Analyse temps réel des micro-expressions par vision artificielle

Reda Belaiche

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A l'UNIVERSITE DE BOURGOGNE

Ecole doctorale n^{o} 37

Sciences Physiques pour l'Ingénieur et Microtechniques

Par

Mr. BELAICHE Reda

Analyse temps réel des micro-expressions par vision artificielle

Thèse présentée et soutenue à Dijon, le 30/09/2022

Composition du Jury :

Pr. Eva DOKLADALOVA	Professeure, ESIEE	Rapportrice
Pr. Sabir JACQUIR	Professeur, Université Paris-Saclay	Rapporteur
Pr. Frédéric MORAIN-NICOLIER	Professeur, Université de Reims	Président / Examinateur
Pr. Fan YANG	Professeure, ImViA, UBFC	Directrice de thèse
Pr. Dominique GINHAC	Professeur, ImViA, UBFC	Co-directeur de thèse
Dr. Cyrille MIGNIOT	Maître de Conférences, ImViA, UBFC	Co-encadrant de thèse

Remerciements

Il me serait très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement Madame Fan YANG, professeure à l'IUT de Dijon, Monsieur Cyrille MIGNIOT, maitre de conférences à l'UFR sciences et techniques de l'université de Bourgogne, ainsi que Monsieur Dominique GINHAC, professeur à l'ESIREM, qui m'ont encadré tout au long de cette thèse et qui m'ont fait partager leurs savoir-faire et brillantes intuitions. J'aimerais aussi et surtout les remercier pour leur gentillesse, leur disponibilité permanente et pour les nombreux encouragements qu'ils m'ont prodigué.

Je remercie Monsieur Franck MARZANI, directeur du laboratoire de Recherche ImViA, ainsi que Julien DUBOIS, Directeur de l'équipe CoReS de ImViA.

Je tiens à remercier Madame Eva DOKLADALOVA et Monsieur Sabir JACQUIR pour m'avoir fait l'honneur d'accepter d'être rapporteurs et pour leurs retours sur mon manuscrit. Je voudrais aussi remercier Monsieur Frédéric MORAIN-NICOLIER pour m'avoir fait l'honneur d'accepter de participer à mon jury de thèse lui aussi.

Je tiens aussi à remercier toutes les personnes que j'ai rencontré et avec qui je suis devenu amis à Dijon. Que ce soit au niveau de mon îlot dans le laboratoire, mais aussi les collègues un peu plus distants géographiquement. Je tiens aussi à remercier les amis que j'ai rencontré en dehors du laboratoire qui m'ont eux aussi soutenu tout au long de cette these. Je tiens particulièrement à remercier mes parents et ma famille pour leur soutien émotionnel et leur présence continue.

Sections et Chapitres

BELAICHE Reda

Table des matières

Abréviations

3DHOG	3D Extended Histogram of Oriented Gradients
3D FFT	3D Fast Fourier Transform
ANS	Autonomous Nervous System
AU	Action Unit
CASME	Chinese Academy of Sciences Micro-expression
CFD	Collaborative Feature Difference
CL	Convolutional Layer
CLM	Constraint Local Model
CLBP	Compound Local Binary Pattern
CNN	Convolutional Neural Network
DMME	Durée Moyenne d'une Micro-Expression
DSSN	Dual-stream shallow network
EVM	Eulerian Video Magnification
FACS	Facial Action Coding System
fps	frames per second
FHOFO	Fuzzy Histogram of Optical Flow Orientations
GLMM	Global Lagrangian Motion Magnification
HF	High Frequency
HOG	Histogram of Oriented Gradients
HOOF	Histogram Of Optical Flow
HRV	Hear Rate Variability
IA	Intelligence Artificielle
ICE-GAN	Identity-aware and Capsule-Enhanced Generative Adversarial Network
KNN	K-Nearest Neighbors
LBP	Local Binary Pattern
LBP-TOP	Local Binary Pattern on three orthogonal planes
LF	Low Frequency
LGCP	Local Gray Code Pattern
LOOCV	Leave One Out Cross-Validation
LOSO	Leave One Subject Out
LR-GACNN	Landmark Relations with Graph Attention Convolutional Network
LSTM	Long short-term memory
LTP	Local Temporal Pattern
LTP-ML	Local Temporal Pattern of Facial Movements
M	Magnitude
MAE	Mean Absolute Error
M/M-FE	Micro/Macro-Facial Expressions
ME	Micro Expression

MEGC	Micro-Expression Grand Challenge
MER-GCN	Micro-Expression Recognition Graph Convolutional Networ
MDMO	Main Directional Mean Optical-flow
OFF-ApexNet	Optical Flow Features from Apex frame Network
OS	Optical Strain
PPI	Pulse-Pulse Interval
PRV	Pulse Rate Variability
RBF	Radial Basis Functio
RCNN	Region-based Convolutional Neural Networks
ROI	Region Of Interest
RPPG	Remote photoplethysmography
SAMM	Spontaneous Micro-Facial Movement Dataset
SFED	Subtle Facial Expression Database
SMIC	Spontaneous MICro-expression Database
SVM	Support Vector Machine
TIM	Time interpolation model
VGG	Visual Geometry Group

Chapitre 1

Introduction

Les technologies de l'interaction homme-machine se concentrent de plus en plus sur l'être humain, que ce soit sur son identité, ou bien sur son état physique et mental. Des progrès conséquents ont été réalisés depuis quelques décennies. Par exemple, l'imagerie médicale apporte une aide précieuse pour le diagnostic des maladies. De même, la biométrie permet de reconnaître de façon certaine un individu en utilisant une modalité biologique comme le visage, l'empreinte digital ou la paume. Cependant l'étude des pensées et des émotions reste encore un domaine peu développé.

L'activité physiologique d'un individu est étroitement liée à ses états émotionnels. Par exemple, des modifications du rythme cardiaque, de la pression artérielle, de la température corporelle, des rythmes EEGs, et de la conductance cutanée peuvent intervenir à la suite d'événements émotionnellement chargés. Une branche moderne de l'informatique, appelée "Informatique Affective" et basée sur les travaux de Rosalind Picard datés de 1995, permettrait d'élaborer des machines capables de lire nos émotions en temps réel : les jeux vidéo sauront quand vous vous ennuyez, les annonceurs sauront lorsque vous êtes influencés par une publicité, et surtout les professionnels de la santé seront avertis lorsque vous aurez besoin de l'aide d'un psychologue. Dans ce contexte, il est primordial d'arriver à acquérir ces données physiologiques, à comprendre ou à anticiper le comportement des êtres humains.

Nous pouvons distinguer trois modes d'expression de l'émotion chez l'homme : la voix, l'expression corporelle et l'expression faciale. La voix, facile à imiter et usurper, ne peut pas servir de moyen d'authentification. De plus, les différents accents au sein d'une même langue posent problème lors de la compréhension. En ce qui concerne l'expression corporelle, il n'y a pas d'interprétation clairement définie ; bien qu'il existe des similitudes, les expressions corporelles n'ont pas toujours la même signification selon les cultures. En revanche, les expressions faciales possèdent de nombreux atouts. Tout d'abord elles sont révélatrices de l'émotion, et exprimées de façon compréhensible d'un point de vue humain. En effet, nous pouvons sans difficulté associer un visage à une émotion. Mais surtout elles peuvent être capturées de manière non-intrusive : les expressions sont visibles extérieurement et sont capturées par une simple caméra.

Dans ce manuscrit, nous proposons une étude pour décrire, analyser et reconnaître les expressions faciales en utilisant l'intelligence artificielle. Plus particulièrement, nos contributions portent sur le traitement des micro-expressions avec la technologie de la vision par ordinateur.

1.1 L'affective computing : l'étude de la psyché humaine

Les émotions jouent un rôle essentiel pour la santé et le bien-être d'une personne. En effet, de fortes émotions peuvent modifier significativement le rythme cardiaque, la pression artérielle ou même la température du corps. De plus, les troubles psychologiques comme le stress, la dépression ou l'anxiété chronique présentent un véritable risque pour la santé.

L'informatique affective (en anglais Affective Computing et aussi appelée intelligence artificielle émotionnelle) consiste à concevoir et à développer des systèmes pouvant reconnaître, interpréter, analyser et synthétiser les affects humains (sentiment, humeur, émotion et sensation). Cette branche moderne de l'informatique permet d'analyser les états émotionnels d'une personne et de fournir une réponse adaptée et appropriée de la part de l'interface homme machine.

De multiples signaux primaires peuvent exprimer l'émotion [1], ou de façon plus générale l'affect en réponse à une situation : la voix, le style d'écriture, la posture du corps ou bien l'expression du visage. Ces signaux peuvent être internes au corps humain comme le rythme cardiaque ou la température de la peau : ce sont des signaux physiologiques. Ils peuvent être relevés à partir de procédés complexes comme l'électrocardiographie ou l'électromyographie.

L'informatique affective est ainsi un très vaste domaine entremêlant de multiples domaines allant de l'informatique à la psychologie. Elle engendre alors des études nombreuses et variées : l'acquisition des signaux, le débruitage de ces signaux, l'extraction des caractéristiques descriptives, la reconnaissance d'une émotion, la quantification de son intensité ou même sa prédiction. Autant d'explorations possibles et de recherches à effectuer, chacune définie par une suite spécifique de contraintes et d'objectifs. L'informatique affective peut être utilisée dans un but utile et vertueux pour participer au "Mouvement pour une informatique positive". Ce mouvement spécule que le bien-être des personnes devrait être l'objectif premier de l'innovation technologique. Selon lui, des ordinateurs doux et bienveillants pourraient offrir une assistance personnalisée aux personnes souhaitant améliorer leurs compétences et leur confiance en elles.

1.2 Applications basées sur l'affective computing

En lien avec le concept de l'informatique positive, de nombreuses applications de l'analyse des émotions ont vu le jour pour promouvoir l'autonomie humaine. Introduisons ici quelques exemples dans des domaines variés de la vie courante.

Dans le domaine de l'e-santé, l'estimation de l'affect peut aider les services psychologiques à déterminer l'état émotionnel d'un patient ou bien permettre, par exemple, à des personnes souffrant d'autisme de bénéficier de technologies de communication adaptées. De même, la définition de la souffrance est un sujet sensible. Le degré de souffrance est une donnée très recherchée par les praticiens pour établir un diagnostic. Mais il s'agit souvent d'un niveau subjectif décrit par chaque patient. Un score déterministe calculé par un ordinateur serait un atout significatif à l'aide au diagnostic. Dans le domaine de la formation en ligne, le style de présentation peut être modifié si l'apprenant s'ennuie, se sent frustré ou bien si son attention baisse. Dans le domaine de l'aide à la conduite, un retour direct du ressenti du conducteur fournit un critère de choix pour évaluer ses capacités à conduire en toute sécurité. Les différents outils d'aide pourraient alors s'adapter à ses besoins immédiats et réduire considérablement les dangers encourus. Enfin parlons du domaine du marketing : connaître le ressenti profond d'une personne face à un produit ou une annonce permet de cibler au mieux les besoins d'un individu et de lui fournir un produit adéquat. Les enquêtes de satisfaction étant biaisées par le média écrit et l'interprétation du client de son besoin/désir.

Les nouvelles technologies apportent diverses solutions dans le domaine de l'informatique affective. Les mesures des signaux physiologiques produisent une évaluation honnête et quantifiable du fonctionnement du corps humain et donc du ressenti de la personne. Ces signaux peuvent être acquis avec le matériel adéquat. Par exemple, l'ECG consiste à enregistrer l'activité du cœur sur une période du temps en utilisant des électrodes placées sur la peau. L'EMG retranscris le degré de contraction des muscles face à une activité ou un stimulus électrique. L'EEG est une technique d'imagerie médical donnant une image du cerveau en mesurant la fluctuation du voltage résultant d'un flux de courant à travers les neurones du cerveau. Ces mesures sont fiables et révélatrices d'un état interne à la personne, mais elles se révèlent intrusives. Par exemple, le placement des électrodes sur la peau peut être ressenti négativement et représenter une source de stress. Pour évaluer des éléments si subtils que les émotions, un tel protocole ne semble pas très pertinent puisque l'installation même du capteur provoquerait le stimulus émotionnel.

Les solutions basées sur la vision par ordinateur apportent un confort inédit vis-à-vis de ces mesures invasives : les données sont acquises par une simple caméra.

Il peut être perturbant de se savoir filmé, mais cela reste négligeable par rapport aux systèmes avec contact physique. Le signal reçu, bien que bruité par beaucoup d'information redondante, met cependant à disposition une multitude d'indicateurs.

La vision par ordinateur est aussi capable de fournir des données physiologiques à travers une nouvelle technologie nommée rPPG (remote PhotoPlesthymoGraphy, PPG sans contact en français). Le PPG est une technique pour détecter les changements de volume microvasculaire du sang dans les tissus. Elle ne requière qu'une source de lumière et un photodétecteur. La source de lumière illumine le tissu et le photodétecteur mesure la faible variation de lumière réfléchie ou transmise associée au changement en perfusion dans le tissu. À la différence du PPG, le rPPG permet de réaliser ce procédé à distance à l'aide d'une simple caméra pour capturer une zone de la peau (souvent le visage d'un individu). Plusieurs indicateurs physiologiques peuvent être extraits du signal rPPG, comme la variation du rythme cardiaque (Heart Rate Viriation – HRV en anglais) et le rythme respiratoire.

Nous pouvons aussi noter que certains indicateurs des émotions qui s'expriment de façon extérieure au corps humain sont relativement facile à capter par une caméra. On peut citer les expressions corporelles et plus particulièrement celles faciales, les plus révélatrices des émotions. L'analyse des expressions faciales est une branche très importante de la vision par ordinateur depuis une dizaine d'années.

De nos jours, la technologie du deep-learning a fortement révolutionné le domaine de l'intelligence artificielle. De nombreux problèmes ont trouvé une solution adéquate avec la forte hausse des performances et aussi la rapidité d'exécution des algorithmes sur des réseaux de neurones profonds. Dans ce contexte, cette technologie devient une réponse pertinente et compétitive tant en termes de précision que de temps de calcul pour l'informatique affective.

1.3 Les macro- et micro-expressions faciales

Comme vu précédemment, l'expression corporelle d'une émotion peut se traduire de plusieurs façons : la voix, le gestuel, la température de la peau, etc. L'expression faciale reste l'une des plus pertinentes. Tout d'abord elle exprime une grande palette d'émotions. Ensuite, comme c'est une expression extérieure, elle peut être acquise de façon non-invasive. Techniquement la détection de visage est un problème maintenant associé à des algorithmes légers et très efficaces. Ainsi se concentrer sur cette partie du corps n'engendre pas de complications méthodologiques.

Les expressions faciales représentent un élément crucial de la communication entre les êtres humains. Il s'agit même de l'une des parties les plus importantes de la communication non verbale. De ce fait, elles peuvent être envisagées comme une

porte ouverte sur nos pensées et nos émotions. Les expressions faciales sont aussi communes à tous les êtres humains, et même aux mammifères. En effet, Darwin a démontré en 1872 que les expressions faciales étaient universelles [2], mais aussi qu'il existait des émotions quasi-instinctives, relevant donc plus de l'inné que de l'acquis.

Du fait de leur importance notable dans les interactions sociales, les chercheurs en psychologie se sont intéressés aux expressions faciales depuis les balbutiements du domaine. Ce n'est cependant qu'en 1971 que Ekman and Friese ont énoncé le postulat de l'existence de six émotions primaires [3]. Elles sont universelles et communes à tous les êtres humains, indépendamment de leur culture ou de leur origine. Ces six émotions sont : la joie, la tristesse, la colère, la surprise, le dégoût et la peur.

Ekman a travaillé sur les moyens les plus objectifs de reconnaître les émotions ; principalement en se concentrant sur les mouvements musculaires spontanées. Il a grandement contribué au domaine en concevant le Facial Action Coding System (FACS) qui attribue des mouvements spécifiques des muscles présents sur le visage humain aux six émotions primaires. Le FACS a très vite été pris en considération pour représenter et distinguer les expressions faciales. Certaines déformations du visage traduisent explicitement le ressenti d'une émotion. Par exemple un sourire traduit la joie, une grimace le dégoût ou des sourcils froncés la colère. Ces expressions visibles à l'œil, sont appelées macro-expressions (Figure 1.1).

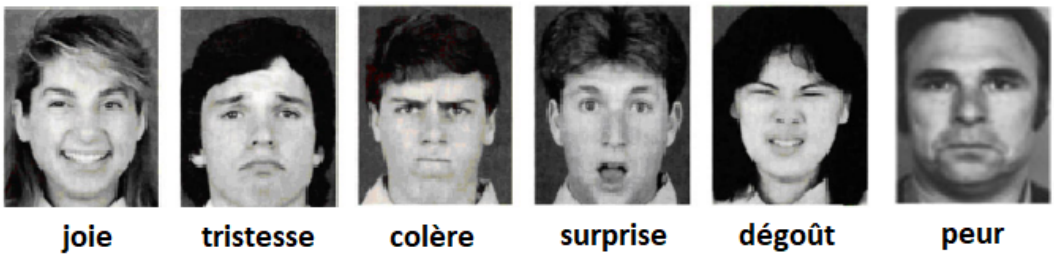

FIGURE 1.1 – Exemples de macro-expressions sur la base de données "The Extended Cohn-Kanade Dataset" [4]. On y retrouve les 6 types d'émotions tels que défini par Ekman : la joie, la tristesse, la colère, la surprise, le dégoût et la peur.

Initialement découvertes par Haggard et Isaacs [6], les micro-expressions sont un type d'expressions faciales involontaires extrêmement rapides et de très faible intensité. Ces micro-expressions peuvent se produire dans deux situations : la suppression consciente et la répression inconsciente. La suppression consciente apparaît lorsqu'une personne essaie intentionnellement de s'empêcher de montrer ses véritables émotions ou de les cacher. La répression inconsciente se manifeste lorsque le sujet lui-même ne se rend pas compte de ses véritables émotions. Dans

FIGURE 1.2 – Exemples de micro-expressions sur la base de données SAMM [5]. On y retrouve les 6 types d'émotions tels que définis par Ekman.

les deux cas, les micro-expressions trahissent les véritables émotions du sujet indépendamment de sa conscience de leur existence.

Dans la littérature, les macro-expressions sont définies par des mouvements volontaires, caractérisées par des mouvements de forte intensité d'une durée comprise entre 0.5 et 4 secondes. Quant aux micro-expressions, elles sont souvent involontaires et durent une fraction de seconde, en moyenne entre 40 ms et 250 ms. Bien que ces mouvements soient très rapides et généralement non perceptibles pour l'œil humain, les micro-expressions apportent de précieux renseignements sur l'état émotionnel d'une personne.

Si les macro-expressions peuvent être imitées ou simulées pour falsifier l'émotion (par exemple, les acteurs reproduisent l'état émotionnel du personnage qu'ils interprètent), il est très difficile de dissimuler ou feindre une micro-expression. De ce fait, on considère que les micro-expressions d'un individu représentent ses vrais ressentis.

Si la reconnaissance et la classification des macro-expressions font l'objet de recherches scientifiques dans le domaine de l'intelligence artificielle depuis déjà une dizaine d'années, l'étude sur les micro-expressions est très récente et de nombreuses questions restent ouvertes.

1.4 Objectifs de la thèse

On peut recenser deux tâches principales dans l'étude des expressions faciales : la reconnaissance et le spotting. La reconnaissance consiste à analyser le contenu de la séquence vidéo d'une expression et à estimer le type d'émotion qui lui est lié. Le spotting consiste quant à lui à détecter temporellement une expression dans une

séquence. La reconnaissance se base sur une classification alors que le spotting a pour objectif la détection et la localisation d'évènement (l'expression). Les deux tâches sont intimement liées puisque le spotting fournit les expressions à classer par la reconnaissance. En fait la sortie du spotting correspond à l'entrée de la reconnaissance. Bien qu'une utilisation pratique implique la combinaison des deux tâches, les études dans le domaine les séparent généralement en 2 étapes.

L'analyse des micro-expressions (ME) est un sujet d'actualité dans le domaine de la vision par ordinateur car elle constitue une passerelle importante pour saisir et comprendre les émotions humaines quotidiennes. Il s'agit néanmoins d'un problème difficile, car la micro-expression est généralement transitoire (étant donné qu'elle dure moins de 250 ms) et subtile.

Les récents progrès du machine learning permettent d'adopter de nouvelles méthodes efficaces pour accomplir diverses tâches de la vision par ordinateur. En particulier, l'utilisation de techniques d'apprentissage profond sur de grands ensembles de données surpasse les approches classiques basées sur l'apprentissage classique avec des caractéristiques « artisanales ». Même si les ensembles de données disponibles pour la ME spontanée sont rares et beaucoup plus réduits, l'utilisation de réseaux neuronaux convolutionels (CNN) dans ce domaine donne aussi des résultats de classification relativement satisfaisants. Cependant, ces réseaux sont gourmands en termes de consommation de mémoire et de ressources de calcul. Cela pose de grands défis lors du déploiement de solutions basées sur les CNN dans de nombreuses applications grand public, telles que la surveillance des conducteurs et la reconnaissance de l'émotion dans les classes virtuelles (e-learning), qui exigent une analyse précise, rapide et portable sur des systèmes embarqués.

Basée sur les nouvelles avancées techniques du deep learning, l'objectif de cette thèse est d'établir un système de reconnaissance des émotions sans contact et temps réel. Il devra respecter plusieurs contraintes liées aux applications grand-public : robustesse et flexibilité, faible coût, usage simple et capacité à être embarqué à moindre coût énergétique.

1.5 Contributions et plan du manuscrit

Le travail réalisé dans le cadre de cette thèse a abouti à l'élaboration de plusieurs méthodes faisant avancer la problématique de l'analyse des micro-expressions faciales. Ces méthodes ont été expérimentalement validées et évaluées pour fournir une base d'étude solide en vue des travaux à venir.

Nos principales contributions sont :

— Nous avons repris les principaux descripteurs utilisés pour la reconnaissance

des micro-expressions. Une comparaison poussée a permis de mettre en exergue les qualités et défauts de chacun. Nous avons alors introduit un cadre d'uniformité temporel pour développer une méthode originale se basant sur le descripteur le plus performant : le LBP-TOP. Ce cadre a permis une comparaison plus juste entre les micro-expressions et une hausse des performances de reconnaissance.

— Suivant l'ordre chronologique des avancées dans le domaine, nous avons ensuite repris les méthodes les plus performantes de reconnaissance d'émotion utilisant l'apprentissage profond. Dans un objectif de diminution de la complexité du système, en accord avec un système embarqué crédible, nous avons présenté un nouveau réseau calibré, dans son architecture, pour optimiser les performances avec le réseau le moins profond.

— Toujours dans l'objectif de produire un réseau performant et léger, nous avons aussi proposé un réseau prenant en entrée des éléments de moindre dimension. À la place du format d'image classique 3D (séquence vidéo), nous avons introduit des données liées au flot optique (mouvement) codées sur une et deux dimensions. Les expériences réalisées ont confirmé le meilleur compromis obtenu entre la précision et le gain en complexité.

— Pour une mise en pratique des méthodes pour des scénarios plus réalistes, nous avons évalué les performances des algorithmes de reconnaissance en les combinant avec ceux de spotting, c'est à dire de détection de micro-expressions dans des séquences de vidéo. Nous avons ainsi testé le protocole complet d'analyse des micro-expressions. Nous avons alors développé un nouveau paradigme où l'utilisateur indique le début de la micro-expression. Ce paradigme permet de rendre le problème plus accessible tout en restant sur un contexte cohérent avec l'application pratique. Nous avons ainsi évalué plusieurs procédés d'estimation du spotting en fonction de leur effet sur le protocole général.

Ce manuscrit est organisé comme suit :

— **Chapitre 2** : Les scientifiques distinguent deux types d'expressions faciales : les macro-expressions et les micro-expressions. Cette thèse se concentre principalement sur les micro-expressions, et donc, dans ce chapitre, nous présentons le système le plus utilisé pour décrire les mouvements des muscles faciaux liés aux différentes expressions dans la littérature scientifique pour décrire les micro-expressions. Par la suite, nous exposons les avancées dans le domaine de l'analyse des micro-expressions en distinguant deux principales tâches : le spotting, qui consiste en une détection de la présence de la micro-expression, et deuxième tache qui consiste en la reconnaissance de celle-ci. Nous présentons des bases de données publiques et des protocoles d'évaluation couramment utilisés.

— **Chapitre 3** : La caractérisation des données par un descripteur permet de maîtriser les éléments choisis pour définir les classes recherchées. Leur

association avec un classifieur formait la majorité des premières méthodes de reconnaissances de ME. Dans ce chapitre nous reprendrons dans un premier temps les méthodes les plus significatives de l'état de l'art utilisant ce procédé pour évaluer et comparer leur performance. Nous présentons ensuite deux contributions : l'unification temporelle au niveau du descripteur le plus reconnu, afin de rendre l'étape de comparaison plus légitime, et la simple utilisation de corrélation avec des gabarits de mouvement. Les améliorations apportées en précision et en simplicité sont expérimentalement explicitées.

— **Chapitre 4** : Dans ce chapitre, nous nous intéressons aux méthodes basées sur l'apprentissage profond pour la classification de ME. Premièrement nous reprenons les méthodes les plus efficaces de l'état de l'art pour bien en révéler le comportement et servir de base de référence. Puis nous présenterons deux méthodes originales. L'objectif est de vérifier la capacité du problème à s'accommoder d'un système embarqué et de ses contraintes (temps-réel, consommation, ...). Les deux méthodes se basent sur le réseau CNN ResNet18. Dans la première, la structure du réseau est modifiée : nous avons réduit le nombre de couches du réseau pour qu'il s'adapte au mieux au cas bien particulier de la classification de ME. La seconde méthode fait varier le format des données d'entrée du réseau : nous entrons des représentations du mouvement par flot optique codées sur une, deux ou trois dimensions. Les résultats obtenus montrent une amélioration notable de la complexité du réseau sans dégradation significative de ses capacités de classification.

— **Chapitre 5** : Pour un scénario réaliste d'applications, la reconnaissance de ME se doit d'être associée à leur détection préalable (spotting). Dans ce chapitre nous présentons une étude sur le système complet associant ces deux problématiques. Après avoir évalué plusieurs méthodes de spotting ainsi que l'influence des imprécisions qu'elles génèrent sur les résultats de classification, nous avons proposé un nouveau paradigme où le début de la ME est connu. Cette simplification du problème le rend certes moins général mais permet de rester dans un cadre pratique acceptable tout en ouvrant la voie à des résultats bien plus encourageants. Ainsi, le spotting seul mais aussi le système complet sont évalués. Enfin nous présentons une implantation matérielle sur un petit système embarqué.

— **Chapitre 6** : Les performances des méthodes de classification sont intimement liées aux données utilisées pour l'étape d'apprentissage. Dans ce chapitre, nous présentons une étude de l'influence des caractéristiques des bases de données sur les résultats obtenues. En fait, les bases de données disponibles présentent de forts déséquilibres interclasse sur les données. Nous en discutons alors les effets en utilisant des bases équilibrées tout d'abord au niveau des tests puis au niveau de l'apprentissage.

— **Chapitre 7** : Dans ce chapitre, nous réalisons une synthèse des travaux menés au cours de cette thèse, puis nous soulignons les limitations relatifs à la fois à nos travaux mais aussi au domaine d'étude. Finalement, nous proposons des perspectives de recherche afin de prolonger les travaux menés et de surmonter les limitations rencontrées.

Chapitre 2

État de l'art

Les expressions faciales ont toujours joué un rôle prépondérant dans l'expression des émotions des êtres humain et dans leur capacité à communiquer [7, 8]. Les scientifiques distinguent deux types d'expressions faciales : les macro-expressions qui sont définies par leur amples mouvements musculaires et leur relativement longues durées (entre 0.5 et 4 secondes) [9, 2], et les micro-expressions caractérisées quant à elles par leur mouvements musculaires infimes et leur durée de temps très courtes (entre 0.25 et 0.04 secondes) [2, 10, 11].

Dans ce chapitre, nous présentons d'abord le système FACS décrivant les mouvements des muscles faciaux liés aux différentes expressions. Ensuite, nous passons en revue les avancées dans le domaine de l'analyse des micro-expressions en distinguant deux principales tâches : le spotting et la reconnaissance. Nous introduisons aussi les bases de données publiques couramment utilisées ainsi que les protocoles d'évaluation des performances.

2.1 Le Facial Action Coding System

Il existe deux voies principales dans les recherches psychologiques qui se concentrent sur la mesure des expressions faciales : le jugement du message et le jugement du signe. Le jugement par message vise à décrire les expressions faciales sous la forme d'un ensemble d'étiquettes affectives discrètes telles que les émotions de base ou d'un autre ensemble d'étiquettes émotionnelles. L'approche du jugement par le signe vise quand à lui à décrire les expressions faciales affichées (mouvement du visage ou forme de la composante faciale) en termes de mouvements musculaires faciaux activés ou d'unités d'action (Action unit – AU en anglais) par le biais du Facial Action Coding System (FACS) [2, 3].

Ekman et Friesen ont développé le FACS pour décrire les expressions faciales à partir des AUs. Sur les différents AUs du FACS qu'ils ont définies, 30 sont anatomiquement liées aux contractions de muscles faciaux spécifiques (12 pour le haut du visage et 18 pour le bas du visage). En outre les AUs peuvent être isolées ou

combinées. Lorsque les AUs sont combinées, elles peuvent être additives ou non. Des AUs additives signifient que la combinaison ne modifie pas l'apparence des AUs qui les composent. Bien que le nombre d'unités d'action atomique soit relativement faible, plus de 7000 combinaisons différentes d'AUs ont été observées [12]. Le FACS fournit le pouvoir descriptif nécessaire pour décrire les détails de l'expression faciale. Dans le Tableau 2.1 nous pouvons voir les AUs les plus significatives pour l'expression des émotions sur le visage. Le Tableau 2.2 quant à lui représente différentes émotions basiques et les AUs desquelles elles résultent.

L'un des points les plus intéressant concernant les ME est qu'elles transparaissent principalement dans des scénarios où le sujet arbore des émotions refoulées par son inconscient, ou lorsqu'il essaye de dissimuler des informations dont il est bel et bien conscient [11, 13, 14]. Ceci engendre des mouvements des muscles faciaux avec de petites amplitudes et de très courtes durées. Les AUs sont ainsi difficilement détectables. Bien que la grande majorité des êtres humains soient capable de percevoir les macro-expressions, très peu sont capable d'apercevoir les micro-expressions. En effet, il faut avoir une très bonne acuité visuelle, et avoir suivi un entraînement assez rigoureux pour pouvoir espérer reconnaître les ME à l'œil nu. Mêmes après un entraînement intensif, seulement 47% des ME en moyenne peuvent être repérées et classées par les êtres humains [15].

La reconnaissance de ME est une sous-discipline de la reconnaissance d'émotion relativement récente. Avant l'année 2010, la communauté scientifique s'était principalement concentrée sur la reconnaissance de macro-expressions. La recherche sur la reconnaissance de ME a commencé en 2011 après la publication de la première base de données relatant des ME spontanées. Dans les sections suivantes de ce chapitre, nous détaillons les méthodes de spotting et de classification de ME les plus courantes dans la littérature scientifique en passant par les techniques d'extraction de caractéristiques les plus utilisées.

2.2 Spotting de micro-expressions

Le processus de spotting consiste à déterminer l'occurrence d'une micro-expression et de la localiser dans le temps. La reconnaissance de ME a donné lieu aux premières études. La faisabilité d'une différenciation automatiquement entre plusieurs émotions a déjà dû être démontrée et ce, pour un plus grand contrôle et une évaluation facilitée, à partir d'une base de données annotée. Cependant, suite aux premiers résultats, l'étape de spotting s'est imposée comme un traitement à étudier d'urgence. En effet, pour une application pratique, il n'est pas raisonnable de demander à l'utilisateur de spécifier précisément la présence et le moment d'une ME.

On distingue deux approches différentes pour le spotting de ME. La première consiste à extraire sur chaque frame une caractéristique spécifique de l'apparition

Num	nature de l'AU
1	Remontée de la partie interne des sourcils
2	Remontée de la partie externe des sourcils
4	Abaissement et rapprochement des sourcils
5	Ouverture entre la paupière supérieure et les sourcils
6	Remontée des joues
7	Tension de la paupière
8	Lèvres collées
9	Plissement de la peau du nez vers le haut
10	Remontée de la partie supérieure de la lèvre
11	Ouverture du nasolabial
12	Étirement du coin des lèvres
13	Étirement et rentrée des lèvres
14	Plissement externe des lèvres (fossettes)
15	Abaissement des coins externes des lèvres
16	Ouverture de la lèvre inférieure
17	Élévation du menton
18	Froncement central des lèvres
19	Sortie de la langue
20	Étirement externe des lèvres
21	Tension du cou
22	Lèvres en "O"
23	Tension refermante des lèvres
24	Lèvres pressées (pincement des lèvres)
25	Ouverture de la bouche et séparation légère des lèvres
26	Ouverture de la mâchoire
27	Bâillement
28	Succion interne des lèvres
29	Poussée de la mâchoire
30	Déplacement de côté de la mâchoire
31	Serrement de la mâchoire
32	Morsure des lèvres
33	Gonflement des joues
34	Bouffée des joues
35	Aspiration des joues
36	Bombement de la langue
37	Essuyage des lèvres
38	Dilatation des naseaux
39	Compression des naseaux
41	Abaissement de la glabelle
42	Abaissement interne des sourcils
43	Yeux fermés Relaxation du levator
44	Rapprochement des sourcils
45	Clignotement
46	Clignement de l'œil

TABLE 2.1 – Liste des AUs basiques.

Émotion	Action Units
Joie	AU6, AU12, AU6+AU12, AU6+AU7+AU12, AU7+AU
Surprise	AU1+AU2, AU5, AU25, AU1+AU2+AU25, AU25+AU26, AU5+AU24
Colère	A23, AU4, AU4+AU7, AU4+AU5, AU4+AU5+AU7, AU17+AU24, AU4+AU6+AU7, AU4+AU38
Dégoût	AU10, AU9, AU4+AU9, AU4+AU40, AU4+AU5+AU40, AU4+AU7+AU9, AU4+AU9+AU17, AU4+AU7+AU10, AU4+AU5+AU7+AU9, AU7+AU1
Tristesse	AU1, AU15, AU1+AU4, AU6+AU15, AU15+AU
Peur	AU1+AU2+AU4, AU2

TABLE 2.2 – Les émotions basiques et les AUs qui leurs sont associées.

d'une ME. Son évolution au cours du temps doit révéler des motifs descriptifs de la présence d'une ME. Cette évolution peut venir d'un descripteur de quantité de mouvement dont les maximas locaux vont être considérés comme des ME, mais aussi de méthodes un peu plus fines. Aucun apprentissage n'est demandé ; la lecture des données donne seule le résultat sans comparaison à une base d'exemples ou à un modèle.

La seconde approche réalise d'abord un apprentissage. Des motifs récurrents entre descripteurs classiques sont recherchés pour bien différentier des exemples positifs et négatifs d'une base de données. C'est l'étude des données plutôt que l'estimation des éléments descriptifs qui sont mis en valeur. Les méthodes basées sur la variation des caractéristiques ont, dans un premier temps, été les plus présentes dans la littérature. Cependant la tendance commence à s'inverser avec l'apparition de plus en plus de méthodes utilisant l'apprentissage automatique.

2.2.1 Méthodes basées sur la différence de caractéristiques

Ces méthodes se basent sur un a-priori fort concernant le comportement d'une ME. Elles se basent généralement sur la même suite d'opérations : les caractéristiques spécifiques sont extraites à partir des différentes images composant une ME ; puis la différence entre ces caractéristiques en deux instants (par rapport à une référence ou entre deux instants séparés par une durée fixe) est calculées ; enfin les mouvements les plus significatifs sont isolés (par exemple en fixant un seuil) et considérés comme une ME. De nombreux travaux ont bien entendu approfondi et nuancé ce cadre d'étude (Tableau 2.3).

La première étape consiste en la sélection des caractéristiques. Comme une ME est un mouvement rapide et léger, un descripteur de déformation local réduit est alors à favoriser. Une différence peut se faire à partir d'un descripteur de formes comme le HOG (Histogram of Oriented Gradient) [5] ou le LBP (Local Binary Pattern) [16]. Davinson [17] utilise les 3DHOG, la transposition du HOG selon 3 plans pour une représentation spatio-temporelle semblable au LBP-TOP (Local Binary Pattern

histograms from Three Orthogonal Planes) pour le LBP. Ces descripteurs sont parfois combinés (Han [18] propose une stratégie collaborative pour associer LBP et HOOF (la version du HOG adaptée au flot optique)) ou les comparer [19, 20].

