Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine

Khadija Lekdioui

SPIM

Thèse de Doctorat

école doctorale sciences pour l'ingénieur et microtechniques
UNIVERSITÉ DE TECHNOLOGIE BELFORT-MONTBÉLIARD

Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine

Khadija LEKDIOUI

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

جامعة ابن طفيل Université Ibn Tofaïl

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat en Sciences pour l'Ingénieur

par

KHADIJA LEKDIOUI

Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine

Thèse présentée et soutenue à Kenitra-Maroc, le 29 décembre 2018

Composition du Jury :

M. HAMAD DENIS	Professeur à l'Université du Littoral Côte d'Opale	Président/Examinateur
M. TALEB-AHMED ABDELMALIK	Professeur à l'Université de Valenciennes	Rapporteur
M. OULAD HAJ THAMI RACHID	Professeur à l'Université Mohamed V, ENSIAS, Maroc	Rapporteur
M. SBIHI MOHAMED	Professeur à l'Université Mohamed V, EST, Maroc	Examinateur
M. RUICHEK YASSINE	Professeur à l'Université Bourgogne Franche-Comté	Directeur de thèse
M. MESSOUSSI ROCHDI	Professeur à l'Université Ibn Tofail, Maroc	Co-directeur de thèse

N° 2018UBFCA042

Remerciements

Cette thèse a fait l'objet d'une co-tutelle entre l'Université Ibn Tofail de Kénitra, Maroc et l'Université Bourgogne Franche-comté, France. Le travail a été effectué conjointement entre le laboratoire Systèmes des Télécommunications et Ingénierie de la Décision à Kénitra et le laboratoire d'Electronique, d'Informatique et de l'Image à Belfort.

C'est un plaisir de remercier ceux qui ont rendu cette thèse possible. J'adresse d'abord mes vifs remerciements à **M. Rachid OULAD HAJ THAMI** et **M. Abdelmalik TALEB-AHMED** pour avoir accepté d'être les rapporteurs de ces travaux. Je remercie également **M. Mohamed SBIHI** et **M. Dennis HAMAD** d'en être les examinateurs.

Je tiens à adresser mes plus sincères remerciements au **Professeur Rochdi MESSOUSSI**, mon Directeur de thèse au Maroc. Il m'a fait confiance lorsqu'il m'a accepté en doctorat et c'est grâce à ses conseils et sa patience que ces travaux ont pu arriver à leurs termes. Il est un Directeur de thèse très particulier, ayant des valeurs humaines exceptionnelles et doté d'un grand savoir-faire. Je le remercie pour son aide et son écoute aux moments les plus difficiles. Je le remercie pour son soutien scientifique et surtout moral. J'ai appris tellement de choses de lui que je suis fière et ravie d'avoir été un jour son étudiante.

Je tiens à exprimer ma profonde reconnaissance au **Professeur Yassine RUICHEK**, mon Directeur de thèse en France, pour l'honneur qu'il m'a fait en acceptant de diriger cette thèse. Il donne de la valeur à chaque travail scientifique avec ses précieux conseils et ses nombreuses corrections. Je le remercie infiniment pour l'appui scientifique et moral qu'il m'avait octroyé. Merci pour son encouragement continu et en particulier durant la dernière période de la rédaction de cette thèse. J'aimerai bien lui dire à travers ce remerciement qu'il est un Directeur de thèse exceptionnel, doté d'une perfection, d'un dynamisme, d'une ouverture d'esprit et d'une exigence scientifique fascinante.

Je remercie **Mme Raja TOUAHNI**, Professeur à l'université Ibn Tofail, qui a contribué au bon déroulement de cette thèse et qui est à la fois une professeur et une amie de tous. Merci chère madame pour votre disponibilité, votre écoute et vos précieux conseils. J'ai appris d'elle beaucoup de choses utiles dans la vie personnelle et la vie professionnelle.

Un merci spécial à **Mme Caroline DELAMARCHE**, gestionnaire du bureau doctoral à l'université Bourgogne Franche-Comté, pour les efforts qu'elle a fournis pour me faciliter les démarches administratives.

J'adresse tous mes remerciements à tous les membres des deux laboratoires pour la vivante et conviviale atmosphère au bureau et pour nos échanges généreux et agréables, au laboratoire et en dehors. Leur soutien a été inestimable tout au long de mon étude de doctorat, rendant mon temps à la fois mémorable et agréable. J'ai beaucoup apprécié les pauses-café quotidiennes avec mes collègues, ainsi qu'avec mes chers professeurs.

Enfin et surtout, je voudrais adresser mes plus sincères remerciements à mes parents pour leur soutien continu qui m'a aidé à rester forte et concentrée sur le travail de thèse. Merci mon père d'avoir minutieusement relu ma dissertation. J'adresse aussi mes profonds remerciements à mes sœurs, à mon frère, à mon mari et à tous mes proches pour leur affection et leurs encouragements.

Résumé

L'expression faciale est l'un des moyens non verbaux les plus couramment utilisés par les humains pour transmettre les états émotionnels internes et, par conséquent, joue un rôle fondamental dans les interactions interpersonnelles. Bien qu'il existe un large éventail d'expressions faciales possibles, les psychologues ont identifié six expressions fondamentales (la joie, la tristesse, la surprise, la colère la peur et le dégoût) universellement reconnues. Il est évident qu'un système capable de réaliser une reconnaissance automatique des émotions humaines est une tâche souhaitable pour un ensemble d'applications telles que l'interaction homme-machine, la sécurité, l'informatique affective, etc. Le travail de cette thèse vise à concevoir un système robuste de reconnaissance d'expressions faciales (REF). Un système de REF peut être divisé en trois modules, à savoir l'enregistrement du visage, l'extraction de caractéristiques et la classification. Dans cette thèse, nous nous sommes intéressés à chaque module du système de REF. Dans le premier module, nous présentons une nouvelle méthode efficace de représentation de l'image d'un visage, grâce à une décomposition automatique de l'image du visage en régions d'intérêt (ROI) en se basant sur des points faciaux. Cette méthode consiste à extraire sept ROIs représentant plus précisément des composantes faciales impliquées dans l'expression des émotions (sourcil gauche, sourcil droit, œil gauche, œil droit, entre sourcils, nez et bouche). Ceci permet de garantir un meilleur enregistrement du visage et par la suite une représentation faciale appropriée. Dans le deuxième module, chaque ROI est caractérisée par plusieurs descripteurs de texture, de forme, de géométrie et de leur combinaison. Enfin, dans le troisième module, deux classifieurs (SVM et Random Forest) ont été mis en œuvre pour classer une image d'entrée en une de six expressions faciales de base et l'état neutre. En terme d'évaluation, la décomposition faciale proposée est comparée aux méthodes existantes pour montrer son efficacité, en utilisant plusieurs jeux de données publics d'émotions posées et spontanées. Les résultats expérimentaux ont montré la supériorité de notre décomposition faciale par rapport à celles existantes. Ensuite, une comparaison avec les méthodes REF de l'état de l'art est réalisée à l'aide des jeux de données CK+ et SFEW.

L'analyse des résultats a démontré que notre méthode surpasse ou concurrence les résultats obtenus par les méthodes comparées. Par la suite, en se basant sur certains modules de la méthode précédente, une nouvelle technique REF a été proposée pour classer les émotions à partir d'une multi-observation (une séquence d'images ou un ensemble d'images), en utilisant un SVM multi-classe avec plusieurs stratégies. Cette technique a été évaluée sur plusieurs jeux de données publics, et comparée avec les méthodes de la littérature en utilisant deux jeux de données publics CK+ et Oulu-CASIA. Les résultats issus de cette approche multi-observation (ou dynamique) de REF dépassent généralement ceux obtenus par les méthodes de l'état de l'art de la même catégorie (c'est-à-dire à base de multi-observation).

Abstract

Facial expression is one of the most commonly used nonverbal means by humans to transmit internal emotional states and, therefore, plays a fundamental role in interpersonal interactions. Although there is a wide range of possible facial expressions, psychologists have identified six fundamental ones (happiness, sadness, surprise, anger, fear and disgust) that are universally recognized. It is obvious that a system capable of performing automatic recognition of human emotions is a desirable task for a set of applications such as human-computer interaction, security, affective computing, etc. The work of this thesis aims to design a robust facial expression recognition system (FER). FER system can be divided into three modules, namely facial registration, feature extraction and classification. In this thesis, we are interested to all the modules of FER system. In the first module, we present a new and effective method to represent the face image, thanks to an automatic decomposition of the face image into regions of interest (ROI) based on facial points. This method consists of extracting seven ROIs representing more precisely facial components involved in the expression of emotions (left eyebrow, right eyebrow, left eye, right eye, between eyebrows, nose and mouth). This ensures better face registration and therefore an appropriate facial representation. In the second module, each ROI is characterized by several descriptors of texture, shape, geometry and their combination. Finally, in the third module, two classifiers (SVM and Random Forest) have been trained and used to classify an input image into one of the six basic facial expressions and the neutral state. In terms of evaluation, the proposed facial decomposition is compared with existing ones to show its effectiveness, using several public datasets of posed and spontaneous emotions. The experimental results showed the superiority of our facial decomposition against existing ones. Then, a comparison with the state-of-the-art FER methods is carried out using the CK+ and SFEW datasets. The comparison analysis demonstrated that our method outperforms or competes the results achieved by the compared methods. Thereafter, based on modules of the previous method, a new FER technique was proposed to classify emotions from a multi-observation (image sequence

or subset of images), using a multi-class SVM with several strategies. This technique has been evaluated on several public datasets, and compared with the state-of-the-art methods using CK+ and Oulu-CASIA public datasets. The results of the proposed multi-observation (or dynamic) based FER approach generally exceed those obtained by the state-of-the-art methods of the same category (i.e. based on multi-observation).

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

AAM	Active Appearance Model
ACP	Analyse en composantes Principales
ASM	Active Shape Model
AU	Action Unit
BDBN	Boosted Deep Belief Network
BN	Bayesian Network
CDPM	Cascade Deformable Part Models
CK	Cohn Kanade
CK+	Extended Cohn Kanade
CLBP	Compound Local Binary Pattern
CLQP	Completed Local Quantized Pattern
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
CV	Cross-Validation
DCN	Deep Convolutional Network
DCNNs	Deep Convolutional Neural Networks
DBN	Deep Belief Network
DNN	Deep Neural Networks
DPM	Deformable Partsbased Model
DyBN	Dynamic Bayesian Network
EOH	Edge Orientation Histograms
FACS	Facial Action Coding System
FEED	Facial Expressions and Emotion Database

GL Gauss–Laguerre

GLTP Gradient Local Ternary Patterns

GSDM Global Supervised Descent Method

GSRRR Group Sparse Reduced-Rank Regression

HCRF Hidden Conditional Random Fields

HMM Hidden Markov Model

HOG Histogram of Oriented Gadient

ICA Independent Component Analysis

IHM Interaction Homme Machine

JAFFE Japanese Female Facial Expression

KDEF Karolinska Directed Emotional Faces

KNN K Nearest Neighbor

LBP Local Binary Pattern

LBP-TOP Local Binary Pattern on three orthogonal planes

LDA Linear Discriminant Analysis

LDTP Local Directional Ternary Pattern

LGBP Local Gabor Binary Pattern

LMP Local Motion Patterns

LOSO Leave-One-Subject-Out

LOSeqO Leave-One-Seqence-Out

LPQ Local Phase Quantization

LSiBP Local Saliency-inspired Binary Pattern

LTP Local Ternary Pattern

LTP-TOP Local Ternary Pattern on three orthogonal planes

MKL Multiple KerneL

MRL Multinomial Regression Logistic

MVFER (Multi-view Facial Expression Recognition

NB Naive Bayes

NCM Normalized Central Moments

NN Neural Networks

PDM Point Distribution Model

PHOG Pyramid Histogram of Oriented Gadient

PLS Partial Least Squares

RBF Radial Basis Function

RBM Restricted Boltzmann Machine

REF Reconnaissance des Expressions Faciales

RF Random Forest

ROI Region Of Interest

SDM Supervised Descent Method

SFEW Static Facial Expressions in the Wild

SIFT Scale-Invariant Feature Transform

SSS Stochastic Structure Search

STCLQP SpatioTemporal Completed Local Quantized Pattern

SVM Support Vector Machines

TAN Tree Augmented Naïve Bayes

VJ Viola-Jones

Introduction

Contexte

Que savez-vous de votre visage ? Avez-vous déjà utilisé ces expressions "Rire du bout des dents", "Avoir le front de faire quelque chose", "Faire bonne figure", "S'en mordre les lèvres", "Avoir les paupières lourdes", "Rester bouche bée" ?. Ces expressions sont largement utilisées par les peuples du monde pour faire passer une sensation, une émotion ou une idée par le biais des traits du visage. De là, nous pouvons imaginer le rôle important du visage dans la communication et sa puissance de transmettre un message spécifique à un interlocuteur. Donc, il n'est pas surprenant que l'expression du visage ait été un des domaines de recherche centraux sur le comportement humain depuis plus de cent ans. Le visage est l'un des canaux les plus puissants de la communication non verbale [45]. Il est difficile de comprendre l'état d'esprit interne de la personne sans voir son visage. Par conséquent, les expressions faciales jouent un rôle irremplaçable dans la communication non verbale. Elles communiquent l'émotion et signalent les intentions, la vigilance, la douleur et les traits de personnalité [54]. Les émotions peuvent être exprimées à la fois verbalement et non verbalement. Il existe de nombreux canaux tels que la voix, le visage et les gestes corporels à travers lesquels l'information non verbale est transmise aux observateurs. En outre, Mehrabian et Ferris [140] ont indiqué que l'expression faciale du locuteur contribue à 55% à l'effet du message parlé, alors que la partie verbale et la partie vocale qu'avec 7% et 38% respectivement. Ainsi, le visage a tendance à être la forme la plus visible de la communication de l'émotion. Il fait de la reconnaissance d'expression faciale un moyen largement utilisé pour mesurer l'état émotionnel des êtres humains. A cet égard, les expressions faciales fournissent des informations aux observateurs sur l'expérience émotionnelle d'une personne. Par exemple, la tristesse émotionnelle est signalée aux observateurs en soulevant les coins intérieurs des sourcils, en soulevant légèrement les joues et en tirant les coins des lèvres vers le bas [54]. Lors d'une interaction sociale, les individus évaluent leurs émotions respectives, puis ajustent leur comportement en conséquence [73]. Par conséquent, si une émotion n'est pas reconnue correctement, le comportement approprié qui devrait suivre cette expression n'est

pas manifesté, perturbant ainsi le flux naturel de l'interaction sociale [145]. Autrement dit, les expressions faciales peuvent non seulement changer le flux de la conversation [25] mais aussi fournir aux auditeurs un moyen de communiquer une grande quantité d'informations au locuteur sans même prononcer un seul mot [243]. Selon [27, 68], lorsque l'expression faciale ne coïncide pas avec les mots parlés, alors l'information véhiculée par le visage prend plus de poids dans le décodage des informations.

Du fait que les machines et les personnes commencent, progressivement, à coexister et à partager en commun diverses tâches, le besoin de canaux de communication efficaces entre eux devient de plus en plus important. Jusqu'à présent, les appareils que nous utilisons sont indifférents à nos états affectifs et émotionnellement aveugles. Si tel est le cas, l'Interaction Homme-Machine (IHM) devrait être perfectionnée de façon à simuler plus étroitement l'interaction homme-homme. Dans cette optique, une communication homme-homme réussie doit reposer sur la capacité de lire des signaux affectifs et émotionnels. Cependant, l'IHM qui ne tient pas compte des états affectifs de ses utilisateurs ignore l'essentiel de l'information disponible dans l'interaction. Quoiqu'il en soit, les humains sont capables de détecter et interpréter, aisément, l'information dans une scène. Par contre, cette même tâche est très difficile à réaliser par les machines. En conséquence, les changements dans le visage et le corps d'une personne (interlocuteur) doivent être modélisés en utilisant des caractéristiques correctement choisies et obligatoirement suivies, et classées en temps réel. De ce fait, un système typique pour l'analyse automatique des expressions faciales est basé sur des algorithmes informatiques qui tentent d'interpréter les mouvements du visage et les changements de caractéristiques faciales à partir des images faciales [204, 154, 79, 205].

Objectifs et Contributions

L'objectif ultime de cette recherche est de concevoir, implémenter et évaluer un nouveau système de reconnaissance d'expression faciale en proposant diverses approches basées principalement sur la définition de régions faciales susceptibles d'avoir des changements d'apparence, lors de l'expression de différentes émotions, constituant des formes très discriminatives et distinctives. Ces régions doivent être adéquates pour la reconnaissance des émotions posées et spontanées à partir d'images statiques et de séquences d'images.

Les principales contributions de cette thèse peuvent être résumées comme suit :

- Une nouvelle décomposition faciale basée principalement sur le système de codage des actions faciales [59] a été définie. Cette décomposition faciale est comparée empiriquement avec celle utilisant le visage entier et les différentes décompositions de la littérature. Ensuite, une étude comparative de la représentation faciale basée sur

des descripteurs d'apparence (forme :HOG, Texture :LBP, ses variantes, ainsi que la combinaison de la forme et la texture) est effectuée. Deux différentes méthodes d'apprentissage automatique (SVM et forêts aléatoires) sont systématiquement examinées dans plusieurs bases de données publiques.

- Un système dynamique simple et efficace, basé uniquement sur les probabilités estimées par le classifieur SVM, est proposé afin de classer une multi-observation représentant une émotion dans un ensemble d'images. Cette multi-observation peut correspondre à un groupe de personnes (chaque personne représente une observation) ou une personne (séquence d'images ou des images de différents points de vue de la même personne). L'approche proposée est largement évaluée en utilisant plusieurs bases de données publiques.
- Etude comparative de descripteur géométrique, descripteur d'apparence et leur combinaison.
- Une nouvelle approche basée sur le regroupement des émotions qui partagent les mêmes représentations/déformations de composantes faciales (par exemple, la surprise et la peur) est proposée. En fait, les six émotions de base sont regroupées en deux ou quatre groupes d'émotions. L'idée est d'identifier des sous-ensembles d'expressions faciales caractérisées par des confusions spécifiques similaires sur les plans sémantique et physique. Le fondement de cette approche est basé principalement sur l'étude des matrices de confusion et le travail de Rachael at al. [91].

Organisation de la thèse

Le reste de cette thèse est structuré comme suit :

Le chapitre 1 présente tout d'abord un bref aperçu des principales théories émotionnelles, avec un accent particulier sur les expressions faciales. Il décrit ensuite les étapes de traitement qui composent un système d'analyse d'expression faciale, tout en discutant diverses techniques utilisées à chaque étape.

Le chapitre 2 fournit des exemples de divers domaines d'application. Il donne ensuite un aperçu général sur quelques bases de données utilisées dans la littérature pour évaluer la reconnaissance des expressions faciales. Le chapitre se termine par une comparaison des systèmes existants de reconnaissance d'expression faciale.

Le chapitre 3 propose une nouvelle décomposition faciale, puis présente une étude empirique complète de l'enregistrement du visage en considérant différentes décompositions faciales d'une part, et d'autre part la représentation faciale basée sur la texture, en utilisant

l'opérateur LBP et ses variantes, la forme, en utilisant le descripteur HOG, et leur combinaison. Enfin, nous examinons deux différentes méthodes d'apprentissage automatique SVM et forêts aléatoires pour la reconnaissance d'expression faciale.

Le chapitre 4 présente une méthode de reconnaissance de l'expression faciale en exploitant des images issues de la multi-observation (sous-ensemble d'images, vidéo,etc.), en utilisant les probabilités estimées par SVM.

Le chapitre 5 présente quelques résultats préliminaires de nos futures recherches. Le chapitre est divisé en deux parties. La première partie introduit un nouveau descripteur géométrique. Une comparaison de trois différentes méthodes d'extraction de caractéristiques (méthode géométrique, méthode d'apparence et méthode de combinaison) est ensuite effectuée. La deuxième partie est dédiée à l'analyse des expressions faciales, en regroupant les émotions en deux ou quatre groupes d'émotions selon les critères retenus.

Enfin, une conclusion résume les travaux de cette thèse et expose les limites actuelles ainsi que les orientations futures dégagées.

Publications

1. Lekdioui, K., Messoussi, R., Ruichek, Y., Chaabi, Y., and Touahni, R. (2017). Facial decomposition for expression recognition using texture/shape descriptors and svm classifer. Signal Processing : Image Communication.
2. Lekdioui, K., Ruichek, Y., Messoussi, R., Chaabi, Y., and Touahni, R. (2017). Facial expression recognition using face-regions. In Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on, pages 1–6. IEEE.
3. Lekdioui, K., Messoussi, R., and Chaabi, Y. (2015). Etude et modélisation des comportements sociaux d'apprenants à distance, à travers l'analyse des traits du visage. In 7ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2015), pages 411–413

Chapitre 1

Analyse des expressions faciales : Etat de l'art

Sommaire

1.1 Introduction

Au cours de la dernière décennie, la communauté de la recherche en vision par ordinateur a montré beaucoup d'intérêt pour l'analyse et la reconnaissance automatique des expressions faciales. Initialement inspirée par les découvertes des chercheurs en sciences cognitives, la communauté de la vision par ordinateur et de la recherche scientifique envisageait de développer des systèmes capables de reconnaître les expressions faciales dans des vidéos ou

des images statiques. La plupart de ces systèmes d'analyse des expressions faciales tentent de classer les expressions en quelques grandes catégories émotionnelles, telles que la joie, la tristesse, la colère, la surprise, la peur et le dégoût.

Ce chapitre présente un bref aperçu des principales théories émotionnelles, avec un accent particulier sur les expressions faciales (Section 1.2). Ensuite, il décrit les étapes de traitement pertinentes au cours desquelles un système d'analyse d'expression faciale peut être décomposé tout en discutant les diverses techniques utilisées à chaque étape (Section 1.3).

1.2 Expressions faciales vs émotions

Les expressions faciales émotionnelles sont des changements faciaux traduisant des états émotionnels internes, des intentions ou des communications sociales d'une personne. L'analyse de l'expression faciale était un sujet de recherche actif pour les scientifiques du comportement depuis le travail de Darwin en 1872 [43, 53, 56, 175]. Il est important de souligner dès le début qu'il y a une distinction d'un point de vue de la vision par ordinateur entre la reconnaissance de l'expression faciale (REF) et la reconnaissance des émotions humaines. Comme l'expliquent Fasel et Luettin [65], la première traite la classification du mouvement facial et la déformation des traits faciaux en classes abstraites basées sur des informations visuelles. La deuxième est le résultat de nombreux facteurs différents et peut être révélée sur plusieurs canaux, par exemple, la voix, la pose, les gestes, la direction du regard et l'expression faciale. En effet, la correspondance d'une expression faciale à une émotion implique la connaissance des catégories d'émotions humaines auxquelles des expressions faciales peuvent être attribuées.

1.2.1 Emotion

Les émotions sont sophistiquées et subtiles et peuvent être analysées à partir de différentes perspectives. Les théories sociales expliquent les émotions comme étant des produits du conditionnement culturel et social, puisque la manière dont nous, les humains, exprimons nos émotions reflète notre environnement social. Les premiers travaux sur les émotions ont été faits par Darwin qui avait recueilli du matériel depuis 1838. Son intention était de montrer comment les expressions des émotions chez l'homme étaient analogues à celles chez les animaux. L'émotion est un processus, un type particulier d'évaluation automatique influencé par notre passé évolutionnaire et personnel, dans lequel nous sentons que quelque chose d'important pour notre bien-être se produit, puis un ensemble de changements physiolo-

giques et comportementaux commence à gérer la situation. Les émotions produisent des changements dans les parties de notre cerveau qui nous mobilisent pour faire face à ce qui a déclenché l'émotion, ainsi que des changements dans notre système nerveux autonome, qui régule notre fréquence cardiaque, notre respiration, notre transpiration et bien d'autres changements corporels. Les émotions envoient aussi des signaux, des changements dans nos expressions, notre visage, notre voix et notre posture corporelle. Nous ne choisissons pas ces changements ; ils arrivent spontanément.

1.2.2 Expressions faciales

L'expression faciale est le moyen le plus expressif pour l'être humain de communiquer des émotions et de signaler des intentions, ce qui véhicule des signaux de communication non verbaux dans les interactions face à face. Une expression faciale est une manifestation visible de l'activité, de l'intention, de la personnalité et de la psychopathologie d'une personne [49]. L'expression faciale, ainsi que d'autres gestes, transmettent des signaux de communication non verbaux dans l'interaction face à face humaine. Ces indices peuvent également compléter la parole en aidant l'auditeur à obtenir la signification voulue des mots parlés. Elles jouent un rôle important dans nos relations. Elles peuvent révéler l'attention, la personnalité, l'intention et l'état psychologique d'une personne [191]. Ce sont des signaux interactifs qui peuvent réguler nos interactions avec l'environnement et d'autres personnes dans notre voisinage [159]. Basée sur le travail de Darwin [44], ainsi que sur celui d'Ekman [51, 60, 59], la recherche dans le domaine de l'analyse automatique de l'expression faciale se concentre plus sur six expressions faciales émotionnelles prototypiques (joie, surprise, peur, tristesse, dégoût et colère). Cependant, ces expressions de base ne représentent qu'un petit ensemble d'expressions faciales humaines. En fait, l'émotion humaine est composée de centaines d'expressions, bien que la plupart d'entre elles diffèrent par des changements subtils de quelques traits du visage. Les expressions faciales pour l'émotion résultent principalement des signaux faciaux rapides. Ces mouvements temporaires des muscles faciaux tirent la peau, changeant temporairement la forme des yeux, des sourcils et des lèvres, ainsi que l'apparition de plis, et de sillons dans différentes parties de la peau. Ces changements ne durent que quelques secondes.

1.2.3 Système FACS

Du point de vue physiologique, l'expression faciale est une conséquence de l'activité musculaire faciale. Ces muscles sont également appelés muscles mimétiques ou muscles des expressions faciales. Ils font partie du groupe des muscles de la tête, qui contiennent en outre

des muscles du cuir chevelu, des muscles de la mastication responsables du déplacement de la mâchoire et de la langue. Les muscles du visage sont innervés par le nerf facial, qui se ramifie dans le visage et son activation provoque des contractions ce qui se traduit par divers mouvements observables. Les actions musculaires habituellement visibles sont des blocs de mouvement de la peau (par ex. les sourcils, les lèvres, les joues) et les rides (par ex. sur le front, entre les sourcils ou sur le nez). L'étude de l'expression faciale ne peut se faire sans l'étude de l'anatomie du visage et de la structure sous-jacente des muscles faciaux. Les chercheurs ont concentré leur attention sur un système de codage pour les expressions faciales. Facial Action Coding System (FACS), initialement développé par Ekman et Friesen en 1978 [59], est le système de codage le plus largement utilisé dans les sciences du comportement. Le système a été initialement développé en analysant des séquences vidéo d'une gamme d'individus et en associant les changements d'apparence faciale avec les contractions des muscles sous-jacents. Cette étude a permis le codage de 44 unités d'action distinctes (AUs), c'est-à-dire anatomiquement liées à la contraction de muscles faciaux spécifiques, chacune étant intrinsèquement liée à un petit ensemble d'activations musculaires localisées (voir Table 1.1). En utilisant FACS, on peut coder manuellement presque n'importe quelle expression faciale anatomiquement possible, en la décomposant en AUs spécifiques et leurs segments temporels qui ont produit l'expression. Toutes les expressions résultantes peuvent être décrites en utilisant les 44 AUs décrites par Ekman ou une combinaison des 44 AUs. Ekman et Friesen [57] ont également postulé six émotions primaires qu'ils considèrent comme universelles à travers les ethnies et les cultures humaines. Ces six émotions universelles, communément appelées émotions de base, sont : la joie, la colère, la surprise, le dégoût, la peur et la tristesse (voir Figure 1.1). L'étude principale d'Ekman et Friesen [57] a formé l'origine de l'analyse d'expression faciale, quand les auteurs ont proclamé que les six expressions faciales prototypiques de base sont universellement reconnues. La plupart des chercheurs affirment que ces catégories d'expressions ne sont pas suffisantes pour décrire toutes les expressions faciales en détail. Cependant, la plupart des analyseurs d'expression faciale existants utilisent encore la théorie d'Ekman et Friesen. Il y a un lien entre FACS et les émotions de base. Chaque émotion de base a été décrite par Ekman et Friesen [59] en utilisant des indices spécifiques décrivant l'activité faciale. Cette description est résumée dans la table 1.2. Bien que les humains reconnaissent les expressions faciales sans effort ni retard, la reconnaissance fiable de l'expression faciale par une machine reste un défi.

Dans notre travail, nous étudierons la reconnaissance d'expression faciale en utilisant ces catégories d'émotions de base. L'analyse des émotions universelles nous donne la possibilité d'évaluer notre travail par rapport au travaux d'autres chercheurs qui utilisent le même cadre. Dans notre travail, nous considérons chaque classe particulière d'expressions faciales comme

la source ou la représentation directe de l'état émotionnel humain équivalent. Ainsi, les termes «expression faciale» et «émotion faciale» sont utilisés de manière interchangeable dans le reste de cette thèse.

TABLE 1.1 Les AUs fréquentes décodées par le manuel FACS sur la base d'images CK+

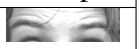
AU	Nom	Exemple
1	Remontée de la partie interne des sourcils	
2	Remontée de la partie externe des sourcils	
4	Abaissement et rapprochement des sourcils	
5	Ouverture entre la paupière supérieure et les sourcils	
6	Remontée des joues	
7	Tension de la paupière	
9	Plissement de la peau du nez vers le haut	
12	Étirement du coin des lèvres	
15	Abaissement des coins externes des lèvres	
16	Ouverture de la lèvre inférieure	
20	Étirement externe des lèvres	
23	Tension refermante des lèvres	
26	Ouverture de la mâchoire	

TABLE 1.2 Combinaison des AUs correspondant à chaque émotion

Émotion	Action Units
Joie	6+12
Tristesse	1+4+15
Surprise	1+2+5+26
Peur	1+2+4+5+20+26
Colère	4+5+7+23
Dégoût	9+15+16

1.2.4 Psychologie des émotions basales

1.2.4.1 Joie

La joie est une émotion positive souvent associée à un sourire sur le visage. L'émotion de joie apparaît lors de la réalisation des objectifs. Les caractéristiques typiques sur le visage sont :

(a) Neutralité (b) Joie (c) Tristesse (d) Surprise

(e) Colère (f) Peur (g) Dégoût

FIGURE 1.1 Exemples d'expressions faciales de la base CK+.

- Les coins des lèvres tirés vers le haut.
- La bouche peut être ouverte et les dents peuvent être visibles.
- Les joues peuvent être soulevées.
- Les rides sous la paupière inférieure ainsi que les rides aux coins des yeux peuvent apparaître.

1.2.4.2 Surprise

La surprise est une émotion soudaine. Cela vient sans réfléchir et ne dure que peu de temps. Le début est une situation inattendue ou faussement attendue. Par conséquent, une émotion surprenante peut être positive ou négative. La surprise ne peut être anticipée. Si la situation s'accompagne d'un temps de réflexion, la réaction ultérieure ne sera pas une surprise. La surprise entraîne souvent d'autres émotions, en l'occurrence, la joie ou la tristesse. Les caractéristiques typiques de la surprise sont :

- La remontée des parties interne et externe des sourcils.
- Des rides horizontales pouvant apparaitre sur le front.
- L'ouverture de la bouche et des yeux.

1.2.4.3 Peur

La peur est induite par des situations dangereuses ou stressantes. On peut ressentir la peur des événements futurs, par ex. peur de la misère. Pour éprouver la peur, le corps de

la personne se prépare à une évasion ou à une défense contre toute attaque possible. Les caractéristiques typiques sur le visage sont :

- Les yeux sont ouverts et les pupilles deviennent larges.
- Les sourcils sont soulevés et tirés vers l'intérieur.
- Les paupières supérieures sont soulevées, ce qui a pour effet d'afficher un sanpaku supérieur (c-à-d, un blanc des yeux visible).
- Les muscles de la bouche se tendent et abaissent souvent la lèvre inférieure.

1.2.4.4 Colère

La colère est une réaction émotionnelle forte et peut également être une émotion dangereuse car elle pourrait provoquer la violence. La source de la colère a de nombreuses raisons (par ex. nous pourrions ressentir de la colère contre un obstacle sur la voie du succès, quand quelqu'un veut nous faire du mal, les menaces verbales, les réclamations provoquent aussi l'émotion de la colère). La colère a un impact important sur tout le corps. L'augmentation de la pression sanguine, le visage rouge et la tension dans les muscles sont généralement reflétées. La réponse physiologique induit les signaux suivants :

- Les sourcils s'abaissent et se serrent ensemble, ce qui entraine l'apparence des rides verticales entre les sourcils.
- Les lèvres sont fermées hermétiquement ou doucement ouvertes (se préparant à crier). Ensemble, ces caractéristiques préparent souvent le corps à une éventuelle attaque physique ou verbale.

1.2.4.5 Dégoût

Le dégoût est une émotion négative généralement évoquée par l'odorat, le goût ou la vision. Contrairement aux autres émotions, les objets évoquant le dégoût ne sont pas universels, mais culturels ou personnels, par ex. aliments. La réponse physiologique extrême est le vomissement. Les caractéristiques les plus significatives du visage sont dans la bouche et le nez :

- La lèvre supérieure est levée, plus le dégoût est grand plus la dentition supérieur est visible.
- Les rides apparaissent sur le nez.

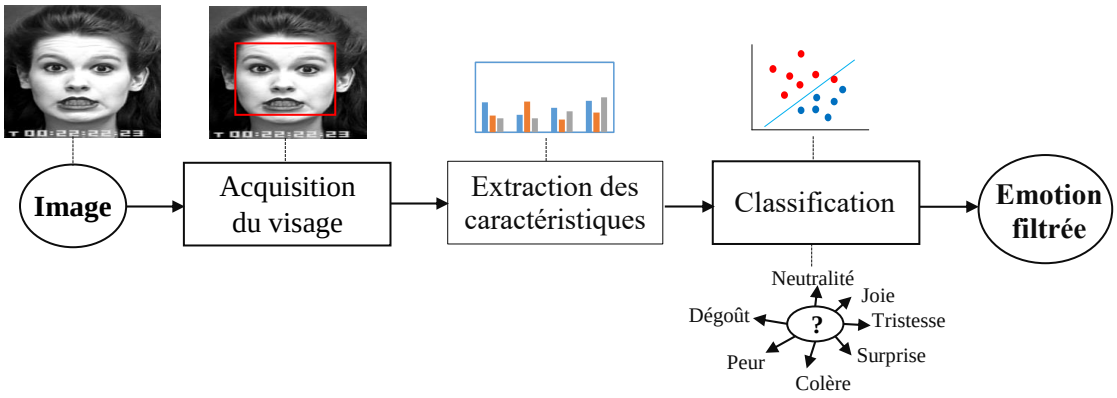


FIGURE 1.2 Modules du système d'analyse des expressions faciales

1.2.4.6 Tristesse

La tristesse apparaît quand une personne souffre. L'origine de la tristesse est typiquement une perte de quelque chose. Cette émotion calme et non impulsive, est souvent accompagnée de larmes. Pendant l'émotion, les muscles du visage perdent la tension qui peut entraîner les caractéristiques physiologiques typiques suivantes :

- Les parties internes des sourcils sont abaissées.
- Les coins des lèvres s'abaissent.

1.3 Modules du système d'analyse des expressions faciales

Un système qui effectue une reconnaissance automatique des expressions faciales est généralement composé de trois modules principaux, comme illustré dans la figure 1.2. Le premier module consiste à détecter et enregistrer la région du visage dans les images ou les séquences d'images d'entrée. Il peut s'agir d'un détecteur pour détecter le visage dans chaque image ou simplement détecter le visage dans la première image, puis suivre le visage dans le reste de la séquence vidéo. Le deuxième module consiste à extraire et représenter les changements faciaux causés par les expressions faciales. Le dernier module détermine une similarité entre l'ensemble des caractéristiques extraites et un ensemble de caractéristiques de référence. D'autres filtres ou modules de prétraitement de données peuvent être utilisés entre ces modules principaux pour améliorer les résultats de détection, d'extraction de caractéristiques ou de classification.

Un aperçu plus détaillé de chaque module sera donné dans le reste de ce chapitre.

1.3.1 Acquisition du visage

Le problème de détection et d'enregistrement des visages implique l'identification de la présence de visages dans une image et la détermination des emplacements et des échelles des visages. La précision de la détection et l'enregistrement du visage est particulièrement importante dans des conditions réalistes, où la présence du visage dans une scène et sa localisation globale ne sont pas connues a priori. Un système complet de localisation du visage devrait faire face à certains défis décrits ci-dessous.

Pose : Les traits du visage, y compris les yeux, le nez et la bouche, peuvent être partiellement invisibles ou déformés en raison de la pose relative du visage ou de la caméra.

Occlusion : Les traits du visage peuvent être obstrués par une barbe, une moustache et des lunettes. De même, le maquillage peut provoquer l'apparition de régions artificielles sur le visage ou cacher les limites faciales normales.

Expression : les traits du visage montrent de grands changements dans leur forme sous différentes expressions. Certaines caractéristiques peuvent devenir invisibles ou d'autres ne sont visibles que sous différentes expressions.

Conditions d'acquisition de l'image : L'éclairage et les changements dans les caractéristiques de la caméra affectent de manière significative la chrominance des régions du visage. Certaines caractéristiques peuvent être masquées ou combinées avec des ombres ou des brillances sur le visage provoquant ainsi une perte d'informations sur les couleurs.

1.3.1.1 Détection du visage

La première étape d'un système d'analyse des expressions faciales entièrement automatique consiste à localiser la région du visage et ses limites. Le but de la détection du visage est de déterminer si un visage est présent ou non sur l'image et, le cas échéant, de localiser son emplacement (voir Figure 1.2, le rectangle rouge englobe le visage détecté).

Une étude exhaustive sur les algorithmes de détection de visages dans [236] a regroupé les différentes méthodes en quatre catégories : des méthodes basées sur la connaissance, des approches invariantes, des méthodes d'appariement de modèles et des méthodes basées sur l'apparence.

1. Les méthodes basées sur la connaissance utilisent des règles prédéfinies basées sur la connaissance humaine afin de détecter un visage (par ex. un visage comprend deux yeux, un nez et une bouche avec des positions relatives clairement définies entre elles).
2. Les approches basées sur les caractéristiques invariantes visent à trouver des caractéristiques de structure du visage robustes aux conditions de pose et d'éclairage.

3. Les méthodes basées sur la correspondance avec les modèles utilisent des modèles de visage pré-stockés. Les valeurs de corrélation entre le modèle et l'image d'entrée sont calculées et la présence du visage est déterminée en utilisant ces valeurs de corrélation,
4. Les méthodes basées sur l'apparence utilisent l'apprentissage automatique et les techniques statistiques pour modéliser le visage. Les modèles appris se présentent généralement sous la forme de distributions et de fonctions discriminantes, et lesdits modèles sont utilisés pour distinguer les objets faciaux ou non-faciaux.

La catégorisation ci-dessus, cependant, ne s'applique guère aux méthodologies récentes développées depuis [216]. Par conséquent, selon [246], les algorithmes de détection du visage peuvent se décomposer en deux grandes catégories.

1. La première catégorie est basée sur des modèles rigides et comprend les variations de boosting. Les principaux algorithmes de cette catégorie comprennent l'algorithme de détection de visage Viola-Jones (VJ) et ses variations [216, 217], les algorithmes basés sur des réseaux neuronaux convolutionnels (Convolutionnel Neural Network, CNN) et CNN profonds (Deep CNN, DCNN) [248], et les méthodes [182] qui appliquent des stratégies inspirées de l'extraction d'images (image-retrieval) et la transformée généralisée de Hough [13].
2. La deuxième catégorie est basée sur l'apprentissage et l'application d'un modèle des parties déformables (Deformable Part Model, DPM) [66, 67] pour modéliser une déformation potentielle entre les parties faciales. Ces méthodes peuvent également combiner détection de visage et localisation de la partie faciale [258]. Cette famille d'algorithmes s'articule principalement autour des extensions et des variations de la méthodologie générale de détection d'objets [66, 67].

Une étude détaillée sur la détection des visages peut être trouvée dans le document [246].

Pour terminer cette section, nous allons nous concentrer sur l'algorithme VJ en raison de sa dominance dans la littérature récente. Le détecteur VJ [217] est incontestablement le plus couramment utilisé dans l'analyse automatique de la reconnaissance faciale et de ses expressions. L'idée de base de l'approche est d'entrainer un classifieur en cascade pour les caractéristiques rectangulaires pseudo-haar. Le détecteur balaye ensuite une image à différentes échelles et positions par une sous-fenêtre tandis que les régions acceptées par le classifieur sont déclarées faisant partie du visage. Les caractéristiques rectangulaires pseudo-haar peuvent être efficacement calculées avec des images intégrales [217], ce qui permet à l'approche d'obtenir une vitesse de détection en temps réel. Pour augmenter encore la vitesse de détection tout en conservant la précision, AdaBoost [70] a été utilisé pour sélectionner les caractéristiques pseudo-haar représentatives. De plus, au lieu d'entraîner un seul classifieur

fort, un certain nombre de classifieurs peu complexes (faibles) sont construits. Ces classifieurs faibles sont combinés en une cascade. La motivation est que les classifieurs simples au début de la cascade peuvent rejeter efficacement les régions sans visage, tandis que les classifieurs plus forts, plus tard dans la cascade, doivent simplement classer les régions plus semblables à des visages. Le détecteur de visage final à 38 couches atteint une précision impressionnante et une vitesse de détection très rapide. Au cours de la dernière décennie, il y a eu plusieurs tentatives d'extension de l'algorithme de détection de visage VJ bien établi pour résoudre le problème de la détection de visages à vues multiples [96, 121, 86].

1.3.1.2 Détection des points caractéristiques du visage

Les points caractéristiques du visage sont principalement situés autour des composants faciaux tels que les yeux, la bouche, les sourcils, le nez et le menton. La détection des points caractéristiques du visage commence habituellement à partir d'une boîte englobante rectangulaire renvoyée par un détecteur de visage qui localise ce dernier (voir Section précédente 1.3.1.1). Bien qu'optionnelle, cette étape de détection des points faciaux est importante car elle facilite la décomposition du visage (voir Section 1.3.1.3), l'extraction de caractéristiques géométriques telles que les contours des composants faciaux, les distances faciales, etc (voir Section 1.3.2.1) et fournit les emplacements où les caractéristiques d'apparence peuvent être calculées (voir Section 1.3.2.2). Les méthodes de détection et de suivi des points faciaux peuvent être classées soit en méthodes basées sur des modèles de formes paramétriques soit en méthodes basées sur des modèles de formes non paramétriques, selon le modèle de forme (paramétrique ou non paramétrique) utilisé dans la méthode.

1. Les méthodes basées sur un modèle de forme paramétrique sont en outre divisées en deux classes : les méthodes utilisant un modèle d'apparence basé à son tour sur des parties locales, par exemple, les Modèles de Formes Actifs (ASM) et les méthodes utilisant un modèle holistique, par exemple, les Modèles Actifs d'Apparence (AAM).
2. Les méthodes basées sur les modèles de formes non paramétriques sont en outre divisées en quatre catégories selon leur processus de construction de modèle : les méthodes basées sur des exemples [19, 186], les méthodes basées sur des modèles graphiques [40], les méthodes basées sur la régression en cascade, par exemple, Supervised Descent Method (SDM) [231] et les méthodes basées sur l'apprentissage profond, par exemple, Deep Convolution Network (DCN) [195].

Pour une étude approfondie de ces méthodes, voir [222].

Par la suite, nous décrivons brièvement les méthodes basées sur ASM, AAM et SDM qui sont très populaires pour la tâche de détection des points faciaux.

Les ASMs [39] sont des modèles statistiques de la forme des objets qui se déforment itérativement pour s'adapter à un exemple de l'objet dans une nouvelle image. Par ailleurs, l'inconvénient de l'ASM est qu'il n'utilise que des contraintes de forme (avec quelques informations sur la structure de l'image près des points faciaux), et ne profite pas de toutes les informations disponibles (par ex. la texture). Par contre, Un AAM [38] peut être découplé en un modèle de forme linéaire (Point distribution model, PDM) et un modèle de texture linéaire. Pour construire le modèle de texture, toutes les images des visages d'entraînement doivent être déformées dans le cadre de la forme moyenne par interpolation linéaire par morceaux en utilisant un maillage triangulaire ou interpolation plan utilisant des splines ; les images résultantes doivent être exemptées de variations de forme (textures sans forme ou shape-free textures). Le modèle de texture peut être généré en appliquant une Analyse en Composantes Principales (ACP) sur toutes les textures normalisées.

L'objectif principal de l'AAM est d'abord de définir un modèle puis de trouver les meilleurs paramètres correspondants entre la nouvelle image donnée et le modèle construit en utilisant un algorithme d'ajustement. En effet, l'algorithme d'ajustement est répété jusqu'à ce que les paramètres de forme et d'apparence satisfassent des valeurs particulières. Quant au modèle de forme, il est créé en combinant les vecteurs construits à partir des points des images étiquetées :

$$s = s_0 + \sum_{i=1}^{m} p_i s_i \tag{1.1}$$

où p_i sont des paramètres de forme, s_0 est la forme moyenne et les m premiers vecteurs propres de forme s_i sont obtenus par une ACP. Avant d'appliquer l'ACP, les points caractéristiques sont normalisés. La variation d'apparence est représentée par une combinaison linéaire d'une apparence moyenne $A_0(x)$ et n vecteurs de base d'apparence $A_i(x)$ comme suit :

$$A(x) = A_0(x) + \sum_{i=1}^{n} \alpha_i A_i(x) \tag{1.2}$$

où α_i est le paramètre d'apparence. Après avoir trouvé les paramètres de forme et d'apparence, une déformation affine par morceaux est appliquée pour construire l'AAM en localisant chaque pixel d'apparence sur le côté interne de la forme actuelle. L'objectif est de minimiser la différence entre l'image déformée et l'image d'apparence.

A propos de la stratégie non-paramétrique, la méthode la plus utilisée récemment est SDM. Elle a été développée par Xiong et Torre [231] pour résoudre une série de problèmes

de moindres carrés linéaires, exprimés de la manière suivantes :

$$\arg\min_{R_k,b_k} \sum_{d^i} \sum_{x_k^i} \left\| \Delta x_*^i - R_k \phi_k^i - b_k \right\|^2 \tag{1.3}$$

où $\Delta x_*^i = x_*^i - x_k^i$ est l'écart de vérité terrain entre la forme optimal x_*^i de la i^{eme} image de l'ensemble d'apprentissage et la forme x_k^i obtenue à partir de la i^{me} itération. ϕ_k^i représente les caractéristiques Scale-Invariant Feature Transform (SIFT) extraites autour de la forme x_k^i sur l'image d'apprentissage. R_k est une direction générique de descente. b_k est un terme de biais. Cette méthode possède un processus de dérivation naturelle basé sur la méthode de Newton. Des séries de $\{R_k, b_k\}$ sont apprises dans la phase d'apprentissage, et dans la phase de test, elles sont appliquées aux caractéristiques SIFT extraites de l'image de test pour mettre à jour la forme de manière séquentielle. SDM est un algorithme local et il est probable que les moyennes des directions de gradient soient contradictoires. Pour résoudre ce problème, les auteurs ont proposé le SDM global (GSDM) dans leurs travaux ultérieurs [232], c-à-d, une extension de SDM qui divise l'espace de recherche en régions de directions de gradient similaires. Puisque le calcul majeur de la méthode SDM est l'extraction de caractéristiques et la multiplication linéaire simple, elle a attiré de grandes attentions dans le domaine d'alignement de points caractéristiques du visage [165, 229, 230, 20].

