement des modèles de reconnaissance des émotions sont de taille limitée et peuvent ne pas représenter la variabilité réelle des expressions émotionnelles.

— **Difficultés liées aux langues :** Les systèmes de reconnaissance des émotions doivent faire face à des défis supplémentaires lorsqu'ils sont appliqués à différentes langues en raison de variations linguistiques et culturelles.

— **Fusion de modalités :** Combiner efficacement les informations provenant de différentes modalités (parole, visage, gestes, etc.) pour améliorer la reconnaissance des émotions est un défi complexe.

I.7 Structure du systeme de reconnaissance des emotions

I.7.1 Phases du système de reconnaissance des émotions

Comme tout système de reconnaissance des émotions, un système de reconnaissance des émotions à partir de la voix et des expressions faciales fonctionne en deux phases principales : l'apprentissage et le test Chanel et al. [2011], Gračanin et al. [2021].

— **Phase d'apprentissage** : Au cours de cette étape, le signal est transformé en un ensemble de coefficients adaptés pour modéliser les paramètres liés à la reconnaissance des émotions. La modélisation consiste à décrire les caractéristiques émotionnelles des paramètres. Le modèle obtenu à partir de cette opération doit fournir des moyens de comparer ces caractéristiques à celles d'une expression émotionnelle inconnue.

— **Phase de test** : Les similitudes entre les caractéristiques de test et les modèles émotionnels enregistrés dans la base de données sont évaluées. Ensuite, un module basé sur une stratégie de décision spécifique fournit la réponse du système en termes de reconnaissance émotionnelle.

I.7.2 Étape du système de reconnaissance des émotions

Comme mentionné dans la section précédente sur les phases du système de reconnaissance, chaque phase comprend différentes étapes telles que le prétraitement, l'extraction des caractéristiques, la classification et la prise de décision (voir la figure **??** pour une explication des étapes de la reconnaissance des émotions) [Ahmed et al., 2023, Khammari et al., 2023].

Prétraitement

La phase de prétraitement est cruciale dans le processus global de reconnaissance des émotions. Son but est d'améliorer la qualité des données d'entrée brutes et de mettre en évidence les zones pertinentes, afin de préparer ces données pour les étapes suivantes d'extraction de caractéristiques et de classification. Les techniques de prétraitement utilisées varient selon le type de données traité, que ce soit des signaux de parole ou des images faciales.

Extraction des caractéristiques

L'étape d'extraction des caractéristiques vise à extraire des informations riches et pertinentes à partir des données brutes, afin de faciliter la tâche de classification ultérieure. Dans cette section, nous aborderons les concepts théoriques sous-tendant les différentes techniques d'extraction de caractéristiques que nous avons utilisées dans notre projet, tant pour les signaux de parole que pour les images faciales.

Classification

À cette étape, les caractéristiques extraites sont utilisées comme vecteurs d'entrée pour des algorithmes de classification, qu'ils soient traditionnels ou basés sur l'apprentissage automatique. L'objectif est d'associer une émotion aux données d'entrée.

Décision

L'étape de décision consiste à entraîner un modèle prédictif à discriminer différentes expressions faciales à partir d'un ensemble de données annotées. Ce modèle apprend dans l'espace de représentation choisi lors de l'extraction des caractéristiques. Ainsi, lorsqu'une nouvelle donnée est soumise, le modèle prédictif utilise les connaissances acquises sur la base de données annotée pour prédire le type de signal ou l'expression faciale correspondant .

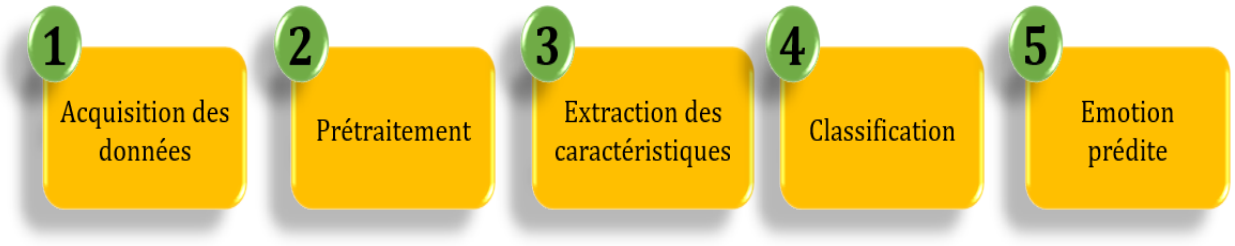

FIGURE I.8 – Etapes de la Reconnaissance automatique d'emotion.

I.8 Conclusion

Dans ce chapitre, nous avons donné un aperçu de la reconnaissance automatique de l'emotion à partir du visage et de la voix. Nous avons également abordé certaines applications qui peuvent utiliser ce type de systèmes ainsi que les défis auxquels sont confrontés les chercheurs.

Dans le prochain chapitre, nous présenterons quelques notions et plusieurs définitions relatives à la conception des systèmes de reconnaissance d'emotion.

Chapitre II

Méthode Pour La Reconnaissance Des Emotions

II.1 Introduction

La reconnaissance des émotions est un domaine fascinant qui utilise plusieurs méthodes pour détecter et interpréter les émotions humaines à partir de signaux tel que la parole, le langage corporel ou les expressions faciales. Ces sources sont intégrées à l'aide d'approches d'intégration au niveau des caractéristiques ou au niveau décisionnel [Chen et al., 2020].

Dans ce chapitre, nous allons concentrer sur les processus de la reconnaissance des émotions qui sont défini par les étapes suivantes :travaux connexe, base de données, prétraitement, extraction des caractéristiques, classification et décisions.

II.2 Travaux Connexes

Dans leur étude [Patil and Veni, 2019], les auteurs se sont concentrés sur cinq émotions spécifiques pouvant avoir un impact négatif sur le temps de réponse des conducteurs et potentiellement conduire à des accidents mortels, tels que la colère, la peur, la joie, la neutralité et la tristesse. Pour répondre à cette problématique, ils ont proposé une méthode novatrice combinant des caractéristiques géométriques et texturales, notamment une fusion des motifs binaires locaux (LBP) et des traits faciaux, pour la détection des émotions. Pour classifier les différentes émotions, ils ont utilisé un algorithme d'apprentissage automatique supervisé, la Machine à Vecteurs de Support (SVM). La méthode proposée a utilisé un classificateur basé sur des caractéristiques de Haar pour la détection de visage et pour extraire les repères faciaux, qui ont été concaténés avec les caractéristiques LBP. Les performances de la méthode proposée ont été évaluées sur l'ensemble de données étendu Cohn-Kanade (CK+), ce qui a donné de meilleurs résultats avec une précision de 86,7%.

Pour résoudre le problème de la rage au volant et réduire le risque d'accidents, des mesures d'atténuation efficaces sont cruciales. Dans cet article [Azman et al., 2019], les auteurs ont proposé un système de reconnaissance en temps réel des émotions faciales des conducteurs qui détecte les expressions de colère et alerte le conducteur. L'algorithme utilisé dans le système utilise la méthode de Viola-Jones Haar pour la détection frontale des visages, et la classification des émotions faciales est réalisée à l'aide d'un modèle de SVM. Le modèle SVM a été entraîné sur la base de données JAFEE et a atteint une précision moyenne d'environ 97%. Le système utilise une webcam installée devant le conducteur pour détecter la colère sur le visage du conducteur en temps réel et active un son d'alerte une fois qu'une émotions de colère est identifiée.

Dans leur article,[Sudha and Suganya, 2023] ont présenté la méthode de Reconnaissance des Émotions Faciales du Conducteur (REFC) pour aborder les occlusions partielles du visage lors de la reconnaissance des émotions faciales des

conducteurs. Cette méthode implique le calcul du flux optique entre deux images partiellement masquées en utilisant l'approche de Farneback. Ensuite, un Optimiseur Multi-Univers Parallèle (OMUP) entraîné sur des données de flux optique est utilisé pour reconstruire le flux optique, rétablissant ainsi les expressions faciales bloquées. Les flux optiques reconstruits sont ensuite employés pour reconnaître les émotions faciales des conducteurs sur la route en utilisant un classificateur de Réseaux Convolutifs Très Profonds (VGGNet) . Pour évaluer la performance de REFC, divers ensembles de données ont été utilisés pour analyser la reconnaissance des émotions faciales dans des séquences d'images, y compris les bases de données CK+ et KMU-FED. Le classificateur VGGNet proposé a été comparé à différentes méthodes de reconnaissance en termes de précision, de rappel, de score F1 et d'exactitude. Les résultats ont démontré que la méthode proposée a obtenu des performances élevées, avec une précision de 95,92 % et de 94,78 % sur les ensembles de données CK+ et KMU-FED respectivement.

Le rôle des émotions des conducteurs dans les tâches de conduite est crucial car il améliore la sécurité routière et réduit les risques de conduite. Ce sujet suscite un vif intérêt parmi les chercheurs en vision par ordinateur, qui ont obtenu des résultats remarquables en utilisant diverses approches de reconnaissance des émotions faciales. Cependant, ces approches ont principalement été testées sur des ensembles de données contrôlés en laboratoire ou basés sur Internet, et non sur des données réelles de scénarios de conduite. Pour répondre à cette lacune Dans cette article,[Xiao et al., 2022] ont proposé une nouvelle approche de reconnaissance des émotions faciales appelée FERDERnet. Cette méthode a été développée et testée sur un ensemble de données recueillies auprès de conducteurs sur la route. La méthode comprend trois modules : un module de détection faciale (FD) pour localiser le visage du conducteur, un module de rééchantillonnage basé sur l'augmentation (ABR) pour équilibrer l'ensemble de données d'entraînement et améliorer la généralisation du modèle, et un module de reconnaissance des émotions (ER) pour classifier les émotions des conducteurs en situation de conduite . L'article a réalisé diverses expériences en utilisant différents réseaux de référence, tels que Googlenet, Resnet50, InceptionV3 et Xception, pour évaluer l'approche proposée sur un ensemble de données d'émotions faciales de conducteurs sur la route. Le modèle proposé, utilisant le réseau Xception comme base, a surpassé les autres et a atteint une précision de reconnaissance élevée de 96.6 %

Dans cette étude ,[Zaman et al., 2022]ont proposés une méthode pour reconnaître l'état émotionnel d'un conducteur sans nécessiter d'efforts supplémentaires de la part du conducteur. Le modèle de reconnaissance des émotions faciales (FER) proposé a été utilisé pour identifier les émotions faciales du conducteur. Pour ce faire, un détecteur de visage Faster R-CNN amélioré a été proposé, qui remplace un bloc d'apprentissage de caractéristiques CNN personnalisé par le

modèle CNN de base à 11 couches. Cela améliore la précision et l'efficacité de la détection de visages, permettant une détection haute vitesse des visages des conducteurs. Pour reconnaître les différentes émotions du conducteur, le transfert d'apprentissage a été utilisé dans le modèle CNN NasNet Large. La précision des modèles proposés de détection de visages et de reconnaissance des eémotions faciales a été évaluée à l'aide de jeux de données de reconnaissance des émotions faciales de référence, comprenant JAFEE, CK+, FER-2013, AffectNet, et un jeu de données personnalisé. Le modèle proposé a atteint une précision élevée, avec des taux de 98,48%, 99,73%, 99,95%, 95,28% et 99,15% respectivement, surpassant ainsi certaines techniques de pointe. Cette étude démontre l'efficacité de la méthode proposée dans la reconnaissance précise de l'état émotionnel du conducteur.

La hausse des accidents de la route en a fait la première cause de décès chez les jeunes. Il est bien établi que les émotions influent considérablement sur les performances au volant. Afin d'améliorer la reconnaissance de l'état émotionnel des conducteurs, une étude récente [Varma et al., 2022] a proposé une approche hybride d'apprentissage de caractéristiques spatio-temporelles profondes. Cette méthode se concentre non seulement sur les informations spatiales dans les données de séquences vidéo/image, mais également sur la classification au niveau de la trame plutôt qu'au niveau de la séquence. Pour ce faire, les chercheurs ont mis au point une architecture de Réseau de Neurones Long Court Bidirectionnel Convolutionnel (CBiLSTM) qui capture efficacement les caractéristiques spatio-temporelles des données vidéo. Ils ont extrait la région faciale des images vidéo en utilisant le Réseau d'Alignement Facial (FAN), encodé ces régions à l'aide d'un modèle CNN SqueezeNet léger, puis alimenté la sortie à un réseau BiLSTM à deux couches pour apprendre les caractéristiques spatio-temporelles. Une couche entièrement connectée a ensuite généré les probabilités softmax des classes d'émotion. Pour prouver l'efficacité et la fiabilité de leur modèle, les chercheurs ont utilisé deux ensembles de données différents : KMU-FED, DMD, et leur propre ensemble de données vidéo expérimentales. Les deux derniers ensembles de données ont été annotés manuellement à l'aide d'un outil d'annotation interactif. Les résultats ont démontré que le modèle proposé offre une solution rapide, précise et applicable pour divers capteurs de caméra, atteignant un score F1 de 0,958 sur l'ensemble de données KMU-FED .

II.3 Base de Données

1. L'Expression Faciale Féminine Japonaise (JAFFE) : L'ensemble de données JAFFE se compose de 213 images présentant sept expressions faciales – six émotions de base et une expression neutre – qui ont été posées par dix femmes

Japonaise. Les images sont de taille 256 × 256 pixels. Cependant, l'ensemble de données pose un défi car il offre des exemples limités par sujet et expression, et toutes les images sont généralement utilisées pour évaluation [Lyons et al., 1998].

2. Karolinska Directed Emotional Faces (KDEF) : L'ensemble de données KDEF se compose de 4 900 données de haute images de 70 individus affichant sept expressions faciales (dont neutre) sous cinq angles différents [Lundqvist et al., 1998]. Chaque image a une résolution de 562 pixels × 762 pixels. Pour évaluer la performance de leurs méthodes, les chercheurs utilisent fréquemment l'intégralité de l'ensemble de données KDEF, en utilisant soit la répartition des tests de train, soit la validation croisée k-fold.

3. Cohn-Kanade étendu (CK+) : Est une base de données contrôlée en laboratoire qui est devenue largement utilisée pour évaluer les systèmes de reconnaissance des expressions faciales. L'ensemble de données comprend 593 séquences vidéo et images fixes comprenant sept différentes expressions faciales, qui vont des émotions de base au mépris. Les images de l'ensemble de données ont une résolution de 640 × 480 pixels ou 640 × 490 pixels, avec une précision de 8 bits valeurs en niveaux de gris. Les séquences CK+ affichent un décalage d'une expression neutre à une expression maximale, et la norme [Lucey et al., 2010]. La méthode de sélection des données consiste à extraire la première image de chaque séquence et la dernière à trois images avec formation de pics.Les chercheurs ont généralement utilisé l'ensemble de données avec une validation croisée k fois expériences, où les valeurs k de 5, 8 ou 10 sont courantes aux choix.

4. La petite base de données d'expressions faciales (MPI) : expression facial du visage MPI est une base de données comprend des vidéos mettant en vedette six acteurs, trois hommes et trois femmes, démontrant neuf expressions faciales différentes y compris l'accord, désaccord, heureux, triste, désemparé, pensant, confus, dégoût et surprise. Les vidéos ont été capturées à partir de cinq points de vue différents de la caméra et ont été enregistrés dans une conversation contexte. L'ensemble de données a été utilisé pour explorer les facteurs qui contribuent aux expressions faciales dans des milieux naturels, et les vidéos ont été ajustées par les chercheurs pour isoler et analyser des composants des expressions [Cunningham et al., 2005].

5. Ensemble de données de reconnaissance des expressions faciales 2013 (FER2013) : Le FER2013 est une base de données des images de visage, niveaux de gris, collectées via l'API de recherche d'images Google. Il comprend 35 887 images, dont 28 709 dans l'ensemble de formation, 3 589 dans l'ensemble de validation et 3589 dans l'ensemble de test. Les images sont redimensionnées à 48 × 48 pixels et les cadres mal étiquetés sont supprimés. Cet ensemble de données est principalement utilisé pour évaluer les méthodes DFER à l'aide d'une division spécifique de l'ensemble de données, bien que certains chercheurs aient

utilisé d'autres fractionnements de test d'apprentissage [Goodfellow et al., 2013].

6. Base de données audiovisuelle Ryerson sur la parole émotionnelle et Chanson (RAVDESS) : Le RAVDESS l'ensemble de données comprend 24 participants issus de diverses origines ethniques tels que les Caucasiens, les Asiatiques de l'Est et les Métis (Caucasiens d'Asie de l'Est, et des Premières nations noires-canadiennes de race blanche). L'ensemble de données capture les expressions émotionnelles à travers de véritables performances d'acteurs qui ont été chargés d'induire l'émotion souhaitée dans leur état. Cet ensemble de données est bien adapté aux applications d'apprentissage automatique qui utilisent des techniques d'apprentissage supervisé [Livingstone and Russo, 2018].

7. Expression faciale des conducteurs de l'Université Keimyung (KMU-FED) : L'ensemble de données KMU-FED se compose de 55 séquences d'images des conducteurs capturés en temps réel à l'aide d'un proche infrarouge (NIR) caméra montée sur le tableau de bord ou sur le volant. L'ensemble de données comprend 12 sujets présentant diverses expressions faciales en conduisant, avec occlusions partielles causées par les cheveux ou les lunettes de soleil et différents conditions d'éclairage telles que l'éclairage avant, gauche, droit et rétro-éclairage certains échantillons sur. Les images ont une résolution en pixels de 1600× 1200, ce qui en fait une ressource idéale pour étudier la reconnaissance des expressions faciales dans des conditions de conduite réalistes. Comme les chercheurs n'utilisent pas l'apprentissage et le test dans cet ensemble de données, il ont utilisé une technique pli k de validation croisée [Jeong and Ko, 2018].

II.4 Approches pour la reconnaissance d'émotion

Au cours des dernières années, plusieurs recherches physiologiques portant sur la capacité humaine à reconnaître et vérifier les visages et la voix ont été menées. Ces études ont conduit les chercheurs en vision par ordinateur et en apprentissage automatique à définir différentes approches automatiques avec des performances variables. Il est important de noter que ces performances ne peuvent pas être directement comparées car les approches sont testées sur des ensembles de données différents [Lu et al., 2014].

II.4.1 Prétraitement

Le prétraitement revêt une importance capitale dans le processus global de reconnaissance des émotions . Cette phase vise à améliorer la qualité des données d'entrée et à mettre en évidence les régions d'intérêt en vue de les préparer pour les étapes ultérieures d'extraction de caractéristiques et de classification [Khammari et al., 2023]. Les techniques de prétraitement varient en fonction du type

de données traitées. Dans cette section, nous examinerons en détail ces méthodes pour les signaux de la parole et les images faciales.

II.4.1.1 Prétraitement de la Parole

Les différentes étapes de prétraitement du signal vocal, telles qu'illustrées dans la figure II.1, facilitent la préparation du signal pour une extraction précise des informations qu'il renferme.Ces étapes sont : l'application du filtre de préaccentuation, la segmentation et le fenêtrage [Nema and Abdul-Kareem, 2018].