L'espace de représentation peut varier pour mieux correspondre au contexte. Par exemple Moilanen [16] utilise une transformé de Riesz pour magnifier le mouvement. Lu [21] applique la projection intégrale pour réduire le coût de calcul tandis que Li [22] intègre une FFT 3D sur la vidéo partant du principe que les petits mouvements sont plus visibles dans le domaine fréquentiel.

Le mouvement ou la déformation locale doit ensuite être mis en exergue par un calcul de différence. Liong [19] a proposé une simple corrélation avec la première frame de la séquence. Néanmoins, c'est la distance χ^2 qui sert de référence à la majorité des travaux de recherche [5, 21, 20, 17] pour comparer des signaux de hautes dimensions. Utilisant le domaine fréquentiel, Li [22] extrait les différences pertinentes par un filtre de bande haute fréquence.

Le flot optique permet d'extraire directement le mouvement. Calculé entre les frames relatives à deux instants, il produit une carte de déplacement. La différence est donc cette fois réalisée en amont de la description. Wang [23] extrait la magnitude de la différence maximale selon la direction principale du flot optique. La méthode *Optical Strain* (OS) est un dérivé du flot optique [19] efficace dans le contexte. Le HOOF, lui, permet de décrire la forme du flux optique comme le HOG décrit la forme d'une image RGB [20, 18].

Pour localiser l'information sur l'image, cette dernière est d'abord découpée en blocs (Figure 2.1(a)) sur lesquels le descripteur est calculé avant que le tout soit concaténé. Mais, en particulier au niveau des ME, l'information pertinente est localisée sur des parties bien spécifiques du visage. Un pré-traitement est alors souvent réalisé sur l'image pour localiser un certain nombre de points relatifs à des repères biologiques du visage (landmark) (Figure 2.1(b)). Le système FACS et la liste des AUs permettent alors d'isoler certaines parties de l'image (Region of Interest -ROI en anglais) les plus pertinentes (Figure 2.1(c-e)).

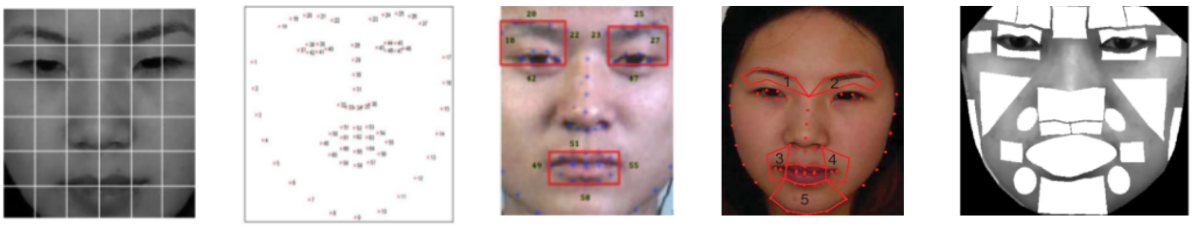

FIGURE 2.1 – Sélection des régions d'intérêts : découpage en blocs (a) ; points du visage (b) ; ROIs utilisés par [19] (c), [24] (d) et [5] (e).

Le processus de description est alors uniquement réalisé au niveau des ROIs,

évitant ainsi la perturbation des parties non actives du visage. Selon l'étude, il peut y avoir trois [19], cinq [25, 24] voire beaucoup plus [5] de ROIs. Ma [24] l'applique sur le flot optique pour créer un nouveau descripteur qu'il nomme RHOOF. La différence entre le nombre de pixels appartenant aux directions descendantes et la somme des pixels appartenant aux directions ascendantes y est extrait. Patel [26] réalise une intégration spatio-temporelle en étudiant le flot optique relatif à des points du visage regroupés selon leur appartenance à des AUs. Han [18] pondère l'influence des ROIs à partir du discriminant linéaire de Fischer.

La dernière étape du spotting réalise la détection de ME en se basant sur un signal temporel de vraisemblance ou de différence. Si le signal est correct, un seuillage simple [18], ou un algorithme de détection de pics [22] peut suffire. Liong [19] propose une recherche binaire pour sélectionner un apex (la frame où la ME est plus prononcée) plus centré au niveau de la ME. Un seuillage adaptatif permet de filtrer les maxima locaux erronés en imposant un certain motif dans l'évolution [23, 25]. Davinson [5] intègre les annotations de la base de données test avec un ABT (Adaptative Baseline Threshold). Pour améliorer la qualité du signal avant la recherche des maxima, Lu [21] réalise un ajustement polynomial, Li [20] joue sur le contraste et Ma [24] ajoute un filtrage médian.

ref	année	descripteur	distance	détection	ROI
[16]	2014	LBP	χ^2	seuillage	
[19]	2015	CLM + LBP + OS	corrélation	recherche binaire	✓
[26]	2015	Flot optique sur points du visage		seuillage	
[24]	2017	RHOOF		seuillage + filtrage médian	✓
[21]	2017	projection intégrale	χ^2	seuillage + ajustement polynomial	
[23]	2017	flot optique		seuillage adaptatif	
[20]	2017	LBP et HOOF	χ^2	seuillage sur signal contrasté	
[17]	2018	3DHOG	χ^2	seuillage adaptatif	✓
[25]	2018	Transformé de Riesz	variation 1D	seuillage adaptatif	✓
[18]	2018	LBP / HOOF	CFD	seuillage	✓
[5]	2018	HOG	χ^2	ABT	
[27]	2018	3D FFT	MAE	détection de pique	

TABLE 2.3 – Méthodes de spotting basées sur la différence de caractéristiques.

Les méthodes de ce type suivent un protocole assez similaire. Les principales variations résident sur le choix des caractéristiques et surtout les traitements permettant de nettoyer au mieux le signal temporel. Les méthodes se basant sur la différence de caractéristiques sont utiles pour étudier les mouvements du visage relativement importants, mais leur capacité à distinguer les ME d'autres mouvements reste faible, en particulier dans les vidéos longues. De ce fait, la dernière étape du spotting peut engendrer beaucoup de faux positifs : un seuil bas est aussi sensible aux micro-mouvements causés par les ME qu'aux clignotements d'œil, aux tics nerveux ou encore aux légers mouvements de la tête. C'est pourquoi certains chercheurs ont commencé à privilégier

les méthodes se basant sur l'apprentissage automatique afin de laisser les modèles eux même différencier les micro-mouvements relatifs aux ME des autres événements présents dans les vidéos.

2.2.2 Méthodes basées sur l'apprentissage automatique

En se basant sur une base de données, la classification est moins contrôlée explicitement mais des associations de caractéristiques inattendues peuvent être implicitement sélectionnées pour leur efficacité. Xia [28] utilise l'apprentissage pour extraire un signal de vraisemblance semblable à ceux de la section précédente. La valeur de vraisemblance relative à une courte séquence est obtenue à partir d'un RWM (Random Walk Model) initialisé par Adaboost sur la déformation géométrique des points caractéristiques du visage, suivi d'un seuillage du signal temporel pour définir les positions des ME.

L'apprentissage peut aussi se concentrer sur l'étape de classification (Tableau 2.4) : un descripteur donne une représentation du contexte de la ME pour chaque instant et un classifieur décide quels instants correspondent à une ME. Dans ce cas, l'apprentissage sert alors plus à séparer les données qu'à les sélectionner.

La ME est définie par un mouvement spécifique. Elle est donc décrite par une variation de l'apparence dans le temps. Borza [29] prédéfinit la durée moyenne d'une ME et utilise comme descripteur la magnitude du mouvement sur dix ROIs entre deux frames séparées de cette durée. Pour Husak [30], une ME correspond à la différence d'intensité du mouvement sur douze ROIs extraites sur une suite de frames d'une fenêtre glissante. Quant à Tran [31], il introduit une fenêtre glissante multi-échelle associée à une méthode de suppression de non-maximum afin de rendre la détection plus robuste à partir de plusieurs descripteurs.

Le flot optique est une représentation courante et efficace du mouvement. Li [20] utilise les HOOF sur douze ROIs et Liong [32] associe l'optical strain au LBP-TOP. Li [33] a remporté la partie spotting du challenge MEGC2019 à partir d'un descripteur particulier : le Local Temporal Pattern of Facial Movements (LTP-ML). Une Analyse en Composantes Principales (ACP) a été appliquée sur l'axe temporelle d'une fenêtre glissante pour isoler le mouvement principal au niveau des ROIs avant de fusionner la classification de chaque ROI.

En dehors des méthodes d'apprentissage classiques, l'apprentissage profond a aussi été récemment utilisé pour réaliser le spotting de ME malgré les quantités limitées de données disponibles (Tableau 2.5). Zhang et al.[34] furent les premiers à utiliser un CNN pour effectuer le repérage de la ME avec le SMEConvNet. Ils ont constaté que les données présentes ne sont pas assez diversifiées pour tirer le plein parti des propriétés des CNNs dans le cadre du spotting. Néanmoins, comme les caractéristiques extraites à partir d'une ME sont entraînées pour la classification, le processus d'apprentissage automatique améliore la capacité de distinction entre une

ref	année	descripteur	classifieur	ROI
[32]	2016	OS + LBP-TOP	SVM	
[29]	2017	différence de magnitude	adaboost	✓
[30]	2017	différence d'intensité	SVM	✓
[20]	2017	HOOF	SVM	✓
[31]	2018	LBP-TOP, HOG-TOP et HIGO-TOP	SVM	
[33]	2019	LTP-ML	SVM	✓

TABLE 2.4 – Méthodes de spotting basées sur le machine learning.

ME et un autre mouvement du visage.

Pan et al. [35] ont proposé une méthode basée sur le réseau de neurones convolutif bilinéaire local (LBCNN) afin d'atténuer le problème de la localisation des mouvements de faible intensité sur le visage. L'originalité vient de l'utilisation du réseau à la fois sur l'image complète et sur les ROIs pour une double perception. Tran [36] utilise la dimension temporelle du réseau de neurones récurrent LSTM. Il présente en entrée les différents descripteurs (LBP-TOP, HOG-TOP et HIGO-TOP) sur une fenêtre glissante.

ref	année	réseau	ROI
[34]	2018	SMEConvNet	
[36]	2019	LSTM	✓
[35]	2020	LBCNN	✓

TABLE 2.5 – Méthodes de spotting basées sur l'apprentissage profond.

Les méthodes basées sur l'apprentissage automatique sont de plus en plus utilisées dans la littérature pour résoudre le problème de spotting. Cependant, compte tenu de la taille relativement limitée des bases de données, nous pouvons spéculer que ces méthodes n'ont pas encore atteint leur plein potentiel.

2.3 Reconnaissance de micro-expressions

La grande majorité des applications visant l'étude des ME ignore la détection et la localisation temporelle des ME : savoir quand apparaît une ME n'induit pas vraiment l'état émotionnel de la personne. Seules les indices contenus dans ces ME peuvent nous donner cette information. Au final, le spotting constitue un pré-traitement à l'étape principale : la reconnaissance de ME. Ainsi, la plupart des méthodes traitant de la reconnaissance des ME se basent sur le postulat d'une connaissance exacte de la localisation de ces ME (soit l'apex, soit l'onset – début des ME , soit les deux).

Cela a du sens dans une certaine mesure car pour évaluer les performances d'une méthode il faut ne retenir que son influence. En effet les erreurs résultent de la méthode et non des données utilisées. Cependant, comme nous le développeront dans le chapitre 5, la reconnaissance des ME est fortement dépendante du spotting.

Nous ne parlerons ici que de la seule étape de la reconnaissance, dissociée du reste du processus.

Comme pour le spotting, et d'ailleurs de façon plus générale pour la reconnaissance de la forme, les méthodes développées se sont dans un premier temps basées sur l'association du descripteur et du classifieur avant d'appliquer massivement l'apprentissage profond. Notre présentation ci-dessous suit cet ordre chronologique.

2.3.1 Les descripteurs de ME

Les caractéristiques (descripteurs) couramment utilisées dans la reconnaissance des micro-expressions seront discutées dans cette sous-section. Il est à noter que le 3DHOG fut le premier à être utilisé pour décrire une ME avant de perdre en popularité au profit du HOOF et du LBP_TOP. Ces dernières années le LBP_TOP était devenu le descripteur le plus courant avant l'apparition des méthodes d'apprentissage profond.

3D Histograms of Oriented Gradients (3DHOG)

Partons d'une séquence vidéo considérée comme un objet 3D, un cube spatio-temporel. Le 3DHOG est une description du contenu de cette séquence qui extrait la répartition, sur l'ensemble des pixels, de l'orientation du gradient. La forme mais aussi sa déformation dans le temps (accélération du mouvement) sont ainsi décrits. Polikovsky et al. [37] sont les premiers à appliquer ce descripteur pour les ME. Le visage est d'abord divisé en douze ROIs sur chacune desquels un 3DHOG est calculé (Figure 2.2). Le descripteur final, résultat de la concaténation des 3DHOG sur l'ensemble des ROIs est transmis à une classification à partir d'un k-means. Les auteurs ont perfectionné la représentation dans l'espace dans [38] avec en plus une procédure de vote pour la classification. Chen et al [39] obtiennent des performences plus élevées en intégrant un système de pondération des caractéristiques de l'histogramme isolant la réelle contribution de chacune d'elles.

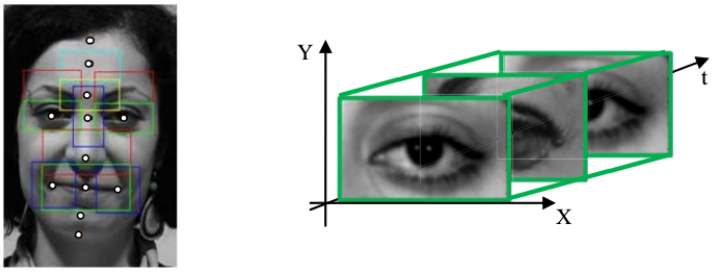

FIGURE 2.2 – Les 3DHOG dans [37] : plusieurs parties du visage sont détectées (à gauche) ; ensuite une représentation spatio-temporelle (cube) est extraite sur chaque partie (à droite) sur laquelle un histogramme de gradient orienté 3D est calculé.

Histogram of Oriented Optical Flow (HOOF)

Pour réduire la complexité et ainsi simplifier l'apprentissage, il peut être avantageux de prendre juste deux instants (état de repos et activation de l'expression) et d'évaluer le mouvement entre ces deux instants plutôt que de prendre en compte toute son évolution au cours du temps. Le flot optique est une représentation efficace du mouvement entre deux instants. Le HOOF [40] permet de décrire spatialement le flot optique et est un descripteur fréquemment utilisé pour la reconnaissance de ME.

Liu et al. [41] ont proposé un descripteur nommé Main Directional Mean Optical-flow (MDMO). Contrairement aux HOOF, ce descripteur ne garde que l'orientation principale pour réduire la dimensionalité et améliorer l'effet de l'apprentissage. Happy et al. [42] utilisent le Fuzzy Histogram of Optical Flow Orientations (FHOFO), un dérivé du HOOF où la magnitude du flot optique n'est pas prise en compte afin d'être plus robuste aux variations d'intensité de l'expression. Mais cette méthode donne un descripteur de haute dimensionalité. La méthode est alors améliorée dans [43] à partir d'une sélection des caractéristiques par proximité des caractéristiques par paire.

Le LBP_TOP et ses variations

Le LBP est un descripteur de texture qui se base sur l'évolution des nuances de gris dans un voisinage spatial. Le LBP_TOP [44] est son adaptation au format vidéo à partir de la concaténation de son résultat selon les trois plans spatio-temporels. Pfister et al. [45] ont été les premiers à les utiliser pour la reconnaissance des ME. Le modèle d'interpolation temporelle (TIM) a été utilisé pour augmenter le nombre d'images afin d'obtenir des histogrammes statistiquement plus stables. Dans le but de différencier les expressions faciales spontanées de celles posées, ils proposent dans [46] le CLBP_TOP qui est la version dynamique du CLBP [47] (lui même une adaptation du LBP où le signe et la magnitude du motif sont pris en compte). Le principe peut ensuite être optimisé. Guo et al. [48] associent le LBP_TOP au kNN pour interpréter l'influence de chaque plan et trouver la combinaison du nombre de points voisins à prendre sur chaque plan qui maximise l'efficacité de la méthode. Zhang et al.[49] combinent le LBP_TOP avec le flot optique. Le traitement se réalise sur des zones spécifiques (ROI) plus descriptives. Dans le même ordre d'idée, Duan et al. [50] ont extrait le LBP-TOP uniquement sur la région des yeux, la plus représentative du comportement recherché. La qualité des données de départ limitent les performances. Pour renforcer leur descriptivité, le STLBP-IP [51] se base sur la projection intégrale. Pour amplifier le mouvement caractéristique, Talukder et al. [52] magnifient préalablement le mouvement, tandis que Wang et al. [53] réalisent en pré-traitement un agrandissement vidéo eulérien. Même sans parler de l'apprentissage profond, la quantité de données disponibles est un frein. Il est alors courant d'utiliser plusieurs bases de données. Pour résoudre le problème des variations entre bases de données, Zong et al. [54] ont proposé de régénérer l'échantillon cible dans le processus de reconnaissance afin d'avoir des distributions de caractéristiques similaires à celles de l'échantillon source.

2.3.2 L'apprentissage profond pour définir la nature des ME

Motivé par la robustesse des réseaux de neurones résiduelles (principalement Resnet), plusieurs études les ont utilisé pour aborder le problème de la reconnaissance de ME [55, 56, 57, 58]. Le modele InceptionNet [59] développé par les équipes de google a aussi été assez utilisé [60] avec des variantes qui en ont découlé [61, 62, 63].

Du point de vue de l'architecture, la grande majorité des études applique des réseaux de neurones à convolution 2D. Cependant certaines études ont aussi recours à des réseaux 3D comme les modèles GAM [64], MERANet [65] et CBAMNet [66]. Cette approche traite simultanément un nombre plus important de donnée, mais cela engendre aussi des besoins accrus en termes de temps et de puissance nécessaires pour la classification. Nous pouvons également citer des approches utilisant des réseaux de neurones convolutifs récurent (Recurrent Convolutional Network -RCN) [67, 68] et des CapsuleNets [69, 70].

Le concept MSN (Multiple Stream Networks) a aussi été adapté à l'analyse des ME. L'idée est de fusionner les caractéristiques extraites par plusieurs réseaux avant de passer à la classification. La façon la plus directe est d'utiliser deux fois la même architecture de CNN avec des données d'entrée différentes. Les réseaux OFF-ApexNet (Optical Flow Features from Apex frame Network) [71] et DSSN (Dual-Stream Shallow Network) [72] sont des CNNs à double flux alimentés par le flux optique extrait entre Onset et Apex de ME. Ensuite, Liong et al. [73] ont étendu OFF-ApexNet à plusieurs flux avec différentes composantes de flux optique comme données d'entrée. [74, 75] ont conçu des modèles CNN à trois flux avec trois types d'entrées. Plus précisément, la première approche utilise l'image d'apex, le flux optique et l'image d'apex masquée par le seuil de flux optique, tandis que la seconde utilise l'image d'apex, le flux optique entre Onset et Apex et l'image de décalage pour étudier les informations spatio-temporelles. En outre, She et al. [76] ont proposé un modèle à quatre flux considérant trois ROIs et la région globale pour explorer simultanément les informations locales et globales. En plus des CNNs 2D à flux multiples, CNN3D comme 3DFCNN [77], SETFNet [78] et [79] ont aussi été développés pour la reconnaissance de ME.

Les MSN ci-dessus permettent de traiter séparément les différents types de données avec la même architecture du CNN. Pour améliorer la représentation des caractéristiques des ME, certains travaux [76, 64, 80, 81, 20] ont étudié la combinaison de différentes convolutions. Liong et al. ont conçu un CNN tridimensionnel à triple flux peu profond adoptant plusieurs CNN 2D avec différents noyaux. Lo et al. [20] ont construit un réseau nommé AffectiveNet à quatre chemins avec quatre champs réceptifs différents afin d'obtenir des caractéristiques multi-échelles pour mieux décrire les ME. D'autre part, Kumar and Bhanu [82] avec leur LR-GACNN (Landmark Relations with Graph Attention Convolutional Network) et Lo et al. [83] avec leur MER-GCN (Micro-Expression Recognition Graph Convolutional Network)

ont établi des réseaux de graphes à deux flux pour explorer les relations entre les points de repère et les patchs locaux, ainsi que celles entre les AUs et la séquence. En outre, [64, 81] ont intégré un réseau CNN 2D et un réseau CNN 3D pour extraire des informations spatiotemporelles.

Depuis quelques temps, le *Transfert Learning* est largement généralisé dans le domaine de l'apprentissage profond. Il consiste à entraîner les CNNs sur une autre tâche pour laquelle il existe une large quantité de données avant d'effectuer l'apprentissage spécifique sur la tâche à accomplir. L'approche d'apprentissage par transfert pour la reconnaissance de ME consiste à affiner la spécificité aux ensembles de données de ME à partir de modèles pré-entraînés, la plupart du temps sur ImageNet [84, 66, 75] ou bien sur des macro-expressions faciales [85, 86, 87, 88]. Une forme plus poussée de Transfer Learning fait appel à la DA (Domain Adaptation) [89, 87].

Pour que l'apprentissage profond soit efficace, une grande quantité de données disponibles est primordiale. La technique du *data augmentation* est incontournable lorsque les bases de données sont de taille relativement réduites par rapport à la tâche et au nombre de classes que l'on veut différentier. C'est bien le cas pour les bases de données dédiées à la classification de ME. Plusieurs études [90, 91, 92] ont donc exploré les possibilités de data augmentation pour renforcer et rendre plus robuste les modèles de classification.

La magnification du mouvement offre des possibilités pour augmenter les données d'entraînement, en plus de rendre les ME plus faciles à discerner. Ainsi, Xia et al. [67] ont pu multiplier le nombre d'exemples de leur base de données en appliquant différents niveaux de magnification. Le GAN (Generative Adversarial Network) [93] est aussi une piste bien explorée par les chercheurs en Deep Learning. Dans le domaine des ME, Xie et al. [94] ont utilisé leur GAN contrôlable par l'intensité de l'AU (ou AU-ICGAN pour AU Intensity Controllable GAN) pour créer des ME synthétisées à partir de données d'entraînement. Liong et al. [73] ont utilisé leur GAN conditionnel pour synthétiser des flots optiques de ME.

2.4 Bases de données de micro-expressions

2.4.1 Bases de données de ME non spontanées

Avant la création et la mise en circulation de bases de données composées de vraies ME, la communauté a d'abord travaillé avec des bases de données utilisant des micro-expressions dites "posées" où les sujets essayaient de simuler les ME. Parmi ces bases de donnés de ME non spontanées, nous pouvons citer :

La base de données USF-HD [95, 96] est constituée de 47 séquences et contient 181 macro-expressions (sourire, surprise, colère, tristesse) et 100 micro-expressions.

Les vidéos ont été collectées soit par un caméscope JVC-HD100, soit par un caméscope Panasonic AG-HMC40, à une résolution de 1280×720 et avec une fréquence d'images de 29,7 frames per second (fps). La durée de chaque vidéo est en moyenne d'environ 1 minute. Pour capturer et enregistrer les micro-expressions, il a été demandé aux sujets d'imiter les exemples de ME. Les constructeurs de cette base de données ne divulguent pas le nombre de sujets, ni leur âge, ni leur sexe, ni leur origine ethnique. Elle n'est pas accessible au public.

Polikovsky [37, 97] propose une base de données constituée de 42 échantillons de ME (contenant les 6 émotions de base en plus du mépris). Les vidéos ont été collectées par une caméra Point Grey Grasshopper à une résolution de 640×480 et une fréquence d'images de 200 fps. 10 étudiants de différentes ethnies ont été recrutés comme sujets. Les auteurs de cet base ne divulguent pas le sexe des participants. Les participants devaient d'abord exprimer les émotions de base avec une faible intensité des muscles faciaux et ensuite revenir à l'expression neutre du visage aussi vite que possible, en simulant le mouvement d'une ME. Cette base de données n'est pas accessible non plus au public.

SFED2007 [98] (The Subtle Facial Expression Database) présente 20 sujets exécutant 4 expressions faciales (neutre, sourire subtil, surprise subtile et colère subtile). Tous les sujets sont asiatiques (les auteurs de ce jeu de données ne divulguent ni le sexe, ni l'âge des participants). La taille de l'image est de 640×480. Il n'y a pas d'information disponible sur les appareils ou la procédure de capture des données. Cette base de données est disponible sur la page Web du laboratoire des médias intelligents de l'Université des sciences et technologies de Pohang (POSTECH).

2.4.2 Bases de données de ME spontanées

Depuis une dizaine d'année, des bases de données de ME spontanées ont été créées en conditions réelles et mise en circulation permettant de réaliser le spotting et la reconnaissance de ME.

SMIC (Spontaneous Micro-expression database) [45] consiste en trois ensembles de données contenant 164 échantillons de ME étiquetés selon trois types d'émotion (positive, négative et surprise). Les vidéos ont été collectées par une caméra PixeLINK PL-B774U Highspeed avec une résolution de 640×480 et une fréquence d'images de 100 fps mais aussi par une caméra visuelle normale et une caméra proche infrarouge, toutes les deux avec 25 fps et une résolution de 640×480. 16 sujets de différents âges, sexes et ethnies ont été recrutés comme participants. Les sujets ont été invités à regarder des vidéos émotionnelles (Figure 2.3) tout en cachant leurs véritables émotions (sous peine de devoir remplir un questionnaire long et ennuyeux). Cette base de données est disponible sur la page web du centre de recherche Center for Machine Vision and Signal Analysis (CMVS) de l'Université d'Oulu.

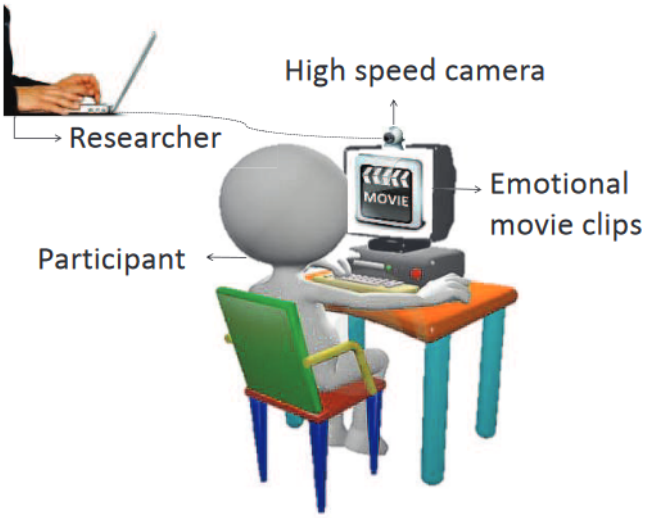

FIGURE 2.3 – Protocole d'acquisition pour la base de donnée SMIC [45].

CASME [99] (The Chinese Academy of Sciences Micro-expression database) comprend 195 échantillons de ME étiquetés comme les sept émotions de base. Les vidéos ont été collectées avec deux caméras différentes : une caméra BenQ M31 avec une résolution de 1280×720 et une fréquence d'images de 60 fps ; et une autre caméra Point Grey GRAS-03K2C avec une résolution de 640×480 et une fréquence d'images de 60 fps. 19 sujets de différents âges et sexes (tous les sujets ont la même origine ethnique) ont participé à la collecte de données. Il a été demandé aux sujets de regarder des vidéos émotionnelles tout en cachant leurs véritables émotions [1].

CASME II [100] est une version améliorée de la base de données CASME. Elle se compose de 247 échantillons de ME étiquetées comme 5 types d'émotions (joie, dégoût, surprise, répression et autres). Les vidéos ont été collectées par une caméra Point Grey GRAS-03K2C avec une résolution de 640×480 et une fréquence d'images de 200 fps. 26 sujets d'âge et de sexe différents (tous les sujets ont la même origine ethnique) ont été recrutés comme participants. Il a été demandé à certains participants de garder un visage neutre lorsqu'ils regardaient des clips vidéo. D'autres devaient seulement essayer de supprimer les mouvements du visage lorsqu'ils se rendaient compte de l'apparition d'une expression faciale.

CAS(ME)^2 [101] est constituée de longues vidéos contenant 287 échantillons de macro-expressions et 54 échantillons de ME étiquetées comme quatre types d'émotions (positive (111), négative (125), surprise (23) et autres (82)). Les vidéos ont été recueillies par une caméra Logitech Pro C920 à une résolution de 640×480 et à une fréquence d'images de 30 fps. 22 sujets ont été enregistrés. Les participants ont été invités à supprimer leurs expressions au mieux de leurs capacités [1].

SAMM (Spontaneous Micro-Facial Movement dataset) [46] contient 159 échantillons de ME étiquetées comme les sept émotions de base. Les vidéos ont été collectées par une caméra LBasler Ace acA2000-340km à une résolution de 2040×1088 et une fréquence d'images de 200 fps. Les auteurs ont voulu créer un ensemble de données très diversifié, ainsi, 30 participants de différents âges (de 19 à 57 ans), sexes et ethnies (13) ont été recrutés. Les participants ont été invités à supprimer leurs expressions au mieux de leurs capacités [1]. Cette base de données peut être téléchargée à partir de l'article [46].

Le Tableau 2.6 résume les caractéristiques techniques des différentes bases de données de ME spontanées.

Base de donnée	Participants	Résolution	FPS	Nb. Classes	AU
SMIC	20	640×480	100, 25	3	non
SAMM	32	2040×1088	200	7	oui
CASME	35	640×480 1280×720	60	7	oui
CASME II	35	640×480	200	5	oui
CAS(ME)2	22	640×480	30	4	oui

TABLE 2.6 – Caractéristiques des bases de données de ME spontanées avec notamment le nombre de classes (Nb. Classes), la fréquence d'acquisition (FPS) et la disponibilité des AUs dans l'annotation.

2.4.3 Bases de données composites

Même sans évoquer l'apprentissage profond, la quantité de données disponibles de ME est un frein. Il est alors courant d'utiliser plusieurs bases de données. Cela représente plusieurs avantages. Par exemple des modèles construits avec des images venant de plusieurs bases de données différentes sont plus robustes au changement des conditions de saisies. De plus les résultats obtenus sur des grandes plages de données sont statistiquement plus crédibles.

Pour former des bases de données composites, nous pouvons discerner 2 combinaisons différentes utilisées lors des MEGC (Facial Micro-Expression Grand Challenge) de 2018 et 2019. MEGC2018 combine Casme II et SAMM. Ces deux bases de données furent choisies car bien que les annotations de ces bases ne soient pas compatibles, elles fournissent toutes les deux les AUs observées sur chaque ME. Les dites AUs ont permis de donner une vérité terrain unifiée pour les bases de données, classifiant chaque ME selon l'une des 5 émotions basiques : joie, tristesse, peur, surprise et colère. Les deux bases de données ont aussi la même cadence de

1. Les participants ont été informés que leurs récompenses monétaires seraient réduites s'ils produisaient une expression perceptible.

capture (200 fps).

MEGC2019 combine CASME II, SAMM, et SMIC. Au contraire de SAMM et CASME II fournissant les AUs de chaque ME en plus de leur propre étiquetage, SMIC ne fournit que des ME étiquetées selon 3 types d'émotion (positive, négative et surprise). Pour unifier les annotations, les classes de CASME II et SAMM sont ajustées à celles de SMIC pour avoir uniquement les 3 classes (positive, négative, et surprise). La cadence de la base SMIC est de 25 images par seconde.

Pour résoudre le problème des variations entre bases de données, Zong et al. [54] ont d'abord proposé de régénérer l'échantillon cible dans le processus de reconnaissance afin d'avoir des distributions de caractéristiques similaires à celles de l'échantillon source. Ensuite ils ont amélioré leur travail dans [102] pour des distributions de caractéristiques plus similaires avec une génération à partir de la source et de la cible.

2.5 Protocole d'évaluation

Il existe plusieurs protocoles de validations appliqués dans le domaine de la reconnaissance de formes. Quand la quantité de données est assez conséquente, les bases de données sont généralement divisées en une partie dédiée à l'entraînement, et une autre partie réservée au test. Cela a pour avantage d'assurer aux chercheurs travaillant sur ces bases de données qu'ils effectuent tous leurs tests sur les mêmes données, facilitant ainsi la comparaison entre les différentes études.

Quand la quantité de données est limitée, et/ou que l'on veut maximiser l'utilisation des données, les chercheurs ont recours à l'une des nombreuses variantes de la cross-validation. La k-fold cross-validation est la méthode la plus utilisée par les études dans le domaine d'analyse de ME. Elle divise la base de données en k parts égales. S'en suit alors k apprentissages avec à chaque fois l'une des parts de la base de données utilisée comme jeu de test et le reste comme jeu d'entraînement. Une version encore plus populaire est la Leave One Out cross-validation : celle-ci préconise de prendre un seul échantillon comme jeu de test à chaque fois. Cette méthode permet les jeux d'entraînement les plus conséquents, mais demande aussi un volume de calcul beaucoup plus élevé.

2.6 Conclusion

Si la reconnaissance et la classification des macro-expressions font l'objet de recherches scientifiques dans le domaine de l'intelligence artificielle depuis une dizaine d'années, l'étude sur les micro-expressions est très récente et de nombreuses questions restent ouvertes. Grâce à l'avancement de la technologie et notamment avec l'intelligence artificielle par la vision, il serait alors possible d'apporter des

éléments de réponses concernant l'étude de ces micro-expressions.

Il est à noter que plusieurs facteurs ont participé aux gains en popularité du domaine comme la mise en circulation de bases de données des ME posées, puis celles avec des ME spontanées. Ensuite la mise en place des MEGC ont aussi donné une plus grande visibilité de la communauté. Les MEGC de 2020 et surtout celle de 2021 ont bien guidé les chercheurs à s'intéresser à la tâche difficile du spotting de ME qui avait été délaissée au profit de la reconnaissance de ME jusque là.

Malgré des études récentes sur l'analyse de ME utilisant l'apprentissage profond, il n'existe pas encore un système satisfaisant vis à vis de multiples contraintes d'applications grand publique : robustesse, faible complexité, transportabilité, basse consommation d'énergie, etc.

Chapitre 3

Reconnaissance de ME utilisant des descripteurs

Lors du début de cette thèse en 2018, la problématique de la reconnaissance de ME s'est principalement concentrée sur l'association du descripteur et du classifieur. Le descripteur a pour but de représenter une donnée dans un espace qui mettra bien en exergue les caractéristiques les plus descriptives des catégories à classifier. Le descripteur transforme aussi la donnée de départ de haute dimensionalité en un vecteur de caractéristiques de dimension plus petite.

L'objectif est de produire un vecteur de caractéristiques de petite dimension pour réduire au maximum les informations redondantes ou non pertinentes. Un bon descripteur permet d'extraire des caractéristiques discriminantes maximisant la distance interclasse et minimisant celle entre les échantillons de la même classe. L'établissement du descripteur se base sur des connaissances a-priori fortes des classes. Les caractéristiques sont sélectionnées en fonction des besoins et les descripteurs peuvent se concentrer sur les formes, les textures ou les couleurs.

Lors de l'utilisation d'un descripteur d'information locale, l'image d'origine est souvent divisée en plusieurs zones avant la concaténation en un seul et unique vecteur des vecteurs de caractéristiques extraits. La Figure 3.1 représente le protocole général.