1.3.1.3 Enregistrement du visage : visage entier vs régions spécifiques

L'enregistrement du visage est une étape fondamentale pour la REF. En général, il vise à trouver la région d'intérêt (Region Of Interest, ROI), principalement le visage entier, à partir de l'image d'entrée, et à normaliser la ROI trouvée en détectant certains composants faciaux internes tels que les yeux. Par exemple, le visage peut être aligné et normalisé en fonction des emplacements des yeux détectés et de la distance entre eux [179, 202], ce qui entraîne la suppression de la translation et la différence d'échelle. Cependant, cette approche simple reste sensible aux rotations de la tête et à la variation des sujets.

les stratégies d'enregistrement peuvent être séparées en trois classes principales qui sont la méthode holistique [179, 26, 4, 63], la méthode locale [82, 80, 79, 71, 244, 50, 221, 225, 32] et la méthode basée sur l'enregistrement de l'emplacement des points caractéristiques [252, 212, 181]. La première peut être simplement réalisée par la détection du visage en utilisant par exemple l'algorithme VJ [217]. La deuxième méthode consiste à enregistrer des régions locales (par ex. les yeux, le nez, la bouche) en se basant sur des points caractéristiques ou sur la géométrie du visage. La troisième méthode peut être effectuée en utilisant les coordonnées des points faciaux estimés par un détecteur de points caractéristiques (voir Section 1.3.2.1).

De nombreuses méthodes ont été appliquées pour reconnaître des expressions sur la totalité de la zone du visage détecté. En effet, certaines régions du visage sont totalement indépendantes de la production d'expression. Elles peuvent être supprimées de la zone détectée sans influence sur la reconnaissance des expressions. Ainsi, la plupart des régions du visage ne participent pas à la production d'expressions faciales. Par conséquent, différents travaux ont sélectionné les sous-régions du visage qui subissent un changement durant ou lors d'une expression faciale. Par exemple, Youssif et Asker [244] ont considéré que le visage détecté par VJ est frontal ou quasi-frontal, et en se basant sur certaines contraintes géométriques (la position à l'intérieur du visage, la taille et la symétrie par rapport à l'axe de symétrie faciale), le visage a été segmenté en trois ROIs (bouche, nez et les yeux/les sourcils). De même dans [32], après l'acquisition de la région du visage par VJ, les auteurs détectent les yeux et extraient les autres composants en fonction de leurs positions relatives. Puisque les sourcils sont au dessus des yeux, ces auteurs ont agrandi les régions des yeux détectées pour contenir les sourcils. Quant aux nez et la bouche qui se situent juste au-dessous des yeux, il n'est plus difficile de localiser la région qui les contient puisque les images du visage de la base utilisée correspondent toutes à vue frontale. Dans [50], les auteurs ont divisé le visage manuellement en six principales ROIs (sourcils, œil gauche, œil droit, entre les yeux, le nez et la bouche) puisque selon FACS [58], ce sont les régions les plus représentatives des expressions faciales. Dans [221], après avoir détecté le visage par VJ, trois régions faciales sont segmentées, extraites et normalisées par le détecteur des points faciaux AAM. Dans [82], la région du visage détectée par VJ est insérée dans l'étape d'extraction de ROIs qui estime d'abord les dimensions du visage. Ensuite, les ROIs sont automatiquement segmentées pour obtenir la région bouche et la région front/yeux en utilisant les moments de l'image et les intégrales projectives. Happy et Routray [80] ont localisé avec précision certains des points faciaux tels que les yeux, le nez et les coins des lèvres puis, en tenant compte de la géométrie du visage humain, les emplacements des ROIs peuvent être dérivés en utilisant la largeur du visage comme paramètre. Ils ont défini dix ROIs qui sont situées sur les parties du visage qui subissent un changement majeur lors d'une expression faciale (autour des yeux, des lèvres, des sourcils et du nez). Dans leur travail ultérieur, Happy et Routray [79] ont proposé une approche sans apprentissage pour détecter des points faciaux, puis ont défini dix-neuf petites régions situées autour des points faciaux. Aussi dans [225], en fonction de la position des points faciaux détectés par SDM, trois ROIs ot été extraites autour des régions du nez, de la bouche et des yeux, considérées comme contribuant davantage à la variance d'expression. De plus, pour un meilleur enregistrement du visage, Ghimire et al. [71] ont proposé de diviser le visage en régions locales de forme arbitraire tout en se basant sur les positions des points faciaux estimées en utilisant la méthode de détection de points caractéristiques présentée par

Kazemi et al. [104]. Par ailleurs, l'utilisation de petites parties du visage au lieu du visage entier pour extraire les caractéristiques réduit le coût de calcul et empêche le surajustement des caractéristiques pour la classification.

1.3.2 Extraction des caractéristiques

Une fois l'enregistrement du visage effectué, l'étape suivante consiste à extraire et représenter les changements faciaux causés par une expression faciale. L'obtention de caractéristiques efficaces d'expression faciale, à partir de ROI(s) extraite(s), est cruciale pour une reconnaissance d'expressions faciales réussie. Les expressions faciales sont définies principalement par la contraction des muscles faciaux qui produisent des changements dans l'apparence et la forme du visage [155]. De ce fait, les méthodes d'extraction des caractéristiques pour l'analyse d'expression peuvent être séparées en deux types d'approches : les méthodes basées sur les caractéristiques géométriques et les méthodes basées sur l'apparence.

Les caractéristiques géométriques représentent la forme et l'emplacement des composants du visage (y compris la bouche, les yeux, les sourcils et le nez). Les composants faciaux ou les traits faciaux sont extraits pour former un vecteur de caractéristiques représentant la géométrie du visage.

Les caractéristiques d'apparence représentent les changements d'apparence (texture de la peau) du visage, tels que les rides et les sillons. Ces caractéristiques d'apparence peuvent être extraites sur tout le visage ou sur des régions spécifiques du visage (voir Section précédente 1.3.1.3). Selon les différentes méthodes d'extraction des caractéristiques, les effets de la rotation de la tête dans le plan et les différentes échelles de prise de vue du visage peuvent être éliminés par une normalisation de ce dernier avant l'extraction des caractéristiques ou par une représentation des caractéristiques avant l'étape de reconnaissance d'expression. La reconnaissance de l'expression faciale étant la dernière étape d'un système d'analyse des expressions faciales.

1.3.2.1 Caractéristiques géométriques

Les caractéristiques géométriques décrivent la forme des composants faciaux (i.e. la bouche, les yeux, les sourcils, le nez) et leur emplacement (i.e. les coins des yeux, les coins de la bouche, etc). Elles sont représentées par des composantes faciales ou des traits faciaux, formant un vecteur caractéristique qui représente la géométrie du visage. Ainsi, la motivation pour employer une méthode basée sur la géométrie est que les expressions faciales affectent la position relative et la taille des divers traits faciaux et que, en mesurant le mouvement de certains points faciaux, l'expression faciale sous-jacente peut être déterminée. Pour que les

méthodes géométriques soient efficaces, les emplacements de ces points fiduciaires doivent être déterminés avec précision (voir Section 1.3.1.2). Les approches utilisant seulement les caractéristiques géométriques reposent principalement sur la position des points caractéristiques du visage comme des informations visuelles [158, 209, 252], ou le déplacement géométrique des points caractéristiques du visage [109], ou le paramétrage de la forme du composant du visage [35, 102].

Dans un premier travail, Zhang et al. [252] ont utilisé 34 points fiduciaires pour représenter une image de visage. Les points fiduciaires ont été sélectionnés manuellement et les coordonnées de ces points ont été utilisées comme des caractéristiques, ce qui donne un vecteur de caractéristiques à 68 dimensions. Tian et al. [201] ont proposé un modèle de composant facial multi-états pour détecter et suivre les changements des composants faciaux dans les images de visage. Ce modèle représente les mouvements du visage en mesurant les transitions d'état des composants faciaux correspondants. Dans une séquence d'images, les mouvements faciaux peuvent être modélisés en mesurant le déplacement géométrique des points caractéristiques faciaux entre l'image courante et l'image initiale. Les travaux de [158, 159] utilisent un ensemble de 20 points caractéristiques faciaux qui sont détectés en utilisant le détecteur de points faciaux proposé dans [218], en vue de décrire les expressions faciales. De même, Valstar et Pantic [212] ont utilisé les mêmes emplacements de points pour calculer des caractéristiques supplémentaires en fonction des distances et des angles entre eux, ainsi que la vitesse des déplacements des points dans les images. Ces caractéristiques ont été utilisées pour décrire le développement temporel des unités d'action (AUs). Un modèle de forme défini par 58 points faciaux a été adopté dans [31], où l'analyse des catégories d'expression de base a été effectuée sur une variété de points faciaux. Les emplacements de 68 sommets d'un ASM, faisant partie de l'AAM [137], ont été utilisés dans [99, 133] pour décrire la variation d'intensité des AUs et les expressions faciales de la douleur. L'ASM a été utilisé aussi dans [181] où des points faciaux fiables ont été extraits en appliquant la technique d'ajustement ASM. Le déplacement géométrique entre les coordonnées des points caractéristiques ASM projetés et la forme moyenne de l'ASM ont été utilisés pour évaluer l'expression faciale. De même, Lei et al. [116] se sont servi de l'ASM pour calculer la distance euclidienne entre le centre de gravité de la forme du visage et les points faciaux. Enfin, ils extraient l'information géométrique déformable discriminante entre les caractéristiques de l'expression neutre et les autres expressions. Dans [208, 210], les auteurs ont suivi 49 points faciaux en utilisant SDM [231] et les ont alignés avec une forme moyenne à partir de points stables, situés sur les coins des yeux et sur la région du nez. Pour l'extraction des caractéristiques, ils ont calculé la différence entre les coordonnées des points faciaux alignés et ceux de la forme moyenne, ainsi qu'entre les points faciaux alignés dans l'image

précédente et l'image courante. Cette procédure a fourni 196 caractéristiques au total. Ensuite, ils ont divisé les points faciaux en groupes selon trois régions différentes (œil gauche/sourcil gauche, œil droit/sourcil droit et la bouche). Pour chacun de ces groupes, les distances euclidiennes et les angles entre les points sont calculés, fournissant 71 caractéristiques. Ils ont également calculé la distance euclidienne entre la médiane des points faciaux stables et chaque point facial aligné dans une image vidéo. Au total, l'ensemble géométrique comprend 316 caractéristiques.

Par ailleurs, l'extraction des caractéristiques géométriques nécessite généralement une détection et un suivi précis et fiables des caractéristiques faciales. La détection automatique et le suivi des traits faciaux est toujours un problème ouvert dans de nombreuses situations du monde réel.

Les caractéristiques géométriques ont l'avantage de la faible dimension et de la simplicité. Néanmoins, toutes les méthodes de construction des caractéristiques géométriques souffrent de problèmes causés par la variation de l'éclairage et du mouvement non rigide. De plus, elles sont sensibles à l'erreur d'enregistrement de l'image du visage et aux discontinuités de mouvement. Par conséquent, il est difficile de concevoir un modèle physique déterministe des expressions faciales représentant exactement les propriétés géométriques faciales et les activités musculaires pour toutes les expressions faciales. C'est pourquoi, les caractéristiques basées sur l'apparence pour l'analyse de l'expression faciale ont été également étudiées.

1.3.2.2 Caractéristiques d'apparence

Les caractéristiques d'apparence décrivent la texture de la peau comme les rides et les sillons. Ces caractéristiques peuvent être extraites soit sur tout le visage, soit sur des régions spécifiques d'une image de visage (voir 1.3.1.3). La dernière décennie a connu le développement de nombreuses approches basées sur l'extraction des caractéristiques d'apparence. Les plus communément utilisées pour l'analyse d'expression faciale comprennent les ondelettes de Gabor [252, 203, 24, 138, 1, 233], les caractéristiques pseudo-Haar [154, 174, 240], l'Analyse en Composantes Principales (ACP) [46, 1], Analyse Discriminante Linéaire (LDA) [46, 177], Analyse en Composantes Indépendantes (ICA) [24, 130], les descripteurs basés sur le gradient [190, 233, 26], et les motifs binaires locaux (Local Binary Pattern, LBP) [179, 1, 79] et ses variantes comme Local Ternary Pattern (LTP) [199], Compound Local Binary Pattern (CLBP) [4] et LBP on three orthogonal planes (LBP-TOP) [254]. Les détails sur LBP et ses variantes peuvent être trouvés dans les papiers [87, 192].

Dans ce qui suit, nous décrivons brièvement les méthodes d'extraction des caractéristiques mentionnées ci-dessus.

Les ondelettes de Gabor [115] sont obtenues en modulant une onde sinusoïdale 2D avec une enveloppe gaussienne. Les représentations basées sur les résultats des filtres de Gabor à de multiples échelles spatiales, orientations et localisations se sont avérées efficaces pour l'analyse d'images faciales, car elles peuvent être sensibles à des structures d'images plus fines comme celles correspondant aux rides. Elles sont également robustes au désalignement du visage. Cependant, le calcul des ondelettes de Gabor est coûteux car il implique la convolution d'images de visage avec un ensemble de filtres, et il peut également entraîner un grand nombre de caractéristiques redondantes. Des techniques telles que ACP [206] ou LDA [18] peuvent être appliquées pour réduire le nombre de caractéristiques.

Le détecteur de visages VJ [216] propose des caractéristiques pseudo-Haar en raison de leur simplicité de calcul pour l'extraction de caractéristiques. L'avantage principal des caractéristiques pseudo-Haar par rapport à la plupart des autres caractéristiques est sa vitesse du calcul [161]. Grâce à l'utilisation d'images intégrales, une caractéristique pseudo-Haar de n'importe quelle taille peut être calculée en temps constant [216]. En raison de ces avantages, les chercheurs l'ont également appliquée à l'analyse de l'expression faciale [240] et ont obtenu des résultats prometteurs.

Les descripteurs basés sur le gradient tels que Histogramme de Gradient Orienté (HOG) [41] comptent les occurrences d'orientations de gradient dans une partie localisée d'une image. De même, les descripteurs SIFT [131] sont calculés à partir du vecteur de gradient pour chaque pixel du voisinage afin de construire un histogramme normalisé des directions de gradient. Ces caractéristiques peuvent capturer des changements faciaux subtils et sont particulièrement robustes à l'éclairage et aux variations d'échelle. Les descripteurs LBP ont été initialement proposés pour l'analyse de texture [148], mais récemment ils ont été utilisés avec succès pour l'analyse d'expression faciale [254, 179]. Les propriétés les plus importantes des caractéristiques LBP sont leur tolérance aux changements d'éclairage et leur simplicité de calcul [148, 149]. Ces caractéristiques sont construites en assignant à chaque pixel un code dépendant des niveaux de gris des pixels de son voisinage. Un histogramme est ensuite calculé, où chaque bin correspond à l'un des motifs binaires. Les LBP sont généralement extraites de blocs d'image. Les LBP sont robustes aux changements d'illumination, computationnellement simple, et peuvent bien représenter les détails de la texture, même en présence de changements spatiaux [179]. Cependant, comparées aux descripteurs basés sur le gradient, elles sont moins robustes aux rotations d'images. Shan et al. [179] ont appliqué LBP pour représenter des expressions faciales. Ils ont montré que les caractéristiques LBP sont plus discriminantes et efficaces que celles de Gabor. A la suite de cette démonstration, de nombreux efforts ont tenté d'utiliser directement des variantes de LBP pour la REF. LTP [199] avec un niveau de discrimination supplémentaire et des codes ternaires ont été introduits

pour répondre aux limitations du LBP et pour lutter contre le bruit non uniforme. Ahmed et al. [5] proposent Gradient Local Ternary Patterns (GLTP) qui combinent les avantages de LTP et ceux de LBP en codant des valeurs d'amplitude de gradient plus robustes dans un schéma à trois niveaux pour obtenir des motifs de texture cohérents dans un bruit aléatoire et une illumination variable. Une version améliorée de GLTP avec un meilleur opérateur de gradient et une technique de réduction de dimensionnalité a été utilisée par Holder et al. [84] et a montré de meilleurs résultats. Des applications plus récentes des variantes de LBP dans la reconnaissance d'expression faciale sont rapportées dans [84, 196, 171, 11].

Les méthodes présentées ci-dessus sont des méthodes d'extraction de caractéristiques statiques qui consistent à coder image-par-image (appelées des méthodes spatiales). Dans le reste de cette section, nous allons présenter les méthodes spatio-temporelle qui codent les informations d'apparence d'un ensemble d'images consécutives plutôt que seulement celles d'une seule image.

Dans une méthode bien connue, Yacoob et Davis [234] ont appliqué le flot optique dense à la reconnaissance d'expression faciale ou d'actions faciales dans les années 1990. La procédure d'utilisation du flot optique dense consiste à calculer le mouvement dans des régions rectangulaires pour estimer l'activité de la région du visage.

Différentes extensions spatio-temporelles des caractéristiques basées sur l'image ont été conçues. Notamment, les LBP ont été étendus pour représenter les volumes spatio-temporels [254]. Pour simplifier l'approche, un volume spatio-temporel est décrit en calculant les caractéristiques LBP uniquement sur trois plans orthogonaux (TOP) : XY, XT et YT. Le descripteur LBP-TOP résulte de la concaténation de ces trois vecteurs caractéristiques. La même stratégie a ensuite été suivie pour étendre d'autres caractéristiques, telles que les caractéristiques LTP [146], Local Phase Quantization (LPQ) [94] et Local Gabor Binary Pattern (LGBP) [7]. LTP-TOP quantifie les différences d'intensité des pixels voisins et le pixel central en trois niveaux pour augmenter la robustesse vis-à-vis du bruit. LGBP-TOP utilise LBP pour coder les modèles de filtres de Gabor multi-échelle et multi-orientation pour améliorer les résultats de reconnaissance des émotions dans des conditions non restreintes. LPQ-TOP décrit l'information temporelle pour les actions faciales. Les représentations qui en résultent tendent à être plus efficaces, comme le montre l'amélioration significative des performances rapportée systématiquement dans [254, 146, 7, 94, 7]. Egalement motivé par LBP-TOP, Huang et al. [89] ont étendu Completed Local Quantized Pattern (CLQP) [88] au domaine spatio-temporel pour l'analyse de micro-expression dynamique, appelé SpatioTemporal Completed Local Quantized Pattern (STCLQP). Dans leur travail, STCLQP rend les motifs locaux plus compacts et discriminants en développant des codebooks basés sur le critère de Fisher.

Une stratégie alternative a été utilisée pour étendre les caractéristiques de type Haar afin de représenter les volumes spatio-temporels [238]. Dans ce cas, chaque caractéristique dynamique Haar code la variation temporelle dans une séquence d'images avec un modèle de valeurs binaires, où chaque valeur binaire est obtenue en seuillant la sortie de la caractéristique Haar dans l'image correspondante.

1.3.2.3 Caractéristiques hybrides

Il est connu que les caractéristiques basées sur la géométrie et l'apparence ont des avantages et limitations spécifiques respectifs. Par exemple, les caractéristiques géométriques sont efficace dans le calcul alors qu'elles sont sensibles au bruit ; en revanche, les caractéristiques basées sur l'apparence sont robustes au désalignement de l'image, mais cela prend beaucoup de temps de calcul. Par conséquent, les caractéristiques hybrides décrites par des caractéristiques géométriques et d'apparence sont devenues un sujet de recherche actif [201, 162, 33, 183, 97, 71].

Plusieurs chercheurs utilisent les caractéristiques hybrides pour améliorer la précision du décodage d'expression. Tian et al. [201] utilisent des caractéristiques hybrides pour analyser les expressions faciales, basées sur les traits permanents (i.e. les sourcils, les yeux et la bouche) et les traits transitoires du visage (i.e. l'approfondissement des sillons faciaux). Des modèles multi-états de visage et de composant facial sont proposés pour le suivi et la modélisation des traits faciaux. Dans [162], les caractéristiques d'apparence et les caractéristiques géométriques ont été ensuite extraites en utilisant Gauss–Laguerre (GL) et les coordonnées de 18 points faciaux détectés par AAM respectivement. En se basant sur ces points faciaux, 15 distances euclidiennes ont été calculées. A la fin, ces caractéristiques d'apparence et géométrique ont été combinées en un seul vecteur. Rapp et al. [164] ont proposé une combinaison originale de deux descripteurs hétérogènes. Le premier utilise LGBP afin d'exploiter la représentation multi-résolution et multi-directionnelle entre les pixels. Le deuxième descripteur est basé sur AAM qui fournit une information importante sur la position des points clés du visage. Chen et al. [33] ont proposé l'opérateur HOG-TOP et l'ont utilisé pour extraire les caractéristiques dynamique d'apparence, puis ils l'ont combinées avec des caractéristiques géométriques, en utilisant un noyau multiple pour trouver la combinaison optimale de ces descripteurs hétérogènes. Dans [71], les caractéristiques d'apparence sont extraites en utilisant LBP et en se basant sur une nouvelle définition des régions locales. Ensuite, les caractéristiques de forme géométriques de ces régions locales sont calculées à l'aide de descripteurs de moments centrés normalisés (NCM). Enfin, les caractéristiques géométriques et celles d'apparence sont concaténées en un seul vecteur de caractéristiques.

Il existe plusieurs comparaisons dans la littérature entre les différents types de caractéristiques. La recherche montre que les caractéristiques basées sur l'apparence peuvent obtenir de meilleurs résultats que les caractéristiques géométriques [252, 15]. De plus, les caractéristiques hybrides surpassent les caractéristiques d'apparence et les caractéristiques géométriques [158, 14].

Les caractéristiques présentées dans les sections précédentes 1.3.2.1 et 1.3.2.2 sont extraites manuellement (en anglais : handcrafted features). Des caractéristiques sont extraites au travers d'un apprentissage automatique, notamment l'apprentissage profond (en anglais : learned features). Ces caractéristiques seront décrites dans la section suivante 1.3.2.4.

1.3.2.4 Caractéristiques basées sur le deep learning

L'apprentissage profond ou deep learning est un paradigme qui permet d'apprendre des représentations hiérarchiques multicouches à partir de données d'apprentissage [197]. La représentation globale contient généralement au moins deux couches de bas niveau. La première couche convolue l'image d'entrée avec un certain nombre de filtres locaux appris à partir des données, et la seconde couche agrège la sortie de convolution par des opérations telles que le pooling [167]. Des couches de niveau supérieur peuvent être conçues à diverses fins telles que la lutte contre les occlusions partielles [197]. Les filtres dans les couches de bas niveau sont généralement des filtres lisses qui calculent la différence locale, donc ils sont robustes contre l'éclairage et les erreurs d'enregistrement dans une certaine mesure. Les opérations de regroupement (par ex. max-pooling [167]) améliorent la robustesse aux erreurs d'enregistrement.

Kim et al. [108] ont exploré plusieurs architectures Convolutional Neural Networks (CNN) et méthodes de prétraitement pour l'analyse des expresions faciales. Jung et al. [97] ont utilisé un réseau profond basé sur deux modèles différents. Le premier réseau profond, qui est basé sur CNN, extrait des caractéristiques d'apparence temporelles à partir de séquences d'images, tandis que le deuxième réseau profond, qui est basé sur Deep Neural Networks (DNN) entièrement connecté, extrait des caractéristiques géométriques temporelles à partir de points temporels faciaux. Ces deux modèles sont combinés en utilisant une méthode d'intégration afin de booster les performances de la reconnaissance d'expression faciale. Dans [52], Egede et al. codent des informations de forme et d'apparence dans les deux types d'extraction de caractéristiques handcrafted et deep learning. Pour les caractéristiques handcrafted, HOG et un nombre de métriques ont été extraits à partir de 49 points faciaux pour représenter, respectivement, les caractéristiques d'apparence et les caractéristiques géométriques. Puis, en se basant sur CNN, des caractéristiques sont apprises à partir d'une combinaison des pixels de l'image d'origine (apparence) et de masques binaires (forme

de visage). Les auteurs de [129] ont proposé une approche appelée Boosted Deep Belief Network (BDBN). Leur approche effectue les trois étapes d'apprentissage (apprentissage des caractéristiques, sélection des caractéristiques et construction du classifieur) de manière itérative dans un système unifié. Liu et al. [126] ont construit une architecture profonde en utilisant des noyaux convolutionnels pour apprendre les variations d'apparence locales provoquées par les expressions faciales et extraire des caractéristiques basées sur Deep Belief Network (DBN). De même, Zhao et al. [256] se sont basés sur DBN pour extraire les caractéristiques. Khorrami et al. [107] ont montré empiriquement que les caractéristiques apprises par les CNNs correspondent complètement au FACS développé par Ekman et Friesen [58]. Jaiswal et Valstar [93] ont intégré des réseaux de neurones de mémoire à long terme bidirectionnels (bi-directional long-term memory neural networks) avec le CNN pour extraire des caractéristiques temporelles.

1.3.3 Apprentissage automatique

La dernière étape d'un système automatique d'analyse d'expression est la reconnaissance de l'expression faciale en fonction des caractéristiques extraites. Certains systèmes classent directement les expressions tandis que d'autres classent les expressions en reconnaissant d'abord des unités d'action (AUs) particulières (voir la section 1.2.3 pour la description de FACS et AUs). De nombreux classifieurs ont été appliqués à la reconnaissance d'expression tels que :

- réseaux de neurone (Neural Networks, NN),
- machines à vecteurs de support (Support Vector Machine, SVM),
- analyse Discriminante Linéaire (Linear discriminant analysis, LDA),
- K-plus proche voisin (K Nearest Neighbor, KNN),
- régression logistique multinomiale (Multinomial Regression Logistic, MRL),
- modèles de Markov cachés (Hidden Markov Model, HMM),
- réseaux bayésiens (Bayesian Network, BN), et d'autres.

Ici, nous résumons les méthodes de reconnaissance d'expression à des méthodes basées sur des images statiques et sur des séquences vidéo. La méthode de reconnaissance basée sur des données statiques utilise uniquement l'image courante avec ou sans image de référence (il s'agit principalement d'une image de visage neutre) pour reconnaître l'expression d'une seule image. La méthode de reconnaissance basée sur des données dynamiques utilise les informations temporelles des séquences pour reconnaître les expressions d'une ou plusieurs images.

La reconnaissance d'expression basée sur des images statiques. La reconnaissance d'expression basée sur des images n'utilise pas d'information temporelle pour les images d'entrée. Elle utilise seulement les informations de l'image d'entrée courante. L'image d'entrée peut être une image statique ou une image d'une séquence traitée indépendamment des autres image de la séquence. Plusieurs méthodes peuvent être trouvées dans la littérature pour la reconnaissance d'expressions faciales telles que les NNs [29, 201, 203, 106, 205], les SVMs [179, 26, 79, 71], KNN [187, 162, 111, 224] et le BN [36].

Tian et al. [203] ont utilisé un système de reconnaissance basé sur un NN pour reconnaître les AUs. Ils ont utilisé des NNs à trois couches avec une couche cachée pour reconnaître les AUs par une méthode de rétropropagation standard [168]. Des réseaux séparés sont utilisés pour les parties supérieure et inférieure du visage. Les entrées peuvent être des caractéristiques géométriques, d'apparence ou les deux. Les sorties sont les AUs reconnus. Le réseau est entrainé pour répondre aux AUs désignées, qu'elles se produisent seules ou en combinaison. Lorsque les AUs sont combinées, plusieurs nœuds de sortie sont excités. Yang et al. [239] emploient RankBoost avec la régularisation l1 pour la reconnaissance d'expression. Ils évaluent également l'intensité des expressions en utilisant les scores de classement en sortie. Kotsia et al. [110] ont fusionné les scores des caractéristiques de la forme et des caractéristiques de texture en utilisant un NN à fonction de base radiale (RBF). Cohen et al. [36] ont observé que bien que les données étiquetées sont disponibles en petites quantités, il existe un énorme volume de données non étiquetées disponibles. Ils ont donc utilisé des classifieurs de BNs comme Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN) et Stochastic Structure Search (SSS) pour l'apprentissage semi-supervisé avec un certain nombre de données étiquetées et de grandes quantités de données non étiquetées.

Comme le SVM s'avère très puissant pour les tâches de classification, il est considéré comme la méthode de pointe et est utilisé dans presque tous les systèmes les plus récents/révisés pour la reconnaissance d'expression [179, 26, 79, 71, 194]. Rapp et al. [164] ont combiné deux descripteurs hétérogènes en utilisant un SVM à plusieurs noyaux (multiple Kernel, MKL) pour atteindre la classification des émotions. Le travail de Zhang et al. [250] a présenté un nouveau framework pour le problème MKL en développant l'algorithme HessianMKL en SVM multi-classes avec une règle un-contre-un. Ce cadre a également été utilisé pour reconnaître sept expressions faciales en combinant trois fonctions du noyau et deux représentations d'image. Dans [194], Sun et al. ont entrainé, pour chaque caractéristique, les classifieurs SVM et Partial Least Squares (PLS) individuellement qui ont des capacités discriminantes différentes pour la classification des expressions faciales. Ils ont ensuite proposé un réseau de fusion pour exploiter ces caractéristiques. Ainsi, ils ont noté que certaines caractéristiques sont plus performantes lorsqu'elles sont classées par PLS. Par conséquent,

un réseau de fusion combinant PLS et SVM ensemble peut obtenir de meilleurs résultats que l'utilisation d'un SVM seul.

La faiblesse commune à toutes les méthodes de classification basées sur l'image est qu'elles ignorent la dynamique des expressions faciales ou des AUs cibles. Bien que certaines méthodes basées sur des images (par ex. [95]) utilisent des caractéristiques extraites de plusieurs images pour coder la dynamique des expressions faciales, les modèles d'apprentissage automatique pour la classification dynamique fournissent une méthode plus raisonnée pour ce faire.

La reconnaissance d'expression basée sur des séquences vidéos. La plupart des approches dynamiques de classification des expressions faciales sont basées sur les variantes des réseaux bayésiens dynamiques (Dynamic Bayesian Network, DyBN). Les DyBN sont des modèles probabilistes graphiques qui codent les dépendances entre des ensembles de variables aléatoires évoluant dans le temps, capables de représenter des relations probabilistes entre différentes expressions faciales, et de modéliser la dynamique de leur développement [178]. Les modèles les plus couramment utilisés pour la classification des séquences, Hidden Markov Models (HMM) [163] et Conditional Random Fields (CRF) [113], sont des versions génératives et discriminantes, respectivement, des DyBN avec une structure de graphe linéaire.

Le DyBN est une extension de la méthode d'inférence bayésienne à un réseau de graphes, où les nœuds représentent des modalités différentes et les arêtes désignent leurs dépendances probabilistes. Le DyBN est appelé par différents noms dans la littérature tels que les modèles probabilistes génératifs, les modèles graphiques, etc. L'avantage de ce réseau par rapport aux autres méthodes est que la dynamique temporelle des données multimodales peut facilement être intégrée. La forme la plus populaire de DyBN est le HMM. Diverses approches basées sur les HMMs ont été proposées pour la classification dynamique des expressions faciales [153, 124, 150, 35, 241, 180, 105, 212, 183, 100]. Par exemple, [153, 124, 150, 241] ont entrainé des HMMs indépendants en utilisant des séquences d'images de chaque catégorie d'émotions, puis ont effectué une catégorisation des émotions en comparant les probabilités d'observation des HMMs spécifiques à une expression. Pour mieux tenir compte de la variabilité des sujets, Otsuka et al.[153] ont modélisé la probabilité d'observation des états cachés dans les HMMs en utilisant des mélanges gaussiens. De plus, [241] a proposé une approche fondée sur les HMMs en deux étapes pour la classification des expressions correspondant aux six émotions de base. Premièrement, une banque de classifieurs linéaires a été appliquée au niveau des images, et la sortie a été fusionnée pour produire une signature temporelle pour chaque observation. Deuxièmement, des HMMs discrets ont été utilisés pour apprendre les signatures temporelles pour chaque catégorie d'expression. Pour modéliser les AUs, Olivier et al. [124]

ont utilisé des HMMs pour modéliser des séquences d'images de chaque AU indépendamment des autres. Valstar et Pantic [212] ont utilisé des HMMs pour effectuer un lissage temporel des sorties de SVM spécifiques aux émotions/AU, entraînées par image. La principale critique de ces approches est qu'elles ne sont pas totalement discriminantes, car elles effectuent la modélisation des catégories d'expression faciale (et AU) indépendamment les unes des autres. Plus récemment, dans [100], un HMM correspondant à chaque classe d'expression est entrainé en utilisant les données de l'ensemble d'apprentissage, puis l'expression avec la plus grande probabilité est identifiée pour prédire la classe d'expression à laquelle une vidéo appartient. Contrairement aux méthodes précédentes qui apprennent un HMM pour chaque classe, Sikka et al. [183] ont entrainé un modèle HMM pour chaque exemple. Pour ce faire, ils ont proposé d'utiliser des modèles HMMs entièrement bayésiens qui utilisent des probabilités antérieurs pour apprendre avec de petites quantités de données (par vidéo). Ils ont calculé ensuite les distances entre ces modèles exemplaires en utilisant un noyau probabiliste qui mesure efficacement la même chose entre les composants statiques et dynamiques des HMMs individuels. Puis, ces distances ont été utilisées pour apprendre un classifieur SVM pour chaque classe.

Des modèles discriminatoires basés sur les CRF ont également été proposés [214, 92, 30, 219, 2]. Dans [214], les auteurs ont entrainé un CRF à chaîne linéaire par AU, et chaque image a été associée à un nœud dans le graphe. L'état d'un tel noeud est une variable binaire indiquant si l'AU est présente ou non dans l'image courante. La classification d'AU est effectuée par image en seuillant la probabilité d'état pour chaque image dans la séquence de test. Acevedo et al. [2] ont aussi utilisé un CRF à chaîne linéaire pour modéliser les dépendances séquentielles entre les images d'une vidéo. Hidden Conditional Random Fields (HCRFs) est une variante de CRF qui a été appliquée avec succès pour la reconnaissance de gestes. Elle consiste à étiqueter toute la séquence comme un tout [223]. Récemment, Walecki et al. [219] ont proposé une variante de HCRF pour modéliser la dynamique cachée des expressions faciales séquentielles et sélectionner automatiquement le modèle optimal qui peut mieux discriminer entre les différentes expressions faciales.

Très récemment, les approches basées sur le deep learning sont devenues de plus en plus dominantes dans le domaine de la vision par ordinateur. Par conséquent, plusieurs travaux basés sur CNN et Deep Belief Networks (DBN) ont été proposés pour la classification dynamique des expressions faciales [129, 207, 81, 78]. Par exemple, une approche semi-supervisée pour la reconnaissance d'expression faciale à partir d'une vidéo utilisant un CNN spatio-temporel est proposée dans [78]. Dans [207], un DBN avec une machine de Boltzmann restreinte (RBM) [83] a été utilisé pour concevoir un système de reconnaissance d'expression faciale. Hasani et al. [81] ont proposé un réseau spatio-temporel en deux parties

qui utilise DNN et CRF pour reconnaître des expressions faciales dans une séquence d'images. Le réseau basé sur DNN contient trois modules, Inception-ResNet [198] et deux couches entièrement connectées qui capturent les relations spatiales de l'expression faciale dans les images. Le module CRF capture la relation temporelle entre les images.

Les lecteurs intéressés peuvent trouver plus de détails sur les trois modules d'analyse des expressions faciales dans les documents [173] et [136].

1.4 Conclusion

Dans ce chapitre, nous avons d'abord évoqué le contexte de l'analyse faciale, puis présenté un aperçu général du développement dans ce domaine, et nous avons décrit brièvement quelques techniques de pointe proposées dans la littérature. Dans le chapitre suivant, nous allons nous intéresser à la description de certaines bases de données de visages pour la reconnaissance d'expression utilisées pour l'évaluation des méthodes de la littérature. Nous donnons ensuite une comparaison synthétique des systèmes existants de reconnaissance d'expression faciale.

Chapitre 2

Systèmes d'analyse des expressions faciales et Bases d'images

Sommaire

2.1 Introduction

Ce chapitre fournit d'abord la motivation pour l'analyse et la détection automatique des émotions, en donnant des exemples de divers domaines d'application. Ensuite, il présente une brève description des bases de données d'expression faciale, utilisées tout au long de la dissertation, et qui ont joué un rôle majeur dans l'évaluation de nos approches proposées. Ceci est suivi d'un aperçu global des différents défis majeurs rencontrés dans le cadre de la Reconnaissance des Expressions Faciales (REF). Le chapitre se termine par une analyse comparative des travaux existants sur la REF pour pouvoir donner une base et un contexte au travail présenté dans cette thèse.

2.2 Applications possibles

La reconnaissance et la classification de l'expression faciale peuvent être applicables dans divers aspects de notre vie quotidienne ; ci après une liste non exhaustive des domaines d'application susceptibles de bénéficier de la capacité de détection des expressions faciales émotionnelles.

Interaction homme-machine : Les expressions faciales constituent un moyen de communication comme de nombreuses autres manières (par ex. le signal vocal). La détection des émotions est naturelle pour les humains mais c'est une tâche très difficile pour les machines. Comme ces dernières s'impliquent de plus en plus dans la vie quotidienne des hommes et participent à la fois à ses espaces de vie et de travail, elles doivent évoluer pour devenir plus intelligentes en termes de compréhension des humeurs et des émotions de l'être humain. En effet, l'intégration d'un système capable de reconnaitre les émotions dans ces machines pourrait les rendre plus humains et par conséquent améliorer la communication homme-machine [247, 147, 237].

Interaction homme-robot : Pour les robots sociaux, il est également important qu'ils puissent reconnaître différentes expressions et agir en conséquence afin d'avoir des interactions efficaces [176, 176, 166]. Pour atteindre une telle interaction homme-robot, il est primordial que le robot comprenne les expressions faciales des humains. C'est pourquoi la mise en œuvre de nouveaux algorithmes pour la reconnaissance des émotions permettrait le développement de robots capables de comprendre le comportement humain dans des environnements intérieurs et extérieurs (par ex. conception de robots compagnons, robots guides touristiques, etc.).

Neuromarketing : La mesure automatisée des préférences des consommateurs à partir de leurs expressions faciales en réponse aux publicités des produits aurait un impact profond sur l'analyse des études de marché, car cela permettrait aux entreprises de mieux connaître le consommateur et proposer de nouveaux produits et services à leur clients. McDuff et al. [139]) ont réussi à déterminer si les gens aimaient certaines publicités en analysant leur comportement facial.

Surveillance visuelle et sécurité : Le décodage des micro-expressions est crucial pour établir ou nuire à la crédibilité, et pour déterminer toute tromperie de suspects lors des interrogatoires. En effet, la micro-expression est une expression faciale involontaire momentanée que les gens affichent inconsciemment lorsqu'ils cachent une émotion. Ainsi, les systèmes d'analyse des expressions faciales permettraient l'évaluation et la détection du stress, l'ennui, l'inattention et le micro-sommeil dans des situations où une attention ferme est essentielle (par ex. la surveillance du conducteur [75]).

E-éducation, jeux vidéo et chat : Les expressions faciales des apprenants informent l'enseignant de la nécessité d'ajuster le message pédagogique envers les apprenants. Par conséquent, l'intégration d'un système de REF dans une plateforme pédagogique faciliterait le développement de systèmes de tutorat intelligents en permettant une évaluation automatique du niveau d'intérêt des apprenants, leur compréhension et leur plaisir d'apprendre en ligne [237, 147]. Dès lors, les systèmes de tutorat automatisés, capables de reconnaître les états émotionnels et cognitifs des apprenants, pourraient adapter le niveau de difficulté d'un cours au degré de confusion ou perplexité apparues sur le visage de l'apprenant. Ainsi, des jeux vidéo pourraient adapter leur niveau de difficulté en fonction des informations provenant des expressions faciales du joueur [144]. Dans [117], nous avons proposé d'intégrer un système de REF dans des plateformes pédagogiques afin d'enrichir la communication textuelle par une certaine perception de l'état émotionnel des individus, dans un contexte d'apprentissage collaboratif.

Dans une application de chat réalisée par Anderson et McOwen [10], les utilisateurs peuvent se connecter et commencer à chatter, au même moment un système de reconnaissance d'expression faciale est connecté à cette application de chat afin d'insérer automatiquement des émoticônes basées sur les expressions faciales de l'utilisateur.

Médecine : L'expression faciale est le moyen direct d'identifier quand des processus mentaux spécifiques (par ex. la douleur, la dépression) se produisent. Par conséquent, les systèmes de REF peuvent également être utilisés pour surveiller les patients dans les hôpitaux lorsque le personnel médical n'est pas disponible ou surchargé. Il pourrait également être utilisé dans des scénarios d'assistance à domicile pour surveiller les patients et informer le personnel médical en cas d'urgence. En effet, il existe des déve-

loppements prometteurs de l'informatique affective dans les applications médicales. Parmi ces développements, nous pouvons citer l'exemple de la détection automatique de la douleur, comme proposé dans [12, 99, 52]. Il a été démontré qu'il était possible de dériver une mesure de la douleur et de faire la distinction entre différents types de douleur à partir des expressions faciales d'un patient [226]. Un autre développement prometteur est celui de la détection automatique de la dépression à partir des signaux faciaux et auditifs [37].

Psychologie : La détection d'expressions est extrêmement utile pour l'analyse de la psychologie humaine. Ainsi, la reconnaissance de l'incapacité d'une personne à exprimer certaines expressions faciales peut aider à diagnostiquer les troubles psychologiques précoces.

2.3 Bases de données d'expressions faciales

L'un des aspects les plus importants du développement de tout nouveau système de reconnaissance ou de détection d'expression faciale est le choix de la base de données qui sera utilisée pour tester ce nouveau système. De plus, des bases de données communes sont nécessaires pour évaluer les algorithmes de manière comparative. Dans cette section, nous allons présenter quelques bases de données d'expressions faciales populaires qui sont publiquement et librement disponibles. D'autres bases de données sont disponibles et leur couverture n'est pas faite ici. Ainsi, nous nous intéressons qu'aux bases de données qui ont été utilisées pour évaluer les travaux de cette dissertation.

Les bases de données disponibles peuvent être classées en deux catégories, comme il sera mentionné dans la section suivante 2.4 : les bases d'expressions faciales spontanées et les bases d'expressions faciales posées.

2.3.1 Cohn-Kanade (CK) et son extension Cohn-Kanade (CK+)

La base de données CK [101] est très populaire et a été largement utilisée par la communauté de reconnaissance d'expressions faciales. La base CK contient 97 étudiants universitaires âgés de 18 à 30 ans. Ces sujets sont constitués de 69% de femmes, 31% d'hommes, 81% d'euro-Américains, 13% d'afro-américains et 6% d'asiatiques et de latinos. Les sujets ont été instruits par un expérimentateur pour effectuer une série de 23 expressions comprenant des unités d'action unique (par ex. AU12, i.e. coins de lèvre tirés obliquement) et des combinaisons d'unités d'action (par ex. AU1 + AU2, i.e. sourcils intérieurs et extérieurs levés). Avant d'effectuer chaque expression, un expérimentateur a décrit et modélisé l'expression souhai-

tée. Alors, six de ces expressions étaient basées sur des descriptions d'émotions basiques (joie, surprise, colère, peur, dégoût et tristesse). Les séquences d'images ainsi présentées commencent par l'expression neutre et se terminent par le pic de l'expression demandée. Par ailleurs, la dernière image de la séquence est toujours codée par des experts. Ces images ont été numérisées en résolution de 640×490 pixels avec une précision de 8 bits pour les niveaux de gris. A noter, l'orientation de la caméra est frontale et les petits mouvements de la tête sont présents.

Plus tard, la base de données CK a été étendue à la base de données Cohn-Kanade étendue (CK+) [132]. L'ensemble de données a été augmenté de 107 séquences et 26 sujets. Ceci donne au total 593 séquences de 123 sujets. Parmi ces 593 séquences vidéo, seulement 327 séquences (118 sujets) ont des étiquettes d'émotions validées et classées en sept expressions faciales de base (joie, tristesse, surprise, colère, peur, dégoût et mépris). Les séquences d'images sont organisées de la même manière que les séquences de la base CK, chaque vidéo commence par un visage neutre, puis progressivement se développe dans l'une des sept expressions faciales. La figure 2.1 présente des sujets appartenant aux deux versions de la base CK simulant les six émotions.

FIGURE 2.1 Exemples d'images extraites de la base CK/CK+. De gauche à droite : Neutralité, Joie, Tristesse, Surprise, Colère, Peur, Dégoût.

2.3.2 Karolinska Directed Emotional Faces (KDEF)

La base de données KDEF [134] a été initialement développée pour la recherche en neurosciences. Cependant, elle a depuis été utilisée dans le domaine de la vision par ordinateur en raison de son applicabilité. La base KDEF contient 4900 images prises à partir de 70 personnes (35 hommes et 35 femmes), leur âge allant de 20 à 30 ans. Chaque individu affiche 7 expressions (colère, dégoût, peur, joie, neutralité, tristesse, surprise), qui sont capturées deux fois à partir de 5 angles différents (-90, -45, 0, +45, +90 degrés) et enregistrées au format JPEG avec une résolution de 562×762 pixels (Figure 2.2).

Les auteurs de cette base ont gardé un œil sur plusieurs questions importantes lors de l'élaboration de la base de données. L'environnement (éclairage, arrière-plan, distance de la caméra) a été maintenu constant tout au long du processus de capture et les sujets ont été invités à retirer tous les accessoires (chapeaux, lunettes, poils du visage et maquillage).

FIGURE 2.2 Exemples d'images extraites de la base KDEF. De gauche à droite : Neutralité, Joie, Tristesse, Surprise, Colère, Peur, Dégoût. De haut en bas : 0°, 45°,$-45°$,90°,$-90°$.

2.3.3 Japanese Female Facial Expression (JAFFE)

La base de données JAFFE [135] comprend 213 images d'expressions faciales de dix femmes japonaises. Ces dernières ont simulé 3 à 4 exemples pour chacune des six émotions de base, ainsi que l'émotion neutre. La résolution des images est de 256×256 pixels. La figure 2.3 montre quelques exemples d'images de la base de données JAFFE. En outre, des images statiques ont été capturées dans un environnement contrôlé. Les évaluations sémantiques (vérité de terrain) des expressions ont été effectuées à partir d'expériences psychologiques par 60 autres femmes japonaises. Selon le créateur de la base, Lyons [135], une expression n'est jamais une pure expression mais un mélange d'émotions différentes. Ainsi, une échelle de 5 niveaux a été utilisée pour évaluer chacune des images d'expression (5 pour le haut niveau et 1 pour le bas niveau). Deux de ces évaluations ont été données, l'une avec des images d'expression de peur et l'autre sans images d'expression de peur. Les images d'expression sont étiquetées selon l'expression prédominante dans chacune des images.

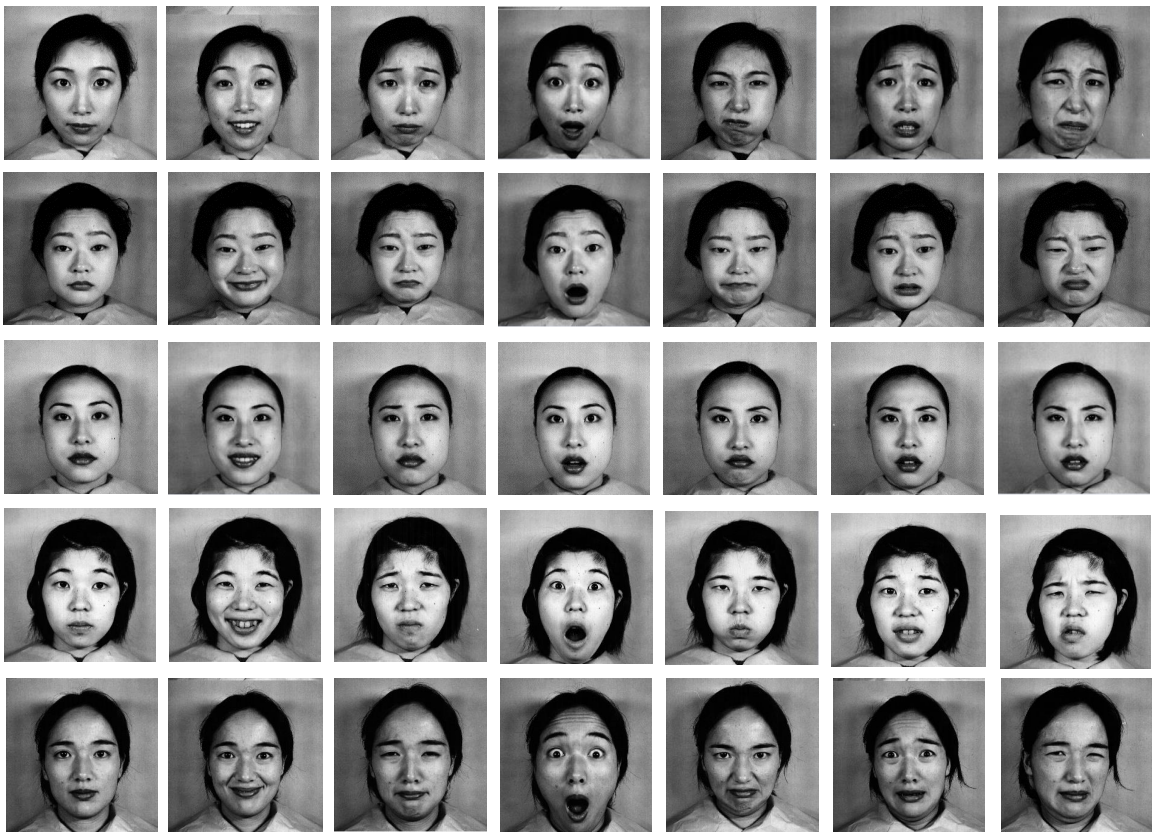


FIGURE 2.3 Exemples d'images extraites de la base JAFFE. De gauche à droite : Neutralité, Joie, Tristesse, Surprise, Colère, Peur, Dégoût.