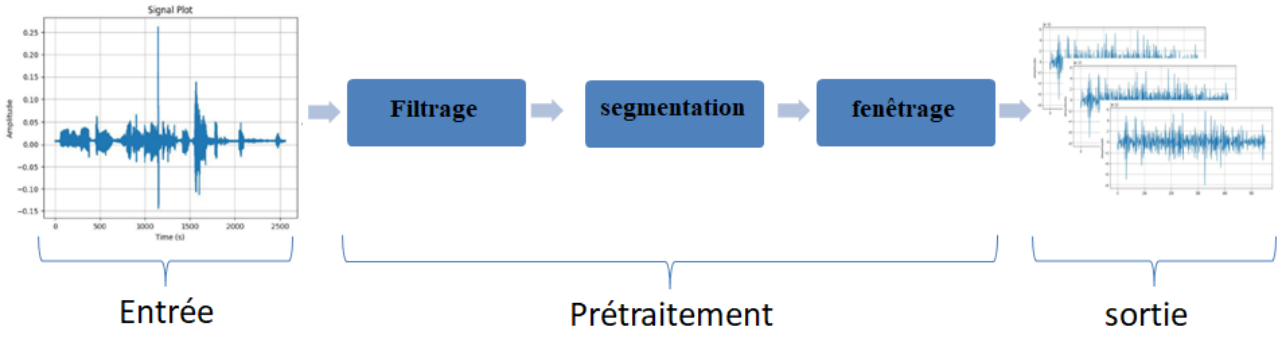

FIGURE II.1 – Étapes de prétraitement.

Filtre de préaccentuation : C'est une technique utilisée dans le domaine du traitement du signal audio pour améliorer la netteté et la clarté des hautes fréquences d'un signal. Il s'agit d'un filtre passe-haut qui atténue les basses fréquences et amplifie les hautes fréquences. Il est défini comme suit :

$$y(n) = x(n) - \alpha \cdot x(n-1)$$

où :

$y(n)$: Signal de sortie du filtre.
$x(n)$: Échantillon d'entrée du signal audio à l'instant n.
$x(n-1)$: Échantillon précédent du signal audio.
α : Coefficient de préaccentuation ayant une valeur comprise entre 0,9 et 1.

L'idée derrière le filtre de préaccentuation est de compenser la perte d'énergie dans les hautes fréquences qui peut se produire lors de l'enregistrement ou de la transmission d'un signal audio due à leur sensibilité aux distorsions [Nema and Abdul-Kareem, 2018].

Segmentation : Le signal vocal est caractérisé par sa nature non stationnaire, ce qui rend sa manipulation complexe. Cependant, sur de courtes périodes, il peut être considéré comme stationnaire et invariant. C'est à ce moment que la technique de segmentation entre en jeu. La segmentation est une étape cruciale du

traitement du signal audio qui implique de diviser le signal continu en segments courts appelés trames ou frames, ce qui facilite son analyse et son traitement ultérieur. En segmentant le signal, il devient possible d'extraire des caractéristiques spécifiques à chaque partie du signal, fournissant ainsi des informations plus précises et détaillées sur ses différentes composantes [Sakran et al., 2017].

Elle est définie par les paramètres suivants :

— Taille de trame : Elle est généralement choisie pour garantir la stationnarité du signal à l'intérieur de chaque trame. Les valeurs couramment utilisées se situent entre 20 et 30 ms, permettant ainsi de capturer des portions du signal où les propriétés acoustiques sont relativement constantes.

— Décalage entre trames (overlap) : Le décalage entre les trames détermine le chevauchement entre les segments adjacents. Souvent, un chevauchement de 50% est utilisé, ce qui signifie que chaque trame se chevauche avec la moitié de la trame précédente et la moitié de la trame suivante. Cela garantit une meilleure transition entre les segments tout en préservant les variations temporelles.

Fenêtrage : Le fenêtrage est une étape spécifique qui intervient après la segmentation pour atténuer les discontinuités aux extrémités de chaque trame, susceptibles de causer des distorsions dans les caractéristiques spectrales et temporelles du signal, et ainsi affecter la précision des analyses ultérieures. Cette étape consiste à appliquer une fonction de fenêtre à chaque trame individuellement. Cette fonction est généralement une fonction mathématique comme la fenêtre de Hamming ou la fenêtre de Hann, qui réduit l'amplitude des échantillons aux bords de la trame. Le choix de la fonction de fenêtre dépend du contexte et des objectifs spécifiques de l'analyse [Aparna and Chithra, 2017].

II.4.1.1.1 Spectrogramme : L'intensité du signal évolue au fil du temps à différentes fréquences de la forme d'onde, comme le montre le spectrogramme [Kamp]. Le spectrogramme peut se présenter sous la forme d'un graphique à deux dimensions comportant une variable couleur (voir figure II.2.a) ou d'un graphique à trois dimensions comportant une variable couleur (voir figure II.2.b).

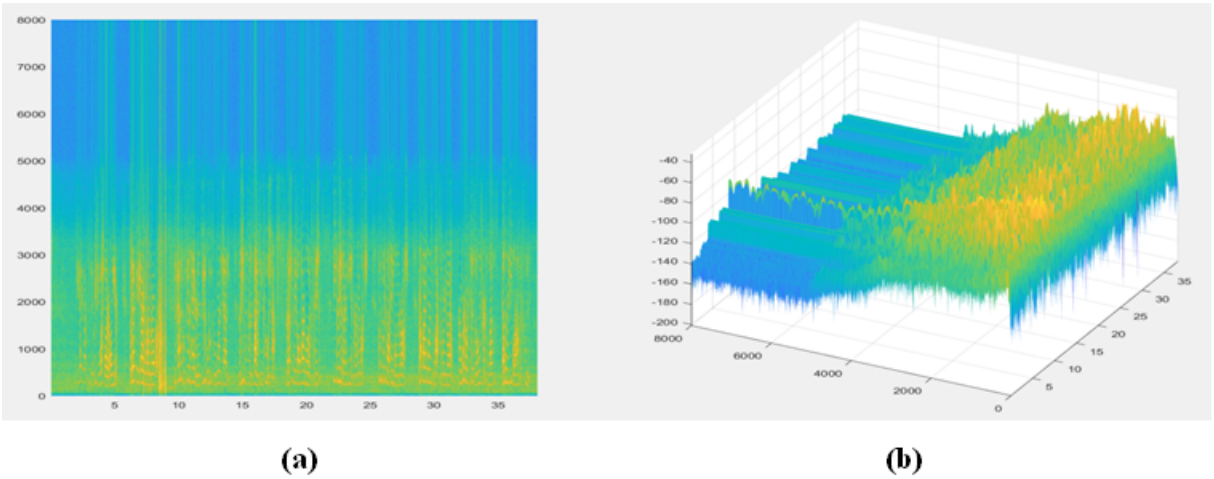

FIGURE II.2 – Spectrogramme en 2D (a) et 3D (b).

II.4.1.2 Prétraitement de l'image facial

II.4.1.2.1 Détection de visage : est une première procédure de notre méthode de reconnaissance d'expressions faciales des images qui consiste à délimiter la zone d'intérêt par un rectangle. Pour réaliser cette tâche, nous avons utilisé un détecteur de visage rapide et robuste. Il a été initialement élaboré par P. Viola et M. Jones [Viola, 2001]. Ce détecteur utilise les descripteurs de Haar et des classifieurs en cascade.

FIGURE II.3 – Détection de visage (Viola et Jones).

-Les descripteurs de Haar : Le descripteur de HAAR est un détecteur qui sert à détecter les objets dans des images ou les régions d'intérêt en temps réel. Les descripteurs de Haar sont des modelés rectangulaires simples de taille et de position spécifique chaque rectangle compose de partie blanche et noire (voir figure II.4).

-Cascade de classifieur : Le descripteur de Haar est destiné pour la détection des caractéristiques des image utilisant des classifieur de cascade qui est une méthode effective pour détection des objets.

Le principe de cette fonction de cascade et de faire l'apprentissage avec des images positives (with face) et des images négatives (without face) et donc appliquer pour la détection des objets d'autre image. Utilisant l'algorithme (AdaBoost) qui accélère les caractéristiques de Haar.

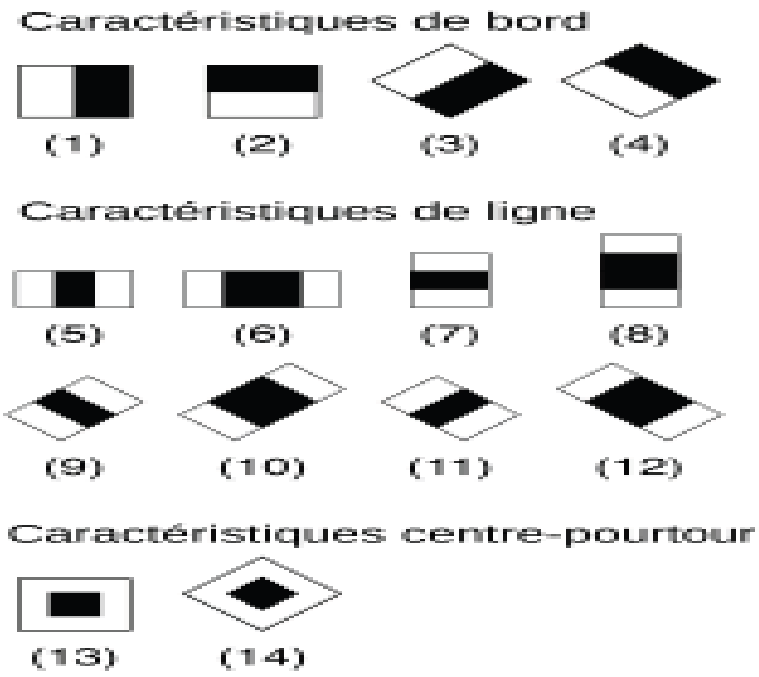

FIGURE II.4 – Exemples des caractéristiques de Haar [Viola, 2001].

II.4.1.2.2 Multiscale Retinex (MSR) : L'algorithme Multiscale Retinex (MSR) est appliqué comme une technique d'amélioration d'image [Jiang et al., 2015]. Cet algorithme améliore la qualité des images en élargissant leur plage dynamique et en maintenant la précision des couleurs. La procédure de mise en œuvre des étapes de MSR est décrite comme suit.

Étapes de MSR

1. Pour chaque canal de couleur R, G, B, calculer la sortie Retinex pour N échelles de fonctions de voisinage gaussiennes.
2. Combiner les sorties Retinex multi-échelles pour chaque canal.
3. Appliquer la restauration des couleurs pour améliorer la fidélité des couleurs.
4. Normaliser la sortie à la plage affichable.

Formulation Mathématique

Étant donné une image d'entrée $I(x, y)$, l'algorithme MSR est appliqué à chaque canal de couleur indépendamment. Le calcul du Retinex pour une seule échelle s est donné par :

$$R_s(x, y) = \log(I(x, y)) - \log(F_s(x, y) * I(x, y)) \tag{II.1}$$

où $F_s(x, y)$ est la fonction de voisinage gaussienne à l'échelle s, et $*$ désigne la convolution.

La sortie Retinex multi-échelle pour un canal de couleur est la somme pondérée des sorties à échelle unique :

$$R_{MSR}(x,y) = \sum_{s=1}^{N} w_s R_s(x,y) \tag{II.2}$$

où w_s sont les poids pour chaque échelle, satisfaisant $\sum_{s=1}^{N} w_s = 1$.

La restauration des couleurs est appliquée pour améliorer la fidélité des couleurs, définie comme :

$$C(x,y) = \beta \log\left(\alpha I(x,y)\right) \tag{II.3}$$

où α et β sont des paramètres contrôlant la force de la restauration des couleurs.

La sortie finale de MSR est obtenue en combinant la sortie Retinex multi-échelle avec la fonction de restauration des couleurs :

$$I_{MSR}(x,y) = R_{MSR}(x,y) \cdot C(x,y) \tag{II.4}$$

II.4.2 Extraction de Caractéristiques

Nous pouvons catégoriser ces méthodes en deux types : celles qui se fondent sur l'extraction de caractéristiques handcrafted et celles qui s'appuient sur l'apprentissage profond.

II.4.2.1 Méthode basé sur les CNN

L'apprentissage profond est en effet une branche de l'intelligence artificielle qui se base sur des réseaux de neurones, s'inspirant du fonctionnement des neurones du cerveau humain. Ces techniques sont utilisées pour résoudre des problèmes complexes tels que la reconnaissance d'image ou de la parole [Huang et al., 2014, Altaher et al., 2020]. L'objectif principal de ces techniques est d'identifier des caractéristiques communes dans les données d'entraînement.

Dans notre travail, nous avons utilisés les Réseaux de Neurones Convolutifs (CNN). Ces réseaux sont particulièrement adaptés à la reconnaissance d'images grâce à leur capacité à apprendre des motifs et des structures à partir des pixels d'une image. Les CNN sont composés de couches de convolution, de pooling et de couches entièrement connectées, ce qui leur permet d'extraire des caractéristiques pertinentes à partir des données d'entrée [Chauhan et al., 2018].

II.4.2.1.1 Réseau de Neurones Convolutif (CNN) :

Le CNN se compose de deux blocs extraction des caractéristiques et classification. Ils dépendent entièrement des quatre couches principales : Convolution, la correction ReLU, Pooling et Couche entièrement connectée (voir la figure II.5).

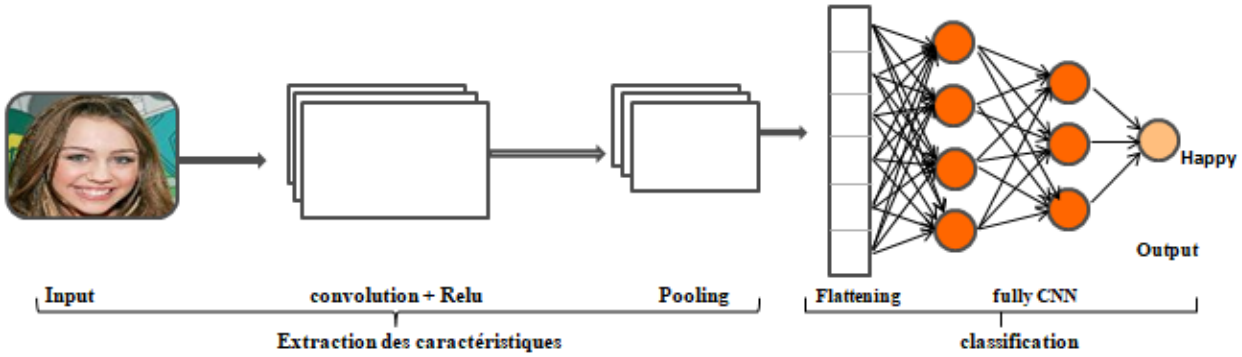

FIGURE II.5 – Réseau de Neurones Convolutif.

La couche convolution : Les opérations mathématiques importantes, telles que la phase de prétraitement des données utilisées dans les images, permettent l'extraction des caractéristiques les plus pertinentes en utilisant des filtres (par exemple : gaussiens). La convolution a quatre hyperparamètres :

— Le nombre de filtres : définit la profondeur du volume de sortie.

— La taille du filtre (ou noyau) : chaque filtre a des dimensions $F \times F \times D$ pixels.

— Stride (pas) : contrôle la manière dont le filtre se déplace sur l'image d'entrée pendant la convolution.

— Padding (remplissage) : cette technique consiste à ajouter des valeurs de zéro autour des bords de l'image ou de la carte des caractéristiques avant d'appliquer la convolution.

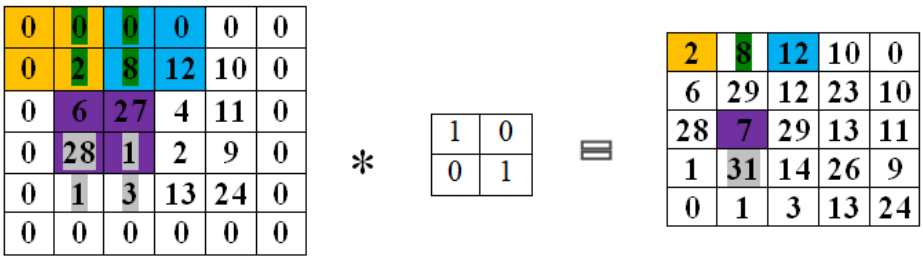

FIGURE II.6 – Exemple convolution avec un filtre (2*2) et pas égale à 1.

la correction ReLU (abréviation de Unité Linéaire Rectifiée) :

fonction est couramment en deep learning permet d'appliquer un filtre en sortir de couche elle laisse passer les valeurs positives et bloque les valeurs négatives, Cette fonction s'appelle "fonction d'activation"

$$\int(x) = \max(0, x) \tag{II.5}$$

$$RelU(x) = \begin{cases} x, x > 0. \\ 0, sinon. \end{cases} \tag{II.6}$$

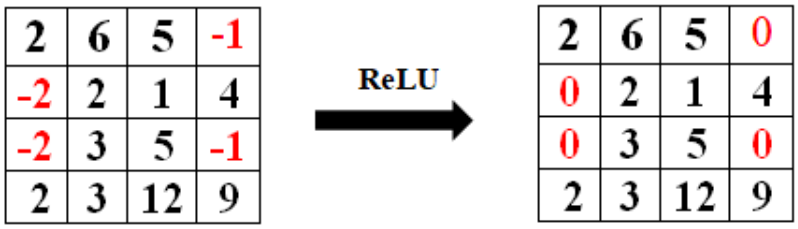

FIGURE II.7 – Exemple de la correction ReLU.

La couche Pooling : il permet de réduire la taille spatiale de la représentation en conservant uniquement les informations les plus importantes, il y a trois types de Pooling le MaxPooling (prend le maximum des valeurs de chaque région), MinPooling (prend le minimum des valeurs de chaque région) et MeanPooling (prend la moyenne des valeurs de chaque région).

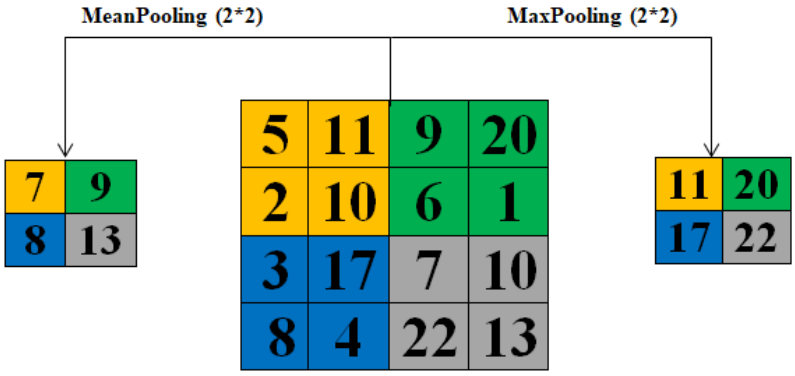

FIGURE II.8 – Exemple Pooling avec un filtre (2*2) et pas égale à 2.

Flattening : après l'extraction des caractéristique (Convolution, Pooling), les sorties sont transformées (mise à plat) en un seul vecteur linéaire ou unidimensionnel qui porte toutes les caractéristiques de l'image.

Couche entièrement connectée :

généralement, cette couche est située à la fin de chaque architecture CNN. A l'intérieur de cette couche, chaque neurone est connecté à tous les neurones de la couche précédente, l'approche dite Fully Connected (FC). elle est utilisé comme le classifieur CNN. elle suit la méthode de base du multicouche conventionnel réseau neuronal perceptron, car elle s'agit d'un type d'ANN. L'apport de la couche FC

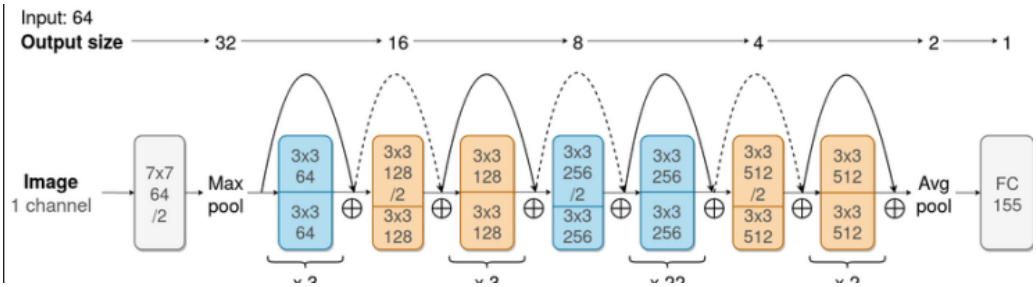

FIGURE II.9 – Architecture resnet-101.

provient de la dernière couche de mise en commun ou de convolution. Cette entrée est sous la forme d'un vecteur qui va être la sortie de la couche FC représente la sortie CNN finale.