Ici, le classifieur prend en entrée les caractéristiques définies par le descripteur et décide la classe à associée à un échantillon. La procédure classique fonctionne en deux phases :

1. Lors de la phase d'apprentissage, le modèle du classifieur est entraîné. Tout d'abord, le descripteur produit un vecteur de caractéristiques pour les échantillons d'une base d'apprentissage annotée (la classe de chaque échantillon est connue). Puis le classifieur construit un modèle permettant de séparer au mieux les éléments de chaque classe.
2. Lors de la phase de test, un nouvel échantillon est testé. Le descripteur calcule

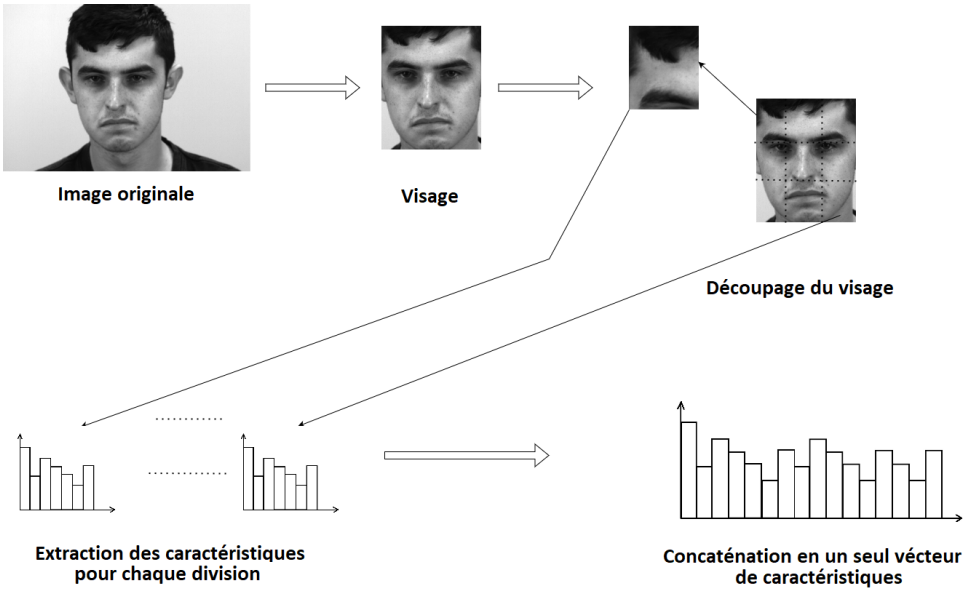

FIGURE 3.1 – Le protocole suivi lors de l'extraction de caractéristiques.

son vecteur de caractéristiques. Puis ce vecteur est comparé à ceux de la base d'apprentissage par le biais du modèle du classifieur. La décision de sa classe est alors prise.

Par exemple, avec les SVM (Support Vector Machine) [103], un hyperplan (séparateur) est établi pour séparer les échantillons de différentes classes en maximisant la marge entre l'hyperplan et les échantillons les plus proches. Ensuite, selon le côté de l'hyperplan où se trouve l'échantillon de test, il sera assigné à une classe ou à une autre. Le descripteur et le classifieur sont souvent définis indépendamment. Le descripteur intègre la connaissance sur les classes et le classifieur est choisi pour optimiser les résultats obtenus.

Dans ce chapitre, nous étudierons les méthodes basées sur les descripteurs pour reconnaître les ME. La section 3.1 présente d'abord les descripteurs les plus couramment utilisés dans la littérature avant d'effectuer une évaluation et comparaison expérimentale. La section 3.2 introduit nos contributions dans le domaine. Nous reprenons d'abord le descripteur le plus reconnu, le LBP_TOP, pour l'adapter à une meilleure cohérence temporelle lors de la phase de reconnaissance. Puis nous présentons une méthode originale de comparaison de gabarits. Cette méthode, très simple et économe, se base sur une observation faite à partir d'une manipulation des bases de données des ME. La section 3.4 tire finalement un bilan de toutes ces études.

3.1 Description d'une micro-expression

Dans cette section, nous présentons les descripteurs basiques pour représenter des ME. Il s'agit d'isoler de très faibles mouvements localisés au niveau de zones

spécifiques du visage (les Action Units par exemple). Le LBP (section 2.3.1) est un descripteur classique de reconnaissance de forme. Les LBP_TOP [44] sont plus utilisés dans notre domaine d'étude, car ils considèrent la déformation des formes. Les HOOF (section 2.3.1) permettent quant à eux de représenter l'orientation du mouvement localisé. Ces descripteurs sont représentatifs des différentes familles de description des ME.

3.1.1 LBP (Local Binary Pattern)

Le LBP est en un descripteur visuel représentant la variation de la forme dans un voisinage local. Créé initialement pour décrire les textures [104], il a ensuite été utilisé pour décrire les visage [105, 106] puis pour la reconnaissance des ME [32]. Nous le décrivons ci-dessous.

Soit un pixel p. Un voisinage est défini autour de p en sélectionnant régulièrement N_v pixels sur un cercle de rayon r_v centré sur p. Nommons $V_p(i)$ le $i^{\text{ième}}$ pixel correspondant à ce voisinage. Chacun de ces pixels génère une valeur qui vaut 1 si la valeur liée à ce pixel est supérieure à celle de p, et 0 dans le cas contraire. Ces N_v valeurs forment un code binaire qui peut être convertie en une valeur décimale par l'équation suivante :

$$LBP_{N_v,r_v}(p) = \sum_{k=0}^{N-1} \mathcal{T}\big(g(V_p(k)) - g(p) > 0\big)\, 2^k \tag{3.1}$$

Ici $g(p)$ correspond à la nuance de gris du pixel p et $\mathcal{T}$ est la fonction de seuillage suivante :

$$\mathcal{T}(X) = \begin{cases} 1, & \text{si X est vraie} \\ 0, & \text{sinon} \end{cases} \tag{3.2}$$

L'image de départ est divisée spatialement en cellules. Pour chaque cellule, un histogramme d'occurrence de tous les codes produits est ensuite calculé :

$$H(b) = \sum_{p} \mathcal{T}\big(LBP_{N_v,r_v}(p) = b\big) \tag{3.3}$$

Le descripteur LBP lié à une image est alors la concaténation des histogrammes normalisés de toutes les cellules.

Nous pouvons considérer les LBP comme des primitives de texture qui comprennent différents types d'arêtes, de courbes, de tâches, de zones plates, etc. Le nombre de cellules et la taille de chaque cellule déterminent le niveau d'information spatiale retenu. Une fois que les histogrammes d'occurrence LBP basés sur les blocs sont concaténés, la classification est effectuée en calculant les similarités entre eux.

Le descripteur LBP permet de décrire les textures contenues dans une image. Le découpage en cellule permet d'introduire une représentation spatiale. Pour interpréter le mouvement, qui est la base des ME, le descripteur obtenu au niveau de l'apex est comparé à celui de l'onset. L'idée est alors d'extraire les zones de déformation et donc de reconnaître les AUs activées.

3.1.2 LBP_TOP

Le LBP_TOP est une variante du LBP prenant en compte le mouvement temporel. La représentation de la texture n'est plus que spatiale mais spatio-temporelle. La séquence vidéo est considérée comme un objet tri-dimensionnel. Trois coupes sont alors réalisées selon les axes XY (spatial), XT et YT comme sur la Figure 3.2. Chaque coupe donne un histogramme et la concaténation des trois histogrammes constitue le descripteur LBP_TOP.

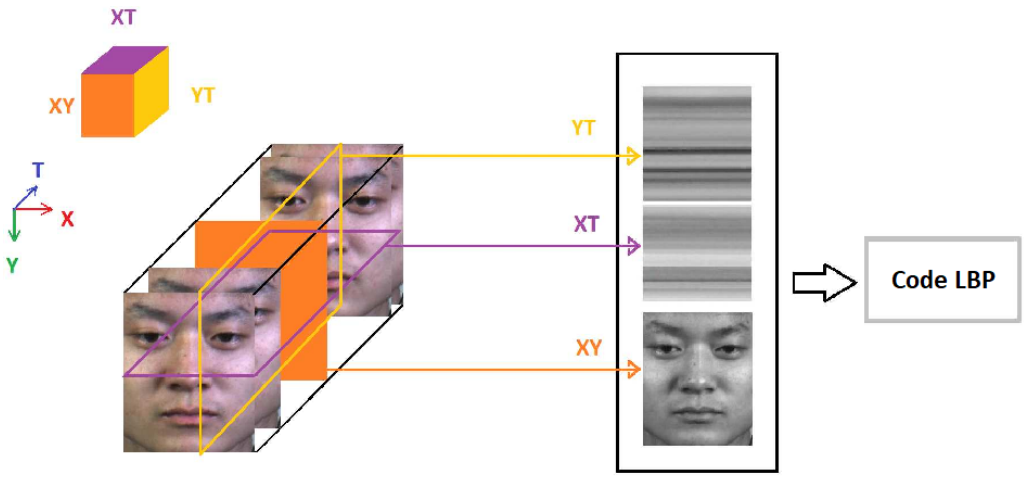

FIGURE 3.2 – Découpage spatio-temporel du LBP_TOP.

Bien que plus complexe et de plus haute dimensionalité, ce descripteur est plus riche puisqu'il décrit la déformation au cours du temps. Après avoir été proposé à l'origine pour la description de textures dynamiques [44], il a été utilisé pour la première fois par Pfister et al.[107] pour la reconnaissance de ME. Il est à ce jour considéré comme le descripteur de référence dans le domaine pour ses hautes performances.

3.1.3 HOOF

Le flot optique [108] entre deux frames donne une représentation dense du mouvement. Chaque pixel est associé à un vecteur de déplacement. Le flot optique entre l'onset et l'apex produit une représentation riche de l'ensemble du mouvement

présent entre ces deux instants.

L'histogramme de gradients orientés (Histogram of Oriented Gradients – HOG) [109] est un descripteur de forme à partir d'une image 3D. Il a été créé pour la reconnaissance de personne avant de devenir une méthode populaire et d'être utilisé dans bien d'autres domaines.

Le HOOF (Histogramm of Oriented Optical Flow) [110] est un descripteur associant le HOG au flot optique. L'image de départ est découpée en cellule et un algorithme de calcul du flot optique lui est appliqué. Pour chaque pixel, le flot optique représente un vecteur pouvant être défini par une orientation et une norme. L'espace des orientations est découpé en intervalles réguliers (bin). Dans chaque cellule, l'histogramme des vecteurs de flot optique selon les différents bins d'orientation est calculé. Le descripteur HOOF est finalement la concaténation de l'ensemble de ces histogrammes normalisés.

L'avantage de ce descripteur est qu'il utilise des caractérisations du mouvement plutôt que de déformation. L'association du flot optique et des HOGs permet de combiner la fine représentation du flot optique avec l'interprétation condensée du HOG.

3.1.4 Reconnaissance de micro-expressions en utilisant des descripteurs basiques

Avant d'aller plus loin, nous voulons d'abord évaluer ces trois descripteurs pour la reconnaissance de micro-expressions. Nous décrivons le protocole utilisé ci-dessous :

— Nous avons utilisé CAS(ME)^2 comme base de données. Elle contient à la fois des ME et de très courtes macro-expressions annotées selon quatre classes : émotion positive (111 échantillons), émotion négative (125 échantillons), surprise (23 échantillons) ou autre (82 échantillons).
— Pour la classification, nous utilisons un SVM avec le même schéma d'optimisation automatique des hyperparamètres pour les différents descripteurs.
— Nous avons choisi le protocole LOSO pour la validation.

La Figure 3.3 montre la matrice de confusion obtenue en associant un descripteur LBP à un classifieur SVM. Le LBP a été appliqué à l'image de l'apex où la micro-expression est plus prononcée. Nous pouvons observer que le taux de reconnaissance correcte (accuracy) est de 43.30%. Cela nous pousse à croire que bien que les descripteurs de texture 2D ne soient pas optimaux pour les ME, ils contiennent quand même quelques informations descriptives. Cependant, une meilleure représentation est nécessaire.

La Figure 3.4 affiche les résultats obtenus avec le descripteur LBP_TOP et un SVM. Ici, le score est de 57.48%, nettement supérieure au LBP. Cela est dû

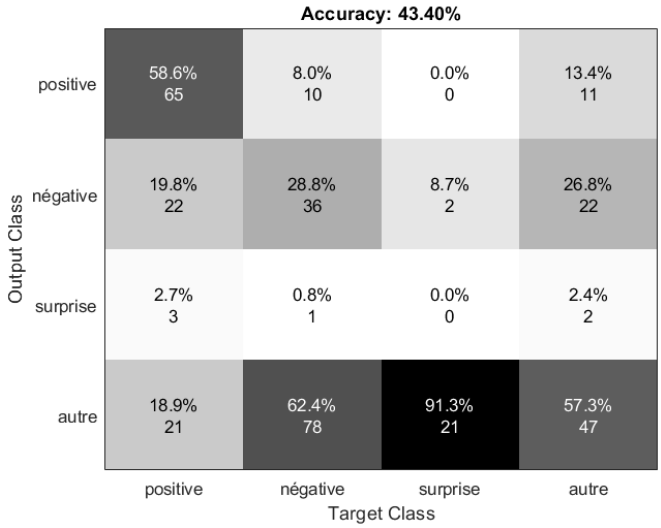

FIGURE 3.3 – Matrices de confusion obtenues à partir de la méthode des LBP/SVM.

à la nature spatiotemporelle du descripteur, qui encrypte des informations sur 3 dimensions : les axes X et Y pour représenter le visage dans l'espace et l'axe T qui capte les mouvements des muscules du visage. D'ailleurs, contrairement au LBP qui ne s'effectue que sur une seule image (l'apex), le LBP_TOP calcule la variation des pixels sur toute la durée de la ME.

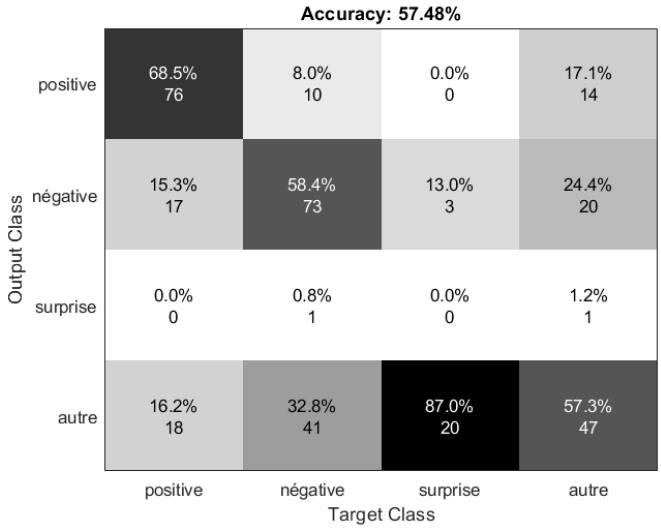

FIGURE 3.4 – Matrices de confusion obtenues à partir de la méthode des LBP_TOP/SVM.

La Figure 3.5 montre la matrice de confusion obtenue en associant un descripteur HOOF à un classifieur SVM. La performance de reconnaissance est de 53.80%. Un score légèrement moins bon que celui atteint par le LBP_TOP. Ce descripteur a cependant l'avantage de mieux détecter la surprise. Cela dit, sa tendance à donner de faux positifs pour cette classe est elle aussi assez grande.

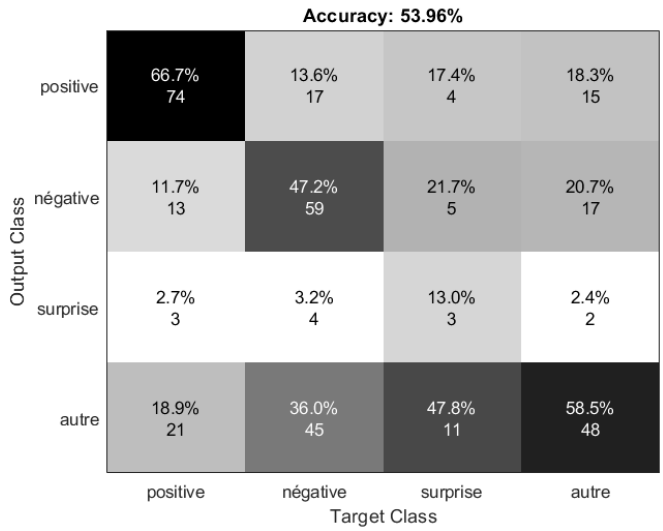

FIGURE 3.5 – Matrices de confusion obtenues à partir de la méthode des HOOF/SVM.

Parmi ces trois descripteurs, le LBP a le score le moins élevé pour la reconnaissance de ME. Cela nous pousse à conclure que le LBP, étant un descripteur de textures et de formes locales d'une seule image, n'est pas le plus adapté à cette tâche. Le HOOF, quant à lui, obtient une performance nettement plus élevée. Cela peut principalement être attribué au fait que ce descripteur se base sur le flot optique, qui quantifie le mouvement entre l'onset et l'apex de la ME.

Le LBP_TOP obtient le score le plus élevé pour la reconnaissance de ME. Parmi les trois descripteurs, il encode le plus d'information sur les ME. Il décrit la ME dans son entièreté (toute la durée), de l'onset à l'offset, en passant par l'apex, en encodant les caractéristiques spatiotemporelles présentes dans toutes les images.

Comme nous pouvons le voir, la reconnaissance automatique de micro-expressions est une problématique récente et difficile. S'il est compliqué, même pour un être humain, de repérer une micro-expression et de l'associer à une émotion à l'œil nue, c'est encore plus difficile de trouver les caractéristiques pour les représenter. Cependant, comme les performances obtenues sont relativement base, la marge de progression au niveau des descripteurs est bien présente.

Nous proposons par la suite deux contributions à la description des ME. Premièrement, nous reprenons le descripteur donnant les meilleurs résultats, c'est à dire le LBP_TOP, et nous appliquons une unification temporelle pour améliorer la performance de reconnaissance. Deuxièmement, nous introduisons une corrélation des images de flot optique avec des gabarits représentatifs de chaque type d'expression dans un objectif de simplification de la complexité et d'optimisation du temps de traitement.

3.2 Unification temporelle des LBP_TOP

Comme vu précédemment, le LBP_TOP est le descripteur le plus efficace pour la reconnaissance de ME. Pour une projection suivant les directions données par les trois plans XT, YT et XY, les séries d'images sont considérées comme une matrice tridimensionnelle. Les composantes spatiales (X et Y) ne varient pas entre les séquences de ME : les frames sont recadrées autour du visage et redimensionnées pour toujours avoir la même taille. Ce n'est pas le cas, pour la composante temporelle T. Chaque expression a une durée propre. Par exemple sur la base de données $CAS(ME)^2$, la durée des ME (différence entre l'onset et l'offset) va de 4 frames à 118 frames.

Lors de l'étape de classification, il faut comparer des éléments de même dimension et de même format. Cela revient à calculer la distance entre le vecteur de caractéristique de l'échantillon de test avec ceux des échantillons d'apprentissage. Il n'y a pas ici d'incompatibilité puisque le vecteur de caractéristiques est toujours de même dimension. En effet si l'objet vidéo change de taille entre deux ME, le descripteur prend en compte le même nombre de cellules. Le LBP_TOP utilise des histogrammes, donc ce sont des fréquences d'occurrence des motifs qui sont considérés.

La normalisation de ces histogrammes rend le descripteur stable vis à vis de la durée des ME. Cependant il ne représente pas la même donnée de départ. Selon la durée de la ME une cellule contiendra plus ou moins de pixels. La répartition en valeur est donc statistiquement plus ou moins représentative. Finalement la comparaison reste-t-elle vraiment équitable ?

Nous proposons de combiner le sous-échantillonnage et l'interpolation linéaire afin d'avoir le même nombre d'images sur toutes les expressions faciales présentes dans le jeu de données. Le sous-échantillonnage permet de se débarrasser des informations superflues en n'incorporant pas d'images trop redondantes. Il réduit également le calcul nécessaire à l'extraction des caractéristiques. L'interpolation temporelle augmente les échantillons trop courts. En comparant les performances obtenues avec et sans unification temporelle, nous pourrons conclure sur l'influence de ce format temporel.

3.2.1 Unification temporelle classique

Nous nommerons dans la suite $desired\_size$ le nombre de frame que nous voulons appliquer à l'ensemble des ME et $original\_size$ le nombre de départ de frames d'une séquence de ME. Un sous-échantillonnage est appliqué à la séquence de ME si $original\_size > desired\_size$ et une interpolation temporelle si $original\_size < desired\_size$.

L'interpolation est réalisée de la façon suivante. La position interpolée de la frame d'indice i est obtenu par :

$$p(i) = \frac{(original\_size - 1)i + desired\_size - original\_size}{desired\_size - 1} \tag{3.4}$$

Dans le cas d'un sous-échantillonnage, il suffit alors de prendre la frame juste avant cette instant. Soit $F(i)$ la $i^{\text{ième}}$ frame de la séquence de départ, $F_u(i)$ la $i^{\text{ième}}$ frame de la séquence unifiée et E(x) la partie entière de x ; la séquence unifiée est définie par :

$$F_u(i) = F\Big(E(p(i))\Big) \tag{3.5}$$

Dans le cas d'une interpolation temporelle, nous voulons éviter d'avoir plusieurs fois la même frame dans la séquence. Notons $\Delta^{int}(i) = p(i) - E(p(i))$ la distance à la partie entière. La séquence unifiée est alors donnée par :

$$F_u(i) = \Big(1 - \Delta^{int}(i)\Big).F\Big(E(p(i))\Big) + \Delta^{int}(i).F\Big(E(p(i)) + 1\Big) \tag{3.6}$$

Au niveau du descripteur, le LBP_TOP est configuré à partir des paramètres suivants : un découpage en 5 × 5 cellules, un rayon de 1 pixels, un voisinage de 8 points. Cette configuration a été choisie car elle correspond à celle utilisée par [101]. Le vecteur de caractéristique est finalement de dimension 19200. Le classifieur choisi est un SVM avec un noyau RBF dont les options d'optimisation des hyperparamètres sont réglées sur $expected - improvement - plus$ pour la classification.

Nous testons cette méthode pour différentes valeurs de $desired\_size$ allant de 10 à 60 (Tableau 3.1). Nos tests commencent à la valeur 10 car les paramètres temporels du LBP_TOP (TimeLength et TInterval) étant tous les deux égaux à 4, la longueur minimum que l'on puisse calculer est égale à 10. Le meilleur score est obtenu avec la plus petite valeur d'unification temporelle (10 frames) qui correspond à un taux de reconnaissance correcte de 58.65%. Quand la valeur d'unification égale à 60 frames, le sore baisse à 49.27%.

desired_size	**10**	**20**	**30**	**40**	**50**	**60**
accuracy (%)	58.65	53.95	50.14	52.49	53.96	49.27

TABLE 3.1 – Performances atteintes par le modèle selon la durée choisie pour l'unification temporelle classique.

Pour mieux illustrer l'évolution de la précision du modèle selon le nombre de frames choisie pour l'unification temporelle, nous en traçons la courbe dans la Figure 3.6. Nous observons une claire détérioration des résultats lorsque le nombre de frames augmente.

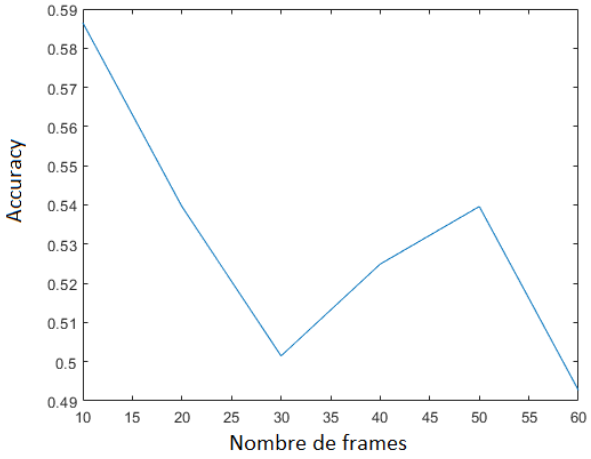

FIGURE 3.6 – Évolution des performances (accuracy) en fonction du nombre de frame des séquences unifiées avec une unification temporelle classique.

L'unification temporelle classique rend les données plus uniformes. Cependant les informations autour de l'apex peuvent échapper à l'échantillonnage. Or c'est bien dans cette région que se concentre les données les plus descriptives d'une ME.

3.2.2 Unification temporelle centrée sur l'apex

Pour ce mode d'unification temporelle, l'attention est tout particulièrement mise sur la conservation de l'information autour de l'apex.

Le sous-échantillonnage se réalise si $original\_size > desired\_size$. Il faut donc sélectionner $desired\_size$ frames parmi les $original\_size$ frames disponibles dans la séquence. Dans un premier temps, les frames sont sélectionnées de façon régulière selon un intervalle correspondant au premier entier supérieur à $\frac{original\_size}{desired\_size}$. La séquence est alors complétée par des frames se trouvant autour de l'apex. En effet, c'est à ce niveau que le mouvement est le plus pertinent car d'intensité la plus élevée. Si $original\_size < desired\_size$, alors toutes les frames de la séquence originale sont gardées. Dans un premier temps nous calculons l'intervalle sur lequel positionner les frames. Celui-ci est égale à $\frac{original\_size}{desired\_size}$. Nous choisissons alors les frames selon cet intervalle. Suite à cela nous sélectionnons les frames autour de l'apex dans la suite de frame originale pour les insérer dans la nouvelle suite de frames jusqu'à atteindre la taille désiré.

Nous utilisons ici le même protocole de test que pour l'unification temporelle classique (c'est-à-dire la base de données $CAS(ME)^2$ avec 4 classes, LBP_TOP avec 5 × 5 cellules et LOSO) pour évaluer les performances de reconnaissance. Nous avons fait varier le nombre de frame désiré d'une ME $desired\_size$ au cours des expériences pour en évaluer l'influence et déterminer empiriquement sa valeur optimale.

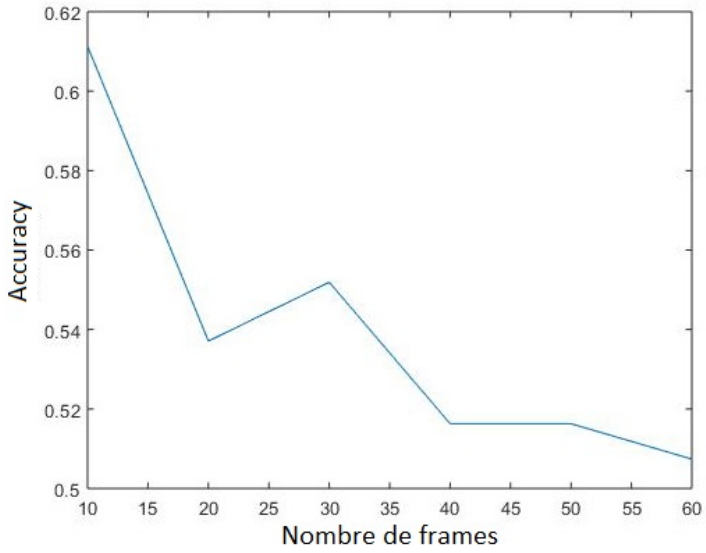

FIGURE 3.7 – Évolution des performances (accuracy) en fonction du nombre de frame des séquences unifiées avec une unification temporelle centrée sur l'apex.

La Figure 3.7 montre l'évolution du critère d'accuracy moyen obtenu avec notre méthode. L'allure générale tend à présenter une dynamique où les performances réduisent avec la valeur de *desired_size*. Pour une valeur de 10 frames (le minimum), l'accuracy est le plus élevé (61,13%).

Dans un second temps, nous comparons les performances obtenues avec et sans l'unification temporelle pour estimer son influence sur la reconnaissance de ME. Les matrices de confusion obtenues à partir de ces deux configurations se trouvent sur la Figure 3.8. Tout d'abord, notons que l'accuracy moyenne augmente avec l'introduction de l'unification temporelle. Elle passe de 57.48% à 61.29%.

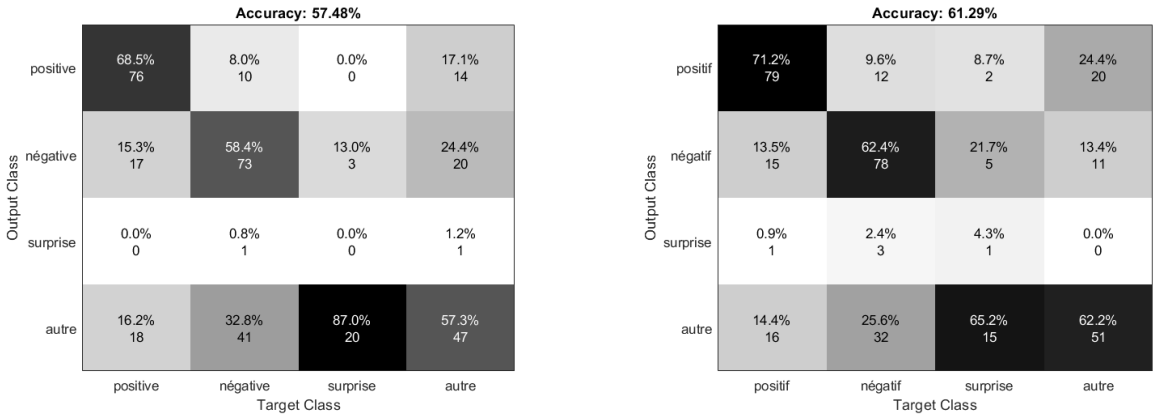

FIGURE 3.8 – Matrices de confusion obtenues par l'association LBP_TOP/SVM sans (à gauche) et avec unification temporelle (à droite).

L'amélioration est présente sur l'ensemble des classes. Nous pouvons donc en conclure que l'unification temporelle exerce une influence réelle sur la reconnais-

sance. Même si la durée des séquences de ME n'influe pas sur la dimension du vecteur de caractéristiques produit par le descripteur, la classification est plus juste avec ce pré-traitement et les performances sont améliorées.

La surprise est toujours mal reconnue même si l'unification temporelle permet certaines améliorations. Cela s'explique par la très faible représentation de cette classe sur l'ensemble de la base de données. En effet la surprise représente seulement 6.74% de $CAS(ME)^2$ contre 32.55% du jeu de données pour positif, 36.66% pour négatif et 24.05% pour autres. Sur ces trois dernières classes, l'accuracy obtenue est assez équilibrée (au moins 62%), ce qui démontre une bonne capacité de séparation. La répartition des classes dans l'apprentissage est importante car elle pondère l'influence de chacune d'elle. Une classe peu représentée donne moins de variation et influe moins dans l'estimation de l'hyperplan séparateur.

3.3 Corrélation avec des gabarits

Nous allons présenter ici une méthode de corrélation pour reconnaître les ME. Mais avant cela, il nous faut décrire un phénomène particulier.

Le flot optique est une représentation dense du mouvement d'une image à l'autre. C'est un outil couramment utilisé dans le domaine de la reconnaissance de ME, car il décrit le mouvement local entre deux instants (par exemple entre l'onset et l'apex). Nous l'avons déjà introduit pour le descripteur HOOF. Le flot optique renvoie un vecteur 2D associé à chaque pixel.

La magnitude correspond à la norme de ce vecteur. Pour visualiser les zones du visage où les mouvements sont les plus courants lors d'une ME, nous avons calculé la valeur moyenne de la magnitude pour tous les types d'expressions sur l'ensemble des échantillons de la base de données $CAS(ME)^2$.

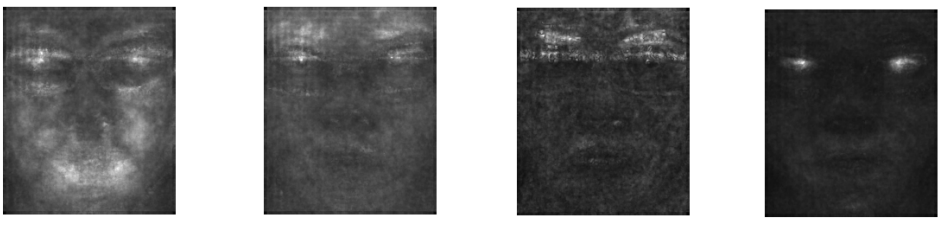

FIGURE 3.9 – Valeur moyenne de la magnitude du flot optique par type d'expression à partir de la base de données $CAS(ME)^2$.

Le résultat est affiché sur la Figure 3.9. On remarque des motifs bien différenciés pour chaque type d'expression, c'est-à-dire qu'il y a des zones du visage significative vis à vis de l'expression exprimée. Par exemple, l'émotion négative s'exprime principalement par des mouvements autour des yeux (les sourcils) ; l'émotion positive engendre de forts mouvements au niveau des lèvres et des joues.

Il s'agit là d'une représentation très intéressante. Nous nous demandons si ces motifs peuvent directement être utilisés comme descripteur de ME ? Sont-ils en eux même un modèle suffisant en tant que classifieur ? La méthode que nous présentons ici consiste simplement à comparer chaque ME test à ces motifs pour réaliser la classification.

Le procédé est simpliste et sans optimisation de la part d'un classifieur. Il y a fort à parier que l'on garde trop de redondance pour être efficace. L'objectif est cependant de proposer une description légère basée sur une observation pratique et d'en évaluer la pertinence expérimentalement.

3.3.1 Méthode

Nous appelons gabarits les images obtenues dans la Figure 3.9. Il y a quatre gabarits, un par type de ME, sous la forme d'une image 2D de la taille d'une frame recadrée autour du visage. Pour classifier une ME, on calcule d'abord sa carte de magnitude du flot optique. Ensuite, une simple corrélation est réalisée entre cette carte et chacun des gabarits :

$$\Gamma_e(S) = \frac{\sum_p M^S(p).G_e(p)}{\sum_{p'} G_e(p')} \tag{3.7}$$

où M^S est l'image de magnitude du flot optique entre l'onset et l'apex pour la séquence test S contenant une ME, G_e est le gabarit du type de ME e et p est un pixel de l'image. Plus l'image test et le gabarit sont semblables (avec le même motif), plus la corrélation est importante. La normalisation permet d'avoir une valeur comparable entre type de ME. Sinon un type de ME générant plus de mouvement (dont la valeur moyenne des intensités du gabarit serait plus élevée) donnerait des valeurs de corrélation plus élevées.

Finalement, la séquence de test est associée au type de ME dont le gabarit produit la valeur de corrélation la plus élevée :

$$\tau(S) = \underset{e}{\operatorname{argmax}}\,(\Gamma_e(S)) \tag{3.8}$$

3.3.2 Résultats

Nous appliquons le protocole LOSO pour la cross-validation. Le jeu d'entraînement participe à la constitution des gabarits.

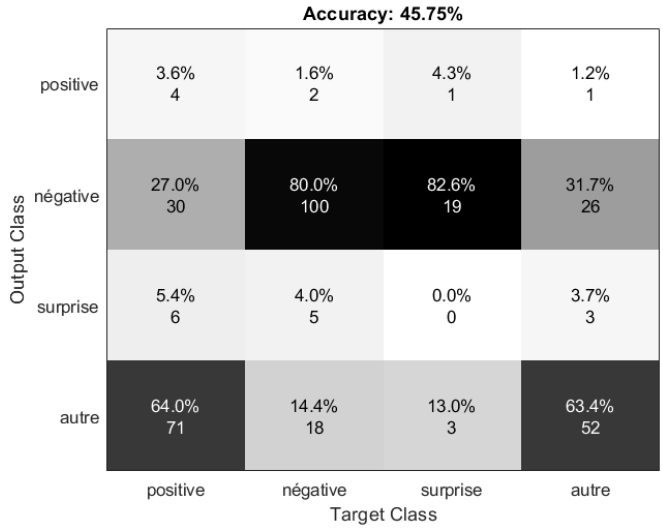

FIGURE 3.10 – Matrices de confusion obtenue à partir de la méthode de corrélation avec des gabarits.

La Figure 3.10 montre la matrice de confusion obtenue par la méthode de gabarits sur l'ensemble de la base de donnée. Nous pouvons observer que le modèle est relativement performant pour classifier les émotions des classes *négatif* et *autre*. Mais le score sur l'ensemble des classes (45.75%) peine à dépasser celui de LBP à lui tout seul.

3.4 Conclusion

L'association descripteur-classifieur possède de nombreux avantages. Le principal consiste à avoir un contrôle sur la caractérisation des données. En observant l'expression des ME à travers la déformation de certaines zones du visage, la littérature a appliqué des descripteurs de forme (LBP) et de mouvement (HOOF). C'est finalement en intégrant le second dans le premier que l'on observe les meilleures performances à travers les LBP_TOP.

Cette association nécessite un fort a-priori sur les caractéristiques recherchées. Autrement dit, pour chaque descripteur, nous maîtrisons plus ou moins les défauts à éviter et les améliorations à apporter. Nous avons par exemple pu observer que la variation de durée des ME exerçait une influence sur la reconnaissance et qu'une unification temporelle améliorait les performances de reconnaissance de ME avec les LBP_TOP. Il est à noter que le meilleur taux de réussite a été obtenu avec un petit

nombre d'images (10 frames) à travers nos évaluations expérimentales. Ce résultat est très intéressant pour le contexte d'applications grand publique. En effet moins le nombre d'images nécessaires est important, plus la vitesse de traitement est élevée. De ce fait, la portabilité sur des systèmes embarqués possédants des ressources de calcul et de stockage limitées est augmentée.

Pour extraire des caractéristiques, il est aussi possible de se baser sur une observation, comme nous l'avons fait avec la méthode des gabarits. Bien que les performances obtenues soient limitées, cette méthode a le mérite d'être simple et compacte.

Si l'unification temporelle réduit l'influence de la variation de durée entre ME, il y a fort à parier que d'autres déséquilibres existent. Dans ce contexte, il est alors logique de se tourner vers les méthodes d'apprentissage profond devenues très populaire ces dernières années. En effet, puisque la caractérisation est compliquée à établir par les êtres humains, il est raisonnable de présenter les ME aux réseaux de neurones artificielles afin de leur demander de les classer après la phase d'apprentissage.