2.3.4 Oulu-CASIA

La base de données Oulu-CASIA [253] est composée de 80 sujets, âgés de 23 à 58 ans, avec six émotions de base (colère, dégoût, peur, joie, tristesse et surprise). 50 sujets proviennent de l'université d'Oulu et les 30 autres de CASIA, dont 73.8% sont des hommes. Chaque image a une résolution de 320×240 pixels. Les images ont été prises dans trois conditions d'éclairage différentes : normale, faible et sombre. L'éclairage normal signifie que les séquences d'images ont été prises dans des bonnes conditions d'éclairage. L'éclairage faible signifie que seul l'écran de l'ordinateur était allumé et que le sujet était assis devant l'ordinateur lors de l'enregistrement de l'expression faciale dynamique. L'éclairage sombre signifie qu'aucune (ou presque) lumière n'était présente (proche de l'obscurité). Les séquences vidéo contiennent des images allant de la phase neutre à la phase apex des expressions faciales. Le nombre de séquences vidéo est de 480 pour chacune des conditions d'éclairage. Des exemples d'images des six expressions de la base de données Oulu-CASIA sont illustrés dans la figure 2.4.

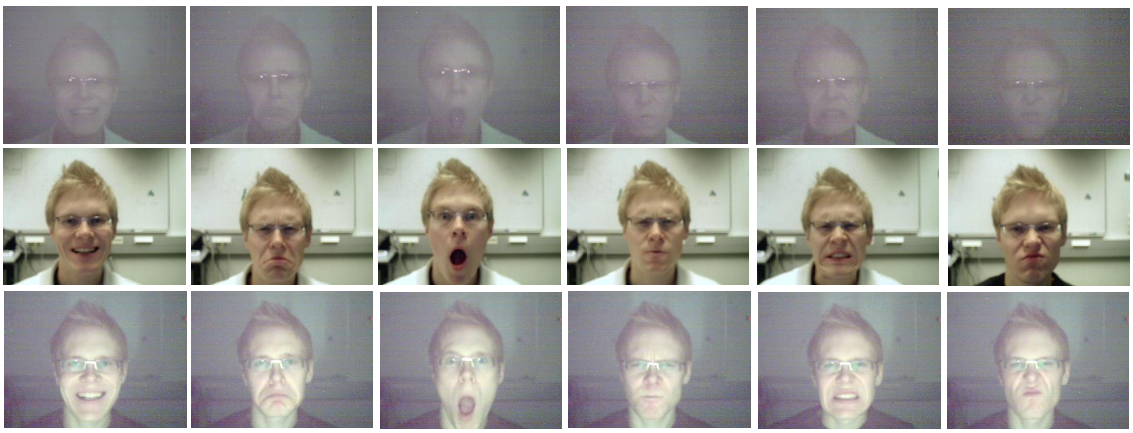


FIGURE 2.4 Exemples d'images extraites de la base Oulu-CASIA. De gauche à droite : Joie, Tristesse, Surprise, Colère, Peur, Dégoût. De haut en bas : condition d'éclairage sombre, normal, faible.

2.3.5 Facial Expressions and Emotion Database (FEED)

La base de données d'expressions faciales FEED [220] a été créée dans le cadre du projet de l'Union européenne FG-NET (Face and Gesture Recognition Research Network). Elle comprend 320 vidéos de 18 sujets exprimant les 6 émotions basiques (joie, surprise, peur, colère, dégoût, tristesse) définies par Ekman et Friesen [57], ainsi que l'expression neutre.

L'ensemble des données présente des expressions naturelles (ou spontanées), qui ont été suscitées en montrant aux sujets plusieurs stimulus, sous forme de vidéos, soigneusement sélectionnés. Chaque sujet visualise trois vidéos pour chacune des émotions. Les réactions des sujets sont alors enregistrées et labellisées suivant la vidéo stimulante. Par exemple, pour une vidéo censée stimuler la joie, les expressions faciales du sujet sont enregistrées et labellisées comme des expressions de joie. Ces expressions sont considérées comme spontanées. Ceci est différent des bases de données comme CK/CK+ [101, 132], où les sujets ont été invités à effectuer des mouvements faciaux spécifiques. D'ailleurs, les sujets, appartenant à la base d'images CK/CK+, montrent des séries d'expressions exagérées avec des changements d'intensité progressifs. Alors que les expressions spontanées affichées dans la base d'images FEED sont dépourvues de toute exagération et sont très lisses en termes de changements d'intensité (voir Figure 2.5).

FIGURE 2.5 Exemples d'images extraites de la base FEED. De gauche à droite : Neutralité, Joie, Tristesse, Surprise, Colère, Peur, Dégoût.

2.3.6 Static Facial Expressions in the Wild (SFEW)

La base de données SFEW [47] contient des captures d'écran extraites des films (voir les exemples de la figure 2.6). Cette base est différente des bases de données d'expression faciale présentées précédemment et qui sont générées dans des environnements de laboratoire hautement contrôlés. Cette base de données décrit les conditions du monde réel ou simulées du monde réel pour la reconnaissance d'expression, en supposant que les films fournissent

des environnements «proches des environnements réels». La base de données est divisée en trois ensembles. Chaque ensemble contient sept sous-dossiers correspondant à sept catégories d'expression (colère, dégoût, peur, neutralité, joie, tristesse et surprise). Ces ensembles ont été créés de manière strictement indépendante de la personne, de sorte qu'il n'y ait pas de chevauchement entre les personnes appartenant à l'ensemble d'apprentissage et celles appartenant à l'ensemble de test. La base SFEW comprend 891, 427 et 372 images couleur pour les ensembles d'apprentissage, validation et test, respectivement.

FIGURE 2.6 Exemples d'images extraites de la base SFEW. De gauche à droite : Neutralité, Joie, Tristesse, Surprise, Colère, Peur, Dégoût.

2.4 Expression spontanée vs délibérée

Les expressions faciales peuvent être classées en deux catégories : spontanée/naturelle ou posée/délibérée. Par exemple, les sourires posés impliquent souvent seulement un mouvement de la bouche (les muscles zygomatiques) alors que les sourires spontanés sont plus symétriques et peuvent également être caractérisés par une action additionnelle d'autres muscles faciaux, en particulier les muscles orbiculaires de l'œil qui produit des rides au coin extérieur de l'œil (appelées les rides de la patte d'oie) [61, 69]. De même, il a été montré que les différences entre les unités d'actions sourcilières spontanées et délibérées (AU1, AU2, AU4) résident dans la durée, la rapidité de déclenchement, le décalage et le timing des actions [213]. Les expressions posées sont généralement enregistrées dans des environnements plus

contraints en demandant aux sujets de simuler l'expression de l'état affectif cible. Le contenu sémantique et la réalisation physique des expressions faciales spontanées et posées diffèrent considérablement [62, 55]. Une étude psychologique [62] indique que les expressions posées peuvent différer en apparence et en timing par rapport aux expressions spontanées. La même chose a été prouvée par [16]. Alors, bien que la majorité des systèmes d'analyse des expressions faciales aient été développés pour classer les expressions posées, principalement en raison de la disponibilité des données, leur performance devrait diminuer considérablement lorsqu'ils sont appliqués aux expressions spontanées. Cela a également été souligné par les spécialistes de la cognition et de l'informatique. Ils considèrent que la principale critique des travaux existants est que les méthodes conçues en utilisant des expressions posées ne sont pas applicables dans des situations réelles, où apparaissent des changements subtils dans les expressions faciales plutôt que des changements exagérés caractérisant les expressions posées [157]. De plus, les effets de la pose de la tête (des mouvements des têtes des sujets durant l'enregistrement) et les changements d'éclairage sont beaucoup plus prononcés pour les expressions faciales spontanées [204].

La plupart des bases de données d'expressions faciales ont été collectées en demandant aux sujets d'effectuer une série d'expressions (par ex. CK/CK+, JAFFE, KDEF et Oulu-CASIA décrites dans la section 2.3). Récemment, les efforts se sont portés sur l'analyse automatique des expressions spontanées. FEED est un exemple de base de données d'expressions faciales spontanées (voir Section 2.3).

2.5 Expression Faciale statique vs dynamique

La REF basée sur l'image n'utilise pas d'information temporelle pour les images d'entrée. En effet, elle utilise seulement des informations de l'image d'entrée courante. Cette dernière peut être une image statique ou une image d'une séquence traitée indépendamment. La REF basée sur une séquence d'images utilise des informations temporelles extraites de la séquence pour reconnaître une expression à partir de plusieurs images. C'est-à-dire, il ne s'agit plus de décrire une expression à partir d'une image mais à partir d'une séquence d'images. Ces dernières années, la REF basée sur l'analyse de séquences d'images (REF dynamique) est devenue un sujet de recherche actif. Les caractéristiques dynamiques des séquences d'images peuvent fournir beaucoup plus d'informations (apparence + mouvement) que les images statiques qui ne mettent à disposition que des informations relatives à l'apparence [251].

Les manifestations faciales de l'émotion sont un phénomène très dynamique qui évolue dans le temps depuis l'onset (le début de l'émotion), l'apex (le pic de l'émotion) et l'offset (la fin d'émotion : retour à l'état neutre) [22]. Les différences entre les expressions faciales sont

souvent transmises plus puissamment par des transitions dynamiques entre différentes étapes d'expressions plutôt que par un seul état représenté par une image statique [178]. Alors que la REF basée sur l'image est basée sur la configuration statique du visage à partir d'images fixes, la REF basée sur la séquence modélise les comportements temporels des expressions faciales à partir de séquences d'images. A cet effet, des expériences psychologiques [17] suggèrent que la dynamique des expressions faciales est cruciale pour une interprétation réussie des expressions faciales. Ceci est particulièrement vrai pour les expressions faciales naturelles sans aucune pose délibérée [9]. De plus, des études d'électromyographie indiquent que les expressions dynamiques ont tendance à susciter des réponses plus intenses du mimique facial et sont liées à une activation physiologique plus élevée [170]. Ces résultats soutiennent l'hypothèse selon laquelle les expressions faciales dynamiques sont plus valables sur le plan écologique et donc plus appropriées à la recherche sur les émotions [8].

L'étude comportementale dans [9] indique que le mouvement facial améliore la reconnaissance, tandis que d'autres ne trouvent pas de différences entre les conditions statiques et dynamiques [103]. Dans ce contexte, la question qui se pose est de savoir si les expressions faciales dynamiques produisent des résultats différents de ceux des expressions statiques. Si oui, quel serait l'impact sur la recherche en reconnaissance émotionnelle ?

2.6 Reconnaissance des expressions faciales à vues multiples

Les systèmes de REF à vues multiples (Multi-view Facial Expression Recognition, MVFER) étendent les approches de reconnaissance d'expression de visage frontal afin de traiter des images ou des séquences vidéo de visage expressif sous différents angles de vue. De ce fait, la reconnaissance automatique de l'expression faciale à partir de vues non frontales est un sujet de recherche difficile qui a récemment commencé à attirer l'attention de la communauté scientifique.

La plupart des approches existantes fonctionnent sur des vues frontales ou quasi-frontales alors que dans les applications du monde réel, une vue frontale est une hypothèse irréaliste (les images de vue frontale ne sont pas toujours disponibles) et limite l'applicabilité. Pour cette raison, l'analyse non frontale est maintenant l'un des défis actifs liés à la REF, qui nécessite non seulement une approche de reconnaissance efficace, mais aussi une méthode pour compenser les informations manquantes. C'est un problème difficile puisque certaines des caractéristiques faciales, nécessaires pour la reconnaissance, ne sont pas ou pas complètement disponibles en raison de l'orientation du visage. Par exemple, les sourcils, qui sont très

importants pour reconnaître l'expression faciale, peuvent ne pas être visibles sur une face non frontale.

Le problème de MVFER n'a pas été beaucoup abordé dans la littérature. Les raisons de cette situation sont multiples. Comparée à la REF à partir d'une vue frontale ou quasi-frontale, la MVFER est beaucoup plus difficile en raison des vastes variations intra-classes introduites par les différentes poses faciales. Plus important encore, il y a un manque de création de base de données d'expression faciale multi-vue. Sans la disponibilité de telles bases, la recherche sur la MVFER a été sérieusement freinée. Contrairement à l'abondance des bases de données disponibles pour l'analyse des expressions faciales à partir de vues frontales ou quasi-frontales, les bases de données conçues pour l'analyse des expressions faciales à partir de vues multiples pourraient se compter sur les doigts de la main, selon nos connaissances. Ci-dessous une brève description des quatre bases de données disponibles dans la littérature :

1. La base de données **Binghamton University 3D facial expression (BU-3DFE)** [242] est la base de données la plus couramment utilisée pour la reconnaissance d'expressions faciales multi-vues. Elle a été conçue pour échantillonner les comportements faciaux 3D avec différents états émotionnels prototypiques. Les sujets de la base de données sont d'âges différents, allant de 18 à 70 ans, et d'une grande variété d'ethnies. 56% des sujets sont des femmes et 44% sont des hommes. Pendant la session d'enregistrement, chaque sujet a effectué six expressions faciales universelles, à savoir : la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Des modèles de texture du sujet ont été capturés. Pour une description détaillée de cette base de données, les lecteurs peuvent se référer à [242].
2. La base de données **Carnegie Mellon University Multi- Pose, Illumination et Expression (CMU Multi-PIE)** [74] se compose de plus de 750 000 images de 337 personnes enregistrées dans jusqu'à quatre sessions sur une période de cinq mois. Des photographies de sujets ont été prises à partir de 15 points de vue et de moins de 19 conditions d'illumination, tout en affichant une gamme d'expressions faciales (neutre, souriante, strabisme, surprise, dégoûtée et hurlante). Les données démographiques de cette base de données comprenaient 69.7% d'hommes, une distribution raciale consistant en 60% d'européens, 35% d'asiatiques et 5% d'autres.
3. **Radboud Faces Database (RaFD)** [114] contient des images faciales de 67 sujets de différents âges (adultes et enfants), sexes (hommes et femmes) et races (caucasiennes et marocaines). Il y a 120 images pour un sujet, où les images ont été prises, simultanément, à partir de cinq angles de caméra. Dans ces images, les sujet ont été formés pour montrer leurs émotions dans trois différentes directions de regard. Les émotions

enregistrées sont la colère, le dégoût, la peur, la joie, la tristesse, la surprise, le mépris et l'état neutre.

4. La base de données **KDEF** (voir la section 2.3.2).

2.7 Protocoles d'expérimentation

Pour évaluer la fiabilité et l'efficacité des algorithmes de REF et puis les comparer, ils doivent être soumis à un protocole expérimental bien défini. Ce dernier consiste en deux points importants : le jeu de données et la validation croisée. Ci-dessous, nous allons discuter la procédure relative à chacun d'entre eux.

2.7.1 Jeu de données (dataset)

Presque toutes les bases de données de la littérature (Section 2.3) nécessitent une sélection d'un sous-ensemble d'images disponibles avant d'être utilisées. Par exemple, la base de données CK+ contient 593 séquences et parmi elles seulement 327 séquences sont étiquetées et peuvent ainsi être utilisées pour évaluer un système de reconnaissance d'émotion. En revanche, chaque séquence se compose d'environ (30 frames) et juste la dernière image est codée par un expert en utilisant FACS (Facial Action Coding System) [58]. Par conséquent, cela permet de sélectionner seulement les trois dernières images ou juste la dernière image pour représenter une émotion.

Dans la plupart de temps, la façon dont cette sélection est effectuée n'est pas bien indiquée et les comparaisons rapportées sont toujours biaisées par ce manque d'information. Dans les chapitres 3, 4 et 5, nous allons aborder puis tester les différentes sélections aux fins d'une comparaison équitable.

2.7.2 Validation croisée

La validation croisée (Cross-Validation, CV) est une méthode statistique d'évaluation et de comparaison des algorithmes d'apprentissage en divisant l'ensemble de données en deux parties : La première est utilisée pour entrainer le modèle (l'ensemble d'apprentissage) et la seconde pour le valider (l'ensemble de test). Dans une CV, les ensembles d'apprentissage et de validation doivent être croisés par des cycles successifs de sorte que chaque point de données ait une chance d'être validé. La forme de base de la CV est la k-fold CV. D'autres formes de CV sont des cas spéciaux de k-fold CV qui impliquent des cycles répétés de k-fold CV. Parmi les CV les plus connues, nous trouvons k-fold (k-fold CV) et leave-one-out (LOOCV).

K-fold : Dans la CV de k-fold, les données sont d'abord partitionnées en k échantillons de taille égale (ou presque). Par la suite, k itérations d'entraînement et de validation sont effectuées de sorte qu'à chaque itération, un échantillon différent des données est mis en attente de validation tandis que les échantillons k-1 restants sont utilisés pour l'apprentissage.

Leave-One-Out La CV de Leave-one-out est un cas particulier de k-fold CV où k est égal au nombre d'instances dans les données. En d'autres termes, à chaque itération, presque toutes les données, sauf une observation unique, sont utilisées pour l'apprentissage, et le modèle est ainsi testé sur cette seule observation. A savoir, la LOOCV est coûteuse en temps de calcul.

La performance de chaque modèle sur chaque échantillon peut être calculée en utilisant une métrique de performance prédéterminée comme la précision (accuracy), le rappel (recall) et la F-mesure (Fscore) ou représentée au moyen de la matrice de confusion. Pour plus de détails sur les métriques de performance, le lecteur peut se référer à [188].

2.8 Systèmes d'analyse des expressions faciales

La reconnaissance automatique de l'expression faciale a attiré beaucoup d'attention de la part des scientifiques du comportement depuis le travail de Darwin [44]. Sown [189] a mené la première tentative d'analyse automatique des expressions faciales à partir de séquences d'images en 1978. Ainsi, au cours de la dernière décennie, beaucoup de progrès ont été réalisés. Nous résumons dans la table 2.1 quelques travaux antérieurs afin de mettre notre travail en contexte. Cette table est lu de la manière suivante :

Colonne 1

Système

Colonne 2

Année

Colonne 3, 4 et 5

Les trois modules d'un système de REF :

- **Module 1** : Enregistrement du visage
- **Module 2** : Extraction des caractéristiques
- **Module 3** : Classification

Colonne 6

Le type d'information à traiter :

- Vidéo/MV ou image/MV (**MV** quand il s'agit d'une vue multiple)

Colonne 7

Les sorties reconnues par le système (Output Recognition, **OR**), à savoir :

- Des AUs et leurs combinaisons
- Emotions :
  - **Ne** (Neutralité, Neutrality)
  - **An** (Colère, Anger)
  - **Ha** (Joie, Happiness)
  - **Sa** (Tristesse, Sadness)
  - **Su** (Surprise, Surprise)
  - **Fe** (Peur, Fear)
  - **Di** (Dégout, Disgust)
  - **Co** (Mépris, Contempt)
  - **Sm** (Sourire, Smile)
  - **Sq** (Strabisme, Squint)
  - **Sc** (Hurlement, Scream)

Colonne 8

Les bases de données (Databases, **DB**) utilisées pour l'évaluation

Colonne 9

le type de la **CV** (validation croisée)

Colonne 10

Le taux de reconnaissance (Recognition Rate, **RR**)

TABLE 2.1 Tableau comparatif des différents systèmes de REF de la littérature.

1	2	3	4	5	6	7	8	9	10
Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[252]	1998	Point spécifiques	Hybride (Gabor et 34 points fiduciaire)	NN	Images	Ne, Ha, Sa, Su, Fe, An, Di	JAFFE	10-fold	90.10
Suite à la page suivante									

TABLE 2.1 – suite de la page précédente

Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[203]	2002	ROIs	Hybride (Ondelettes de Gabor et Géométrique)	NN	Vidéos	9 AUs	CK	1-fold	92.7
[158]	2006	Point spécifiques	Géométrique	Temporal rules	Vidéos /Profil	27 AUs et leurs combinaisons	MMI	——	87
[254]	2007	Visage	Apparence (LBP-TOP)	SVM	Vidéos	Ha, Sa, Su, Fe, An, Di	CK	10-fold	96.26
[179]	2009	Visage	Apparence (Boosted LBP)	SVM	Images	Ne, Ha, Sa, Su, Fe, An, Di	CK JAFFE MMI	10-fold	91.4 81 86.9
[116]	2009	Visage	Géometrique	SVM	Images	Ne, Ha, Sa, Su, Fe, An, Di	JAFFE	3-fold	68.5
[240]	2010	ROIs	Apparence (Pseudo-Haar)	Adaboost	Images Vidéos	Ha, Sa, Su, Fe, An, Di	CK	1-fold	92.3 80
[143]	2011	Visage	Apparence (LGBP)	SVM	Images /MV	Ne, Ha, Sa, Su, Fe, An, Di Di, Ne, Sc, Sm, Sq, Su	BU-3DFE Multi-PIE	10-fold	67.96 80.6
[244]	2011	ROIs	Hybride (EOH et géométriques)	NN		Ne, Ha, Sa, Su, Fe, An, Di	CK	1-fold	93.5

Suite à la page suivante

TABLE 2.1 – suite de la page précédente

Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[200]	2012	Visage	Apparence (SIFT)	SVM	Images /MV	Di, Ne, Sc, Sm, Sq, Su Ha, Sa, Su, Fe, An, Di	Multi-PIE BU-3DFE	5-fold	81.7 76.1
[162]	2012	Visage	Hybride (GL et géométriques)	KNN	Images	Ha, Sa, Su, Fe, An, Di	CK MMI JAFFE	10-fold	86.1 80.16 91.12
[257]	2014	Visage	Apparence (LBP)	GSRRR	Images /MV	Ha, Sa, Su, Fe, An, Di Di, Ne, Sc, Sm, Sq, Su	BU-3DFE Multi-PIE	10-fold	66 81.7
[32]	2014	ROIs	Apparence (HOG)	SVM	Images	Ha, Sa, Su, Fe, An, Di, Co Ne, Ha, Sa, Su, Fe, An, Di	CK+ JAFFE	LOSO LOSO	88.7 94.3
[50]	2014	ROIs	Apparence (HOG)	SVM	Images	Ha, Sa, Su, Fe, An, Di	CK+	—-	95
[129]	2014	Visage	BDBN	BDBN	Images	Ha, Sa, Su, Fe, An, Di	CK+	8-Fold	96.7
[126]	2015	Visage	DBN	SVM	Images	Ne, Ha, Sa, Su, Fe, An, Di	CK+ MMI	10-fold	93.7 75.85
[33]	2015	Visage	Hybride (HOG-TOP et géométrique)	MKL-SVM	Vidéos	Ne, Ha, Sa, Su, Fe, An, Di	CK+	10-fold	93.6

Suite à la page suivante

TABLE 2.1 – suite de la page précédente

Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[26]	2015	Visage	Apparence (HOG)	SVM	Images	Ne, Ha, Sa, Su, Fe, An, Di	CK+	10-fold	94.1
						Ne, Ha, Sa, Su, Fe, An, Di, Co	RaFD		92.9
[183]	2015	Visage	Apparence (distribution gaussienne multivariée)	Exemplar-HMM	Vidéos	Ha, Sa, Su, Fe, An, Di, Co	CK+	LOSO	94.6
						Ha, Sa, Su, Fe, An, Di	OULU-CASIA FEED	10-fold	75.62
								LOSO	54.14
[225]	2015	ROIs	Apparence (LBP-TOP)	SVM	Vidéos	Ha, Sa, Su, Fe, An, Di	CK+	LOSO	87.74
[80]	2015	ROIs	Apparence (PHOG et LBP)	SVM	Images	Ha, Sa, Su, Fe, An, Di	CK+	5-fold	94.63
							JAFFE		83.86
[79]	2015	ROIs	Apparence (LBP)	SVM	Images	Ha, Sa, Su, Fe, An, Di	CK+	10-fold	94.39
							JAFFE		92.22
[97]	2015		CNN et DNN	CNN et DNN	Vidéos	Ha, Sa, Su, Fe, An, Di, Co	CK+	10-fold	97.25
[194]	2016	Visage	Apparence (SIFT) et CNN	SVM et PLS	Images	Ne, Ha, Sa, Su, Fe, An, Di	SFEW	1-fold	56.32

Suite à la page suivante

TABLE 2.1 – suite de la page précédente

Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[82]	2016	ROIs	Apparence (Fonctions de Gabor)	Clustering et Logique Floue	Images	Ne, Ha, Sa, Su, Fe, An, Di	KDEF	—-	98.8
[172]	2016	Visage	LSiBP	SVM	Images /MV	Ne, Ha, Sa, Su, Fe, An, Di	KDEF SFEW	10-fold 1-fold	80.5 29.7
[249]	2016	Point fiduciaires	Apparence (SIFT)	DNN		Ha, Sa, Su, Fe, An, Di Di, Ne, Sc, Sm, Sq, Su	BU-3DFE Multi-PIE	1-fold	80.1 85.2
[89]	2016	Visage	Apparence (STCLQP)	SVM	Vidéos	Di, Su, Répression, micro-expressions tendues	CASME [235]	LOSO	57.31
[71]	2017	ROIs	Hybride (LBP et NCM)	SVM	Images	Ha, Sa, Su, Fe, An, Di Ne, Ha, Sa, Su, Fe, An, Di	CK+	5-fold	97.25 91.95
[84]	2017	ROIs	Apparence (Improved GLTP)	SVM	Images	Ha, Sa, Su, Fe, An, Di Ne, Ha, Sa, Su, Fe, An, Di	CK+	LOSO	86.5 83.1

Suite à la page suivante

TABLE 2.1 – suite de la page précédente

Système	Année	Module 1	Module 2	Module 3	Images ou Vidéo	OR	DB	CV	RR (%)
[171]	2017	Visage	Apparence (LDTP)	SVM	Images	Ne, Ha, Sa, Su, Fe, An, Di Ha, Sa, Su, Fe, An, Di An, Fe, Ha, Sa, Relief	CK+ JAFFE BU-3DFE GEMEP-FERA [211]	LOSO	94.2 67.61 72.7 71.3
[100]	2017	Visage	Géométrique	HMM	Vidéos	Ha, Sa, Su, Fe, An, Di	CK+	10-fold	82.4
[6]	2018	ROIs	Apparence (LMP)	SVM	Vidéos	Ha, Sa, Su, Fe, An, Di, Co Ha, Sa, Su, Fe, An, Di	CK+ Oulu-CASIA	10-fold 10-fold	97.25 84.58
[112]	2018	Visage	CNN	CNN	Vidéos	Ha, Sa, Su, Fe, An, Di, Co Ha, Sa, Su, Fe, An, Di	CK+ Oulu-CASIA	10-fold 10-fold	98.47 91.67

2.9 Conclusion

Dans ce chapitre, nous avons d'abord présenté les différents domaines d'application de l'analyse automatique des expressions faciales, puis nous avons évoqué les divers défis qui en découlent et qui attirent l'attention de la communauté de la vision par ordinateur. A la fin, nous avons parcouru les différents travaux sur la REF pour pouvoir disposer d'une vision claire à la fois sur les techniques, les protocoles expérimentaux utilisés, et sur les performances obtenues jusqu'à présent.

A partir du chapitre suivant, nous présenterons notre approche et le système proposé pour la REF.

Chapitre 3

Analyse de forme et de texture des régions faciales pour la reconnaissance d'expressions

Sommaire

3.1 Introduction

Une première étape cruciale dans un système d'analyse des expressions faciales est l'enregistrement du visage. Cette étape nécessite une meilleure détection de certains points caractéristiques du visage, permettant ainsi une meilleur définition des composants du visage (i.e. les sourcils, les yeux, le nez et la bouche). En outre, la définition des régions représentant les composantes faciales nous permet d'analyser la forme et la texture de chaque composante indépendamment de la forme et l'expression du visage. Nous avons émis, à partir de différents exemples (voir Figure 3.4b), l'hypothèse que les caractéristiques de forme et de texture des différentes régions faciales peuvent avoir de fortes capacités discriminantes. En effet, nous pouvons observer que différentes régions faciales sont susceptibles d'avoir des changements de texture constituant des formes très discriminatives et distinctives lorsque nous montrons des expressions différentes. Dans cette perspective, nous allons utiliser le descripteur HOG pour extraire l'information de forme et le descripteur LBP et ses variantes pour extraire l'information de texture.

Dans ce chapitre, nous allons définir une nouvelle décomposition faciale qui sera ensuite comparée avec différentes décompositions de la littérature. Ensuite, une expérimentation sera menée, plus largement, en utilisant les descripteurs HOG, LBP et ses variantes, et leur combinaison. Nous allons examiner deux différentes méthodes d'apprentissage automatique : SVM et forêts aléatoires (Random Forest, RF) pour la reconnaissance d'expressions faciales indépendamment de la personne sur plusieurs bases de données publiques. Au travers les différentes expériences, nous allons montrer que la meilleure performance de reconnaissance est obtenue en utilisant le classifieur SVM avec des caractéristiques hybrides LTP+HOG extraites à partir de notre décomposition faciale proposée.

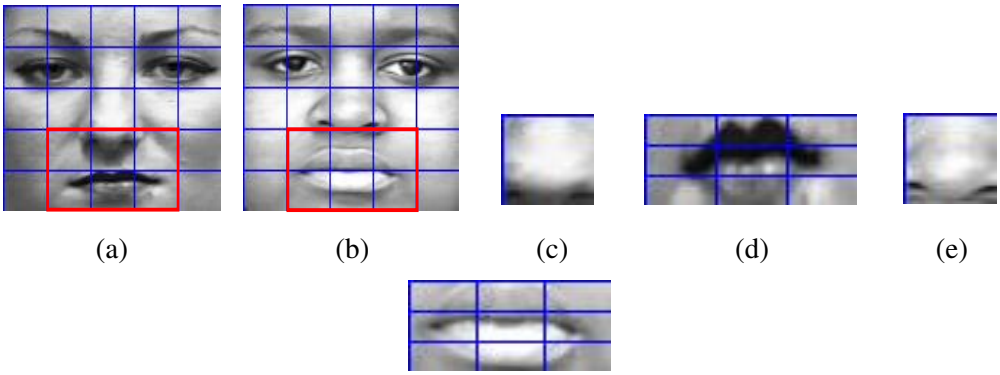
3.2 Aperçu de la méthodologie proposée

L'objectif de la méthodologie proposée est de caractériser les six expressions faciales universelles (Joie, Peur, Dégoût, Surprise, Tristesse, Colère) et l'état neutre, en analysant des régions spécifiques du visage bien définies à partir desquelles des caractéristiques faciales sont extraites. L'étude proposée pour la reconnaissance d'expressions faciales (REF) est basée sur deux étapes principales : (i) la détermination de régions spécifiques et plus précises du visage en utilisant la méthode Supervised Descent Method (SDM) [231], permettant la

détection de points faciaux ; (ii) l'évaluation de différents descripteurs de texture et de forme (la texture à travers LBP et ses variantes, et la forme à travers HOG) et leur combinaison. Pour expliquer l'intérêt de la décomposition faciale en différentes ROI, représentant les principales composantes du visage (yeux, sourcils, nez, bouche, entre sourcils), la figure 3.1 montre que lorsque le visage entier est utilisé comme une seule ROI, les composantes du visage ne sont pas situées dans la même région. En effet, la région rouge de la figure 3.1a contient la bouche et le nez, alors que la même région rouge de la figure 3.1b ne contient que la bouche. Cela est dû à la différence de forme de visage d'une personne à l'autre. Comme indiqué dans [72], la méthode holistique, qui utilise l'ensemble du visage comme une seule ROI, ne permet pas un meilleur enregistrement du visage en raison des différentes formes et tailles des composants faciaux dans la population et selon l'expression. Par conséquent, une composante faciale peut ne pas se trouver dans la même région. D'où l'intérêt de la décomposition faciale en différentes ROIs, représentant les principales composantes du visage (œil gauche, œil droit, sourcil gauche, sourcil droit, nez, bouche, entre sourcils). Ce problème affecte les performances des systèmes de REF, comme nous le montrerons dans la section 3.8.2. Lorsque la décomposition faciale est appliquée, on peut extraire les composants du visage et les analyser globalement pour effectuer la reconnaissance d'expression. Les figures (3.1c et 3.1d) et les figures (3.1e et 3.1f) illustrent des ROIs, représentant le nez et la bouche, extraites respectivement des visages des figures 3.1a et 3.1b. La figure 3.2 illustre les différentes étapes du système proposé : (a) le détecteur Viola-Jones (VJ) est utilisé pour détecter la position du visage dans l'image ; (b) les points faciaux sont détectés en utilisant SDM ; (c) les composantes faciales (ROIs) sont extraites en utilisant des coordonnées de points faciaux spécifiques ; (d) les ROIs définis (sourcil gauche, sourcil droit, entre les sourcils, l'œil gauche, l'œil droit, le nez et la bouche) sont découpées ; (e) chaque ROI est redimensionnée et partitionnée en blocs (cette étape permet d'extraire plus d'informations locales) ; (f) pour chaque bloc, le descripteur de caractéristique est extrait. Les descripteurs de tous les blocs sont alors concaténés pour construire le descripteur du ROI. Enfin, les descripteurs de toutes les ROIs sont concaténés pour obtenir le descripteur du visage ; (g) le descripteur obtenu est introduit dans un classifieur (SVM, RF) multiclasse pour accomplir la tâche de reconnaissance.

3.3 Détection du visage

La détection de visages est un domaine évolué en vision par ordinateur avec un certain nombre de détecteurs de visage disponibles dans diverses bibliothèques. Le détecteur de visage VJ [217] est incontestablement le plus populaire à ce jour. Il est plus rapide que



(a) (b) (c) (d) (e)

(f)

FIGURE 3.1 (a) Le visage entier en tant que ROI, la région rouge incorpore deux composants faciaux qui sont la bouche et une partie du nez. (b) Le visage entier en tant que ROI, la région rouge incorpore seulement un composant facial qui est la bouche. (c) et (d) les ROIs nez et bouche extraites de (a). (e) et (f) les ROIs nez et bouche extraites de (b).

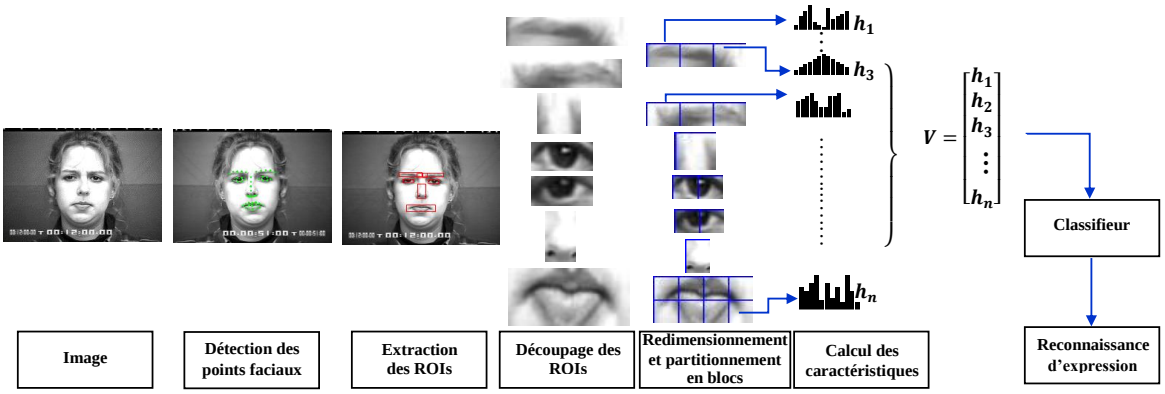


FIGURE 3.2 Système automatique de REF proposé.

la plupart des autres approches et ses implémentations sont facilement disponibles dans la plupart des bibliothèques de vision par ordinateur. La majorité des implémentations disponibles du détecteur VJ sont destinées aux faces frontales, mais des détecteurs de profils existent également (par ex. [151]). Cependant, la performance de ces détecteurs de profil est généralement faible. Ceci est probablement dû à un manque de données d'entraînement disponibles, ou la détection de visages de profil étant une tâche intrinsèquement plus difficile.

Dans ce travail, l'implémentation du détecteur VJ disponible dans la bibliothèque OpenCV 2.4.8 est utilisée [23].

3.4 Détection des points caractéristiques

La méthode SDM [231] s'est révélée efficace pour la localisation des points faciaux dans des environnements incontrôlés. Elle a ainsi obtenu de bonnes performances en temps réel. SDM peut être appliquée en utilisant deux modes : interactif ou automatique. En mode interactif, l'utilisateur est invité à créer un rectangle pour localiser le visage souhaité. Pour obtenir de bonnes performances, les limites supérieures et inférieures doivent dépasser les sourcils et les lèvres. En mode automatique, les visages sont détectés par le détecteur VJ d'OpenCV. Il faut noter que le détecteur VJ fournit une liste de boîtes englobantes de visages détectés dans une image, si plusieurs visages sont détectés le plus grand est choisi. Par la suite, la zone de délimitation détectée est utilisée pour initialiser les paramètres de forme nécessaires pour l'ajustement du SDM [231]. La forme initiale pourrait être simplement la forme moyenne initialisée dans le cadre de la délimitation renvoyée par VJ. Dans ce travail, VJ n'a pas pu détecter la majorité des visages de la base SFEW (voir Section 2.3.6). De même pour la base KDEF dans le cas où l'angle de vue est différent de 0 (voir Section 2.3.2). C'est pourquoi, nous nous sommes penchés sur le détecteur du visage Cascade Deformable Part Models (CDPM) [151] qui est destiné à la détection des visages à vue multiple. Malgré l'utilisation de ce détecteur CDPM, la plupart des visages reste non détectable. Dans l'objectif d'éviter ce problème, la méthode SDM a été appliquée en utilisant ses deux modes. Le mode automatique est utilisé dans toutes les expérimentations effectuées tout au long de cette thèse sauf pour les bases SFEW (voir Section 3.8.3.2) et KDEF (voir Section 4.3.1.3) dans le cas multi-view où le mode interactif a été appliqué.

SDM peut détecter 49 points caractéristiques autour des sourcils, des yeux, du nez et de la bouche (voir Figure 3.3). Les 49 points faciaux détectés par la méthode SDM sont représentés sur la figure 3.3, où chaque point a un numéro d'étiquette. Par conséquent, la forme du visage est décrite par $X_* = (X_1, X_2, \cdots, X_p)$, où X_i est le i^{eme} point, et p indique le nombre de points faciaux (ici $p = 49$). $X_i = (x_i, y_i)$, où x_i et y_i sont respectivement les

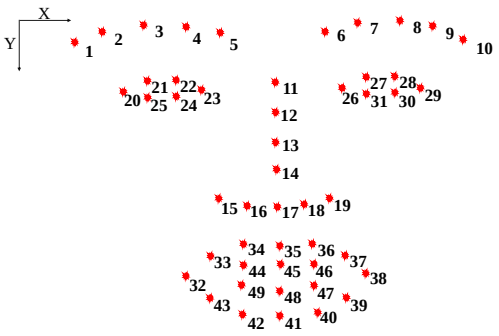


FIGURE 3.3 49 points faciaux détectés par SDM.

coordonnées horizontal et vertical du i^{eme} point. Dans ce travail, nous utilisons certains de ces points (voir Table 3.1) pour définir la décomposition du visage en ROIs. Ces dernières se trouvent principalement autour des sourcils, des yeux, du nez et de la bouche (voir Figure 3.3).

3.5 Extraction des ROIs

Comme indiqué dans [72], l'utilisation du visage entier comme une seule région pour extraire les caractéristiques affecte les performances des systèmes de REF. De plus, pour différentes personnes, la taille et la forme des organes faciaux ne sont pas les mêmes, ainsi, il ne peut être assuré que la même position faciale est toujours présente dans un bloc particulier dans toutes les images. Pour résoudre ce problème, notre approche propose d'extraire des régions spécifiques du visage, représentant les principales composantes du visage (yeux, sourcils, nez, bouche, entre sourcils), à partir desquelles les caractéristiques seront extraites. L'objectif est d'augmenter la performance de REF à travers des descripteurs de caractéristiques. Dans la section 3.8.2, nous allons montrer à travers des expériences approfondies la performance de la décomposition faciale proposée, en comparaison avec l'utilisation du visage entier en tant que ROI unique et les décompositions faciales de l'état de l'art.

Pour atteindre l'étape d'extraction des ROIs, nous commençons par détecter les points faciaux en utilisant la méthode SDM. Après les avoir détectés, sept ROIs, susceptibles de changer avec l'expression du visage, sont extraites (sourcil gauche, sourcil droit, œil gauche, œil droit, entre les sourcils, le nez et la bouche). La figure 3.4b montre un exemple d'extraction de ROIs. Ces ROIs sont les régions les plus représentatives de l'expression faciale selon le Facial Action Coding System (FACS) [59]. L'idée derrière le système de REF proposé est d'analyser le visage localement en mettant l'accent sur les caractéristiques

permanentes et transitoires. Les caractéristiques permanentes sont les yeux, les sourcils, le nez et la bouche. Leurs formes et leurs textures sont exposées à changer avec l'expression faciale, qui produit différentes rides et sillons appelés traits transitoires du visage (par ex. les rides verticales entre les sourcils en raison de leur convergence les unes vers les autres, surtout quand le visage exprime la tristesse et la colère selon FACS). Cela nous amène à choisir méticuleusement le point de départ, la largeur et la longueur de chaque ROI (voir Table 3.1) afin de capturer plus exactement les caractéristiques permanentes et transitoires (voir Figure 3.4b). Dans notre travail, la REF est spécialement basée sur des informations de texture. En outre, nous sommes intéressés à analyser des zones spécifiques du visage qui sont susceptibles de présenter des changements, relatifs aux informations de texture, avec l'expression faciale. A cet effet, il est plus utile de définir la largeur de la région de la bouche à la distance horizontale entre les points x_{25} et x_{30}, au lieu de choisir les points x_{32} et x_{38} qui délimitent, horizontalement, seulement la bouche. Ce choix permet de détecter les changements d'informations de texture dans la zone entre les points x_{25} et x_{32} et la zone entre les points x_{30} et x_{38}, surtout quand le visage exprime la joie, tel que rapporté par FACS. En effet, l'émotion de joie s'exprime sur le visage en étirant les lèvres des coins. Pour la hauteur, nous avons choisi de le mettre à la différence entre le maximum des points y_{32}, y_{38} et y_{41} et le minimum des points y_{32} , y_{38} et y_{35}, au lieu d'utiliser simplement la différence entre les points y_{41} et y_{35} qui ne délimitent verticalement que la bouche. En effet, quand certaines personnes sourient, les coins X_{32} et X_{38} de leur bouche se déplacent au dessus du point X_{35}. De plus, toujours basé sur FACS, la tristesse s'exprime en abaissant les coins extérieurs des lèvres, ce qui justifie le choix du maximum des points y_{32}, y_{38} et y_{41}. Notre but en utilisant le minimum des points y_{32}, y_{38} et y_{35}, et le maximum des points y_{32}, y_{38} et y_{41}, consiste à considérer respectivement le point le plus haut et le point plus bas afin d'extraire la région d'intérêt qui apporte l'information complète nécessaire pour une REF précise et fiable. Pour la région des sourcils, nous définissons la largeur de la région du sourcil gauche à la distance horizontale entre les points x_1 et x_{11}, au lieu de choisir les points x_1 et x_5 qui délimitent, horizontalement, seulement le sourcil gauche. Ce choix permet de détecter les changements d'informations de texture dans la zone située entre les points x_5 et x_{11}, surtout lorsque le visage exprime la peur, la colère et la tristesse comme le rapporte FACS. La même stratégie est adoptée pour le sourcil droit.

Les régions bouche et sourcils contiennent les composantes faciales (i.e. la bouche, sourcil gauche, sourcil droit) ainsi que d'autres petites zones autour d'elles. Ces zones sont affectées par la déformation de la bouche et des sourcils lorsque le visage exprime une émotion. En revanche, pour les yeux, nous n'avons pas besoin d'analyser une zone supplémentaire car les informations requises sont à l'intérieur des yeux. Par exemple, lorsque le visage exprime

TABLE 3.1 Extraction de composants faciaux en utilisant des points faciaux détectés avec la méthode SDM.

		Dimensions de ROI		
		Point de départ	Largeur	Longueur
Composantes faciales	Sourcil gauche	(x_1, y_4)	$x_{11} - x_1$	$max(x_4 - x_3, y_3 - y_1)$
	Sourcil droit	(x_{11}, y_7)	$x_{11} - x_{10}$	$max(x_7 - x_6, y_7 - y_{10})$
	Entre les sourcils	$(x_5, min(y_5, y_6))$	$x_6 - x_5$	$y_{13} - y_{12}$
	Œil gauche	(x_{20}, y_{22})	$x_{23} - x_{20}$	$y_{24} - y_{22}$
	Œil droit	(x_{26}, y_{27})	$x_{29} - x_{26}$	$y_{31} - y_{27}$
	Nez	(x_{15}, y_{12})	$x_{19} - x_{15}$	$y_{12} - y_{17}$
	Bouche	$(x_{25}, min(y_{32}, y_{35}, y_{38}))$	$x_{30} - x_{25}$	$max(y_{32}, y_{38}, y_{41}) - min(y_{32}, y_{35}, y_{38})$

la peur, les paupières supérieures se soulèvent et les paupières inférieures se resserrent, ce qui a pour effet d'afficher un Sanpaku (blanc des yeux visible) supérieur. D'où notre choix d'utiliser les points qui délimitent seulement les yeux.

3.6 Extraction des caractéristiques

Une fois que les ROIs sont extraites d'une image faciale, l'étape suivante consiste à appliquer une procédure de redimensionnement en modifiant la taille des ROIs. Cette étape est cruciale car chaque composante faciale (représenté par une ROI) peut avoir des tailles différentes en fonction de la forme du visage dans l'image. L'idée derrière le redimensionnement est d'obtenir des ROIs d'une même composante avec une même taille. Pour ce faire, nous appliquons différentes tailles de ROI, puis sélectionnons celle qui offre de meilleures performances de REF, comme nous le montrerons dans la section 3.8.2. Après la procédure de redimensionnement, nous partitionnons chaque ROI en blocs réguliers dans lesquels les caractéristiques seront extraites. L'étape de partitionnement est intéressante car elle permet d'extraire des informations locales. Une fois que l'extraction de caractéristiques est réalisée dans chaque bloc d'une ROI, les caractéristiques de tous les blocs sont concaténées pour construire le descripteur de caractéristiques de la ROI. Enfin, les descripteurs de caractéristiques de toutes les ROIs sont concaténés pour créer le descripteur des caractéristiques faciales. Dans la suite, nous présenterons les différents descripteurs de caractéristiques que nous avons utilisés dans cette étude.

3.6.1 Local Binary Pattern (LBP)

L'opérateur LBP a été initialement proposé par Ojala et al. [148] afin de caractériser la texture d'une image. Il consiste à comparer les pixels voisins au pixel central afin d'obtenir

(a)

(b)

FIGURE 3.4 (a) Exemples de 49 points caractéristiques détectés par SDM. (b) Exemples d'extraction de ROIs.

un motif binaire (binary pattern). Ce motif binaire est généré comme suit : tous les voisins prennent la valeur "1" si leur valeur est supérieure ou égal au pixel central ou "0" sinon. Ensuite, les pixels de ce code binaire sont multipliés par les poids correspondants et sommés afin d'obtenir le code LBP du pixel central (Figure 3.5). Formellement, l'opérateur LBP est exprimé comme suit :

$$LBP_{P,R}(x_c, y_c) = \sum_{p=1}^{P} s(g_p - g_c)2^{p-1} \tag{3.1}$$

avec,

$$s(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{3.2}$$

où x_c et y_c sont les coordonnées du pixel courant (pixel central), P est le nombre de pixels du voisinage, R est le rayon du voisinage, g_c est le niveau de gris du pixel central, et g_p est le niveau de gris du pixel p.