II.4.2.1.2 ResNet101 : ResNet-101 est un réseau neuronal convolutif composé de 101 couches de profondeur [He et al., 2016]. Une version préconfigurée de ce réseau, formée sur une base de données de plus d'un million d'images provenant d'ImageNet, est disponible au téléchargement. Ce réseau pré-entraîné est capable de classifier des photographies dans 1000 catégories différentes, incluant des objets tels que des claviers, des souris, des crayons et divers animaux. En conséquence de cet entraînement sur une vaste gamme d'images, le réseau a appris à représenter de manière exhaustive les caractéristiques présentes dans ces images. La taille d'entrée des images pour ce réseau est de 224 par 224 pixels RGB.

II.4.2.1.3 VGG19 : C'est une architecture typique des réseaux neuronaux modernes pour la vision par ordinateur.il se base sur un reseau neuronal convolutif(visual geometry group) dévéloppé par le groupe de recherche de visual geometry a l'université d'Oxford, Les blocs bleus et verts représentent différentes couches du réseau, avec leurs dimensions respectives indiquées [He et al., 2016]. Les blocs blues sont des couches de convolution qui extraient les caractéristiques de l'image, tandis que les blocs verts sont des couches de pooling qui réduisent la taille des caractéristiques. L'architecture commence par des couches de convolution et de pooling de faible dimension, puis augmente progressivement la profondeur et la taille des filtres dans les couches ultérieures. Cela permet d'extraire des caractéristiques de plus en plus complexes et abstraites à partir de l'image d'entrée. À la fin, il y a deux couches entièrement connectées (FC1 et FC2) de taille 4096, qui combinent les caractéristiques extraites pour effectuer la classification finale. Cette architecture en forme de pyramide est courante dans les réseaux neuronaux convolutionnels modernes, car elle permet une extraction efficace des caractéristiques à partir des images tout en réduisant progressivement la dimension spatiale des représentations pour une classification finale. Malgre le VGG19 effectue la classification d'images avec precision,sa profendeur peut egalement rendre l'entrainement

plus couteux en termes de temps et de ressouces.

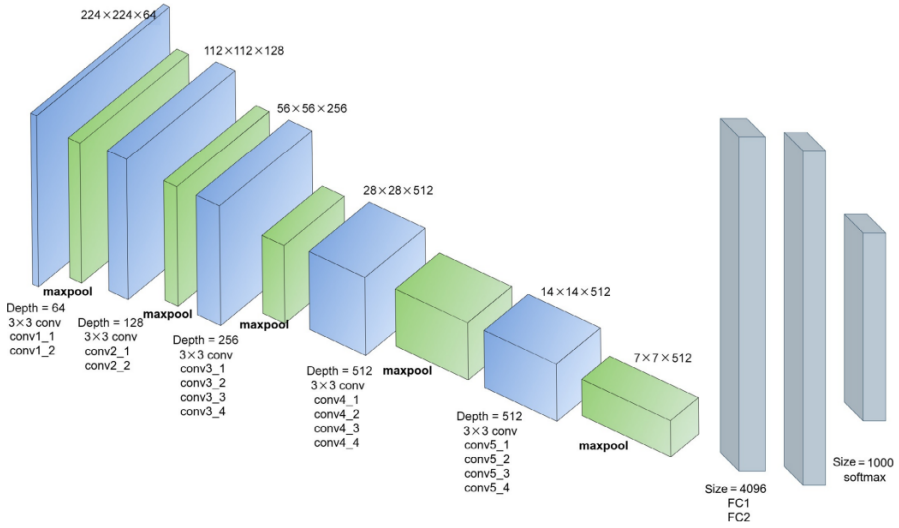

FIGURE II.10 – Architecture VGG19.

II.4.2.1.4 VGGish : L'architecture VGGish, introduite en 2017 [Hershey et al., 2017], a été conçue pour des tâches de classification audio à grande échelle. Ce modèle a été entraîné sur le dataset Youtube-100M, qui comprend 5,24 millions d'heures de vidéos. Pour traiter les données audio, celles-ci sont segmentées en trames non chevauchantes de 960 millisecondes. Ensuite, des spectrogrammes log-mel bidimensionnels de 96 × 64 sont générés via un processus de conversion temps-fréquence utilisant la transformée de Fourier à court terme et l'intégration de 64 bandes de fréquences espacées mel. L'architecture VGGish, illustrée dans la Figure II.11, se compose de 62 millions de poids.

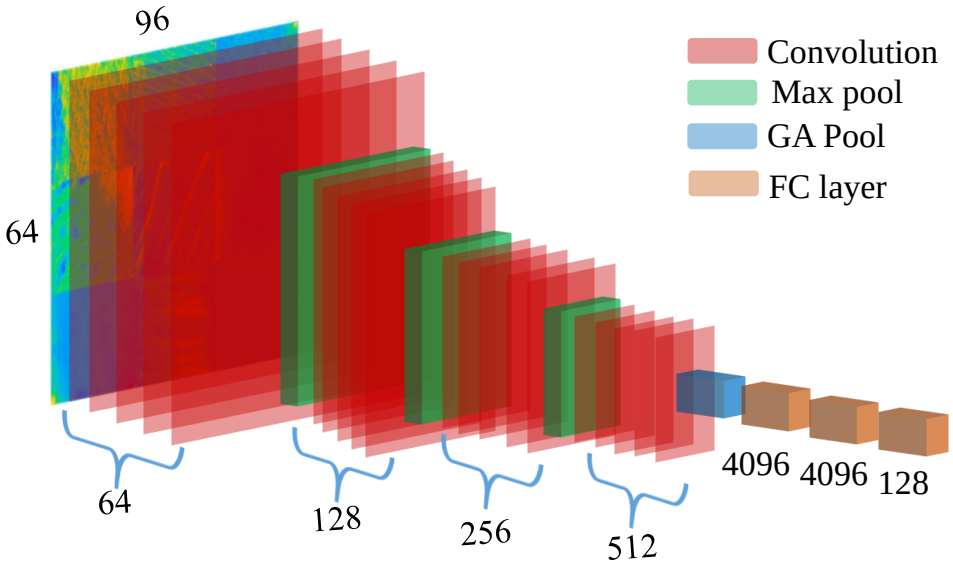

FIGURE II.11 – Architecture VGGish.

II.4.2.2 Caractéristiques Handcrafted

Malgré les performances remarquables des systèmes de reconnaissance des émotions faciales basés sur les caractéristiques profondes, ces derniers exigent un grand nombre de paramètres, un ajustement minutieux et des ensembles de données d'entraînement étendus [Fang et al., 2010]. Ces limitations posent des défis significatifs pour un traitement en temps réel. Pour cette raison, de nombreuses méthodes s'appuient sur des descripteurs handcrafted, qui ont fait leurs preuves dans plusieurs domaines des systèmes de reconnaissance, comme les descripteurs utilisés dans nos travaux dans cette section .

II.4.2.2.1 Modèles binaires locaux (LBP) : L'opérateur LBP (Local Binary Pattern) est un descripteur de texture puissant introduit dans les années 90. Il caractérise la texture locale autour d'un pixel en lui attribuant un code basé sur la comparaison de son niveau de gris avec celui de ses 8 voisins.

Pour chaque pixel central, on calcule son code LBP en :

1. Soustrayant sa valeur de celle de chaque voisin.
2. Affectant un 0 binaire si le voisin est plus sombre, 1 s'il est plus clair.
3. Concaténant ces 8 bits dans le sens horaire à partir du voisin en haut à gauche.
4. Convertissant ce nombre binaire en décimal pour obtenir le code LBP.

Un histogramme des codes LBP de tous les pixels de l'image sert alors de signature et descripteur de texture très discriminant. Cet opérateur LBP encode donc efficacement l'information de texture de manière compacte, en attribuant à chaque pixel un code qui capture les différences locales d'intensité avec son voisinage [Ahonen et al., 2004].

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{p} s(i_n - i_c) \cdot 2^n \tag{II.7}$$

Où s(x) est une fonction de signe définie par : a fonction $s(x)$ est définie comme suit :

$$s(x) = \begin{cases} 0 & \text{si } x < 0. \\ 1 & \text{si } x \geq 0. \end{cases}$$

II.4.2.2.2 Caractéristiques des images statistiques binarisées (BSIF) : Contrairement à LBP et LPQ, le descripteur local nommé BSIF (Binarized Statistical Image Features), introduit récemment par Kannala et Rahtu, offre une approche alternative pour le calcul des statistiques d'étiquettes dans les voisins de pixels locaux [Kannala and Rahut, 2012]. Il se base sur un ensemble prédéfini de filtres

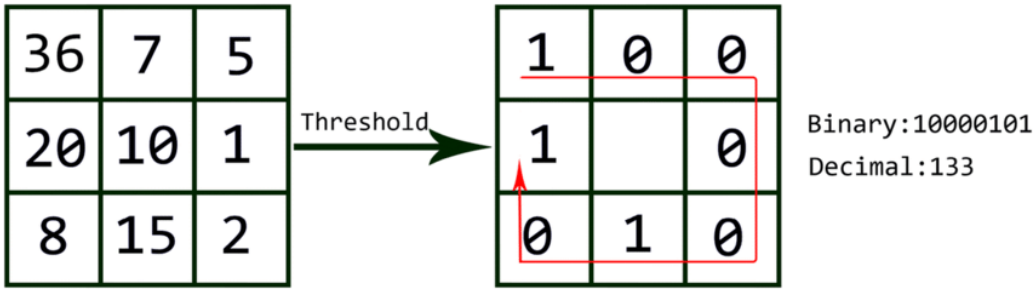

FIGURE II.12 – Un exemple de descripteur LBP.

linéaires manuels et une étape de binarisation pour filtrer les réponses . Pour une image patch X de taille l x l pixels et un filtre linéaire Wi de même taille, la réponse du filtre Si est obtenue par :

$$s_i = \sum_{u,v} W_i(u,v)X(u,v) = w_i^T x \tag{II.8}$$

Étant donné un ensemble de n filtres linéaires W_i, on peut les regrouper dans une matrice W et calculer toutes les réponses simultanément. À cette dernière étape, la notation vectorielle est introduite, où les vecteurs w et x contiennent respectivement les pixels de W_i et X. La fonction de binarisation b_i est obtenue par :

$$b_i = \begin{cases} 1 & \text{if } s_i > 0. \\ 0 & \text{otherwise.} \end{cases} \tag{II.9}$$

Ceci signifie que b_i représente le ième élément de b. Ainsi, un code binaire de série b de n bits peut être calculé pour chaque pixel, permettant ensuite de représenter la région de l'image à l'aide d'histogrammes binaires des codes de pixel.

II.4.2.2.3 Quantification de phase locale (LPQ) : Cette méthode est utilisée pour extraire un motif de texture local qui reste invariant par rapport au flou et à l'illumination. La technique LPQ, proposée par repose sur la caractéristique d'invariance au flou du spectre de phase de Fourier. LPQ examine un voisinage rectangulaire autour de chaque pixel de l'image afin de calculer la transformée de Fourier à court terme 2D (STFT), fournissant ainsi l'information de phase locale de l'image [Ojansivu and Heikkila, 2008].

Le flou d'une image est une méthode visant à atténuer le contenu des bords de l'image, permettant une transition en douceur d'une couleur à une autre. Il est représenté par une fonction spatiale, notée $g(x)$, qui résulte de la convolution entre l'image d'origine $f(x)$ et une fonction d'étalement de point (PSF) $h(x)$. En domaine fréquentiel, cette opération se manifeste par :

$$G(u) = F(u) * H(u) \tag{II.10}$$

Les transformées de Fourier discrètes (DFT) de l'image floue, de l'image originale et de la fonction de transfert d'image (PSF) sont notées respectivement $G(u)$, $F(u)$ et $H(u)$. Ici, u représente l'ensemble des coordonnées vectorielles $[u, v]^T$. Les composantes de l'amplitude et de la phase peuvent être séparées et représentées comme suit :

$$|G(u)| = |F(u)| \cdot |H(u)| \quad \text{et} \quad < G(u) =< F(u) + < H(u) \tag{II.11}$$

En notant $< G(u)$ la phase de $G(u)$, lorsque la PSF de la fonction est à symétrie centrale, sa transformée de Fourier H est toujours réelle, ce qui signifie que :

$$< H(u) = \begin{cases} 0 & \text{si } H(u) \geq 0. \\ \pi & \text{si } H(u) < 0. \end{cases} \tag{II.12}$$

L'équation II.13 illustre la propriété d'invariance du flou, où $< G(u) =< F(u)$ lorsque $H(u) = 0$. L'algorithme LPQ (Local Phase Quantization) extrait les informations de phase en examinant le voisinage local $N(x)$ de taille $M \times M$ à chaque position de pixel x de l'image $f(x)$:

$$F(u, x) = \sum_{y \in N_x} f(x - y) e^{-2j\mathbf{t}_u^T y} = w_u^T f_x \tag{II.13}$$

Ici, w_u représente le vecteur de base pour la transformée de Fourier discrète 2D à la fréquence u, et $f(x)$ est un vecteur contenant tous les échantillons m^2 de $N(x)$.

Les coefficients de Fourier locaux sont calculés à quatre points de fréquence : $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, $u_4 = [a, -a]^T$. Pour chaque position de pixel, cela produit un vecteur :

$$F_{cx} = [F(u_1, x), F(u_2, x), F(u_3, x), F(u_4, x)] \tag{II.14}$$

L'information de phase dans les coefficients de Fourier est enregistrée en observant les signes des parties réelles et imaginaires de chaque composante en f_{cx}. Ceci est réalisé en utilisant un quantificateur scalaire simple $q_j(x) = 1$ si $g_i(x) \geq 0$, et 0 sinon, où $g_i(x)$ est la composante j-ème du vecteur $G_x = [\mathrm{Re}\{F_{cx}\}, \mathrm{Im}\{F_{cx}\}]$. Les coefficients quantifiés $q_j(x)$ résultants sont représentés sous forme de valeurs entières comprises entre 0 et 255 grâce à un encodage binaire :

$$f_{lpq} = \sum_{i=1}^{8} q_i 2^{i-1} \tag{II.15}$$

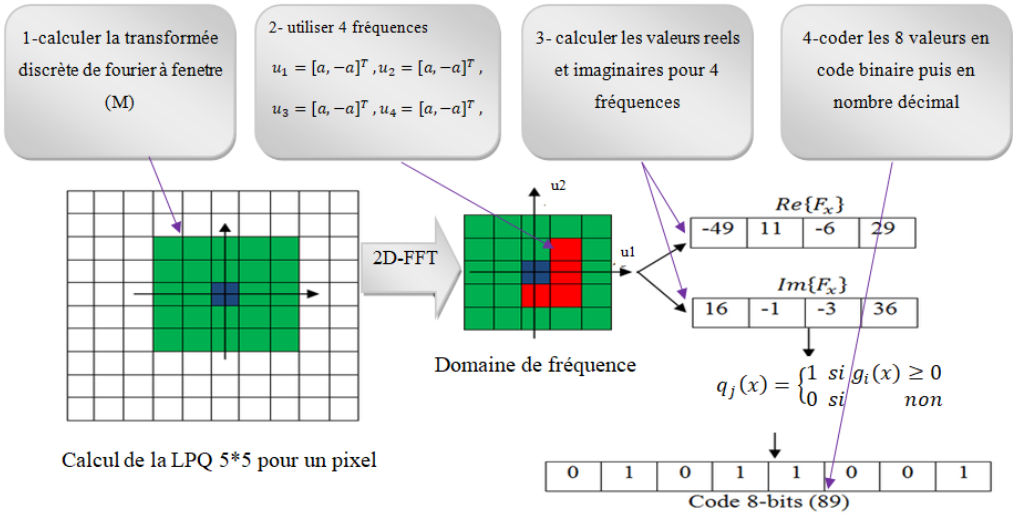

FIGURE II.13 – Les étapes nécessaires pour extraire les caractéristiques LPQ.

L'histogramme de ces valeurs entières est ensuite utilisé comme vecteur caractéristique.

II.4.2.2.4 Motif Directionnel local (LDP) : Les opérateurs de modèle directionnel local (LDP) utilisent les valeurs de réponse de bord des pixels du voisinage et coder la texture de l'image Bordure Kirsch. Le détecteur est utilisé pour trouver les réponses de bord. Les matrices suivantes montrent huit masques de Kirsch pour détecter la réponse des bords valeurs [Jabid et al., 2010].

$$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$$

$$\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

Le LDP est calculé comme suit : Le LDP attribue un code binaire de 8 bits pour chaque pixel d'une entrée image. Ce modèle est ensuite calculé par comparer les valeurs relatives de réponse de bord d'un pixel en utilisant le détecteur de bord Kirsch. Donné un pixel central de l'image, le huit directionnel les valeurs de réponse de front m_i $(i = 0, 1, 2....., 7)$ sont calculé par les masques de Kirsch. Depuis qu'un coin ou un bord présente des valeurs de réponse élevées dans certaines

directions particulières, les directions les plus importantes du nombre k avec des valeurs de réponse élevées sont sélectionnées pour générer le code LDP. En d'autre mots, réponses binaires directionnelles top-k, b_i , sont fixés à 1, et les bits restants (8 k) sont mis à 0. Enfin, le code LDP est dérivé de l'équation II.16.

$$LDP_k = \sum_{i=0}^{b} b_i(m_i - m_k) \times 2^i \tag{II.16}$$

$$b_i(x) = \begin{cases} 1 & \text{si } X \geq 0. \\ 0 & \text{si } X < 0. \end{cases} \tag{II.17}$$

Où est la $k\breve{}i$ème réponse directionnelle la plus significative. cette deux matrices represent une Réponse de bord et positions des bits binaires LDP

$$\begin{bmatrix} m3 & m2 & m1 \\ m4 & x & m0 \\ m5 & m6 & m7 \end{bmatrix} \begin{bmatrix} b3 & b2 & b1 \\ b4 & x & b0 \\ b5 & b6 & b7 \end{bmatrix}$$

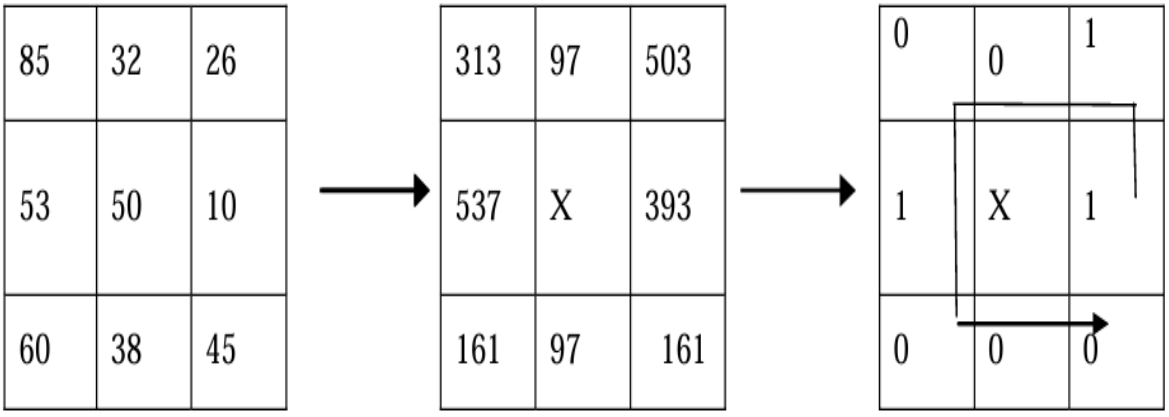

FIGURE II.14 – Un exemple de descripteur LDP de base.