Dans le chapitre suivant, nous quitterons l'étude des descripteurs de ME pour nous intéresser à l'utilisation des techniques d'apprentissage profond. Les objectifs se concentrent alors au niveau des performances, que ce soit en précision, en robustesse, mais aussi en réduction de complexité (système léger) et en temps de traitement.

Chapitre 4

Méthodes basées sur l'apprentissage profond

Bien que les méthodes de machine learning classiques offrent de nombreux avantages, elles sont de plus en plus remplacées par les méthodes dites *Deep Learning* (ou apprentissage profond en français) quand il s'agit de résoudre des problèmes de la vision par ordinateur (les problèmes de classification automatique en général). La reconnaissance de ME ne déroge pas à la règle. Depuis quelques années, les méthodes d'apprentissage profond deviennent prédominantes pour cette tâche malgré le problème récurrent du manque de bases de données.

4.1 Contexte scientifique

Un des principaux inconvénients des méthodes classiques est à sa faible capacité à généraliser la classification aux bases de données mixes. Concrètement, elles obtiennent des performances relativement élevées sur des bases de données individuelles qui chutent rapidement sur des données collectées différemment, même sous des conditions similaires. Cela se voit notamment quand les méthodes classiques sont testées sur des agrégations de bases de données. Pour résoudre ce problème, la communauté scientifique tend à élargir et à diversifier la taille des bases de données pour l'apprentissage profond.

L'analyse de ME n'échappe pas à cette tendance. Les chercheurs fusionnent principalement les bases de données existantes ou appliquent des techniques d'augmentation de données. La plupart des publications dans le domaine utilisent l'une des méthodes de fusion proposée dans les challenges MEGC (Micro-Expression Grand Challenge) de 2018 et 2019. Le MEGC'2018 fusionne deux bases de données ayant toutes les deux fournis les AUs du FACS (section 2.1) pour chacune de leur ME. Le MEGC'2019, quant à lui, essaye d'intégrer la troisième base de données SMIC à l'ensemble. Cette dernière ne fournit pas les AUs, mais présente la possibilité d'avoir une vérité terrain avec la polarisation des émotions sur 3 classes : positive, négative

et surprise. En utilisant les AUs inclues dans CASME II et SAMM, il est aussi possible d'obtenir la polarisation des ME présentes. Le MEGC de 2019 se concentre donc sur la classification en 3 classes. Un MEGC a aussi été organisé en 2020, mais il ne portait que sur le spotting de ME (à partir bases de données dédiées à cette tâche : "SAMM long vidéos" et $CAS(ME)^2$).

Il existe plusieurs réseaux de neurones pré-entraînés dans la littérature. Les plus connus et utilisés sont Alexnet, VGG16, GoogleNet et Resnet :

- Alexnet est le premier réseau à avoir remporté le challenge ILSRV [1] sur la reconnaissance d'objet utilisant ImageNet (qui est un ensemble de données de plus de 14 millions d'images appartenant à 1000 classes). C'est cette première victoire qui a permis de faire connaître les méthodes de Deep Learning au public en 2012.
- VGG16 est un modèle de réseau neuronal convolutif proposé dans [111]. Le réseau atteint une précision de 92,7 % dans le test top-5 d'ImageNet. C'est l'un des célèbres modèles soumis à l'ILSVRC-2014. Il améliore AlexNet en remplaçant les grands filtres à noyau (11 et 5 dans la première et la deuxième couche convolutive, respectivement) par de multiples filtres à noyau 3×3, l'un après l'autre.
- Le réseau Inceptionet a été l'une des principales percées dans le domaine des réseaux neuronaux, en particulier pour les CNN. Jusqu'à présent, il existe trois versions des réseaux Inception nommées Inception Version 1, 2 et 3. La première version est entrée dans le domaine en 2014 et, comme son nom *GoogleNet* le suggère, a été développée par une équipe de Google. Ce réseau a été chargé d'établir un nouvel état de l'art en matière de classification et de détection au sein de l'ILSVRC.
- Après la célèbre victoire d'AlexNet au concours de classification ILSVRC2012, le réseau profond ResNet a sans doute été le travail le plus révolutionnaire de ces dernières années dans la communauté de la vision par ordinateur et de l'apprentissage profond. ResNet permet d'entraîner jusqu'à des centaines, voire des milliers de couches tout en obtenant des performances remarquables. Par ses propriétés et ses résultats, il s'agit du CNN le plus efficace quand il s'agit de Transfert Learning, et en Deep Learning en général sur plusieurs tâches.

Différentes structures de CNN sont utilisées dans la littérature relative à la classification de ME, mais Resnet18 reste l'architecture la plus prisée et l'une des plus performantes. Liu et al. [89] ont d'ailleurs remporté le défi de MEGC 2019 en utilisant un Capsulenet qui s'appuyait sur la structure de Resnet18. Peng et al. [112] ont aussi remporté le MEGC de 2018 grâce à Resnet18.

La nature des ME les rends très difficile à reconnaître. En effet les mouvements

1. ImageNet Large Scale Visual Recognition Challenge

musculaires lors d'une ME sont tellement peu intenses qu'il est difficile même pour un être humain de reconnaître une ME à partir d'une seule image. Il faut en général visionner une vidéo au ralenti pour pouvoir les entrevoir. Pour l'instant, dans l'optique d'une reconnaissance automatique, le flot optique est considéré comme la représentation la plus efficace d'une ME.

Notre objectif est de concevoir des systèmes compacts de reconnaissances de ME pour des applications grand public satisfaisant de multiples contraintes : robustesse, mémoire réduite, bonne autonomie, implémentation facile sur des architectures matérielles, etc. Basée sur les travaux existants, nous présentons dans ce chapitre une architecture CNN légère, possédant une performance de reconnaissance similaire à l'état de l'art tout en utilisant beaucoup moins de ressource de calcul et de mémoire. Elle fonctionne de surcroît en temps réel.

Ces résultats ont été obtenus en exploitant l'architecture du réseau Resnet 18 et les propriétés intrinsèques du flot optique. Sachant qu'il est plus intéressant pour nous de classifier les émotions spécifiques plutôt que d'avoir une polarisation de l'émotion, nous avons évalué nos méthodes en nous basant sur les conditions expérimentales du MEGC de 2018.

Dans la suite de ce chapitre, les différentes architectures étudiées sont présentées dans la section 4.2. Nous avons dans un premier temps choisi Resnet18 au vu de ses capacités d'extraction des caractéristiques et de ses bonnes performances en transfert d'apprentissage. Pour obtenir le meilleur compromis entre la vitesse de traitement, le besoin de mémoire et la précision, nous lui avons apporté plusieurs modifications. Nous avons ensuite conçu nos propres architectures CNN pour exploiter au maximum les propriétés du flot optique extrait de ME afin de créer des structures spécifiques dédiées à la tâche de classification. Dans la section 4.3, nous exposons les résultats expérimentaux correspondant aux différentes étapes de notre étude. La section 4.4 est consacrée à l'analyse des performances en termes de précision, d'espace mémoire nécessaire et de vitesse de traitement ; ainsi qu'à une comparaison avec l'état de l'art. Finalement une conclusion est dressée dans la section 4.5.

4.2 Conception d'architectures à mémoire réduite

La plupart des hommes n'arrivent pas à reconnaître une ME sans la visionner entièrement au ralenti. Cependant, encouragé par les progrès de l'apprentissage profond par rapport aux êtres humains pour certaines tâches très difficiles, les chercheurs ont établi plusieurs modèles dans ce domaine. En effet il existe plusieurs façons de représenter le mouvement subtil d'une ME et de la traiter par intelligence artificielle. Certaines méthodes utilisent les LSTM (Long Short-Term Memory) ou les RCNN (Recurrent CNN) pour traiter les ME, mais avec moins de succès que les méthodes utilisant le flot optique couplé à un CNN.

4.2.1 Données d'entrée

Le flot optique est un moyen très efficace dont les machines disposent pour caractériser les ME en portant l'attention sur les mouvements des pixels lors d'un laps de temps précis. La grande partie des publications, ainsi que la totalité des méthodes présentées lors du MEGC de 2019 l'utilisent dans leur chaîne de traitement.

À partir de l'hypothèse d'invariance de la luminosité, le mouvement de chaque pixel entre les images sur une période de temps est estimé et représenté sous forme de vecteur (Figure 4.1c) indiquant la direction et l'intensité du mouvement.

La projection du vecteur sur l'axe horizontal correspond au champ Vx (Figure 4.1d) tandis que sa projection sur l'axe vertical est le champ Vy (Figure 4.1e). La magnitude (M) est la norme du vecteur (Figure 4.1f). La Figure 4.2 illustre cette représentation d'un vecteur de flux optique.

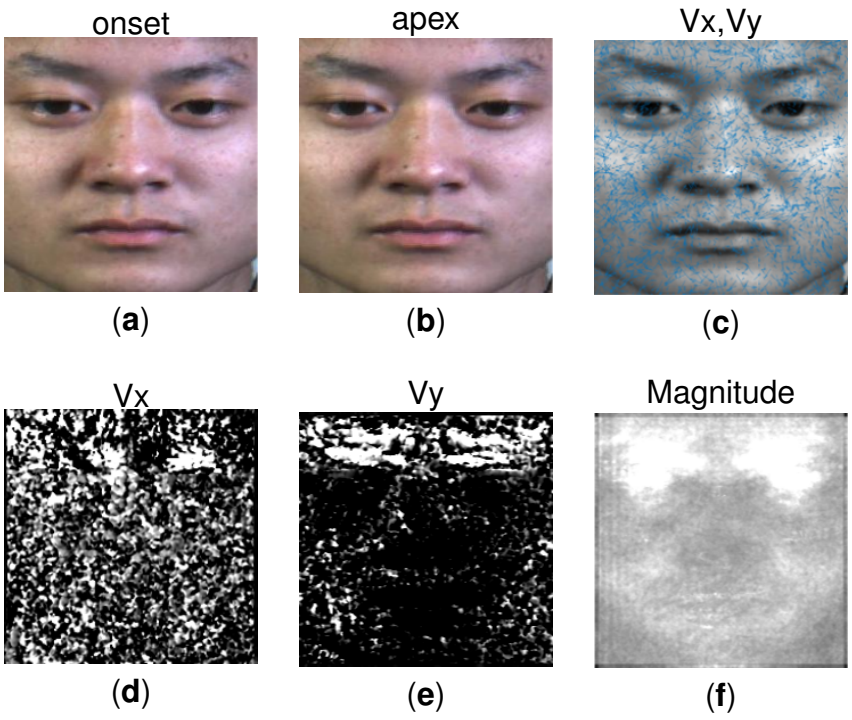

FIGURE 4.1 – Le flux optique est calculé entre l'onset (**a**) et l'apex (**b**) : vecteurs obtenus pour un échantillon aléatoire de pixels (**c**), champ Vx (**d**), champ Vy (**e**) et champ Magnitude (**f**).

La méthode de Horn-Schunck [108] a été choisie pour calculer le flux optique. Cet algorithme a été largement utilisé pour l'estimation du flux optique dans de nombreuses études récentes en raison de sa robustesse et de son efficacité.

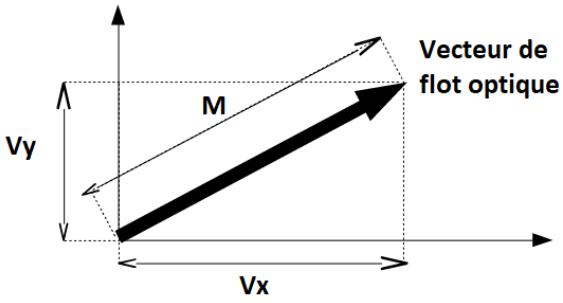

FIGURE 4.2 – Visualisation de M, Vx et Vy pour un vecteur de flot optique.

4.2.2 Étude sur la profondeur du CNN

Les études appliquant l'apprentissage profond pour résoudre le problème de la classification des ME [113, 22, 114, 71] ont généralement utilisé des CNN pré-entraînés tels que ResNet [115] et VGG [111] et ont appliqué l'apprentissage par transfert pour obtenir les caractéristiques de ME. Dans notre travail, nous avons d'abord choisi ResNet18 parce qu'il offrait le meilleur compromis entre la précision et la vitesse sur la classification difficile d'ImageNet et qu'il était reconnu pour ses performances en apprentissage par transfert.

ResNet18 possède une profondeur de 18 couches : 17 couches convolutionnelles (convolutional layers - CL) successives suivies d'une couche de neurones complétement connectés. Celle-ci possède un nombre de neurones égale au nombre de classe qu'on désire classifier, tous connectés à la totalité des neurones de la dernière CL qui la précède. Les liens résiduels après chaque paire d'unités convolutionnelles successives sont utilisés et la taille du noyau après chaque lien résiduel est doublée. Comme ResNet18 est conçu pour extraire des caractéristiques des images de couleur RGB, il implique que les entrées aient 3 canaux.

Afin d'accélérer la vitesse de traitement et de réduire les besoins en mémoire, la principale tendance actuelle visant à diminuer la complexité des CNNs est de réduire le nombre de paramètres apprenables relatifs au nombre de neurones artificielles et aux connexions qui existent entre eux. Par exemple, Hui et al. [116] ont proposé un CNN très compact, nommé LiteFlowNet, qui est 30 fois plus petit dans la taille du modèle et 1,36 fois plus rapide en terme de vitesse d'exécution par rapport aux CNNs de pointe pour l'estimation du flux optique. Dans [117], Rieger et al. ont exploré les réseaux résiduels à paramètres réduits sur des ensembles de données, en ciblant l'estimation en temps réel de la pose de la tête. Ils ont expérimenté diverses architectures ResNet avec un nombre variable de couches pour gérer différentes tailles d'images (y compris des images à basse résolution). ResNet optimisé a atteint une précision élevée avec une vitesse en temps réel.

Il est bien connu que les CNNs sont créés pour des problèmes spécifiques et donc

sur-calibrés ou parfois sous-calibrés lorsqu'ils sont utilisés dans d'autres contextes. ResNet18 a été conçu pour la reconnaissance d'objets de bout en bout dans des images RGB : le jeu de données utilisé pour l'entraînement comporte des centaines de milliers d'images pour chaque classe et plus de mille classes au total.

Nous utilisons ici ce réseau pour la classification de ME. Il s'agit d'une application ayant des implications bien différentes de celles de l'étude générale des images en couleur. Nous pouvons citer deux principales différences : une étude de reconnaissance de ME considère au maximum 5 classes, et les jeux de données de ME spontanées sont rares et contiennent beaucoup moins d'échantillons. D'ailleurs, le flux optique est une caractéristique de haut niveau contrairement aux valeurs RGB considérées comme des caractéristiques de bas niveau et nécessite des réseaux moins profonds.

La classification d'objets dans des images RGB nécessite de reconnaître des formes et des courbures bien spécifiques à cette tâche. La reconnaissance de ME à partir du flot optique, quant à elle, tient plus de la reconnaissance de mouvements musculaires légers exprimés par le flot optique au niveau du visage. Les ME que nous devons analyser ont moins de variabilité. Mais nous disposons aussi de moins de données d'entraînement. Ces différents points nous poussent à penser que nos CNNs n'ont pas besoins d'autant de paramètres apprenables que Resnet18 en possède.

Pour réduire le nombre de ces paramètres, nous proposons dans un premier temps de retirer des couches de Resnet18. Nous avons donc réduit les besoins en mémoire de l'architecture de ResNet18 en supprimant itérativement des couches résiduelles. Nous avons pu ainsi observer graduellement le changement des performances du modèle au fur et à mesure que nous enlevons les couches. Cela nous permet d'évaluer l'influence de la profondeur du réseau sur ses capacités de classification dans notre contexte et donc d'estimer la calibration pertinente du réseau.

La Figure 4.3 illustre le protocole de réduction : à chaque étape, la dernière couche résiduelle avec deux CL est supprimée et la précédente est connectée à la couche entièrement connectée. Seuls les réseaux avec un nombre impair de CL sont donc évalués. Les poids de tous les CNNs sont pré-entraînés en utilisant ImageNet. Comme le soulignent le Tableau 4.1 et la Figure 4.4 , la diminution du nombre de CL a un impact significatif sur le nombre de paramètres apprenables du réseau, ce qui affecte directement le temps d'apprentissage et de classification.

4.2.3 Étude sur la dimensionnalité des données d'entrée

La dimensionnalité des entrées du CNN participe aussi à la détermination de sa complexité, puisque le nombre des canaux d'entrée dicte le nombre de filtres à utiliser dans toutes les couches suivantes du CNN. Dans le domaine de la reconnaissance

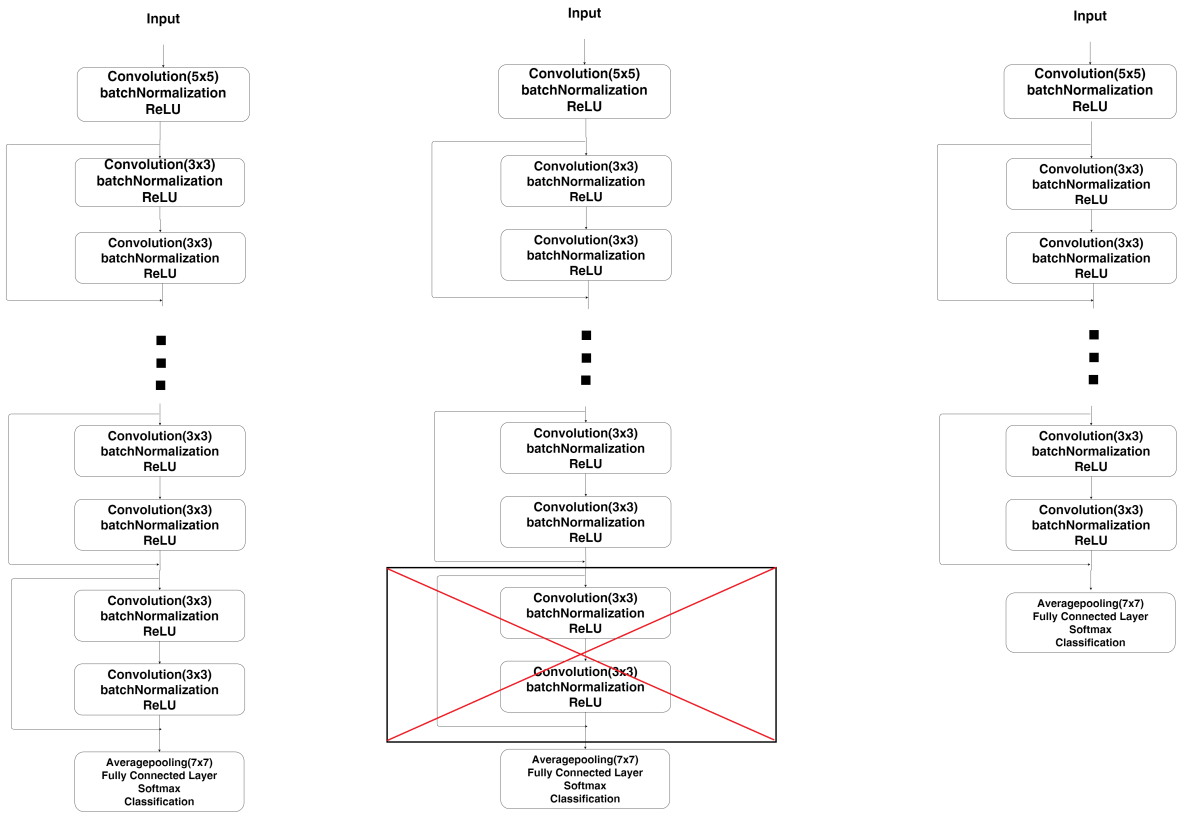

FIGURE 4.3 – Réduction de la profondeur d'un réseau de neurones profond : dans le réseau initial, chaque couche résiduelle contient deux couches convolutives (CL) (**gauche**) ; la dernière couche résiduelle est supprimée (**millieu**) pour obtenir un réseau moins profond (**droite**).

CL	Nb. de param.
17	10 670 932
15	5 400 725
13	2 790 149
11	1 608 965
9	694 277
7	398 597
5	178 309
3	104 197
1	91 525

TABLE 4.1 – Nombre de CL et nombre de paramètres apprenables dans les architectures proposées.

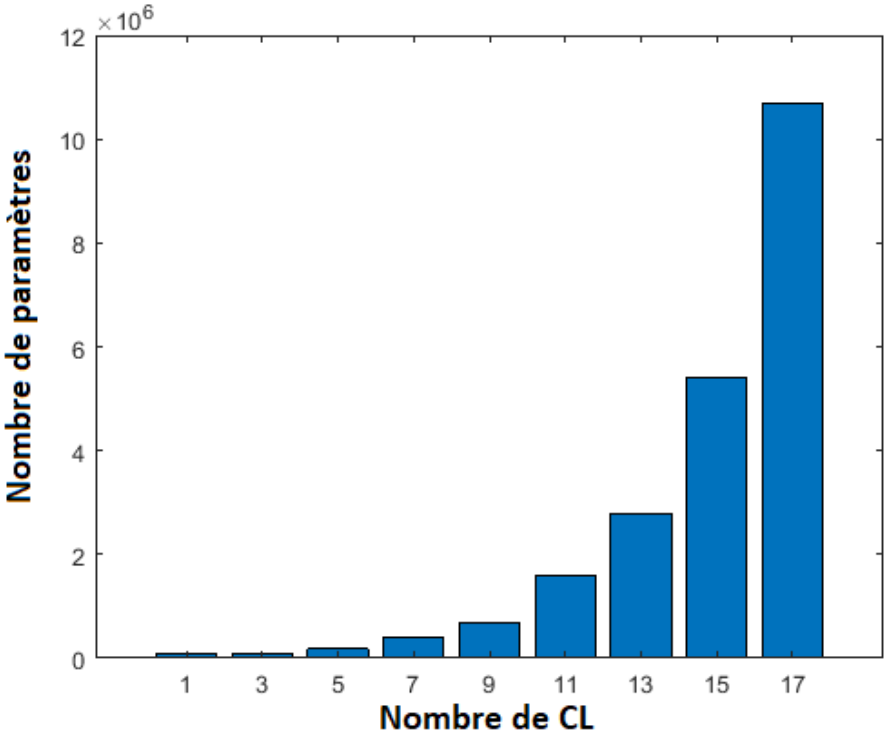

FIGURE 4.4 – Nombre de paramètres apprenables en fonction du nombre de CL.

de ME avec apprentissage profond, le flux optique entre l'onset (Figure 4.1a) et l'apex (Figure 4.1b) est souvent utilisé pour alimenter les réseaux CNN comme données d'entrée. Il est représenté sous la forme de 3 canaux Vx, Vy et M pour remplacer les canaux RGB des images en couleur.

Le flot optique est une représentation efficace d'une ME mais pas parfaite. Par exemple, il peut être bruité par des mouvements ne correspondant pas aux ME comme le clignotement des yeux ou le mouvement de la tête. D'ailleurs, il faut garder à l'esprit que le flot optique est moins riche en dimension que les images RGB pour lesquels les CNNs ont originalement été conçus. Utiliser des CNNs tels qu'ils ont été créés ne tire pas partie des propriétés des ME et du flot optique utilisé pour les représenter. Nous proposons donc une nouvelle étude pour concevoir des CNNs plus rapides et moins gourmands en espace mémoire. Ainsi, dans un soucis d'embarquabilité, nous pouvons proposer des approches compactes de reconnaissance de ME capables de tourner sur des cibles matérielles avec des ressources limitées.

Lors de la classification de ME, nous estimons d'abord le flot optique entre les images de l'onset et de l'apex. C'est entre ces deux moments que le mouvement est susceptible d'être le plus fort. Ensuite, les matrices Vx, Vy et M sont traditionnellement présentées comme entrées du réseau Resnet. Pourtant, le troisième canal est intrinsèquement redondant puisque M est calculé à partir de Vx et Vy. Un flux optique composé des champs Vx et Vy à deux canaux pourrait déjà fournir toutes les informations pertinentes.

En outre, nous voulons savoir si un seul champ de mouvement pourrait être suffisamment descriptif d'une ME. Nous avons donc créé et évalué des réseaux configurés pour prendre en entrée une représentation du flux optique à deux canaux (Vx-Vy) mais aussi à un canal (M, Vx ou Vy). À cette fin, les réseaux proposés commencent par un certain nombre de CL liés à l'optimisation de la profondeur, suivis d'une normalisation par lots et d'une fonction ReLU. Puis les réseaux se terminent par une couche de *maxpooling* et une couche entièrement connectée.

La Figure 4.5 présente les architectures proposées avec de 1 à 4 CL selon les résultats des expériences de la section 4.3. Comme illustré dans le Tableau 4.2, une entrée de faible dimension conduit à une réduction significative du nombre de paramètres apprenables et donc de la complexité du système. Avec cette Figure, nous observons aussi une évolution exponentielle du nombre de neurones par rapport au nombre de couches. La taille du CNN à deux canaux contient également beaucoup plus de neurones que celui à un seul canal d'entrée.

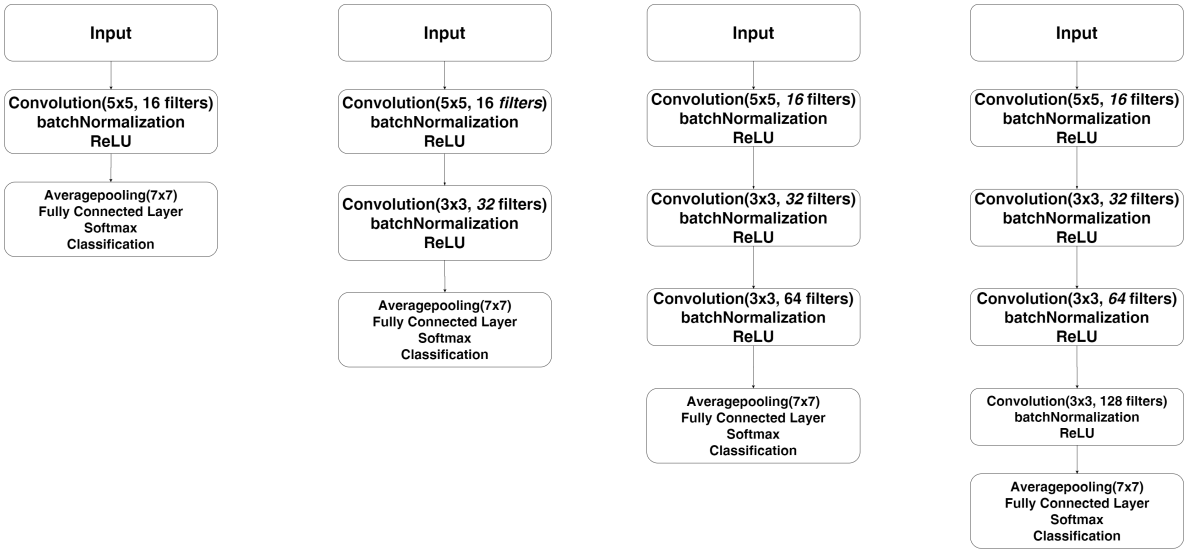

FIGURE 4.5 – Les architectures proposées avec de 1 à 4 CL.

Entrée	1 CL	2 CL	3 CL	4 CL
Un canal	82 373	168 997	333 121	712 933
Deux canaux	165 541	348 005	709 477	1 620 197

TABLE 4.2 – Nombre de paramètres apprenables en fonction de la dimensionnalité de l'entrée du réseau.

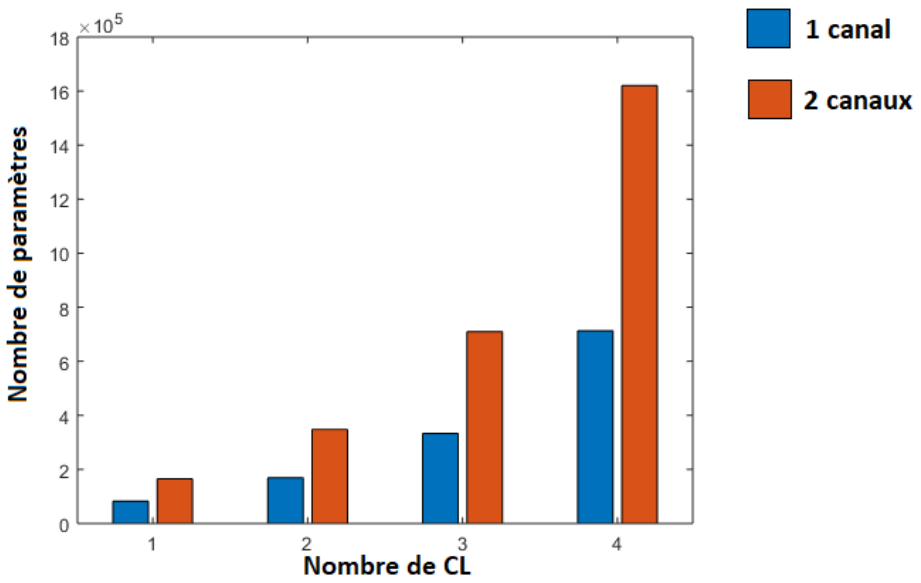

FIGURE 4.6 – Nombre de paramètres apprenables en fonction du nombre de CL sur les deux architectures (un seul canal d'entrée en bleu et 2 canaux d'entrée en orange).

4.3 Résultats expérimentaux

Dans cette section, nous étudions différentes architectures proposées pour réaliser d'une manière optimale la tâche de la reconnaissance de ME. Nous présentons premièrement le protocole expérimental avant de tester les différentes profondeurs de ResNet. Dans la deuxième partie, nous illustrons les performances des réseaux compacts avec les entrées de basse dimensionalité. Finalement, nous effectuons une analyse comparative sur les capacités des architectures à extraire les caractéristiques discriminatives à partir du flot optique.

4.3.1 Protocole expérimental

Deux bases de données de ME ont été utilisées dans nos expériences. CASME II (Chinese Academy of Sciences Micro-Expression II) [100] est une base de données complète de ME spontanées contenant 247 échantillons vidéo, recueillis auprès de 26 participants asiatiques dont l'âge moyen est de 22,03 ans. Par rapport à la première base de données, la base de données Spontaneous Actions and Micro-Movements (SAMM) [5] est plus récente et se compose de 159 micromouvements (une vidéo pour chacun). Ces vidéos ont été capturés auprès d'un groupe démographiquement diversifié de 32 participants avec un âge moyen de 33,24 ans et une répartition équilibrée entre les sexes. Conçue à l'origine pour étudier les micromouvements faciaux, SAMM a initialement annoté les sept émotions de base.

Les bases de données de CASME II et de SAMM ont toutes les deux étés enregistrées à une fréquence d'images de 200 fps. Elles contiennent également toutes deux des "classes objectives", comme indiqué dans [118]. Pour cette raison,

le MEGC'2018 [119] a proposé de combiner tous les échantillons de ces deux bases de données en un seul ensemble de données composites de 253 vidéos avec cinq classes d'émotion. Il convient de noter que la répartition n'est pas très bien équilibrée. En effet, dans cette base de données, la classe *joie* (numéro I) représente 19,92% des échantillons, la classe *surprise* (numéro II) en représente 11,62% des échantillons, la classe *colère* (numéro III) en représente 47,30% des échantillons, la classe *dégoût* (numéro IV) en représente 11,20% des échantillons, et la classe *tristesse* (numéro V) en représente 9,96% des échantillons.

Dans toutes les expériences, nous avons entraîné les modèles de CNN avec une taille de mini-batch de 64 pour 150 époques en utilisant l'optimisation RMSprop. Une simple augmentation des données a été appliquée pour doubler la taille de l'entraînement. Plus précisément, pour chaque échantillon utilisé pour l'entraînement, nous utilisons le flot optique entre les images de l'onset et de l'apex mais aussi entre celle de l'onset et celle suivant l'apex (apex+1).

Comme pour le MEGC'2018 [119], nous avons appliqué le protocole de validation croisée LOSO (Leave One Subject Out) pour la classification de ME, dans lequel l'ensemble des données d'un sujet est utilisé comme échantillons de test. Ceci afin de mieux reproduire des scénarios réalistes où les sujets rencontrés ne sont pas présents lors de la formation du modèle. Dans toutes les expériences, la performance de reconnaissance est mesurée à partir du critère d'accuracy, qui est le pourcentage d'échantillons vidéo correctement classés par rapport au nombre total d'échantillons dans la base de données.

4.3.2 Étude sur la profondeur de ResNet

Afin de trouver la profondeur optimale de ResNet qui permet le meilleur compromis entre la performance de reconnaissance de ME et le nombre de paramètres apprenables, nous avons testé différentes profondeurs de CNN en utilisant la méthode décrite dans la section 4.2.2. Les précisions obtenues sont données dans le Tableau 4.3.

Nous pouvons observer que les meilleures performances ont été obtenues par la version de ResNet8 avec 7 CL. Cependant la variation des scores en fonction du nombre de CL est limitée. En outre, au-delà de 7 CL, l'ajout de CL supplémentaires n'améliore pas la précision du modèle. Cela confirme que de multiples CL successives ne sont pas nécessaires pour obtenir une meilleure précision.

Le phénomène le plus intéressant se révèle être qu'avec une seule CL nous avons obtenu un score qui n'est pas très éloigné du score optimal alors que la taille du modèle est beaucoup plus réduite. Cela suggère qu'au lieu d'un apprentissage profond, une approche plus "classique" exploitant des réseaux neuronaux peu profonds présente un champ intéressant à explorer pour optimiser la portabilité et l'efficacité des

CL	Nb. of param.	Accuracy
17	10 670 932	57.26%
15	5 400 725	57.26%
13	2 790 149	60.58%
11	1 608 965	59.34%
9	694 277	60.17%
7	398 597	61.00%
5	178 309	58.51%
3	104 197	60.17%
1	91 525	58.92%

TABLE 4.3 – Les performances (accuracy) varient en fonction du nombre de couches convolutionnelles (CL) et du nombre associé de paramètres apprenables.

calculs pour des systèmes embarqués. C'est la raison principale pour laquelle nous concentrons nos études sur des CNNs compacts.

4.3.3 Étude sur la dimensionnalité des données d'entrée

Dans cette sous-section, nous étudions l'impact des représentations du flux optique sur les performances de reconnaissance de ME. Deux types de CNN ont été étudiés : l'un avec une entrée à un canal (Vx, Vy, ou M) et l'autre utilisant la paire Vx-Vy à deux canaux. Étant donné que les CNNs standards prennent généralement des entrées à 3 canaux et sont préformés en conséquence, l'application de l'apprentissage par transfert pour adapter nos modèles aurait été une tâche difficile et inappropriée. Nous avons donc créé des CNNs personnalisés et les avons entraîné à partir de zéro.

Le Tableau 4.4 montre les précisions de reconnaissance sous différentes configurations utilisant un petit nombre de couches CNN. Nous pouvons observer que la paire Vx-Vy et Vy seul ont donné les meilleurs résultats. Les deux représentations atteignant une précision de 60,17%.

	1 CL	2 CL	3 CL	4 CL
Vx	52.24%	54.34%	53.92%	53.50%
Vy	58.09%	59.34%	**60.17**%	**60.17**%
Vx-Vy	58.51%	59.75%	**60.17**%	58.09%
M	58.09%	58.92%	59.34%	59.34%

TABLE 4.4 – Précisions sous différentes architectures CNN et représentations du flux optique.

Cela nous amène à penser que les caractéristiques les plus utiles pour la reconnaissance de ME pourraient être présentes dans les mouvements verticaux donnés par Vy. Cette hypothèse est logique si l'on pense aux mouvements musculaires qui

se produisent dans chaque expression faciale connue.

D'autre part, l'utilisation de la Magnitude seule conduit à une précision similaire à celle de Vy et de la paire Vx-Vy avec un score de 59,34%. Vx a obtenu les plus mauvais résultats dans l'ensemble, avec un score maximal de 54,34%. Cette observation indique que les caractéristiques les plus importantes pour la classification de ME pourraient en effet être plus dominantes dans le mouvement vertical que dans le mouvement horizontal.

Pour mieux visualiser la différence entre les caractéristiques de haut niveau présentes dans Vx, Vy et la Magnitude, nous avons fait une moyenne sur tous les différents échantillons en fonction de leurs classes. Le résultat est visible sur la Figure 4.7. Nous pouvons observer que Vx présente une quantité non négligeable de bruit. En revanche, Magnitude et Vy présentent des régions d'activité claires pour chaque classe. Les régions d'activité sont alignées avec les muscles responsables de chaque expression faciale.