Une extension de l'opérateur LBP d'origine est appelée LBP uniforme (LBP_u). LBP uniforme est défini comme un motif avec exactement 0 ou 2 transitions 0-1 ou 1-0 (par ex. 10000001 ou 00011000 est un motif uniforme, mais pas 00011001). Ojala et al [148] ont trouvé que seulement 58 des 256 modèles de LBP sont uniformes. Par conséquent, la taille de l'histogramme LBP peut être considérablement réduite à 59, au lieu de 256. Chacun

des 58 premiers motifs contient le nombre d'occurrences d'un motif uniforme. Le dernier contient le nombre d'occurrences de tous les motifs non uniformes. Cette variante LBP permet de réduire la taille du descripteur sans perte significative d'information, comme nous le montrerons dans la section 3.8.4. Ces caractéristiques LBP sont appelées LBP-Uniforme (LBP_u).

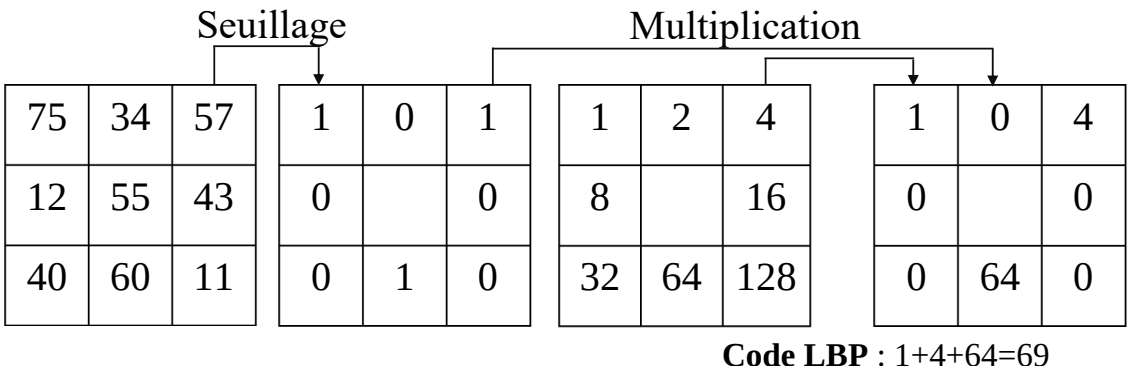


FIGURE 3.5 L'opérateur LBP.

3.6.2 Compound Local Binary Pattern (CLBP)

Ahmed et al. [4] ont tenté d'augmenter la robustesse du descripteur LBP en incorporant des informations locales supplémentaires qui sont ignorées par l'opérateur original LBP . Ils ont étendu le LBP de base à une autre variante, appelée CLBP. Dans CLBP, la fonction d'indicateur $s(x)$ est définie comme suit :

$$s(x) = \begin{cases} 00 & \text{si } x < 0, |x| \leq M_{avg} \\ 01 & \text{si } x < 0, |x| > M_{avg} \\ 10 & \text{si } x \geq 0, |x| \leq M_{avg} \\ 11 & \text{sinon} \end{cases} \tag{3.3}$$

où M_{avg} est l'amplitude moyenne des différences x entre la valeur de gris du pixel central et les valeurs de gris des pixels voisins. Le premier bit représente le signe de la différence, comme dans le codage LBP de base. Le deuxième bit est utilisé pour coder l'amplitude de la différence x par rapport à la valeur de seuil M_{avg}. Afin de réduire le nombre de caractéristiques, un schéma de codage est également proposé par Ahmed et al. [4] en divisant le modèle binaire CLBP en deux modèles sous-CLBP, comme illustré dans la figure 3.6.

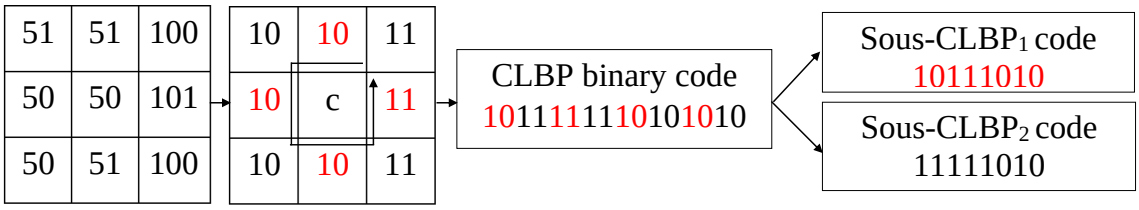


FIGURE 3.6 L'opérateur CLBP et la génération de deux sous-CLBP.

3.6.3 Local Ternary Pattern (LTP)

LTP est une simple généralisation de LBP introduite par Tan et Triggs [199] qui est considérablement plus discriminante que LBP, et plus robuste au bruit dans des régions uniformes. Décrivant les variations locales de texture, la mesure LTP est basée sur l'opérateur LBP, qui donne une mesure de texture invariante calculée à partir de l'analyse d'un voisinage local. Au lieu de binariser les valeurs échantillonnées, celles-ci peuvent prendre trois valeurs en fonction de leur distance à la valeur du pixel central. Dans LTP, la fonction d'indicateur $s(x)$ est définie comme suit :

$$s(x) = \begin{cases} 1 & \text{si } x \geq t \\ 0 & \text{si } |x| < t \\ -1 & \text{si } x \leq -t \end{cases} \tag{3.4}$$

où t est un seuil défini manuellement et x est la différence entre la valeur du pixel central et la valeur du pixel de voisinage. LTP est plus résistant au bruit, bien qu'il ne soit pas invariant aux transformations du niveau de gris en raison de la naïveté du choix du seuil. Pour surmonter ce problème, de nombreuses études ont proposé des techniques [228, 142, 90] pour calculer un seuil dynamique basé sur les valeurs des pixels voisins. Néanmoins, à notre connaissance, toutes les formules proposées dans la littérature pour calculer un seuil dynamique utilisent un paramètre appelé facteur d'échelle dont la valeur est constante pour tout le voisinage. Dans tous les travaux existants, pour définir ce paramètre, les auteurs ont choisi une valeur comprise entre 0 et 1. Dans [90], les auteurs ont évalué les performances de la méthode proposée en considérant différentes valeurs pour le facteur d'échelle. Dans notre étude, nous avons réalisé de vastes expériences en utilisant le LTP original avec différentes stratégies pour le calcul du seuil. La première stratégie consiste à définir la valeur du seuil de 0 à 5 [199], comme proposé dans de nombreux travaux de recherche. La deuxième stratégie proposée dans [90] utilise la formule (F1) (voir Table 3.2), en considérant différentes valeurs pour le facteur d'échelle δ. Dans cette formule, p_c indique la valeur du pixel central. Dans cette thèse, nous introduisons d'autres stratégies pour réaliser un seuillage dynamique avec des

formules simples (F2, F3, F4 et F5), données dans la table 3.2, où p_i est la valeur du i^{eme} pixel voisin. Une autre considération importante abordée dans cette thèse est la manière de définir LTP. En effet, la modification de la définition d'origine de l'opérateur LTP [199] pourrait avoir un impact sur les résultats de la classification. La définition originale de LTP (Eq. (3.4)) consiste à assigner 0 lorsque la valeur absolue de la différence est strictement inférieure au seuil. Alors que dans d'autres travaux [141, 123], les auteurs ont inconsciemment changé la définition originale de LTP en affectant 0 lorsque la valeur absolue de la différence est inférieure ou égale au seuil (Eq. (3.5)). Puisque la représentation de LTP est basée sur trois valeurs, l'attribution du signe égal à une ou plusieurs des trois conditions pourrait affecter les résultats de la classification. Cette déclaration nous a conduit à poser la question suivante : quelle définition de LTP serait adaptée à notre application ? Pour répondre à cette question, nous avons mené des expériences approfondies en considérant les différentes définitions (Eq (3.4), (3.5), (3.6) et (3.7)).

$$s(x) = \begin{cases} 1 & \text{si } x > t \\ 0 & \text{si } |x| \leq t \\ -1 & \text{si } x < -t \end{cases} \tag{3.5}$$

$$s(x) = \begin{cases} 1 & \text{si } x \geq t \\ 0 & \text{si } x \geq -t \quad \text{et} \quad x < t \\ -1 & \text{si } x < -t \end{cases} \tag{3.6}$$

$$s(x) = \begin{cases} 1 & \text{si } x > t \\ 0 & \text{si } x > -t \quad \text{et} \quad x \leq t \\ -1 & \text{si } x \leq -t \end{cases} \tag{3.7}$$

TABLE 3.2 Les formules du seuil dynamique de l'opérateur LTP (N : nombre de voisinage (dans ce travail N = 8)).

Formule 1 (F1)	**Formule 2 (F2)**	**Formule 3 (F3)**	**Formule 4 (F4)**	**Formule 5 (F5)**
$t = p_c \times \delta$	$t = \frac{\sum_{i=0}^{N-1} \sqrt{p_i}}{N}$	$t = \frac{\sum_{i=0}^{N-1} p_i}{N}$	$t = \frac{\sqrt{\sum_{i=0}^{N-1} p_i}}{N}$	$t = \sqrt{\sum_{i=0}^{N-1} \frac{p_i}{N}}$

3.6.4 Histogram of Oriented Gradient (HOG)

Dalal et Triggs [41] ont proposé un descripteur HOG, largement inspiré de SIFT, pour surmonter les limites de ce dernier dans le cas des grilles denses. Le but de HOG est de représenter l'apparence locale et la forme d'un objet dans une image par la distribution des gradients d'intensité ou de directions des contours. Ceci est accompli en divisant l'image

en petites régions connectées, appelées cellules, et, pour chaque cellule, en calculant un histogramme local (avec 9 bins) de directions de gradient pour les pixels appartenant à cette cellule. La concaténation de tous les histogrammes locaux forme le descripteur HOG. Les histogrammes locaux sont normalisés au contraste local en calculant une mesure d'intensité sur des régions spatiales plus grandes, appelées blocs. Cette normalisation conduit à une meilleure invariance des changements d'illumination et d'ombrage. Les paramètres HOG appliqués pour coder les composantes faciales dans les ROIs extraites sont : la taille des cellules, la taille des blocs et la taille des histogrammes.

3.7 Méthodes de classification

3.7.1 Support Vector Machine (SVM)

SVM est parmi les méthodes de classification les plus connues. Il est basé sur la théorie d'apprentissage statistique [215].

Étant donné une donnée d'entraînement x_i étiquetée $y_i, i = 1, \cdots, p, y_i \in \{-1, 1\}, x_i \in \Re^d$, l'algorithme SVM construit un hyperplan avec la plus grande marge qui sépare les exemples positifs et négatifs. Tous les points situés d'un côté de l'hyperplan sont étiquetés comme 1, et tous les points situés de l'autre côté sont étiquetés comme -1. Les points les plus proches du plan de séparation de données sont appelés vecteurs de support. SVM peut alors être défini comme :

$$\begin{cases} x_i.w + b \geq +1 & \text{pour } y_i = +1 \\ x_i.w + b \leq -1 & \text{pour } y_i = -1 \end{cases} \tag{3.8}$$

La limite de décision est un hyperplan de la forme : $x.w + b = 0$, où w est le vecteur normal à l'hyperplan, $|b|/||\ w\ ||$ est la distance de l'hyperplan par rapport à l'origine, et $||\ w\ ||$ est la norme euclidienne de w. Le problème de trouver l'hyperplan optimal peut être résolu en résolvant le problème d'optimisation suivant :

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{p} \xi(w, x_i, y_i) \tag{3.9}$$

où $C > 0$ est un paramètre de pénalité. $\sum_{i=1}^{p} \xi(w, x_i, y_i)$ peut être considéré comme une erreur totale de classification. La classification d'une nouvelle donnée de test est effectuée en calculant la distance entre les données de test et l'hyperplan.

SVM est un classifieur binaire. Pour le cas multiclasse, comme dans notre cas, le problème est décomposé en plusieurs problèmes de classification binaire, chacun d'entre eux est ensuite géré par un SVM. Parmi les méthodes SVM multiclasses existantes, nous avons opté pour

la méthode un-contre-un, qui est une approche compétitive selon la comparaison détaillée menée dans [85]. La méthode SVM multiclasse, basée sur un-contre-un, consiste à construire un classifieur pour chaque paire de classes, puis à entrainer $k(k-1)/2$ classifieurs (k est le nombre de classes) afin de distinguer les exemples d'une classe de ceux d'une autre classe. La classification d'une nouvelle donnée de test est déterminée par un vote majoritaire sur l'ensemble des classifieurs. Dans ce travail, pour implémenter la classification SVM, nous avons utilisé la bibliothèque LIBSVM [28], qui propose des codes sources de la méthode SVM multiclasse basée sur la stratégie un-contre-un. Nous avons utilisé un classifieur SVM linéaire dans toutes nos expériences pour éviter la sensibilité des paramètres. Les autres noyaux, RBF et polynomial ont été utilisés aux fins de comparaison (voir Section 3.8.4).

3.7.2 Random Forest (RF)

RF est un classifieur constitué de nombreux arbres de décision. Chaque arbre de décision est construit récursivement en assignant un test binaire à chaque nœud non-feuille en fonction des échantillons d'apprentissage. Pour la classification, la forêt aléatoire combine les résultats des arbres de décision pour voter pour la classe la plus populaire. La figure 3.7 illustre l'architecture du modèle RF pour une classification. Le classifieur RF peut être défini comme :

$$H(x) = arg\max_{Y} \sum_{1}^{k} I(h_i(x) = Y) \tag{3.10}$$

où $H(x)$ est le classifieur combiné final, k est le nombre d'arbres de décision, $h_i(x)$ représente un arbre de décision, Y est l'étiquette de classe, $I(h_i(x) = Y)$ indique x appartient à la classe Y.

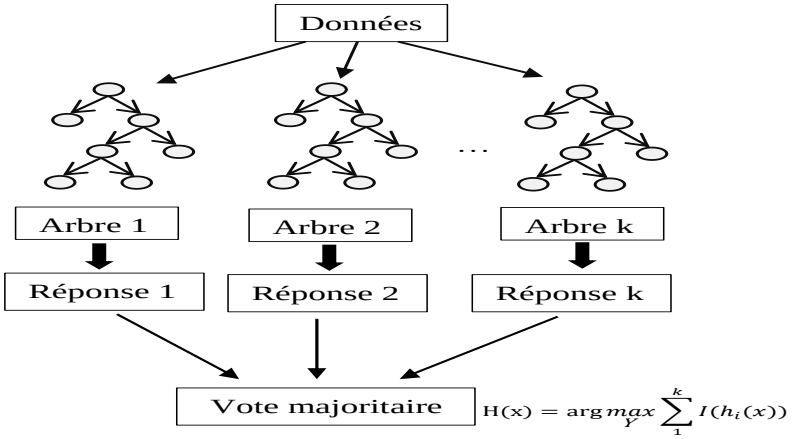


FIGURE 3.7 Architecture du modèle de RF.

TABLE 3.3 Propriétés des jeux de données utilisés CK, FEED, KDEF et JAFFE

Jeu de données	CK	FEED	KDEF	JAFFE
Type	posé	spontané	posé	posé
Conditions d'éclairage	uniforme	uniforme	uniforme	uniforme
Normalisation du visage	aucune	aucune	aucune	aucune
Angle du visage	frontal	frontal	frontal	frontal
# des images	610	630	280	213
Emotions	Ne, Ha, Sa, Su, An, Fe, Di	Ne, Ha, Sa, Su, An, Fe, Di	Ne, Ha, Sa, Su, An, Fe, Di	Ne, Ha, Sa, Su, An, Fe, Di
# de sujets	9 hommes/ 23 femmes	8 hommes/ 8 femmes	20 hommes/ 20 femmes	10 femmes
Ethnicité	euro-Américaines, afro-Américaines, asiatiques et latinos	européenne	caucasienne	Japonaise

3.8 Expérimentations

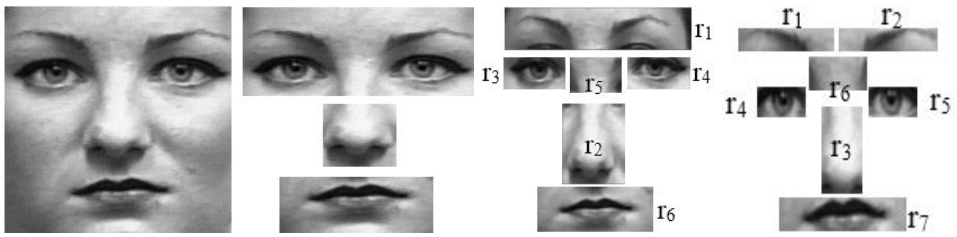
3.8.1 Bases de données

Pour évaluer la méthodologie proposée, quatre bases de données publiques différentes sont utilisées pour effectuer des tests dans différentes conditions. La première base de données, appelée FEED [220], représente des émotions spontanées. Les autres bases de données CK [101], KDEF [134] et JAFFE [135] contiennent des émotions posées. La table 3.3 résume les propriétés de chaque jeux de données utilisés.

3.8.2 Résultats et discussion

Nous avons appliqué la validation croisée 10-fold pour tester notre méthode proposée. Chaque ensemble de données est divisé en 10 sous-ensembles indépendamment de la personne (person-independent) avec un nombre à peu prés égal d'images pour effectuer 10 expériences. Dans chaque expérience, neuf sous-ensembles sont utilisés pour l'apprentissage et le sous-ensemble restant est utilisé pour le test. La validation croisée 10-fold est utilisée pour remplir la matrice de confusion, puis le taux de reconnaissance est obtenu en calculant le F-score, défini [188] comme suit :

$$F - score = \frac{2.Rappel.Precision}{Rappel + Precision} \tag{3.11}$$



(a) (b) (c) (d)

FIGURE 3.8 Régions d'intérêt (ROIs). (a) le visage entier comme une seule ROI (b) 3 ROIs déjà utilisées dans la littérature [244] (c) 6 ROIs déjà utilisées dans la littérature [50] (d) Nos 7 ROIs proposées.

avec

$$Precision = \frac{Nb\ vrai\ positif}{Nb\ vrai\ positif + Nb\ faux\ positif} \tag{3.12}$$

et

$$Rappel = \frac{Nb\ vrai\ positif}{Nb\ vrai\ positif + Nb\ faux\ ngatif} \tag{3.13}$$

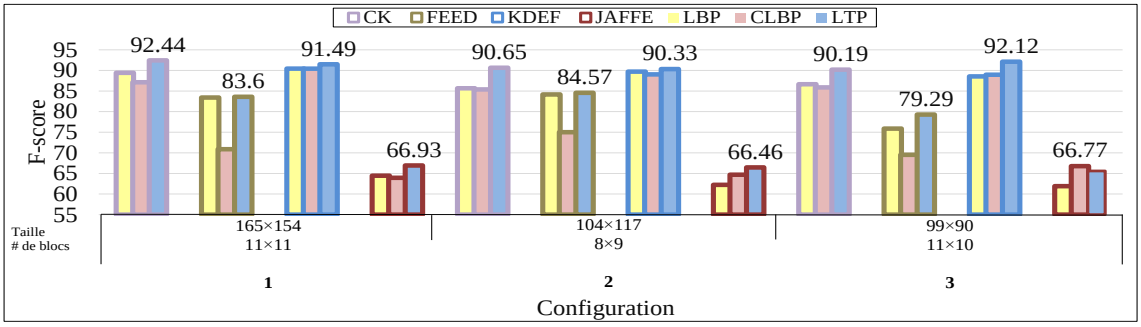
Le choix de la taille et du nombre de blocs de ROI ainsi que des descripteurs affecte généralement les performances de reconnaissance. Dans nos expériences, chaque ROI est divisée en blocs $W \times H$ pour le calcul des descripteurs. Les meilleurs paramètres sont définis en effectuant plusieurs tests. Pour le test et l'évaluation, nous avons considéré trois décompositions du visage en ROIs avec différentes configurations de paramètres, à savoir la taille des ROIs et le nombre de blocs dans chaque ROI, comme indiqué dans les figures 3.9, 3.10 et 3.11. La première décomposition considère le visage entier comme une seule ROI (voir Figure 3.8a), comme proposé dans [4, 26]. Dans la deuxième décomposition, le visage est composé de six ROIs (voir Figure 3.8c), comme proposé dans [50]. La troisième décomposition est celle que nous avons proposée, elle considère sept ROIs, comme le montre la figure 3.8d. Dans [244], les auteurs ont proposé une autre décomposition faciale avec trois ROIs (voir Figure 3.8b). Les performances des décompositions faciales avec trois ROIs (voir Figure 3.8b) et six ROIs (voir Figure 3.8c) sont proches, avec un léger avantage à la deuxième décomposition (six ROIs). En effet, les meilleurs taux de reconnaissance fournis par la décomposition du visage avec trois ROIs contre six ROIs sont 93.34 % (vs. 93.55 %), 82.17 % (vs. 83.04 %), 92.6 % (vs. 92.53 %) et 73.65% (vs. 76.46%) sur les jeux de données CK, FEED, KDEF et JAFFE, respectivement. Pour des raisons de concision, nous avons choisi de ne pas fournir de détails sur les résultats de le décomposition en trois ROIs comme nous l'avons fait pour les autres décompositions.

Les paramètres de l'expérience sont les suivants :

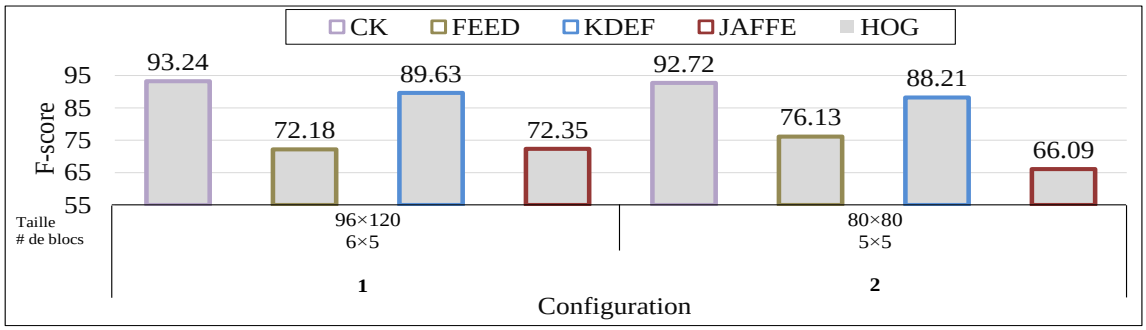
- Paramètres de ROI : taille de ROI et nombre de blocs dans chaque ROI (voir Figures 3.9, 3.10, 3.11). Figures 3.9, 3.10 et 3.11 fournissent, pour chaque jeu de données testé, uniquement les configurations de paramètres donnant des taux de reconnaissance élevés : pour chaque descripteur testé, nous conservons la configuration des paramètres qui donne le taux de reconnaissance le plus élevé.
- Paramètres du descripteur de caractéristiques : pour LBP, CLBP et LTP, nous avons utilisé un rayon de 1 pixel, un 8-voisinage et un histogramme de 256 bin. Concernant le descripteur LTP, le seuil est défini manuellement ou automatiquement en utilisant les formules données dans la table 3.2. Nous mentionnons dans l'analyse des résultats comment nous définissons le seuil. Lorsqu'il est défini manuellement, la valeur du seuil est donnée. Pour HOG, nous avons utilisé des blocs de 8×8 pixels, des cellules de 2×2 pixels et un histogramme de 9 bins pour chaque bloc.
- Paramètres de SVM : nous avons utilisé un noyau linéaire avec le paramètre C = 0.3 sélectionné en effectuant la méthode de grille de recherche (grid-search) basée sur une validation croisée de 10-fold.

3.8.2.1 Résultats expérimentaux en utilisant le visage entier comme une seule ROI

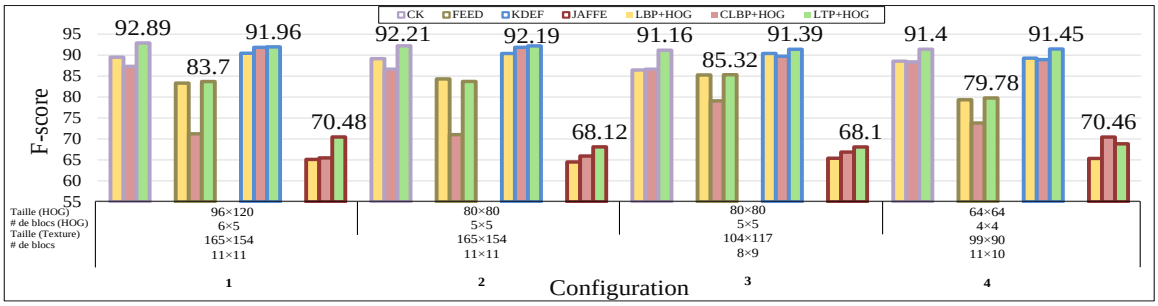
La figure 3.9 présente les résultats obtenus en utilisant le visage entier comme une seule ROI. Pour chaque jeu de données, nous avons considéré les descripteurs de texture (LBP, CLBP et LTP) (voir Figure 3.9a), le descripteur de forme (HOG) (voir Figure 3.9b) et leur concaténation (voir Figure 3.9c). Comme nous pouvons le voir dans la figure 3.9, le descripteur HOG fournit les meilleurs taux de reconnaissance (93.24% et 72.35%) lorsque les jeux de données CK et JAFFE sont respectivement testés avec la configuration 1 des paramètres 1 (voir Figure 3.9b). Nous pouvons également voir pour le jeu de données CK, que lorsque le descripteur LTP est impliqué (seul ou sous forme hybride avec le descripteur HOG), le taux de reconnaissance est intéressant avec presque toutes les configurations de paramètres et souvent proche du meilleur taux de reconnaissance. Cependant, lorsque les jeux de données FEED et KDEF sont testés, le descripteur hybride LTP + HOG montre un taux de reconnaissance élevé, par rapport aux autres descripteurs. En effet, nous obtenons 85.32% sur le jeu de données FEED avec la configuration 3 des paramètres (voir Figure 3.9c). Notons que le LTP utilisé ici est défini en utilisant Eq. (3.4) avec un seuil fixe sélectionné expérimentalement ($t = 1$). Pour l'ensemble de données KDEF, nous obtenons un taux de reconnaissance de 92.19%, avec la configuration 2 des paramètres (voir Figure 3.9c). Ici, le LTP est défini en utilisant Eq. (3.4) avec une procédure de seuil basée sur la formule 3 (voir Table 3.2).



(a)



(b)



(c)

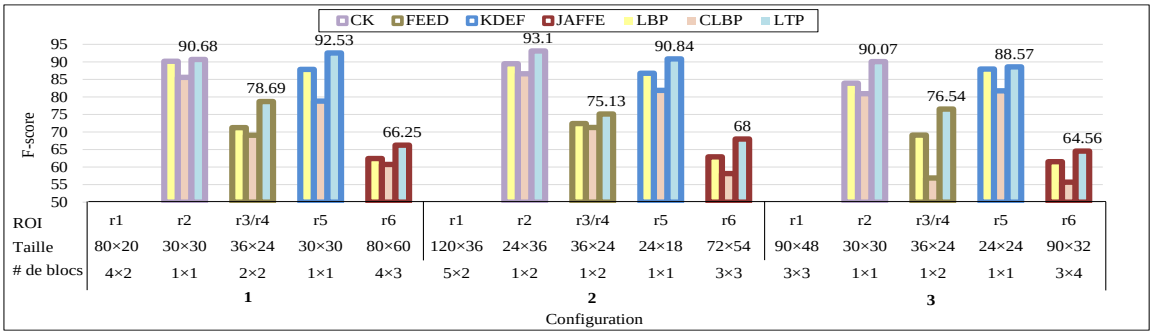
FIGURE 3.9 Taux de reconnaissance pour ROI = 1 en utilisant différents descripteurs avec différentes configurations de taille et nombre de blocs. (a) Descripteur de texture (trois configurations). (b) Descripteur HOG (deux configurations). (c) Méthode hybride (trois configurations).

3.8.2.2 Résultats expérimentaux en utilisant six régions d'intérêt (nombre de ROIs = 6)

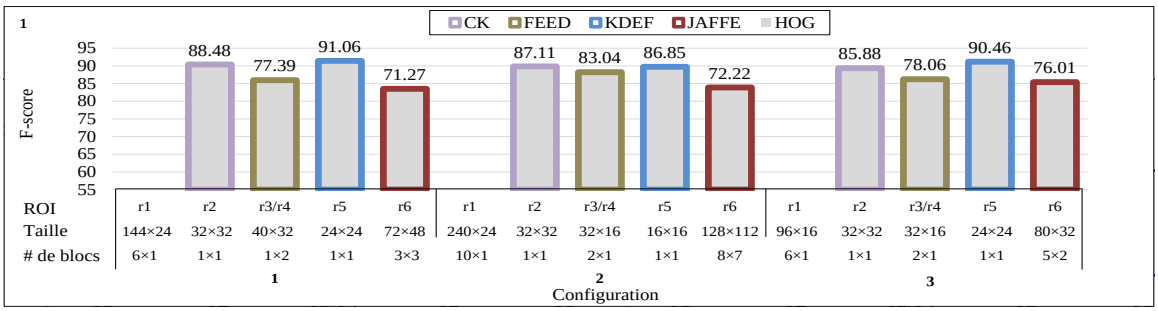
Dans cette section, nous analysons les performances des descripteurs en utilisant la décomposition du visage avec 6 ROIs. Comme nous pouvons le voir dans la figure 3.10, appliqué sur le jeu de données CK, le descripteur hybride (LTP + HOG) surpasse tous les autres descripteurs dans toutes les configurations de paramètres. Ici, le LTP est calculé en utilisant Eq. (3.7) avec le seuil donné par la formule 1 (voir Table 3.2). Le descripteur hybride (LTP + HOG) était capable d'augmenter légèrement le taux de reconnaissance à 93.55% (voir Figure 3.10c), comparé aux résultats du cas précédent, c'est-à-dire en considérant le visage comme une seule ROI (nombre de ROI = 1), où le descripteur HOG a montré son efficacité face à tous les autres descripteurs (voir Figure 3.9). Concernant le jeu de données KDEF, le descripteur LTP, défini avec Eq. (3.5) et le seuil de la formule 4 (voir Table 3.2), fournit des meilleurs résultats (voir Figure 3.10a) . Il a légèrement augmenté le taux de reconnaissance à 92.53%, comparé aux résultats du cas précédent (nombre de ROI = 1), où les descripteurs hybrides ont donné les meilleurs résultats, en particulier la combinaison de HOG et LTP (voir Figure 3.9c). A propos du jeu de données JAFFE, cette décomposition a montré son efficacité par rapport à ROI = 1 presque pour tous les descripteurs utilisés (par ex. en utilisant HOG le taux a basculé de 72.35% à 76.01%), un meilleur taux de 76.46% est obtenu en utilisant le descripteur hybride LBP+HOG. Pour le jeu de données FEED, on peut noter globalement à partir de la Figure 3.10 une diminution significative du taux de reconnaissance à 83.04%, par rapport aux résultats du cas précédent où ROI = 1 (voir Figure 3.9). Le taux de reconnaissance décroissant concerne les descripteurs de texture et hybrides. Cependant, le descripteur HOG augmente le taux de reconnaissance, comparé aux résultats du cas précédent (nombre de ROI = 1) avec le même descripteur, et fournit le meilleur taux de reconnaissance (83.04%) par rapport à tous les descripteurs testés (voir Figure 3.10).

3.8.2.3 Résultats expérimentaux utilisant la décomposition proposée (nombre de ROIs = 7)

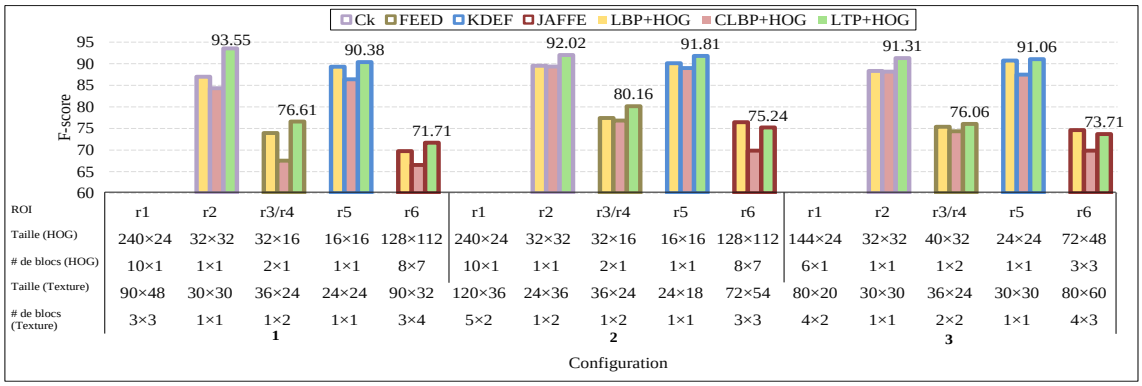
Comme nous pouvons le voir dans la figure 3.11, une augmentation significative du taux de reconnaissance pour tous les jeux de données testés est obtenue avec notre décomposition faciale proposée (nombre de ROIs = 7) comparée aux précédentes (nombre de ROI = 1 et nombre de ROIs = 6). En effet, notre décomposition faciale a fourni les meilleurs taux de reconnaissance, obtenus principalement avec les descripteurs hybrides. En particulier, la combinaison de LTP et HOG atteint les taux de reconnaissance élevés de 96.06%, 92.03%, 93.34% et 77.08% pour les jeux de données respectifs CK, FEED, KDEF et JAFFE. Dans



(a)



(b)

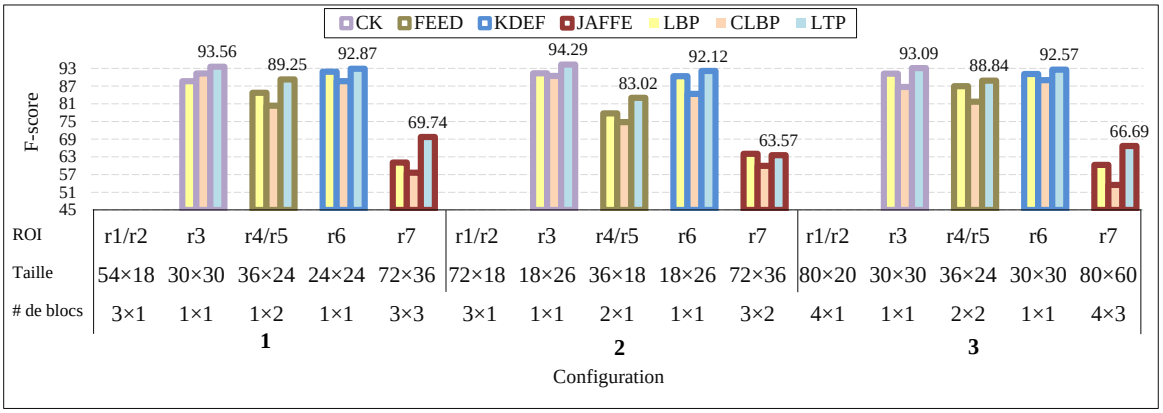


(c)

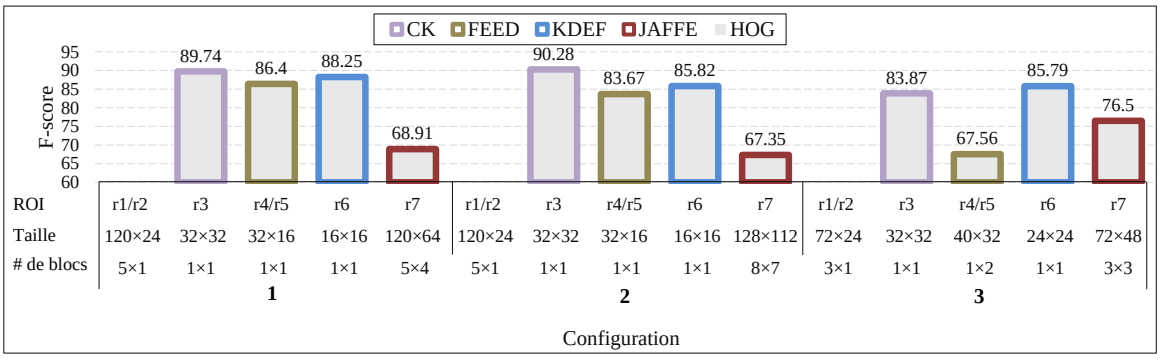
FIGURE 3.10 Taux de reconnaissance pour ROIs = 6 en utilisant différents descripteurs avec différentes configurations de taille et nombre de blocs. (a) Descripteur de texture (trois configurations). (b) Descripteur HOG (trois configurations). (c) Méthode hybride (trois configurations).

la combinaison des descripteurs, le LTP qui a permis ces meilleurs résultats est défini avec Eq. (3.7) associé au seuil de la formule 1, Eq. (3.7) associé au seuil de la formule 3, Eq. (3.5) associé au seuil de la formule 1, et Eq. (3.5) associé au seuil de la formule 4, respectivement pour CK, FEED, KDEF et JAFFE. Les améliorations apportées par notre décomposition faciale ont été possibles grâce aux ROIs sélectionnées (emplacement précis et au bon recadrage des composants du visage). Les tables 3.4, 3.5, 3.6 et 3.7 résument les meilleurs résultats obtenus respectivement pour les jeux de données CK, FEED, KDEF et JAFFE à travers les différents descripteurs et les différentes décompositions faciales (nombre de ROIs $= 1$, 6 et 7). A partir de ces tables, nous pouvons voir que la méthode proposée (décomposition faciale avec le nombre de ROIs $= 7$) surpasse toutes les autres en utilisant presque tous les descripteurs et pour tous les jeux de données testés. Les tables montrent que les descripteurs hybrides présentent généralement de meilleurs résultats en particulier pour la décomposition faciale proposée.

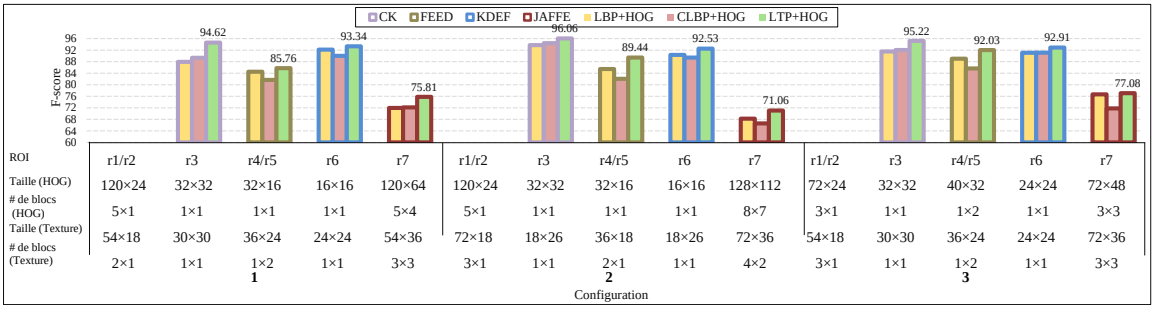
Concernant la variabilité des jeux de données, on peut voir dans les figures 3.9, 3.10 et 3.11, que d'une part, les taux de reconnaissance pour le jeu de données FEED, composé d'expressions faciales spontanées, sont inférieures à ceux obtenus pour les jeux de données CK et KDEF, composés d'expressions faciales posées. La raison principale est que les expressions faciales posées sont faciles à détecter, par rapport aux expressions spontanées, qui sont généralement acquises avec un protocole expérimental non contrôlé. D'un autre côté, les taux de reconnaissance pour l'ensemble de données KDEF sont faibles, comparés à ceux obtenus pour l'ensemble de données CK. Cela est dû à la différence de construction de chaque base de données. En effet, l'ensemble de données CK est composé de séquences d'images, alors que l'ensemble de données KDEF est composé d'images indépendantes. Le taux de reconnaissance de la base de données JAFFE est inférieur à ceux des autres bases de données. La raison principale réside dans l'étiquetage incorrect de certaines images d'expressions faciales dans la base de données JAFFE. La figure 3.12 montre des exemples d'étiquetage incorrect des images d'expression dans la base de données JAFFE.



(a)



(b)



(c)

FIGURE 3.11 Taux de reconnaissance pour ROIs = 7 en utilisant différents descripteurs avec différentes configurations de taille et nombre de blocs. (a) Descripteur de texture (trois configurations). (b) Descripteur HOG (trois configurations). (c) Méthode hybride (trois configurations).

TABLE 3.4 Comparaison des performances de ROIs = 7, ROI = 1 et ROIs = 6 pour l'ensemble de données CK et tous les descripteurs testés.

	ROI=1	ROIs=6	ROIs=7
LBP	89.42	90.17	**91.33**
CLBP	87.13	86.6	**91.26**
LTP	92.44	93.1	**94.29**
HOG	**93.24**	88.48	90.28
LBP+HOG	89.5	89.53	**93.75**
CLBP+HOG	87.31	89.58	**94.43**
LTP+HOG	92.89	93.55	**96.06**

TABLE 3.5 Comparaison des performances de ROIs = 7, ROI = 1 et ROIs = 6 pour l'ensemble de données FEED et tous les descripteurs testés.

	ROI=1	ROIs=6	ROIs=7
LBP	84.17	72.41	**86.96**
CLBP	75.01	71.29	**81.7**
LTP	84.75	78.69	**89.25**
HOG	76.13	83.04	**86.4**
LBP+HOG	85.26	77.43	**89.89**
CLBP+HOG	79.09	76.84	**85.66**
LTP+HOG	85.32	80.16	**92.03**

TABLE 3.6 Comparaison des performances de ROIs = 7, ROI = 1 et ROIs = 6 pour l'ensemble de données KDEF et tous les descripteurs testés.

	ROI=1	ROIs=6	ROIs=7
LBP	90.44	87.95	**91.87**
CLBP	**90.43**	81.87	88.66
LTP	92.12	92.53	**92.87**
HOG	89.63	**91.06**	88.25
LBP+HOG	90.44	90.74	**92.2**
CLBP+HOG	91.85	89	**92.59**
LTP+HOG	92.19	91.81	**93.34**

TABLE 3.7 Comparaison des performances de ROIs = 7, ROI = 1 et ROIs = 6 pour l'ensemble de données JAFFE et tous les descripteurs testés.

	ROI=1	ROIs=6	ROIs=7
LBP	**64.45**	62.92	64.04
CLBP	**66.77**	60.75	60.07
LTP	66.93	68	**69.74**
HOG	72.35	76.01	**76.08**
LBP+HOG	65.36	**76.46**	73.68
CLBP+HOG	66.08	71.86	**73.97**
LTP+HOG	70.48	75.24	**77.08**

(a) (b) (c)

FIGURE 3.12 Etiquettage incorrect des images d'expression dans la base de données JAFFE. (a) Surprise. (b) Joie. (c) Tristesse.

3.8.2.4 Matrices de confusion

Cette section analyse les résultats à travers des matrices de confusion. Les tables 3.8, 3.9, 3.10 et 3.11 affichent les matrices de confusion (correspondant au taux de reconnaissance élevé atteint) respectivement pour les jeux de données CK, FEED, KDEF et JAFFE. Comme nous pouvons le voir dans la table 3.8, nous avons obtenu de très bons résultats pour la majorité des émotions avec un taux moyen proche de 100%, à l'exception des émotions neutre et colère reconnues respectivement avec des taux moyens de 86% et 90%. L'expression neutre est mal classée en tant qu'expression de peur avec un taux d'erreur de 9%. Pour le jeu de données FEED (voir Table 3.9), nous pouvons observer que l'une des principales raisons de la diminution du taux de reconnaissance est la mauvaise classification des expressions tristesse, colère et peur comme expression neutre, reconnue avec une précision moyenne de 82.13% (calculée à partir de la colonne neutralité comme $95/(95+7+5+6.66+2)$ en utilisant Eq. 3.12), et la confusion entre les expressions peur et surprise et entre les expressions de dégoût et joie/colère. Pour le jeu de données KDEF (voir Table 3.10), les émotions neutre, joie et surprise sont reconnues avec un taux maximum de 100%. Les autres émotions sont classées

avec un taux de reconnaissance moyen compris entre 85% et 90%. Nous pouvons voir aussi que la confusion d'expressions est plus présente, avec des pourcentages plus ou moins faibles. Les tables 3.9 et 3.10 montrent aussi que l'expression peur est confondue avec l'expression surprise avec un taux d'erreur de 10% et de 7.5% respectivement pour les jeux de données FEED et KDEF. Cette classification erronée est peut-être due à une déformation similaire du visage causée par ces expressions. Concernant le jeu de données JAFFE (voir Table 3.11), presque toutes les émotions sont confondues avec l'émotion neutralité avec un taux d'erreur supérieur à 3%. Les émotions neutralité et surprise sont reconnues respectivement avec des taux acceptables de 90% et 93.33%. Nous pouvons aussi remarquer que les émotions dégoût, tristesse et peur sont classées avec un taux de reconnaissance bas, compris entre 55% et 68%.

TABLE 3.8 Matrice de confusion pour l'ensemble de données CK (associée au taux de reconnaissance élevé : 96.06%)

	Neutralité	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Neutralité	**86**	0	1	0	3	9	1
Joie	0	**100**	0	0	0	0	0
Tristesse	1	0	**99**	0	0	0	0
Surprise	3	0	0	**97**	0	0	0
Colère	2.22	0	4.44	0	**90**	3.33	0
Peur	0	0	0	0	0	**100**	0
Dégoût	0	0	0	0	0	0	**100**

TABLE 3.9 Matrice de confusion pour l'ensemble de données FEED (associée au taux de reconnaissance élevé : 92.03%)

	Neutralité	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Neutralité	**95**	0	2	0	0	3	0
Joie	0	**94**	0	6	0	0	0
Tristesse	7	0	**93**	0	0	0	0
Surprise	0	0	0	**98**	0	2	0
Colère	5	0	0	0	**93**	2	0
Peur	6.66	0	0	10	0	**83.33**	0
Dégoût	2	6	1	0	5	0	**86**

TABLE 3.10 Matrice de confusion de l'ensemble de données KDEF (associée au taux de reconnaissance élevé : 93.34%)

	Neutralité	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Neutralité	**100**	0	0	0	0	0	0
Joie	0	**100**	0	0	0	0	0
Tristesse	5	0	**90**	0	2.5	0	2.5
Surprise	0	0	0	**100**	0	0	0
Colère	2.5	0	2.5	2.5	**87.5**	0	5
Peur	0	0	2.5	7.5	0	**90**	0
Dégoût	7.5	0	5	0	2.5	0	**85**

TABLE 3.11 Matrice de confusion de l'ensemble de données JAFFE (associée au taux de reconnaissance élevé : 77.08%)

	Neutralité	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Neutralité	**90**	3.33	0	0	6.66	0	0
Joie	9.67	**87.09**	3.22	0	0	0	0
Tristesse	3.22	3.22	**61.29**	0	9.67	16.12	6.45
Surprise	3.33	3.33	0	**93.33**	0	0	0
Colère	6.66	0	3.33	0	**80**	0	10
Peur	12.5	0	0	0	9.37	**68.75**	9.37
Dégoût	6.89	0	13.79	0	24.13	0	**55.17**

3.8.2.5 Temps de traitement

Le temps de traitement (en ms), nécessaire pour traiter une image avec toutes les étapes de la méthode proposée (détection de visage, détection de points, extraction de régions, conversion RVB en gris, redimensionnement et partitionnement, calcul des caractéristiques, prédiction SVM), est reporté dans la table 3.12. Il a été calculé en considérant la moyenne du traitement sur 100 images, sur une machine Windows Intel(R) Core(TM) i5-2430M CPU 2.40GHz. Les codes d'algorithme sont écrits en C++.

TABLE 3.12 Temps de traitement pour les jeux de données CK,FEED et KDEF.