Ensuite, le code binaire LDP est calculé à partir de la position du bit binaire 0 de droite à gauche, calculé comme suit : Code binaire LDP : 00010011 Code décimal LDP :19

II.4.2.2.5 Coefficients cepstraux en fréquence Mel (MFCC) Les coefficients cepstraux sont calculés en utilisant la transformée en cosinus discrète DCT combinée avec le coefficient d'énergie. La conversion du spectre de puissance du signal en coefficients d'énergie par bande de fréquence est effectuée par l'analyse du banc de filtres à l'échelle Mel, puis une compression logarithmique est appliquée à ces coefficients [Klautau, 2005].

Les différentes étapes de cette extraction, illustrées dans la figure II.15, sont exposées ci-dessous :

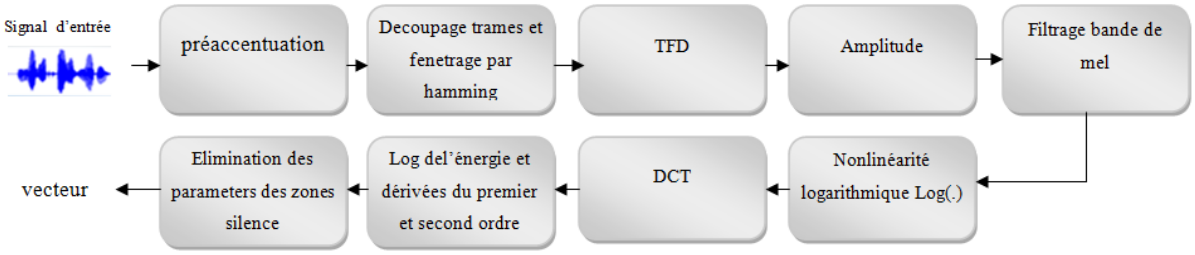

FIGURE II.15 – Extraction des paramètres MFCC.

L'accentuation préalable permet d'accentuer la partie supérieure de la fréquence. Afin de réduire la distorsion spectrale et pour des raisons mentionnées précédemment (paragraphe II.4.1.1) une multiplication par une fenêtre de Hamming (figure II.16) (la plus appropriée pour la voix) est réalisée.

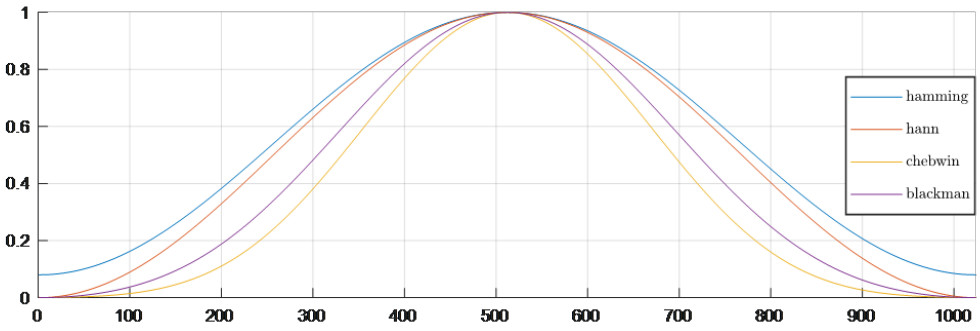

FIGURE II.16 – des examples des fenêtres.

On divise alors le signal parole en différentes trames telles que :

$$y(n,t) = x(n,t) \times w(n), n = 0,1,2,\ldots N-1 \; ; t = 0,1,2,\ldots,T-1 \quad \text{(II.18)}$$

Avec $x(n,t)$: le signal parole original, N : nombre d'échantillons, T : nombre de trames,$w(n)$: est la fenêtre de Hamming donnée par :

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi}{N-1}n\right), \; 0 \leq n \leq N \quad \text{(II.19)}$$

— Dans notre travail, nous utilisons un chevauchement de 10ms entre les trames consécutives pour maintenir les paramètres au maximum et éviter les interruptions entre les trames.

— La deuxième étape implique la transition du domaine temporel vers le domaine fréquentiel en utilisant une transformation discrète telle que :

$$Y(k,t) = \frac{1}{N}\sum_{n=0}^{N-1} y(n,t)\exp\left(\frac{-2\pi jkn}{N}\right), \quad k = 0,1,\ldots,M-1;\ t = 0,1,\ldots,T-1. \tag{II.20}$$

Dans le cas complexe, on ne prend en compte que la valeur absolue. L'intervalle de fréquence $0 \leq f \leq f_e$ Correspond à $0 \leq k \leq \frac{M}{2} - 1$ et la gamme de $-\frac{f_e}{2} \leq f \leq 0$ correspond à $\frac{M}{2} + 1 \leq k \leq M - 1$.

— Le module est utilisé pour générer le spectre de puissance de chaque trame, ce qui représente la troisième étape du processus.

— Étant donné que l'échelle de perception de fréquence de l'oreille humaine n'est pas linéaire, la prochaine étape de traitement consiste à filtrer le spectre de puissance du signal vocal à l'aide d'un ensemble de k filtres placés selon l'échelle de Mel (Figure II.17).

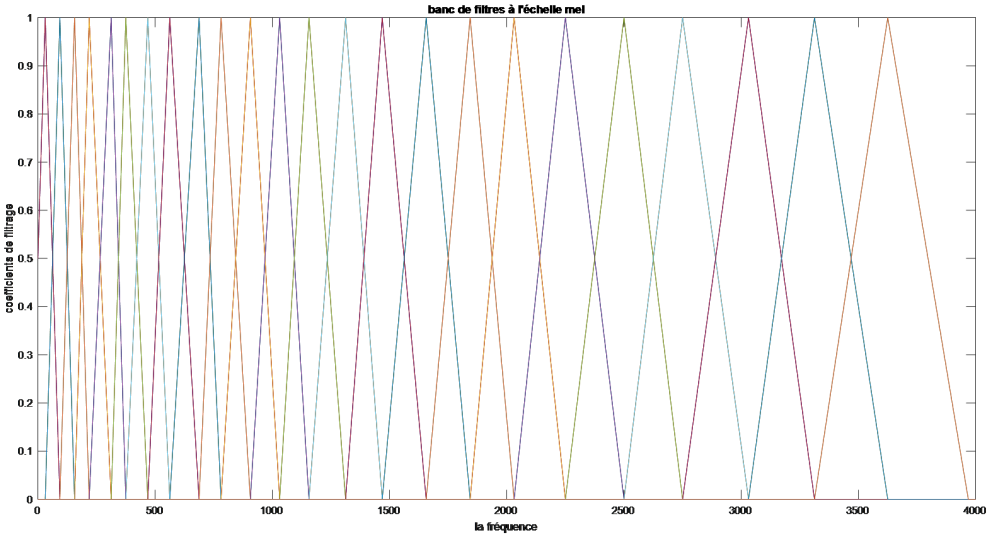

FIGURE II.17 – Filtre Mel.

Les filtres passe-bande triangulaires constituent l'échelle, permettant ainsi un positionnement linéaire de basse fréquence (<1000 Hz) et un positionnement logarithmique de haute fréquence. Le premier filtre signale une très faible quantité d'énergie au voisinage de 0 Hz, et cette largeur augmente jusqu'aux fréquences élevées. Ceci reproduit le fonctionnement de l'oreille humaine, qui n'est pas sensible à ces fréquences et ne peut distinguer deux fréquences très proches. Après cette étape, un vecteur indicant la quantité d'énergie présente dans chaque filtre est généré. [Klautau, 2005]. Le passage à cette échelle se fait par l'équation suivante :

$$m = \ln(1 + \frac{f}{700}) \times \frac{1000}{\ln(1 + \frac{1000}{700})} \tag{II.21}$$

— L'application de la transformée en cosinus discrète (DCT) sur le vecteur E_W d'énergies spectrales logarithmiques est la dernière étape de ce processus, afin d'éviter toute compression de l'information [Klautau, 2005]. Tous ces coefficients cepstraux produits sont désignés sous le nom de vecteur acoustique. La DCT équation est la suivante :

$$C_m = \sum_{w=1}^{k} \cos\left(m\,(w - 0.5)\,\frac{\pi}{k}\right) E_w \ , m = 1, 2, \ldots, L \tag{II.22}$$

Avec k le nombre de filtres passe-bande triangulaires, L le nombre de coefficients cepstraux à l'échelle de Mel.

II.4.3 Reduction de dimensionnalité

Le principal défi associé au vecteur réside dans sa haute dimensionnalité. Ainsi, il est essentiel de projeter le vecteur de caractéristiques dans un espace de dimension inférieure contenant uniquement des informations discriminantes. Cette étape de réduction de dimensionnalité améliore l'efficacité du calcul du système de reconnaissance et prévient les problèmes techniques, tels que la malédiction de la dimensionnalité. La réduction de dimensionnalité peut contourner ce problème en réduisant le nombre d'entités dans l'ensemble de données avant le processus de formation. De plus, cela peut réduire le temps de calcul et la taille de stockage des classificateurs. Cependant, le principal inconvénient de la réduction de dimensionnalité est le risque de perte d'informations. Une mauvaise exécution de cette étape peut entraîner la suppression d'informations pertinentes plutôt que d'informations non pertinentes, et cette perte d'informations est irrécupérable quel que soit le traitement ultérieur effectué. Il existe de nombreuses méthodes de réduction de dimensionnalité utilisées en science des données pour différents types d'applications, telles que l'analyse en composantes principales (ACP), l'analyse discriminante linéaire (LDA), l'analyse discriminante des informations secondaires (SIDA), l'ACP multilinéaire (MPCA), le SIDA multilinéaire (MSIDA), l'analyse discriminante quadratique croisée XQDA [Van Der Maaten et al., 2009].

II.4.3.1 Analyse des Composants Principaux (PCA)

L'Analyse en Composantes Principales (ACP) est une méthode qui permet de définir un sous-espace à partir d'un ensemble de données d'apprentissage, ce qui permet de sauvegarder des informations distinctives tout en éliminant les

informations secondaires (non informatives). Cette approche consiste à trouver une nouvelle base dans l'espace de données où tous les vecteurs sont orthogonaux les uns aux autres. Le premier vecteur de cette base correspond à la direction de la variance maximale des données d'apprentissage. Les autres composantes sont déterminées en respectant des contraintes orthogonales entre les vecteurs, tout en tenant compte de la direction de la variance maximale.

Dans la méthode de l'ACP, la standardisation de l'éclairage est toujours essentielle. Cette technique est largement utilisée en reconnaissance de formes en raison de sa rapidité et de sa simplicité. Elle est considérée comme la meilleure approche pour reconstruire une base de dimension réduite, car les projections de l'ACP sont optimales. L'ACP implique la recherche des vecteurs propres de la matrice de covariance formée à partir des différentes images de notre ensemble d'apprentissage, selon la procédure suivante :

Étape 1 : Sélectionnez la matrice de données X, avec X^T ayant une moyenne nulle.

Étape 2 : Calculer la moyenne :

$$\Psi = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Étape 3 : Soustraire la moyenne de la distribution à partir de l'ensemble de données :

$$X_i = X^T - \Psi$$

Étape 4 : Calculer la matrice de covariance XX^T :

$$C = \sum_{i=1}^{N} X_i X_i^T$$

Étape 5 : Calculer les valeurs propres et les vecteurs propres V de la matrice de covariance, où $i = 1, ..., N$.

Étape 6 : Ordonner les vecteurs propres V_i (avec $i = 1, ..., N$) par leurs valeurs propres correspondantes λ_i par ordre décroissant.

Étape 7 : Ne conserver que les vecteurs propres avec les valeurs propres les plus importantes (les composants principaux), k (où $k \ll N$) :

$$X_k = V_k \cdot X$$

Étape 8 : Résoudre pour l'Analyse en Composantes Principales (ACP) :

$$\lambda V^T X^T = C X V^T$$

II.4.3.2 Analyse discriminante linéaire (LDA)

La méthode LDA [Ambikairajah et al., 2012] est employée pour atténuer l'impact du canal. Son objectif est de minimiser la variabilité intra-classe tout en maximisant la variabilité inter-classes. La matrice de projection

A de cette méthode est déterminée en résolvant le problème des valeurs propres suivant

$$S_b * v = \lambda * S_w, \tag{II.23}$$

où s_b et s_w représentent respectivement la matrice de variabilité inter-classes et la matrice de variabilité intra-classes.

$$S_b = \sum_{s=1}^{S} (w_s - w)(w_s - w)^T \tag{II.24}$$

$$S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_{si} - w_s)(w_{si} - w_s)^T \tag{II.25}$$

Ici, S est le nombre de classes, n_s est le nombre de I-Vecteurs pour chaque classe, w_s est la moyenne des I-Vecteurs pour chaque classe, et w est la moyenne de tous les I-Vecteurs (CM 2007). La moyenne sur tous les locuteurs est définie par :

$$w_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_{si} \tag{II.26}$$

II.4.4 Classification

Cette section se concentre sur les algorithmes utilisés pour la classification des émotions faciales et vocales. Nous abordons deux approches distinctes : la première repose sur l'algorithme d'apprentissage automatique SVM, tandis que la seconde utilise la distance euclidienne et la distance cosinus, des mesures couramment employées pour l'apprentissage métrique dans divers domaines [Khammari et al., 2023].

II.4.4.1 Machines à Vecteurs de Support (SVM)

Les machines à vecteurs de support (SVM) cherchent l'hyperplan optimal pour séparer deux classes de données en maximisant la marge, c'est-à-dire la distance aux points les plus proches de chaque classe. Un hyperplan avec une large marge offre une meilleure généralisation et robustesse, tandis qu'une petite marge rend le modèle plus sensible au bruit (voir la figure II.18 montrant le fonctionnement SVM). L'objectif est de trouver le séparateur linéaire maximisant la marge entre les classes pour améliorer les performances de classification [Jakkula, 2006].

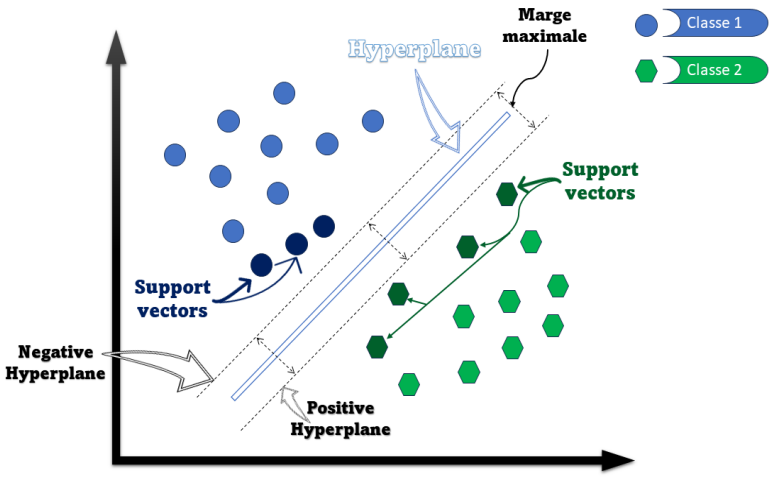

FIGURE II.18 – Machines à Vecteurs de Support (SVM).

II.4.4.2 K-plus proche voisin classifieur (KNN)

Le classificateur K-NN (K-Nearest Neighbors) est un système qui permet de classer des données inconnues en les reliant à des données connues en utilisant des mesures de similarité. Le fonctionnement de cet algorithme consiste à identifier la classe dominante parmi les k voisins les plus proches d'une instance donnée. C'est une approche simple qui repose sur le principe du vote majoritaire parmi les voisins les plus proches [Wani et al., 2022].

KNN est adapté aux problèmes de classification et peut également être utilisé pour des tâches de régression. En pratique, il est souvent privilégié dans l'industrie pour les problèmes de classification (voir la figure II.19 exemple pour K-NN). Les mesures de similarité utilisées pour déterminer les voisins les plus proches peuvent être basées sur différentes distances telles que la distance euclidienne, de Manhattan, de Minkowski ou de Hamming (pour les variables catégorielles) [Wani et al., 2022].

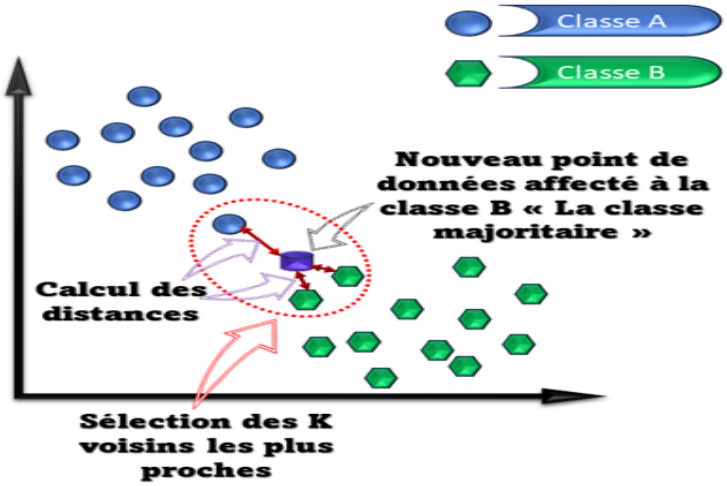

FIGURE II.19 – k plus proches voisins (K-NN).

II.4.5 Décision

La reconnaissance automatique des émotions (RAE) est conçue de manière à ce que la comparaison entre le modèle, représentant les émotions préalablement identifiées, et le signal de test donne un résultat sous forme de valeur scalaire . Cette valeur indique si les émotions détectées dans le signal de test correspondent à celles du modèle. Dans le cas de l'identification, la phase de décision détermine l'identité des émotions reconnues dans le signal de test [Chen et al., 2020]. En revanche, dans le processus de vérification, la décision est binaire et vise à confirmer ou infirmer la cohérence des émotions exprimées dans la session de test avec celles revendiquées par le modèle. Si le résultat obtenu lors de la comparaison dépasse (ou ne dépasse pas) le seuil prédéfini, le système accepte (ou rejette) les émotions détectées dans le signal de test comme correspondant au modèle.

II.5 Métrique évaluation

Les métriques d'évaluation jouent un rôle essentiel dans le développement et l'amélioration des systèmes de reconnaissance faciale. Elles agissent comme des mesures quantifiables qui nous permettent de comprendre la performance des systèmes dans différentes conditions [Tolba et al., 2015]. Les principales métriques pour évaluer les systèmes de reconnaissance faciale comprennent :

— **Accuracy :** La précision est une métrique d'évaluation fondamentale et très courante pour tout système de classification, y compris la reconnaissance faciale. Elle mesure la justesse globale du modèle en quantifiant le nombre de prédictions correctes effectuées sur le nombre total de prédictions. Mathématiquement, la précision est calculée comme suit (Ainsi que le démontre cette équationII.27) :

$$accuracy = \frac{Nombre\,de\,Prdictions\,Correctes}{Nombre\,Total\,de\,Prdictions} \tag{II.27}$$

— **Precision :** La précision mesure la proportion d'identifications positives qui étaient effectivement correctes. Dans le contexte des systèmes de reconnaissance faciale, elle fait référence à la proportion de correspondances positives correctes parmi toutes les correspondances positives identifiées par le système. Mathématiquement, la précision peut être calculée comme suit (Ainsi que le démontre cette équationII.28) :

$$Prcision = \frac{Vrais\,Positifs}{Vrais\,Positifs + Faux\,Positifs} \tag{II.28}$$

— **Recall :** Il s'agit de la capacité d'un système à trouver toutes les instances pertinentes au sein d'un ensemble de données. Dans les systèmes de recon-

naissance faciale, le rappel fait référence à la proportion de correspondances positives réelles qui ont été correctement identifiées par le système. Mathématiquement, le rappel est calculé comme suit (Ainsi que le démontre cette équation II.29) :

$$Racall = \frac{Vrais\,Positifs}{Vrais\,Positifs + Faux\,Ngatifs} \tag{II.29}$$

— **F1-score :** Le score F1 est la moyenne harmonique de la précision et du rappel. La moyenne harmonique est utilisée ici plutôt que la moyenne arithmétique simple car elle pénalise davantage les valeurs extrêmes. En termes mathématiques, il est défini comme suit(Ainsi que le démontre cette équationII.30) :

$$F1 = \frac{2 \times Prcision \times Rappel}{Prcision + Rappel} \tag{II.30}$$

— **Taux de fausse acceptation (FAR) :**Le FAR est la mesure de la probabilité que le système identifie ou vérifie incorrectement un individu comme quelqu'un d'autre. C'est une mesure significative de l'efficacité de sécurité du système (Ainsi que le démontre cette équationII.31) :

$$FAR = \frac{\text{Nombre de Faux Acceptations}}{\text{Nombre de Tentatives d'Identification}} \tag{II.31}$$

Un FAR élevé dans un système indique qu'il est moins sécurisé, car il permet souvent l'accès à des individus non autorisés. Dans les applications de reconnaissance faciale telles que les systèmes de sécurité ou le déverrouillage de smartphones, un FAR plus faible est toujours préféré pour garantir une haute sécurité.