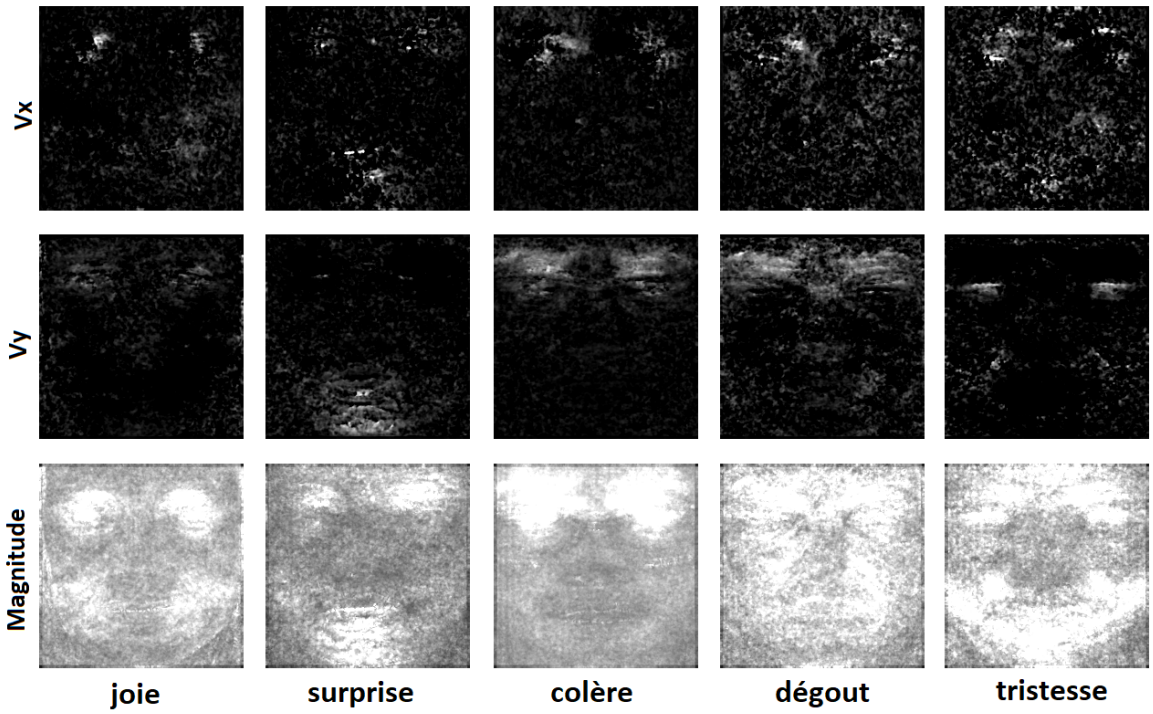

FIGURE 4.7 – Moyenne sur le jeu de données du flux optique par classe de ME. Les classes étudiées sont dans l'ordre de gauche à droite : joie, surprise, colère, dégoût et tristesse.

4.3.4 Analyse de la similarité des caractéristiques extraites

La capacité des CL à extraire des caractéristiques permettant à distinguer les différentes classes est primordiale pour la reconnaissance des ME. De ce fait, analyser

la capacité du CNN à produire des caractéristiques permet de mieux comprendre les résultats qu'il est capable d'obtenir. Dans notre cas, analyser la similitude entre les caractéristiques peut aussi nous indiquer à quel point la réduction des calculs à altérer la capacité du modèle à discerner les différentes ME.

Habituellement, les couches convolutionnelles des CNNs sont considérées comme différents extracteurs de caractéristiques. Seule la dernière couche entièrement connectée effectue directement la tâche de classification. Les caractéristiques juste avant la classification peuvent être représentées sous forme d'un vecteur. Nous nous arrêtons donc aux valeurs d'activation des neurones dans la dernière couche convolutionnelle (avant le softmax) pour en extraire les valeurs et calculer leurs similarités.

Afin de comprendre les résultats obtenus, nous avons mesuré la similarité en cosinus (Equation 4.1) des caractéristiques extraites par trois CNNs : ResNet8 (Section 4.3.2), Vx-Vy-3-CL et Vy-3-CL (Section 4.3.3). La similarité en cosinus mesure la similarité entre deux vecteurs a et b en utilisant l'équation suivante :

$$cosine(a, b) = \frac{a^T b}{\|a\| \, \|b\|} \tag{4.1}$$

La similarité en cosinus prend valeur dans l'intervalle $[-1, 1]$. Plus la valeur se rapproche de 1, plus cela implique une forte similarité entre les deux vecteurs. Les Tableaux 4.5, 4.6 et 4.7 affichent les valeurs de similarité en cosinus avec 3, 2 et 1 canaux. Nous avons calculé l'intra-similarité et l'inter-similarité moyenne de chaque classe ME en utilisant la même configuration pour trois CNNs.

	Joie	**Surprise**	**Colère**	**Dégoût**	**Tristesse**
Joie	0.8464	0.3966	0.3860	0.3126	0.2960
Surprise	0.3966	0.8159	0.4040	0.3362	0.3324
Colère	0.3860	0.4040	0.8344	0.3654	0.3307
Dégoût	0.3126	0.3362	0.3654	0.8598	0.2363
Tristesse	0.2960	0.3324	0.3307	0.2363	0.9343

TABLE 4.5 – Similarité en cosinus pour ResNet8 avec 3 canaux en entrée (Vx, Vy et M).

	Joie	**Surprise**	**Colère**	**Dégoût**	**Tristesse**
Joie	0.5615	0.1700	0.1171	0.1155	0.1195
Surprise	0.1700	0.5831	0.1432	0.1502	0.1618
Colère	0.1171	0.1432	0.5672	0.1176	0.1503
Dégoût	0.1155	0.1502	0.1176	0.5447	0.1225
Tristesse	0.1195	0.1618	0.1503	0.1225	0.5443

TABLE 4.6 – Similarité en cosinus pour le CNN 3-CL avec 2 canaux en entrée (Vx-Vy).

	Joie	Surprise	Colère	Dégoût	Tristesse
Joie	0.6007	0.1320	0.0574	0.0146	0.1154
Surprise	0.1320	0.5572	0.0485	0.0667	0.1415
Colère	0.0574	0.0485	0.5260	0.0318	0.0698
Dégoût	0.0146	0.0667	0.0318	0.5663	0.0159
Tristesse	0.1154	0.1415	0.0698	0.0159	0.5099

TABLE 4.7 – Similarité en cosinus pour le CNN 3-CL avec un canal en entrée (Vy).

Tout d'abord, nous pouvons remarquer que les valeurs sur la diagonale (intra-classe) des trois CNNs sont significativement plus élevées par rapport aux autres valeurs (inter-classe). Cela montre que les trois CNNs sont capables de séparer les différentes classes de ME. Deuxièmement, la similarité cosinus intra-classe de ResNet est plus élevée. Ce qui suggère que les caractéristiques de ResNet sont plus discriminantes. Nous supposons que nos CNNs avec des couches réduites extraient des caractéristiques moins raffinées ; ce qui entraîne une légère diminution des performances (61,00 % contre 60,17 %).

4.4 Analyse des performances

Dans cette section, nous effectuons une analyse des performances de notre méthode sous trois aspects : la précision pour la reconnaissance, l'espace de mémoire nécessaire et la vitesse de traitement. Comme nous avons obtenu des résultats optimaux en utilisant le champ Vy et un CNN à 3 couches, les évaluations ultérieures se concentrent sur cette configuration particulière.

4.4.1 Comparaison avec l'état de l'art

Pour évaluer la reconnaissance, nous effectuons une comparaison avec les méthodes présentes dans la littérature qui suivent le même protocole expérimental et sont évaluées sur la même agrégation de bases de données et selon les même vérités de terrain.

Le résultat comparatif est affiché sur le Tableau 4.8. Nous pouvons constater que les méthodes utilisant le deep learning ont de meilleurs résultats que le LBP_TOP. Ce dernier est pourtant le descripteur de texture spatio-temporel produisant le meilleur résultat parmi les méthodes dites classiques. Comme précisé auparavant, le LBP_TOP a une capacité de généralisation plus faible que celle des réseaux de neurones profonds.

Notre meilleur CNN obtient des performances similaires à celles d'autres études utilisant le même protocole de validation comme celle de Khor et al. [120]. Leur méthode utilise aussi le flot optique et le deep learning. Cependant, ils n'explorent la

Méthode	Accuracy
LBP_TOP [120]	42.29%
Khor et al. [120]	57.00%
Peng et al. [112]	74.70%
Notre méthode	60.17%

TABLE 4.8 – Comparaison entre notre méthode et d'autres méthodes de la littérature utilisant la même base de données.

data augmentation ce qui laisse une marge de progression.

Peng et al. [112] possède la meilleure performances. Mais il est à noter qu'ils emploient un modèle ResNet10 avec l'apprentissage de transfert des macro-expressions vers les micro-expressions. Ce modèle utilise quatre ensembles de données de macro-expressions (>10 K images) et certains prétraitements, tels que le changement de couleur, la rotation et le lissage. Ces opérations supplémentaires rendent leur méthode difficile à déployer sur des systèmes embarqués.

La Figure 4.8 affiche la matrice de confusion obtenue par notre méthode et celle de [120]. Nous pouvons également remarquer que la distribution des évaluations correctes pour Vy était plus équilibrée que celles obtenues par [120].

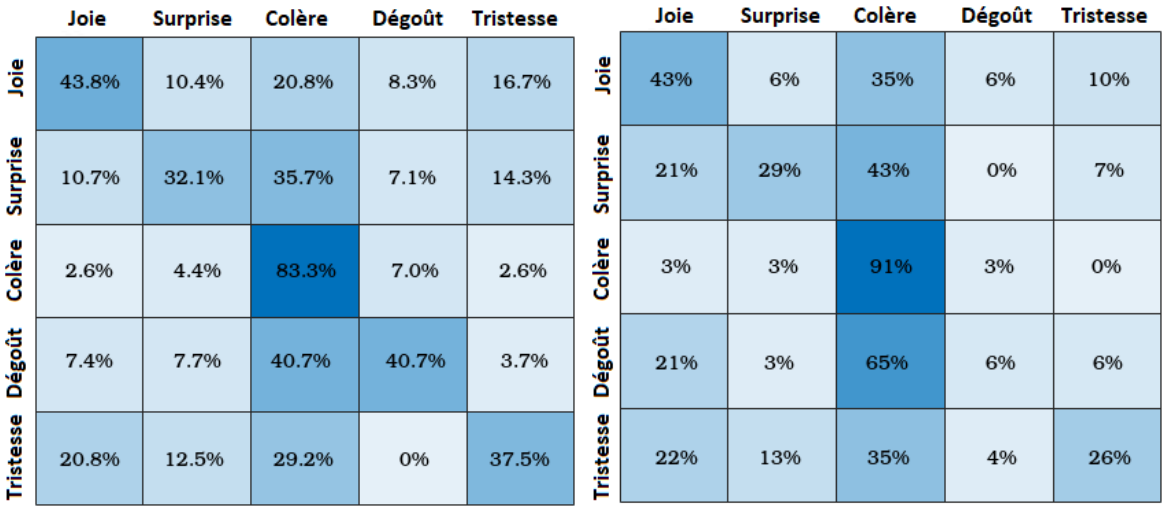

FIGURE 4.8 – Matrices de confusion correspondant à notre réseau avec 3 CL et Vy comme entrée (à gauche) et à l'étude de [120] (à droite).

Dans les deux matrices, il est clair que la classe colère est la mieux reconnue en général. La principale raison pour cela est très probablement la répartition des classes dans la base de données. Cette classe représente 47,30% de la base de données. Il est donc logique que les classifieurs tendent à mieux se spécialiser sur cette classe en particulier. Avoir un bon score sur une classe n'est cependant pas idéale si cela se répercute sur le score des autres classes.

4.4.2 Évaluation de l'espace mémoire nécessaire

Quand il s'agit de concevoir des algorithmes de traitement d'images et de vidéos pour des systèmes embarqués, l'un des éléments les plus importants est le besoin en mémoire. Pour les CNNs, la quantification de ce besoin est relativement facile. En fait, dans un réseau de neurones, le nombre de neurones, le nombre de connexions ainsi que leur type se reflètent directement le besoin en mémoire. La relation est quasi linéaire : il faut toujours un espace mémoire précis pour coder un paramètre apprenable.

Le Tableau 4.9 résume le nombre de paramètres apprenables et de filtres utilisés en fonction de la dimensionnalité des entrées du réseau. L'espace mémoire minimal requis pour notre méthode correspond à 333 121 paramètres, soit moins de 3,12 % de celui du réseau standard ResNet18.

Entrée	**1 CL**	**2 CL**	**3 CL**	**4 CL**
un canal	82 373 (16)	168 997 (48)	333 121 (112)	712 933 (240)
double canaux	165 541 (32)	348,005 (96)	709 477 (224)	1 620 197 (480)

TABLE 4.9 – Nombre de paramètres, et de filtres apprenables (entre parenthèses) de diverses architectures de réseau sous différentes dimensions d'entrée.

Nous pouvons observer que Resnet18 dans sa forme basique a besoin de plus de 10 millions de paramètres (voir section 4.2.3). À titre de comparaison, notre CNN à deux canaux d'entrée avec 4 CL est constitué de 1.6 millions de paramètres. Cela est bien entendu dû à la profondeur réduite du CNN, mais aussi au noyau de plus petite taille dont il a besoin pour effectuer ses calculs. Le CNN à un seul canal d'entrée avec 4 CL ne possède que 712 933 paramètres, c'est à dire un peu moins de la moitié du CNN à deux canaux. En comparant les méthodes produisant les meilleurs compromis précision/mémoire, Resnet8 a besoin de 398 597 paramètres alors que notre CNN à 1 canal et 3 CL utilise 333 121 paramètres.

Tout au long des expérimentations, nous avons pu remarquer que les scores restent relativement bon même avec très peu de couches convolutionnels. Si nous comparons les versions comportant une seule CL pour les trois possibilités en termes de dimension d'entrée, nous observons que Resnet2 nécessite 91 525 paramètres (voir Tableau 4.1) et le CNN à 1 canal et 1 CL a besoin de 82,373. Autrement dit, si nous considérons le fait qu'il traite moins d'information et ne nécessite que le calcule Vy et non de [Vx, Vy, M], le CNN à 1 canal est le plus moins coûteux en termes de besoins en mémoire.

La meilleure performance en terme de précision revient à Resnet8 avec 61% et une configuration qui nécessite 398 597 paramètres. Si nous voulons faire une concession au niveau de la précision et accepter un score de 58.09%, nous pouvons considérer

le CNN à 1 canal et 1 CL qui n'a que de 82 373 paramètres. Sachant qu'il représente moins de 21% des besoins de Resnet8 et qu'il sollicite trois fois moins de prétraitement, il semble un bon compromis entre la robustesse de reconnaissance et le besoin en mémoire.

4.4.3 Évaluation de la vitesse de traitement

Si la relation entre le besoin de mémoire et le nombre de paramètres apprenables est presque linaire, ce n'est pas le cas pour la vitesse de traitement. En effet la vitesse d'exécution d'un programme utilisant un CNN dépend de beaucoup de facteurs. Le nombre de paramètres est l'un des éléments qui impacte le plus la vitesse de propagation interne d'un CNN, mais l'agencement et le type des neurones le composant jouent aussi un rôle primordial. Avec l'exécution parallèle d'une partie de la propagation des données au sein du CNN, la relation entre le nombre de paramètres apprenables et la vitesse d'exécution ne peut pas être linéaire et est même difficile à modéliser.

Dans ce contexte, nous mesurons le temps nécessaire à nos CNNs pour effectuer la propagation des données de la couche d'entrée à la couche de sortie. Toutes nos expériences ont été réalisées sur un ordinateur de milieu de gamme avec un processeur Intel Xeon et une carte graphique Nvidia GTX 1060. La chaîne de traitement complète a été implémentée en utilisant le logiciel MatLAB 2018a avec sa boîte à outils d'apprentissage profond.

Notre modèle qui a obtenu le meilleur score est le CNN avec une entrée à un seul canal et 3 CL successifs. Il faut 12,8 ms pour classer la composante verticale Vy. Le calcul du flux optique entre deux images nécessite 11,8 ms, ce qui conduit à un temps d'exécution total de 24,6 ms pour classifier un clip vidéo contenant une ME. À notre connaissance, la méthode proposée surpasse la plupart des systèmes de reconnaissance de ME en termes de vitesse de traitement.

En considérant le fait que les ME les plus courtes durent 40 ms, ce temps d'exécution semble raisonnable dans un contexte de reconnaissance en temps réel. Ces résultats sont encourageants en vue des perspectives qu'ils laissent entrevoir. Par ailleurs, nous parvenons à ces résultats avec un calcul du flot optique relativement lourd. Des optimisations apportées à ce niveau pourraient encore réduire le temps de traitement nécessaire.

4.5 Conclusion

Dans ce chapitre, nous proposons des architectures CNNs pour reconnaître les ME spontanées. Nous avons d'abord effectué une étude sur la profondeur du réseau ResNet18 pour démontrer que l'utilisation d'un petit nombre de couches est suffisante

pour notre tâche. Sur la base de cette observation, nous avons ensuite expérimenté plusieurs représentations au niveau l'entrée du réseau. Ces études nous ont permis de concevoir plusieurs architectures compactes de CNNs pour la reconnaissance de ME.

Concrètement, nous avons alimenté les CNNs avec le flux optique estimé à partir de l'onset et de l'apex des ME. Différentes représentations du flux (Vx horizontal, Vy vertical, Magnitude M et la paire Vx-Vy) ont été testées et évaluées sur un jeu de données composite (CASME II et SAMM) pour la reconnaissance de cinq classes objectives. Les résultats obtenus sur l'entrée Vy seule sont les plus convaincants. Cela est probablement dû au fait qu'une telle orientation est plus apte à décrire le mouvement de la ME et ses variations entre les différentes classes d'expression. Les résultats expérimentaux ont montré que la méthode proposée peut atteindre un taux de reconnaissance similaire à celui des approches de l'état de l'art.

Enfin, nous avons obtenu une précision de 60,17% avec un CNN léger et composé de 3 CL avec une entrée à canal unique Vy. Cette configuration permet de réduire le nombre de paramètres apprenables d'un facteur 32 par rapport à ResNet18. De plus, nous avons obtenu un temps de traitement de 24,6 ms, ce qui est plus court que les ME (40 ms). Notre étude ouvre une voie intéressante pour trouver le compromis entre la vitesse et la précision pour la reconnaissance de ME.

La conception d'une architecture compacte, rapide en termes de calculs, et peut vorace en mémoire constitue une étape importante dans la conception de dispositifs de reconnaissance automatique de ME. Le système complet étant bien plus conséquent que la partie classification à elle seule. Plusieurs autres étapes doivent être ajoutées à la chaîne de traitement pour avoir un système complet.

La littérature sépare souvent les différentes étapes d'analyse de ME ; le processus complet est très rarement adressé. Cela peut principalement être attribué au fait que le spotting de ME reste toujours une tâche très difficile produisant des F1-scores dépassant rarement 15%. Cela dit, la communauté œuvre activement à concevoir différentes techniques afin de surmonter ou à défaut contourner les obstacles les plus entravants à la réalisation d'un système complet. Nous présentons une étude exploratoire dans ce sens au chapitre suivant.

Chapitre 5

Analyse complète : du spotting à la classification

Le développement des méthodes basées sur la vision par ordinateur pour l'étude des ME se décompose, selon la communauté scientifique, en deux sous-disciplines : le spotting et la reconnaissance (Figure 5.1). La reconnaissance, étudiée dans les chapitres 3 et 4, consiste à assigner une classe à une séquence contenant une ME. Cette classe correspond au type de ME (lié à l'émotion ressentie) auquel elle appartient. En pratique la classification se réalise souvent entre 5 classes : la joie, la tristesse, la colère, la surprise et le dégoût. La description de chaque classe se base sur les mouvements spécifiques séparant les différentes classes plutôt que sur les mouvements communs d'une ME.

Le spotting, quant à lui, part d'une séquence complète relativement longue dans laquelle il localise temporellement, s'il y en a, la ou les ME. Il n'y a alors que deux classes pour chaque instant : "est une micro-expression" ou "n'est pas une micro-expression".

La répartition est souvent inégale car, en pratique, les périodes contenant une ME sont ponctuelles, isolées et relativement rares vis à vis de la séquence complète. En reprenant les notations du chapitre 1, l'onset, l'offset et parfois l'apex sont connus pour la tâche de la reconnaissance, et c'est le type d'expression qui est recherché. Dans le spotting, rien n'est connu et il faut estimer où trouver l'onset et soit l'offset soit l'apex. Si ces deux étapes se suivent naturellement dans un processus complet, elles sont généralement étudiées séparément depuis les toutes premières études [37, 40, 45].

La reconnaissance est beaucoup plus étudiée que le spotting. Par exemple, en 12 ans, nous estimons qu'il y a eu environs 80 articles scientifiques parus pour la reconnaissance contre seulement 30 pour le spotting [121, 122]. Cela s'explique par la très haute complexité du spotting et des relativement faibles résultats obtenus. Ceci est préjudiciable car les deux étapes sont nécessaires pour une application pratique. En effet un classifieur peut n'être efficace que sur une segmentation parfaite des ME.

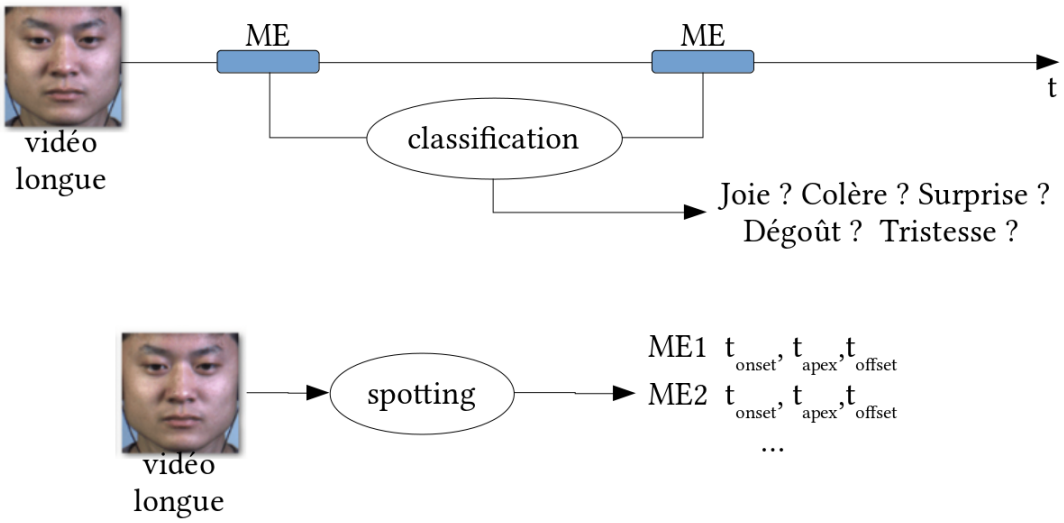

FIGURE 5.1 – Différence entre la reconnaissance et le spotting : en haut, la reconnaissance se base sur une ME et détermine quelle est l'émotion qui lui est associée ; en bas, le spotting part d'une vidéo longue et localise dans le temps les ME s'il y en a.

Dans le cas contraire, sa performance baisse drastiquement si les ME sont isolées avec peu de précision. Ainsi les résultats de classification peuvent paraître artificiels puisqu'ils sont obtenus à partir de données très difficiles à estimer.

Le spotting reste un défi car il faut détecter les mouvements musculaires subtils émanant des ME tout en les distinguant des autres mouvements musculaires non lié aux ME tels que les tics nerveux, les clignements des yeux et autres mouvements rapides. Étant donné leur faible occurrence d'apparition et la quantité réduite de mouvement pouvant apparaître (rappelons que l'étude se concentre sur de très petits intervalles de temps), les ME sont excessivement difficiles à isoler correctement.

Traditionnellement, les descripteurs usuels de classification de ME (LBP_TOP, HOG, IA) sont utilisés pour réaliser le spotting. Récemment dans sa thèse, Jingting Li a proposé un descripteur lié non plus à la nature des émotions, mais à la présence d'une ME [123]. Ce nouveau descripteur, pertinent pour la détection de ME, est appelé le motif temporel local (Local Temporal Pattern -LTP). Bien que ce transfert de l'étude des caractéristiques des classe d'émotion vers celles de l'occurrence des ME soit prometteur, nous en sommes encore aux balbutiements de la description recherchée.

Le deep learning a considérablement modifié le domaine de la vision par ordinateur et particulièrement les problématiques de la classification. Les nouveaux algorithmes sont devenus bien plus performants et ont ouvert de nombreuses perspectives. Ces avancées impressionnantes ont ruisselé vers les différents sous-domaines ce qui a grandement profité à l'étude de la reconnaissance de ME.

Le spotting n'a cependant pas autant profité de cette révolution. Les seules méthodes traitant du spotting à partir du deep learning se base sur des apprentissages liés à la reconnaissance. Cela est en partie dû à la faible quantité de données disponible. Les bases de données existantes ne contiennent souvent que des ME labélisées, mais pas de longues séquences appropriées. Or la localisation temporelle étant encore plus complexe que la classification, il faudrait bien plus de données d'entraînement. Les résultats que peuvent atteindre les algorithmes de spotting contemporains sont donc peu convaincants. Par exemple, le meilleur score obtenu pour le défi MSFG 2020 se trouve aux alentours de 15%.

Malgré toutes les difficultés liées à la conception d'un système complet d'analyse de ME, dans ce chapitre nous voulons effectuer une étude exploratoire pour associer le spotting à la reconnaissance et ceci dans des conditions les plus proches de l'application réelle.

Devoir travailler avec de longues vidéos, sans aucun à priori sur le nombre de ME ni sur leurs positions, engendre un nombre d'erreurs trop important avec les algorithmes actuels pour pouvoir envisager de les utiliser pour concevoir un système complet apte à être déployé en conditions réelles. Pour notre étude exploratoire, nous avons décidé de contourner ces difficultés en proposant une simplification forte mais cohérente avec l'application recherchée. Plutôt que de proposer un nouveau descripteur pour lever ces verrous, nous intégrons la méthode présentée dans le chapitre précédent dans une étude globale de l'application.

De par leur nature, les ME ont tendance à apparaître plus fréquemment dans des situations stressantes ou en relation directe avec un dialogue, un son ou une image. Les sentiments intrinsèques vis-à-vis d'un contexte sont très difficiles à estimer automatiquement mais ils donnent un a priori fort sur la présence de ME. Devant une situation ou une ambiance spécifique, un sujet pourrait générer une ME qui traduit son émotion ressentie.

Rappelons que la ME est une représentation ponctuelle de l'émotion et non d'un état émotionnelle de la personne sur la durée. Il est donc possible de supposer qu'un être humain puisse estimer l'apparition d'une ME. Un homme ayant l'acuité visuelle et la vitesse de réaction et de penser pour trouver l'apex ou l'offset aurait probablement aussi la capacité d'inférer de par lui-même sur la nature de la ME et n'aurait donc pas particulièrement besoin du système d'aide à la classification de ME. Mais cette partie de la population reste très peu nombreuse si elle existe.

Contrairement à l'offset et l'apex, l'onset pourrait être déduit du contexte d'une situation par la plupart des personne. Il est donc plus pragmatique de partir du principe que l'utilisateur ne peut qu'estimer le temps de début de la ME. Cela ne garantit pas une détection exhaustive des ME mais permet de limiter grandement l'espace de recherche.

En pratique dans ce chapitre, nous supposerons que le début des ME (onset) ainsi que leur nombre sont connus. Mais pour appliquer la méthode de reconnaissance dans un système complet, il faut encore estimer la position de l'apex. Dans la section 5.1, nous introduirons plusieurs méthodes de bas-niveaux pour détecter l'apex dans ce nouveau contexte. Ensuite, nous évaluerons l'efficacité à la fois au niveau du spotting seul, mais aussi au niveau du système complet dans la section 5.2. Notre objectif est ici d'évaluer l'incidence des défauts d'estimation du spotting sur le protocole complet qui répond au seul besoin pratique.

En section 5.3, nous allons présenter et comparer plusieurs implémentations du système complet proposé sur une cible matériel afin de tester son utilité pratique. Sur ce système, l'utilisateur actionnera un bouton pour estimer l'apex et recevra une classification de la ME depuis la cible électronique.

5.1 Méthode de pseudo-spotting

Dans cette section, nous partons du postulat que nous connaissons le début (l'onset) de chaque ME. Donc, le nombre de ME est aussi connu et ceci réduit drastiquement le nombre de segments d'une séquence vidéo à étudier. Pour chaque ME, il reste maintenant à estimer les moments d'occurrence de l'apex et de l'offset pour le spotting. En pratique, nous avons vu dans les chapitres précédents que les descripteurs utilisés pour la reconnaissance de ME se basent principalement sur la connaissance de l'apex mais très rarement sur celle de l'offset. C'est pour cette raison, nous nous sommes concentré sur la seule détection de l'apex par la suite.

5.1.1 Métriques d'évaluation du spotting

Puisque nous voulons établir un système complet, c'est-à-dire le spotting suivi par la reconnaissance de ME, l'évaluation des performances s'effectue à la sortie de leur association et non à celle de chacune de ces deux étapes. L'idée sous-jacente est de pondérer le jugement des écarts du spotting obtenus face à l'attendu par une mise en condition pratique.

Les deux problématiques, le spotting et la reconnaissance, répondant à des difficultés et des contraintes différentes, ils ne peuvent être évalués de façon identique. Dans le cas de la reconnaissance, il y a de multiples classes et chacune doit être approximativement représentée dans les tests de façon équitable. Bien que ce ne soit, en pratique, pas toujours le cas (nous en parlerons dans le chapitre 6), l'accuracy (ou exactitude) s'impose alors comme un critère pertinent :

$$\text{accuracy} = \frac{VC}{VC + FC} \tag{5.1}$$

où VC est le nombre de classifications justes et FC le nombre de fausses classifications. En effet le taux de bonne reconnaissance est descriptif de l'application associée.

Par contre, l'accuracy est un critère insuffisant pour évaluer le spotting. En effet, rappelons qu'il n'y a que deux classes et que leurs occurrences respectives dans les tests ne sont absolument pas équilibrées. Sur les séquences vidéo de la base de données CASME II, les ME représentent moins de 0.4% du temps.

Sur une classification binaire, le critère associé est souvent la précision (Equation 5.2) proche de l'accuracy. Mais de même ce critère n'est pas pertinent au vu du déséquilibre déjà évoqué. Imaginons le cas d'un descripteur associant "pas une ME" à toutes les périodes de temps testées. Alors un score de précision très élevé (beaucoup de bonnes détections) serait obtenu alors qu'aucune ME ne serait détectée.

Il faut donc aussi associer le rappel (Equation 5.3) qui assure de ne pas oublier de vrai positif. Pour combiner les deux notions, un autre critère est introduit dans l'analyse statistique de la classification binaire : le F-score ou $F_{measure}$ (Equation 5.4) vise à favoriser le compromis entre les précision et rappel.

$$\text{précision} = \frac{VP}{VP + FN} \tag{5.2}$$

$$\text{rappel} = \frac{VP}{VP + FP} \tag{5.3}$$

$$F_{measure} = \frac{2.\text{précision}.\text{rappel}}{\text{précision} + \text{rappel}} \tag{5.4}$$

où VP correspond au nombre de vrais positifs (séquence contenant une ME et estimée comme une ME), FN correspond au nombre de faux négatifs (séquence contenant une ME mais non estimée comme une ME) et FP correspond au nombre de faux positifs (séquence ne contenant pas de ME mais estimée comme une ME).

Sur les algorithmes de spotting les plus élaborées, la $F_{measure}$ peinent à surpasser 15% [124] même sur des bases de données composées conjointement de macros-expressions et de ME. Pour mesurer le score final de la chaîne de traitement proposée, nous devons mesurer les performances de reconnaissance sans utiliser sur la base de données les informations relatives ni à la taille de la ME ni à la position de l'apex.

Dans la section 5.1.2, nous introduirons une étude centrée autour de la problématique de spotting. Connaissant l'onset nous cherchons à estimer au plus juste la position de l'apex. L'objectif est alors de minimiser l'écart temporel entre l'instant d'apparition de l'apex et son estimation.

5.1.2 Approximation de la position de l'apex

Pour estimer la position de l'apex, celle de l'onset étant connue, nous avons testé un certain nombre de méthodes classiques. Il s'agit de descripteurs de bas niveau calculés sur chaque frame à partir de l'onset sur une durée supérieure à celle d'une ME. Il faut ensuite introduire une distance qui va comparer le descripteur obtenu pour l'onset avec celui obtenu pour les autres frames.

Le balayage de la vidéo engendre alors une courbe représentant l'évolution de la déformation du visage lors de la ME. La plus grande variation correspond à l'expression la plus importante de la ME, c'est à dire l'apex. La valeur fournie par la distance pour une frame correspond à sa vraisemblance en tant qu'apex.

Le LBP [104, 125] et ses variants sont les outils de base pour décrire une expression comme nous l'avons vu dans le chapitre 2. C'est pourquoi il est souvent utilisé non seulement pour la classification [122, 121], mais aussi pour le spotting [124]. Il est ainsi naturel de le tester pour notre approche. Le LBP est un descripteur statique qui s'applique sur une seule image.

Comme les ME consistent en une déformation du visage, il est aussi logique de s'intéresser au mouvement. Le flot optique [108] est alors tout indiqué car il propose une estimation localisée du mouvement. C'est pour cette raison que nous l'avions utilisé dans le chapitre 4. Nous calculerons par la suite en tant que descripteur lié à une frame, le flot optique entre l'onset et cette frame.

Nous introduisons quatre méthodes d'estimation de l'apex :

— χ^2 **du LBP** : Avec l'Histogramme du LBP (HLBP), l'image est découpée en blocs (3 × 3 dans notre étude). Chaque pixel produit un code représentatif du motif généré par la comparaison de ses voisins avec lui-même. Pour chaque bloc, un histogramme d'occurrence des différents codes possibles est alors calculé. La forme, et surtout l'évolution des contours, sont ainsi décrits. Le descripteur final est alors la concaténation de ces histogrammes sur l'ensemble des blocs. La distance que nous utilisons pour définir la frame F est :

$$\chi^2_{HLBP}(F) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(HLBP_F(i) - HLBP_{onset}(i)\right)^2}{HLBP_F(i) + HLBP_{onset}(i)} \tag{5.5}$$

où N est la dimension du descripteur HLBP et $HLBP_F(i)$ la $i^{\text{ème}}$ valeur du descripteur obtenue pour la frame F.

— **corrcoefLBP** : Pour définir un indice de corrélation plus fin entre les valeurs du descripteur LBP pour l'onset et pour la frame F, et ainsi déterminer une distance

pertinente, nous pouvons aussi utiliser le coefficient de corrélation corrcoef :

$$corrcoef_{LBP}(F) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{HLBP_F(i)-\mu_F}{\sigma_F}\right)\left(\frac{HLBP_{onset}(i)-\mu_{onset}}{\sigma_{onset}}\right) \quad (5.6)$$

où μ_F correspond à la moyenne des valeurs du descripteur obtenu sur la frame F et σ_F à son écart-type.

— **sommeFO** : Le calcul du flot optique donne une valeur 2D du mouvement par pixel (un vecteur). Cette information est précise et pertinente mais de très haute dimensionnalité. Pour cette méthode, l'intensité du flot optique pour chaque pixel, puis la moyenne des intensités sur l'ensemble de l'image sont calculés. Il n'y a alors plus qu'une valeur en sortie. Nous perdons de cette manière toute information de localisation du mouvement. Mais cette valeur nous renseigne sur la quantité de mouvement. L'idée ici et de vérifier la présence du mouvement.

$$somme_{FO}(F) = \sum_{p\in F} FO(p) \quad (5.7)$$

où $FO(p)$ est l'intensité du flot optique calculée pour le pixel p de la frame F.

— χ^2 **histFO** : Les descripteurs HOG [109] sont très efficaces pour la représentation des formes (silhouettes) dans une image. Le HOOF [40] (Histogramm of Oriented Optical Flow) est une variation des descripteurs HOG adaptée au mouvement. Nous l'utiliserons ici pour décrire la déformation du visage. L'image est découpée en blocs (3×3 dans notre étude) dans lesquels la fréquence d'occurrence pour chaque orientation du vecteur de flot optique relatif aux pixels contenus dans ce bloc est calculée. Un histogramme (une valeur par orientation) est alors obtenu. La concaténation des histogrammes de chaque bloc génère le descripteur. La représentation du mouvement est moins précise spatialement mais plus caractéristique d'un mouvement significatif. Pour avoir une mesure de distance, le χ^2 du descripteur est effectué :

$$\chi^2_{HOOF}(F) = \sum_{i=1}^{N}\frac{\left(HOOF_F(i)-HOOF_{onset}(i)\right)^2}{HOOF_F(i)+HOOF_{onset}(i)} \quad (5.8)$$

où N est la dimension du descripteur HOOF et $HOOF(i)$ sa $i^{\text{ème}}$ valeur.