	CK	FEED	KDEF
HOG	196.24	201.30	228.38
LTP	214.96	220.02	242.89
HOG+LTP	236.18	244.61	266.60

3.8.3 Comparaison avec l'état de l'art

Cette section est dédiée à la comparaison avec les méthodes les plus récentes de l'état de l'art, en utilisant deux bases de données CK+ [132] et SFEW [47]. Le processus de comparaison, dans le domaine de REF, fait face à différentes difficultés telles que l'absence d'évaluation commune, les bases de données partagées nécessitent un paramétrage expérimental pour sélectionner des images et les codes d'algorithmes des méthodes existantes ne sont pas accessibles et leur réimplémentation peut générer des erreurs.

3.8.3.1 Base de données CK+

Pour faire face aux difficultés présentées ci-dessus et faire une comparaison équitable, nous avons effectué différentes expériences sur la base de données CK+ [132], utilisée dans toutes les méthodes comparées, en suivant les mêmes protocoles pour la sélection d'images et la validation croisée k-fold comme effectué dans les travaux considérés pour la comparaison [72, 79, 107, 129, 126]. Le premier protocole, appliqué dans [107, 129, 126], consiste à sélectionner dans chaque séquence la première image pour l'expression neutre et trois images de pic (les trois dernières images de la séquence) pour l'expression cible. Nous rappelons que chaque séquence représente l'expression cible en commençant par l'expression neutre. Le second protocole, utilisé dans [79], sélectionne dans chaque séquence la dernière image de pic pour l'expression cible. Dans [79], l'expression neutre n'est pas prise en compte dans le processus de reconnaissance. Le troisième protocole, appliqué dans [72], prend deux images de pics pour les expressions colère, peur et tristesse, le dernier pic pour les expressions dégoût, joie et surprise et la première image pour l'expression neutre dans quelques séquences. Les tables 3.13 et 3.14 rapportent les taux de reconnaissance de notre méthode proposée et les méthodes comparées, appliquées respectivement sur CK+7 (toutes les expressions émotionnelles y compris neutre) et CK+6 (excluant l'expression neutre). La table 3.15 résume les valeurs des paramètres qui ont permis à notre méthode d'atteindre les meilleurs résultats. Pour toutes les méthodes en comparaison, nous avons considéré les performances de reconnaissance rapportées dans les articles référencés (voir Table 3.13 et 3.14). Comme indiqué dans la table 3.13, notre méthode proposée atteint les meilleurs taux de reconnaissance parmi toutes les méthodes comparées indépendamment de la catégorie d'extraction de caractéristiques et de la procédure utilisée pour l'enregistrement du visage. En effet, comparée à [126], notre méthode offre un taux de reconnaissance global obtenu sur k-fold CV (désigné par gTR) de 96.03% VS 93.7%. Par rapport à [72], la méthode proposée atteint 94.48% en gTR et 94.52% en F-score VS 90.08% et 90.64%. À partir de la table 3.14, nous pouvons observer que notre méthode dépasse la méthode qui extrait les caractéristiques

TABLE 3.13 Comparaison de différentes méthodes sur la base de données CK+ avec 7 expressions. a et b sont les références des configurations (voir Table 3.15) ayant permis à notre méthode d'atteindre les meilleurs résultats en utilisant les protocoles expérimentaux [126] et [72].

Méthode	Catégorie	Enregistrement du visage	Protocole expérimental	F-score	gTR
Liu et al. [126]	Deep learning	Visage	[126]	N/A	93.7
Notre méthode	Apparence	ROI		94.63^{a} (94.19^{b})	$\mathbf{96.03}^{a}$ ($\mathbf{95.47}^{b}$)
Ghimire et al. [72]	Apparence et Géométriques	ROI	[72]	90.64	90.08
Notre méthode	Apparence	ROI		$\mathbf{94.52}^{b}$	$\mathbf{94.48}^{b}$

des petits patchs situés autour des points faciaux [79] et surpasse aussi celles utilisant BDBN [129] et les caractéristiques d'apparence et géométrique [72]. Nous notons que la méthode BDBN [129] atteint un F-score de 83.4%, ce qui est inférieur à celui fourni par notre méthode (96.9%). Comparée à notre méthode, la méthode [107] obtient le meilleur gTR de 98.3% VS 96.77% grâce à la procédure d'augmentation de données en appliquant une transformation aléatoire à chaque image d'entrée (translations, retournements horizontaux, rotations, mise à l'échelle et augmentation d'intensité de pixel). Il convient de noter que notre méthode, contrairement à toutes les méthodes comparées, n'utilise aucun prétraitement (l'alignement des visages ou l'augmentation des données). En outre, notre méthode nécessite moins de mémoire et de coût de calcul contrairement aux méthodes de deep learning [129, 125].

Comme nous pouvons le voir dans la table 3.15, trois configurations (a, b et c) ont été considérées dans notre méthode. Ces configurations ont permis d'obtenir les meilleurs résultats, surpassant toutes les méthodes de l'état de l'art comparées, sauf une d'entre elles à laquelle la méthode proposée reste compétitive. Cependant, si l'on considère uniquement la configuration b (voir Table 3.15) comme une configuration standard pour notre méthode, les résultats sont toujours meilleurs que ceux des méthodes de l'état de l'art comparées, quel que soit le protocole expérimental utilisé et quel que soit le nombre d'expressions considéré (6 ou 7), comme nous pouvons le voir dans les tables 3.13 et 3.14.

3.8.3.2 Base de données SFEW

La base de données SFEW couvre des expressions faciales non contraintes des différentes poses de tête et des occlusions. En fait, ses conditions d'acquisition d'images sont proches

TABLE 3.14 Comparaison de différentes méthodes sur la base de données CK + avec 6 expressions. b et c sont les références des configurations (voir Table 3.15) ayant permis à notre méthode d'atteindre les meilleurs résultats en utilisant les protocoles expérimentaux([107],[129], [72]) et [79].

Méthode	Catégorie	Enregistrement du visage	Protocole expérimental	F-score	gTR
Khorrami et al. [107]	Deep learning	Visage	[107]	N/A	**98.3**
Notre méthode	Apparence	ROI		96.01^{b}	96.77^{b}
Liu et al. [129]	Deep learning	Visage	[129]	83.4	96.7
Notre méthode	Apparence	ROI		**96.9**^{b}	**97.52**^{b}
Ghimire et al. [72]	Apparence et Géométrique	ROI	[72]	94.24	94.1
Notre méthode	Apparence	ROI		**95.8**^{b}	**95.58**^{b}
Happy and Routray [79]	Apparence	Patch	[79]	94.39	94.14
Notre méthode	Apparence	ROI		**96.3**^{c} (**95.13**^{b})	**97.18**^{c} (**96.25**^{b})

TABLE 3.15 Configurations optimales de notre méthode pour atteindre les meilleurs résultats (voir Tables 3.13 et 3.14) sur l'ensemble de données CK+.

Référence	Descripteur	LTP/ seuil	Configuration de LTP	Configuration du HOG
a	LTP+HOG	Eq. (3.6) / formule 3 (voir Table 3.2)	Configuration 1 (voir Figure 3.11a)	Configuration 1 (voir Figure 3.11b)
b	LTP+HOG	Eq. (3.4) / formule 3 (voir Table 3.2)	Configuration 1 (voir Figure 3.11a)	Configuration 1 (voir Figure 3.11b)
c	LTP+HOG	Eq. (3.7) / seuil (t=2)	Configuration 2 (voir Figure 3.11a)	Configuration 1 (voir Figure 3.11b)

du monde réel. Ainsi, les visages ne sont pas facilement détectable. Dans un premier temps, nous avons utilisé le détecteur du visage VJ qui malheureusement, n'a pas pu détecter que quelques visages de la base SFEW. Ensuite, nous avons utilisé le détecteur du visage CDPM destiné à la détection des visages à vue multiple [151]. Bien que les résultats de détection du visage fournis par [151], à l'aide des modèles des parties déformables en cascade, soient précis dans de nombreux scénarios difficiles, ils contiennent une quantité non négligeable de faux positifs et de faux négatifs. Les faux positifs correspondent aux images dans lesquelles un objet est identifié comme un visage et les faux négatifs correspondent aux images évaluées à tort comme ne contenant aucun visage. Par conséquent, nous avons utilisé le mode interactif de la méthode SDM (voir Section 3.4) pour assurer la précision de la détection du visage (c-à-d, nous avons procédé à la détection manuelle du visage). A l'issue de cette étape, nous appliquons SDM pour détecter les 49 points faciaux. SDM n'a pas pu détecter ces points sur la plupart des visages de la base SFEW. La table 3.16 montre le nombre d'images de visage de la base SFEW et le nombre d'images où un visage est détecté. Ce sont ces images qui sont utilisées dans notre expérience.

Toutes les méthodes considérées pour la comparaison sont des méthodes récentes, qui utilisent le deep learning, sauf [42] qui est basée sur des caractéristiques handcrafted. Nous pouvons remarquer à partir de la table 3.17 que le taux de reconnaissance de notre méthode surpasse les taux de la méthode handcrafted [42] et les méthodes de deep learning [152, 126]. Notre méthode reste compétitive avec [120] mais très inférieure à [194, 245], en précisant que ces méthodes [194, 245] sont très gourmandes en temps de calcul, puisque [194] utilise quatre méthodes pour l'extraction de caractéristiques et un réseau de fusion en utilisant SVM et Partial Least Squares (PLS) pour la classification, et [245] utilise un deep learning multiple.

TABLE 3.16 Le nombre d'images où un visage est détecté, pour chaque expression dans SFEW.

	# images de SFEW		# images de SFEW avec détection de visage	
	Apprentissage	Validation	Apprentissage	Validation
Neutralité	150	86	142	86
Joie	198	73	184	69
Tristesse	172	73	163	71
Surprise	96	57	93	54
Colère	178	77	161	73
Peur	98	47	84	42
Dégoût	66	23	64	23
Total	958	436	891	418

TABLE 3.17 Comparaison de différentes méthodes sur la base de données SFEW avec 7 expressions.

	Méthode proposée	[42]	[152]	[194]	[126]	[245]	[120]
gTR (%)	38.27	37.1	38	56.32	30.14	55.96	40

3.8.4 Expériences de comparaison de SVM vs RF et LBP/LTP vs LBP_u/LTP_u

Dans cette section, quant à la tâche de reconnaissance des expressions faciales, deux techniques d'apprentissage automatique sont examinées (voir Table 3.20). La première est la technique SVM avec différents noyaux à savoir linéaire, polynomial et RBF. La deuxième est la technique RF. Les descripteurs hybrides utilisant LBP/LTP+HOG et LBP_u/LTP_u+HOG sont considérés et comparés (voir Table 3.21).

La table 3.20 présente les taux de reconnaissance du descripteur hybride LTP+HOG calculé par un SVM linéaire, un SVM gaussien (RBF), un SVM polynomial et une RF sur les jeux de données CK, FEED, KDEF et JAFFE. Les tables 3.18 et 3.19 détaillent les paramètres utilisés respectivement pour les classifieurs SVM et RF. Nous précisons que nous avons utilisé la méthode de grille de recherche (grid-search) basée sur une validation croisée pour trouver les paramètres optimaux des noyaux RBF et polynomial.

TABLE 3.18 Les paramètres des noyaux RBF et polynomial.

Noyau	RBF		Polynomial			
CV	5-fold		2-fold			
Paramètres	gamma	C	gamma	degrés	coef	C
	5.96E-08	256	0.00015	0.49	274.4	12.5

TABLE 3.19 Paramètres utilisés pour la construction de RF

Nombre de caractéristiques à considérer pour la séparation	50
Profondeur maximal des arbres	20
Nombre minimal d'échantillons par nœud à diviser	5
Précision minimale de la forêt	99%
nombre d'échantillons aléatoires par arbre	$\sqrt{p}$ (p est la taille du descripteur)
Nombre d'arbres de la forêt	300

A partir de la table 3.20, nous observons que les SVM produisent des taux de reconnaissance globaux élevés (91-96%) pour les jeux de données CK, FEED et KDEF et un taux

moyen (77.5%) pour le jeu de données JAFFE. Le noyau linéaire et le noyau polynomial fournissent presque la même performance, meilleure que celle du noyau RBF. La technique RF produit des performances inférieures à SVM atteignant respectivement 87.72%, 79.86%, 90.06% et 72% pour les jeux de données CK, FEED, KDEF et JAFFE.

TABLE 3.20 Performance de reconnaissance des classifieurs SVM avec différents noyaux et RF à base du descripteur hybride LTP+HOG.

	CK	FEED	KDEF	JAFFE
SVM (Lineair)	**96.06**	92.03	93.34	77.08
SVM (Plynomial)	96.04	**92.46**	**93.36**	**77.5**
SVM (RBF)	91.07	91.24	92.66	77
RF	87.72	79.86	90.06	72

Les comparaisons résumées dans la table 3.21 montrent que les descripteurs uniformes peuvent améliorer les performances. Plus important encore, l'avantage de ces descripteurs uniformes réside dans la réduction de leur taille, sans perte d'informations, et par la suite ils permettent d'accélérer le temps de calcul.

TABLE 3.21 Comparaisons entre les descripteurs hybride LBP/LTP+HOG et LBP_u/LTP_u+HOG en utilisant SVM avec un noyau linéaire.

	LBP+HOG	LBP_u+HOG	LTP+HOG	LTP_u+HOG
CK	93.75	93.18	96.06	**96.48**
FEED	89.89	83.76	**92.03**	91.61
KDEF	92.2	93.22	93.34	**94.32**
JAFFE	73.68	74.96	77.08	**77.92**

3.8.5 Evaluation des bases de données croisées

Nous avons évalué la capacité de généralisation de notre méthode à travers différentes bases de données en effectuant neuf expériences. Dans chaque expérience, nous avons effectué l'apprentissage en utilisant un jeu de données et nous avons fait le test sur les trois autres jeux de données (voir Table 3.22). Comme nous pouvons le voir dans la Table 3.22, notre méthode peut aboutir à des résultats encourageants. En particulier, lorsque le modèle est entraîné en utilisant le jeu de données KDEF (émotions posées), les résultats sur les trois autres jeux de données (émotions spontanées ou posées) sont très intéressants. Cela permet de prétendre que l'entrainement du modèle avec l'ensemble de données KDEF est utile pour reconnaître les émotions spontanées et celles posées. Nous pouvons également voir que le

modèle se comporte relativement bien lorsqu'il est entraîné et testé en utilisant des émotions posées (cas CK/KDEF et KDEF/CK).

TABLE 3.22 Performance avec des bases de données croisées sur les ensembles de données CK, KDEF, FEED et JAFFE.

Apprentissage	CK			KDEF			FEED		
Test	FEED	KDEF	JAFFE	CK	FEED	JAFFE	CK	KDEF	JAFFE
gTR	68.41	79.28	51.17	78.85	79.52	50.7	58.36	67.85	43.66
F-score	70.41	79.35	49.02	77.14	74.17	57.01	57.83	70.04	51.89

3.9 Conclusion

Dans ce chapitre, nous avons présenté une étude empirique complète de l'enregistrement du visage en considérant différentes décompositions faciales d'une part et d'autre part la représentation faciale basée sur la texture en utilisant LBP et ses variantes ainsi que la forme en utilisant HOG et leur combinaison. Les questions clés peuvent être résumées comme suit :

1. La représentation basée sur des régions locales fournit un meilleur enregistrement de visage contrairement à la représentation holistique. En effet, dans la méthode locale (la décomposition du visage), la ROI définie contient toujours une seule composante faciale quelque soit la forme et l'expression du visage. En revanche, dans la représentation holistique, la ROI, à partir de laquelle les descripteurs de caractéristiques sont extraits, peut contenir plusieurs composantes faciales en fonction de l'expression et la forme du visage (voir Figure 3.1).
2. L'évaluation des descripteurs a démontré que les hybrides construits par une concaténation hétérogène à partir des caractéristiques de texture et de forme sont les meilleurs, en particulier la concaténation de LTP et du HOG.
3. l'utilisation des descripteurs uniformes produit une meilleure précision par rapport aux autres descripteurs. Les motifs uniformes suppriment les estimations bruitées dans l'image en les accumulant dans un bin d'histogramme, augmentant ainsi la précision de la reconnaissance. Ces descripteurs se caractérisent par une taille réduite et par la suite ils permettent d'accélérer le temps de calcul.
4. Après plusieurs expériences avec les noyaux polynomial, RBF et linéaire, nous avons retenu ce dernier pour ses performances de classification supérieures et pour éviter la sensibilité des paramètres.

Toutes les expériences de reconnaissance sont effectuées sur des images statiques sans tenir compte de la dynamique des expressions faciales. Dans le chapitre suivant, nous présentons

des méthodes pour capturer et représenter l'expression faciale en exploitant des images issues de la multi-observation (sous-ensemble d'images, vidéo,etc.).

Les travaux présentés dans ce chapitre sont publiés dans le journal "Signal processing : image communication" [118] et "International Conference on Advanced Technologies for Signal and Image Processing" [119].

Chapitre 4

Reconnaissance d'expressions faciales multi-observations basée sur SVM

Sommaire

4.1 Introduction

Le présent chapitre traite le problème de la reconnaissance des expressions faciales (REF) à partir de séquences vidéos. Nous proposons un schéma dynamique de reconnaissance d'expression faciale. Nous fournissons des évaluations de performances en utilisant des

séquences vidéos de test avec un nombre d'observations varié. Le chapitre précédent était consacré à la reconnaissance de l'expression faciale dans des images statiques. A cet effet, de nombreuses techniques ont été appliquées : SDM (Supervised Descent Method) pour la localisation des composantes faciales [231], le descripteur hybride "LTP+HOG" pour l'extraction des caractéristiques et SVM pour la classification. Une limitation très importante de cette stratégie est le fait que les images statiques captent habituellement le pic de l'expression, c'est-à-dire, l'instant auquel les indicateurs d'émotion sont les plus marqués. Plus récemment, l'attention s'est déplacée particulièrement vers la modélisation des expressions faciales dynamiques. Cela est dû au fait que les différences entre les expressions sont modélisées plus puissamment par des transitions dynamiques entre les différentes étapes d'une expression plutôt que par leurs images clés statiques correspondantes. Les expressions faciales sont naturellement dynamiques et peuvent être segmentées en quatre segments temporels : neutre, onset, apex et offset [65, 64]. Neutre signifie qu'aucune expression n'est affichée, l'onset est l'instant où la contraction musculaire se produit et augmente en intensité, l'apex est le pic de l'expression, et offset est l'instant lorsque l'expression commence à disparaître. La dynamique des expressions faciales constitue l'information cruciale requise pour interpréter le comportement facial [169].

Contrairement aux travaux existants qui ont utilisé des méthodes plus ou moins complexes [183, 97, 77, 63, 100, 78, 255, 98, 184, 6, 112], nous proposons dans ce travail un système dynamique simple et efficace, basé uniquement sur les probabilités a posteriori produites par le classifieur SVM, afin de classer une multi-observation représentant une émotion dans un ensemble d'images. Les résultats montrent que le système proposé peut être considéré comme une méthode puissante pour reconnaître les expressions faciales à partir de séquences d'images.

Les performances de notre système sont évaluées sur cinq bases de données disponibles publiquement : CK, CK+, KDEF, FEED, et Oulu-CASIA. Une description plus détaillée des jeux de données utilisés peut être trouvée dans la section 4.3.1. Les expériences montrent que notre méthode de reconnaissance des expressions faciales à partir des vidéos en utilisant la multi-observation surpasse de manière significative les approches existantes.

4.2 Méthodologie proposée

Cette section décrit les détails de la méthodologie globale pour l'approche de REF proposée.

La méthodologie proposée vise à reconnaître sept expressions faciales de base (joie, tristesse, surprise, colère, peur, dégoût, mépris) outre la neutralité à partir d'une multi-

observation qui peut correspondre à un groupe de personnes (chaque personne représente une observation) ou une personne (séquence d'images ou des images de différents points de vue de la même personne). L'approche proposée est largement évaluée sur les bases de données CK, CK+, KDEF, FEED et Oulu-CASIA suivant différents contextes d'expérimentation. La première expérience, appliquée sur CK+ et Oulu-CASIA, consiste à utiliser un ensemble de test contenant des séquences vidéos commençant par la phase neutre et terminant par la phase apex d'une personne en considérant différents nombres d'observations (de 1 à 7). Les résultats obtenus par cette expérience sont comparés avec plusieurs méthodes dynamiques de REF de la littérature. La deuxième expérience, appliquée sur KDEF, prend en compte les observations de test de la même personne avec différents angles de vue en utilisant trois nombres d'observations (1 à 3). La dernière expérience, appliquée sur CK, KDEF et FEED, montre l'impact de la taille des ensembles apprentissage/test variant de 10% à 90% et le nombre d'observations (variant entre 1 et 9). Dans cette expérience, l'ensemble de test contient des personnes différentes. Les évaluations sont effectuées en utilisant un protocole indépendant de la personne (person-independent), dans lequel les images d'une personne ne peuvent apparaître que dans un ensemble d'apprentissage ou un ensemble de test.

La figure 4.1 décrit les différentes étapes du système proposé. D'abord, selon le contexte d'expérimentation, l'entrée du système de REF proposé peut être soit une séquence d'images (onset–>apex), soit un ensemble d'images (apex). La première étape détecte le visage en utilisant l'algorithme VJ [217] et les points faciaux en appliquant la méthode SDM [231], puis la seconde étape extrait des régions spécifiques de visage, plus précisément ses composantes (les sourcils, entre-sourcils, les yeux, le nez et la bouche). La troisième étape caractérise chaque ROI en utilisant un descripteur hybride (LTP_u + HOG) comme cela a été fait dans le système précédemment proposé (voir le chapitre 3 pour plus de détails). Après l'extraction des ROIs et leurs caractéristiques, ces dernières sont ensuite partitionnées en ensembles d'apprentissage et de test. Ces ensembles d'apprentissage et de test contiennent des vecteurs de caractéristiques de différentes expressions faciales. Durant l'apprentissage, si l'entrée est une séquence d'images, nous choisissons seulement les vecteurs de caractéristiques apex pour construire le modèle en utilisant un classifieur SVM basé sur des estimations de probabilité. Une fois le modèle entraîné, durant le test et selon le contexte d'expérimentation, un ensemble de test sera utilisé en appliquant l'une des stratégies présentées dans la section 4.2.3, pour attribuer une expression à une multi-observation d'entrée (sous-ensemble d'images).

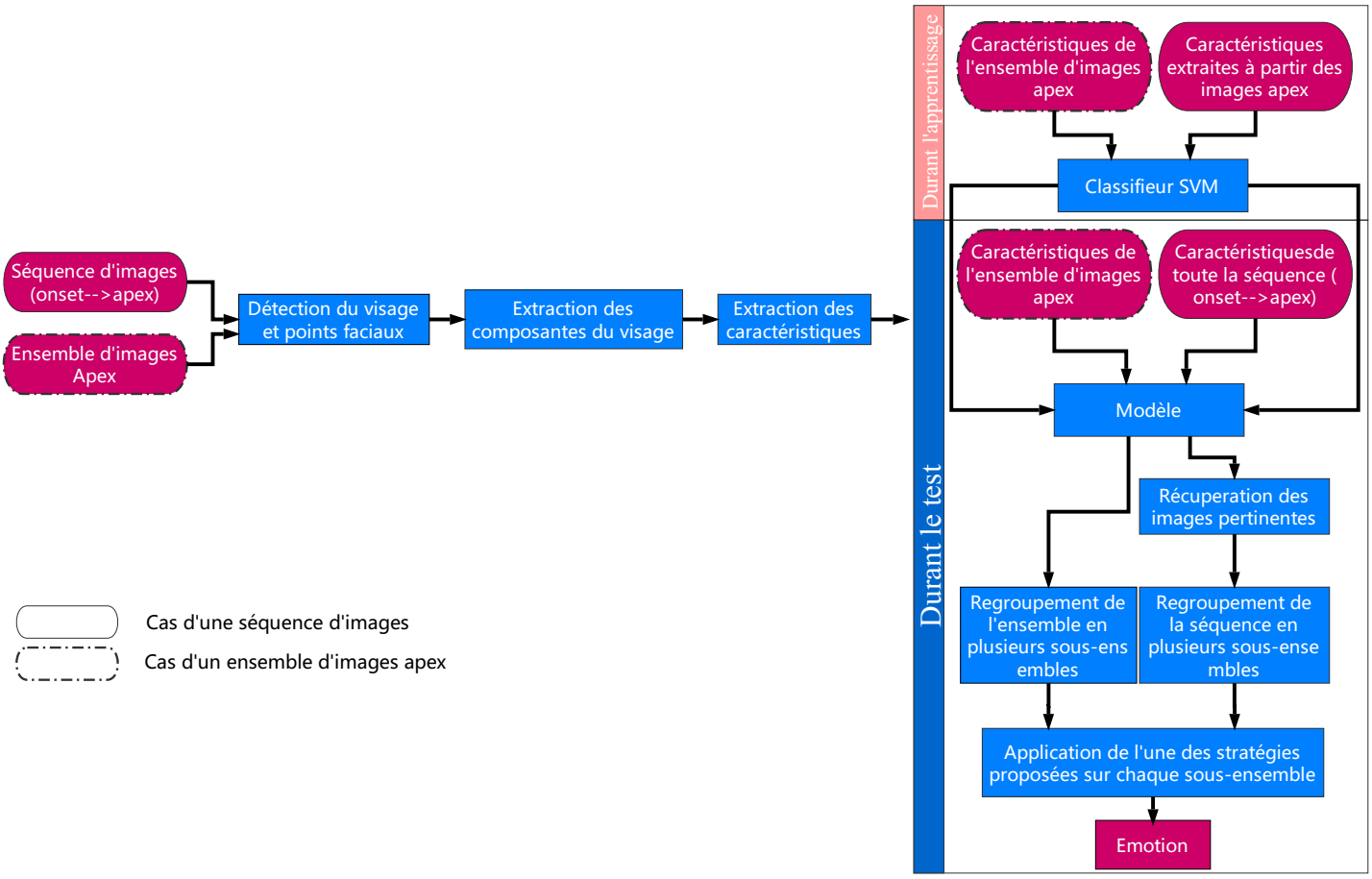


FIGURE 4.1 Architecture du système proposé pour la REF dynamique.

L'approche proposée est basée essentiellement sur des probabilités a postériori. Alors, puisque le classifieur SVM classique, utilisé dans le chapitre précédent, ne prédit que l'étiquette de la classe sans information de probabilité, dans ce chapitre, nous utilisons le SVM étendu qui peut estimer la probabilité [227]. Étant donné k classes de données, pour toute image x de la séquence de test où $x \in \{1,...,t\}$ avec t est le nombre d'images de la séquence de test, nous obtenons la probabilité p_{xj} contre les k classes,

$$p_{xj} = P(y = j|x), j = 1,...,k. \tag{4.1}$$

En se basant sur les probabilités estimées en appliquant SVM [227] sur chaque image x de la séquence de test, nous pouvons extraire un vecteur de décision correspondant à chaque image x. Chaque vecteur de décision contient k probabilités selon le nombre de classes

utilisé. Comme expliqué ci-dessus, cette approche a été évaluée sous différents contextes d'expérimentation : séquences d'images de la même personne, ensemble d'images issus de différents points de vue de la même personne ou ensemble d'images contenant différentes personnes. Le traitement effectué pour construire l'ensemble d'apprentissage et l'ensemble de test dépend du contexte d'expérimentation.

4.2.1 Ensemble d'apprentissage

Traitement à effectuer lors du premier contexte : une séquence d'images de la même personne. Parmi toutes les images de chaque séquence, nous choisissons seulement les images les plus adéquates pour former le modèle d'apprentissage. Bien évidemment, ces images doivent bien représenter les expressions faciales. En effet, selon les créateurs de la base CK+, seulement la dernière image de chaque séquence est codée par un expert en utilisant le système FACS. A cet égard, nous choisissons les trois dernières images de chaque séquence pour constituer l'ensemble d'apprentissage.

Traitement à effectuer lors du deuxième et troisième contextes : ensemble d'images issus de différents points de vue de la même personne ou ensemble d'images contenant différentes personnes. En ce qui concerne le deuxième contexte, le cas multi-vue, les images sont statiques et chaque vue correspond à une image représentant le pic de l'expression. En effet, les images de différents points de vue sont utilisées pour former l'ensemble d'apprentissage.

Le jeu de données considéré dans le troisième contexte est le même que celui utilisé dans le chapitre précédent. Nous rappelons que ce dernier a été consacré à la REF statique, c'est à dire que la construction des jeux de données est basée seulement sur les images apex qui représentent le pic de l'expression.

4.2.2 Ensemble de test

Traitement à effectuer lors du premier contexte : une séquence d'images de la même personne. Étant donnée une séquence de test avec t nombre d'images, nous définissons $V(x)$ comme vecteur de décision correspondant à l'image x. Ce vecteur est composé des sorties du SVM. Chaque valeur de ce vecteur de décision est une probabilité p_{xj} que l'image x appartient à la j^{eme} classe où $x \in \{1,\ldots,t\}$ et $j \in \{1,\ldots,k\}$, k étant le nombre de classes. Sachant que la séquence de test utilisée commence par le segment temporel "neutre" et finit par le segment temporel "apex", les probabilités de la classe dominante auront probablement des valeurs croissantes du premier vecteur de décision au dernier vecteur de la séquence.

En effet, pour chaque vecteur $V(x)$, nous cherchons la probabilité maximale et la classe j^* correspondante. Ensuite, nous calculons la différence entre la valeur de cette probabilité maximale et la moyenne des probabilités correspondant à la même classe j^* des trois derniers vecteurs de la séquence représentant le pic de l'expression. La différence est calculée comme suit :

$$diff(p_x) = p_x - \frac{\sum_{y=t-2}^{t} p_{yj^*}}{3}, \tag{4.2}$$

Nous pouvons aussi faire l'analyse en considérant la différence entre la valeur de la probabilité maximale et le produit des probabilités correspondant à la classe j^* des trois derniers vecteurs de la séquence représentant le pic de l'expression. Cette différence s'exprime comme suit :

$$diff(p_x) = p_x - \prod_{y=t-2}^{t} p_{yj^*}, \tag{4.3}$$

où

$$p_x = max(p_{xj});\ x \in \{1,...,t\}\ et\ j = 1,\ldots,k.$$

Dans les sections 4.3.2 et 4.3.3, nous étudierons l'impact du choix du calcul de la différence $diff(p_x)$ sur la REF. Selon la différence calculée par Eq. 4.2 (ou Eq. 4.3), une décision doit être faite pour garder le vecteur $V(x)$ ou l'éliminer de la séquence de test. Alors, si la différence est inférieure à un certain seuil S (le seuil est défini expérimentalement dans les sections 4.3.2 et 4.3.3), le vecteur $V(x)$ est accepté. Sinon, le vecteur $V(x)$ est rejeté. Dans le dernier cas, nous considérons que la probabilité de la classe dominante j^* a considérablement baissé dans les derniers vecteurs de décision de la séquence de test. Cela signifie que la classe j^* n'est plus considérée comme dominante pour les images représentant le pic de l'expression. Cette étape fournit une information très importante : la classe d'expression la plus probable à laquelle appartient la séquence de test (Voir l'exemple de mise en œuvre Figure 4.8).

Le but final est d'assigner une classe à une multi-observation (sous-ensemble d'images). Soient E_r la nouvelle séquence de test formée grâce aux vecteurs de décisions $V(x)$ sélectionnés en utilisant Eq. 4.2 (ou Eq. 4.3), et L le nombre d'éléments de E_r. Avant de commencer le processus d'affectation d'une classe à la nouvelle séquence E_r, nous allons regrouper les images de cette séquence en plusieurs sous-ensembles. Chaque sous-ensemble contient un nombre R d'observations. Le nombre R d'observations doit être inférieur ou égal à L, sinon il faut alimenter l'ensemble E_r en utilisant les vecteurs de décisions éliminés précédemment. Pour ce faire, soit E_e l'ensemble des vecteurs éliminés, nous trions les vecteurs

de décisions de l'ensemble E_e par ordre croissant des différences calculées par Eq. 4.2 (ou Eq. 4.3). Puis, nous récupérons les $(R-L)$ premiers vecteurs de décisions. Le nombre C de sous-ensembles à considérer dans la suite du processus vaut le nombre de combinaisons de R observations parmi L ($C = C_L^R$). L'utilisation de toutes les combinaisons possibles permet de ne pas manquer des observations et des informations pertinentes favorisant la reconnaissance des émotions. La figure 4.2 présente le logigramme de REF d'une séquence d'images.

Traitement à effectuer lors du deuxième et troisième contextes : ensemble d'images issus de différents points de vue de la même personne ou ensemble d'images contenant différentes personnes. Ce cas correspond au cas de REF statique. Il est à noter que les images constituant l'ensemble de test sont toutes des images apex représentant différents point de vue (même personne) ou différentes personnes (vue frontale) respectivement dans le deuxième et le troisième contexte. Nous précisons que le but derrière le troisième contexte d'expérimentation est de mesurer la performance de l'approche proposée lorsqu'il s'agit de la REF avec plusieurs personnes différentes. Après l'extraction des caractéristiques des images apex et l'obtention du vecteur de décision de chaque image en utilisant le classifieur SVM, les images sont regroupées en sous-ensembles comme expliqué ci-dessus (sans l'étape basée sur le vecteur de décision). Le nombre C de sous-ensembles à considérer dans la suite du processus vaut le nombre de combinaisons de R observations parmi t ($C = C_t^R$), t étant le nombre d'images de l'ensemble initial à tester (pour lequel on cherche à assigner une classe). L'idée derrière l'utilisation de toutes les combinaisons possibles est de considérer toutes les personnes/vues (selon le contexte d'expérimentation utilisé) dans le processus de REF.

Le reste du processus est le même pour les trois contextes d'expérimentation.

4.2.3 Stratégies proposées pour la REF

Étant donné un sous-ensemble d'images avec un nombre R d'observations, la matrice des probabilités peut être définie comme suit :

$$M_p = \overbrace{\begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{R1} & \cdots & p_{Rk} \end{pmatrix}}^{k \text{ classes}} \Bigg\updownarrow R \text{ images} \tag{4.4}$$

où k dénote le nombre d'expressions (nombre de classes) et p_{ij} est la probabilité que la i^{eme} image appartient à la classe j.

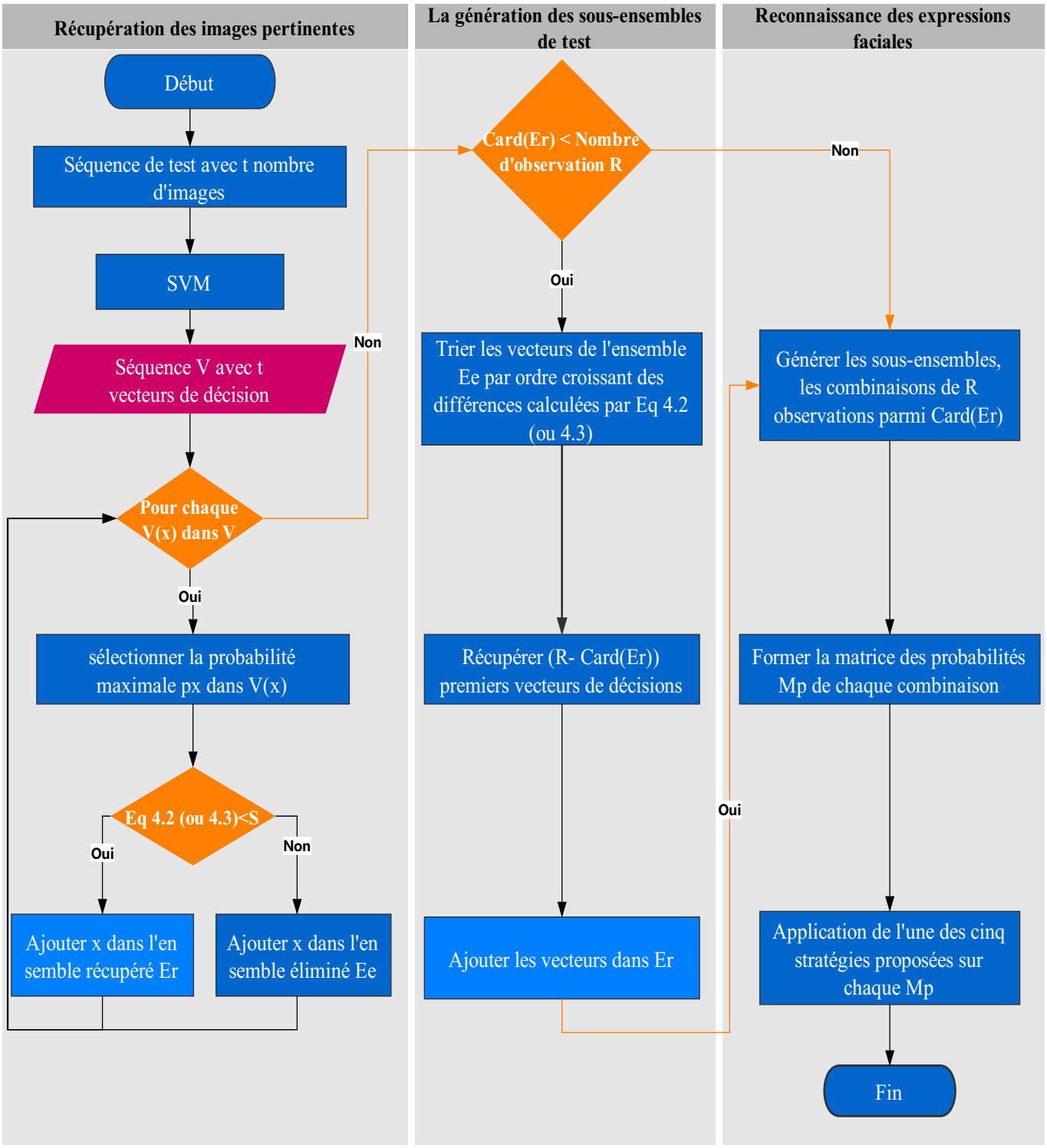


FIGURE 4.2 Logigramme de REF d'une séquence d'images.

Stratégie 1 (dénotée comme S1). La première stratégie est simple. Elle consiste à assigner à chaque sous-ensemble d'images la classe correspondant à la plus grande probabilité. Cela peut être exprimé comme suit :

$$c^* = arg\max(p_{ij});\ i = 1,\ldots,R\ et\ \ j = 1,\ldots,k. \tag{4.5}$$

La figure 4.3 représente le logigramme de la première stratégie.

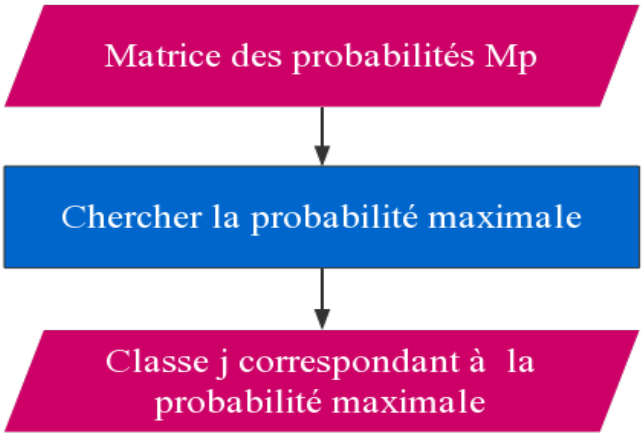


FIGURE 4.3 Logigramme de la stratégie 1.

Stratégie 2 (dénotée comme S2). La seconde stratégie consiste à attribuer au sous-ensemble d'images la classe correspondant à la moyenne maximale des probabilités sur les images du sous-ensemble. Formellement, la classe prédite est donnée comme suit :

$$c^* = arg\max(p_j);\ j = 1,\ldots,k. \tag{4.6}$$

où

$$p_j = \frac{\sum_{i=1}^{R} p_{ij}}{R};\ j = 1,\ldots,k \tag{4.7}$$

La figure 4.4 représente le logigramme de la deuxième stratégie.

Stratégie 3 (dénotée comme S3). La troisième stratégie consiste à assigner au sous-ensemble d'images la classe dominante avec le vote le plus élevé, c-à-d, la classe prédite est obtenue en se basant sur le vote majoritaire comme le montre Eq. 4.8. Dans le cas où il y a plus d'une classe dominante, la décision est prise en utilisant Eq. 4.6 appliquée sur les

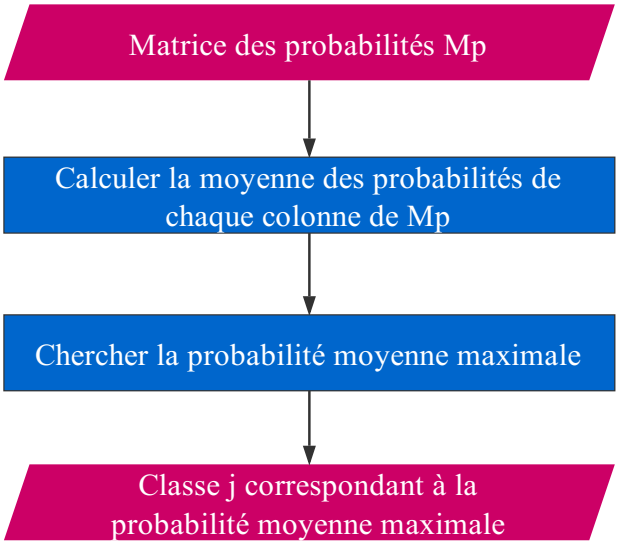


FIGURE 4.4 Logigramme de la stratégie 2.

classes dominantes avec uniquement les probabilités ayant été impliquées dans ce vote. Etant donnée une classe dominante k, on ne prend en compte que les probabilités maximales de la colonne k de M_p, obtenues lors de la recherche des probabilités maximales dans les lignes de M_p.

$$c^* = arg\max(Card(V_c));\ c = 1,\ldots,k \tag{4.8}$$

où

$$V_c = \{i/\forall j, j \neq c, p_{ic} > p_{ij}\};\ i = 1,\ldots,R\ et\ j = 1,\ldots,k. \tag{4.9}$$

La figure 4.5 représente le logigramme de la troisième stratégie.

Stratégie 4 (dénotée comme S4). Similaire à S3, la quatrième stratégie consiste à attribuer au sous-ensemble d'images la classe dominante avec le vote le plus élevé en utilisant Eq. 4.8. Dans le cas où il y a plus d'une classe dominante, la décision est prise en utilisant cette fois-ci Eq. 4.5 appliquée sur les classes dominantes. La figure 4.6 représente le logigramme de la quatrième stratégie.

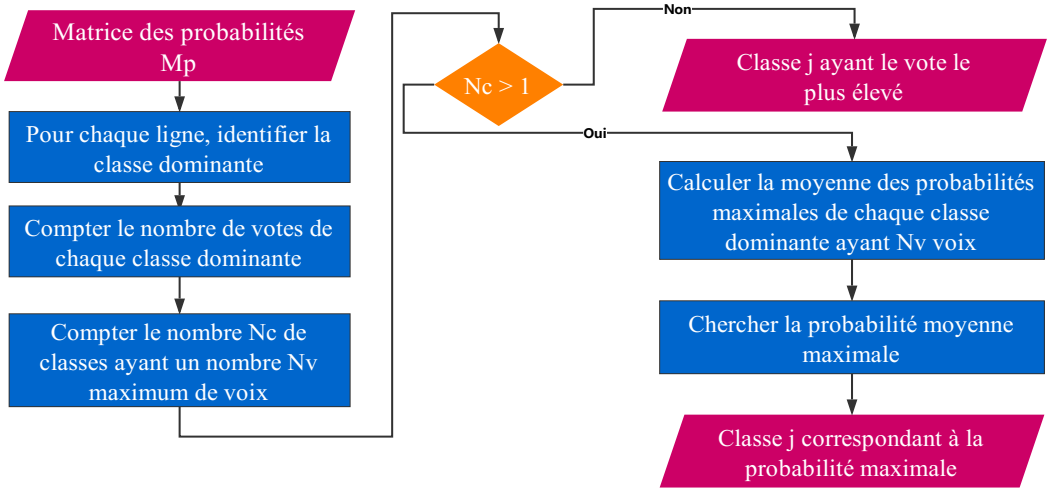


FIGURE 4.5 Logigramme de la stratégie 3.

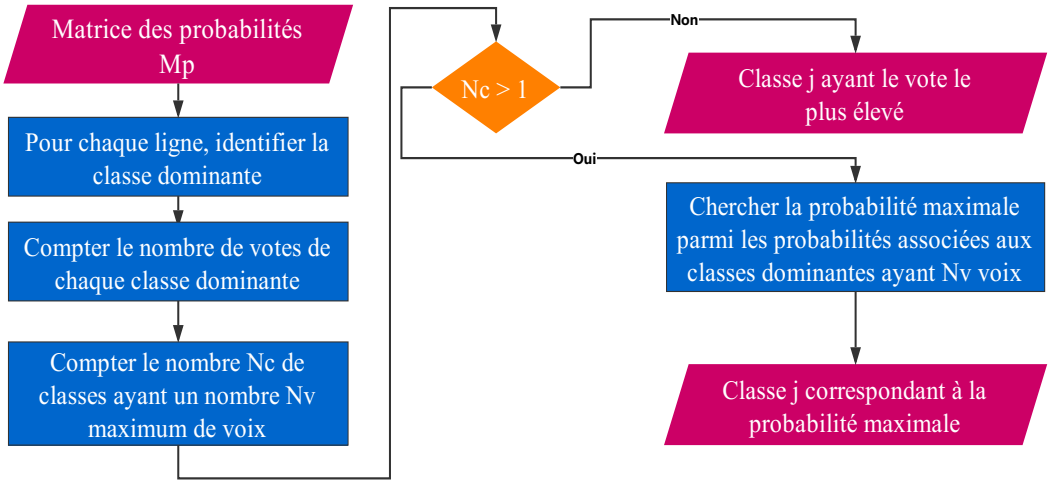


FIGURE 4.6 Logigramme de la stratégie 4.

Stratégie 5 (dénotée comme S5). La cinquième stratégie consiste à sélectionner la classe ayant le produit des probabilités maximum. Formellement, elle est donnée comme suit :

$$c^* = arg\max(p_j);\ j = 1,\ldots,k. \tag{4.10}$$

où

$$p_j = \prod_{i=1}^{R} p_{ij};\ j = 1,\ldots,k \tag{4.11}$$

La figure 4.7 représente le logigramme de la cinquième stratégie. L'implémentation de cette stratégie a été réalisée dans la période de rédaction du manuscrit. Ainsi, nous n'avons malheureusement pas, par manque de temps, analyser les résultats dans le cadre des deuxième et troisième contextes d'expérimentation (voir Sections 4.3.5.2 et 4.3.6). Par conséquent, cette stratégie a été évaluée uniquement dans le cadre du premier contexte (voir Sections 4.3.2 et 4.3.3)

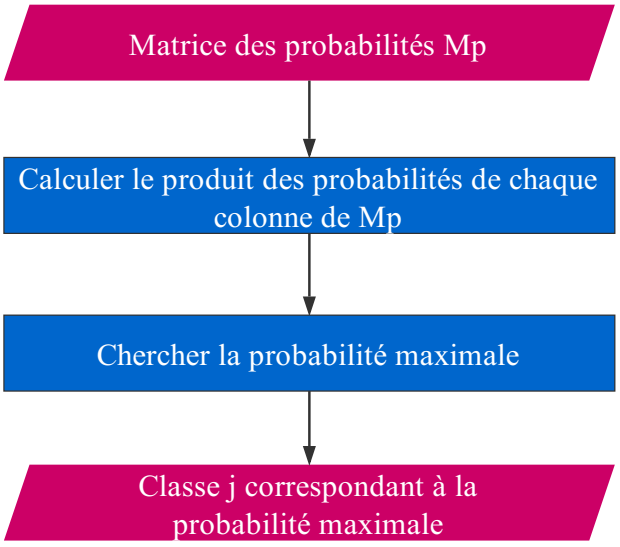


FIGURE 4.7 Logigramme de la stratégie 5.