— **Taux de faux rejets (FRR) :** Le FRR indique la probabilité que le système rejette incorrectement une identité qui correspond. Le FRR est une mesure cruciale de la convivialité du système(Ainsi que le démontre cette équationII.32) :

$$FRR = \frac{\text{Nombre de Faux Rejets}}{\text{Nombre de Tentatives d'Identification}} \tag{II.32}$$

Un FRR élevé indique que le système n'est pas convivial, car il refuse souvent l'accès à des individus autorisés. Pour la satisfaction de l'utilisateur, un FRR plus faible est généralement souhaité.

— **Taux d'erreur égal (EER) :**C'est le point où le FAR et le FRR sont égaux. Plus l'EER est bas, meilleure est la précision du système.

— **Caractéristiques de fonctionnement du récepteur(ROC) :** Elles per-

mettent de comparer l'efficacité de différents modèles et aident les concepteurs de systèmes à choisir le seuil optimal pour une performance maximale. Cependant, dans les ensembles de données déséquilibrés où les instances positives (individus autorisés) sont nettement plus nombreuses que les instances négatives (individus non autorisés), les courbes Précision-Rappel pourraient fournir une image plus informative que les courbes ROC.

II.6 Conclusion

Le but de ce chapitre était de présenter une vue d'ensemble des solutions mises en œuvre jusqu'à présent pour la reconnaissance des emotions,Nous avons commencé par évoquer les différentes bases de données publiques disponibles et à définir plusieurs approches utilisées par les chercheurs dans ce domaine. Dans la première section, nous avons examiné les méthodes de prétraitement du visage et d'extraction des caractéristiques, et nous avons conclu le chapitre en détaillant les méthodes de réduction des caractéristiques ainsi que les algorithmes d'apprentissage et de classification.

Dans le chapitre suivant, nous aborderons la conception de notre système et les approches proposées que nous allons utiliser.

Chapitre III

Présentation de la Solution

III.1 Introduction

Actuellement, les modèles d'intelligence artificielle peuvent atteindre une précision comparable à celle des humains dans l'analyse et la reconnaissance d'images et de la voix. Motivés par le succès impressionnant des approches d'intelligence artificielle dans la représentation et la classification de diverses images et de la voix, nous avons proposé une contribution pour la reconnaissance automatique des émotions. Notre travail consiste à combiner deux systèmes différents basés sur des caractéristiques faciales et vocales.

Dans ce chapitre, nous détaillerons les différentes étapes nécessaires à la mise en œuvre de notre système automatique de reconnaissance des émotions

III.2 Méthodologie proposée

Notre système multimodal de reconnaissance des émotions, illustré à la Figure III.4, résulte de la fusion de deux systèmes distincts : le premier basé sur l'audio et le second sur les données visuelles (voir les Figures III.1, III.3 et III.4, respectivement). Chaque système comprend deux phases : la phase d'apprentissage et la phase de test. Chacune de ces phases se divise en quatre étapes essentielles.

Premièrement, il y a le prétraitement des données. Ensuite, vient l'extraction des caractéristiques. Après cela, la réduction de la dimensionnalité et la classification sont effectuées à l'aide de la projection non linéaire dans un sous-espace. Enfin, la correspondance est réalisée par la similarité cosinus, et les scores des deux systèmes sont fusionnés pour tirer parti de leur complémentarité.

Les sections suivantes fournissent une explication détaillée de chaque étape des systèmes audio et visuel.

III.2.1 Système audio pour la reconnaissance des émotions

Notre système de reconnaissance des émotions basé sur le signal de parole est illustré à la Figure III.1. Les étapes essentielles sont les suivantes : premièrement, le prétraitement du signal de parole, puis l'extraction des caractéristiques audio. Nous détaillons chaque étape dans les sous-sections suivantes.

III.2.1.1 Audio vers représentation visuelle (spectrogramme)

Dans cette sous-section, nous décrivons les étapes clés du prétraitement des données. L'étape initiale avant de convertir l'audio en image spectrogram consiste à normaliser le signal. Des différences notables existent dans l'amplitude audio pour la même espèce vocale en raison de conditions d'enregistrement variables [Ji et al., 2021]. Pour normaliser, l'audio a été normalisé en soustrayant la moyenne

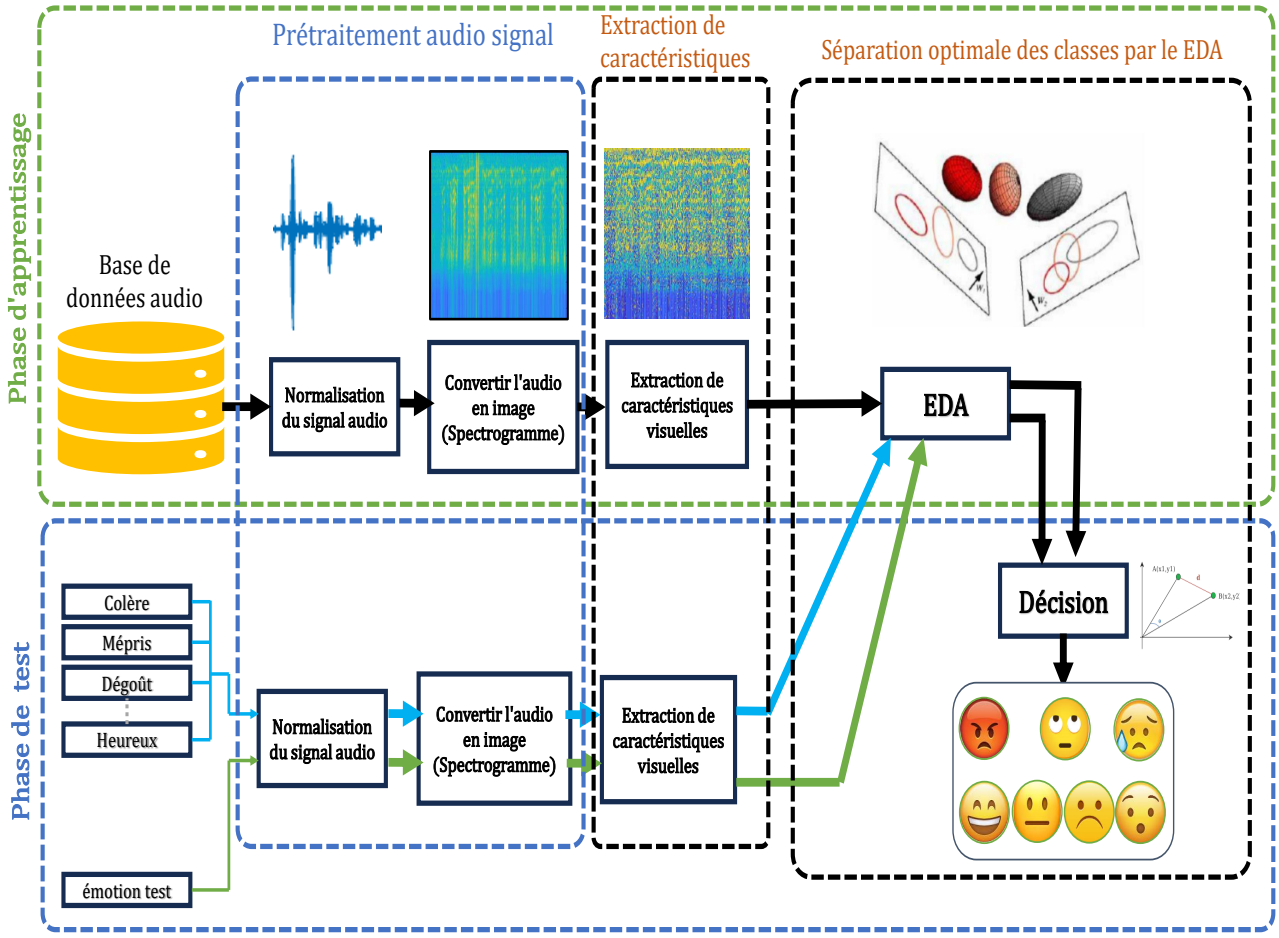

FIGURE III.1 – Schéma proposé de la reconnaissance des émotions basée sur l'audio.

et en mettant l'amplitude à l'échelle de l'intervalle [-1, 1], comme indiqué dans l'équation.III.1 .

$$\mathrm{X} = \frac{\mathrm{X} - \mathrm{mean(X)}}{\mathrm{max(abs(X))}} \tag{III.1}$$

Une fois le signal audio normalisé, il est transformé en représentations d'images à l'aide de deux méthodes distinctes : le spectrogramme à transformée de Fourier à court terme et le spectrogramme Mel.

III.2.1.1.1 STFT Spectrogramme : Le spectrogramme $X(k,t)$ est produit en soumettant le signal d'entrée à la Transformée de Fourier à Court Terme (STFT) à fenêtre [Boashash, 2015], comme défini dans l'Équation (III.2).

$$X(m,t) = \sum_{n=0}^{N-1} x[n]w[n-t]e^{-\frac{2\pi imn}{N}}, \quad m = 0, \ldots, N-1 \tag{III.2}$$

où $x[n]$ représente le signal vocal d'entrée, N désigne la longueur de la fenêtre, $w[n]$ fait référence à la fonction de fenêtre de Hamming, et m est l'indice de fréquence mesuré en hertz (Hz).

III.2.1.1.2 Mel Spectrogramme : Le spectre Mel comprend la transformée de Fourier à court terme (STFT) appliquée à chaque trame, transformant le spectre d'énergie/amplitude d'un échelle de fréquence linéaire à un échelle logarithmique Mel. Ensuite, il passe par un banque de filtres pour obtenir les vecteurs propres. Ces vecteurs propres peuvent être approximés comme la distribution de l'énergie du signal à travers les bandes de fréquence de l'échelle Mel [Shen et al., 2018] . Par conséquent, pour chaque tonalité avec une fréquence réelle f, mesurée en Hertz (Hz), une hauteur subjective est quantifiée sur une échelle connue sous le nom d'échelle Mel (Équation III.3).

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700\text{Hz}}\right) \tag{III.3}$$

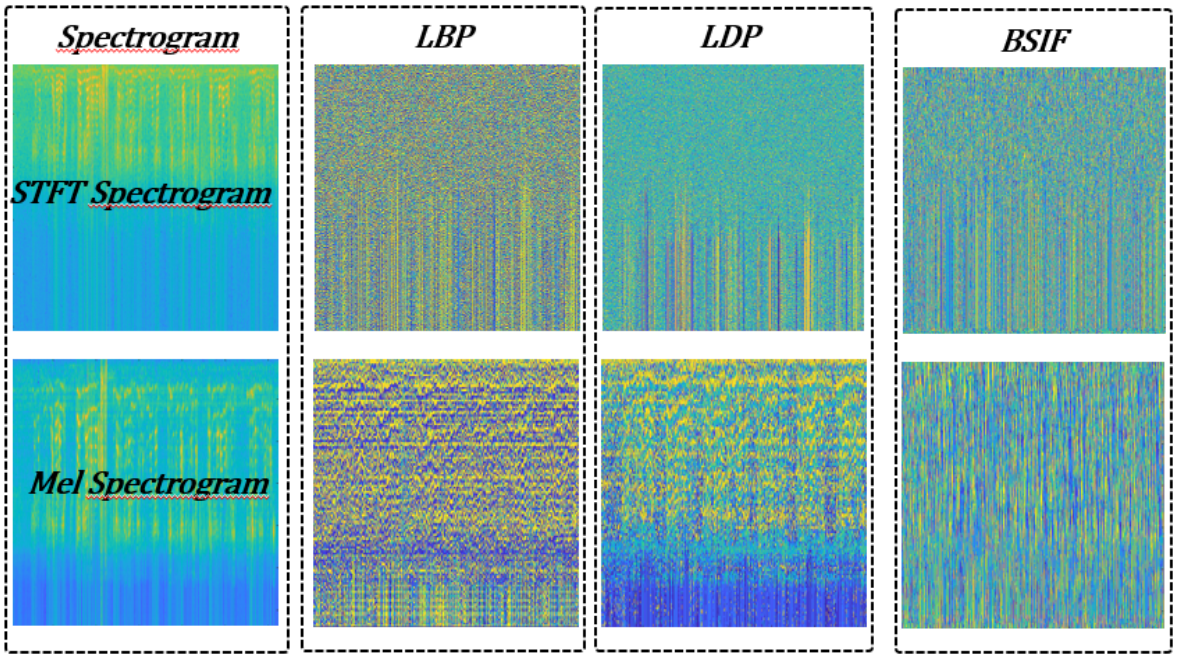

FIGURE III.2 – Visualisation comparative des caractéristiques du signal audio à l'aide de différentes méthodes de spectrogramme et des résultats de techniques de descripteurs visuels de texture.

III.2.1.2 Extraction de caractéristiques acoustique

L'importance de l'extraction des caractéristiques visuelles à partir des images est largement reconnue dans diverses tâches de reconnaissance. Cette extraction joue un rôle crucial en tant qu'étape initiale dans les systèmes de reconnaissance et les applications de vision par ordinateur, en particulier celles qui se concentrent sur l'analyse audio. Dans la classification des images, la qualité de l'encodage représentationnel influence de manière critique l'efficacité de la méthode. Ces encodages peuvent inclure des détails texturaux locaux ou des caractéristiques apprises. Des

études récentes ont démontré que les caractéristiques visuelles surpassent les caractéristiques audio, telles que les coefficients cepstraux prédictifs linéaires (LPCC) [Hermansky, 1990] et la prédiction linéaire perceptuelle (PLP) [O'Shaughnessy, 1988], dans les tâches de reconnaissance basées sur l'audio. Cette méthodologie s'est révélée efficace dans divers domaines, notamment la reconnaissance de locuteurs, la classification des scènes acoustiques, la détection des attaques par usurpation d'identité et la classification des sons d'oiseaux, entre autres.

Nous utilisons des approches pour extraire des caractéristiques efficaces et discriminatives à partir des images de spectrogrammes. Celles-ci incluent les motifs binaires locaux (LBP), les caractéristiques statistiques binarisées des images (BSIF) et les motifs directionnels locaux (LDP), en plus du fameux audio descripteur Coefficients cepstraux en fréquence Mel (MFCC) avec 12 parameter et modèle acoustique préentraîné VGGish basé sur un CNN (Convolutional Neural Network). La Figure III.2 présente une visualisation comparative des caractéristiques des signaux audio utilisant diverses méthodes de spectrogramme, ainsi que les résultats des techniques de descripteurs de texture visuelle. Nous détaillons chacun de ces descripteurs au chapitre II, section II.4.2.

III.2.2 Système visuel pour la reconnaissance des émotions

La Figure III.3 présente notre système de reconnaissance des émotions à partir de l'image de visage. Les principales étapes comprennent : d'abord, le prétraitement de l'image faciale, puis l'extraction des caractéristiques visuelles et profondes. Chaque étape est expliquée en détail dans les sous-sections qui suivent :

III.2.2.1 Pretraitement de visage

Le Multiscale Retinex (MSR) [Jiang et al., 2015] a été utilisé comme méthode d'amélioration d'image afin d'accroître la plage dynamique des images numériques tout en maintenant leur précision des couleurs. La renommée de l'algorithme MSR réside dans sa capacité à extraire des caractéristiques insensibles à l'éclairage pour les images de visages dans différentes conditions d'éclairage, ainsi que dans sa capacité à réduire les effets du bruit visuel. Par la suite, la méthode de Viola et Jones [Viola and Jones, 2001] a été employée pour détecter la zone du visage. La méthode de Viola et Jones est reconnue pour son efficacité et sa rapidité dans la détection des visages. (voir les détails de cette technique au chapitre II, sous-section II.4.1.2).

III.2.2.2 Extraction des caractéristiques visuels

Pour l'extraction des caractéristiques à partir d'images faciales, nous utilisons deux méthodes déférent telles que :

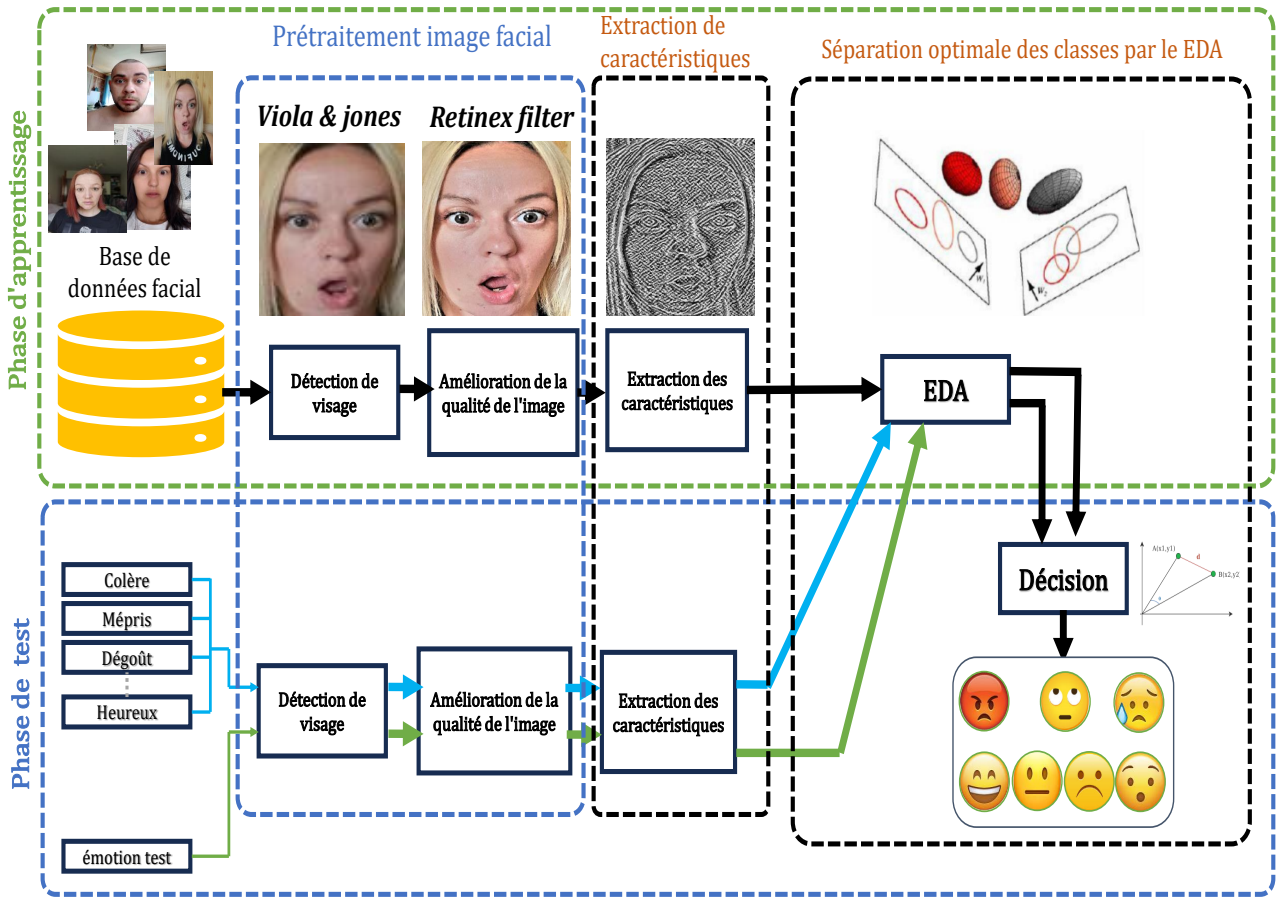

FIGURE III.3 – Schéma proposé de la reconnaissance des émotions basée sur les images faciales.