5.1.3 Évaluation de l'estimation de l'apex

Pour évaluer l'efficacité de ces méthodes et les comparer, nous les avons testé sur les vidéos de la base de données composite CASME II & SAMM. Cette base de données fournit l'onset et l'apex de chaque ME. L'onset sert alors de frame de référence et l'apex est considéré comme la vérité de terrain.

Compte-tenu de la durée maximum d'une ME et de la fréquence d'acquisition (framerate) des caméras utilisées lors de l'enregistrement des bases de données, nous prenons en compte 50 frames successifs pour notre évaluation. Pour chacune, la valeur de vraisemblance est calculée et la frame donnant le score le plus élevé est considérée comme l'apex estimé. De cette manière, une valeur d'apex est estimée pour chaque ME de la base de données.

La position de l'apex est très difficile à estimer précisément même manuellement à l'œil nu. En effet elle correspond à la plus grande expression de la ME qui peut s'étendre sur plusieurs frames. Pour cette raison nous accordons une tolérance. Pour une analyse plus fine et une meilleure représentation des capacités d'estimation de nos algorithmes, nous calculons des courbes ROC illustrant le pourcentage de ME dont l'apex a été bien estimé en fonction de la tolérance fixée. Ces courbes sont exprimées par l'équation suivante :

$$ROC(d) = \frac{1}{N_{ME}} \{m \in E_{ME} \mid |a_m - \widehat{a}_m| < d\} \tag{5.9}$$

où $\{E\}$ est la cardinalité de l'ensemble E, N_{ME} et le nombre de ME dans l'ensemble E_{ME} de la base de données, a_m est l'apex réel de la ME m et $\widehat{a}_m$ son apex estimé.

L'évolution des capacités de détection peut ainsi être analysée en fonction de la tolérance que l'on accepte. Avec une forte tolérance le taux de réussite sera plus élevé mais l'intérêt pratique sera moindre. La Figure 5.2 affiche les évolutions obtenues pour les quatre méthodes présentées précédemment. Nous pouvons observer que la somme du flot optique donne des performances bien inférieures aux autres méthodes. C'était à anticiper car il s'agit d'une représentation très simple (juste le quantité moyenne de mouvement dans l'image). Il est néanmoins rassurant de constater que ce n'est pas juste le mouvement mais la localisation du mouvement qui définit la position de l'apex.

Au niveau du LBP, nous pouvons constater que le χ^2 est plus performant que la corrélation. Nous pouvons en conclure qu'il n'y a pas vraiment une composante significative relative à la composition globale de la frame mais qu'au contraire la représentation en code du LBP plus que son évolution est suffisante pour décrire son contenu.

Les HOOF donnent finalement des résultats très proches du χ^2 des LBP. La description proposée par ces deux méthodes sont pourtant très différentes : la forme contre le mouvement. Ces deux informations semblent donc aussi efficaces.

Dans ce type de courbes, le point d'inflexion définit généralement le compromis le plus avantageux. Ici nous obtenons un point d'inflexion autour d'une tolérance de

30 frames. Avec notre framerate cela représente 150ms. Or la durée moyenne d'une ME est égale à 170ms sur la base de données CASME II & SAMM. En pratique cette tolérance n'est donc pas acceptable.

Prenons une tolérance de 10 frames qui est plus raisonnable. Le χ^2 du HOOF et du LBP donnent alors un taux de réussite inférieur à 40%. C'est une valeur très basse qui s'explique par la grande variation et l'extrême subtilité du mouvement des ME. D'ailleurs, l'évolution au cours du temps des vraisemblances obtenues avec les quatre méthodes est peu régulière. L'estimation du flot optique par exemple est même assez bruitée.

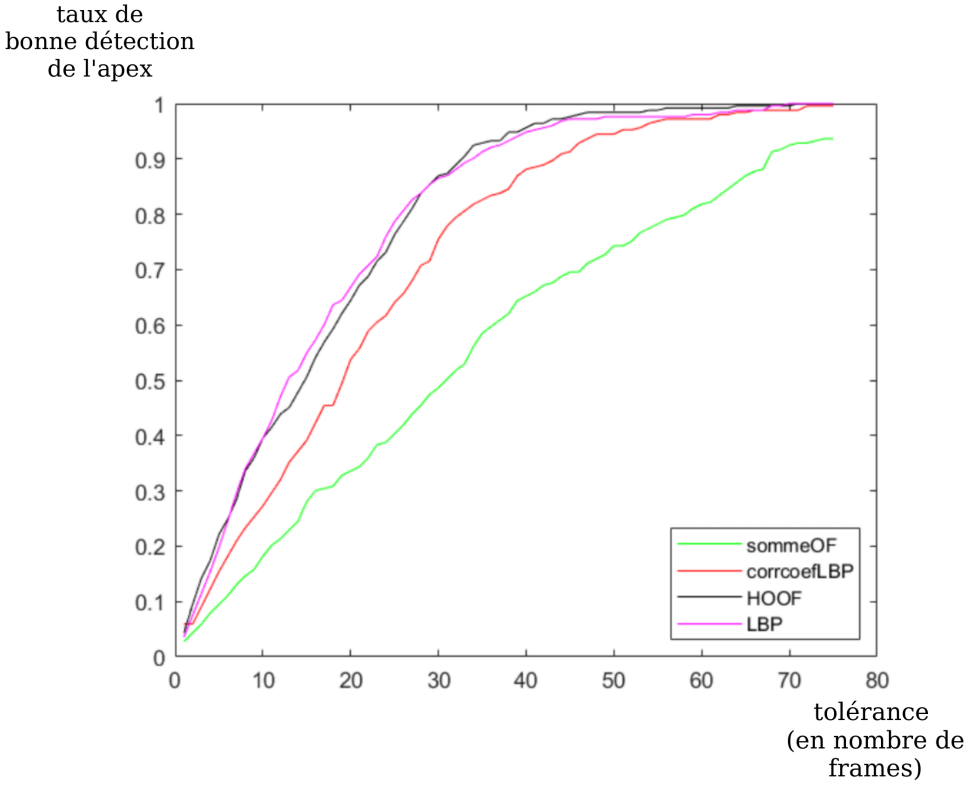

FIGURE 5.2 – Évolution du taux de bonne détection de l'apex en fonction de la tolérance pour les quatre méthodes de détection proposées.

À partir de ces constats, nous avons alors essayé d'inclure une donnée statistique en prenant en compte la durée moyenne d'une micro expression t_{ME}. Les mêmes tests que précédemment ont ensuite été effectués en prenant comme apex estimé la position $\widehat{t}_{apex} = t_{onset} + t_{ME}$. Nous nommerons par la suite cette méthode DMME (durée moyenne d'une ME). La Figure 5.3 affiche une comparaison entre cette méthode et les quatre précédemment développées.

Cette nouvelle méthode donne des performances bien supérieures à celles obtenues avec des descripteurs. Pour une tolérance de 10 frames, elle obtient un taux de reconnaissance de 70%. Cela confirme le fait que les descripteurs sont trop

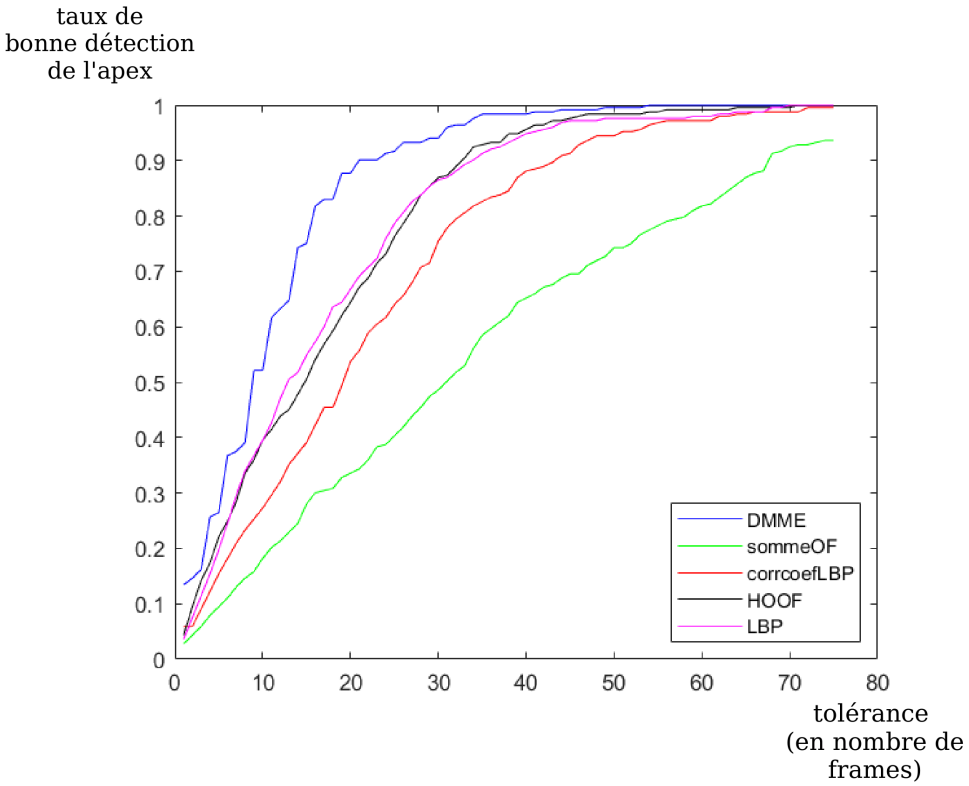

FIGURE 5.3 – Comparaison entre les taux de réussite obtenus tout d'abord avec les quatre méthodes précédentes puis en supposant une durée fixe (t_{ME}) entre l'onset et l'apex estimé.

imprécis et bien trop sensibles à diverses sources de bruit. De plus cette méthode est bien moins coûteuse en calcul.

Nous avons vu que l'estimation automatique de la position de l'apex est loin d'être précise. Une désignation statistique sans prise en compte des caractéristiques particulières à la séquence donne même les meilleurs résultats. Ces résultats semblent très insuffisants. Mais ce n'est pas le spotting en lui-même qui est le cœur de notre étude. Des écarts d'estimation ne sont pas vraiment problématiques en soi s'ils n'entraînent pas une dégradation significative de la classification qui est notre objectif principal. Dans la section suivante, nous allons présenter une étude sur le protocole complet. L'évaluation portera alors sur l'association spotting-classification (la classification prenant en tant qu'entrées les sorties du spotting). En attendant la découverte d'une méthode efficace de spotting, il est important d'évaluer son impact sur la performance de reconnaissance.

5.2 Pipeline complet pour l'analyse de ME

Dans cette partie, nous associons les étapes de spotting et de classification pour constituer un pipeline complet pour l'analyse de ME. Rappelons que la classification prend, en entrée d'un réseau CNN, le flot optique calculé entre l'onset et l'apex.

Nous estimons les apex à partir des 5 méthodes présentées précédemment. Nous les utilisons ensuite en reprenant le processus de classification du chapitre 4 avec la configuration qui procurait les meilleurs performances : le CNN qui considère Vy comme entrée avec 3 couches convolutionnelles.

Partir d'une estimation de l'apex proche de celle réelle est forcément optimale pour maximiser les performances de la classification. Mais ici notre objectif est de quantifier la dégradation au niveau de la classification du fait d'une estimation automatique et donc moins précise de l'apex.

La Figure 5.4 montre les matrices de confusion et les scores d'accuracy obtenus à partir des différentes méthodes d'estimation de l'apex. Puisque la méthode de correcoef_LBP a donné le plus mauvais résultat, elle n'est pas considérée dans la comparaison.

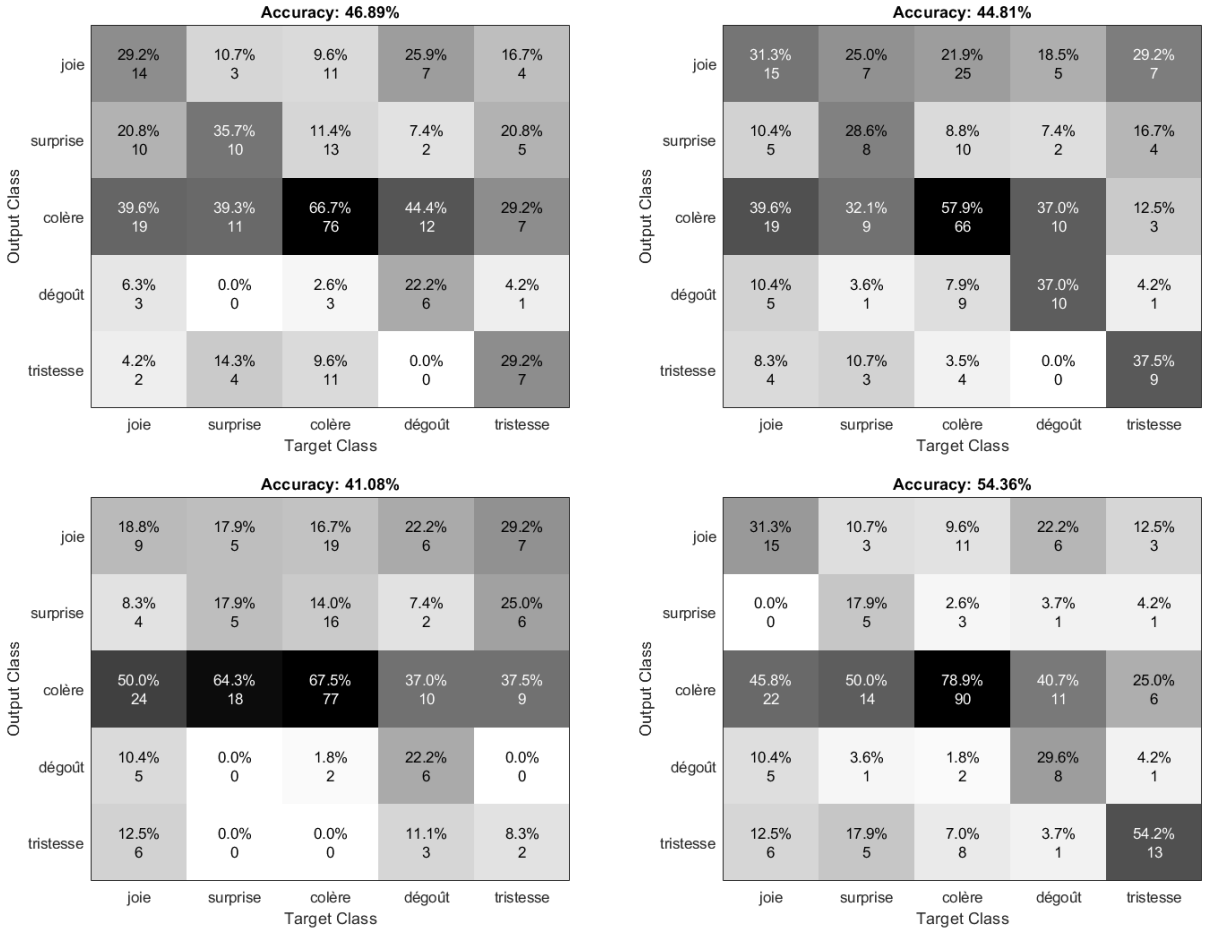

FIGURE 5.4 – Matrices de confusion obtenues sur le système complet à partir de quatre méthodes d'estimation de l'apex : (en haut, à droite) le χ^2 du LBP, (en haut, à gauche) le χ^2 HOOF, (en bas, à droite) la somme du flot optique et (en bas, à gauche) la durée moyenne d'une ME (DMME).

Concentrons-nous tout d'abord uniquement sur les scores d'accuracy. Ils

confirment les remarques déjà observées sur l'évaluation de la détection de l'apex réalisée en 5.1.3 : le DMME donne les meilleurs résultats et la somme du flot optique obtient les moins bons. Cependant nous remarquons cette fois une distinction plus nette entre les HOOF et le LBP. La description basée sur la déformation de silhouette semble donc plus fructueuse.

L'écart entre la somme du flot optique et les autres méthodes est ici moins prononcé. Cela est probablement dû à la difficulté croissante d'amélioration du critère (une amélioration de 1% est plus significatif vers les hautes valeurs que vers les basses). Malgré cet aspect, l'amélioration apportée par le DMME est confirmée. Cette donnée statistique (la durée moyenne d'une ME) est donc bien significative.

À titre de comparaison, le score atteint en utilisant un apex détecté manuellement est proche de 60.17%. Le score de 54.36% en utilisant le DMME est inférieur mais reste proche des performances atteintes par l'être humain en ce qui concerne la détection et la reconnaissance de ME. Bien sûr, il reste des cas où l'apex réel est bien loin de celui estimé par une durée moyenne. Mais ces cas particuliers sont relativement rares.

Les performances en classification sont-elles donc directement et uniquement liées à la distance à l'apex ? Pour obtenir la Figure 5.5, nous définissons un décalage fixe (de 0 à 30 frames) puis nous calculons le score d'accuracy obtenu avec une estimation de l'apex correspondant à sa valeur réelle à laquelle ce décalage est ajouté. Ainsi nous observons l'effet moyen de la dégradation en classification en fonction de la dégradation en estimation de l'apex. La variation entre l'onset réel et l'apex estimé par DMME est de 11 frames. Sur la courbe nous observons qu'une dégradation de 11 frames provoque en moyenne une accuracy en classification de 53,11%. Cependant le score d'accuracy obtenu sur le pipeline complet avec le DMME est de 54.36%. Les deux valeurs sont proches. Nous pouvons en conclure que l'estimation statistique par DMME est cohérente ; ce qui valide son utilisation.

Considérons maintenant les matrices de confusion. La méthode des LBP donne une très bonne répartition des résultats qui caractérise la robustesse de la classification. C'est à dire que le score de reconnaissance de chaque classe est suffisamment correct et équilibré. Les HOOF par exemple ont une reconnaissance efficace principalement sur la classe "colère" qui est la plus représentée. La supériorité du HOOF sur le LBP au niveau de l'accuracy est donc à mettre en perspective : l'amélioration étant plus provoquée par la caractérisation de la base de test que de l'efficacité de l'algorithme.

Il en est de même pour le DMME qui catégorise très mal la surprise (la classe la moins bien représentée dans la base de test). La méthode DMME se caractérisant par une durée statistique, il est cependant logique qu'elle profite de la composition globale de la base de test.

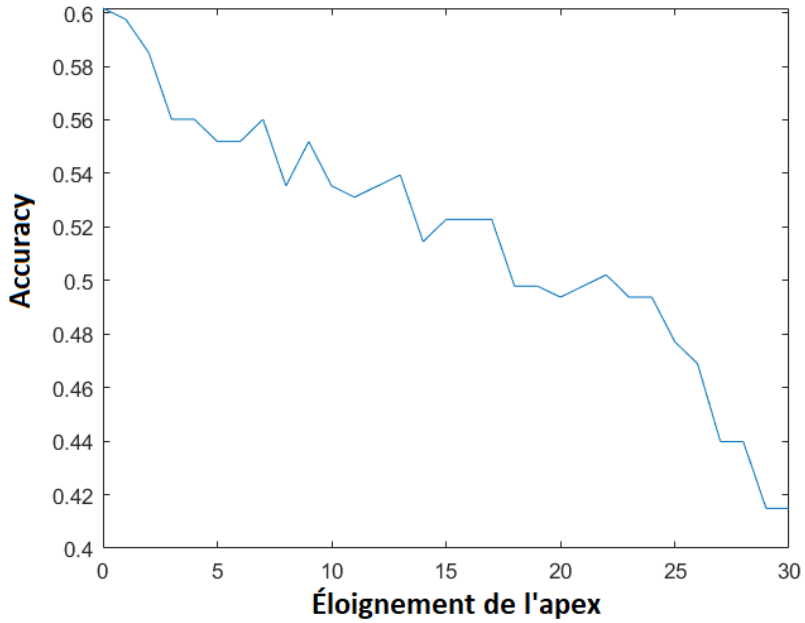

FIGURE 5.5 – Effet moyen de la dégradation de la classification en fonction de la dégradation de l'estimation de l'apex.

La somme du flot optique est la méthode qui produit la répartition la plus déséquilibrée avec la classe "colère" majoritaire sur la reconnaissance de chaque classe. L'influence du déséquilibre de la base de test sera discuté dans le chapitre 6.

5.3 Implantation matérielle

Dans les sections précédentes, nous avons présenté une chaîne complète couplant le processus de détection de l'apex à la classification de ME pour proposer une méthodologie prête à l'emploi (proche des scénarios réels). Nous avons montré que le frame se trouvant 170ms (soit la durée moyenne d'une ME) après l'onset est la méthode d'estimation de l'apex la plus appropriée à la fois en termes d'efficacité de la classification qu'en terme de temps de calcul et de complexité. Nous avons vérifié que, malgré la difficulté de l'étape de spotting, cette chaîne d'analyse de ME permet d'avoir des résultats relativement corrects. Pour ces expériences, nous supposons connaitre l'onset, ce qui est une hypothèse acceptable puisque l'utilisateur peut anticiper l'apparition d'une ME selon le contexte de l'application.

Par la suite, nous allons décrire une implantation matérielle pour démontrer l'usage pratique de l'approche proposée. L'objectif est d'avoir un système temps-réel et embarqué qui soit utilisable et efficace en conditions réelles.

5.3.1 Présentation du système

La cible matérielle consiste en un microcontrôleur Raspberry Pi 4 Modèle B (4B) 4Go avec une alimentation 5,1V (3A). La carte électronique est connectée à une caméra LABISTS B01 Module 5M 1080 P qui possède trois options pour capturer les images avec un framerate et une résolution prédéfinie : 30 fps (1080 P), 60 fps (720P) et 90 fps (480 P). Nous utilisons la dernière configuration, c'est-à-dire que l'acquisition se fait en 640 x 480 pixels avec une vitesse de 90 images par seconde.

L'utilisateur participe au processus puisque c'est lui qui détermine le début de la ME. En effet il connaît l'élément déclencheur qui doit générer la ME. Il faut donc lui fournir le moyen d'en informer le système. Dans cet optique, l'action d'une touche lui permet de générer un signal au processus.

Les actions s'enchaînent selon plusieurs étapes affichées sur la Figure 5.6. Une personne (le sujet) se place en face de la caméra puis lance l'exécution de l'algorithme sur le système. La caméra capture les informations tandis que le visage du sujet est affiché en continu. Le système se met alors en attente d'une entrée de la part de l'utilisateur.

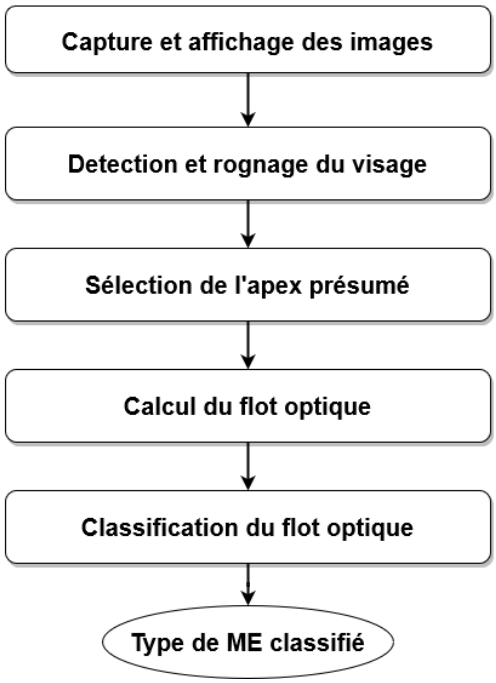

FIGURE 5.6 – La chaîne des actions à réaliser par le processus.

Lorsque celui-ci juge qu'une ME est susceptible de se produire (après avoir posé une question lors d'un moment de tension par exemple), il donne le signal à l'application. La position du visage du sujet est estimée par la méthode de Viola-Jones [126]. La frame est recoupée autour du visage détecté afin de recentrer l'étude. Cette capture est sauvegardée et considérée comme l'onset de la ME. Elle servira de

référence comme état neutre du visage.

Une nouvelle capture est réalisée 170ms après la première selon le même protocole. Elle représente l'apex estimé. Le calcul du flot optique est réalisé entre ces deux images. Le résultat obtenu est alors entré dans le réseau CNN présenté dans le chapitre 4. La sortie du réseau donne le type de ME reconnu qui produit un signal renvoyé à l'utilisateur. Nous avons programmé ces tâches sous Unix avec le langage Python. Une exécution séquentielle nous permet de mesurer le temps nécessaire pour chaque partie de la chaîne de traitement (voir le Tableau 5.1).

capture	affichage	détection visage	calcul flot optique	classification CNN
0,0111s	0,0344s	0,0108s	0,0241s	0,3126s

TABLE 5.1 – Répartition du temps nécessaire pour chaque étape de la chaîne complète.

5.3.2 Implémentation en parallèle

La carte Raspeberry 4B que nous utilisons supporte très bien la parallélisation. Avec son CPU à 4 cœurs, elle peut réellement soutenir 4 threads matériels à la fois (les 4 threads s'exécutent de façon purement parallèle et indépendante, chacun des cœurs faisant les calculs séparément). À cela s'ajoute la capacité d'exécuter des threads software en portant l'attention de chaque cœur CPU sur plusieurs threads selon l'ordonnancement décidé par le système d'exploitation.

Pour cette première implémentation en parallèle, nous exécutons simultanément les trois étapes principales nécessaires à la classification des ME d'une manière pipeline : la capture d'image, l'affichage des différentes informations essentielles à l'utilisateur et le traitement (sélection de l'apex, calcul du flot optique et classification). Ce système d'exécution multi-thread est présenté sur la Figure 5.7.

Nous avons estimé dans le chapitre 5.1.3 que la durée moyenne entre l'onset et l'apex d'une ME est de 170ms. Avec un framerate de 90 FPS, nous pouvons capturer une frame toutes les 11ms. Il nous est donc possible de capturer la frame correspondant à l'apex présumé à 6ms prés.

Ce système est plus confortable avec affichage des éléments contextuels et des informations sur l'exécution des tâches. Cela peut en effet garantir le bon usage des données : une acquisition correcte, une personne cible bien placée ou son visage bien détecté. Une situation basée sur cet affichage où l'utilisateur du système déclenche toute l'opération est envisageable. La visualisation de la personne cible est alors nécessaire pour définir le signal contextuel correspondant à l'onset.

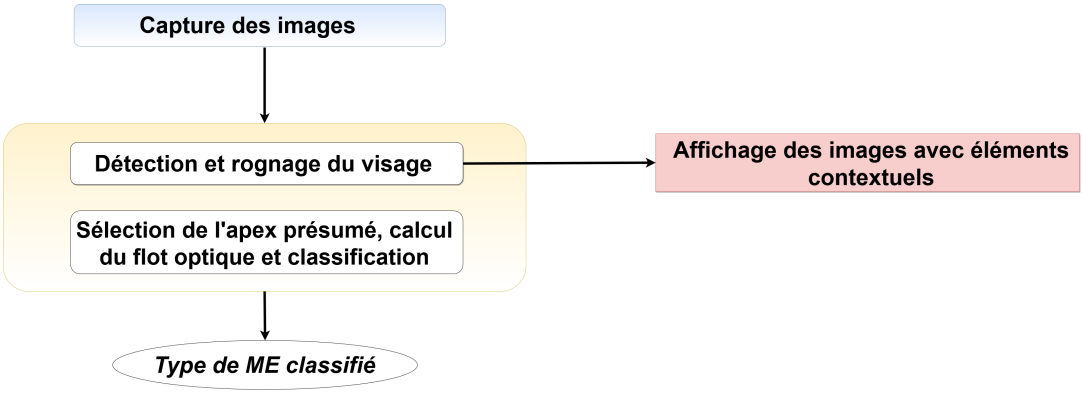

FIGURE 5.7 – La chaîne d'exécution du système en multi-thread : chaque rectangle de couleur représente un thread.

Pour optimiser l'économie d'énergie et la rapidité d'exécution, le retour à l'utilisateur peut se limiter à un signal ponctuel correspondant à la classe de l'émotion reconnue. Le programme reste en attente du signal du départ de l'utilisateur (l'onset), lance l'exécution de la classification en arrière-plan et fournit juste un signal de retour. Pour cette première implémentation en parallèle, le temps de la classification d'une ME correspond à 523 ms (voir le Tableau 5.1 et la section suivante).

5.3.3 Amélioration du système

Pour augmenter l'ergonomie du système, nous avons ensuite réalisé une version avec affichage des informations supplémentaires. La Figure 5.8 illustre le fonctionnement du pipeline. Une représentation du système est visible sur la Figure 5.9.

Nous mettons en place trois fenêtres. La première affiche le visage du sujet en temps réel et en continu. Un cadre est ajouté autour du visage pour vérifier que cette étape est bien réalisée. La seconde fenêtre est activée lorsque l'utilisateur enclenche le signal. À ce moment, elle affiche le visage du sujet, en faisant défiler les frames au ralenti, jusqu'à la sélection de la frame correspondant à l'apex estimée. Le résultat de la classification s'affiche alors à son tour. La troisième fenêtre quant à elle affiche le flot optique calculé dans la zone du visage.

La capture des images s'effectue avec une fréquence de 90 FPS (soit une période de 11ms). L'apex estimé arrive à la 16ème frame suivant l'onset ($16 \times 11ms = 176ms$). Le traitement suivant (c'est à dire la détection du visage, le calcul du flot optique et la classification avec CNN) est opéré quant à lui sur une période de 347ms (à titre de comparaison, Resnet18 nécessite 1,3 s).Le temps complet pris pour la classification d'une ME par notre système est donc de 523 ms. Avec une exécution aussi rapide, nous pouvons très bien envisager le déploiement et l'utilisation de ce système en conditions réelles lors d'un entretien afin de servir d'outil d'aide à la reconnaissance de ME.

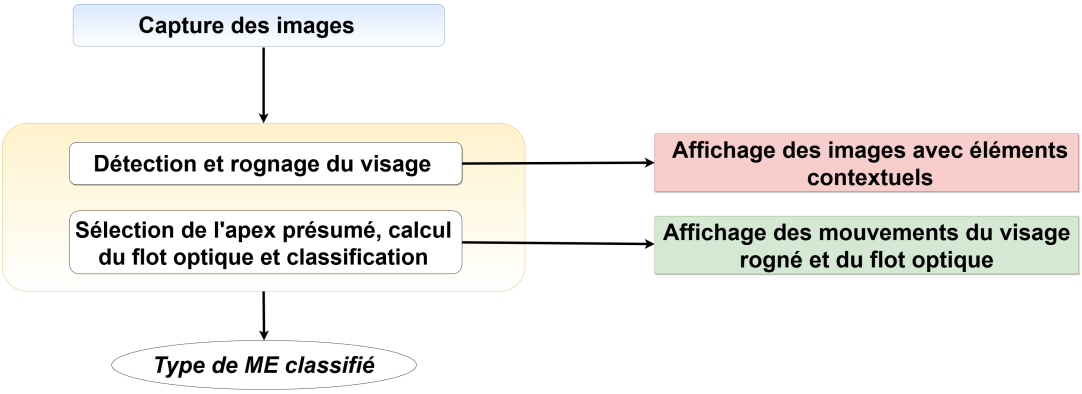

FIGURE 5.8 – La chaîne d'exécution du système en multi-thread avec affichage des informations supplémentaires.

5.4 Conclusion

Un système complet couple les étapes de spotting et de classification. Nous avons observé une incidence directe entre la précision de l'estimation de l'apex et les performances de la classification. Motivé par le niveau de performance des méthodes de la littérature pour le spotting de vidéos longues, nous avons étudié la possibilité d'implémenter une méthode de pseudo-spotting sur des ME pré-localisées. Ce problème plus simple reste en adéquation avec notre application pratique. Après comparaisons des performances obtenues à partir des méthodes classiques, nous avons constaté qu'une approche statistique basée sur la durée moyenne d'une ME nous procure les meilleurs résultats. Nous en concluons que cet a-priori statistique est plus déterminant que les descriptions du mouvement étudiées.

Nous avons ensuite implémenté le système complet sur une carte Raspberry Pi 4 modèle B dans le but de faire tourner le pipeline en temps réel sur un système embarqué avec une consommation en mémoire réduite. Une parallélisation de l'utilisation des ressources est nécessaire pour garantir une vitesse d'exécution suffisante pour capturer une ME. Notre système fonctionne alors selon les conditions recherchées, et ce même en utilisant une partie de la puissance de calcul pour le confort de l'utilisateur en ajoutant l'affichage des signaux visuels.

Il serait intéressant d'approfondir cette étude notamment en incorporant des caméras plus rapides et en optimisant le système implémenté. Cependant nous avons démontré ici la faisabilité d'utiliser nos algorithmes au niveau d'une application réelle.

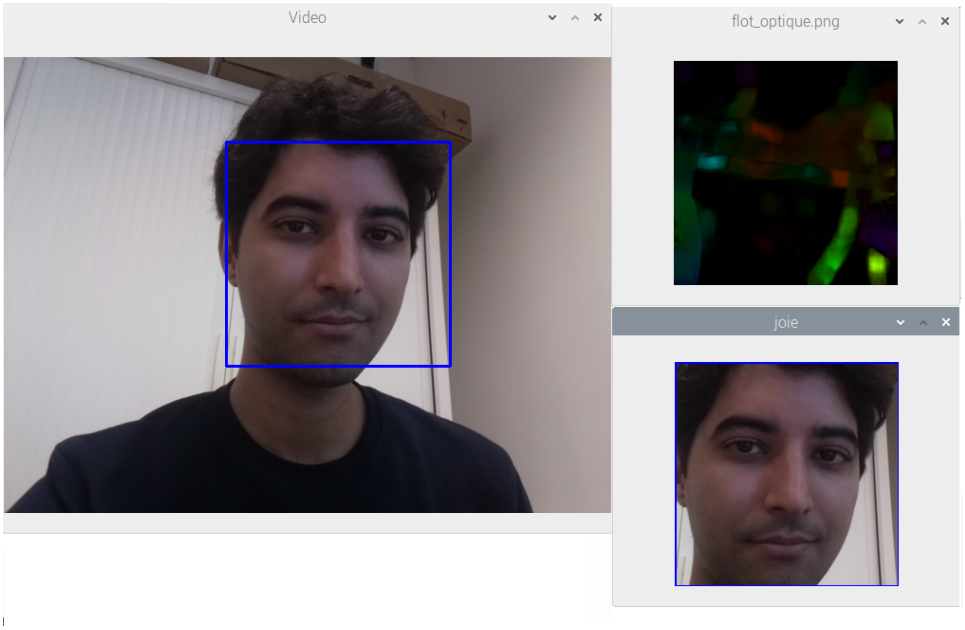

FIGURE 5.9 – Notre système génère trois fenêtres : la première restitue le flux vidéo auquel est ajouté un cadre englobant autour du visage ; la seconde donne le résultat de la classification ; enfin la dernière affiche le flot optique.

Chapitre 6

Discussions sur l'utilisation des données

Nous avons, dans les chapitres précédant, proposé un certain nombre de méthodes de classification des ME. Pour évaluer leur capacité et les comparer à l'état de l'art, nous avons suivi le protocole d'évaluation traditionnellement utilisé dans les challenges (comme le Micro-Expressions Grand Challenge [119]) et la littérature. Dans ce chapitre, nous présentons une réflexion sur les limites de ce protocole d'évaluation. L'objectif est de prendre du recul sur les résultats que nous avons présenté pour en tirer la plus juste signification.

Le domaine des ME étant particulièrement complexe, nous proposons de nous éloigner des expertises classiques à la classification afin d'analyser les résultats à travers le prisme de notre application. Le premier point à prendre en compte est l'évaluation de la classification de façon isolée en connaissance de la position des ME. La connaissance du début (onset), de la fin (offset) voire même du moment d'intensité maximale de l'expression (apex) est une hypothèse forte.

Nous avons déjà bien détaillé les effets de ce premier point dans le chapitre 5 en étudiant un système complet spotting/classification. Nous avons notamment vu que le manque de précision au niveau de l'étape de spotting entraînait une dégradation significative des résultats de reconnaissance.

Le second point à prendre en considération est le contenu des bases de données utilisées (SAMM et CASMEII). Pour l'évaluation, les échantillons sont répartis en deux ensembles. L'ensemble d'apprentissage permet de créer le modèle de classification. L'ensemble de test permet d'évaluer la méthode.