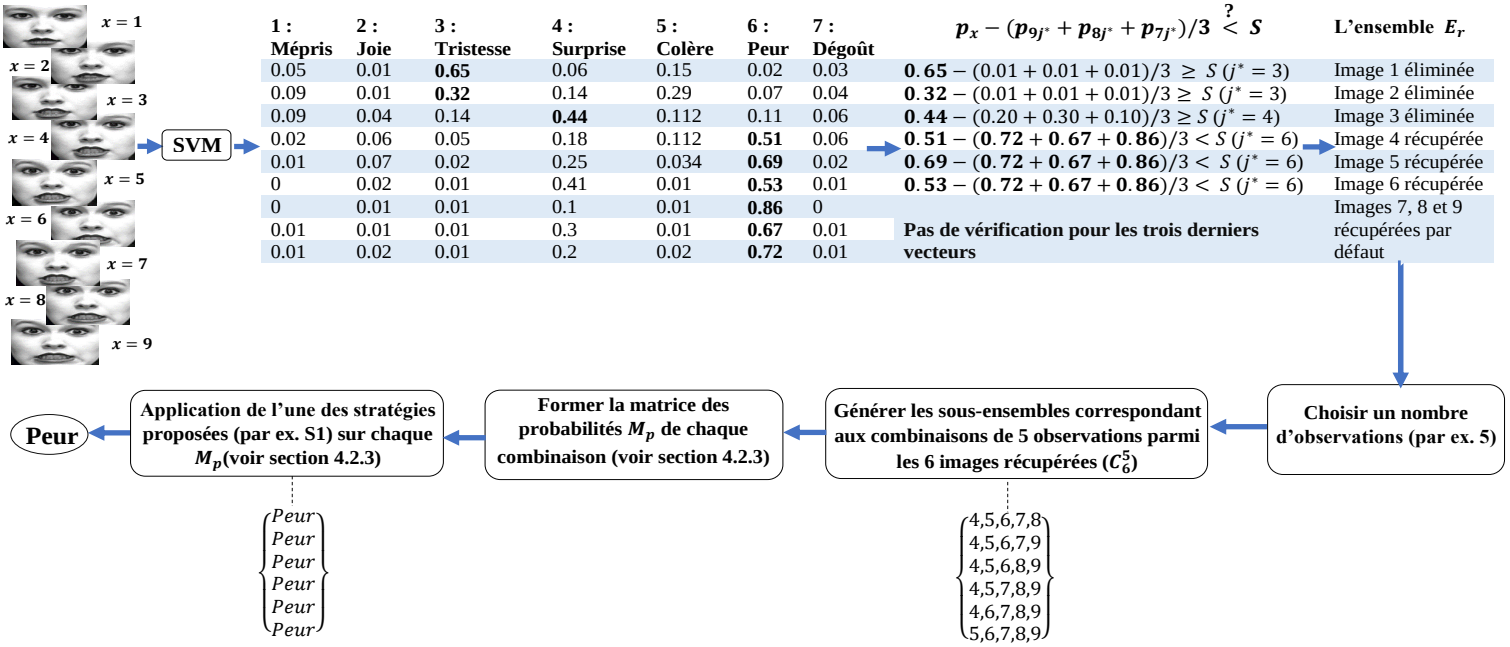


FIGURE 4.8 Exemple de mise en œuvre de la méthode proposée.

4.3 Expérimentations

4.3.1 Bases de données et protocole d'expérimentation

Nous évaluons la méthode proposée en utilisant des images faciales issues de cinq bases de données accessibles publiquement : CK [101], CK+ [132], Oulu-CASIA [253], KDEF [134] et FEED [220].

4.3.1.1 CK+

327 séquences ont été sélectionnées à partir de la base de données CK+. En effet, les étiquettes d'expression validées ne sont attribuées qu'à 327 séquences répondant aux critères d'une des sept émotions de base (colère, mépris, dégoût, peur, joie, tristesse et surprise), basés sur le système FACS (Facial Action Coding System). Les séquences sélectionnées appartiennent à 118 sujets avec les sept expressions de base. Les séquences vidéo contiennent des images de la phase "neutre" jusqu'à la phase "apex" des expressions faciales. Quant à la taille des séquences, elle varie de 10 à 60 images.

Deux expériences ont été menées pour évaluer les performances de reconnaissance de la méthode proposée. Elles ont été effectuées en utilisant le protocole indépendant de la personne (person-independent). La première expérience adopte le protocole validation croisée de 10-fold (10-fold CV). Plus précisément, les 118 sujets concernés par les séquences sélectionnées ont été séparés en dix groupes, chaque groupe ayant à peu près le même nombre de sujets. Ensuite, neuf groupes de sujets ont été utilisés comme ensemble d'apprentissage, tandis que le groupe restant a servi d'ensemble de test. Ce processus a été répété jusqu'à ce que chaque groupe ait été servi comme ensemble de test une fois. La deuxième expérience utilise le protocole validation croisée de leave-one-subject-out (LOSOCV), c-à-d 118-fold, où chaque ensemble de test consistait en des données provenant d'un seul sujet.

4.3.1.2 Oulu-CASIA

Cette base comprend 480 séquences d'images de 80 sujets, prises dans des conditions d'éclairage normales (c-à-d, un éclairage fort et bon). Les séquences sont étiquetées avec l'un des six émotions de base (colère, dégoût, peur, joie, tristesse et surprise). Chaque séquence commence par une expression faciale neutre et se termine par le pic de l'expression, comme dans le cas de la base CK+. La taille des séquences varie de 9 à 72 images.

De manière similaire à [253, 76], nous avons adopté le protocole 10-fold CV indépendant de la personne sur les 480 séquences. La tâche d'évaluation consiste à prédire la classe d'un échantillon de test qui n'appartient pas à l'ensemble d'apprentissage.

4.3.1.3 KDEF

Nous avons effectué deux expériences sur l'ensemble de données KDEF. La première expérience consiste à reconnaitre l'expression faciale multi-vue d'une seule personne. Comme expliqué dans la section 3.4 du chapitre 3, l'algorithme VJ de détection de visage ne peut pas détecter tous les visages pris sous des angles de vue différents de 0°. Par conséquent, nous avons utilisé le mode interactif de la méthode SDM (voir Section 3.4) pour assurer la précision de la détection du visage (c-à-d, nous avons procédé à la détection manuelle du visage). Après cette étape, nous appliquons SDM pour détecter les 49 points faciaux. SDM n'a pas pu détecter ces points sur tous les visages pris sous différents angles de la base KDEF. Pour cette raison, nous n'avons considéré que les images avec les angles de vue $-45°$, 0° et 45° dans lesquels les 49 points caractéristiques sont détectés par SDM [231]. Nous ferons référence à l'ensemble de données KDEF dans cette première expérience comme KDEF_MV. Le protocole LOSOCV (68-fold) est réalisé sur 1224 images de 68 sujets afin de reconnaître des expressions faciales multi-vues (MVFE) d'une seule personne. Dans la seconde expérience, comme dans le précédent chapitre, nous utilisons le protocole 10-fold

CV indépendant de la personne sur un ensemble de données de 280 images représentant 40 sujets avec les six expressions faciales de base, ainsi que l'état neutre afin de reconnaître l'expression faciale d'une multi-observation d'un groupe de personnes.

4.3.1.4 CK

Pour évaluer notre méthode sur la base de données CK, le protocole 10-fold CV indépendant de la personne est adopté sur un total de 98 séquences appartenant à 32 sujets, configuration similaire à celle du chapitre précédent.

4.3.1.5 FEED

En procédant de manière identique à celle décrite au chapitre précédent, le protocole 10-fold CV indépendant de la personne est considéré avec un total de 81 séquences d'images de 16 sujets, représentant les six expressions faciales universelles, ainsi que l'état neutre.

Les détails des bases de données CK/CK+, KDEF, Oulu-CASIA, FEED sont présentés respectivement dans les sections 2.3.1, 2.3.2, 2.3.4 et 2.3.5 du chapitre 2.

4.3.2 Expérience sur la base de données Cohn-Kanade étendu (CK+)

Dans cette section, nous présentons les expériences menées sur le jeu de données CK+. Ces expériences ont été réalisées sur des séquences qui commencent par un visage neutre et se terminent par un visage expressif, comme dans plusieurs travaux de la littérature [132, 184, 6, 112]. Dans notre travail, l'ensemble de test est regroupé en séquences de différentes tailles (de 1 à 7), c'est à dire différents nombres d'observations, comme expliqué dans la section 4.2.2. Rappelons que la taille $R = 1$ fait référence à la reconnaissance mono-observation (c-à-d, la REF statique).

Les cinq stratégies proposées dans la section 4.2.3 se sont comparées en faisant varier le nombre d'observations R entre 1 et 7, sous différents seuils S = 0, 0.01 et 0.1. Les résultats d'évaluation sont présentés dans les tables 4.1 et 4.2 en utilisant le protocole LOSOCV, ainsi que dans les tables 4.3 et 4.4 en appliquant le protocole 10-fold CV. Pour une comparaison approfondie notamment celle relative à la construction de l'ensemble de test E_r, les résultats des tables 4.1 et 4.3 ont été obtenus avec Eq. 4.2, et les résultats présentés dans les tables 4.2 et 4.4 ont été obtenus en utilisant Eq. 4.3.

A partir des tables 4.1, 4.2, 4.3 et 4.4, nous pouvons voir que les meilleurs résultats ont été obtenus en utilisant Eq. 4.3 et cela quelque soit le protocole utilisé (LOSOCV ou 10-fold CV). Rappelons que cette équation est basée sur le produit des probabilités au lieu la moyenne des probabilités utilisée dans Eq. 4.2. Nous pouvons aussi observer à partir de la table 4.2 lorsque

$R = 1$ (c-à-d, mono-observation) et $S = 0.01$, le taux de reconnaissance chute à 97.03%. Lorsque $R > 1$, les informations d'évolution temporelle ont très bien exploitées et le taux de reconnaissance augmente et vaut 99.68% en $R = 4$, ce qui indique l'importance d'incorporer l'apparence de l'image, à travers le descripteur LTP_u+HOG, avec l'information temporelle (multi-observation). Comme nous pouvons le voir sur la table 4.4 où le protocole 10-fold CV est appliqué, plus la séquence de test est longue, plus le taux de reconnaissance est élevé. Le taux de reconnaissance pour $R = 1$ vaut 97.73% avec $S = 0.01$ comme dans le protocole LOSOCV. Ce taux augmente avec l'augmentation de R, et atteint le taux maximal de 99.99% en $R >= 6$.

A partir des tables 4.1, 4.2, 4.3 et 4.4, nous pouvons constater que l'utilisation d'une stratégie ou une autre influe peu sur les résultats qui restent presque identiques quelque soit la stratégie utilisée.

TABLE 4.1 Taux de reconnaissance obtenus en utilisant Eq. 4.2, avec différents nombres d'observations (R) en appliquant le protocole LOSOCV sur le jeu de données CK+.

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	96.75	98.28	98.98	**99.09**	99.05	98.96	98.91
	S2	96.75	98.26	98.98	**99.09**	99.06	98.11	98.05
	S3	96.75	98.28	98.12	**99.1**	98.21	98.11	98.06
	S4	96.75	98.28	98.12	**99.09**	98.21	98.11	98.06
	S5	96.75	98.23	98.96	**99.08**	99.06	98.96	98.05
0.01	S1	96.84	98.3	98.95	99.08	99.13	99.14	**99.2**
	S2	96.84	98.27	98.93	99.07	**99.12**	98.29	98.35
	S3	96.84	98.3	98.09	**99.09**	98.28	98.3	98.35
	S4	96.84	98.3	98.09	**99.08**	98.28	98.29	98.35
	S5	96.84	98.25	98.91	99.05	99.11	**99.13**	98.34
0.1	S1	96.18	97.29	97.82	98.21	98.51	98.76	**98.84**
	S2	96.18	97.21	97.7	97.96	98.23	97.78	97.86
	S3	96.18	97.29	97.44	97.97	97.7	97.89	97.91
	S4	96.18	97.29	97.44	97.95	97.7	97.88	97.91
	S5	96.18	97.19	97.67	98.02	98.15	98.51	97.72

4.3.2.1 Matrices de confusion

Cette section analyse les résultats à travers des matrices de confusion. Les tables 4.5, 4.6 affichent les matrices de confusion du jeu de données CK+ respectivement pour les protocoles LOSOCV et 10-fold CV. Nous pouvons observer que de très bons résultats ont été obtenus pour toutes les émotions (le taux de reconnaissance de chaque expression est supérieur à 98%). Nous pouvons voir aussi que la confusion d'expressions est présente, avec

TABLE 4.2 Taux de reconnaissance obtenus en utilisant Eq. 4.3, avec différents nombres d'observations (R) en appliquant le protocole LOSOCV sur le jeu de données CK+.

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	96.79	98.45	**99.26**	99.21	98.87	98.69	98.15
	S2	96.79	98.47	**99.33**	99.31	99.16	98.22	97.32
	S3	96.79	98.45	98.47	**99.31**	98.25	98.01	97.32
	S4	96.79	98.45	98.47	**99.31**	98.25	98.01	97.32
	S5	96.79	98.47	**99.33**	99.31	99.16	98.86	97.53
0.01	S1	97.03	98.78	99.59	**99.68**	99.62	99.54	99.12
	S2	97.03	98.79	99.59	**99.68**	99.61	98.67	98.19
	S3	97.03	98.78	98.74	**99.68**	98.76	98.67	98.2
	S4	97.03	98.78	98.74	**99.68**	98.76	98.67	98.2
	S5	97.03	98.79	99.59	**99.68**	99.61	99.31	98.19
0.1	S1	97.27	98.78	**99.31**	99.29	99.24	99.14	98.91
	S2	97.27	98.79	**99.32**	99.3	99.24	98.29	98.07
	S3	97.27	98.78	98.47	**99.3**	98.39	98.29	98.07
	S4	97.27	98.78	98.47	**99.3**	98.39	98.29	98.07
	S5	97.27	98.79	**99.32**	99.3	99.24	98.93	98.07

TABLE 4.3 Taux de reconnaissance obtenus en utilisant Eq. 4.2, avec différents nombres d'observations (R) en appliquant le protocole 10-fold CV sur le jeu de données CK+.

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	97.59	98.95	99.56	99.82	99.93	99.98	**99.99**
	S2	97.59	98.93	99.54	99.81	99.93	99.98	**99.99**
	S3	97.59	98.93	99.53	99.8	99.93	99.98	**99.99**
	S4	97.59	98.93	99.53	99.8	99.93	99.98	**99.99**
	S5	97.59	98.95	99.56	99.82	99.93	99.98	**99.99**
0.01	S1	97.73	99.1	99.7	99.91	99.97	99.99	**99.99**
	S2	97.73	99.09	99.69	99.9	99.97	99.99	**99.99**
	S3	97.73	99.1	99.69	99.9	99.97	99.99	**99.99**
	S4	97.73	99.1	99.69	99.9	99.97	99.99	**99.99**
	S5	97.73	99.1	99.7	99.9	99.97	99.99	**99.99**
0.1	S1	96.11	97.58	98.68	99.36	99.71	99.87	**99.93**
	S2	96.11	97.47	98.53	99.22	99.6	99.8	**99.93**
	S3	96.11	97.47	98.51	99.22	99.63	99.81	**99.93**
	S4	96.11	97.47	98.51	99.22	99.63	99.81	**99.93**
	S5	96.11	97.58	98.67	99.37	99.7	99.84	**99.93**

TABLE 4.4 Taux de reconnaissance obtenus en utilisant Eq. 4.3, avec différents nombres d'observations (R) en appliquant le protocole 10-fold CV sur le jeu de données CK+.

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	97.89	99.41	99.92	99.96	**99.96**	99.95	99.92
	S2	97.89	99.43	99.93	99.96	99.99	99.99	**99.99**
	S3	97.89	99.42	99.91	99.96	99.97	99.96	**99.99**
	S4	97.89	99.42	99.91	99.96	99.97	99.96	**99.99**
	S5	97.89	99.43	99.92	99.96	99.99	99.99	**99.99**
0.01	S1	98.04	99.5	99.94	99.97	**99.97**	99.96	99.93
	S2	98.04	99.52	99.94	99.97	99.99	99.99	**99.99**
	S3	98.04	99.51	99.94	99.97	99.99	99.99	**99.99**
	S4	98.04	99.51	99.94	99.97	99.99	99.99	**99.99**
	S5	98.04	99.52	99.94	99.97	99.97	99.99	**99.99**
0.1	S1	98.51	99.74	99.98	99.99	99.99	99.99	**99.99**
	S2	98.51	99.75	99.98	99.99	99.99	99.99	**99.99**
	S3	98.51	99.75	99.98	99.99	99.99	99.99	**99.99**
	S4	98.51	99.75	99.98	99.99	99.99	99.99	**99.99**
	S5	98.51	99.74	99.98	99.99	99.99	99.99	**99.99**

des pourcentages très faibles. La table 4.6 montre par ailleurs que l'expression mépris est confondue avec les expressions joie, colère et dégoût avec respectivement un taux d'erreur de 0.23%, 0.48% et 0.97%.

TABLE 4.5 Matrice de confusion pour le jeu de données CK+ en appliquant le protocole LOSOCV (associée à la stratégie S1 avec R=4 et S=0.01)

	Mépris	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Mépris	**99.77**	0	0	0	0.23	0	0
Joie	0	**100**	0	0	0	0	0
Tristesse	0	0	**99.94**	0.02	0.03	0	0.01
Surprise	0	0	0	**100**	0	0	0
Colère	0	0	0	0.001	**99.99**	0	0.001
Peur	0	0.005	0	0.04	0	**99.95**	0
Dégoût	0	0	0	0	0.05	0	**99.95**
Moyenne							**99.94**

TABLE 4.6 Matrice de confusion pour le jeu de données CK+ en appliquant le protocole 10-fold CV (associée à la stratégie S1 avec R=4 et S=0.01)

	Mépris	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Mépris	**98.049**	0.487	0	0	0.487	0	0.975
Joie	0.001	**99.998**	0	0	0	0	0.001
Tristesse	0	0	**99.738**	0	0	0.05	0.209
Surprise	0	0	0	**100**	0	0	0
Colère	0	0	0	0	**99.995**	0	0.004
Peur	0	0.007	0.007	0.015	0	**99.953**	0.015
Dégoût	0	0	0	0	0.002	0	**99.997**
Moyenne							**99.67**

4.3.2.2 Comparaison avec l'état de l'art

La table 4.7 fournit les performances de différents algorithmes récents proposés par Lucey et al. [132], Yongqiang et al. [122], Mengyi et al. [127, 128], Sikka et al. [183], Jung et al. [97], Yimo et al. [77], Xijian et al. [63], Khairuni et al. [100], Gupta et al. [78], Zhao et al. [255], Kacem et al. [98], Sikka et Sharma [184], Allaert et al. [6] et Kuo et al. [112] sur l'ensemble de données CK+ pour effectuer une comparaison relative avec notre algorithme. Les jeux de données utilisés dans ces travaux sont identiques à celui adopté dans notre étude, sauf dans [122, 100, 255] où les jeux de données considèrent uniquement les six émotions de base (Joie, Tristesse, Surprise, Colère, Peur, Dégoût). Par conséquent, les résultats présentés dans ces travaux de l'état de l'art peuvent être directement comparés à notre méthode. Les méthodes proposées par [132, 122, 127, 183, 77, 128, 255] ont utilisé la méthode de validation croisée LOSO, la méthode proposée par [63] a utilisé la validation croisée Leave-One-Sequence-Out (LOSeqO) et le reste de méthodes comparées ont utilisé la validation croisée de 10-fold.

Deux mesures de performance ont été considérées pour une comparaison équitable : la première est la moyenne des taux de reconnaissance de sept (ou six) classes (désignée par mTR). La deuxième est le taux de reconnaissance global obtenu sur k-fold CV (désignée par gTR).

Sur la base des résultats présentés dans la table 4.7, la méthode proposée atteint un taux de reconnaissance moyen de 99.94% et un taux global de 99.68% en utilisant le protocole LOSO et atteint un taux de reconnaissance global de 99.99% en appliquant le protocole 10-fold pour les sept émotions de base. Ceci montre que notre méthode est plus efficace que les méthodes récentes de la littérature, mentionnées plus haut. Rappelons que ces travaux emploient des méthodes très complexes, comme la combinaison des caractéristiques

d'apparence et géométriques en utilisant deux réseaux profonds [97], l'utilisation de modèles CNN [112, 78], la méthode Exemplar-HMM [183] et l'utilisation de descripteur spatio-temporel [6], en comparaison avec notre méthode, basée uniquement sur des techniques simples comme le descripteur hybride LTP_u+HOG pour l'extraction des caractéristiques et le classifieur SVM pour la classification.

TABLE 4.7 Performances obtenues par notre méthode et les méthodes récentes de la littérature sur le jeu de données CK+.

Méthode	Année	mTR	gTR	Catégorie	Protocole	# Séquences	# émotions
Lucey et al. [132]	2010	83.32	–	Traditionnelle	LOSOCV	327	7
Yongqiang et al. [122]	2013	85.27	87.43	Traditionnelle	LOSOCV	305	6
Mengyi et al. [127]	2014	–	94.19	Traditionnelle	LOSOCV	327	7
Sikka et al. [183]	2015	94.6	–	Traditionnelle	LOSOCV	327	7
Jung et al. [97]	2015	95.21	97.25	Deep learning	10-fold CV	327	7
Yimo et al. [77]	2016	96.78	97.2	Traditionnelle	LOSOCV	325	7
Mengyi et al. [128]	2016	93.85	95.1	Traditionnelle	LOSOCV	327	7
Xijian et al. [63]	2017	89.35	88.3	Traditionnelle	LOSeqOCV	327	7
Khairuni et al. [100]	2017	70.26	82.4	Traditionnelle	10-fold CV	222	6
Gupta et al. [78]	2017	66	94.18	Deep learning	10-fold CV	327	7
Zhao et al. [255]	2017	95.8	–	Traditionnelle	LOSOCV	309	6
Kacem et al. [98]	2017	95.1	96.87	Traditionnelle	10-fold CV	327	7
Sikka et Sharma [184]	2018	95.1	–	Traditionnelle	10-fold CV	327	7
Allaert et al. [6]	2018	–	97.25	Traditionnelle	10-fold CV	327	7
Kuo et al. [112]	2018	–	98.47	Deep learning	10-fold CV	327	7
Méthode proposée (S1)		<u>99.94</u>	<u>99.68</u>	Traditionnelle	LOSOCV	327	7
Méthode proposée (S1)		**99.67**	**99.99**	Traditionnelle	10-fold CV	327	7

4.3.3 Expérience sur la base de données Oulu-CASIA

Toujours dans l'objectif d'étudier l'efficacité de notre méthode, la deuxième expérience est menée sur le jeu de données Oulu-CASIA qui est plus difficile que le jeu de données CK+. En effet, le jeu de données Oulu-CASIA contient davantage d'expressions de faible intensité qui sont difficiles à distinguer avec une faible résolution des images. La table 4.8 montre les résultats obtenus en utilisant Eq. 4.2 sous différentes valeurs de R et S et la table 4.9 présente ceux obtenus en utilisant Eq. 4.3 sous différentes valeurs de R et S.

A partir des tables 4.8 et 4.9, nous pouvons constater que les meilleures performances ont été obtenues en utilisant l'équation basée sur le produit des probabilités (Eq. 4.3). Cette conclusion est identique à celle obtenue dans la section 4.3.2 pour le jeu de données CK+.

Nous pouvons apprécier que plus la séquence de test est longue, plus le taux de reconnaissance est élevé. Nous pouvons également observer qu'en augmentant le nombre d'images testées, la reconnaissance de l'ensemble de test augmente de manière significative de 84.35% pour $R = 1$ à 94.2% pour $R = 7$ avec un $S = 0$. Le paramètre R est très important dans notre méthode, car il impacte la construction des sous-ensembles de test. Intuitivement et expérimentalement, plus la valeur de R est grande, plus précis est le processus d'évolution de l'expression faciale dans la séquence.

TABLE 4.8 Taux de reconnaissance obtenus en utilisant Eq. 4.2, avec différents nombres d'observations (R) en appliquant le protocole 10-fold CV sur le jeu de données Oulu-CASIA.

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	82.46	85.55	87.65	88.87	89.48	**89.69**	89.63
	S2	82.46	85.54	87.62	88.86	89.48	**89.69**	89.63
	S3	82.46	85.55	87.63	88.86	89.47	**89.69**	89.63
	S4	82.46	85.55	87.63	88.87	89.47	**89.69**	89.63
	S5	82.46	85.51	87.6	88.85	89.48	**89.69**	89.63
0.01	S1	83.25	86.44	88.08	88.79	89.13	89.35	**89.52**
	S2	83.25	86.44	88.06	88.79	89.13	89.35	**89.52**
	S3	83.25	86.44	88.06	88.79	89.13	89.35	**89.52**
	S4	83.25	86.44	88.06	88.79	89.13	89.35	**89.52**
	S5	83.25	86.44	88.08	88.79	89.13	89.35	**89.52**
0.1	S1	81.92	**82.65**	81.34	79.73	78.54	78.34	78.22
	S2	81.92	**82.63**	81.32	79.72	78.53	78.3	78.23
	S3	81.92	**82.63**	81.32	79.72	78.53	78.33	78.23
	S4	81.92	**82.61**	81.32	79.72	78.53	78.33	78.23
	S5	81.92	**82.65**	81.33	79.74	78.54	78.34	78.22

TABLE 4.9 Taux de reconnaissance obtenus en utilisant Eq. 4.3, avec différents nombres d'observation (R) en appliquant le protocole 10-fold CV sur le jeu de données Oulu-CASIA..

Seuil	Stratégie	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	S1	84.35	89.15	91.26	91.98	92.66	93.43	**94.19**
	S2	84.35	89.17	91.26	91.97	92.67	93.43	**94.21**
	S3	84.35	89.15	91.28	91.97	92.66	93.43	**94.2**
	S4	84.35	89.15	91.28	91.97	92.66	93.43	**94.2**
	S5	84.35	89.19	91.27	91.98	92.67	93.44	**94.21**
0.01	S1	85.75	90.57	**91.86**	91.69	91.23	90.79	90.46
	S2	85.75	90.59	**91.86**	91.69	91.23	90.79	90.46
	S3	85.75	90.57	**91.87**	91.69	91.23	90.79	90.46
	S4	85.75	90.57	**91.87**	91.69	91.23	90.79	90.46
	S5	85.75	90.6	**91.86**	91.69	91.23	90.79	90.46
0.1	S1	85.3	**86.87**	85.22	83.35	82.02	81.16	80.05
	S2	85.3	**86.89**	85.22	83.35	82.02	81.16	80.05
	S3	85.3	**86.87**	85.23	83.35	82.02	81.16	80.05
	S4	85.3	**86.87**	85.23	83.35	82.02	81.16	80.05
	S5	85.3	**86.9**	85.22	83.35	82.02	81.16	80.05

4.3.3.1 Matrice de confusion

La table 4.10 montre la matrice de confusion pour six émotions de base sur le jeu de données Oulu-CASIA. A partir de cette table, nous pouvons observer que les taux de reconnaissance des émotions joie, tristesse, surprise, colère et peur sont très élevés (supérieur à 99%), tandis que celui du dégoût est très faible. Nous remarquons que l'émotion tristesse est confondue avec les émotion colère et dégoût avec un taux d'erreur de 0.2% et l'émotion dégoût est incorrectement reconnue en tant qu'émotion tristesse avec un taux d'erreur de 89.28%.

TABLE 4.10 Matrice de confusion pour le jeu de données Oulu-CASIA (associée à la stratégie S1 avec R=7 et S=0)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**99.99**	0	0	0	0.002	0
Tristesse	0.02	**99.58**	0	0.2	0	0.2
Surprise	0	0.003	**99.98**	0	0.01	0
Colère	0.003	0.08	0	**99.89**	0.003	0.02
Peur	0	0.003	0	0	**99.85**	0.15
Dégoût	0	89.28	0	0.54	0.006	**10.16**
Moyenne						**84.88**

4.3.3.2 Comparaison avec l'état de l'art

Pour pouvoir positionner les performances de la méthode proposée, nous la comparons à plusieurs approches récentes de reconnaissance des expressions faciales dynamiques de la littérature, y compris les méthodes traditionnelles et les méthodes de deep learning, proposées par Guo et al. [76], Liu et al. [127], Sikka et al. [183], Jung et al. [97], Mengyi et al. [128], Zhao et al. [255], Kacem et al. [98], Sikka et Sharma [184], Allaert et al. [6] et Kuo et al. [112] sur la base de données Oulu-CASIA. Comme la plupart des méthodes suivent le protocole 10-fold sur six émotions de base et 480 séquences, commençant par le segment temporel "neutre" et finissant par le segment temporel "apex", pour évaluer leurs performances sur la base de données Oulu-CASIA, nous rapportons directement leurs résultats à partir des articles publiés. Les taux de reconnaissance obtenus par ces différentes approches sont présentés dans la table 4.11.

Comme dans la section 4.3.2.2, les deux mesures de performance mTR et gTR ont été considérées pour une comparaison équitable. Elle est montrée que notre méthode atteint les meilleurs performances parmi toutes les méthodes comparées, ce qui reflète son efficacité et sa robustesse.

Au vu des résultats obtenus dans la table 4.11, pendant six ans les performances obtenues sur la base de données Oulu-CASIA ne dépassaient pas 81%, jusqu'en 2018 où ce taux a été battu par [112] en utilisant un réseau profond (le modèle CNN). La table 4.11 montre aussi que notre méthode surpasse toutes les méthodes de la littérature, y compris la méthode de Kuo et al. [112] où le taux de reconnaissance est de 91.67%. Nous avons amélioré ce taux de reconnaissance d'environ 3% (soit 94.21%) en utilisant notre méthode simple, basée sur les probabilités issues du classifieur SVM.

TABLE 4.11 Performances obtenues par notre méthode et les méthodes récentes de la littérature sur le jeu de données Oulu-CASIA.

Méthode	Année	mTR	gTR	Catégorie	Protocole
Guo et al. [76]	2012	75.52	–	Traditionnelle	10-fold CV
Liu et al. [127]	2014	74.59	–	Traditionnelle	10-fold CV
Sikka et al. [183]	2015	75.62	–	Traditionnelle	10-fold CV
Jung et al. [97]	2015	81.46	81.46	Deep Learning	10-fold CV
Mengyi et al. [128]	2016	79.16	79	Traditionnelle	10-fold CV
Zhao et al. [255]	2017	74.37	–	Traditionnelle	10-fold CV
Kacem et al. [98]	2017	83.13	83.13	Traditionnelle	10-fold
Sikka et Sharma [184]	2018	82.1	–	Traditionnelle	LOSOCV
Allaert et al. [6]	2018	–	84.58	Traditionnelle	10-fold CV
Kuo et al. [112]	2018	–	91.67	Deep Learning	10-fold CV
Méthode proposée (S1)		**84.88**	**94.19**	Traditionnelle	10-fold CV

4.3.4 Evaluation des bases de données croisées

La méthode d'évaluation croisée des systèmes d'expression faciale nécessite l'apprentissage du classifieur sur toutes les images d'une base de données et l'évaluation du classifieur sur une base de données différente (dont les images sont inconnues par le classifieur). Comme les images provenant de la même base de données ont des paramètres similaires (éclairage, pose, résolution, etc.), les protocoles d'expérimentation indépendant de la personne sont plus faciles à résoudre que les protocoles croisés. Nous avons évalué la capacité de généralisation de notre méthode à travers les bases de données CK+ et Oulu-CASIA en effectuant deux expériences pour la reconnaissance des six émotions de base.

Dans la première expérience, pour laquelle les résultats sont listés dans la table 4.12, toutes les 480 séquences d'expression faciale sélectionnées dans la base de données Oulu-CASIA (section 4.3.1.2) sont utilisées comme jeu d'apprentissage, et les 327 séquences sélectionnées dans la base de données CK+ (section 4.3.1.1) sont utilisées comme jeu de test, en excluant celles de l'émotion mépris, ce qui nous laisse 309 séquences. Dans la deuxième expérience, pour laquelle les résultats sont présentés dans la table 4.13, le modèle est entraîné en utilisant le jeu de données CK+ contenant 309 séquences de six émotions de base et le jeu de Oulu-CASIA est utilisé comme jeu de test contenant 480 séquences. Pour les deux expériences, l'apprentissage s'est effectué en utilisant uniquement les 3 dernières images de chaque séquence contenant le pic de l'expression.

A partir des tables 4.12 et 4.13, nous pouvons voir que les taux de reconnaissance quand le nombre R d'observations est inférieur à 4 sont systématiquement inférieurs à ceux obtenus lorsqu'on utilise la même base de données pour l'apprentissage et le test (voir Tables 4.2 et 4.9). Cela est principalement dû aux variations plus importantes en termes de conditions d'éclairage, de pose et de forme de visage qui traversent différentes bases de données. Cependant, nous pouvons observer à partir des tables 4.12 et 4.13 que notre méthode atteint toujours des performances de reconnaissance très élevées lorsque le nombre R d'observation dépasse 4. En effet, nous obtenons un taux de reconnaissance global de 99.22% sur le jeu de donnée CK+ avec $R = 7$ et $S = 0.1$ (voir Table 4.12) et sur le jeu de données Oulu-CASIA, le meilleur taux obtenu est de 92.2% lorsque $R = 6$ et $S = 0.1$ (voir Table 4.13).

TABLE 4.12 Performances de la méthode proposée dans le cas de bases de données croisées. Apprentissage : Oulu-CASIA, Test : CK+

Seuil	Stratégies	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	∀ la stratégie	69.55	62.18	73.55	84.71	87.33	86.24	84.06
0.01		71.76	67.41	75.18	83.22	86.28	86	84.33
0.1	(S1, S2, S3, S4 et S5)	71.45	78.84	85.48	91.77	96.01	98.2	**99.22**

TABLE 4.13 Performances de la méthode proposée dans le cas de bases de données croisées. Apprentissage : CK+, Test : Oulu-CASIA

Seuil	Stratégies	R=1	R=2	R=3	R=4	R=5	R=6	R=7
0	∀ la stratégie	50.39	52.69	70.61	84.55	88.58	88.37	86.85
0.01		51.97	56.21	69.69	81.25	86.6	87.66	86.84
0.1	(S1, S2, S3, S4 et S5)	48.5	54.35	61.46	73.18	84.82	**92.2**	86.06

4.3.5 Expérience sur la base de données KDEF_MV

4.3.5.1 Evaluation en utilisant la méthode statique

Afin d'évaluer les bénéfices de la multi-observation dans un contexte de reconnaissance des expressions faciales à vue multiple (MVFE), nous allons tout d'abord effectuer la reconnaissance mono-vue en se basant sur notre méthode précédente (voir chapitre précédent). Pour ce faire, trois expériences ont été adoptées en considérant les vues séparément. Dans la première expérience, les données d'apprentissage et de test sont associées aux images avec l'angle $-45°$. Dans la deuxième expérience, les données d'apprentissage et de test sont associées aux images avec l'angle $0°$. Dans la troisième expérience, les données d'apprentissage et de test contiennent seulement les images avec l'angle $45°$. La table 4.14 rapporte les résultats obtenus en utilisant la méthode précédente (voir chapitre 3), et ceux obtenus avec une méthode de reconnaissance mono-vue de la littérature sur le jeu de données KDEF_MV avec 6 expressions de base ainsi que l'état neutre. Nous pouvons observer que notre méthode surpasse la méthode comparée sur deux angles de vue ($0°$ et $45°$). Il est à noter que la comparaison est faite avec des protocoles expérimentaux différents (voir Table 4.14).

TABLE 4.14 Les taux de reconnaissance obtenus par notre méthode précédente sur le jeu de données KDEF _MV et des comparaisons avec une méthode de l'état de l'art.

Angles de vue	$-45°$	$0°$	$45°$	Moyenne	Protocole	# des images
Notre méthode (chapitre 3)	76.51	86.74	79.63	80.96	LOSOCV	1224
[172]	80.39	84.07	77.03	80.49	k-fold CV	1168

4.3.5.2 Evaluation en utilisant la méthode basée sur la multi-observation

Cette expérience consiste à analyser les expressions faciales à vue multiple où un sous-ensemble d'observations testé doit contenir des images appartenant à la même personne sous des points de vue différents. Le protocole LOSOCV est adopté pour évaluer notre méthode sur le jeu de données KDEF_MV.

Les deux ensembles d'apprentissage et de test contiennent les images prises à partir des trois angles. Les images de l'ensemble de test sont regroupées en lots de taille : 1, 2 et 3. Le

lot pour $R = 1$ signifie que nous testons une vue (”−45°”, ”0°”*et*”45°”) à la fois. Le lot pour $R = 2$ consiste à analyser deux observations à la fois, ce qui va nous amener à tester les trois combinaisons (”−45° et 45°”, ”−45° et 0°” et ”0° et 45°”). Le lot pour $R = 3$ correspond au test de trois vues à la fois comme le montre la figure 4.9.

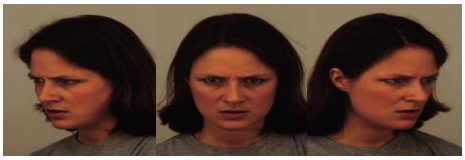


FIGURE 4.9 Exemple de 3 observations prises à partir de différents points de vue représentant la même expression dans la base de données KDEF.

La figure 4.10 montre les résultats obtenus en utilisant nos stratégies proposées pour la multi-observation sur le jeu de données KDEF_MV. Dans cette expérience, les ensembles d'apprentissage et de test contiennent des images provenant de trois vues. Comme nous pouvons le voir sur la figure 4.10, la stratégie (S2) surpasse toutes les autres stratégies proposées. A partir de la figure 4.10 et la table 4.14, nous pouvons observer que l'approche multi-observation a pu améliorer le taux de reconnaissance obtenu par la méthode de chapitre 3 d'environ 3% (83.43 % VS 80.96%).

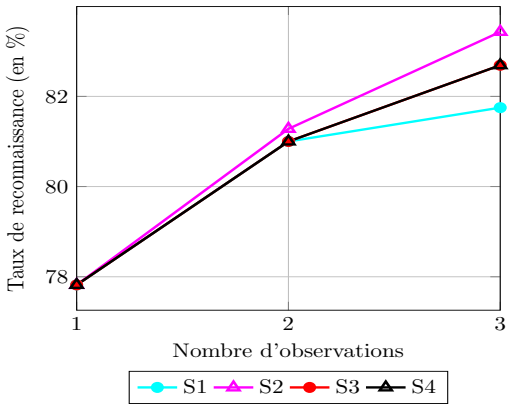


FIGURE 4.10 Taux de reconnaissance obtenus par la méthode proposée (méthode de multi-vue) sur le jeu de données KDEF_MV avec différents nombres d'observations.

4.3.6 Expérience sur les bases de données CK, FEED et KDEF

Dans cette expérience, la méthode de validation croisée de 10-fold a été adoptée pour évaluer les performances de reconnaissance de la méthode proposée. Pour chaque base, les données sont regroupées en dix groupes comme expliqué dans le chapitre précédent, puis

ces groupes servent pour former les ensembles d'apprentissage et de test avec différents ratios : 10%-90% , 20%-80%, 30%-70, 40%-60%, 50%-50%, 60%-40%, 80%-20%, 90%-10%. En commençant par 10%-90%, dans chaque partition, un groupe est utilisé comme ensemble d'apprentissage tandis que les neuf groupes restants sont utilisés pour le test. Puis successivement, nous incrémentons et décrémentons le nombre de groupes utilisés respectivement dans l'ensemble d'apprentissage et l'ensemble de test jusqu'à passer en revu tous les différents ratios apprentissage-test, énumérés ci-dessus.

Comme expliqué ci-dessus, les images de l'ensemble de test sont regroupées en sous-ensembles d'images de taille R (nombre d'observations). Ce regroupement est obtenu en utilisant toutes les combinaisons possibles à partir d'un ensemble de test de taille t. Nous évaluons notre méthode sur toutes les combinaisons possibles (C_t^R). Dans cette expérience, nous nous intéressons à l'analyse d'une multi-observation d'un état émotionnel provenant de personnes différentes et nous étudions également l'impact de la taille de l'ensemble d'apprentissage.

Les figures 4.11, 4.12 et 4.13 montrent les résultats obtenus par les stratégies proposées avec différents ratios apprentissage-test respectivement sur les jeux de données KDEF, CK et FEED. Les axes horizontaux représentent le nombre d'observations testé. Sur la base des résultats présentés sur la figure 4.11, nous pouvons voir que lorsque la taille de l'ensemble d'apprentissage augmente, les performances de la REF s'améliorent. En particulier, la stratégie (S2) atteint un taux de reconnaissance entre 98.18% et 100% lorsque le ratio de l'ensemble d'apprentissage dépasse 20% et le nombre d'observations augmente. Pour le jeu de données CK (voir figure 4.12), le meilleur taux de reconnaissance de 98.35% est obtenu en utilisant la stratégie (S2) en considérant 70% des données totales comme ensemble d'apprentissage et les données restantes comme ensemble de test. Concernant le jeu de données FEED (voir Figure 4.13), les taux de reconnaissance les plus élevés de 97.46% et 97.58% sont atteints en utilisant la stratégie (S2) lorsque les ratios apprentissage-test sont respectivement de 70%-30% et 80%-20%.

Dans cette expérience, les observations représentent un sous-ensemble d'images de taille R appartenant à des personnes différentes comme si nous étudions l'état émotionnel de personnes multiples. Le but principal de cette expérience était de comparer les taux de reconnaissance en utilisant une seule observation ($R = 1$), c-à-d un seul individu, et ceux obtenus en considérant des observations multiples (individus multiples). L'expérience montre clairement que les performances obtenues avec la multi-observation sont plus élevées que celles obtenues avec une seule observation. Puisque l'intensité de l'expression faciale peut varier d'un individu à l'autre, la reconnaissance des émotions à l'aide de visages multiples améliore les performances de reconnaissance.

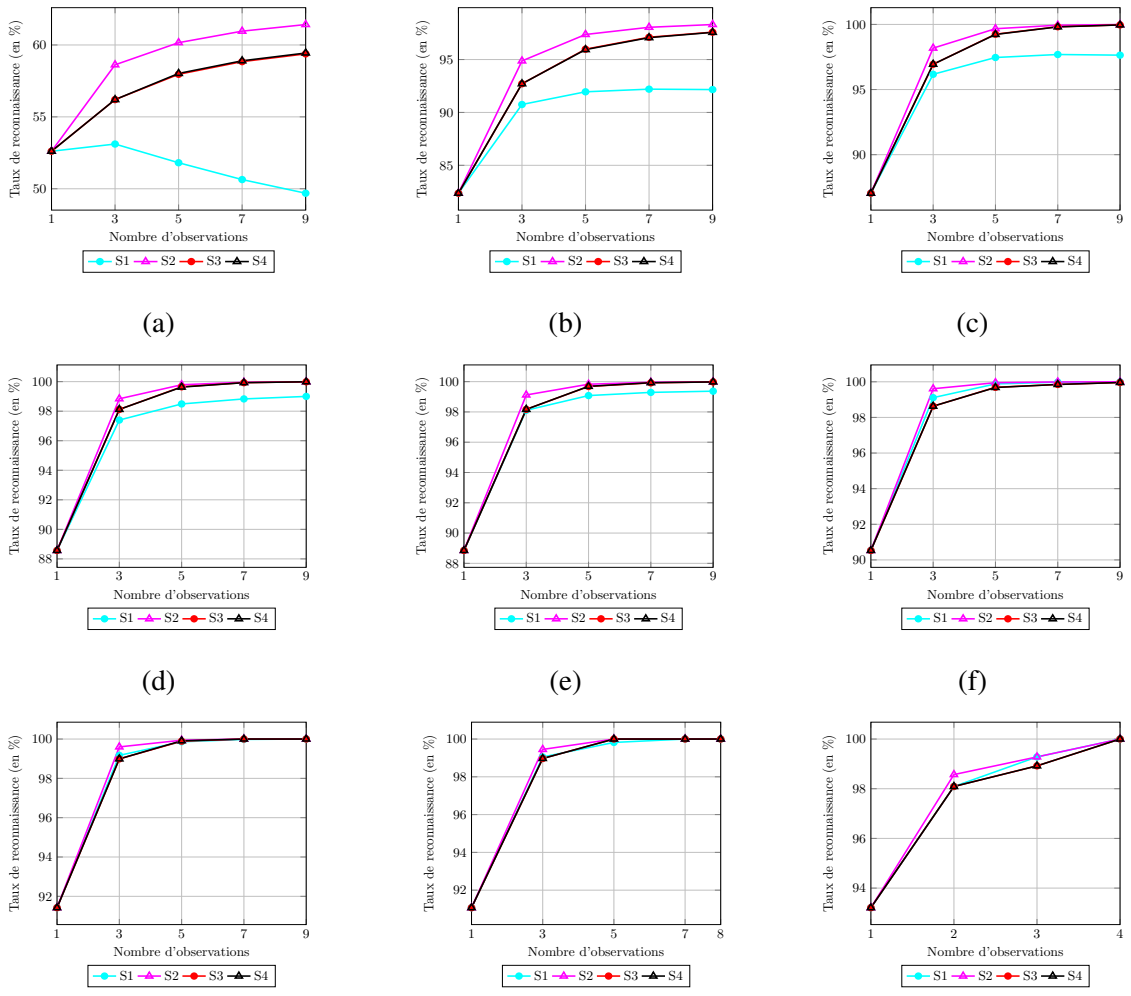


FIGURE 4.11 Taux de reconnaissance obtenus par les stratégies proposées (en %) sur le jeu de données KDEF en utilisant différents nombres d'observations et différent ratios pour les ensembles Apprentissage-Test. (a) 10% – 90%. (b) 20% – 80%. (c) 30% – 70%. (d) 40% – 60%. (e) 50% – 50%. (f) 60% – 40%. (g) 70% – 30%. (h) 80% – 20%. (i)90% – 10%.

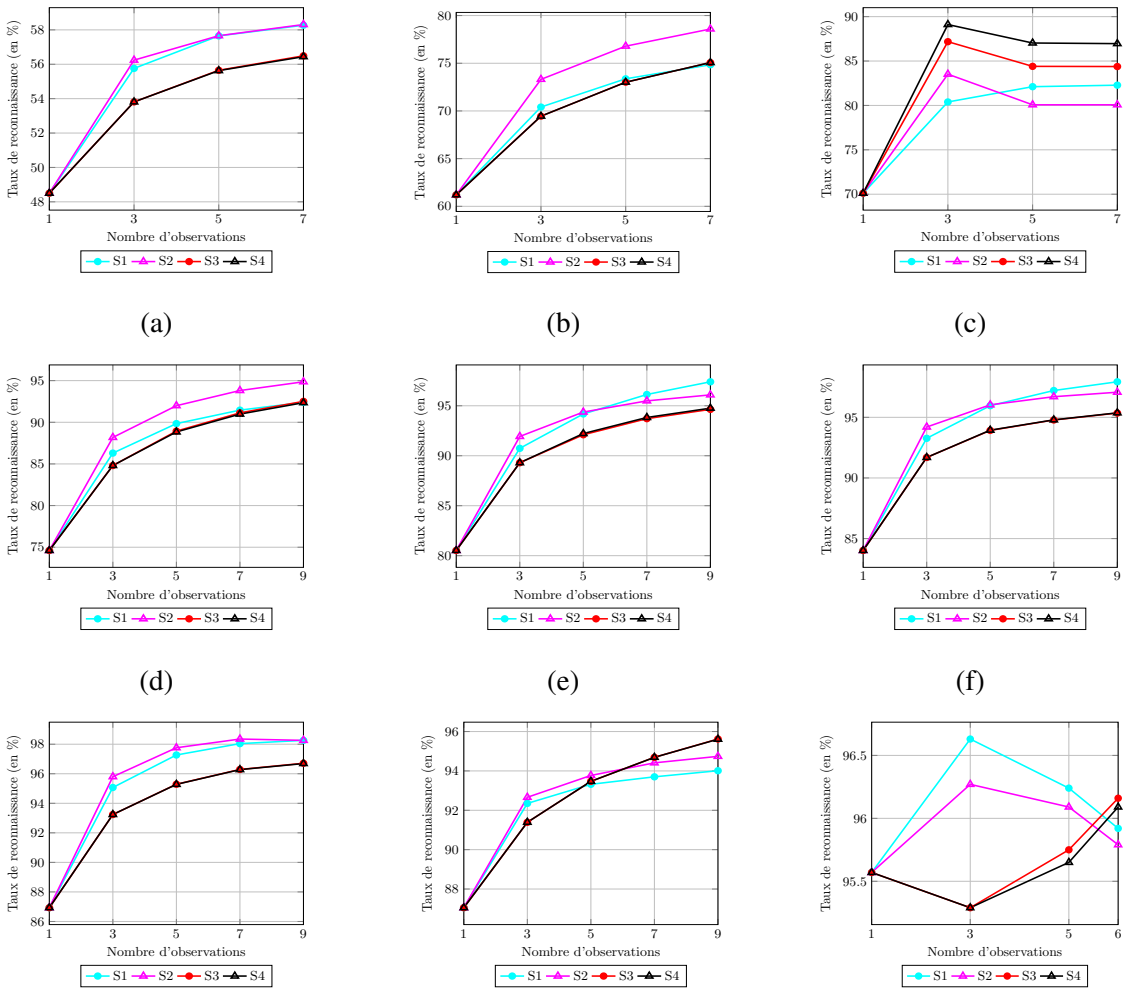


FIGURE 4.12 Taux de reconnaissance obtenus par les stratégies proposées (en %) sur le jeu de données CK en utilisant différents nombres d'observations et différent ratios pour les ensembles Apprentissage-Test. (a) 10% – 90%. (b) 20% – 80%. (c) 30% – 70%. (d) 40% – 60%. (e) 50% – 50%. (f) 60% – 40%. (g) 70% – 30%. (h) 80% – 20%. (i)90% – 10%.