Caractéristiques profondes

Pour les caractéristiques profondes, nous procédons à leur extraction à partir de l'image faciale originale, qui présente une dimension de 224x224x3. Dans cette opération, nous exploitons trois couches du réseau VGG-19, à savoir fc6, fc7 et fc8. De plus, nous intégrons une couche issue du ResNet101, désignée sous le nom de fc1000, pour enrichir notre processus d'extraction de caractéristiques faciales.

Caractéristiques handcrafted

Afin de déterminer les caractéristiques handcrafted, nous employons les descripteurs LBP, BSIF et LDP sur l'image faciale, qui est ensuite séparée en 16 blocs. On combine chaque bloc dans un histogramme avec 256 intervalles, puis on les combine pour créer un vecteur de caractéristiques de taille (16 bloc$\times$(1$\times$256)).

III.2.3 Projection et Classification dans l'éspace par EDA

Après avoir extrait les caractéristiques pour le système basé sur l'audio ainsi que pour le système basé sur l'image faciale, nous utilisons la méthode Exponential Discriminant Analysis (EDA) pour la réduction de la dimensionnalité, la projection et la classification des caractéristiques. L'EDA est une technique supervisée puissante introduite par Ouamane [Ouamane et al., 2014], qui constitue

une extension exponentielle de la Linear Discriminant Analysis (LDA) [Xanthopoulos et al., 2013]. Cette méthode vise à minimiser la variabilité intra-classe et à maximiser la variabilité inter-classe. Comme le montre l'équation II.23, qui représente l'équation de la LDA, nous ajoutons une fonction exponentielle des deux côtés pour obtenir la méthode EDA, illustrée dans l'équation III.4. Cet ajout a montré des résultats remarquables dans de nombreuses études [Adil et al., 2016, Ouamane et al., 2014].

$$exp(S_b)v = exp(S_w)\lambda \tag{III.4}$$

Dans ces équations, S_b et S_w représentent respectivement la matrice de variabilité inter-classe et la matrice de variabilité intra-classe.

III.2.3.1 Comparaison avec CSS

À l'étape suivante de notre processus, après la réduction de la dimensionnalité, la projection dans l'espace pour la séparation optimale des classes, les données vectorielles sont soumises à une procédure de correspondance. Celle-ci utilise la métrique de distance cosinus CSS (Cosine Scoring Similarity) [Dehak et al., 2010] au sein d'un sous-espace discriminant, comme indiqué dans l'Équation III.5.

$$\mathrm{CSS}(A, B) = \frac{A \cdot B}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{III.5}$$

où A et B sont des vecteurs, $A \cdot B$ est le produit scalaire des vecteurs A et B, et $\|\mathbf{A}\|$ et $\|\mathbf{B}\|$ sont les magnitudes des vecteurs A et B respectivement.

Cette métrique est apte à comparer les vecteurs de caractéristiques et elle met en lumière l'amélioration discriminative facilitée par la synergie de l'EDA.

III.2.4 Fusion Pondérée Audio-Visuelle

Dans le cadre de notre système de reconnaissance des émotions audio-visuelles, nous employons la technique de fusion par somme pondérée (WS) [Matin et al., 2017], pour amalgamer les systèmes audio et visuels (voir la Figure III.4). Choisie pour son efficacité prouvée à renforcer les performances du système, la méthode WS amalgame judicieusement les deux types de caractéristiques. Cette stratégie tire parti des avantages distincts inhérents aux systèmes audio et visuels, améliorant ainsi de manière significative la précision de notre système de diagnostic. La formule de la fusion WS est présentée comme suit.

$$WS = \sum_{i=1}^{N} w_i \cdot x_i \tag{III.6}$$

où : WS désigne la somme pondérée agrégée, N signifie le score total, w_i représente le poids attribué à la i-ème entrée et x_i correspond à la valeur ou mesure prise de la i-ème entrée.

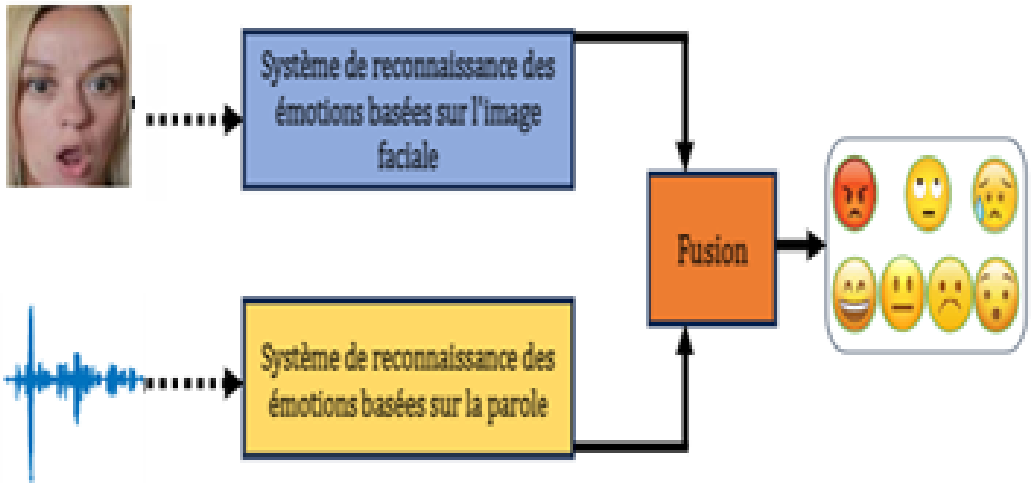

FIGURE III.4 – Schéma proposé de la reconnaissance des émotions Audio-Visuels.

III.3 Conclusion

Ce chapitre a donné une vision claire de notre travail en montrant notre contribution à travers l'aspect conceptuel de notre système et les étapes établies pour concrétiser ce dernier. Nous avons décrit notre solution proposée en expliquant précisément les architectures que nous avons utilisées pour construire notre système final.

Le prochain chapitre présentera les tests effectués, les résultats obtenus ainsi que leurs interprétations.

Chapitre IV

Résultats et Discussions

IV.1 Introduction

Les chercheurs en vision par ordinateur du monde entier cherchent constamment à optimiser la performance des systèmes de reconnaissance des émotions en testant et développant différentes approches et méthodes.

Dans ce dernier chapitre, nous allons mettre en œuvre notre système de reconnaissance des émotions en utilisant deux méthodes différentes : l'une basée sur la parole et l'autre sur l'image faciale. Cela repose sur divers descripteurs audio et visuels ainsi que sur la conversion du signal de parole en image spectrogramme. Afin d'obtenir les meilleures performances pour un système de reconnaissance des émotions, nous mènerons une série d'expériences sur deux bases de données d'émotions (audio et faciale) avec divers réglages et paramètres. Nous utiliserons une nouvelle technique de réduction de dimensionnalité appelée Exponential Discriminant Analysis (EDA). Enfin, nous procéderons à la fusion WS pour combiner les meilleurs scores de chaque type audio et visuel.

IV.2 Environnement de travail

Le langage de programmation utilisé dans ce travail est MATLAB, émulé par l'environnement de programmation du même nom (dans notre cas MATLAB 2023a) et développé par The MathWorks. MATLAB permet la mise en œuvre simple et rapide d'algorithmes, la réalisation de tâches nécessitant une puissance de calcul élevée, la manipulation et l'affichage de courbes, ainsi que la création d'interfaces graphiques. Les expériences ont été réalisées sur un PC équipé d'un processeur Intel(R) Core(TM) i7-1165G7 de 11e génération à 2,80 GHz avec 16 Go de RAM. Nous avons utilisé l'environnement Deep Learning Toolbox Model [MathWorks, 2024] pour les réseaux de neurones convolutifs (CNN).

IV.3 Base de données

Trois bases de données distinctes ont été utilisées. La première a été employée pour détecter les émotions en analysant les expressions faciales sur les images. La deuxième a permis d'identifier les émotions à partir de l'analyse vocale. La troisième base de données audio-visuelle a été utilisée pour combiner l'analyse des deux modalités précédentes

IV.3.1 Base de données de visages

La base de données se compose d'images capturant des personnes exprimant huit émotions distinctes : colère, mépris, dégoût, peur, heureuse, joie, tristesse et

surprise (voir la figure IV.1 des exemples d'images de la base de données TAPAKAH68 avec 8 émotions déférentes). Chaque image représente une de ces émotions spécifiques, permettant aux chercheurs et aux praticiens de l'apprentissage automatique d'étudier et de développer des modèles pour la reconnaissance et l'analyse des émotions. Le lien pour télécharger la base de données est le suivant : https ://www.kaggle.com/datasets/tapakah68/facial-emotion-recognition/data.

FIGURE IV.1 – Exemples d'images extraites des ensembles de données de visage et 8 classes d'émotions déférentes.

IV.3.2 Base de données d'audio

La base de données complète de parole et de chanson, en audio et vidéo [Livingstone and Russo, 2018]. La construction et la validation perceptive du RAVDESS sont décrites dans un article en libre accès sur PLoS ONE. Cette partie du RAVDESS comprend 1440 fichiers, avec 60 essais par acteur, totalisant 24 acteurs professionnels (12 femmes et 12 hommes). Chaque acteur vocalise deux déclarations lexiquement appariées dans un accent nord-américain neutre, couvrant une gamme d'émotions telles que le calme, la joie, la tristesse, la colère, la peur, la surprise et le dégoût, chacune produite à deux niveaux d'intensité émotionnelle (normale et forte), avec une expression neutre supplémentaire. Les fichiers sont nommés selon une convention spécifique, identifiant la modalité (audio ou vidéo), le canal vocal (parole ou chanson), l'émotion, l'intensité émotionnelle, la

déclaration, la répétition et l'acteur. Pour télécharger cette base de données : https ://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

IV.3.3 La base de données Audio-Visuels (IEMOCAP)

est une base de données agie, multimodale et multilocuteur, récemment collectée au laboratoire SAIL de l'USC. Il contient environ 12 heures de données audio-visuelles, dont des vidéos, des paroles, des captures de mouvements de visages et des transcriptions de textes. Il s'agit de séances dyadiques au cours desquelles des acteurs exécutent des improvisations ou des scénarios scénarisés, spécifiquement sélectionnés pour susciter des expressions émotionnelles. La base de données IEMOCAP est annotée par plusieurs annotateurs en étiquettes catégorielles, telles que colère, bonheur, tristesse, neutralité, ainsi qu'en étiquettes dimensionnelles telles que valence, activation et dominance. Les informations détaillées de capture de mouvement, le cadre interactif permettant de susciter des émotions authentiques et la taille de la base de données font de ce corpus un ajout précieux aux bases de données existantes dans la communauté pour l'étude et la modélisation de la communication humaine multimodale et expressive [Busso et al., 2008].

(pour télécharger cette base de donnée : https ://sail.usc.edu/iemocap/

IV.4 Protocole de travail

La base de données est divisée en deux parties distinctes : l'une utilisée pour l'apprentissage des matrices de projection EDA et l'autre pour les tests. 50 % des données sont utilisées pour l'apprentissage et les 50 % restants pour les tests [Saadi et al., 2023]. Pour la base de données faciale, celle-ci contient 8 classes d'émotions, chaque classe comprenant 18 images. Afin de poser un scénario de test rigoureux, le modèle est construit en utilisant une image d'une personne, tandis que les tests sont effectués avec les huit images restantes provenant d'autres personnes. Le même protocole a été utilisé pour le second jeu de données audio, qui contient également 8 classes d'émotions.

IV.5 Expérimentations et Résultats

Dans toutes nos expériences, nous évaluons le système de reconnaissance des émotions en calculant le taux de reconnaissance correcte. Ce taux est déterminé par le rapport entre le nombre de tests corrects et le nombre total de tests effectués. Nous explorons divers scénarios en utilisant deux techniques : la conversion de l'audio en spectrogrammes d'images (STFT-Spectrogramme et Mel-Spectrogramme), en employant cinq caractéristiques de texture visuelle (LBP, LDP, ResNet101 et

TABLE IV.1 – Le taux de reconnaissance (%) du système d'émotion basé sur l'image faciale.

Méthode	**Caractéristique**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
Handcrafted	LBP	74.17	74.17	76.67	78.33	78.33	78.33	78.33
	BSIF	65.00	72.50	76.67	78.33	75.00	75.83	75.83
	LDP	81.67	82.50	88.33	90.00	90.00	90.83	**91.67**
Deep	VGG19	**66.67**	63.33	59.17	61.67	56.67	60.00	58.33
	ResNet101	53.33	60.83	54.17	55.83	55.00	61.67	59.17

VGG19), et en appliquant des méthodes de réduction et de projection dans les sous-espaces, y compris l'EDA. Les résultats obtenus sont présentés dans les Tables IV.1, IV.2 et IV.3 pour les ensembles de données audio, images faciales et fusion entre les deux méthodes, respectivement. Nous menons plusieurs expériences pour évaluer la méthode proposée :

a. Pour les caractéristiques handcrafted Visuels, nous utilisons 4 descripteurs (LBP, LDP, BSIF et LPQ) etle caractéristiques handcrafted audio 1 descripteurs MFCC :
   1. Le nombre de taille de filter pour BSIF est sélectionné comme suit : $W = 3, 5, 7, 9, 11$.
   2. Pour le descripteur audio nous utilisons MFCC avec 12 paramètres.

b. Pour les caractéristiques profondes, nous utilisons troi modèles pré-entraînés (VGG19,VGGish et ResNet101) :
   1. Dans VGG19, seules 3 échelles de caractéristiques sont extraites ($fc6$, $fc7$ et $fc8$).
   2. Dans VGGish, seules 2 échelles de caractéristiques sont extraites ($fc1\_1$ et $fc1\_2$).
   3. Dans ResNet101, il y a 1 échelle de caractéristiques à extraire et à entraîner avec ($fc1000$).

c. Les réglages optimaux pour les caractéristiques audio et visuelles sont combinés en utilisant la méthode de Fusion de Somme Pondérée (WSF).

TABLE IV.2 – Le taux de reconnaissance (%) du système d'émotion basé sur la parole.

Méthode	Caractéristique	10	15	20	25	30	35	40
Handcrafted	MFCC	85.13	85.13	**86.80**	86.03	84.78	84.02	85.00
	S-LDP	55.14	55.78	56.88	56.88	56.88	56.88	56.02
	Mel-LDP	60.12	62.14	62.14	62.14	62.14	62.44	62.44
	S-LBP	52.01	52.90	52.43	52.43	52.43	52.43	52.43
	Mel-LBP	59.09	59.66	59.66	59.66	59.66	57.89	57.43
	S-BSIF	80.45	80.76	80.34	80.65	80.65	80.65	80.65
	Mel-BSIF	83.51	83.44	84.00	84.56	83.89	83.89	83.89
Deep	Mel-VGGish	**91.14**	90.30	89.80	89.34	89.78	88.32	88.32
	S-VGGish	89.56	88.79	86.20	86.20	86.20	86.20	86.20

TABLE IV.3 – Le taux de reconnaissance (%) avec la fusion des meilleurs résultats du système audio et visuel par WS Fusion .

Méthode	Visuels LDP	Audio MelS-VGGish	Taux de Reconnaissance (%)
WS Fusion	0	1	91.14
	0.1	0.9	91.80
	0.2	0.8	92.07
	0.3	0.7	93.17
	0.4	0.6	93.17
	0.5	0.5	93.34
	0.6	0.4	**93.89**
	0.7	0.3	92.87
	0.8	0.2	92.45
	0.9	0.1	91.67
	1	0	91.67

IV.6 Discussion

IV.6.1 L'avantage des caracteristiques handcrafted dans notre système

L'avantage des caractéristiques manuelles dans notre système réside dans leur capacité à extraire des traits spécifiques et discriminants des images, ce qui peut conduire à des performances élevées même avec des modèles moins complexes. Par exemple, les méthodes manuelles LDP, LBP et BSIF ont toutes surpassé les résultats obtenus avec les méthodes de deep learning VGG19 et ResNet101, comme le montre le tableau IV.1, démontrant ainsi l'efficacité de ces techniques pour la reconnaissance des émotions. Les caractéristiques manuelles offrent également une meilleure interprétabilité et peuvent être plus robustes face à certaines variations des données.

IV.6.2 Deep Vs. handcrafetd descripteur

Nous avons utilisé deux méthodes pour la reconnaissance des émotions faciales : la méthode handcrafted, comprenant LBP, LDP, BSIF, et la méthode de deep learning, utilisant les modèles VGG19 et ResNet101 (voir tableu IV.1). Pour la méthode handcrafted, LBP a atteint un taux de reconnaissance de 78,33%, LDP a obtenu 91,67%, et BSIF a également atteint 78,33 %. En ce qui concerne la méthode de deep learning, VGG19 a donné un taux de reconnaissance de 66,67 %, tandis que ResNet101 a obtenu 61,67 %. Malgré les performances de la méthode de deep learning, nous avons obtenu de meilleurs résultats avec la méthode LDP, surpassant VGG19 de 25%.

Pour extraire les caractéristiques audio, nous avons converti le signal vocal en utilisant deux méthodes de spectrogramme différentes : la STFT-S et Mel-S. Ensuite, nous avons appliqué les descripteurs LBP, LDP, BSIF et MFCC sur ces deux méthodes de spectrogramme (voir tableau IV.2). Nous avons un meilleur taux de reconnaissance de 86.80% pour le MFCC avec la méthode handcrafted. En ce qui concerne la méthode deep, le meilleur taux de reconnaissance a été atteint avec Mel-VGGish 91.14%, dépassant ainsi le MFCC avec un taux de reconnaissance de 4.34%.

IV.6.3 L'effet du nombre d'EDA

La manipulation du paramètre N_{EDA} sur la plage de 10 à 45 a exercé une influence significative sur les performances de classification. À mesure que la valeur de N_{EDA} augmentait, il y avait une amélioration constante de la taux de reconnaissance. Cependant, il est crucial de noter qu'au-delà d'un seuil spécifique, des incréments supplémentaires de N_{EDA} n'ont pas conduit à des améliorations sub-

stantielles des performances. Dans la plupart des cas, les résultats les plus élevés ont été observés aux valeurs N_{EDA} de 25 et 45 sur les deux ensembles de données. Cela suggère la présence d'un point de saturation, au-delà duquel de nouveaux ajustements du paramètre N_{EDA} ne produisent pas d'avantages significatifs.