Un échantillon de ces bases de données est une séquence contenant une ME annotée avec le type d'émotion associée : joie, tristesse, colère, surprise et dégoût. Ces annotations servent à la fois de définition des classes recherchées sur l'ensemble d'apprentissage mais aussi de vérité de terrain sur l'ensemble de test. Le Tableau 6.1 donne la répartition des échantillons pour chacune des classes à l'intérieur des prin-

cipales bases de données. Ces bases présentent un net déséquilibre. Par exemple, la classe *colère* représente 47.30% des échantillons sur les deux bases alors que la classe *tristesse* n'en représente que 9.96%.

Base de données	Joie	Surprise	Colère	Dégoût	Tristesse
CASME II [100]	13.71%	8.57%	54.86%	10.86%	12.00%
SAMM [5]	35.82%	19.40%	28.36%	11.94%	4.48%
CASME II + SAMM [127]	19.92%	11.62%	47.30%	11.20%	9.96%

TABLE 6.1 – Répartition des 5 différents types d'émotion dans les bases de données CASMEII [100] et SAMM [5] et leur union.

Pourquoi un tel déséquilibre ? La création d'une base de données de ME est très délicate. En effet les ME sont instinctives et ne peuvent être simulées : impossible donc d'utiliser un acteur pour générer des expressions faciales à la demande en nombre régulier et maîtrisé. Il faut alors déclencher une émotion à partir d'un stimulus et rien ne garantit que cela produira une ME. De plus, la classe de la ME n'est pas désignée à partir du stimulus mais de l'analyse manuelle des muscles mis à contributions (les AUs vus dans la sous-section 2.1). Ainsi, même avec un usage adapté des stimuli, le nombre et la répartition des ME ne sont pas maîtrisables.

L'évaluation peut alors se réaliser de deux façons : soit la base de données est utilisée dans son intégralité malgré un équilibre loin d'être assuré ; soit une sélection de cette base la plus grande possible est préalablement réalisée pour assurer un équilibre sur la représentation de chacune des classes.

Dans l'état de l'art et les challenges, c'est la première solution qui est choisie. Cela s'explique par le besoin de données. En effet comme la classe la moins représentée (la *tristesse*) représente 9.96% des échantillons, la seconde solution implique de rejeter au final 50.21% des échantillons notre base de données. Or les bases de données disponibles sont trop petites pour que n'en n'utiliser qu'une partie n'ait pas d'impact. Cet effet est encore plus critique avec les méthodes utilisant l'apprentissage profond qui requiert un apprentissage sur une grande quantité de données. Le manque de données peut cependant être atténué par des techniques comme le data augmentation ou le transfert learning. Une évaluation en LOSO limite également l'impact de la séparation entre les ensembles d'apprentissage et de test.

La quantité requise est tout juste fournie par les bases de données dans leur intégralité. Cependant nous ne devons pas mettre de côté ce déséquilibre. Parlons en premier lieu de l'ensemble de test. Les différents critères d'évaluation se basent sur le taux de reconnaissance. Sur un ensemble de test non équilibré, un bon résultat peut venir uniquement de la bonne reconnaissance de la classe la plus représentée. L'effet est encore plus trompeur si, comme dans notre cas, le problème est complexe et donne des performances relativement faibles. Avec un taux de reconnaissance supé-

rieur à 90%, la robustesse est démontrée même sur un ensemble de test non équilibré. Mais pour l'analyse des ME, le taux de reconnaissance est compris entre 57% et 61%.

Supposons une méthode où toutes les ME sont associées à la classe *colère*. Cette méthode est évidemment une mauvaise méthode de classification. Malgré tout, du fait du déséquilibre de la base de données, elle donne un taux de reconnaissance de 47.30% relativement proche de celui obtenu par les méthodes de base. Lors des comparaisons entre méthodes, une petite amélioration peut correspondre plus à une spécialisation vis à vis de la classe majoritaire qu'à une meilleure capacité de classification. Prenons par exemple la matrice de confusion de la Figure 4.8, la classe *colère* est reconnue à 91% et la classe *dégoût* à 0%. Mais comme la colère est bien plus représentée que le dégoût, le score d'accuracy reste correcte.

Parlons maintenant du déséquilibre au niveau de l'ensemble d'apprentissage. Son effet est plus difficile à juger. Le modèle généré par l'apprentissage s'adapte aux données qui lui ont été fournies. Ainsi un ensemble d'apprentissage où une classe est très majoritaire va produire un modèle qui va favoriser cette classe. Cependant l'impact négatif de ce déséquilibre est plus discutable. Tout d'abord l'effet présenté plus haut ne sera pas présent si la classe majoritaire est facilement séparable. Ensuite une classe est généralement majoritaire dans une base de données car elle apparaît plus souvent dans le monde réel. Il est alors en pratique pertinent que cette classe soit favorisée dans la phase de classification.

Dans la suite de ce chapitre, nous évaluons quantitativement l'effet du déséquilibre au niveau de l'ensemble de test (section 6.1) et de l'ensemble d'apprentissage (section 6.2). Nous reprenons les résultats obtenus sur les méthodes présentées précédemment mais sur une base de données équilibrée. Les différentes méthodes prises en considération sont :

— La corrélation avec des gabarits de flot optique vu dans la section 3.3 est une méthode basée sur les caractéristiques très simple et très légère.
— L'association LBP_TOP / SVM avec unification temporelle développée dans la section 2.3.1 représente une référence en tant que classification basée caractéristiques à laquelle nous apportons une amélioration : une unification qui rend la comparaison plus cohérente.
— Le réseau ResNet18 nous a servi de base de comparaison pour notre réseau en section 4.3.2. Il est plus lourd et complexe que notre approche mais obtient une haute précision.
— Le réseau CNN optimisé introduit dans la section 4.3.3 est un réseau conçu avec une architecture spécifiquement adaptée à la classification de flot optique de ME : il ne considère que le mouvement vertical pour une vitesse de traitement optimale.

6.1 Effet d'un déséquilibre au niveau de l'ensemble de test

Dans cette partie seule le mode d'évaluation varie : l'apprentissage et donc le modèle de classification sont inchangés. L'objectif est d'introduire une évaluation où chaque classe a le même impact afin de la comparer à la précédente évaluation basée sur un ensemble de test où certaines classes sont surreprésentées.

Dans cette optique, nous introduisons une métrique que nous nommons l'Accuracy Moyenne Par Classe (AMPC) et qui qui se calcule comme suit :

$$\text{AMPC} = \frac{1}{N_c} \sum_{c=1}^{N} \frac{VC(c)}{VC(c) + FC(c)} \tag{6.1}$$

où $VC(c)$ (respectivement $FC(c)$) est le nombre d'échantillons correctement (respectivement faussement) classifiés dans la classe c et N_c représente le nombre total de classes.

Cette métrique est très proche de l'accuracy mais plus juste puisque qu'elle réalise la moyenne du taux de reconnaissance par classe. Remarquons qu'il s'agit également de la moyenne des valeurs se trouvant sur la diagonale des bonnes classifications sur la matrice de confusion. Le Tableau 6.2 permet de comparer l'accuracy et l'AMPC sur les différentes méthodes étudiées.

Méthode	Accuracy	AMPC
Gabarits	42.74%	32.76%
LBP_TOP	42.32%	39.00%
Resnet18	57.26%	40.64%
Réseau optimisé	60.17%	47.48%

TABLE 6.2 – Comparaison des résultats obtenus sur plusieurs méthodes de classification à partir de deux critères d'évaluation : l'accuracy et le AMPC.

Tout d'abord le score d'AMPC est à chaque fois inférieur à l'accuracy. La spécialisation vis à vis des classes les plus représentées est donc bien sensible. Concernant les méthodes classiques, les LBP_TOP ont un score d'AMPC proche de l'accuracy contrairement à la méthode des gabarits. Cela s'explique par le fait que le LBP_TOP donne des résultats proches pour chaque classe (entre 25% et 50%).

La différence entre les deux métriques est bien plus prononcée sur les méthodes utilisant l'apprentissage. En effet, l'architecture profonde du CNN ainsi que l'utilisation d'une descente de gradient rendent le processus plus dépendant de la répartition de la base d'apprentissage que le SVM. Par exemple, la matrice de confusion avec Resnet18 (Figure 6.3 à gauche) montre que seule la classe *colère* obtient un score élevé. Les deux CNNs ont tendance à sur-représenter une classe en particulier aux

dépens des autres.

Notre réseau optimisé est un peu moins impacté (une baisse de 13% contre 17% pour ResNet18). En effet le taux de reconnaissance sur chaque classe est toujours supérieur à 32% contre seulement 15% avec ResNet18. Cela est très probablement dû à la plus grande profondeur et aussi au plus grand nombre de paramètres internes de Resnet18. Finalement il s'agit de la méthode qui donne assez largement le meilleur AMPC.

L'impact de la répartition de l'ensemble de test est manifeste. L'AMPC met en exergue une plus grande robustesse du LBP_TOP et, dans une moindre mesure, de notre réseau optimisé. Cependant il faut prendre en compte que ces comportements sont directement provoqués par la répartition de la base d'apprentissage.

6.2 Effet d'un déséquilibre au niveau de l'ensemble d'apprentissage

Contrairement à l'ensemble de test, l'ensemble d'apprentissage influe sur les modèles de classification et donc la méthode elle-même. Nous testons alors nos méthodes entraînées cette fois sur un ensemble d'apprentissage équilibré. L'objectif est d'estimer l'effet de ce déséquilibre sur la capacité même de classification de la méthode.

Ce procédé est surtout pertinent sur les méthodes basées sur les caractéristiques. En effet, se séparer d'une part importante de la base d'apprentissage, c'est réduire considérablement sa taille. Or le classifieur SVM nécessite un ensemble d'apprentissage moins grand que les méthodes utilisant l'apprentissage profond.

Protocole :

L'approche du LOSO est conservée pour réduire l'effet de la taille de la base de données. Cependant une partie des données disponibles doit être défaussée dans cette partie. Nous sélectionnons pour chaque classe un nombre d'échantillons égal à celui de la classe la moins représentée (soit 24 échantillons). Pour chaque cycle de la cross-validation, nous excluons les données d'un sujet lors de l'entraînement pour les utiliser pour l'ensemble de test. Une fois le parcours fait pour tous les sujets, les résultats de classification sont utilisés pour calculer les matrices de confusion sur un ensemble de 24 échantillons par classe.

Résultats :

Les Figures 6.1, 6.2, 6.3 et 6.4 montrent les matrices de confusions, pour les quatre méthodes de classification évaluées, obtenues avec la base de données

complète puis avec la sélection proposée à l'instant. Notons que comme l'ensemble de test est équilibré, le score d'accuracy pour le second cas est équivalent au AMPC.

Il est intéressant d'observer que même avec un apprentissage équilibré, la classe *colère* est généralement parmi les mieux classées. Nous pouvons en conclure que la classe la plus courante est aussi celle dont le mouvement apparent est le plus descriptif. Pour la méthode des gabarits, nous observons, par rapport à l'utilisation de la base de données complète, une baisse de 3.57% vis à vis de l'accuracy mais une hausse de 8.4% vis à vis du AMPC. Notons aussi que ce rééquilibrage est nettement en faveur des classes *joie* et *surprise*.

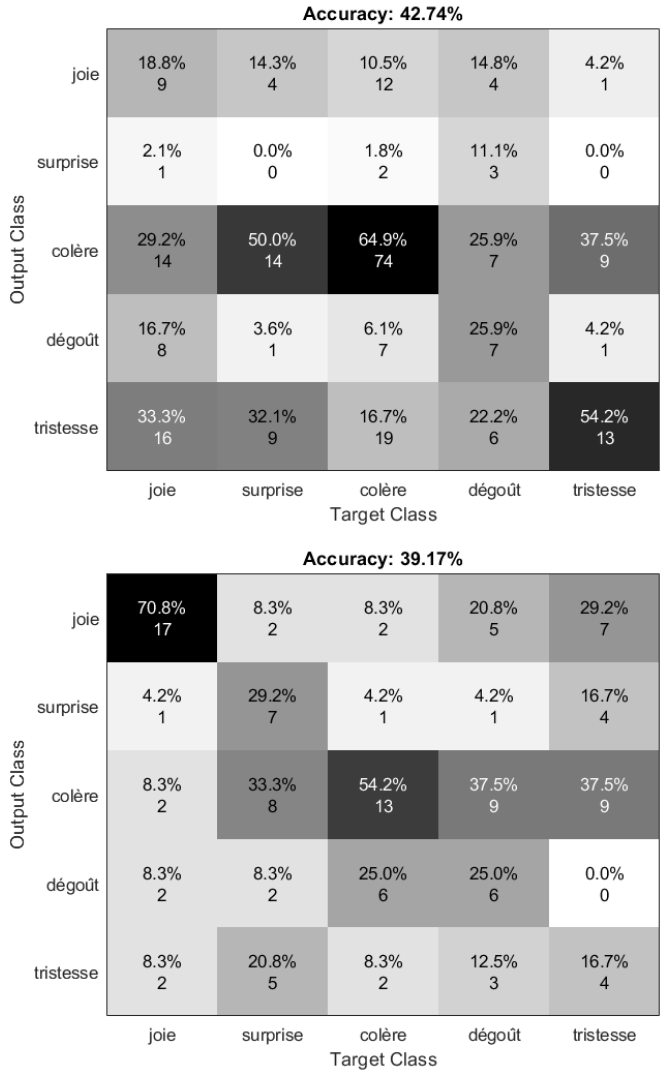

FIGURE 6.1 – Matrices de confusion obtenues à partir de la méthode des gabarits avec la base de données complète (en haut) et avec le nouveau protocole équilibré (en bas).

La méthode des LBP_TOP est celle qui produit les effets les plus minimes. Nous observons, par rapport à l'utilisation de la base de données complète, une hausse de 1.01% vis à vis de l'accuracy et de 4.33% vis à vis du AMPC. Cela s'explique

une nouvelle fois par le fait que les résultats sont équilibrés même avec la base de données complète.

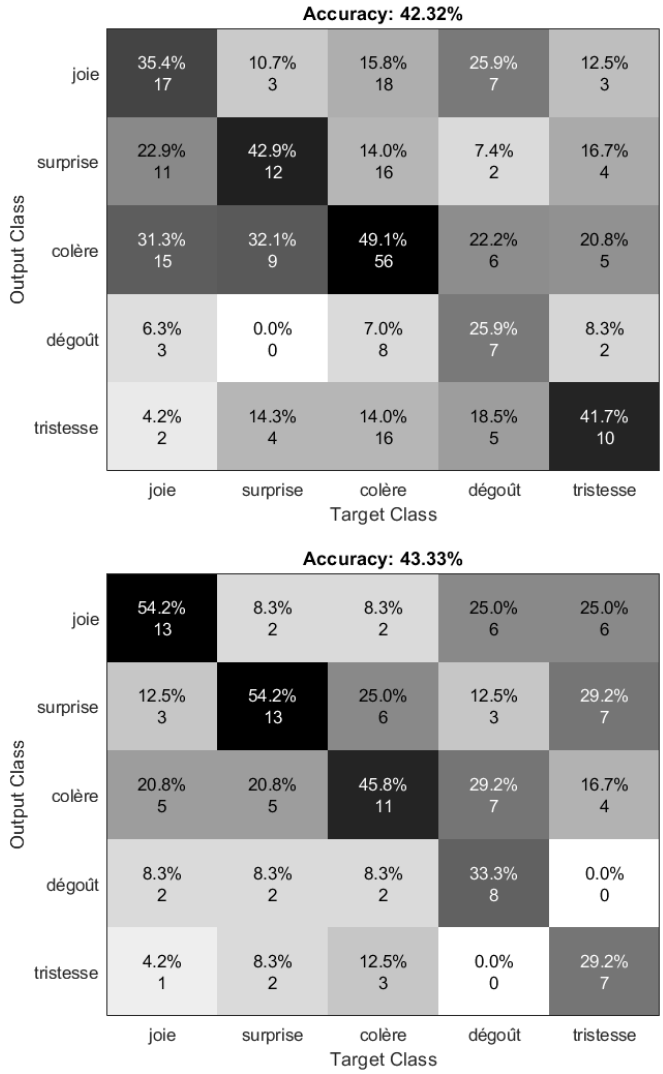

FIGURE 6.2 – Matrices de confusion obtenues à partir de la méthode du LBP_TOP avec la base de données complète (en haut) et avec le nouveau protocole équilibré (en bas).

Le réseau CNN ResNet18 donne une accuracy bien plus faible avec la nouvelle répartition. Resnet18 étant un réseau de neurones extrêmement profond, il a tendance à bien plus facilement être victime d'overfitting ou à s'adapter de façon accrue aux données d'entraînement. La réduction des données d'entraînement a un réel impact, mais est certainement réduit par l'utilisation du transfert learning (Resnet18 est entraîné au préalable sur ImageNet).

Notons aussi que le score d'accuracy obtenu est supérieur de plus de 4% par rapport à l'AMPC sur la base complète. En effet nous observons sur la nouvelle matrice de confusion que les résultats sont bien plus équilibrés. La classe la moins bien classée obtient un taux de reconnaissance de 25% contre moins de 15% avec

la base de données complète. Cela confirme que le réseau s'adapte fortement à la répartition des classes dans l'ensemble d'apprentissage.

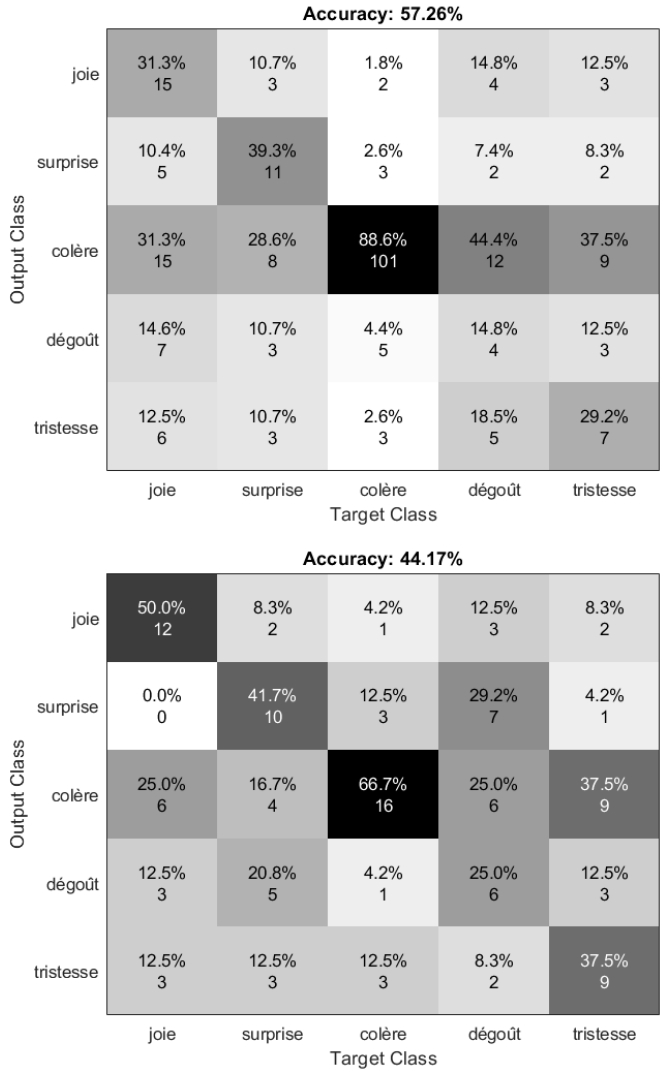

FIGURE 6.3 – Matrices de confusion obtenues à partir de ResNet18 avec la base de données complète (en haut) et avec le nouveau protocole équilibré (en bas).

Notre réseau optimisé est la seule méthode donnant un score d'accuracy plus faible que l'AMPC avec la base complète. Cela peut s'expliquer par deux facteurs. Comme avec ResNet18, le réseau CNN souffre de la réduction de la quantité de données d'apprentissage. Mais, contrairement à ResNet18, les résultats sont relativement équilibrés même avec la base de données complète. Ainsi sur notre réseau optimisé, l'utilisation de la base complète est à favoriser puisque le réseau est fortement dépendant de la quantité de données d'apprentissage tout en étant relativement robuste au déséquilibre de l'ensemble d'apprentissage.

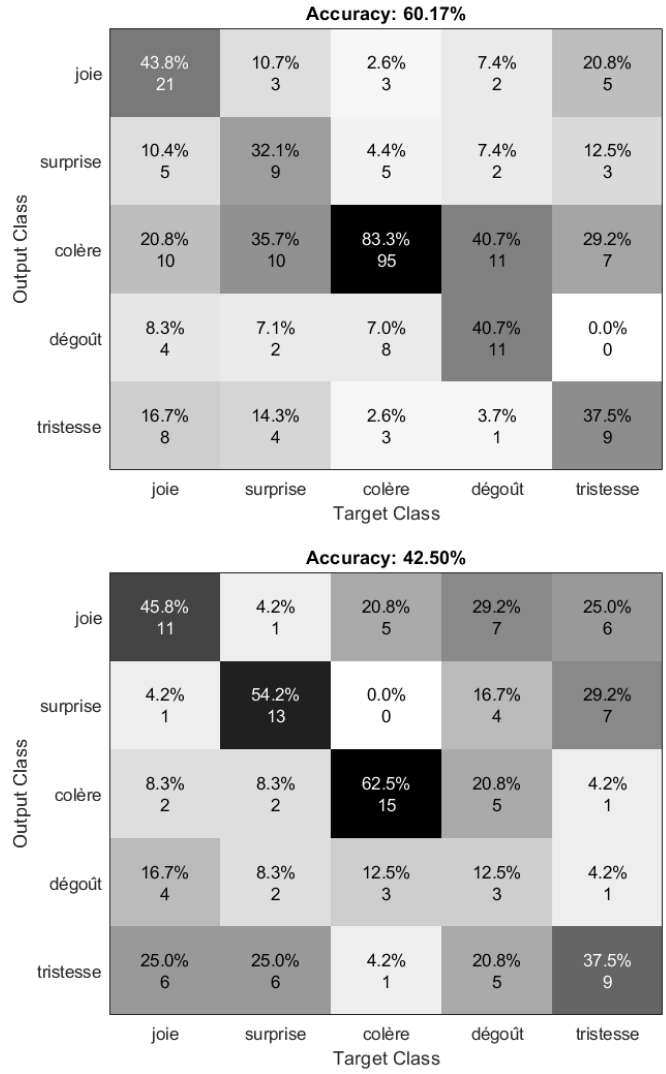

FIGURE 6.4 – Matrices de confusion obtenues à partir de notre réseau optimisé avec la base de données complète (en haut) et avec le nouveau protocole équilibré (en bas).

6.3 Conclusion

Dans ce chapitre, nous avons exploré l'effet de la répartition des classes dans les bases de données. Un équilibre au niveau de l'ensemble de test a un effet significatif. Ainsi le critère d'accuracy ne prend pas vraiment en compte la capacité à bien catégoriser chaque classe. Cela ajoute une nouvelle vision aux comparaisons entre méthodes. Par exemple les performances de ResNet18 sont en grande partie expliquées par sa mise en valeur de la classe la plus représentée. Cependant le modèle s'adapte aux données d'apprentissage, ce qui modère cette remise en question.

Nous avons ensuite observé qu'un apprentissage équilibré génère une classification plus équilibrée. La méthode des LBP_TOP utilise un classifieur SVM moins sensible à la répartition sur l'ensemble d'apprentissage que les méthodes d'apprentis-

sage profond. Les effets de cette mise à l'équilibre sont ainsi réduits sur cette méthode.

Il faut aussi prendre en considération le fait que notre étude est limitée par la quantité de données. La mise en équilibre des ensembles de test et d'apprentissage se fait au dépend de la quantité de données disponibles. Les méthodes utilisant l'apprentissage profond sont celles qui en subissent le plus sensiblement l'impact.

Une plus grande représentation d'une classe peut être intéressante pour intégrer une probabilité d'occurrence dans le cas pratique. Cependant, l'analyse des résultats obtenus doit prendre en compte ce déséquilibre. Ces particularités des bases de données sont peu traitées dans la littérature. Une solution serait de créer de nouvelles bases de données ou de synthétiser des données utilisables pour la data augmentation afin de rééquilibrer les données disponibles sans en réduire le nombre. Cependant il s'agit d'une étape compliquée aux vu de la difficulté à générer des ME.

Chapitre 7

Conclusion

Les ME représentent un indice significatif des émotions des êtres humains. Depuis les quelques décennies que les psychologues ont découvert leur existence et leur importance, de multiple études visent à comprendre leur apparitions, et à trouver comment les déceler et les reconnaître. Dans cet optique, la vision par ordinateur apporte une solution efficace car non invasive. La détection et la classification des ME à partir d'un ordinateur et d'une caméra rapide est alors un sujet très étudié.

Au cours de cette thèse, nous avons dans un premier temps suivi l'évolution historique des ME. Pour cela nous avons d'abord analysé les descripteurs de classification de ME ; anticipant la caractérisation des ME pour en extraire les éléments les plus descriptifs. Puis nous avons changé de types de méthodologie en utilisant les techniques d'apprentissage profond. Nous avons ainsi déterminé la supériorité de ces dernières en terme d'efficacité et cela en dépit des tailles limités des données d'apprentissage. Cela étant, nous avons pris du recul sur nos résultat en considérant une orientation plus dirigée par l'applicatif. Nous avons ainsi étudié le procédé dans son ensemble et non plus réduit à la seule étape de classification. Fort de la prise de conscience du manque de données nous avons expérimentalement évalué son influence dans nos résultats.

Face aux manipulations effectuées, nous pouvons certifier des nombreuses difficultés inhérentes à l'étude des ME. Tout d'abord, le mouvement généré est très rapide et subtile, ce qui fait qu'il est très facilement recouvert par d'autres mouvement du visage. Sa durée est assez variable et donc la comparaison temporelle n'est pas évidente. Elle est même parfois si courte que nous ne pouvons la capturer que sur quelques frames et potentiellement louper l'apex. Ensuite, bien que culturellement invariante, la ME s'exprime sur de nombreuses parties du visage avec une variabilité relativement importante. Enfin, à la fois difficile à générer et impossible à simuler, la capture de ME est complexe ce qui entraîne une quantité de données pour l'apprentissage très limitée. Devant de si nombreuses difficultés, plutôt que de chercher à optimiser l'efficacité nous avons préféré nous concentrer sur une étude de faisabilité en proposant de nouveaux procédés tout en restant toujours dirigé par la vérification d'une application pratique pertinente.

Les principales contributions de cette thèses sont :

— l'introduction de méthodes classiques de classification de ME dont les performances ont été comparées aux méthodes semblables de l'état de l'art.
— l'étude de l'effet des restriction d'un réseau d'apprentissage profond courant au niveau de sa complexité (nombre de couches) mais aussi sur la dimensionnalité des donnés d'entrée pour obtenir un réseau bien plus léger tout en maintenant des performances semblables.
— l'analyse de l'association spotting/classification de ME pour une évaluation plus globale et plus appliquée.
— l'introduction du paradigme du pseudo-spotting ouvrant un problème plus simple mais malgré tout pertinent vis à vis de l'application.
— l'étude de faisabilité quand à l'embarquilité du procédé.
— l'étude de l'influence des caractéristiques des données d'apprentissage sur les résultats obtenus.

7.1 Avancées liées à la thèse

Suite à une observation directe vis à vis du mouvement moyen de chaque expression, nous avons proposé la méthode des gabarits en section 3.3. Cette méthode est très compacte et simple d'exécution. Cependant les résultats sont mitigés d'un point de vue efficacité. En ce domaine, le descripteur LBP_TOP est une référence selon l'état de l'art et comme nous l'avons vérifié. Nous avons proposé une méthode d'unification temporelle afin d'uniformiser la quantité de données traitées (la durée de la ME) avant la classification. Nous avons alors observé une amélioration des résultats, en particulier sur une concentration de l'échantillonnage autour de l'apex, tout en réduisant la quantité de calcul nécessaire.

Nous avons ensuite créé des architectures de réseaux de neurones légères et à exécution rapide en vue de classifier des ME en temps réel à partir du réseau RestNet18 reconnu très efficace. Nous avons dans un premier temps démontré que la profondeur du réseau était trop importante par rapport au besoin de classification des ME. Nous avons alors estimé à quel point il pouvait être réduit. Nous avons également découvert que, sur le flot optique envoyé en entré du réseau, la composante verticale était la plus descriptive. Les résultats obtenus sur cette seule dimension sont les plus convainquants tout en réduisant significativement le modèle.

Ayant observer des résultats très faibles sur les méthodes de l'état de l'art du spotting de ME, nous avons proposé le paradigme du pseudo-spotting où l'onset et connu et où il reste seul l'apex à déterminer. Tout en restant pertinent d'un point de vue pratique, le pseudo-spotting simplifie notablement le problème. Une méthode basée sur la durée statistique d'une ME a donné de meilleurs résultats que les méthodes basées sur les descripteurs classiques nous avons testé. Une évaluation

à partir de l'association entre méthode de spotting et de classification au sein d'un système complet a démontré l'efficacité pratique de la méthode.

Nous avons étudié au long de cette thèse la possibilité de réduire la quantité de calculs nécessaires pour classifier des ME en faisant le minimum de concessions possible sur l'efficacité du système. Nous avons proposé des méthodes originales en mettant l'accent sur la rapidité d'exécution et les faibles besoins en mémoire. En continuité de cela nous avons déployé notre chaîne de traitement complète sur l'un des systèmes embarqués les plus accessibles. Nous avons ainsi démontré la possibilité d'exécution en condition réel de ce système.

7.2 Limitations

Nous avons expérimentalement estimé l'effet des caractéristiques des bases de données utilisés, en particulier leur inégale répartition en fonction des types de ME, sur les résultats obtenus. L'impact est certain. Si les comparaisons entre méthodes doivent être pris avec précautions, le type d'apprentissage influençant le modèle de classification, cette étude démontre le besoin d'agrandir la quantité de données disponibles et en particulier sur certaines classes sous représentées.

Nous avons aussi observé l'influence de l'imprécision du spotting sur la classification (Figure 5.5). Bien qu'une amélioration des performances du spotting apparaît progressivement, elles sont encore insuffisantes pour ne pas dégrader significativement l'étape de classification. Si le pseudo-spotting reste une solution crédible, il nécessite une manifestation de l'utilisateur.

Une ME est une expression instantanée du ressenti de la personne. L'état émotionnelle au contraire est une manifestation sur la durée. Nous avons (Annexe A) analysé l'utilisation de la reconnaissance des ME pour la reconnaissance des état émotionnelles de façon isolée tout d'abord puis en association à une méthode spécialisée. Nous avons observé que la ME est bien dissociée de l'état émotionnelle. Par exemple une réaction ponctuelle de colère face à un stimuli peut apparaître lorsqu'une personne est joyeuse. Il faut donc bien séparer ces deux notions même si chacune est liée aux émotions.

La totalité des bases de données disponibles sont acquises en conditions contrôlées, avec des sujets positionnés parfaitement en face de la caméra. Les méthodes apprises à partir de ces données ne sont donc efficaces que dans des conditions similaires. Des effets provoqués par un éclairage non uniforme ou un arrière-plan chargé ou variant au cours du temps auront également un effet sur les résultats obtenus.

7.3 Perspectives

Au vue des résultats de nos expériences, nous pouvons conclure avec certitude qu'il est tout à fait possible de concevoir un système interactif fonctionnel et utilisable en condition réel. La précision du système est proche de celle qu'un être humain disposant d'une bonne acuité visuelle et ayant suivi un entraînement pour la reconnaissance des ME peut atteindre.

Les pistes d'améliorations sont nombreuses. Les premières sont évidemment liées aux limitations que nous venons de voir. La plus évidente concerne les bases de données. Les méthodes utilisant l'apprentissage profond obtiennent les plus hautes performances mais elles souffrent du peu de données disponibles et il semble évident que des données d'apprentissage plus nombreux garantirait des performances plus élevées. L'augmentation artificielle des données par data augmentation est une piste sérieuse mais la mise à disposition de plus de données reste la solution la plus judicieuse. Vu la difficulté à générer des données, leur création doit se faire en fonction des besoins. Nous avons vu que le déséquilibre entre classe a un impact significatif. Il est donc important de produire des échantillons au niveau des types de ME les moins représentées : c'est à dire la tristesse puis le dégoût. Pour réaliser le spotting, notons qu'il manque de séquences longues suffisamment annotées. Il est intéressant de noter que la Chinese Academy of Science a mis en place une plate-forme pour l'annotation des micro-expressions pour aider la communauté dans la construction de bases de données de ME.

La synthétisation des ME semble aussi être une piste intéressante. Les GAN permettent de générer des images et des vidéos. Cependant ils nécessitent également un apprentissage sur de grandes quantités de données. Il ne s'agit bien évidemment pas d'une solution immédiate mais l'augmentation des données en dupliquant les séquences selon un processus similaire est une piste à envisager.

La ME est directement liée au mouvement. Il serait alors des plus logiques d'utiliser des CNN adaptés à la reconnaissance de vidéo à l'aide de neurones récurrents ou de LSTM. Le frein vient évidemment du manque de données puisque ce type de réseau en nécessite encore plus. Cette solution a cependant déjà été envisagée [128] et deviendra bien plus crédible en cas d'agrandissement de la quantité de données disponible.

Le déploiement d'un système interactif d'aide à la reconnaissance de ME en condition réel est tout à fait possible, nous l'avons démontré. Toutefois, l'interaction entre l'utilisateur et le système n'est pour l'instant pas pris en considération. Pour juger de l'efficacité du système et affiner ses différentes composantes pour une utilisation ergonomique et intuitive, une étude basée sur la psychologie cognitive pourrait être lancée et confirmer par des expériences sur des personnes en conditions réelles.

Mentionnons pour finir les différents pre-traitements pouvant améliorer les performances de nos algorithmes. Une meilleure détection de visage augmenterait la robustesse au mouvement de la tête. Une suppression des mouvement non liés aux ME rendraient le processus applicable dans des conditions moins contraintes. L'estimation du flot optique par apprentissage profond [129] pourrait donner une entrée plus propre est significative à moindre coût.

Bibliographie

[1] R. Arya, J. Singh, and A. Kumar, “A survey of multidisciplinary domains contributing to affective computing,” *Computer Science Review*, vol. 40, p. 100399, 2021.

[2] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception,” *Psychiatry*, vol. 32, pp. 88–106, 1969.

[3] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion. journal of personality and social psychology,” p. 124–129, 1971.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, 2010.

[5] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, “Samm : A spontaneous micro-facial movement dataset,” *Transaction on Affective Computing*, vol. 9, pp. 116–129, 2018.

[6] E. Haggard and K. Isaacs, “Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy,” *Methods of Research in Psychotherapy*, pp. 154–165, 1966.

[7] K. Scherer, “What are emotions ? and how can they be measured,” *Social Science Information*, vol. 44, p. 695–729, 2005.

[8] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition : History, trends, and affect-related applications,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 38, p. 1548–1568, 2016.

[9] P. Ekman, “Emotions revealed : Recognizing faces and feelings to improve communication and emotional life.,” *Holt Paperback*, vol. 128, p. 140–140, 2003.

[10] B. Bhushan, “Study of facial micro-expressions in psychology,” *Understanding Facial Expressions in Communication*, p. 265–286, 2015.

[11] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, “How fast are the leaked facial expressions : The duration of micro-expressions,” *Journal of Nonverbal Behavior*, vol. 37, p. 217–230, 2013.

[12] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge, UK : Cambridge Univ. Press, 1982.

[13] P. Ekman, *Telling Lies : Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton & Company, 2009.

[14] S. Porter and L. T. Brinke, "Reading between the lies : Identifying concealed and falsified emotions in universal facial expressions," *Psychological Science*, vol. 19, p. 508–514, 2008.

[15] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel : training laypeople and professionals to recognize fleeting emotions," *The Annual Meeting of International Communication Association*, 2009.

[16] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," *IEEE International Conference on In Pattern Recognition*, p. 1722–1727, 2014.

[17] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, and M. H. Yap, "Objective micro-facial movement detection using facs-based regions and baseline evaluation," *IEEE International Conference on Automatic Face & Gesture Recognition*, p. 642–649, 2018.

[18] Y. Han, B. Li, Y.-K. Lai, and Y.-J. Liu, "Cfd : A collaborative feature difference method for spontaneous micro-expression spotting," *IEEE International Conference on Image Processing*, 2018.

[19] S.-T. Liong, J. See, K. Wong, A. C. L. Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," *IEEE Asian Conference on Pattern Recognition*, p. 665–669, 2015.

[20] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen., "Towards reading hidden emotions : A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing,*, 2017.

[21] H. Lu, K. Kpalma, and J. Ronsin, "Micro-expression detection using integral projections," *Journal of WSCG*, vol. 25, no. 2, pp. 87–96, 2017.

[22] Y. Li, X. Huang, , and G. Zhao, "Can micro-expression be recognized based on single apex frame ?," *International Conference on Image Processing*, pp. 3094–3098, 2018.

[23] S.-J. Wang, S. Wu, and X. Fu, "A main directional maximal difference analysis for spotting micro-expressions. in, page 449–461. springer„" *Asian Conference on Computer Vision*, 2016.

[24] H. Ma, G. An, S. Wu, and F. Yang, "A region histogram of oriented optical flow (rhoof) feature for apex frame spotting in micro-expression," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, p. 281–286, 2017.

[25] C. Duque, O. Alata, R. Emonet, A.-C. Legrand, and H. Konik, "Micro-expression spotting using the riesz pyramid," *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[26] D. Patel, G. Zhao, and M. Pietikäinen, “Spatiotemporal integration of optical flow vectors for micro-expression detection,” *In International Conference on Advanced Concepts for Intelligent Vision Systems*, 215.

[27] Y. Zhao and J. Xu, “A convolutional neural network for compound micro-expression recognition,” *Sensors*, vol. 19, no. 24, 2019.

[28] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, “Spontaneous micro-expression spotting via geometric deformation modeling,” *Computer Vision and Image Understanding*, 2016.

[29] D. Borza, R. Danescu, R. Itu, and A. Darabant, “High-speed video system for micro-expression detection and recognition,” *Sensors*, vol. 17, p. 2913, 2017.

[30] P. Husák, J. Cech, and J. Matas, “Spotting facial micro-expressions in the wild,” *Computer Vision Winter Workshop*, 2017.

[31] T.-K. Tran, X. Hong, and G. Zhao, “Sliding window based micro-expression spotting : A benchmark,” *International Conference on Advanced Concepts for Intelligent Vision Systems*, p. 542–553, 2017.

[32] J. S. Sze-Teng Liong, R. C.-W. Phan, Y.-H. Oh, A. C. L. Ngo, K. Wong, and S.-W. Tan, “Spontaneous subtle expression detection and recognition based on facial strain,” *Signal Processing : Image Communication*, vol. 47, pp. 170–182, 2016.

[33] J. Li, C. Soladie, and R. Seguier, “Ltp-ml : Micro-expression detection by recognition of local temporal pattern of facial movements,” *IEEE International Conference on Automatic Face & Gesture Recognition*, p. 634–641, 2018.

[34] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, “Sme-convnet : A convolutional neural network for spotting spontaneous facial micro-expression from long videos,” *IEEE Access*, 2018.

[35] H. Pan, “Local bilinear convolutional neural network for spotting macro- and micro-expression intervals in long video sequences,” *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020.

[36] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, “Micro-expression spotting : A new benchmark,” *Neurocomputing*, vol. 443, pp. 356–368, 2021.

[37] S. Polikovsky, Y. Kameda, , and Y. Ohta., “Facial micro-expressions recognition usinghigh speed camera and 3d-gradient descriptor,” *International Conferenceon Crime Detection and Prevention*, pp. 1–6, 2009.

[38] S. Polikovsky and Y. Kameda, “Facial micro-expression detection in hi-speed video based on facial action coding system (facs),” *IEICE transactions on information and systems*, vol. 96, no. 1, p. 81–92, 2013.

[39] M. Chen, H. T. Ma, J. Li, and H. Wang, “Emotion recognition using fixed length micro-expressions sequence and weighting method,” *IEEE International Conference on Real-time Computing and Robotics*, p. 427–430., 2016.

[40] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, “Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for

the recognition of human actions," *Conference on Computer Vision and Pattern Recognition*, pp. 1932–1939, 2009.

[41] Y. J. Liu, J. K. Zhang, W. J. Yan, S. J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *Transaction on Affective Computing*, vol. 7, pp. 299–310, 2015.

[42] S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Transactions on Affective Computing*, 2017.

[43] S. Happy and A. Routray, "Recognizing subtle micro-facial expressions using fuzzy histogram of optical flow orientations and feature selection methods," *Springer, Computational Intelligence for Pattern Recognition*, p. 341–368, 2018.

[44] G. Zhao and M.Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2007.

[45] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro- expressions," *ICCV*, pp. 1449–1456, 2011.

[46] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," *IEEE International Conference on Computer Vision Workshops*, pp. 868–875, 2011.

[47] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, p. 1657–1663, 2010.

[48] Y. Guo, Y. Tian, X. Gao, and X. Zhang, "Micro-expression recogni-tion based on local binary patterns from three orthogonal planes and nearest neighbor method," *International Joint Conference on Neural Networks*, 2014.

[49] S. Zhang, B. Feng, Z. Chen, and X. Huang, "Micro-expression recognition by aggregating local spatio-temporal patterns," *International Conference on Multimedia Modeling*, p. 638–648, 2017.

[50] X. Duan, Q. Dai, X. Wang, Y. Wang, and Z. Hua, "Recognizing spontaneous micro-expression from eye region , vol. 217, pp. „" *Neurocomputing*, vol. 217, p. 27–36, 2016.

[51] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," *IEEE International Conference on Computer Vision Workshops*, 2015.

[52] B. B. Talukder, B. Chowdhury, T. Howlader, and S. M. Rahman, "Intelligent recognition of spontaneous expression using motion magnification of spatio-temporal data," *Pacific-Asia Workshop on Intelligence and Security Informatics*, 2016.

[53] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools and Applications*, p. 1–26, 2016.

[54] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning a target sample re-generator for cross-database micro-expression recognition," *ACM on Multimedia Conference*, p. 872–880, 2017.

[55] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," *ACM International Conference on Multimedia*, p. 2237–2245, 2020.

[56] M. Peng, C. Wang, T. Bi, Y. Shi, X. Zhou, and T. Chen, "A novel apex-time network for cross-dataset micro-expression recognition," *IEEE International Conference on Affective Computing and Intelligent Interaction*, p. 1–6, 2019.

[57] J. Wen, W. Yang, L. Wang, W. Wei, S. Tan, and Y. Wu, "Cross-database micro expression recognition based on apex frame optical flow and multi-head self-attention," *International Symposium on Parallel Architectures, Algorithms and Programming*, 2020.

[58] Z. Lai, R. Chen, J. Jia, and Y. Qian, "Real-time micro-expression recognition based on resnet and atrous convolutions," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2020.

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1–9, 2015.

[60] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," *IEEE International Conference on Automatic Face and Gesture Recognition*, p. 1–5, 2019.

[61] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "Learnet : Dynamic imaging network for micro expression recognition," *IEEE Transactions on Image Processing*, 2019.

[62] M. Verma, S. K. Vipparthi, and G. Singh, "Hinet : Hybrid inherited feature learning network for facial expression recognition," *IEEE Letters of the Computer Society*, 2019.

[63] M. Verma, S. K. Vipparthi, and G. Singh, "Affectivenet : Affective-motion feature learning for micro expression recognition," *IEEE MultiMedia*, 2020.

[64] Y. Wang, H. Ma, X. Xing, and Z. Pan, "Eulerian motion based 3dcnn architecture for facial micro-expression recognition," *International Conference on Multimedia Modeling*, 2020.

[65] V. R. Gajjala, S. P. T. Reddy, S. Mukherjee, and S. R. Dubey, "MERANet : Facial Micro-Expression Recognition using 3D Residual Attention Network," *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–10, 2020.

[66] B. Chen, Z. Zhang, N. Liu, Y. Tan, X. Liu, and T. Chen, "Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition," *Information*, vol. 11, no. 8, p. 380, 2020.

[67] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, “Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions,” *IEEE Transaction on Multimedia*, 2019.

[68] Z. Xia, W. Peng, H. Q. Khor, X. Feng, and G. Zhao, “Revealing the invisible with model and data shrinking for composite-database micro-expression recognition,” *IEEE Transactions on Image Processing*, vol. 29, p. 8590–8605, 2020.

[69] N. V. Quang, J. Chun, and T. Tokuyama, “Capsulenet for micro-expression recognition,” *International Conference on Automatic Face and Gesture Recognition (FG)*, p. 1–7, 2019.

[70] N. Liu, X. Liu, Z. Zhang, X. Xu, and T. Chen, “Offset or onset frame : A multi-stream convolutional neural network with capsulenet module for micro-expression recognition,” *International Conference on Intelligent Informatics and Biomedical Sciences*, 2020.

[71] Y. Gan, S. Liong, W. Yau, Y. Huang, and L. Tan, “Off-apexnet on micro-expression recognition system,” *Signal Processing : Image Communication*, 2019.

[72] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, “Dual- stream shallow networks for facial micro-expression recognition,” *IEEE International Conference on Image Processing*, p. 36–40, 2019.

[73] S.-T. Liong, Y. Gan, D. Zheng, S.-M. Li, H.-X. Xu, H.-Z. Zhang, R.-K. Lyu, and K.-H. Liu, “Evaluation of the spatio-temporal features and gan for micro-expression recognition system,” *Journal of Signal Processing Systems*, p. 1–21, 2020.

[74] K. Li, Y. Zong, B. Song, J. Zhu, J. Shi, W. Zheng, and L. Zhao, “Three-stream convolutional neural network for micro-expression recognition,” *Australian Journal of Intelligent Information Processing System*, p. 41–48, 2019.

[75] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, “Recognizing spontaneous micro-expression using a three-stream convolutional neural network,” *IEEE Access*, vol. 7, p. 184 537–184 551, 2019.

[76] W. She, Z. Lv, J. Taoi, and M. Niu, “Micro-expression recognition based on multiple aggregation networks,” *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020.

[77] J. Li, Y. Wang, J. See, and W. Liu, “Micro-expression recognition based on 3d flow convolutional neural network,” *Pattern Analysis and Applications*, p. 1331–1339, 2019.

[78] L. Yao, X. Xiao, R. Cao, F. Chen, and T. Chen, “Three stream 3d cnn with se block for micro-expression recognition,” *International Conference on Computer Engineering and Application*, p. 439–443, 2020.

[79] H. Yan and L. Li, “Micro-expression recognition using enriched two stream 3d convolutional network,” *International Conference on Computer Science and Application Engineering*, 2020.

[80] S.-T. Liong, Y. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, “Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition,” *International Conference on Automatic Face and Gesture Recognition*, pp. 1–5, 2019.

[81] C. Wu and F. Guo, “Tsnn : Three-stream combining 2d and 3d convolutional neural network for micro-expression recognition,” *Transactions on Electrical and Electronic Engineering*, 2021.

[82] A. J. R. Kumar and B. Bhanu, “Micro-expression classification based on landmark relations with graph attention convolutional network,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, p. 1511–1520, June 2021.

[83] L. Lo, H. Xie, H. Shuai, and W. Cheng, “MER-GCN : micro expression recognition based on relation modeling with graph convolutional network,” *CoRR*, vol. abs/2004.08915, 2020.

[84] Y. Li, X. Huang, and G. Zhao, “Joint local and global information learning with single apex frame detection for micro-expression recognition,” *IEEE Transactions on Image Processing*, vol. 30, p. 249–263, 2020.

[85] B. Sun, S. Cao, D. Li, J. He, and L. Yu, “Dynamic micro-expression recognition using knowledge distillation,” *IEEE Transactions on Affective Computing*, 2020.

[86] B. Xia, W. Wang, S. Wang, and E. Chen, “Learning from macro-expression : a micro-expression recognition framework,” *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[87] B. Xia and S. Wang, “Micro-expression recognition enhanced by macro-expression from spatial-temporal domain,” *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, p. 1186–1193, 2021.

[88] D. Patel, X. Hong, and G. Zhao, “Selective deep features for micro-expression recognition,” *IEEE International Conference Pattern Recognition*, p. 2258–2263, 2017.

[89] Y. Liu, H. Du, L. Zheng, and T. Gedeon, “A neural micro-expression recognizer,” *IEEE International Conference on Automatic Face and Gesture Recognition*, p. 1–4, 2019.

[90] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, “Keepaugment : A simple information-preserving data augmentation approach,” in *Conference on Computer Vision and Pattern Recognition*, pp. 1055–1064, June 2021.

[91] M. Hong, J. Choi, and G. Kim, “Stylemix : Separating content and style for enhanced data augmentation,” in *Conference on Computer Vision and Pattern Recognition*, pp. 14862–14870, June 2021.

[92] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, p. 60, 2019.

[93] S. Liong, Y. S. Gan, D. Zheng, S. Lic, H. Xua, H. Zhang, R. Lyu, and K. Liu, “Evaluation of the spatio-temporal features and GAN for micro-expression recognition system,” *CoRR*, vol. abs/1904.01748, 2019.

[94] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," *ACM International Conference on Multimedia*, 2020.

[95] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," *Workshop onApplications of Computer Vision.*, pp. 1–6, 2009.

[96] M. Shreve, S. Godavarthy, D. Goldgof, , and S. Sarkar, "Macro- and micro-expressionspotting in long videos using spatio-temporal strain," *IEEE InternationalConference on Automatic Face Gesture Recognition and Workshops*, p. 51–56, 2011.

[97] S. Polikovsky, Y. Kameda, , and Y. Ohta, "Detection and measurement of facial micro-expression characteristics for psychological analysis," *Kameda's Publication110*, p. 57–64, 2010.

[98] S. Park and D. Kim, "Subtle facial expression recognition using motion magnification," *Pattern Recognition Letters*, 2009.

[99] W. J. Yan, Q. Wu, Y. J. Liu, S. J. Wang, and X. Fu, "Casme database : A dataset of spontaneous micro-expressions collected from neutralized faces," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[100] W. Yan, X. Li, S. Wang, G. Zhao, Y. Liu, Y. Chen, and X. Fu, "Casmeii : an improved spontaneous micro-expression database and the baseline evaluation," *PLOS One*, vol. 9, pp. 1–8, 2014.

[101] F. Qu and S. W. an W.J. Yan, "Cas(me)^2 : A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transaction on Affective Computing*, 17 January 2017.

[102] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2484–2498, 2018.

[103] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[104] T. Ojala, M. Pietikäinenand, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2002.

[105] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," *European Conference on Computer Vision*, 2004.

[106] A. Hadid, M. Pietikäinen, and T. Ahonen, "A discriminative feature space for detecting and recognizing faces," *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[107] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," *IEEE International Conference on Computer Vision*, 2011.

[108] B. Horn and B. Schunck, "Determining optical flow.," *Artificial Intelligence*, vol. 17, p. 185–203, 1981.

[109] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection. proc. : I :886-893," *Computer Vision and Pattern Recognition*, 2005.

[110] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.

[111] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[112] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From macro to micro expression recognition : Deep learning on small datasets using transfer learning," *International Conference on Automatic Face and Gesture Recognition*, pp. 657–661, 2018.

[113] D. Patel, X. Hong, , and G. Zhao, "Selective deep features for micro expression recognition," *International Conference on Pattern Recognition*, pp. 2258–2263, 2016.

[114] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, , and X. Fu., "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018.

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[116] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet : A lightweight convolutional neural network for optical flow estimation," *Computer Vision and Pattern Analysis*, 2018.

[117] I. Rieger, T. Hauenstein, S. Hettenkofer, and J.-U. Garbas, "Towards real-time head pose estimation : Exploring parameter-reduced residual networks on in-the-wild datasets," *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 122–134, 2019.

[118] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *Journal of Imaging*, vol. 4, 2018.

[119] M. H. Yap, J. See, X. Hong, and S.-J. Wang, "Facial micro-expressions grand challenge 2018 summary," *International Conerence on Automatic Face and Gesture Recognition*, pp. 675–678, 2018.

[120] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," *International Conference on Automatic Face Gesture Recognition*, vol. 1, pp. 667–674, 2018.

[121] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik, "Learning-based video motion magnification," *European Conference on Computer Vision*, 2018.

[122] H. Pan, L. Xie, Z. Wang, B. Liu, M. Yang, and J. Tao, "Review of micro-expression spotting and recognition in video sequences," *Virtual Reality & Intelligent Hardware*, vol. 3, pp. 1–17, 2021.

[123] J. Li, C. Soladie, and R. Seguier, “Local temporal pattern and data augmentation for micro-expression spotting,” *IEEE Transactions on Affective Computing*, 2020.

[124] J. LI, S.-J. Wang, M. H. Yap, J. See, X. Hong, and X. Li, “Megc2020 - the third facial micro-expression grand challenge,” *EEE International Conference on Automatic Face and Gesture Recognition*, 2020.

[125] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns : Application to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[126] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pp. I–I, 2001.

[127] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, “Megc 2019 – the second facial micro-expressions grand challenge,” *IEEE International Conference on Automatic Face Gesture Recognition*, 2019.

[128] M. Bai and R. Goecke, “Investigating lstm for micro-expression recognition,” *International Conference on Multimodal Interaction*, p. 7–11, 2020.

[129] T. Hui, X. Tang, and C. C. Loy, “A lightweight optical flow CNN - revisiting data fidelity and regularization,” *CoRR*, vol. abs/1903.07414, 2019.

[130] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, “The influences of emotion on learning and memory,” *Frontiers in Psychology*, 2017.

[131] G. Vecchiato, L. Astolfi, and F. D. V. Fallani, “On the use of eeg or meg brain imaging tools in neuromarketing research,” *Computational Intelligence and Neuroscience*, 2011.

[132] C. Nass, M. Jonsson, and H. Harris, “Improving automotive safety by pairing driver emotion and car voice emotion,” *Extended Abstracts on Human Factors in Computing Systems*, 2005.

[133] R. McCraty, M. Atkinson, and D. Tomasino, *Science Of The Heart - Exploring the Role of the Heart in Human Performance*. Institute of HeartMath, 2001.

[134] Y. Sun and N. Thakor, “Photoplethysmography revisited : from contact to non contact, from point to imaging,” *IEEE Transaction on Biomedical Engineering*, 2016.

[135] D. McDuff, J. Estepp, A. Piassecki, and E. Blackford, “A survey of remote optical photoplethysmographic imaging methods,” *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015.

[136] A. Sikdar, K. Behera, and D. Dogra, “Computer vision guided human pulse rate estimation : A review,” *IEEE Reviews in Biomedical Engineering*, 2016.

[137] E. Gil, M. Orini, R. Bailon, J. Vergara, L. Mainardi, and P. Laguna, “Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions,” *Physiological Measurement*, 2010.

[138] A. Schafer and J. Vagedes, “How accurate is pulse rate variability as an estimate of heart rate variability ? a review on studies comparing photoplethysmographic technology with an electrocardiogram,” *International Journal of Cardiology*, 2013.

[139] M. Nitzan, A. Babchenko, B. Khanokh, and D. Landau, “The variability of the photoplethysmographic signal - a potential method for the evaluation of the autonomic nervous system,” *Physiological Measurement*, 1998.

[140] C. Conaire, N. O’Connor, and A. F. Smeaton, “Detector adaptation by maximising agreement between independant data sources,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[141] M. Poh, D. McDuff, and R. Picard, “Advancements in noncontact, multiparameter physiological measurements using a webcam,” *IEEE Trans. on Biomedical Engineering*, 2011.

[142] G. De Haan and V. Jeanne, “Robust pulse-rate from chrominance-based rppg,” *IEEE Transaction on Biomedical Engineering*, 2013.

[143] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, “Heart rate variability - standards of measurement, physiological interpretation, and clinical use,” *European Heart Journal*, 1996.

[144] W. Von Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, and D. P. Mandic, “Resolving ambiguities in the lf/hf scatter plots for the categorization of mental and physical stress from hrv,” *Frontiers in Physiology*, 2017.

Annexe A

Reconnaissances des états émotionnels

L'étude des expressions faciales est un domaine prépondérant dans l'étude de l'informatique affective. La reconnaissance de l'état émotionnels en représente une autre facette. L'état émotionnel général des personnes et leur humeur sont beaucoup étudiés dans les domaines de la psychologie et de la médecine. Il a été prouvé que l'état émotionnel d'une personne peut avoir un impact sur son temps de réaction et sa capacité d'apprentissage à court terme, et même sur sa santé à long terme [130]. Pouvoir prédire automatiquement l'état émotionnel d'une personne offre diverses applications dans le monde réel, comme le neuromarketing [131] ou la surveillance des conducteurs automobiles [132].

Contrairement aux émotions simples, les états émotionnels sont des états complexes de l'esprit humain qui sont provoqués par l'environnement ou les pensées internes pendant une certaine période de temps. Pour reconnaître l'état émotionnel d'une personne, les scientifiques ont recherché de nombreux indices : les gestes, les intonations de la voix et, plus évidemment, les variations des expressions macro-faciales dans le temps. Plus récemment, la communauté a commencé à s'intéresser à l'exploration des micro-expressions [2]. Une autre voie de recherche en matière de reconnaissance des émotions est basée sur l'analyse de signaux physiologiques tels que la température de la peau, l'activité électrodermale ou l'électromyographie.

Nous avons pour l'instant présenté une étude sur la reconnaissance de micro-expressions faciales. Il s'agit d'évènement ponctuel révélant un ressenti immédiat de la personne. Cette temporalité représente la principale variation avec un état émotionnel. En effet un état émotionnel consiste en une expression de l'émotion ressenti sur une durée relativement longue. Un état favorisera la présence le ME du type d'émotion correspondant mais ne générera pas son occurrence qui doit être spontanée et en réaction à un évènement. Cependant, il est également possible de prédire l'état émotionnel d'une personne en se basant sur les MEs couplées aux macro-expressions faciales classiques. Il faut alors prendre en compte leur statistique d'occurrence.

Dans cette section, nous proposons une ouverture des descripteurs étudiés précédemment pour la reconnaissance d'états émotionnels. S'ils ne sont pas configurés spécialement pour ce domaine il existe des correspondances et des liens évidents entre ces deux problématiques. Nous comparons alors deux modalités : le LBP_TOP en tant que référence des descripteurs basés reconnaissance de ME et le HRV, un descripteur de signaux physiologique calibré pour la reconnaissance d'état émotionnel. Les deux fonctionnent à partir d'un flux vidéo centré sur le visage de la personne mais ont au départ été créé pour des objectifs différents.

A.1 Descripteur basé sur les signaux physiologiques

Notre état émotionnel a un impact sur les signaux physiologiques émis par notre corps. Lorsqu'une personne subit un stimulus émotionnel, de multiples réactions physiologiques sont générées, telles que des modifications des rythmes respiratoires et cardiaques. Ces réactions sont dues à l'activité du *Système Nerveux Autonome/Autonomous Nervous System* (ANS) [133], composé de deux branches principales : le *Système Nerveux Sympathique* (SNS) et le *Système nerveux parasympathique* (SNP). Le SNS et le SNP fonctionnent de manière complémentaire, afin d'assurer un équilibre entre les systèmes physiologiques et l'état émotionnel [133]. L'un des schémas physiologiques qui a retenu l'attention d'un grand nombre de psychophysiologues et de médecins est le changement des rythmes cardiaques. En effet, la fréquence des battements du cœur fluctue en permanence. La *Variabilité du rythme cardiaque dite Heart Rate Variability* (HRV) est conventionnellement définie comme le changement des intervalles de temps entre des battements successifs [133].

Les signaux HRV sont classiquement extraits d'enregistrements d'électrocardiogrammes. Au cours des dernières décennies, des méthodes sans contact pour évaluer l'activité cardiaque ont été développées. Une méthode particulière est la *Remote Photopletysmography* (RPPG) [134, 135], qui permet d'estimer la fréquence du pouls à partir d'une vidéo. Ceci est réalisé en analysant la quantité de lumière réfléchie par une surface de peau, qui dépend des variations du volume sanguin que l'activité de pompage du cœur provoque.

Le principe de base du RPPG découle de la photopléthysmographie par réflexion, où la lumière atteignant une caméra est modulée par les pulsations sanguines de la peau. Les battements rythmiques du cœur entraînent des modifications du volume sanguin pulsé qui, à leur tour, entraînent des changements infimes de la couleur de la peau qui peuvent être quantifiés à l'aide de différentes techniques de traitement du signal pour générer un signal cardiaque.

A.2 Extraction de caractéristiques

Nous explorons la faisabilité de la classification et de la prédiction de l'état émotionnel d'une personne, sur la base de deux types d'informations cachées uniquement perceptibles par les algorithmes de vision par ordinateur. Ceci est réalisé à travers l'instrumentalisation de la reconnaissance des expressions faciales et sa récente extension à la reconnaissance des micro-expressions d'une part, et l'analyse du signal physiologique mesurable à distance qu'est la *Variabilité du rythme cardiaque* (HRV) d'autre part. Les deux modalités se basent sur une acquisition à partir de simple caméras RGB.

A.2.1 Description à partir de ME

Nous utilisons le LBP_TOP, qui est le descripteur de base utilisé comme référence dans la plupart des articles étudiant les ME [107, 101] comme défini dans la section 3.1. Le LBP_TOP ne peut décrire qu'une seule ME, c'est à dire une classification sur une très courte durée de temps. L'état émotionnel va s'exprimer sur une durée plus importante et doit se classifier à partir d'une vidéo longue. Or, sur une vidéo longue, plusieurs ME peuvent apparaître et correspondre à des type d'émotion différents.

La méthode présentée et évaluée ici suit le protocole suivant. Tout d'abord les différentes MEs contenues dans la vidéo longue sont extraites. Pour chacune, le vecteur de caractéristiques relatif au LBP_TOP est calculé. L'étape de classification donne un résultat par ME qui être, et sont souvent en pratique, non concordant. Un simple vote majoritaire est alors appliqué pour obtenir le sentiment général de la personne sur une longue période. La classe gagnante sera utilisée pour représenter l'état émotionnel de la personne sur l'ensemble de la vidéo. Mais une égalité peut subvenir.

La classification globale nécessite alors quatre SVMs avec des noyaux *Radial Basis Function* (RBF) : le premier est entraîné sur les trois classes, et les trois autres sont entraînés sur chaque paire possible de deux classes à partir des trois classes initiales. Les classifieurs SVM à 2 classes sont utilisés lorsque nous avons une égalité entre 2 classes en utilisant le classificateur à 3 classes.

A.2.2 Description à partir du PRV

À partir d'une onde pulsée RPPG, le signal PRV (Pulse Rate Variability) peut être extrait en mesurant les intervalles de temps entre les impulsions [136]. Il a été démontré dans de nombreuses études que le PRV contient des informations sur le HRV [137, 138] et peut donc être un indicateur pertinent de l'activité autonome [139]. En fait, les composantes spectrales haute et basse fréquence (HF et LF) du PRV

décrivent les interactions entre le SNP et le SNS.

Trois étapes principales sont suivies pour obtenir le signal PRV comme résumé dans la Fig. A.1 : détection et suivi des visages, extraction des impulsions RPPG et estimation du PRV. Tout d'abord, l'algorithme de détection de visage de Viola-Jones est utilisé pour détecter la région d'intérêt, *i.e.* le visage dans notre cas, pour chaque image vidéo. L'emplacement de la région d'intérêt est ensuite suivi et prédit à l'aide d'un filtre de Kalman linéaire. Ensuite, la peau est détectée à l'aide de la méthode de Conaire *et al.* proposée dans [140], permettant de sélectionner des pixels qui sont spatialement moyennés. Cela permet d'obtenir un triplet RVB unique pour chaque image. Les triplets sont ensuite concaténés pour former les traces temporelles RVB.

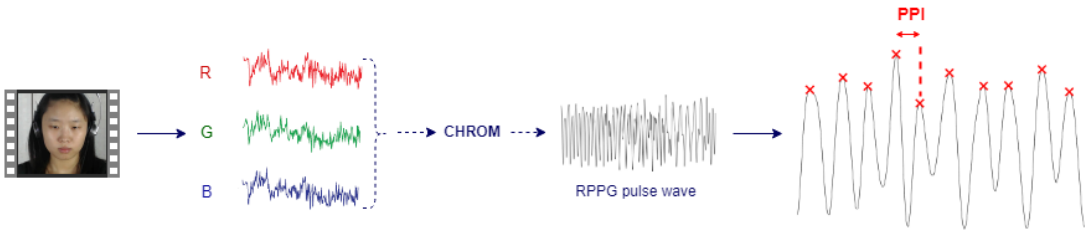

FIGURE A.1 – Cadre pour obtenir un signal PRV à l'aide de RPPG.

La deuxième étape consiste à calculer le signal d'impulsion à partir des traces RVB. De nombreuses techniques avancées et complexes ont été proposées récemment [134, 141]. Dans ce travail, nous utilisons l'algorithme de chrominance proposé par De Haan *et al.* dans [142], et désigné par CHROM dans la Fig.A.1. Le principal avantage de cette méthode réside dans sa simplicité de calcul, du fait de sa formulation analytique. Après normalisation des traces RVB (que R_n, G_n et B_n soient les traces normalisées relatives), deux signaux de chrominance orthogonaux X_s et Y_s sont construits comme suit : $X_s = 3R_n - 2G_n$ et $Y_s = 1,5R_n + G_n - 1,5B_n$.

X_s et Y_s sont ensuite filtrés en passe-bande avec un filtre de Butterworth (fréquences de coupure de 0, 7 et 3, 5 Hz) pour donner deux signaux X_f et Y_f. Le signal d'impulsion S est alors obtenu comme suit : $S = X_f - \alpha Y_f$, où $\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)}$, et $\sigma(.)$ est l'opérateur d'écart type. L'inclusion du rapport α minimise les perturbations dues au mouvement, puisqu'elles modifient de la même manière les amplitudes des signaux de chrominance X_s et Y_s alors que le signal d'impulsion cardiaque ne le fait pas.

La troisième étape consiste à interpoler et à rééchantillonner (nous avons utilisé un taux d'échantillonnage de $125Hz$) l'onde d'impulsion RPPG ; ces étapes sont nécessaires afin d'augmenter la résolution du domaine temporel et de faciliter la détection des pics. La série temporelle *Intervalle d'impulsion à impulsion* (Pulse-Pulse Interval PPI) est ensuite mesurée pour constituer le signal PRV.

Les composantes HF et LF sont obtenues après analyse du spectre de puissance de la forme d'onde PRV. La composante LF est représentée par la puissance spectrale des fréquences comprises entre $0,04$ Hz et $0,15$ Hz, tandis que la gamme HF couvre les fréquences comprises entre $0,15$ Hz et $0,4$ Hz. Les premières études sur les caractéristiques de la HRV ont proposé le rapport LF/HF comme indicateur de l'activité autonome [143]. Cependant, des recherches récentes ont montré des ambiguïtés dans l'interprétation de ce rapport [144]. Par conséquent, nous utilisons la caractéristique bidimensionnelle (HF, LF) telle que proposée dans [144]. La différence est que nous utilisons cette représentation pour la première fois pour l'analyse du signal PRV alors qu'elle a été appliquée au HRV dans [144].

A.3 Expériences

Nous allons ici comparer les performances obtenues à partir de ces deux modalités. L'objectif est différent que précédemment puisque nous cherchons à reconnaître les états émotionnels. Le protocole de test doit donc être adapté.

A.3.1 Jeu de données

L'apprentissage et les tests ont été réalisés à partir de la base de données $CAS(ME)^2$ [101]. Cette base propose 2 types d'annotations pour les M/M-FEs, la première découle des mouvements des muscles faciaux basés sur les *Action Units* (AU) suivant le Facial Action Coding System (FACS) proposé par Ekman. Les secondes annotations recensent les émotions déclarées par les candidats. Ces deux annotations ne sont pas en accord pour toutes les vidéos. De plus, dans certains cas, les états émotionnels et les annotations basées sur les AU sont contradictoires (un sujet montrerait une expression faciale négative devant une vidéo induisant la joie). 24.05% des expressions faciales sont classées comme *autres* (*i.e.* où les AU liées ne sont pas discriminantes). Certains sujets montraient également des expressions faciales contradictoires sur la même vidéo. Ces observations nous ont incités à proposer pour la première fois l'utilisation des vidéos d'excitation comme vérité de terrain. Nous faisons alors l'hypothèse que l'état émotionnel de la personne qui regarde une vidéo est équivalent à l'émotion que cette vidéo est censée induire. Cette décision est motivée par l'utilisation d'un étiquetage qui serait plus simple et moins susceptible de prêter à confusion. Au total, le nombre de vidéos disponibles est de 62, avec 14 de vidéos provoquant la joie, 24 le dégoût et 24 la colère. La durée des vidéos varit de 1 minute à environ 2 minutes et 30 secondes.

A.3.2 Protocole d'évaluation

Comme précédemment nous évaluons les performances à partir du critère d'accuracy en utilisant le LOSO. Pour les LBP_TOP, les paramètres du rayon spatio-temporel

sont équivalents à ceux utilisés dans [101]. Concernant le PRV, les composantes LF et HF ont été obtenues à partir des formes d'onde PRV extraites de chaque vidéo, et concaténées pour former des couples (HF, LF), comme présenté dans [144]. La classification des états émotionnels a ensuite été réalisée sur la base des valeurs (HF, LF) en utilisant un SVM non linéaire, avec un noyau RBF.

A.3.3 Résultats

Les résultats de la classification des états émotionnels à partir des macro et micro-expressions et de la variabilité du pouls sont présentés dans le Tableau A.1. Comme attendu l'utilisation des LBP_TOP est bien moins preformant. L'accuracy chute en effet de 59.79% à 42.74%. Sur les matrices de confusion nous pouvons remarquer que la classe `joie` n'est absolument pas reconnue. Cette émotion est moins instantanée que les autres. La joie est généralement ressenti de façon continu et non une réaction à un stimuli. Dans une phase de joie, le moindre évènement spontané peut générer un léger ressenti autre (une autre émotion comme la surprise ou la peur...) générant une ME venant bruiter notre étude. De l'autre côté le PRV génère de très hautes performances au niveau de chaque classe.

Si nous comparons les résultats de LBP_TOP sur les vidéos d'excitation (pour l'état émotionnel) et son utilisation originale pour les étiquettes basées sur l'AU, nous pouvons voir qu'avec $40,95$ % [101] sur les 4 classes basées sur l'AU et $42,74$ % sur l'étiquetage basé sur les vidéos d'excitation, les scores sont comparables.

	M/M-FEs			**PRV**		
VT \ est	*Dég*	*Col*	*Joi*	*Dég*	*Col*	*Joi*
Dég	57.4	40.4	2.1	75.0	4.0	21.0
Col	44.7	55.3	0.0	0.0	75.0	25.0
Joi	56.7	43.3	0.0	35.0	20.0	45.0
Acc.	**42.74**			**59.79**		

TABLE A.1 – Matrice de confusion (VT désigne la vérité de terrain et est l'estimation faite par la méthode) et valeur d'accuracy de la classification des états émotionnels pour les M/M-FE et les PRV. Les émotions d'excitation présentes dans $CAS(ME)^2$ sont le Dégoût (*Dég*), la Colère (*Col*) et la Joie (*Joi*). Les résultats sont exprimés en pourcentage (%).

A.4 Bilan

Bien que le cadre soit proche, nos résultats démontrent que les domaines de la reconnaissance d'expressions et d'état émotionnel sont franchement séparés dans

leurs objectifs. La temporalité différente engendrent des caractérisations différentes. Il faut alors appliquer des méthodologies spécifiques.

À partir de vidéos simples, il est possible d'extraire des caractéristiques analytiques et physiologiques, y compris les M/M-FE et le PRV. Nos résultats montrent que le PRV peut également être un bon outil pour estimer les états émotionnels avec une précision de classification d'environ 59%. Cette étude est la première, à notre connaissance, à tenter de reconnaître des états émotionnels généraux à l'aide de M/M-FEs et de PRV. Bien que les M/M-FE aient donné des résultats inférieurs, des pistes d'amélioration doivent être explorées.

Publications

Journal international

Reda Belaiche, Yu Liu, Cyrille Migniot, Dominique Ginhac, and Fan Yang, “Cost-Effective CNNs for Real-Time Micro-Expression Recognition”. Applied Sciences 10, 2020

Conférences internationales

Reda Belaiche, Cyrille Migniot, Dominique Ginhac, and Fan Yang, "Time Unification for Local Binary Pattern Three Orthogonal Planes". IEEE International Conference on Signal Image Technology Internet Based Systems (SITIS), November 2019, Sorrento, Italy

Reda Belaiche, Rita Meziati Sabour, Cyrille Migniot, Yannick Benezeth, Dominique Ginhac, Keisuke Nakamura, Randy Gomez, and Fan Yang, "Emotional State Recognition with Micro-Expressions and Pulse Rate Variability". IEEE International Conference on Image Analysis and Processing (ICIAP), September 2019, Trento, Italy

Congrès national

Reda Belaiche, Cyrille Migniot, Dominique Ginhac et Fan Yang, "Reconnaissance des émotions par l’informatique affective". Congrès Annuel de Recherche pour les Instituts Universitaires de Technologie (CNRIUT), 2018