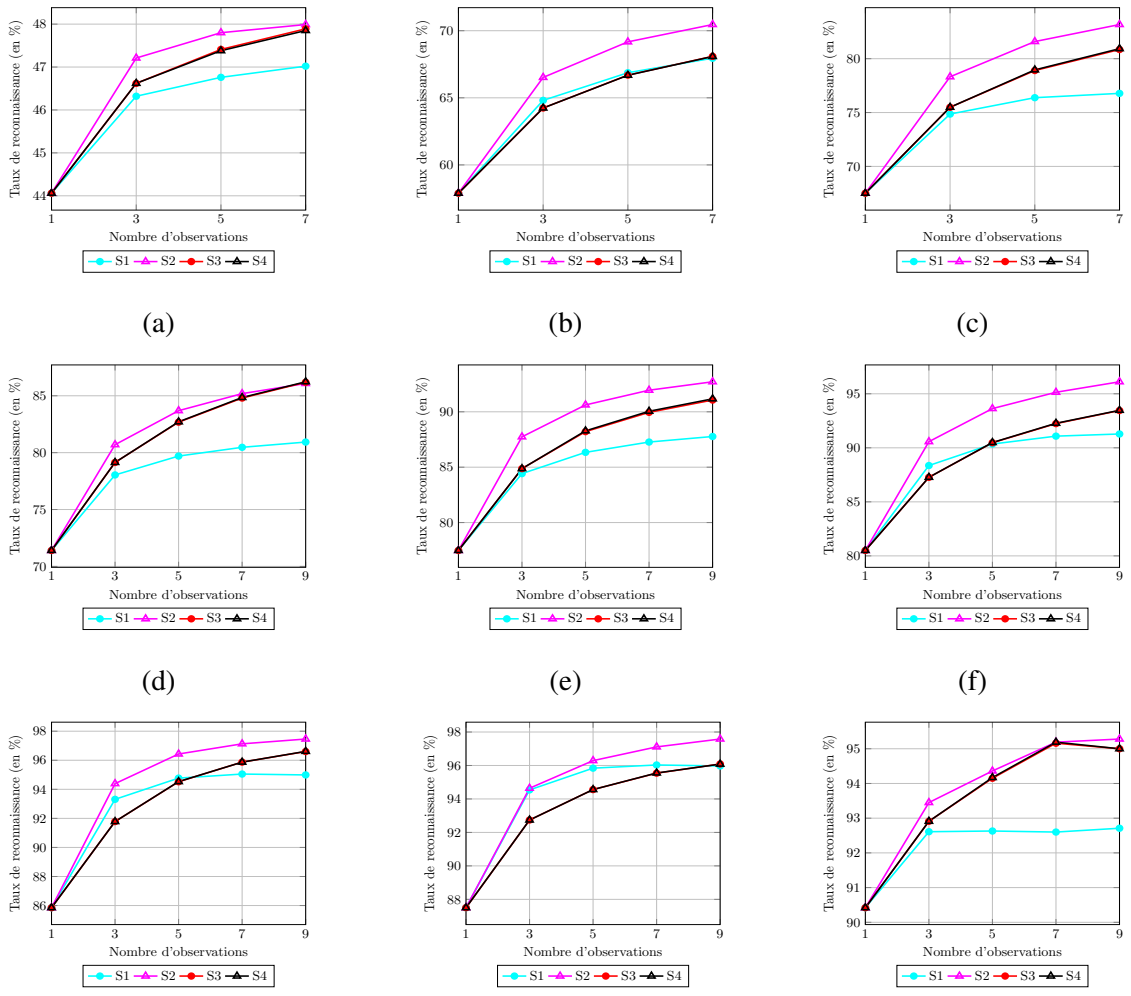


FIGURE 4.13 Taux de reconnaissance obtenus par les stratégies proposées (en %) sur le jeu de données FEED en utilisant différents nombres d'observations et différent ratios pour les ensembles Apprentissage-Test. (a) 10% – 90%. (b) 20% – 80%. (c) 30% – 70%. (d) 40% – 60%. (e) 50% – 50%. (f) 60% – 40%. (g) 70% – 30%. (h) 80% – 20%. (i)90% – 10%.

4.4 Conclusion

Dans ce travail, nous avons présenté une nouvelle méthode de REF dynamique avec un algorithme simple. Cet algorithme est capable de classer une émotion représentée avec plusieurs images. La multi-observation peut représenter une séquence vidéo d'une même personne exprimant une émotion de manière dynamique, un ensemble d'images de différentes personnes exprimant une même émotion ou un ensemble d'images capturées à partir de différents points de vue d'une même personne exprimant une même émotion. Nous avons proposé une étude complète de l'utilisation du classifieur SVM dans la REF montrant que ce dernier est efficace pour les systèmes de REF dynamiques. Les estimations de probabilité en sortie du SVM, associées à chaque observation, sont exploitées en suivant différentes stratégies afin d'affecter à un sous-ensemble d'images une classe (émotion).

La méthode proposée a été largement évaluée sur les bases de données dynamiques d'expression faciale CK+ et Oulu-CASIA. Elle a été comparée aux approches de reconnaissance dynamique des expressions faciales les plus récentes de la littérature. Les résultats expérimentaux montrent que la méthode proposée permet d'obtenir des taux de reconnaissance plus élevés que ceux obtenus par les autres méthodes comparées. Nous espérons que ce travail peut donner lieu à de nouvelles méthodes de reconnaissance d'expressions faciales dynamiques.

Une des limites de la méthode proposée est qu'elle n'est pas conçue pour détecter automatiquement le segment temporel "apex", et pourtant, elle se base principalement sur les images représentant le pic de l'émotion recherchée pour construire l'ensemble d'apprentissage et l'ensemble de test. Cette limitation ne nous a pas permis de tester la base de données MMI [160] ayant des séquences commençant par l'état neutre et finissant par l'état offset. Nous avons testé seulement les bases de données comme CK+ et Oulu-CASIA puisque leurs séquences d'images débutent par un état neutre et se terminent par le pic de l'expression. Une solution possible consiste à chercher automatiquement les images représentant le pic de l'expression dans une séquence donnée puis appliquer la méthode proposée. C'est l'une des orientations possibles pour cette étude, sachant qu'il existe déjà dans la littérature des systèmes qui permettant de détecter chaque segment temporel (Neutre, Onset, Apex et Offset) [212, 193, 185].

Chapitre 5

Reconnaissance des expressions faciales basée sur des caractéristiques géométriques et le regroupement des émotions

Sommaire

5.1 Introduction

Habituellement, il existe deux méthodes pour extraire les caractéristiques du visage : les méthodes géométriques et celles d'apparence. Normalement, les caractéristiques géométriques proviennent des formes des composantes faciales et de l'emplacement des points saillants du visage (coins des yeux, de la bouche, etc.). Tandis que les caractéristiques d'apparence sont extraites en enregistrant les changements d'apparence du visage. Comme mentionné dans la section 1.3.2.3 du chapitre 1, la performance des méthodes utilisant la combinaison des caractéristiques d'apparence et géométriques surpasse celle des méthodes basées sur les caractéristiques géométriques ou d'apparence. Par conséquent, plusieurs travaux [201, 162, 164, 71] utilisent la combinaison de caractéristiques géométriques et d'apparence pour améliorer la précision du codage d'expression. A cet effet, pour améliorer notre premier système basé sur les caractéristiques d'apparence (voir chapitre 3), nous avons étudié les caractéristiques géométriques. Dans un premier temps, nous proposons un nouveau descripteur géométrique basé sur 33 points faciaux détectés par SDM (Supervised Descent Method) [231]. Puis, nous combinons ce descripteur avec les caractéristiques d'apparence, utilisées dans le chapitre 3, pour une meilleure reconnaissance des expressions faciales (REF).

A partir de l'analyse des matrices de confusion obtenues en évaluant les trois types de descripteurs (géométriques, apparence et leur combinaison), nous avons remarqué que les confusions indiquent que les frontières entre certaines expressions de base ne sont pas bien définies. Plus précisément, la peur et le dégoût ont tendance à être classés respectivement en surprise et colère. De là, une approche très intéressante peut être développée en formant quatre ou deux groupes d'expressions faciales au lieu de six. Dans le cas de quatre groupes d'émotions, le premier groupe comprend l'expression joie, le deuxième contient l'expression tristesse, le troisième se compose des expressions surprise et peur tandis que les expressions colère et dégoût sont classées dans le quatrième groupe. Dans le cas de deux groupes d'émotions, le premier comprend la joie, la peur et la surprise, tandis que le deuxième se compose de la tristesse, la colère et le dégoût.

De ce fait, ce chapitre est devisé en deux parties qui contiennent les résultats préliminaires de ces analyses. Dans la première partie, nous allons introduire le descripteur géométrique puis une comparaison entre les descripteurs (géométriques, apparence et leur combinaison) sera faite. La deuxième partie consiste à introduire une nouvelle approche basée sur le regroupement des six émotions en deux ou quatre groupes d'émotions.

5.2 Reconnaissance des expressions faciales basée sur des caractéristiques géométriques

Le système proposé dans ce chapitre suit les étapes suivantes :

1. le visage est détecté en utilisant l'algorithme Viola et Jones (VJ) [217]
2. les point faciaux sont détectés en utilisant SDM [231]
3. la méthode géométrique prend en entrée les coordonnées des points faciaux et fournit en sortie un vecteur de caractéristiques de 15 angles.
4. la méthode d'apparence prend en entrée les 7 ROIs extraites en utilisant les points faciaux (voir la section 3.5 du chapitre 3). Cette méthode permet de caractériser chaque ROI en utilisant les descripteurs LTP et HOG. Enfin, les descripteurs de toutes les ROIs sont concaténés pour obtenir le descripteur du visage
5. Le vecteur de caractéristiques final peut correspondre à un des trois cas possibles suivants :
   (a) Caractéristiques géométriques (en ignorant l'étape 4)
   (b) Caractéristiques d'apparence (en ignorant l'étape 3)
   (c) Combinaison des deux caractéristiques (en considérant les deux étapes 3 et 4)

   Enfin, le vecteur obtenu est introduit dans un SVM multi-classe pour accomplir la tâche de reconnaissance

5.2.1 Descripteur géométrique

Les caractéristiques géométriques décrivent la forme des composantes faciales (i.e. la bouche, les yeux, les sourcils et le nez) et leur emplacement (i.e. les coins des yeux, les coins de la bouche, etc.). Elles représentent des composantes faciales ou des traits faciaux, formant un vecteur de caractéristiques qui décrit la géométrie du visage. Principalement, les déformations sont codées en utilisant la position des points faciaux du visage, ou les distances et les angles entre les points faciaux. Ainsi, la motivation pour employer une méthode basée sur la géométrie est que les expressions faciales affectent la position relative et la taille des diverses traits faciaux et que, en mesurant le mouvement de certains points faciaux, l'expression faciale sous-jacente peut être déterminée.

33 parmi 49 points faciaux détectés sont rassemblés pour construire le descripteur géométrique. Ces points sont extraits automatiquement, en utilisant la méthode SDM. Les coordonnées de ces points faciaux servent à calculer 15 angles. L'extraction des caractéristiques géométriques sélectionnées est effectuées comme suit : une fois la détection des points

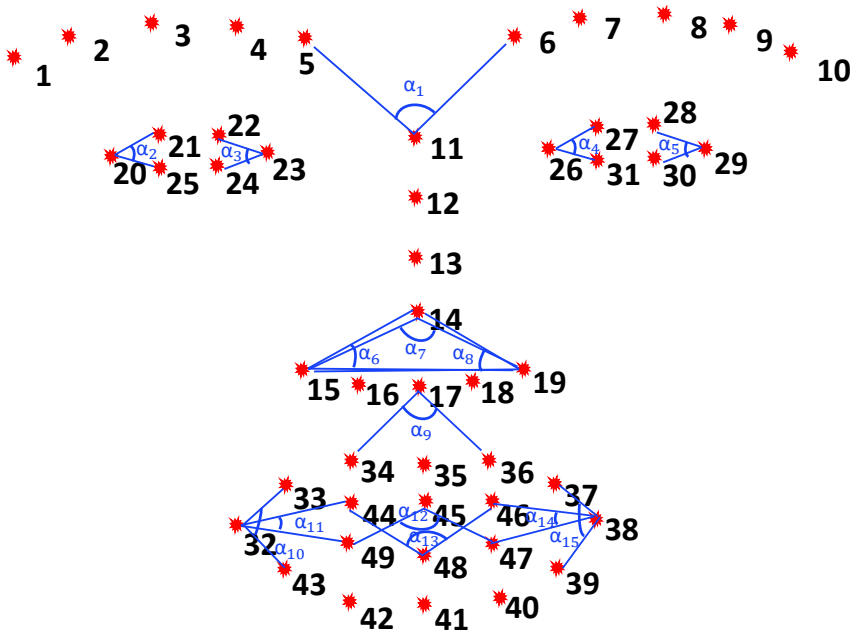


FIGURE 5.1 Sélection des caractéristiques géométriques en utilisant des points faciaux.

faciaux effectuée par la méthode SDM, les angles entre les points sont calculés : Six angles sont utilisés pour coder les mouvements de la bouche, quatre décrivent le mouvement des paupières, un angle code le rapprochement des sourcils et quatre décrivent les déformations du nez. Les angles sont indiqués par α_i comme illustré sur la figure 5.1.

5.2.2 Combinaison des descripteurs d'apparence et géométrique

La forme, la texture et la géométrie des composantes faciales (les yeux, les sourcils, le nez et la bouche) sont exposées à changer avec l'expression faciale. Nous avons aussi montré dans le chapitre 3 que la combinaison des informations d'apparence (forme à travers HOG et texture à travers LTP) améliore les performances d'un système de REF. Par conséquent, la caractérisation des composantes faciales doit comporter les trois informations. Dans cette section, nous proposons une combinaison de différents types de descripteurs pour la reconnaissance des expressions faciales. Nous utilisons le descripteur LTP pour extraire l'information de texture et le descripteur HOG pour extraire les informations de la forme. Ces descripteurs d'apparence sont ensuite combinés à des descripteurs de type géométrique en calculant 15 angles entre les points faciaux (Voir la section précédente 5.2.1).

Nous utilisons une méthode de combinaison assez naïve qui consiste à combiner les trois types de descripteur dans un même vecteur. Ce dernier est ensuite considéré comme entrée pour le classifieur SVM multi-classe.

5.2.3 Expérimentations

Pour évaluer les méthodes d'apparence, géométrique et de combinaison, les bases de données FEED [220], CK [101], CK+ [132] et KDEF [134] sont utilisées. Les jeux de données sélectionnés dans le présent chapitre ainsi que le protocole d'expérimentation 10-fold CV sont les mêmes que ceux utilisés dans le chapitre 3.

5.2.3.1 Comparaison entre les descripteurs d'apparence, les descripteurs géométriques et leur combinaison

Dans cette section, nous comparons les taux de reconnaissance des expressions faciales décrites par chaque type de descripteurs. Les mesures de performance considérées sont : F-score et le taux de reconnaissance global obtenu sur 10-fold CV (désignée par gTR).

Les tables 5.1 et 5.2 montrent les taux de reconnaissance des méthodes géométrique, d'apparence et de combinaison, appliquées respectivement sur six émotions de base et sept émotions (toutes les expressions émotionnelles, y compris la neutralité). Les résultats obtenus par la méthode géométrique sont inférieurs à ceux obtenus par les méthodes de combinaison et d'apparence. Ceci est observé quelque soit le jeu de données testé et le nombre d'émotions utilisé (voir les tables 5.2 et 5.1). Nous pouvons aussi observer, à partir de la table 5.1, que la méthode de combinaison fournit des taux de reconnaissance supérieurs à ceux obtenus par la méthode d'apparence, à l'exception du jeu de données FEED où les résultats obtenus par la méthode de combinaison sont légèrement inférieurs à ceux obtenus par la méthode d'apparence. En revanche, d'après la table 5.2, pour tous les jeux de données utilisés, la méthode de combinaison a montré son efficacité, excepté pour le taux global (gTR) obtenu sur CK+, celui-ci étant inférieur au taux obtenu en utilisant la méthode d'apparence (94.62% VS 96.03%).

TABLE 5.1 Taux de reconnaissance obtenus en utilisant les méthodes d'apparence, géométrique et de combinaison sur les jeux de données KDEF, FEED, CK et CK+ avec 6 émotions de base.

	KDEF		FEED		CK		CK+	
	F-score	gTR	F-score	gTR	F-score	gTR	F-score	gTR
Méthode d'apparence	93.4	93.33	**94.39**	**94.52**	**98.29**	98.03	96.01	96.77
Méthode géométriques	77.43	77.5	66.62	68.11	82.83	81.96	85.6	87.6
Méthode de combinaison	**94.2**	**94.21**	94.24	94.15	**98.28**	**98.43**	**96.43**	**97.29**

TABLE 5.2 Taux de reconnaissance obtenus en utilisant les méthodes d'apparence, géométrique et de combinaison sur les jeux de données KDEF, FEED, CK et CK+ avec 6 émotions de base et la neutralité.

	KDEF		FEED		CK		CK+	
	F-score	gTR	F-score	gTR	F-score	gTR	F-score	gTR
Méthode d'apparence	93.34	93.24	92.03	91.8	96.06	95.7	94.63	**96.03**
Méthode géométriques	73.14	73.21	60.12	62.85	74.88	73.11	80.34	83.17
Méthode de combinaison	**93.7**	**93.57**	**93.21**	**93.01**	**96.73**	**96.39**	**95.71**	94.62

5.2.3.2 Matrices de confusion

Nous analysons, dans cette section, les matrices de confusion obtenues par les trois méthodes (géométrique, apparence et combinaison) sur les jeux de données FEED représentant les émotions spontanées et CK+ représentant les émotions posées. Nous ne considérons que les matrices de confusion de six émotions de base.

Les tables 5.3, 5.4 et 5.5 présentent les matrices de confusion respectives de la méthode d'apparence, de la méthode géométrique et de la méthode de combinaison calculées sur le jeux de données CK+. En comparant seulement les tables 5.3 et 5.4, nous pouvons observer que les taux de reconnaissance, de toutes les émotions, reconnues par la méthode d'apparence sont entre 92% et 99%. Ces taux sont supérieurs à ceux obtenus par la méthode géométrique. En matière de confusion, nous remarquons que les émotions colère et peur se confondent respectivement avec les émotion dégoût et surprise avec des taux d'erreur respectifs de 4.66% et 6.66% pour la méthode d'apparence. En revanche, la méthode géométrique se trompe sur plusieurs émotions avec des taux d'erreur élevés. Par exemple, l'émotion tristesse est mal classée en tant qu'émotion colère avec un taux d'erreur de 18.89%, aussi l'émotion peur est reconnue comme étant l'émotion joie avec un taux d'erreur de 12.2%. Nous observons par ailleurs que les taux de reconnaissance des émotions tristesse, colère et peur sont inférieurs à ceux obtenus pour les émotions joie, surprise et dégoût. Cela peut être expliqué par le fait que la méthode géométrique ignore les informations d'apparence telles que les rides transitoires qui apparaissent entre les sourcils lorsque le visage exprime la peur, la colère ou la tristesse, comme le rapporte le système FACS.

A partir de la table 5.5, nous remarquons une augmentation de 0.95%, 1.11% et 2.8% respectivement dans les taux de reconnaissance des émotions joie, peur et dégoût. Par contre, une faible diminution de performance dans la reconnaissance de la tristesse, la surprise et la colère.

TABLE 5.3 Matrice de confusion de la méthode d'apparence calculée sur le jeu de données CK+ (associée aux F-score :96.01% et gTR : 96.77%)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**98.57**	0	0	1.42	0	0
Tristesse	0	**95.55**	1.11	1.11	0	2.22
Surprise	0	0	**99.58**	0	0.41	0
Colère	0	2	0	**92.66**	0.66	4.66
Peur	1.11	0	6.66	0	**92.22**	0
Dégoût	0	0	0	2.77	0	**97.22**

TABLE 5.4 Matrice de confusion de la méthode géométrique calculée sur le jeu de données CK+ (associée aux F-score :85.6% et gTR :87.6%)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**89.04**	0	0	3.8	7.14	0
Tristesse	0	**80**	0	18.89	0	1.1
Surprise	0	0	**98.33**	0	1.66	0
Colère	6.66	9.33	0	**74**	2	8
Peur	12.22	3.33	1.1	0	**82.22**	1.11
Dégoût	1.11	0.55	1.11	7.22	0.55	**89.44**

TABLE 5.5 Matrice de confusion de la méthode de combinaison calculée sur le jeu de données CK+ (associée aux F-score :96.43% et gTR :97.29 %)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**99.52**	0	0	0.48	0	0
Tristesse	0	**94.44**	0	5.55	0	0
Surprise	0	0	**99.16**	0	0.83	0
Colère	0	2	0	**92**	0	6
Peur	3.33	3.33	0	0	**93.33**	0
Dégoût	0	0	0	0	0	**100**

Les tables 5.6, 5.7 et 5.8 présentent respectivement les matrices de confusion de la méthode d'apparence, la méthode géométrique et la méthode de combinaison sur le jeu de données FEED. A partir de ces tables, nous pouvons remarquer que la méthode géométrique génère beaucoup de confusion entre les émotions avec des taux d'erreur très élevés. Par exemple, la tristesse est reconnue comme étant la peur avec un taux d'erreur de 30% et le dégoût est classé comme émotion surprise avec un taux d'erreur de 28%. Nous pouvons aussi

noter que les taux de reconnaissance obtenus pour les émotions joie et surprise par la méthode géométrique sont supérieurs à 84%. Dans le cas de la base CK+, ces deux émotions sont également reconnues avec les meilleurs taux de reconnaissance par la méthode géométrique par rapport aux autres émotions. Elles semblent ainsi bien définies par les angles proposés.

En comparant les matrices de confusion de la méthode d'apparence (voir Table 5.6) et de la méthode de combinaison (voir Table 5.8), nous pouvons remarquer une diminution de performance dans la reconnaissance de la joie et du dégoût. En revanche, une augmentation de performance respectivement de 3%, 4.67% et 3.3% dans la reconnaissance de la surprise, la colère et la peur, par la méthode de combinaison, peut être observée. D'après la table 5.8, nous pouvons noter que la méthode de combinaison confond les émotions joie et peur avec l'émotion surprise avec un taux d'erreur de 7% et confond le dégoût avec la colère avec un taux d'erreur de 8%. Bien que les expressions faciales de la base FEED sont exprimées avec une faible intensité, les descripteurs de texture LTP et de forme HOG ont montré leur efficacité pour détecter les changements dûs aux expressions faciales. La méthode d'apparence est ainsi plus adaptée pour la reconnaissance des émotions spontanées.

TABLE 5.6 Matrice de confusion de la méthode d'apparence calculée sur le jeu de données FEED (associée aux F-score :94.39% et gTR :94.52%)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**98**	0	0	0	0	2
Tristesse	0	**96**	0	0	4	0
Surprise	0	0	**96**	4	0	0
Colère	0	0	6.67	**93.33**	0	0
Peur	0	3	0	3	**90**	4
Dégoût	5	1	0	0	1	**93**

TABLE 5.7 Matrice de confusion de la méthode géométrique calculée sur le jeu de données FEED (associée aux F-score :66.62% et gTR :68.11%)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**84**	2	1	13	0	0
Tristesse	2	**53**	0	14	30	1
Surprise	4	0	**85**	0	9	2
Colère	14	7	0	**71**	7	1
Peur	6.66	16.66	20	10	**46.66**	0
Dégoût	2	0	28	15	1	**54**

TABLE 5.8 Matrice de confusion de la méthode de combinaison calculée sur le jeu de données FEED (associée aux F-score :94.24% et gTR :94.15%)

	Joie	Tristesse	Surprise	Colère	Peur	Dégoût
Joie	**93**	0	7	0	0	0
Tristesse	0	**95**	0	4	0	1
Surprise	0	0	**99**	0	1	0
Colère	0	1	0	**98**	1	0
Peur	0	0	6.67	0	**93.33**	0
Dégoût	6	0	0	8	0	**86**

5.3 Reconnaissance des expressions faciales basée sur le regroupement des émotions

Les humains ont des émotions très complexes, profondes et nuancées. Ces émotions ont été classées en six catégories différentes : la joie, la tristesse, la surprise, la colère, la peur et le dégoût. Cette catégorisation est basée sur des recherches effectuées par le psychologue Paul Ekman [51, 60, 59], en utilisant la théorie de Charles Darwin [43]. Darwin pensait que les gens du monde entier devaient manifester des émotions de la même manière, et Ekman a voyagé dans le monde entier en demandant aux gens de tous les types de cultures d'exhiber ces émotions et les expressions faciales correspondantes. Selon une nouvelle étude de Rachael et al. [91], les émotions chez tous les humains commencent par les mêmes expressions faciales qui transmettent peu de signaux faciaux d'origine biologique. Tous les autres détails sont ajoutés en fonction de la culture et de la société dans laquelle nous vivons. Le sourire représente toujours la joie, le renfrognement indique la tristesse. Le plissement du nez est le début de la colère et du dégoût et il vient d'un besoin fondamental de montrer que quelque chose de déplaisant ou de dangereux se produit. La peur et la surprise commencent de la même manière, en élargissant les yeux, en intégrant davantage d'informations visuelles, en évaluant la situation et en recherchant une éventuelle évasion. Mis à part ces quatre réactions de base, tout le reste est lié à la culture de la personne.

Dans la première partie de ce chapitre, nous avons analysé les matrices de confusion, en tant qu'indice de discrimination d'expression. Cette analyse nous a permis de dégager le schéma suivant. La tristesse a tendance à être classée comme la colère (5.5% sur CK+ et 4% sur FEED), la peur est confondue avec la surprise (6.6%), le dégoût a tendance à être confondu avec la colère (8%) et inversement la colère est confondue avec le dégoût (6%), la peur se confond avec la joie (3.33%) et avec la tristesse (3.33%) et la joie est confondue

avec la surprise (7%). En tout, ces confusions indiquent que les frontières entre certaines expressions de base, en particulier celles entre surprise et peur, colère et dégoût, ne sont pas bien définies. Fait intéressant, Rachael et al. [91] ont obtenu des résultats cohérents avec ceux issus des données de confusion. Ils ont trouvé la preuve de quatre émotions de base, à savoir, joie, tristesse, peur/surprise et dégoût/colère, au lieu de six. La peur et la surprise, d'une part, et le dégoût et la colère, d'autre part, partageraient les mêmes représentations, au moins au début de la séquence. Cela correspond aux schémas de confusion majeure signalés dans les tâches de catégorisation. Plus précisément, selon [91], la signalisation précoce de l'expression faciale favorise la discrimination en quatre catégories : joie, tristesse, peur/surprise et dégoût/colère. Par exemple, la surprise et la peur transmettent la même unité d'action numéro 5 (AU5, ouverture entre la paupière supérieure et les sourcils) au début de la séquence, entraînant leur confusion précoce.

5.3.1 Système proposé

Dans ce travail, une nouvelle approche de REF a été considérée en se basant sur les résultats issus des matrices de confusion présentées et analysées dans la partie précédente et sur la catégorisation précoce définie par [91]. Deux catégorisations différentes ont été proposées et évaluées. La première consiste à regrouper les six émotions en quatre groupes joie, tristesse, peur/surprise et colère/dégoût (voir Figure 5.2). La deuxième catégorisation repose sur deux groupes : joie/surprise/peur et tristesse/colère/dégoût (voir Figure 5.3).

Le système proposé est illustré sur la figure 5.2 pour une catégorisation en quatre groupes et sur la figure 5.3 pour une catégorisation en deux groupes. Les étapes du système proposé sont les suivantes :

1. Le système localise d'abord les régions faciales saillantes à l'aide de la méthode SDM, comme cela a été effectué dans le système proposé dans le chapitre 3.
2. Ensuite, les caractéristiques LTP+HOG sont extraites de ces régions localisées. Une première classification "Classifieur 1" est effectuée :
   (a) sur la base de caractéristiques d'apparence extraites afin de former quatre groupes d'expressions faciales (voir la figure 5.2). Le premier et le deuxième groupe comprennent respectivement les émotions joie et tristesse. Le troisième contient les deux émotions surprise/peur tandis que le dernier groupe est composé des émotions colère/dégoût.
   (b) sur la base de caractéristiques d'apparence extraites afin de former deux groupes d'expressions faciales (voir la figure 5.3). Le premier est composé de trois émo-

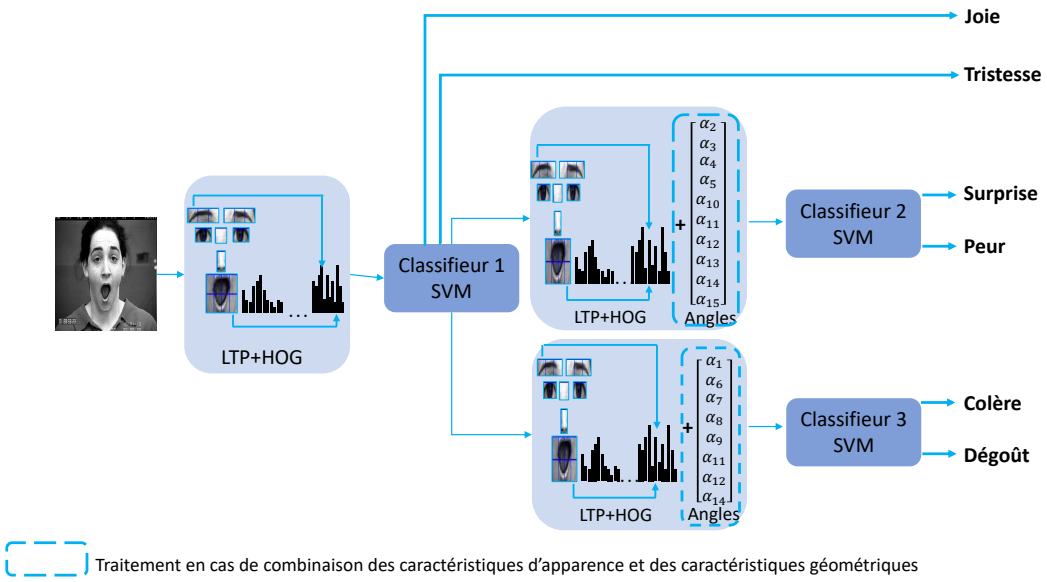


FIGURE 5.2 Aperçu schématique du système proposé en utilisant 4 groupes, à savoir : joie, tristesse, surprise/peur, colère/dégoût.

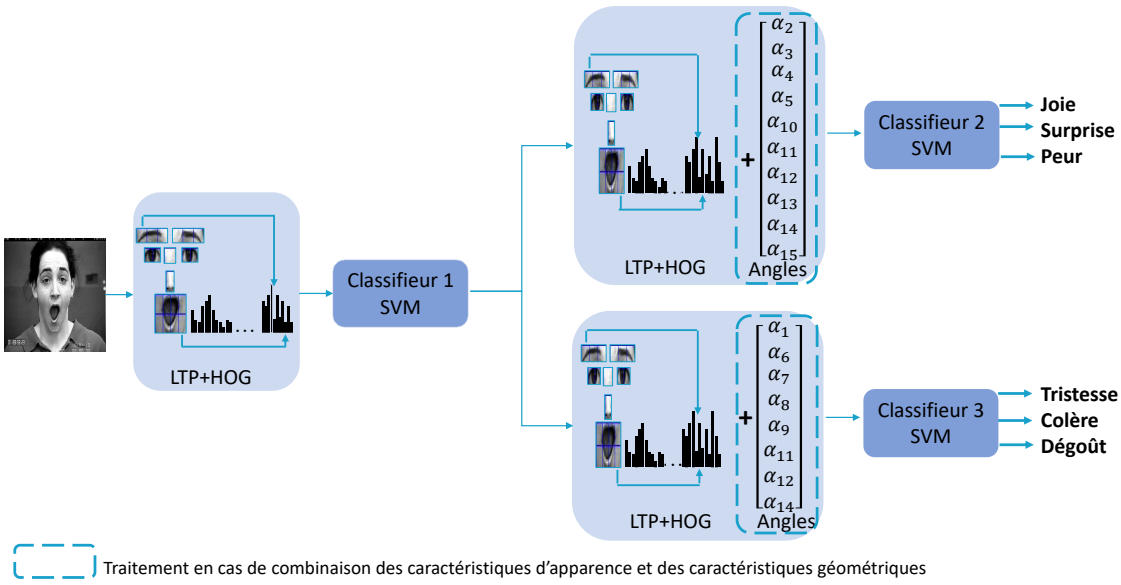


FIGURE 5.3 Aperçu schématique du système proposé en utilisant 2 groupes, à savoir : joie/surprise/peur, tristesse/colère/dégoût.

tions joie/surprise/peur tandis que le second groupe est composé des émotions tristesse/colère/dégoût.

le but de créer des groupes d'émotions est de rassembler celles qui partagent les mêmes représentations/déformations de composantes faciales. L'idée est de reconnaître un sous-ensemble plus restreint d'expressions faciales caractérisées par des confusions spécifiques entre des expressions du visage similaires sur les plans sémantique et physique.

3. (a) Dans la figure 5.2, si l'image d'entrée est reconnue en tant que joie ou tristesse, alors le traitement est fini. Sinon si elle est étiquetée en tant que groupe 3 par le "classifieur 1", alors l'étape suivante consiste à classer l'expression soit en tant que surprise ou peur. Pour effectuer la classification "Classifieur 2", nous utilisons les caractéristiques d'apparence LTP+HOG. En cas de la combinaison de caractéristiques d'apparence et géométriques, nous concaténons les caractéristiques géométriques concernant uniquement les composantes faciales "œil" et "bouche" parce que ces dernières sont des régions saillantes pour la surprise et la peur (voir Table 5.9).

   (b) Dans la figure 5.3, si l'image d'entrée est étiquetée en tant que groupe 1 par le classifieur 1 ", alors l'étape suivante consiste à classer l'expression soit en tant que joie, soit en tant que surprise ou peur. Pour effectuer la classification "Classifieur 2", nous suivons les mêmes étapes que celles présentées pour la figure 5.2.

4. (a) Dans la figure 5.2, si l'image d'entrée est classée dans le quatrième groupe, alors l'étape suivante consiste à classer l'expression soit en tant que colère ou dégoût. Pour effectuer la classification "Classifieur 3", nous utilisons les caractéristiques d'apparence LTP+HOG. En cas de la combinaison de caractéristiques d'apparence et géométriques, nous concaténons les caractéristiques géométriques concernant uniquement les composantes faciales "nez", "entre-sourcils" et "bouche" parce que ces dernières sont des régions saillantes pour la colère et le dégoût (voir Table 5.9).

   (b) Dans la figure 5.3, si l'image d'entrée est classée dans le deuxième groupe 2 par le "classifieur 1", alors l'étape suivante consiste à classer l'expression soit en tant que tristesse, soit en tant que colère ou dégoût. Pour effectuer la classification "Classifieur 3", nous suivons les mêmes étapes que celles présentées pour la figure 5.2.

La table 5.9 montre les caractéristiques géométriques extraites pour chaque groupe d'émotions. Le groupe A présente la catégorie surprise/peur avec ou sans joie, tandis que le groupe

B indique la catégorie colère/dégoût avec ou sans tristesse. Les angles pour chaque groupe sont calculés à partir des régions importantes pour l'identification et la représentation des émotions appartenant au groupe. A titre d'exemples, les quatre angles calculés dans le groupe A, pour décrire le mouvement des paupière, peuvent être très utiles pour mesurer l'élargissement des yeux quand il s'agit de l'émotion peur ou surprise. Tandis que les angles qui décrivent les déformations du nez, dans le groupe B, peuvent être très bénéfiques quand il s'agit de l'émotion dégoût.

TABLE 5.9 Les caractéristiques géométriques considérées pour chaque groupe

Groupe A : surprise/peur avec ou sans joie		**Groupe B :** colère/dégoût avec ou sans tristesse	
Composante Faciale	Angle	Composante Faciale	Angle
Bouche	$\alpha_{10} = \widehat{P_{33}P_{32}P_{43}}$ $\alpha_{11} = \widehat{P_{44}P_{32}P_{49}}$ $\alpha_{12} = \widehat{P_{49}P_{45}P_{47}}$ $\alpha_{13} = \widehat{P_{44}P_{48}P_{46}}$ $\alpha_{14} = \widehat{P_{46}P_{38}P_{47}}$ $\alpha_{15} = \widehat{P_{37}P_{38}P_{39}}$	Sourcil	$\alpha_1 = \widehat{P_6P_{11}P_5}$
		Nez	$\alpha_6 = \widehat{P_{14}P_{19}P_{15}}$ $\alpha_7 = \widehat{P_{15}P_{14}P_{19}}$ $\alpha_8 = \widehat{P_{14}P_{19}P_{15}}$ $\alpha_9 = \widehat{P_{34}P_{17}P_{36}}$
Œil	$\alpha_2 = \widehat{P_{21}P_{20}P_{25}}$ $\alpha_3 = \widehat{P_{22}P_{23}P_{24}}$ $\alpha_4 = \widehat{P_{27}P_{26}P_{31}}$ $\alpha_5 = \widehat{P_{28}P_{29}P_{30}}$	Bouche	$\alpha_{11} = \widehat{P_{44}P_{32}P_{49}}$ $\alpha_{12} = \widehat{P_{49}P_{45}P_{47}}$ $\alpha_{14} = \widehat{P_{46}P_{38}P_{47}}$

5.3.2 Expérimentation

Nous avons évalué notre approche en utilisant quatre jeux de données : CK, CK+, KDEF et FEED. Les détails relatifs à ces jeux de donnée sont présentés dans le chapitre 3.

A partir de la table 5.10, nous pouvons remarquer que l'approche basée sur "2 groupes d'émotions" fournit les meilleurs taux de reconnaissance pour les deux mesures F-score et le taux global (gTR) lorsque les jeux de données KDEF, CK et CK+ sont testés. Le F-score et le taux global (gTR) obtenus sont respectivement 93.78% et 93.75% pour le jeu de données KDEF, 99.27% et 99% pour CK et 96.1% et 96.87% pour CK+. Cependant, lorsque le jeu de données FEED est testé, l'approche basée sur "4 groupes d'émotions" montre un F-score élevée par rapport à celle basée sur "2 groupes d'émotions" (93.75% VS 92.66%), en revanche, son taux global (gTR) est inférieur à celui de l'approche basée sur "2 groupes d'émotions" (93.58% VS 94.15%). Notez que les résultats présentés dans la table 5.10 sont obtenus en utilisant la méthode utilisant les caractéristiques d'apparence "LTP+HOG".

TABLE 5.10 Taux de reconnaissance obtenus en appliquant la méthode d'apparence sur les deux approches proposées "2 groupes" et "4 groupes" sur les jeux de données KDEF, FEED, CK et CK+ avec 6 émotions de base.

	KDEF		FEED		CK		CK+	
	F-score	gTR	F-score	gTR	F-score	gTR	F-score	gTR
2 groupes	**93.78**	**93.75**	92.66	**94.15**	**99.27**	**99.21**	**96.1**	**96.87**
4 groupes	93.48	93.33	**93.75**	93.58	98.2	97.84	95.24	96.04

La table 5.11 donne un aperçu des résultats obtenus par la méthode classique (sans regroupement des émotions), présentée dans le chapitre 3 et dans la première partie de ce chapitre, et les méthodes basées sur le regroupement, en utilisant la méthode de combinaison (Apparence : LTP+HOG, Géométriques :Angles). La table 5.11 montre que l'approche basée sur "2 groupes d'émotions" atteint des performances similaires à la méthode, présentée dans la section 5.2.3.1, lorsque le jeu de données CK+ est testé. Les meilleurs taux globaux (gTR) obtenus pour les jeux de données KDEF, FEED, CK et CK+ sont respectivement 94.58%, 95.09%, 99.21% et 97.29% en utilisant l'approche basée sur "2 groupes d'émotions". Les comparaisons résumées dans la table 5.11 montrent que l'approche basée sur le regroupement des émotions qui se ressemblent peut atteindre des meilleures performances quelque soit la nature de l'émotion (posée ou spontanée).

TABLE 5.11 Comparaison des taux de reconnaissance obtenus en appliquant la méthode de combinaison sur les deux approches proposées "2 groupes" et "4 groupes" et la méthode classique (sans regroupement des émotions) sur les jeux de données KDEF, FEED, CK et CK+ avec 6 émotions de base.

	KDEF		FEED		CK		CK+	
	F-score	gTR	F-score	gTR	F-score	gTR	F-score	gTR
Méthode classique	94.2	94.16	94.24	94.15	98.28	98.43	**96.43**	**97.29**
2 groupes	**94.65**	**94.58**	93.58	**95.09**	**99.27**	**99.21**	**96.47**	**97.29**
4 groupes	94.4	94.16	**94.78**	94.15	97.92	97.64	95.79	97.08

5.3.2.1 Matrices de confusion

Les tables 5.12 et 5.13 présentent les matrices de confusion, respectives des jeux de données CK+ et FEED, associées aux meilleurs taux globaux (gTR) obtenus en utilisant l'approche basée sur "2 groupes d'émotions". Dans cette section, nous comparons les taux de reconnaissance de chaque émotion obtenus par la méthode classique (sans regroupement des émotions) et par l'approche "2 groupes d'émotion", en utilisant la méthode de combinaison pour l'extraction des caractéristiques.

Nous remarquons, à partir de la table 5.11, que les meilleurs taux de reconnaissance obtenus pour le jeux de données CK+ sont similaires dans les deux approches (classique et celle basée sur "2 groupes d'émotions"). En revanche, leurs matrices de confusion sont différentes (voir les tables 5.5 et 5.12). A partir de ces matrices de confusion, nous pouvons remarquer que l'émotion surprise a gardé le même taux de reconnaissance (99.16%). La joie et la peur sont reconnues respectivement avec des taux de reconnaissance de 100% et 96.66%. Par contre, le taux de reconnaissance de la tristesse a chuté d'environ 3%. En effet, la tristesse est mal classée en tant qu'émotion colère avec un taux d'erreur de 7.77%. Nous pouvons déduire que les caractéristiques utilisées dans le "classifieur 2" (voir Figure 5.3) pour reconnaitre les émotions du premier groupe (joie/surprise/peur) ont pu permettre une meilleure séparation entre les trois émotions (joie, surprise, et peur).

TABLE 5.12 Matrice de confusion de la méthode de combinaison appliquée sur le regroupement en "2 groupes" calculée sur le jeu de données CK+ (associée aux F-score :96.47% et gTR :97.29%)

	Joie	Surprise	Peur	Tristesse	Colère	Dégoût
Joie	**100**	0	0	0	0	0
Surprise	0	**99.16**	0.83	0	0	0
Peur	3.33	0	**96.66**	0	0	0
Tristesse	0	0	0	**91.11**	7.77	1.11
Colère	0	0	0	2	**93.33**	4.66
Dégoût	0	0	0	0	1.66	**98.33**

D'après la table 5.13, nous pouvons observer que la joie est reconnue avec un taux de reconnaissance maximal de 100%. Le taux de reconnaissance obtenu pour la surprise est identique à celui atteint en utilisant la méthode classique (sans regroupement des émotions) (voir Table 5.8). Le dégoût et la tristesse ont été reconnus avec des taux élevés (environ 91-96%). Cependant, nous remarquons une chute considérable du taux de la peur (80%). La diminution du taux de la peur vient principalement de la mauvaise classification causée par le "classifieur 1" (voir Figure 5.3) qui n'a pas pu attribuer les images de la peur au premier groupe (joie/surprise/peur). En effet, la peur est mal classée en tant qu'émotions tristesse et colère respectivement avec des taux d'erreur de 10% et 3.3%.

TABLE 5.13 Matrice de confusion de la méthode de combinaison appliquée sur le regroupement en "2 groupes" calculée sur le jeu de données FEED (associée aux F-score :93.58% et gTR :95.09%)

	Joie	Surprise	Peur	Tristesse	Colère	Dégoût
Joie	**100**	0	0	0	0	0
Surprise	0	**99**	1	0	0	0
Peur	0	6.66	**80**	10	3.33	0
Tristesse	0	0	0	**96**	3	1
Colère	0	0	0	6	**94**	0
Dégoût	0	3	0	1	5	**91**

5.3.3 Regroupement des émotions en utilisant les six émotions de base, ainsi que la neutralité

Dans cette section, nous présentons les résultats de reconnaissance des six expressions faciales, ainsi que l'état neutre, en utilisant la nouvelle approche par regroupement des émotions. Pour ce faire, nous introduisons la neutralité suivant différents schémas :

1. "3 goupes d'émotions" : Neutralité, joie/peur/surprise, colère/dégoût/tristesse.
2. "5 goupes d'émotions" : Neutralité, joie, tristesse, peur/surprise, colère/dégoût
3. "2 goupes d'émotions" : Neutralité/joie/peur/surprise, Neutralité/colère/dégoût/tristesse

Les différents schémas proposés, ainsi que la méthode classique proposée dans le chapitre 3 sont comparés en utilisant la méthode d'apparence "LTP+HOG" pour l'extraction des caractéristiques. Les comparaisons résumées dans la table 5.14 montrent que les approches basées sur le regroupement peuvent améliorer les performances de reconnaissance des six émotion de base, outre la neutralité. Nous remarquons que l'approche basée sur "3 groupes d'émotions" offre des performances supérieures en ce qui concerne les jeux de données KDEF et CK+. Cependant, l'approche basée sur "2 groupes" donne les meilleures performances

pour la base CK. Par contre, aucune de ces approches de regroupement n'a pu améliorer le taux de reconnaissance sur le jeu de données FEED composé d'émotions spontanées.

TABLE 5.14 Comparaison des taux de reconnaissance obtenus en appliquant la méthode d'apparence sur les trois approches proposées et la méthode classique présentée dans le chapitre 3 sur les jeux de données KDEF, FEED, CK et CK+ avec 6 émotions de base, ainsi que la neutralité.

	KDEF		FEED		CK		CK+	
	F-score	gTR	F-score	gTR	F-score	gTR	F-score	gTR
Méthode classique	93.34	93.24	**92.03**	**91.8**	96.06	95.7	94.63	96.03
3 groupes^{+}	**95.01**	**95**	85.3	84.76	93.39	91.96	**95.74**	**96.19**
5 groupes^{++}	**95.13**	**95**	88.77	87.93	95.15	94.42	93.83	95.07
2 groupes^{+++}	94.4	94.28	85.03	86.03	**96.65**	**96.39**	94.26	95.71

$^{+}$ (Neutralité, joie/peur/surprise, colère/dégoût/tristesse)

$^{++}$ (Neutralité, joie, tristesse, peur/surprise, colère/dégoût)

$^{+++}$ (Neutralité/joie/peur/surprise, Neutralité/colère/dégoût/tristesse)

5.4 Conclusion

Dans ce chapitre, nous avons dans un premier temps présenté un nouveau descripteur géométrique, ensuite, nous avons combiné ce descripteur avec le descripteur d'apparence LTP+HOG, pour pouvoir atteindre une meilleure reconnaissance des expressions faciales. Ensuite, une nouvelle approche a été proposée en se basant sur le regroupement des émotions dont les caractéristiques se ressemblent et peuvent engendrer des confusions (par exemple, la surprise et la peur). Les principales conclusions de l'étude sont les suivantes :

1. L'utilisation du descripteur géométrique seul n'a pas pu donner de bons résultats parce que nous n'utilisons aucune normalisation et nous calculons les angles proposés directement sur un visage expressif sans considérer l'expression neutre. En effet, l'utilisation de l'expression neutre à partir de laquelle la déformation de la forme des composantes faciales peut être mesurée est une étape indispensable pour obtenir des caractéristiques géométriques fiables et plus discriminantes. Puisque nous cherchons à développer un système indépendant de tous paramètre extérieur, cette étape a été éliminée de notre système pour la raison suivante : la difficulté d'acquérir une expression neutre dans un cas réel.
2. Des expériences sur quatre jeux de données démontrent que notre système peut atteindre des performances de reconnaissance supérieures en matière de discrimination fondée sur le regroupement des émotions et la combinaison de plusieurs caractéristiques.