IV.6.4 Impact de la fusion WS

Le tableau IV.3 présente le taux de reconnaissance d'un système utilisant la méthode "WS Fusion" pour combiner des données visuelles (LDP) et audio (MelS-VGGish). Chaque ligne représente une combinaison différente, allant de poid max égale à 1 d'audio (91.14% de reconnaissance) à poid de 1 pour visuel (91.67%). Les performances s'améliorent généralement en introduisant plus de données visuelles, avec une légère augmentation à 0.1p LDP (91.80%), puis une hausse significative à 0.3p LDP (93.17%). Le taux reste stable à 0.4p LDP, augmente légèrement à 50/50 (93.34%), et atteint son maximum à 0.6p LDP et 0.4p audio (93.89%). Au-delà, trop de données visuelles font baisser les performances : 92.87% à 0.7p LDP, 92.45% à 0.8p LDP, et 91.67% à 0.9p LDP. Ces résultats montrent qu'une combinaison équilibrée, avec un léger avantage pour le visuel, offre les meilleures performances, soulignant l'importance de la fusion multimodale.

IV.6.5 Résultats sur IEMOCAP

À la suite de nos expérimentations menées sur la base de donnée audio RAVDESS en utilisant la méthode deep, nous avons obtenu un taux de reconnaissance optimal de 91,14% en appliquant le descripteur Mel-VGGish. Ce résultat a ensuite été appliqué à la base de données audio-visuel IEMOCAP, produisant un taux de reconnaissance de 63,19%. Pour l'extraction des caractéristiques faciales sur la base de données avec la méthode handcrafted, le descripteur LDP a été identifié comme le plus performant, avec un taux de reconnaissance de 91,67%. L'application de ce descripteur à la base de données IEMOCAP a donné un taux de reconnaissance de 60,14%. Après la fusion des caractéristiques audio-visuelles on employant le descripteur WSFusion sur la base de données IEMOCAP le résultat était 64.01% 5(voir le table IV.4).

IV.7 Conclusion

Dans ce dernier chapitre, nous avons exposé les résultats obtenus suite à la mise en œuvre de nos architectures présentées dans le chapitre précédent (chapitre III). Tout d'abord, nous avons détaillé la mise en œuvre de notre travail, en présentant notre environnement de développement sur lequel le système a été réalisé. Ensuite,

TABLE IV.4 – Le taux de reconnaissance (%) du système d'émotion Audio-Visuels sur la base de donnée IEMOCAP.

Méthode	**Caractéristique**	Taux de Reconnaissance(%)
Audio	Mel-VGGish	63.19
Visuels	LDP	60.34
Fusion	WSF	64.01

nous avons analysé les résultats de chaque expérience en calculant les métriques d'évaluation, facilitant ainsi la comparaison entre les différentes expérimentations.

À partir des résultats obtenus, nous avons démontré que la technique de fusion audio-visuelle permet de concevoir un système multimodal offrant des performances accrues.

Conclusion Générale

Parmi les modalité les plus utilisée dans la reconnaissance des émotions, on trouve : 'la reconnaissance de la parole' et 'la reconnaissance facial'.

L'objectif suivi dans ce mémoire propose une démarche qui consiste à améliorer la performance de la reconnaissance des émotions via la reconnaissance parole et faciale par plusieurs méthodes avec un ensemble de descripteurs et nouvelles techniques.

En premier lieu, dans le cadre de la reconnaissance vocale des émotions (RVE), nous avons entamé la phase de prétraitement en normalisant un signal audio puis en le convertissant en image (STFT-S et Mel-S). Après cette conversion en images, nous avons appliqué une méthode des descripteurs handcrafted (tels que LBP, LDP, BSIF et MFCC), tandis que nous avons utilisé la méthode deep avec VGGish. À travers diverses expériences où nous avons ajusté le nombre (EDA), nous avons réussi à obtenir un taux de reconnaissance de 91.14% avec Mel-VGGish pour la base de données RAVDESS.

En deuxième lieu, dans le cadre de la reconnaissance des émotions faciales (REF), nous avons initié la phase prétraitement en détectant le visage par Viola et Jones puis en améliorant la qualité d'image, une autre méthode d'extraction de cette dernière nous avons appliqué les descripteurs handcrafted (LBP, LDP, BSIF) par contre la méthode deep (VGG19 et ResNet101), nous avons atteint un taux de reconnaissance de 91.67% avec LDP pour la base de donnée TAPAKAH68.

Afin de bénéficier des avantages de deux descripteurs audio-visuel nous avons pris les meilleurs scores LDP 91.67% et Mel-VGGish 91.14% obtenus dans la base de donnés RAVDESS nous les avons appliqué sur une base de données audio-visuel IEMOCAP nous avons obtenu un taux de reconnaissance : Mel-VGGish 63.19% et LDP 60.34. Puis nous avons employé WSFusion qui nous a donnée un resultat 64.01%.

À l'issue de ce travail, nous estimons avoir réalisé un système répondant à l'ob-

jectif que nous nous sommes fixés. Ainsi, l'utilisation des descripteurs audio-visuels pour le système de reconnaissance automatique des émotions permet d'avoir une meilleure robustesse en améliorant les performances de ce système.

Perspectives

Les travaux menés dans le cadre de ce projet représentent un bon début pour plusieurs autres expérimentations futures qui doivent être poursuivi pour parvenir à un système encore plus robuste.

À partir de ce projet, nous prévoyons de continuer nos recherches sur la reconnaissance automatique des émotions, ainsi que sur d'autres projets exploitant la technologie de reconnaissance automatique basée sur la voix et les caractéristiques faciales. Nous nous engageons à améliorer ce système afin de résoudre les véritables défis environnementaux et de garantir un fonctionnement robuste et efficace dans toutes les situations, même les plus critiques.

Bibliographie

M Adil, Muhammad Abid, Abdul Qayyum Khan, Ghulam Mustafa, and Nasir Ahmed. Exponential discriminant analysis for fault diagnosis. *Neurocomputing*, 171 :1344–1353, 2016.

Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17 :200171, 2023.

Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Computer Vision-ECCV 2004 : 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pages 469–481. Springer, 2004.

Ali Altaher, Zahra Salekshahrezaee, Azadeh Abdollah Zadeh, Hoda Rafieipour, and Ahmed Altaher. Using multi-inception cnn for face emotion recognition. *Journal of Bioengineering Research*, 3(1) :1–12, 2020.

Eliathamby Ambikairajah et al. Pncc-ivector-src based speaker verification. *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–7, 2012.

R Aparna and PL Chithra. Role of windowing techniques in speech signal processing for enhanced signal cryptography. *Advanced Engineering Research and Applications*, 5 :446–458, 2017.

Afizan Azman, Kirbana Jai Raman, Imran Artwel Junior Mhlanga, Siti Zainab Ibrahim, Sumendra Yogarayan, Mohd Fikri Azli Abdullah, Siti Fatimah Abdul Razak, Anang Hudaya Muhamad Amin, and Kalaiarasi Sonai Muthu. Real time driver anger detection. In *Information Science and Applications 2018 : ICISA 2018*, pages 157–167. Springer, 2019.

M. Benatia and H. Ouamane. *Identification de reconnaissance faciale avec des expressions*. PhD thesis, Université Mohamed Khider–Biskra, 2012.

Boualem Boashash. *Time-frequency signal analysis and processing : a comprehensive reference*. Academic press, 2015.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42 :335–359, 2008.

Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Emotion assessment for human-computer interaction applications. *IEEE Transactions on Cybernetics*, 41(3) :636–643, 2011.

Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE, 2018.

Luefeng Chen, Min Wu, Witold Pedrycz, and Kaoru Hirota. *Emotion recognition and understanding for emotional human-robot interaction systems*, volume 926. Springer Nature, 2020.

Jonathan Cowie, Karen S Douglas, Anne Wichmann, and Dirk KJ Heylen. Réseau d'interactions dans les dialogues explicatifs : un corpus d'anglais parlé. *Anglais parlé Corpus*, 2001a.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1) :32–80, 2001b.

Douglas W Cunningham, Mario Kleiner, Christian Wallraven, and Heinrich H Bülthoff. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception (TAP)*, 2 (3) :251–269, 2005.

Feriel Debbeche. Système acoustico-anatomique pour l'identification des locuteurs par localisation dans un espace de locuteurs de référence. Mémoire de maîtrise, Université Badji Mokhtar Annaba, 2008.

Najim Dehak, Reda Dehak, James R Glass, Douglas A Reynolds, Patrick Kenny, et al. Cosine similarity scoring without score normalization techniques. In *Odyssey*, volume 15, 2010.

Christian Derbaix and Michel T. Pham. Pour un développement des mesures de l'affectif en marketing : synthèse des prérequis. *Recherche et Applications en Marketing (French Edition)*, 4(4) :71–87, 1989.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4) : 169–200, 1992.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Ran Fang, Kevin D. Tang, Noah Snavely, and Tsuhan Chen. Towards computational models of kinship verification. In *2010 IEEE International Conference on Image Processing*, pages 1577–1580. IEEE, 2010.

S. Gharsalli. *Reconnaissance des émotions par traitement d'images*. PhD thesis, Université d'Orléans, 2016.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning : A report on three machine learning contests. In *Neural Information Processing : 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

Denis Gračanin, Vincenzo Emanuele Vinzi, Mauro Adenzato, and Antonio Rizzi. Emotion recognition for enhancing situational awareness in video surveillance. *Electronics*, 10(14) :1719, 2021.

James J. Gross. Emotion regulation : Current status and future prospects. *Psychological Inquiry*, 26(1) :1–26, 2015.

Sanaul Haq and Philip JB Jackson. Multimodal emotion recognition. In *Machine audition : principles, algorithms and systems*, pages 398–423. IGI global, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4) :1738–1752, 1990.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

Xin Hua, Qiang Tong, Xuelei Wei, and Xuetao Li. Facial expression recognition with boosted weighted sum score. In *ICMR*, 2016.

Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

Taskeed Jabid, Md Hasanul Kabir, and Oksam Chae. Local directional pattern (ldp)–a robust image descriptor for object recognition. In *2010 7th IEEE international conference on advanced video and signal based surveillance*, pages 482–487. IEEE, 2010.

Vikramaditya Jakkula. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5) :3, 2006.

Mira Jeong and Byoung Chul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12) :4270, 2018.

Xunsheng Ji, Kun Jiang, and Jie Xie. Lbp-based bird sound classification using improved feature selection algorithm. *International Journal of Speech Technology*, 24 :1033–1045, 2021.

Bo Jiang, Glenn A Woodell, and Daniel J Jobson. Novel multi-scale retinex with color restoration on graphics processing unit. *Journal of Real-Time Image Processing*, 10 :239–253, 2015.

Jade Vande Kamp. What is a spectrogram ? URL `https://vibrationresearch.com/blog/what-is-a-spectrogram/`.

J. Kannala and E. Rahut. Basif : Binarized statistical image features. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 1363–1366. IEEE, November 2012.

Mohammed Khammari, Ammar Chouchane, Abdelmalik Ouamane, Mohcene Bessaoudi, Yassine Himeur, Mahmoud Hassaballah, et al. High-order knowledge-based discriminant features for kinship verification. *Pattern Recognition Letters*, 175 :30–37, 2023.

Aldebaro Klautau. The mfcc. *Digital Signal Processing*, 2005.

Naman Kohli. *Automatic kinship verification in unconstrained faces using deep learning*. PhD thesis, West Virginia University, 2019.

Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech : a review. *International journal of speech technology*, 15 :99–117, 2012.

Yann LeCun, Lawrence D. Jackel, Bernhard Boser, John S. Denker, Hans P. Graf, Isabelle Guyon, and Wayne Hubbard. Handwritten digit recognition : Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27(11) :41–46, 1989.

Baoli Li and Liping Han. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013 : 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*, pages 611–618. Springer, 2013.

Lu Liu, Chen Xiong, Hongyu Zhang, Zhiqiang Niu, Majun Wang, and Shuicheng Yan. Deep aging face verification with large gaps. *IEEE Transactions on Multimedia*, 18(1) :64–75, 2015.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface : Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess) : A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5) :e0196391, 2018.

J. Lu, J. Hu, X. Zhou, J. Zhou, M. Castrillo-Santana, J. Lorenzo-Navarro, ..., and T. F. Vieira. In ieee international joint conference on biometrics. pages 1–6. IEEE, September 2014.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M Montero, and Fernando Fernández-Martínez. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22) :7665, 2021.

Daniel Lundqvist, Anders Flykt, and Arne Öhman. Karolinska directed emotional faces. *PsycTESTS Dataset*, 91 :630, 1998.

Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality : emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific reports*, 8(1) :13657, 2018.

MathWorks. Deep learning toolbox : Design, train, and analyze deep learning networks, 2024. URL `https://www.mathworks.com/products/deep-learning.html?s_tid=FX_PR_info`. Accessed : 2024-05-30.

Abdul Matin, Firoz Mahmud, Tanvir Ahmed, and Md Sabbir Ejaz. Weighted score level fusion of iris and face to identify an individual. In *2017 international conference on electrical, computer and communication engineering (ECCE)*, pages 1–4. IEEE, 2017.

Steve Nash, Michael Rhodes, and Joanna I Olszewaska. Ifr : Interactively pose corrected face recognition. In *International Conference on Bio-inspired Systems and Signal Processing*, volume 5, pages 106–112. SCITEPRESS, February 2016.

Bashar M Nema and Ahmed A Abdul-Kareem. Preprocessing signal for speech emotion recognition. *Al-Mustansiriyah Journal of Science*, 28(3) :157–165, 2018.

Keith Oatley and Philip N Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1) :29–50, 1987.

V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In *Image and Signal Processing : 3rd International Conference, ICISP 2008*, pages 236–243, Cherbourg-Octeville, France, July 1-3, 2008. Springer Berlin Heidelberg.

Douglas O'Shaughnessy. Linear predictive coding. *IEEE potentials*, 7(1) :29–32, 1988.

Abdelmalik Ouamane, Bengherabi Messaoud, Abderrezak Guessoum, Abdenour Hadid, and Mohamed Cheriet. Multi scale multi descriptor local binary features and exponential discriminant analysis for robust face authentication. In *2014 IEEE International conference on image processing (ICIP)*, pages 313–317. IEEE, 2014.

D. Panić et al. Facial expression recognition with weighted sum score fusion. In *MIPRO*, 2019.

Mrinalini Patil and S Veni. Driver emotion recognition for enhancement of human machine interface in vehicles. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0420–0424. IEEE, 2019.

S. Pigeon, P. Druyts, and P. Verlinde. Applying logistic regression to the fusion of the nist'99 1-speaker submissions. *Digital Signal Processing*, 10(1-3) :237–248, 2000.

Robert Plutchik. Emotions : A general psychoevolutionary theory. *Approaches to emotion*, 1984 :197–219, 1984.

Nicolae-Câtâlin Ristea, Liviu Cristian Du**t**u, and Anamaria Radoi. Emotion recognition system from speech and visual information based on convolutional neural networks. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE, 2019.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6) :1161–1178, 1980.

Ibtissam Saadi, Taleb-Ahmed Abdelmalik, Abdenour Hadid, Yassin El Hillali, et al. Driver's facial expression recognition : A comprehensive survey. *Expert Systems with Applications*, page 122784, 2023.

Soyuj Kumar Sahoo, Tarun Choubisa, and SR Mahadeva Prasanna. Multimodal biometric person authentication : A review. *IETE Technical Review*, 29(1) : 54–75, 2012.

Alaa Ehab Sakran, Sherif Mahdy Abdou, Salah Eldeen Hamid, and Mohsen Rashwan. A review : Automatic speech segmentation. *International Journal of Computer Science and Mobile Computing*, 6(4) :308–315, 2017.

C. Sanderson and K. K. Paliwal. Information fusion and person verification using speech and face information. Technical Report IDIAP-RR, IDIAP Research Institute, 2002.

Conrad Sanderson and Kuldip Paliwal. Information fusion and person verification using speech face information. Technical Report IDIAP-RR 02-33, IDIAP, 2004.

Klaus R. Scherer. What are emotions ? *And What Should They Be ?*, 9(3) : 317–337, 2005.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet : A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

Nicu Sebe, Ira Cohen, and Thomas S Huang. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*, pages 387–409. World Scientific, 2005.

Philip Shaver, Julie Schwartz, Don Kirson, and Cheryl O'Connor. Emotion knowledge : Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6) :1061, 1987.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

S. Shum, N. Dehak, R. Dehak, and J.R. Glass. Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In *Odyssey*, page 16, June 2010.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

V.I. B. R. A. T. N. A. L. Spectroscopies et al. *Pattern Recognition, Analysis and Application*, volume i, page 13. Dr. Mahali, 2016.

S. Sudha and S. Suganya. On-road driver facial expression emotion recognition with parallel multi-verse optimizer (pmvo) and optical flow reconstruction for partial occlusion in internet of things (iot). *Measurement : Sensors*, 26 :Article 100711, 2023.

Thomas Teixeira. Reconnaissance multi-dimensionnelle de l'émotion par apprentissage profond de caractéristiques spatio-temporelles sur séquences vidéo. *10 SEPTEMBRE 2020*, 2020.

Y. L. Tian, T. C. Kanade, and Jeffrey F., editors. *Facial Expression Analysis*. Springer-Verlag, 2004.

A. S. Tolba, A. H. El-Baz, and A. A. El-Harby. Face recognition : A literature review. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(3) :393–420, 2015.

L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction : a comparative. *J Mach Learn Res*, 10(66-71) :13, 2009.

H. Varma, N. Ganapathy, and T. M. Deserno. Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning. In *Medical imaging 2022 : Imaging informatics for healthcare, research, and applications, Vol. 12037*, pages 57–63. SPIE, 2022.

Subhashini Venugopalan, Haizhou Xu, Xavier Domont, and Florian Metze. Audio-visual speech and speaker recognition with deep fenics. In *INTERSPEECH*, 2015.

Paul Viola. Jones,"robust real-time object detection,". In *IEEE Workshop on Statistical and Theories of Computer Vision*, 2001.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, pages I–I. IEEE, 2001.

Javaid Ahmad Wani, Sparsh Sharma, Malik Muzamil, Suhaib Ahmed, Surbhi Sharma, and Saurabh Singh. Machine learning and deep learning based computational techniques in automatic agricultural diseases detection : Methodologies, applications, and challenges. *Archives of Computational methods in Engineering*, 29(1) :641–677, 2022.

Yandong Wen, Kaipeng Ahang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision-ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 14, pages 499–515. Springer International Publishing, 2016.

Petros Xanthopoulos, Panos M Pardalos, Theodore B Trafalis, Petros Xanthopoulos, Panos M Pardalos, and Theodore B Trafalis. Linear discriminant analysis. *Robust data mining*, pages 27–33, 2013.

H. Xiao, W. Li, G. Zeng, Y. Wu, J. Xue, J. Zhang, and et al. On-road driver emotion recognition using facial expression. *Applied Sciences*, 12(2) :807, 2022.

Hui Yan, Jingbin Lu, Wangmeng Deng, and Xiang Zhou. Discriminative multimetric learning for kinship verification. *Information Forensics and Security, IEEE Transactions on*, 9(7) :1169–1178, 2014.

Jie Yu and Bao-Liang Zhang. Speech emotion recognition combining weighted sum scorefusion and diversity ensemble. In *ICALIP*, 2018.

K. Zaman, Z. Sun, S. M. Shah, M. Shoaib, L. Pei, and A. Hussain. Driver emotions recognition based on improved faster r-cnn and neural architectural search network. *Symmetry*, 14(4) :687, 2022.

Nannan Zeng, Hailin Xiao, Jie Dong, Yi Zhang, and Qing Wu. Facial expression recognition via weighting scored sub-region features. In *IAPR Asian Conference on Pattern Recognition*, 2017.

Annexe A

Interface graphique

Dans notre projet, nous avons développé une interface graphique visant à faciliter l'interaction entre l'utilisateur et le système de Reconnaissance des Emotions en :

1. Simplifiant la lecture et la compréhension des résultats grâce à une présentation optimisée.
2. Offrant à l'utilisateur une configuration simplifiée via une liste de choix prédéfinis.

Notre interface, conçue avec GUIDE dans Matlab 2024a, est intuitive et conviviale.

A.1 Fenêtre d'accueil

La fenêtre principale de notre application s'ouvre au démarrage , offrant à l'utilisateur un accès aux différentes fonctionnalités disponibles (figure A.1). Elle permet de choisir entre la configuration de l'extraction basée sur des descripteurs audio, des descripteurs visuels, ou une fusion audio-visuelle.

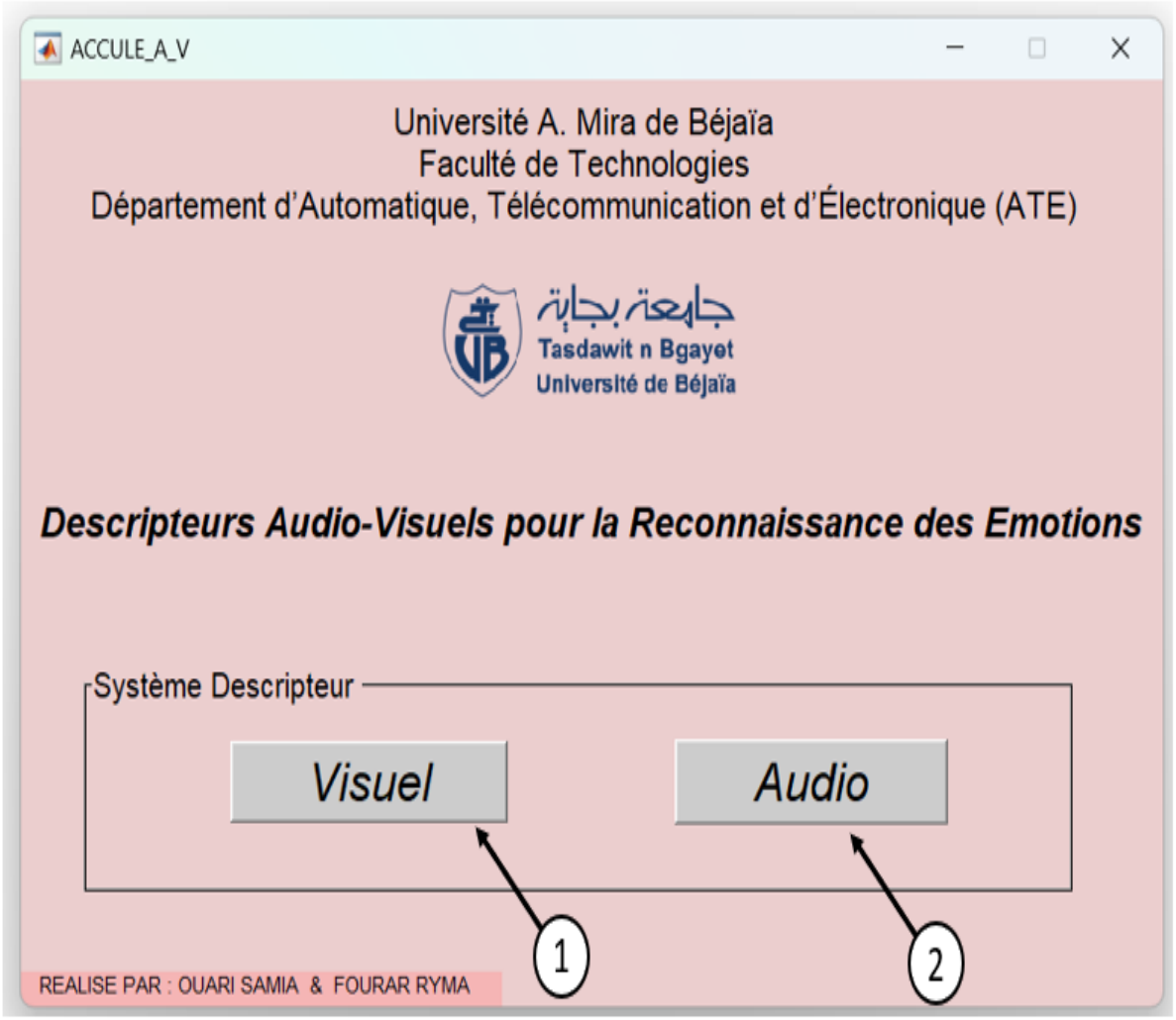

FIGURE A.1 – Fenêtre d'accueil.

1. systeme de reconnaissance d'emotion a partir l'image faciale.
2. systeme de reconnaissance d'emotion a partir la parole.

A.1.1 GUI pour le systeme de reconnaissance d'emotion a partir l'image faciale.

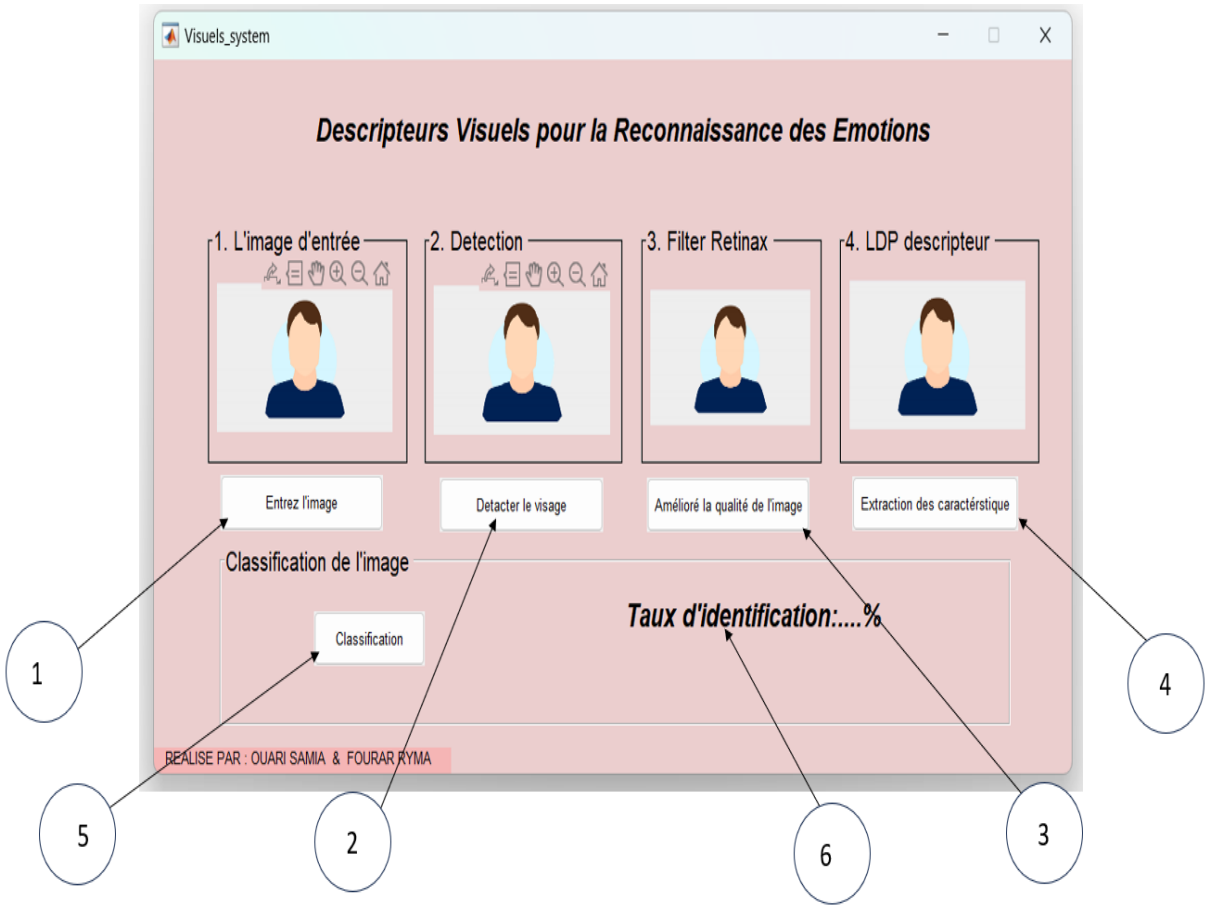

FIGURE A.2 – Système de reconnaissance d'émotion a partir de l'image faciale.

1. Insertion de l'image : Nous procéderons à l'intégration d'une image issue de la base de données.
2. Détection de visage :Nous emploierons l'algorithme de viola-jones.
3. Amélioration de la qualité d'image :Nous aurons recours au filtre retinex pour optimiser la qualité visuelle.
4. Extraction des caracteristiques :Nous utiliserons le descripteur LDP a fin d'extraire les attributs pertinents.
5. Classification :Nous appliquerons la methode de la distance cosinus pour cette etape.
6. Taux d'identification :En activant cette fonction ,nous initierons le systeme ,et le resultat s'affichera sous la forme d'un taux de reconnaissance des émotions(classe d'émotion).

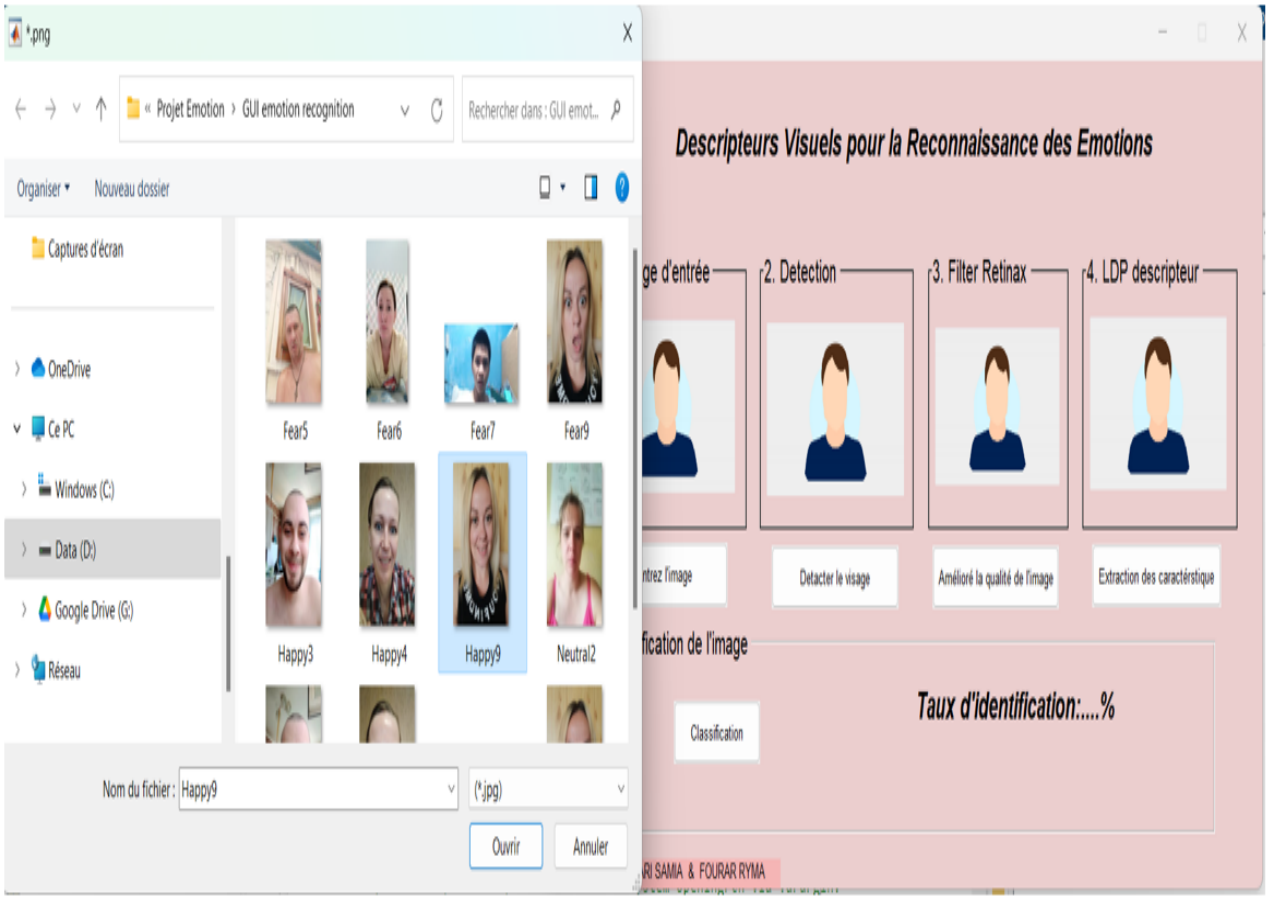

FIGURE A.3 – Système de reconnaissance d'émotion a partir de l'image faciale.

En cliquant sur le button "visuel", il nous permet de charger une image de visage depuis la base de données située dans le repertoire de la machine ,puis l'afficher dans (figure A.3).

A.1.2 Test d'identification

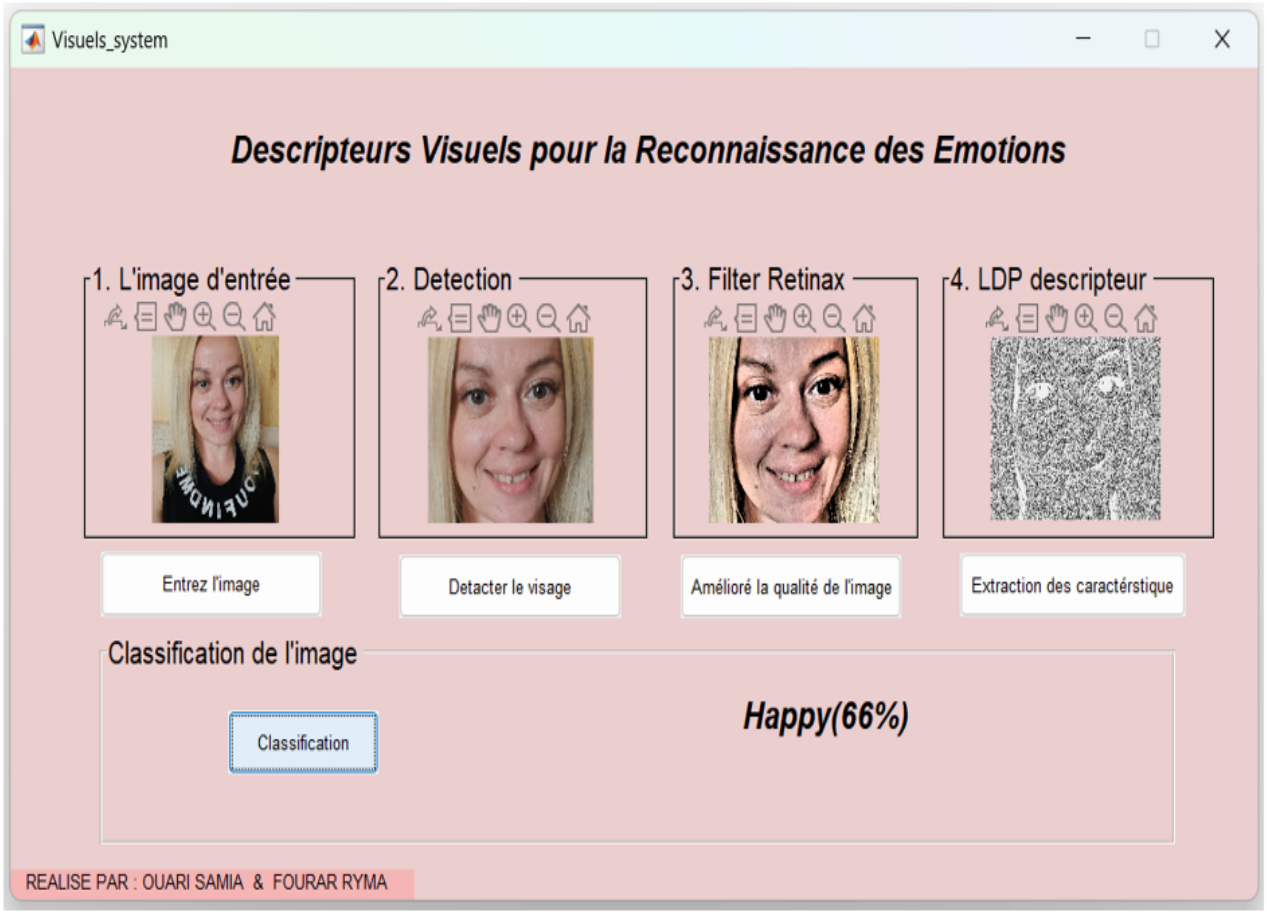

FIGURE A.4 – Test d'identification d'émotion a partir de l'image faciale.

Résumé

Les émotions jouent un rôle crucial dans la communication et l'interaction humaines, permettant aux individus de s'exprimer au-delà du domaine verbal. La capacité de comprendre les émotions humaines est souhaitable pour les ordinateurs dans diverses applications. Les récentes avancées technologiques ont permis aux utilisateurs de communiquer avec les ordinateurs de manière auparavant inimaginable. Cette recherche présente une approche holistique de l'analyse des sentiments et des émotions, intégrant un ensemble diversifié d'algorithmes de machine learning et de deep learning pour analyser de manière exhaustive les données faciales et vocales.

Les contributions de ce travail incluent l'utilisation d'une méthode de prétraitement connue sous le nom de Multiscale Retinex (MSR) pour améliorer la qualité des images et le contraste. De plus, des descripteurs discriminants handcrafted tels que LDP (Local Directional Pattern), BSIF (Binarized Statistical Image Features) et LBP (Local Binary Patterns), ainsi que des descripteurs de deep learning comme VGG19 et ResNet101, sont utilisés pour la reconnaissance des émotions basées sur les images faciales. Pour la reconnaissance des émotions dans la parole, nous utilisons le célèbre descripteur Handcrafted MFCC (Mel Frequency Cepstral Coefficient) et le modèle acoustique préentraîné VGGish basé sur un CNN (Convolutional Neural Network). De plus, la méthode EDA (Exponential Discriminant Analysis) a été utilisée pour la séparation maximale entre les classes. En outre, une fusion au niveau des scores utilisant la somme pondérée (Weighted Sum Fusion, WSF) est employée pour améliorer le processus de correspondance. Des tests ont été effectués en utilisant trois base de données, où la méthode proposée a surpassé l'état de l'art.

Mots clés : Reconnaissance des Emotions Audio-Visuels, EDA, CNN, MSR, MFCC, LDP, WSF.

Abstract

Emotions play a crucial role in human communication and interaction, allowing individuals to express themselves beyond the verbal domain. The ability to understand human emotions is desirable for computers in various applications. Recent technological advancements have enabled users to communicate with computers in previously unimaginable ways. This research presents a holistic approach to sentiment and emotion analysis, integrating a diverse set of machine learning and deep learning algorithms to comprehensively analyze facial and vocal data.

Contributions of this work include the use of a preprocessing method known as Multiscale Retinex (MSR) to enhance image quality and contrast. Additionally, discriminative handcrafted descriptors such as LDP (Local Discriminant Pattern), BSIF (Binarized Statistical Image Features), and LBP (Local Binary Patterns), as well as deep learning descriptors like VGG19 and ResNet101, are used for facial emotion recognition. For speech emotion recognition, we employ the well-known handcrafted descriptor MFCC (Mel Frequency Cepstral Coefficient) and the pretrained acoustical model VGGish based on a Convolutional Neural Network (CNN). Furthermore, Exponential Discriminant Analysis (EDA) method was used for maximal class separation. Additionally, score-level fusion using Weighted Sum Fusion (WSF) is employed to enhance the matching process. Tests were conducted using three datasets, where the proposed method outperformed the state-of-the-art.

Keywords: Audio-Visual Emotion Recognition, EDA, CNN, MSR, MFCC, LDP, WSF.