3. Pour la reconnaissance des six émotions de base, l'approche basée sur "2 groupes d'émotion", en utilisant la combinaison des caractéristiques d'apparence et géométriques, a pu montrer son efficacité sur les quatre jeux de données testés.
4. Pour la reconnaissance des six émotions de base ainsi que la neutralité, nous ne pouvons pas déduire quelle méthode peut être adéquate pour la reconnaissance des expressions faciales, parce que pour chaque jeu de données, une seule méthode (différente à chaque fois) se démarque des autres. Le système peut être amélioré en intégrant des caractéristiques géométriques, parce que dans cette expérience des sept émotions, nous avons utilisé uniquement les caractéristiques d'apparence.
5. Le système basé sur le regroupement des émotions pourrait être considéré comme le pionnier dans la reconnaissance des expressions faciales en regroupant les émotions qui se ressemblent.
6. Nous estimons que la tâche d'analyse et de reconnaissance de l'expression pourrait être effectuée de manière plus favorable, si seules certaines régions (c-à-d, des régions faciales saillantes qui ont de fortes capacités discriminantes pour chaque groupe d'émotions) sont sélectionnées pour l'extraction des caractéristiques. Comme cela est définie par Ekman et Friesen [59], chaque émotion de base est décrite en utilisant des indices spécifiques décrivant l'activité faciale (voir Table1.2 du chapitre 1).

Conclusion

Les expressions du visage jouent un rôle important et fondamental dans la communication interpersonnelle et sociale et révèlent une grande variété d'informations telles que l'émotion, l'identité, le sexe et l'âge. La reconnaissance automatique des expressions faciales est devenue un domaine de recherche actif au cours des dernières décennies, avec un certain nombre d'applications importantes telles que l'interaction homme-machine, neuromarketing, la surveillance visuelle et la sécurité. Bien que les humains détectent et analysent les visages et les expressions faciales dans une scène avec un effort quasi négligeable, le développement d'un système automatique pour accomplir cette tâche présente de nombreuses difficultés [156]. L'approche globale de l'analyse automatique des expressions faciales comprend généralement trois étapes. Étant donné une image d'entrée ou une séquence d'images, la première étape consiste à localiser le visage, détecter un ensemble de points faciaux, qui sont ensuite utilisés pour effectuer l'enregistrement du visage. Une fois le visage enregistré, l'étape suivante concerne l'extraction des caractéristiques du visage. A cet effet, différentes caractéristiques géométriques et/ou d'apparence peuvent être utilisées. L'étape finale prend comme entrée le vecteur de caractéristiques extrait précédemment pour effectuer la tâche de classification en utilisant une technique d'apprentissage automatique.

Dans ce qui suit, nous décrivons nos contributions dans chacune de ces étapes et donnons un aperçu des résultats obtenus.

Etape 1 : Enregistrement du visage

L'enregistrement du visage est une étape fondamentale pour la REF. En général, il vise à trouver la ou les région(s) d'intérêt (ROI) du visage. Dans l'étude de la littérature (pour référence, voir le chapitre 1), différentes méthodes de reconnaissance des expressions faciales sont décrites. Pour analyser le visage et afin d'obtenir des informations discriminantes, certaines méthodes utilisent tout le visage comme une seule ROI et d'autres extraient des ROIs manuellement ou en se basant sur la géométrie du visage. Il est démontré dans ce travail de recherche que la tâche de l'analyse et de la reconnaissance de l'expression faciale

peut être effectuée de manière plus favorable en se basant sur le système FACS. Selon FACS, les principales caractéristiques faciales requises pour analyser les expressions faciales sont situées autour des sourcils, des yeux, du nez et de la bouche. Notre méthode d'extraction des ROIs est différente de celles mentionnées dans le chapitre 1. Elle est basée sur des nouvelles régions faciales définies plus précisément à l'aide de points faciaux (détectés par SDM), permettant d'extraire automatiquement sept ROIs (sourcil gauche, droite sourcil, entre les sourcils, œil gauche, œil droit, nez, bouche). De plus, contrairement aux travaux de l'état de l'art où une région peut contenir plus d'une composante faciale, notre méthode vise à séparer les composantes faciales. Cela garantit un meilleur enregistrement du visage et par la suite une représentation faciale appropriée. Dans la section 3.8.2 du chapitre 3, nous avons évalué à travers des expériences approfondies les performances de la décomposition faciale proposée, en la comparant avec l'utilisation du visage entier et les décompositions faciales de l'état de l'art. L'évaluation a été réalisée en utilisant plusieurs bases de données publiques représentant des émotions spontanées (la base FEED), des émotions posées (les bases CK, CK+, KDEF et JAFFE) et des émotions extraites de films (la base SFEW).

Etape 2 : Extraction des caractéristiques

Une fois l'enregistrement du visage effectué, l'étape suivante consiste à extraire et représenter les changements faciaux causés par les expressions faciales. Pour ce faire, nous avons mené une large expérimentation en utilisant différents descripteurs d'apparence tels que les descripteurs de texture (LBP, LTP, CLBP), le descripteur de forme (HOG) et leur combinaison (LBP+HOG, CLBP+HOG et LTP+HOG). L'évaluation des descripteurs a démontré que les descripteurs hybrides construits par une concaténation hétérogène des caractéristiques de texture et de forme sont les meilleurs, en particulier la concaténation de LTP et du HOG (pour référence, voir le chapitre 3).

D'après la revue de la littérature (voir chapitre 1), nous savons que les performances de la méthode utilisant la combinaison des caractéristiques d'apparence et géométriques surpassent celles de la méthode géométrique ou d'apparence. Par conséquent, dans l'objectif d'augmenter les performances de REF à travers des descripteurs de caractéristiques, nous avons proposé un nouveau descripteur géométrique (Geo) pour l'analyse des caractéristiques faciales (voir chapitre 5). L'utilisation de ce descripteur seul n'a pas donné de bons résultats mais sa combinaison avec les descripteurs d'apparence (LTP+HOG) a montré une bonne adéquation sur toutes les bases de données testées, en atteignant des taux de reconnaissance très prometteurs.

Etape 3 : Reconnaissance des expressions faciales

REF statique : Pour la REF à partir des images statiques, nous avons examiné deux différentes méthodes d'apprentissage automatique : SVM avec ses différents noyaux (polynomial, RBF et linéaire) et RF. Nos expériences ont montré que les meilleures performances de reconnaissance sont obtenues en utilisant le classifeur SVM avec un noyau linéaire (voir chapitre 3).

REF dynamique : Pour la REF à partir des séquences d'images, nous avons présenté une nouvelle méthode de REF dynamique avec un algorithme simple. Cet algorithme est capable d'estimer l'émotion à partir d'une multi-observation. La multi-observation peut représenter une séquence vidéo de la même personne, un sous-ensemble d'images de différentes personnes ou une expression d'une personne capturée à partir de différents points de vue. L'estimation des probabilités en sortie d'un classifieur SVM, associées aux observation, sont exploitées en suivant différentes stratégies afin d'affecter une classe (émotion) à un sous-ensemble d'images.

Pour conclure, nos expériences ont montré que les meilleures performances de reconnaissance sont obtenues en utilisant le classifeur SVM (noyau linéaire) avec la combinaison des caractéristiques géométriques et des caractéristiques d'apparence LTP+HOG extraites à partir de la décomposition faciale proposée (voir Chapitres 3 et 5).

Le système proposé dans la deuxième partie du chapitre 5 est le fruit de l'exploration de toutes les méthodes proposées, les techniques utilisées et les analyses effectuées dans cette thèse pour la REF statique. Il est basé sur la décomposition faciales en sept ROIs, le descripteur hybride (LTP+HOG+Geo) et le classifeur SVM. La conception de ce système est basée principalement sur l'analyse des matrices de confusion issues des émotions posées et spontanées. L'approche proposée consiste à regrouper les émotions qui partageraient les mêmes changements faciaux (par exemple, surprise/peur et colère/dégoût). Ce système a montré son efficacité sur plusieurs jeux de données lorsque l'évaluation est effectuée sur les six émotions de base. Le bilan des résultats obtenus par les méthodes proposées sur différentes bases de données est présenté dans la table ci-dessous

Bilan des résultats obtenus par les différentes méthodes proposées dans ce travail de recherche (+ Approche de regroupement des émotions).

Chapitre de référence	Etape 1	Etape 2	jeu de données	# d'émotions	F-score	gTR
Chapitre 3	ROIs	LTP+HOG	CK+ (images)	7	94.63	96.03
Chapitre 4	ROIs	LTP+HOG	CK+ (Séquences d'images)	7	–	99.99
Chapitre 5	ROIs	Geo	CK+ (images)	7	80.34	83.17
Chapitre 5	ROIs	LTP+HOG+Geo	CK+ (images)	7	95.71	94.62
Chapitre 5^+	ROIs	LTP+HOG	CK+ (images)	7	95.74	96.19
Chapitre 3	ROIs	LTP+HOG	CK (images)	7	96.06	95.7
Chapitre 5	ROIs	Geo	CK (images)	7	74.88	73.11
Chapitre 5	ROIs	LTP+HOG+Geo	CK (images)	7	96.73	96.39
Chapitre 5^+	ROIs	LTP+HOG	CK (images)	7	96.65	96.39
Chapitre 3	ROIs	LTP+HOG	KDEF (images)	7	93.34	93.24
Chapitre 5	ROIs	Geo	KDEF (images)	7	73.14	73.21
Chapitre 5	ROIs	LTP+HOG+Geo	KDEF (images)	7	93.7	93.57
Chapitre 5^+	ROIs	LTP+HOG	KDEF (images)	7	95.13	95
Chapitre 3	ROIs	LTP+HOG	FEED (images)	7	92.03	91.8
Chapitre 5	ROIs	Geo	FEED (images)	7	60.12	62.85
Chapitre 5	ROIs	LTP+HOG+Geo	FEED (images)	7	93.21	93.01
Chapitre 5^+	ROIs	LTP+HOG	FEED (images)	7	88.77	87.93
Chapitre 3	ROIs	LTP+HOG	JAFFE (images)	7	77.08	–
Chapitre 3	ROIs	LTP+HOG	SFEW (images)	7	–	38.27
Chapitre 4	ROIs	LTP+HOG	OULU-CASIA (Séquences d'images)	6	–	94.1
Chapitre 3	ROIs	LTP+HOG	CK+ (images)	6	96.01	96.77
Chapitre 5	ROIs	Geo	CK+ (images)	6	85.6	87.6
Chapitre 5	ROIs	LTP+HOG+Geo	CK+ (images)	6	96.43	97.29
Chapitre 5^+	ROIs	LTP+HOG	CK+ (images)	6	96.1	96.87
Chapitre 5^+	ROIs	LTP+HOG+Geo	CK+ (images)	6	96.47	97.27
Chapitre 5	ROIs	LTP+HOG	CK (images)	6	98.29	98.03
Chapitre 5	ROIs	Geo	CK (images)	6	82.83	81.96
Chapitre 5	ROIs	LTP+HOG+Geo	CK (images)	6	98.28	98.43

Suite à la page suivante

suite de la page précédente

Chapitre de référence	Etape 1	Etape 2	jeu de données	# d'émotions	F-score	gTR
Chapitre 5^+	ROIs	LTP+HOG	CK (images)	6	99.27	99.21
Chapitre 5^+	ROIs	LTP+HOG+Geo	CK (images)	6	99.27	99.21
Chapitre 5	ROIs	LTP+HOG	KDEF (images)	6	93.4	93.33
Chapitre 5	ROIs	Geo	KDEF (images)	6	77.43	77.5
Chapitre 5	ROIs	LTP+HOG+Geo	KDEF (images)	6	94.2	94.21
Chapitre 5^+	ROIs	LTP+HOG	KDEF (images)	6	93.78	93.75
Chapitre 5^+	ROIs	LTP+HOG+Geo	KDEF (images)	6	94.65	94.58
Chapitre 5	ROIs	LTP+HOG	FEED (images)	6	94.39	94.52
Chapitre 5	ROIs	Geo	FEED (images)	6	66.62	68.11
Chapitre 5	ROIs	LTP+HOG+Geo	FEED (images)	6	94.24	94.15
Chapitre 5^+	ROIs	LTP+HOG	FEED (images)	6	92.66	94.15
Chapitre 5^+	ROIs	LTP+HOG+Geo	FEED (images)	6	93.58	95.09

Perspectives

Bien que de nombreux progrès ont été réalisés dans cette thèse, nos travaux présentent cependant quelques limites. Ces limites ainsi que les orientations possibles devant être abordées dans les travaux futurs, sont énumérées comme suit :

- Nous avons remarqué dans la section 3.8.2 du chapitre 3 que le choix de la taille et du nombre de blocs de chaque ROI affecte généralement les performances de reconnaissance. En effet, la taille et le nombre de blocs optimaux des ROIs varient en fonction des données d'apprentissage. Par conséquent, le problème est qu'il est difficile d'aller vers un système générique.

  Le fait que chaque ROI contienne une composante du visage est le facteur principal qui a contribué à améliorer la REF. Autrement dit, la taille des composantes faciales est différente d'une personne à l'autre et d'une base de données à l'autre suivant la forme du visage de chaque personne et la distance entre la caméra et le visage capturé. Ceci a engendré de nombreuses propositions pour définir les tailles optimales, ce qui a rendu complexe le choix d'une configuration standard pour toutes les bases de données.

  Pour remédier à ce problème, nous suggérons que la définition des ROIs soit indépendante de la taille des composantes et nous préconisons d'utiliser des petites régions

(dénotée patchs) entourant uniquement les points caractéristiques. Pour ce faire, nous pouvons nous inspirer du travail proposé par [79].

- La dégradation du taux de reconnaissance, lorsque le descripteur géométrique est utilisé seul, est peut-être due au calcul des angles proposés sur un visage expressif sans considérer l'expression neutre pour mesurer la déformation des composantes faciales, c-à-d, sans procéder à l'étape de normalisation. En effet, l'utilisation d'un visage neutre pour normaliser les caractéristiques géométriques permettrait d'obtenir un descripteur géométrique fiable et plus discriminant. Puisque l'expression neutre n'est pas toujours disponible, nous envisageons de construire un modèle pour l'émotion "neutre" à partir de plusieurs visages neutres de différentes bases de données. Ce modèle sera ensuite considéré dans l'étape de normalisation des caractéristiques géométriques.
- Pour l'approche dynamique proposée dans le chapitre 4, la construction des ensembles de test et d'apprentissage pour effectuer la reconnaissance des expressions faciales à partir des séquences d'images est basée essentiellement sur les images représentant le pic de l'émotion (c-à-d, les images appartenant au segment temporel "apex"). Puisque cette approche n'était pas conçue pour détecter les segments temporels (neutre, onset, apes et offset), elle a été évaluée uniquement sur les bases de données où les séquences d'images débutent par un état neutre et finissent par le pic de l'expression comme dans les bases CK+ et Oulu-CASIA. En revanche, nous n'avons pas pu utiliser la base de données MMI puisque ses séquences d'images commencent par le segment temporel "neutre" et se terminent par le segment "offset". Nous souhaitons alors concevoir un système permettant de détecter les différents segments temporels comme dans les systèmes proposés par [212, 193, 185].
- Nous souhaitons également étendre l'évaluation de l'approche dynamique sur les séquences d'images multi-vues, ainsi que les séquences d'images des expressions faciales non contraintes des différentes poses de tête et des occlusions comme dans la base AFEW [48]. Deux autres paramètres doivent être pris en considération pour s'y faire : l'occlusion et la variation de l'angle de la caméra.
- Nous estimons que la tâche d'analyse et de reconnaissance de l'expression pourrait être effectuée de manière plus favorable, si seules certaines régions (c-à-d, des régions faciales saillantes qui ont de fortes capacités discriminantes pour chaque émotion particulière) sont sélectionnées pour l'extraction des caractéristiques. Comme cela a été défini par Ekman et Friesen [59], chaque émotion de base a été décrite en utilisant des indices spécifiques décrivant l'activité faciale (voir Table 1.2 du chapitre 1). Ainsi, l'extraction des caractéristiques en ne considérant que des régions faciales saillantes réduirait considérablement les dimensions des vecteurs caractéristiques et le temps de

calcul. Cela permettrait de rendre les algorithmes de REF plus adaptés aux applications en temps réel.

- La REF en utilisant des régions faciales saillantes spécifiques à chaque émotion peut être effectuée en procédant, dans un premier temps, à la détection des unités d'actions (AUs). Puis, l'émotion peut être déduite en se basant sur la combinaison des AUs correspondant à chaque émotion (voir Table 1.2 du chapitre 1). Pour la détection des AUs, plusieurs travaux ont été proposés dans la littérature [209, 95, 34, 3, 21].
- Enfin, nous envisageons d'étudier la reconnaissance d'autres émotions telle que la honte, la fierté, la culpabilité, etc. Pour cela, d'autres paramètres doivent être pris en considération comme l'orientation du visage, le suivi du regard, la couleur de la peau, etc.

Bibliographie

[1] Abdulrahman, M., Gwadabe, T. R., Abdu, F. J., and Eleyan, A. (2014). Gabor wavelet transform based facial expression recognition using pca and lbp. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 2265–2268. IEEE.

[2] Acevedo, D., Negri, P., Buemi, M. E., Fernández, F. G., and Mejail, M. (2017). A simple geometric-based descriptor for facial expression recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 802–808. IEEE.

[3] Adams, A. and Robinson, P. (2015). Automated recognition of complex categorical emotions from facial expressions and head motions. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 355–361. IEEE.

[4] Ahmed, F., Bari, H., and Hossain, E. (2014). Person-independent facial expression recognition based on compound local binary pattern (clbp). *Int. Arab J. Inf. Technol.*, 11(2) :195–203.

[5] Ahmed, F. and Hossain, E. (2013). Automated facial expression recognition using gradient-based ternary texture patterns. *Chinese Journal of Engineering*, 2013.

[6] Allaert, B., Bilasco, I. M., and Djeraba, C. (2018). Advanced local motion patterns for macro and micro facial expression recognition. *arXiv preprint arXiv :1805.01951*.

[7] Almaev, T. R. and Valstar, M. F. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 356–361. IEEE.

[8] Alves, N. T. (2013). Recognition of static and dynamic facial expressions : a study review. *Estudos de Psicologia (Natal)*, 18(1) :125–130.

[9] Ambadar, Z., Schooler, J. W., and Cohn, J. F. (2005). Deciphering the enigmatic face : The importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*, 16(5) :403–410.

[10] Anderson, K. and McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(1) :96–105.

[11] Arshid, S., Hussain, A., Munir, A., Nawaz, A., and Aziz, S. (2017). Multi-stage binary patterns for facial expression recognition in real world. *Cluster Computing*, pages 1–9.

[12] Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., and Solomon, P. E. (2009). The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12) :1788–1796.

[13] Ballard, D. H. (1981). Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2) :111–122.

[14] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE.

[15] Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2) :253–263.

[16] Bartlett, M. S., Littlewort, G. C., Sejnowski, T., and Movellan, J. (2003). A prototype for automatic recognition of spontaneous facial actions. In *Advances in neural information processing systems*, pages 1295–1302.

[17] Bassili, J. N. (1979). Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11) :2049.

[18] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7) :711–720.

[19] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12) :2930–2940.

[20] Bian, P., Jin, Y., and Cao, J. (2018). Facial landmark detection via elm feature selection and improved sdm. In *Proceedings of ELM-2016*, pages 217–228. Springer.

[21] Bishay, M. and Patras, I. (2017). Fusing multilabel deep networks for facial action unit detection. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 681–688. IEEE.

[22] Black, M. J. and Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1) :23–48.

[23] Bradski, G. et al. (2000). The opencv library. *Doctor Dobbs Journal*, 25(11) :120–126.

[24] Buciu, I., Pitas, I., et al. (2003). Ica and gabor representation for facial expression recognition. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–855. IEEE.

[25] Bull, P. (2001). State of the art : Nonverbal communication. *The Psychologist*, 14(12) :644–647.

[26] Carcagnì, P., Coco, M., Leo, M., and Distante, C. (2015). Facial expression recognition and histograms of oriented gradients : a comprehensive study. *SpringerPlus*, 4(1) :1.

[27] Carrera-Levillain, P. and Fernandez-Dols, J.-M. (1994). Neutral faces in context : Their emotional meaning and their function. *Journal of Nonverbal Behavior*, 18(4) :281–299.

[28] Chang, C.-C. and Lin, C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3) :27.

[29] Chang, J.-Y. and Chen, J.-L. (1999). A facial expression recognition system using neural networks. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 5, pages 3511–3516. IEEE.

[30] Chang, K.-Y., Liu, T.-L., and Lai, S.-H. (2009). Learning partially-observed hidden conditional random fields for facial expression recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 533–540. IEEE.

[31] Chang, Y., Hu, C., Feris, R., and Turk, M. (2006). Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6) :605–614.

[32] Chen, J., Chen, Z., Chi, Z., and Fu, H. (2014). Facial expression recognition based on facial components detection and hog features. In *International Workshops on Electrical and Computer Engineering Subfields*, pages 884–888.

[33] Chen, J., Chen, Z., Chi, Z., and Fu, H. (2015). Dynamic texture and geometry features for facial expression recognition in video. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4967–4971. IEEE.

[34] Chu, W.-S., De la Torre, F., and Cohn, J. F. (2013). Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522.

[35] Cohen, I., Sebe, N., Garg, A., Chen, L. S., and Huang, T. S. (2003a). Facial expression recognition from video sequences : temporal and static modeling. *Computer Vision and image understanding*, 91(1-2) :160–187.

[36] Cohen, I., Sebe, N., Gozman, F., Cirelo, M. C., and Huang, T. S. (2003b). Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[37] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

[38] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6) :681–685.

[39] Cootes, T. F., Hill, A., Taylor, C. J., and Haslam, J. (1993). The use of active shape models for locating structures in medical images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 33–47. Springer.

[40] Coughlan, J. M. and Ferreira, S. J. (2002). Finding deformable shapes using loopy belief propagation. In *European Conference on Computer Vision*, pages 453–468. Springer.

[41] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

[42] Dapogny, A., Bailly, K., and Dubuisson, S. (2018). Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision*, 126(2-4) :255–271.

[43] Darwin, C. (1872). The expression of emotion in man and animals.

[44] Darwin, C. (1965). *The expression of the emotions in man and animals*, volume 526. University of Chicago press.

[45] De la Torre, F. and Cohn, J. F. (2011). Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer.

[46] Deng, H.-B., Jin, L.-W., Zhen, L.-X., Huang, J.-C., et al. (2005). A new facial expression recognition method based on local gabor filter bank and pca plus lda. *International Journal of Information Technology*, 11(11) :86–96.

[47] Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2011). Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE.

[48] Dhall, A., Goecke, R., Lucey, S., Gedeon, T., et al. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3) :34–41.

[49] Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10) :974–989.

[50] Donia, M. M., Youssif, A. A., and Hashad, A. (2014). Spontaneous facial expression recognition based on histogram of oriented gradients descriptor. *Computer and Information Science*, 7(3) :31.

[51] Eckman, P. (2003). Emotions revealed. *St. Martin's Griffin, New York*.

[52] Egede, J., Valstar, M., and Martinez, B. (2017). Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 689–696. IEEE.

[53] Ekman, P. (1989). The argument and evidence about universals in facial expres-sions. *Handbook of social psychophysiology*, pages 143–164.

[54] Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4) :384.

[55] Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1) :205–221.

[56] Ekman, P. and Friesen, W. (1978a). Facial action coding system : a technique for the measurement of facial movement. *Palo Alto : Consulting Psychologists*.

[57] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2) :124.

[58] Ekman, P. and Friesen, W. V. (1977). Facial action coding system.(1977). *Consulting Psycologists Press.*, 3 :90–28.

[59] Ekman, P. and Friesen, W. V. (1978b). *Manual for the facial action coding system*. Consulting Psychologists Press.

[60] Ekman, P. and Friesen, W. V. (2003). *Unmasking the face : A guide to recognizing emotions from facial clues*. Ishk.

[61] Ekman, P. and O'sullivan, M. (1991). Who can catch a liar ? *American psychologist*, 46(9) :913.

[62] Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[63] Fan, X. and Tjahjadi, T. (2017). A dynamic framework based on local zernike moment and motion history image for facial expression recognition. *Pattern Recognition*, 64 :399–406.

[64] Fang, T., Zhao, X., Ocegueda, O., Shah, S. K., and Kakadiaris, I. A. (2011). 3d facial expression recognition : A perspective on promises and challenges. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 603–610. IEEE.

[65] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis : a survey. *Pattern recognition*, 36(1) :259–275.

[66] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9) :1627–1645.

[67] Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International journal of computer vision*, 61(1) :55–79.

[68] Fernández-Dols, J.-M., Wallbott, H., and Sanchez, F. (1991). Emotion category accessibility and the decoding of emotion from facial expression and context. *Journal of Nonverbal Behavior*, 15(2) :107–123.

[69] Frank, M. G., Ekman, P., and Friesen, W. V. (1993). Behavioral markers and recognizability of the smile of enjoyment. *Journal of personality and social psychology*, 64(1) :83.

[70] Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

[71] Ghimire, D., Jeong, S., Lee, J., and Park, S. H. (2017). Facial expression recognition based on local region specific features and support vector machines. *Multimedia Tools and Applications*, 76(6) :7803–7821.

[72] Ghimire, D., Jeong, S., Yoon, S., Choi, J., and Lee, J. (2015). Facial expression recognition based on region specific appearance and geometric features. In *Digital Information Management (ICDIM), 2015 Tenth International Conference on*, pages 142–147. IEEE.

[73] Gosselin, P., Kirouac, G., and Doré, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. *Journal of personality and social psychology*, 68(1) :83.

[74] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, 28(5) :807–813.

[75] Gu, H. and Ji, Q. (2005). Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, 16(2) :105–115.

[76] Guo, Y., Zhao, G., and Pietikäinen, M. (2012). Dynamic facial expression recognition using longitudinal facial expression atlases. In *Computer Vision–ECCV 2012*, pages 631–644. Springer.

[77] Guo, Y., Zhao, G., and Pietikäinen, M. (2016). Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing*, 25(5) :1977–1992.

[78] Gupta, O., Raviv, D., and Raskar, R. (2017). Multi-velocity neural networks for facial expression recognition in videos. *IEEE Transactions on Affective Computing*.

[79] Happy, S. and Routray, A. (2015a). Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1) :1–12.

[80] Happy, S. and Routray, A. (2015b). Robust facial expression classification using shape and appearance features. In *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*, pages 1–5. IEEE.

[81] Hasani, B. and Mahoor, M. H. (2017). Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 790–795. IEEE.

[82] Hernandez-Matamoros, A., Bonarini, A., Escamilla-Hernandez, E., Nakano-Miyatake, M., and Perez-Meana, H. (2016). Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach. *Knowledge-Based Systems*, 110 :1–14.

[83] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554.

[84] Holder, R. P. and Tapamo, J. R. (2017). Improved gradient local ternary patterns for facial expression recognition. *EURASIP Journal on Image and Video Processing*, 2017(1) :42.

[85] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2) :415–425.

[86] Huang, C., Ai, H., Li, Y., and Lao, S. (2007). High-performance rotation invariant multiview face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4) :671–686.

[87] Huang, D., Shan, C., Ardebilian, M., and Chen, L. (2011). Facial image analysis based on local binary patterns : A survey. *IEEE Transactions on Image Processing*.

[88] Huang, X., Zhao, G., Hong, X., Pietikäinen, M., and Zheng, W. (2013). Texture description with completed local quantized patterns. In *Scandinavian Conference on Image Analysis*, pages 1–10. Springer.

[89] Huang, X., Zhao, G., Hong, X., Zheng, W., and Pietikäinen, M. (2016). Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175 :564–578.

[90] Ibrahim, M., Alam Efat, M., Kayesh, S., Khaled, S. M., Shoyaib, M., and Abdullah-Al-Wadud, M. (2014). Dynamic local ternary pattern for face recognition and verification. In *Proceedings of the International Conference on Computer Engineering and Applications, Tenerife, Spain*, volume 1012.

[91] Jack, R. E., Garrod, O. G., and Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2) :187–192.

[92] Jain, S., Hu, C., and Aggarwal, J. K. (2011). Facial expression recognition with temporal modeling of shapes. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1642–1649. IEEE.

[93] Jaiswal, S. and Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE.

[94] Jiang, B., Valstar, M., Martinez, B., and Pantic, M. (2014). A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE transactions on cybernetics*, 44(2) :161–174.

[95] Jiang, B., Valstar, M. F., and Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE.

[96] Jones, M. and Viola, P. (2003). Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3(14) :2.

[97] Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2983–2991. IEEE.

[98] Kacem, A., Daoudi, M., Amor, B. B., and Paiva, J. C. Á. (2017). A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *ICCV*, pages 3199–3208.

[99] Kaltwang, S., Rudovic, O., and Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*, pages 368–377. Springer.

[100] Kamarol, S. K. A., Jaward, M. H., Kälviäinen, H., Parkkinen, J., and Parthiban, R. (2017). Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. *Pattern Recognition Letters*, 92 :25–32.

[101] Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE.

[102] Kapoor, A., Qi, Y., and Picard, R. W. (2003). Fully automatic upper facial action recognition. In *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pages 195–202. IEEE.

[103] Kätsyri, J., Saalasti, S., Tiippana, K., von Wendt, L., and Sams, M. (2008). Impaired recognition of facial emotions from low-spatial frequencies in asperger syndrome. *Neuropsychologia*, 46(7) :1888–1897.

[104] Kazemi, V. and Josephine, S. (2014). One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society.

[105] Khademi, M., Manzuri-Shalmani, M. T., Kiapour, M. H., and Kiaei, A. A. (2010). Recognizing combinations of facial action units with different intensity using a mixture of hidden markov models and neural network. In *International Workshop on Multiple Classifier Systems*, pages 304–313. Springer.

[106] Kharat, G. U. and Dudul, S. V. (2009). Emotion recognition from facial expression using neural networks. In *Human-Computer Systems Interaction*, pages 207–219. Springer.

[107] Khorrami, P., Paine, T., and Huang, T. (2015). Do deep neural networks learn facial action units when doing expression recognition ? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27.

[108] Kim, B.-K., Roh, J., Dong, S.-Y., and Lee, S.-Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2) :173–189.

[109] Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, 16(1) :172–187.

[110] Kotsia, I., Zafeiriou, S., and Pitas, I. (2008). Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, 41(3) :833–851.

[111] Kumar, V. and Basha, A. S. A. (2014). Facial expression recognition using wavelet and k-nearest neighbour. In *Current Trends in Engineering and Technology (ICCTET), 2014 2nd International Conference on*, pages 48–52. IEEE.

[112] Kuo, C.-M., Lai, S.-H., and Sarkis, M. (2018). A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2121–2129.

[113] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

[114] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8) :1377–1388.

[115] Lee, T. S. (1996). Image representation using 2d gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 18(10) :959–971.

[116] Lei, G., Li, X.-h., Zhou, J.-l., and Gong, X.-g. (2009). Geometric feature based facial expression recognition using multiclass support vector machines. In *Granular Computing, 2009, GRC'09. IEEE International Conference on*, pages 318–321. IEEE.

[117] Lekdioui, K., Messoussi, R., and Chaabi, Y. (2015). Etude et modélisation des comportements sociaux d'apprenants à distance, à travers l'analyse des traits du visage. In *7ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2015)*, pages 411–413.

[118] Lekdioui, K., Messoussi, R., Ruichek, Y., Chaabi, Y., and Touahni, R. (2017a). Facial decomposition for expression recognition using texture/shape descriptors and svm classifier. *Signal Processing : Image Communication*.

[119] Lekdioui, K., Ruichek, Y., Messoussi, R., Chaabi, Y., and Touahni, R. (2017b). Facial expression recognition using face-regions. In *Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on*, pages 1–6. IEEE.

[120] Levi, G. and Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510. ACM.

[121] Li, S. Z. and Zhang, Z. (2004). Floatboost learning and statistical face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 26(9) :1112–1123.

[122] Li, Y., Wang, S., Zhao, Y., and Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, 22(7) :2559–2573.

[123] Liao, W.-H. (2010). Region description using extended local ternary patterns. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1003–1006. IEEE.

[124] Lien, J. J., Kanade, T., Cohn, J. F., and Li, C.-C. (1998). Automated facial expression recognition based on facs action units. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 390–395. IEEE.

[125] Liu, B., Wang, M., Foroosh, H., Tappen, M., and Pensky, M. (2015a). Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814.

[126] Liu, M., Li, S., Shan, S., and Chen, X. (2015b). Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159 :126–136.

[127] Liu, M., Shan, S., Wang, R., and Chen, X. (2014a). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756.

[128] Liu, M., Shan, S., Wang, R., and Chen, X. (2016). Learning expressionlets via universal manifold model for dynamic facial expression recognition. *IEEE Transactions on Image Processing*, 25(12) :5920–5932.

[129] Liu, P., Han, S., Meng, Z., and Tong, Y. (2014b). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812.

[130] Long, F., Wu, T., Movellan, J. R., Bartlett, M. S., and Littlewort, G. (2012). Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing*, 93 :126–132.

[131] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110.

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

[133] Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I. (2011). Painful data : The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE.

[134] Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, pages 91–630.

[135] Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE.

[136] Martinez, B., Valstar, M. F., Jiang, B., and Pantic, M. (2017). Automatic analysis of facial actions : A survey. *IEEE Transactions on Affective Computing*.

[137] Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International journal of computer vision*, 60(2) :135–164.

[138] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa : A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2) :151–160.

[139] McDuff, D., El Kaliouby, R., Demirdjian, D., and Picard, R. (2013). Predicting online media effectiveness based on smile responses gathered over the internet. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE.

[140] Mehrabian, A. and Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3) :248.

[141] Mignon, A. (2012). *Apprentissage de métriques et méthodes à noyaux appliqués à la reconnaissance de personnes dans les images*. PhD thesis, université de caen.

[142] Mohamed, A. A. and Yampolskiy, R. V. (2012). Adaptive extended local ternary pattern (aeltp) for recognizing avatar faces. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 57–62. IEEE.

[143] Moore, S. and Bowden, R. (2011). Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4) :541–558.

[144] Mourão, A. and Magalhães, J. (2013). Competitive affective gaming : winning with a smile. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 83–92. ACM.

[145] Mullins, D. T. and Duke, M. P. (2004). Effects of social anxiety on nonverbal accuracy and response time i : Facial expressions. *Journal of nonverbal behavior*, 28(1) :3–33.

[146] Nanni, L., Brahnam, S., and Lumini, A. (2011). Local ternary patterns from three orthogonal planes for human action classification. *Expert Systems with Applications*, 38(5) :5125–5128.

[147] Nkambou, R. and Heritier, V. (2004). Reconnaissance émotionnelle par l'analyse des expressions faciales dans un tuteur intelligent affectif. In *Technologies de l'Information et de la Connaissance dans l'Enseignement Supérieur et l'Industrie*, pages 149–155. Université de Technologie de Compiègne.

[148] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59.

[149] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7) :971–987.

[150] Oliver, N., Pentland, A., and Bérard, F. (2000). Lafter : a real-time face and lips tracker with facial expression recognition. *Pattern recognition*, 33(8) :1369–1382.

[151] Orozco, J., Martinez, B., and Pantic, M. (2015). Empirical analysis of cascade deformable models for multi-view face detection. *Image and vision computing*, 42 :47–61.

[152] Ossia, S. A., Shamsabadi, A. S., Taheri, A., Rabiee, H. R., Lane, N., and Haddadi, H. (2017). A hybrid deep learning architecture for privacy-preserving mobile analytics. *arXiv preprint arXiv :1703.02952*.

[153] Otsuka, T. and Ohya, J. (1997). Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences. In *Image Processing, 1997. Proceedings., International Conference on*, volume 2, pages 546–549. IEEE.

[154] Panning, A., Al-Hamadi, A. K., Niese, R., and Michaelis, B. (2008). Facial expression recognition based on haar-like feature detection. *Pattern Recognition and Image Analysis*, 18(3) :447–452.

[155] Pantic, M. (2005). Face for interface. *The Encyclopedia of Multimedia Technology and Networking*, 1 :308–314.

[156] Pantic, M. and Bartlett, M. S. (2007). Machine analysis of facial expressions. In *Face recognition*. InTech.

[157] Pantic, M. and Caridakis, G. (2011). Image and video processing for affective applications. In *Emotion-Oriented Systems*, pages 101–114. Springer.

[158] Pantic, M. and Patras, I. (2006). Dynamics of facial expression : recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2) :433–449.

[159] Pantic, M. and Rothkrantz, L. J. (2004). Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3) :1449–1461.

[160] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, page 5. IEEE.

[161] Poghosyan, G. A. and Sarukhanyan, H. G. (2010). Decreasing volume of face images database and efficient face detection algorithm. *Information Theories and Applications*, 17 :36.

[162] Poursaberi, A., Noubari, H. A., Gavrilova, M., and Yanushkevich, S. N. (2012). Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *EURASIP Journal on Image and Video Processing*, 2012(1) :1–13.

[163] Rabiner, L. R. (1990). A tutorial on hidden markov models and selected applications in speech recognition. In *Readings in speech recognition*, pages 267–296. Elsevier.

[164] Rapp, V., Sénéchal, T., Prevost, L., Bailly, K., Salam, H., and Seguier, R. (2012). Combinaison de descripteurs hétérogenes pour la reconnaissance de micro-mouvements faciaux. In *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, pages 978–2.

[165] Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692.

[166] Riek, L. D. and Robinson, P. (2011). Using robots to help people habituate to visible disabilities. In *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, pages 1–8. IEEE.

[167] Rifai, S., Bengio, Y., Courville, A., Vincent, P., and Mirza, M. (2012). Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer.

[168] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1) :23–38.

[169] Rudovic, O., Pavlovic, V., and Pantic, M. (2012). Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *European Conference on Computer Vision*, pages 260–269. Springer.

[170] Rymarczyk, K., Żurawski, Ł., Jankowiak-Siuda, K., and Szatkowska, I. (2016). Do dynamic compared to static facial expressions of happiness and anger reveal enhanced facial mimicry ? *PloS one*, 11(7) :e0158534.

[171] Ryu, B., Rivera, A. R., Kim, J., and Chae, O. (2017). Local directional ternary pattern for facial expression recognition. *IEEE Transactions on Image Processing*.

[172] Santra, B. and Mukherjee, D. P. (2016). Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 624–628. IEEE.

[173] Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect : A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6) :1113–1133.

[174] Satiyan, M. and Nagarajan, R. (2010). Recognition of facial expression using haar-like feature extraction method. In *Intelligent and Advanced Systems (ICIAS), 2010 International Conference on*, pages 1–4. IEEE.

[175] Scherer, K. R. (1985). *Handbook of methods in nonverbal behavior research*. Cambridge University Press.

[176] Scherer, K. R., Bänziger, T., and Roesch, E. (2010). *A Blueprint for Affective Computing : A sourcebook and manual*. Oxford University Press.

[177] Shah, J. H., Sharif, M., Yasmin, M., and Fernandes, S. L. (2017). Facial expressions classification and false label reduction using lda and threefold svm. *Pattern Recognition Letters*.

[178] Shan, C. (2008). *Inferring facial and body language*. PhD thesis, Queen Mary University of London.

[179] Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on local binary patterns : A comprehensive study. *Image and Vision Computing*, 27(6) :803–816.

[180] Shang, L. and Chan, K.-P. (2009). Nonparametric discriminant hmm and application to facial expression recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2090–2096. IEEE.

[181] Shbib, R. and Zhou, S. (2015). Facial expression analysis using active shape model. *Int. J. Signal Process. Image Process. Pattern Recognit*, 8(1) :9–22.

[182] Shen, X., Lin, Z., Brandt, J., and Wu, Y. (2013). Detecting and aligning faces by image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3467.

[183] Sikka, K., Dhall, A., and Bartlett, M. (2015). Exemplar hidden markov models for classification of facial expressions in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–25.

[184] Sikka, K. and Sharma, G. (2018). Discriminatively trained latent ordinal model for video classification. *IEEE transactions on pattern analysis and machine intelligence*, 40(8) :1829–1844.

[185] Simon, T., Nguyen, M. H., De La Torre, F., and Cohn, J. F. (2010). Action unit detection with segment-based svms. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2737–2744. IEEE.

[186] Smith, B. M., Zhang, L., Brandt, J., Lin, Z., and Yang, J. (2013). Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491.

[187] Sohail, A. S. M. and Bhattacharya, P. (2007). Classification of facial expressions using k-nearest neighbor classifier. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 555–566. Springer.

[188] Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4) :427–437.

[189] SOWN, M. (1978). A preliminary note on pattern recognition of facial emotional expression. In *The 4th International Joint Conferences on Pattern Recognition, 1978*.

[190] Soyel, H. and Demirel, H. (2010). Facial expression recognition based on discriminative scale invariant feature transform. *Electronics letters*, 46(5) :343–345.

[191] Stathopoulou, I.-O. and Tsihrintzis, G. A. (2004). An improved neural-network-based face detection and facial expression classification system. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 666–671. IEEE.

[192] Sultana, M., Bhatti, M. N. A., Javed, S., and Jung, S.-K. (2017). Local binary pattern variants-based adaptive texture features analysis for posed and nonposed facial expression recognition. *Journal of Electronic Imaging*, 26(5) :053017.

[193] Sun, B., Cao, S., He, J., Yu, L., and Li, L. (2017). Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks. *Journal of Electronic Imaging*, 26(2) :020501.

[194] Sun, B., Li, L., Zhou, G., and He, J. (2016). Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6) :061407–061407.

[195] Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE.

[196] Sun, Y. and Yu, J. (2017). Facial expression recognition by fusing gabor and local binary pattern features. In *International Conference on Multimedia Modeling*, pages 209–220. Springer.

[197] Susskind, J., Mnih, V., Hinton, G., et al. (2011). On deep generative models with applications to recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2857–2864. IEEE.

[198] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12.

[199] Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6) :1635–1650.

[200] Tariq, U., Yang, J., and Huang, T. S. (2012). Multi-view facial expression recognition analysis with generic sparse coding feature. In *European Conference on Computer Vision*, pages 578–588. Springer.

[201] Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2) :97–115.

[202] Tian, Y.-l. (2004). Evaluation of face resolution for expression analysis. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 82–82. IEEE.

[203] Tian, Y.-l., Kanade, T., and Cohn, J. F. (2002). Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 229–234. IEEE.

[204] Tian, Y.-L., Kanade, T., and Cohn, J. F. (2005). Facial expression analysis. *Handbook of face recognition*, pages 247–275.

[205] Tripathi, A., Pandey, S., and Jangir, H. (2018). Efficient facial expression recognition system based on geometric features using neural network. In *Information and Communication Technology for Sustainable Development*, pages 181–190. Springer.

[206] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1) :71–86.

[207] Uddin, M. Z., Hassan, M. M., Almogren, A., Alamri, A., Alrubaian, M., and Fortino, G. (2017). Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, 5 :4525–4536.

[208] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016 : Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

[209] Valstar, M. and Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE.

[210] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M., and Cohn, J. F. (2015). Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE.

[211] Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., and Scherer, K. (2011). The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE.

[212] Valstar, M. F. and Pantic, M. (2012). Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1) :28–43.

[213] Valstar, M. F., Pantic, M., Ambadar, Z., and Cohn, J. F. (2006). Spontaneous vs. posed facial behavior : automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM.

[214] van der Maaten, L. and Hendriks, E. (2012). Action unit classification using active appearance models and conditional random fields. *Cognitive processing*, 13(2) :507–518.

[215] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5) :988–999.

[216] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[217] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2) :137–154.

[218] Vukadinovic, D. and Pantic, M. (2005). Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1692–1698. IEEE.

[219] Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2016). Variable-state latent conditional random field models for facial expression analysis. *Image and Vision Computing*, 2(4).

[220] Wallhoff, F. (2006). Facial expressions and emotion database. *Technische Universität München*.

[221] Wang, L., Li, R., and Wang, K. (2014). A novel automatic facial expression recognition method based on aam. *Journal of Computers*, 9(3) :608–617.

[222] Wang, N., Gao, X., Tao, D., Yang, H., and Li, X. (2018). Facial feature point detection : A comprehensive survey. *Neurocomputing*, 275 :50–65.

[223] Wang, S. B., Quattoni, A., Morency, L.-P., Demirdjian, D., and Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1521–1527. IEEE.

[224] Wang, X.-H., Liu, A., and Zhang, S.-Q. (2015a). New facial expression recognition based on fsvm and knn. *Optik-International Journal for Light and Electron Optics*, 126(21) :3132–3134.

[225] Wang, Y., Yu, H., Stevens, B., and Liu, H. (2015b). Dynamic facial expression recognition using local patch and lbp-top. In *Human System Interactions (HSI), 2015 8th International Conference on*, pages 362–367. IEEE.

[226] Williams, A. C. d. C. (2002). Facial expression of pain : an evolutionary account. *Behavioral and brain sciences*, 25(4) :439–455.

[227] Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug) :975–1005.

[228] Wu, X., Sun, J., Fan, G., and Wang, Z. (2015). Improved local ternary patterns for automatic target recognition in infrared imagery. *Sensors*, 15(3) :6399–6418.

[229] Wu, Y. and Ji, Q. (2015). Robust facial landmark detection under significant head poses and occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3658–3666.

[230] Wu, Y. and Ji, Q. (2016). Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3400–3408.

[231] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539.

[232] Xiong, X. and De la Torre, F. (2015). Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673.

[233] Xu, X., Quan, C., and Ren, F. (2015). Facial expression recognition based on gabor wavelet transform and histogram of oriented gradients. In *Mechatronics and Automation (ICMA), 2015 IEEE International Conference on*, pages 2117–2122. IEEE.

[234] Yacoob, Y. and Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6) :636–642.

[235] Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J., and Fu, X. (2013). Casme database : a dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on*, pages 1–7. IEEE.

[236] Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images : A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 24(1) :34–58.

[237] Yang, M.-T., Cheng, Y.-J., and Shih, Y.-C. (2011). Facial expression recognition for learning status analysis. In *International Conference on Human-Computer Interaction*, pages 131–138. Springer.

[238] Yang, P., Liu, Q., and Metaxas, D. N. (2009a). Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2) :132–139.

[239] Yang, P., Liu, Q., and Metaxas, D. N. (2009b). Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1018–1025. Ieee.

[240] Yang, P., Liu, Q., and Metaxas, D. N. (2010). Exploring facial expressions with compositional features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2638–2644. IEEE.

[241] Yeasin, M., Bullot, B., and Sharma, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3) :500–508.

[242] Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE.

[243] Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

[244] Youssif, A. A. and Asker, W. A. (2011). Automatic facial expression recognition system based on geometric and appearance features. *Computer and Information Science*, 4(2) :115.

[245] Yu, Z. and Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM.

[246] Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild : past, present and future. *Computer Vision and Image Understanding*, 138 :1–24.

[247] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods : Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1) :39–58.

[248] Zhang, C. and Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1036–1041. IEEE.

[249] Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., and Yan, K. (2016). A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia*, 18(12) :2528–2536.

[250] Zhang, X., Mahoor, M. H., and Voyles, R. M. (2013). Facial expression recognition using hessianmkl based multiclass-svm. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE.

[251] Zhang, Y. and Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5) :699–714.

[252] Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 454–459. IEEE.

[253] Zhao, G., Huang, X., Taini, M., Li, S. Z., and PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9) :607–619.

[254] Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6) :915–928.

[255] Zhao, L., Wang, Z., and Zhang, G. (2017). Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram. *Mathematical Problems in Engineering*, 2017.

[256] Zhao, X., Shi, X., and Zhang, S. (2015). Facial expression recognition via deep learning. *IETE Technical Review*, 32(5) :347–355.

[257] Zheng, W. (2014). Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Transactions on Affective Computing*, 5(1) :71–85.

[258] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE.