Analyse des expressions faciales dans un flux vidéo

Benjamin Allaert

Université de Lille

FACULTÉ DES SCIENCES ET TECHNOLOGIES
Département Informatique

THÈSE

En vue de l'obtention du

DOCTORAT INFORMATIQUE

Discipline Informatique et applications

Délivré par : L'Université de Lille

Présentée et soutenue le 8 juin 2018 par :

BENJAMIN ALLAERT

ANALYSE DES EXPRESSIONS FACIALES DANS UN FLUX VIDÉO

JURY

Pr Laurence DUCHIEN	Université de Lille, France	Président du Jury
Pr Moncef GABBOUJ	Université de Tampere, Finlande	Examinateur
Pr Jenny BENOIS-PINEAU	Université de Bordeaux, France	Rapporteur
Pr Monique NOIRHOMME	Université de Namur, Belgique	Rapporteur
Pr Chaabane DJERABA	Université de Lille, France	Directeur de Thèse
Dr Ioan Marius BILASCO	Université de Lille, France	Co-Encadrant

---

Ecole doctorale et spécialité :

Ecole doctorale Sciences Pour l'Ingénieur

Unité de Recherche :

Centre de Recherche en Informatique, Signal et Automatique de Lille

UMR 9189 - CRIStAL - F-59000 Lille, France

Remerciements

Tout d'abord, je tiens à remercier le professeur Chaabane Djeraba de m'avoir accueilli au sein de son équipe afin de réaliser ma thèse. Je le remercie pour m'avoir conseillé dans la rédaction de mes articles scientifiques et dans la rédaction de la thèse.

Je suis très reconnaissant envers José Mennesson pour son soutien pendant la thèse, notamment pour les nombreuses discussions que nous avons partagées, pour ses idées, ses relectures et sa disponibilité. Mes remerciements vont également à Adel Lablack, pour sa collaboration, et son aide.

Je tiens particulièrement à remercier mon encadrant de thèse Marius Bilasco, qui a accompli un remarquable travail d'encadrement. Ses précieux conseils et nos nombreux échanges ont été déterminants dans la réalisation de la thèse.

Je remercie vivement Madame Laurence Duchien, professeur à l'Université de Lille, d'être le Président de mon jury de thèse. Je tiens à remercier également les professeurs Monique Noirhomme-Fraiture et Jenny Benois-Pineau pour avoir accepté d'être les rapporteurs de mon mémoire de thèse, et le professeur Moncef Gabbouj pour avoir accepté d'être examinateur.

Je remercie le groupe thématique Image, mais aussi l'ensemble des membres du FIL et les étudiants que j'ai encadré, qui ont contribué de près ou de loin à ma thèse.

Je tiens enfin à saluer l'ensemble de l'équipe FOX dans laquelle j'ai réalisé ma thèse. Plus particulièrement, Jean Martinet, Pierre Tirilly et Marius Bilasco pour leurs conseils et la qualité de leur encadrement au long de ces trois dernières années ainsi que Rémi Auguste, Romain Belmonte et Delphine Poux.

Resume

De nos jours, dans des domaines tels que la sécurité, la santé et la communication, une forte demande consiste à pouvoir analyser le comportement des personnes en s'appuyant notamment sur l'analyse faciale. Dans cette thèse, nous explorons de nouvelles approches permettant de répondre à différents problèmes liés à l'analyse des expressions faciales dans une séquence d'images, à destination d'un système d'acquisition peu contraint (variations de pose, changements d'illumination, occultations, expressions spontanées). Plus spécifiquement, nous nous intéressons à la variabilité des intensités des expressions faciales (micro et macro expressions) et à l'analyse des expressions faciales en présence d'occultations induites par des variations de pose du visage.

Notre première contribution s'intéresse à la caractérisation précise des variations d'intensité des expressions faciales. Nous proposons un descripteur innovant appelé LMP (Local Motion Patterns) qui s'appuie sur les propriétés physiques déformables du visage afin de conserver uniquement les directions principales du mouvement facial induit par les expressions faciales. La particularité principale de notre travail est liée au fait que notre descripteur permet d'obtenir de très bonnes performances pour caractériser à la fois les micro et les macro expressions en utilisant, pour les deux catégories d'intensité, le même système d'analyse (descripteur, modèle facial, configuration).

Notre deuxième contribution concerne la prise en compte des variations de pose. Généralement, une étape de normalisation du visage est employée afin d'obtenir une invariance aux transformations géométriques. Cependant, ces méthodes sont utilisées sans forcément connaître leur impact sur les expressions faciales. Nous proposons une nouvelle base d'apprentissage innovant appelé SNaP-2DFe (Simultaneous Natural and Posed 2D Facial expression) pour permettre de mieux caractériser la robustesse des méthodes de normalisation pour l'analyse des expressions faciales. Le système d'acquisition proposé permet de capturer simultanément un visage dans un plan fixe et dans un plan mobile (i.e. suit le déplacement de la tête). Grâce à cela, nous fournissons une connaissance du visage à reconstruire malgré les occultations induites par les rotations 3D (hors plan) de la tête. Nous montrons à travers les données récoltées, que les méthodes de normalisation employées dans les systèmes actuels ne sont pas parfaitement adaptées pour l'analyse des expressions faciales.

Abstract

Facial expression recognition has attracted great interest over the past decade in wide application areas, such as human behavior analysis, video communication, e-health and marketing. In this thesis we explore a new approach to step forward towards in-the-wild expression recognition, dealing with challenges such as various intensity and various expression activation patterns, illumination variation and head pose variations. Special attention has been paid to encode respectively small/large facial expression amplitudes (micro and macro expressions), and to analyze facial expressions in presence of varying head pose.

The first challenge addressed concerns varying facial expression amplitudes. We propose an innovative motion descriptor called local motion patterns (LMP). This descriptor takes into account mechanical facial skin deformation properties (local coherency and local propagation). When extracting motion information from the face, the unified approach deals with inconsistencies and noise, caused by face characteristics (skin smoothness, skin reflect and elasticity). The main originality of our approach is a unified approach for both small and large facial expression recognition, with the same facial recognition framework (descriptor, facial framework, parameters).

The second challenge addressed concerns important head pose variations. In facial expression analysis, the face registration step must ensure that minimal deformation appears. Registration techniques must be used with care in presence of unconstrained head pose as facial texture transformations apply. Hence, it is valuable to estimate the impact of alignment-related induced noise on the global recognition performance. For this, we propose a new database, called Simultaneous Natural and Posed 2D facial expression database (SNaP-2DFe), allowing to study the impact of head pose and intra-facial occlusions on expression recognition approaches. We prove that the usage of face registration approach does not seem adequate for preserving the features encoding facial expression deformations.

Table des matières

Chapitre 1

Introduction

« *Les visages trompent rarement :*
on a l'âme de son visage
et le visage de son âme. »

Paul Brulat

Sommaire

1.1 Contexte

L'essor des systèmes d'acquisition et de traitement de la vidéo joue un rôle important dans la vie quotidienne. Les usages de la vidéo évoluent proportionnellement avec nos besoins d'automatiser le processus d'extraction de l'information depuis la vidéo (analyse du comportement, reconnaissance, suivi, etc). L'objectif général de cette thèse est de concevoir des modèles et des systèmes capables de représenter et d'interpréter le contenu visuel d'une scène. Plus particulièrement, nous nous intéressons à l'analyse des états émotionnels communiqués par le biais des expressions faciales.

Actuellement, il y a un vrai engouement autour des technologies permettant d'analyser le comportement humain à partir de flux vidéo. L'analyse du comportement humain, par le biais d'une séquence d'images, se caractérise principalement par l'extraction de caractéristiques visuelles. La communication s'accompagne de mouvements apparaissant dans des zones corporelles aussi variées que la tête, le visage, les mains, les bras, le torse et les jambes. Chaque communicant émet des comportements corporels (gestes, regards, mimiques faciales, postures) et répond à ceux de l'autre. Ces comportements jouent un rôle important dans l'échange d'informations et la régulation de l'interaction sociale. L'orientation du regard, la posture corporelle et les expressions faciales renseignent les partenaires de l'interaction sur leurs intérêts, leurs intentions, leurs attitudes et leurs états émotionnels respectifs. L'analyse du comportement humain joue un rôle important dans une variété d'applications telles que :

- **La sécurité** : l'analyse du comportement humain est fortement utilisée pour la sécurisation des sites sensibles et de la circulation routière. Les systèmes d'analyse vidéo en temps réel permettent de reconnaître des personnes (âge, genre), de surveiller des flux routiers ou de dépister des comportements suspects, comme illustré dans la Figure 1.1-A. Bien que la vidéosurveillance intelligente ne fournisse pas encore de systèmes totalement autonomes, elle permet d'assister les agents de l'ordre. Actuellement, de nombreuses recherches tentent d'améliorer la robustesse des modèles et d'élargir les conditions de fonctionnement optimales, malgré les changements de luminosité, les angles réduits ou les occultations. L'un des défis majeur dans ce domaine, est la mise en place d'un système capable de reconstituer le déroulement d'un événement comme

un attentat, en analysant les diverses sources vidéo disponibles, issues des systèmes de surveillance. La difficulté de tels systèmes réside dans le fait de pouvoir analyser et identifier une personne malgré les problèmes d'occultations liés à la scène (angle mort, objets occultants, foule d'individus) et aux variations de pose du visage par rapport à la position des caméras.

- **La santé** : l'analyse du comportement humain dans la relation soignant-soigné, permet au soignant (médecin, thérapeute, psychologue) de développer une relation de meilleure qualité avec le soigné, parce qu'il disposera des outils pour mieux le comprendre et le suivre de manière continue. Au-delà d'une attitude de façade, certains indices corporels, lisibles aussi bien dans l'attitude générale que dans des détails de micro-communication, permettent d'évaluer avec davantage de précision le ressenti du patient. D'autre part, le corps d'une personne souffrant de douleurs, donne au soignant des signes (crispation musculaire, réaction involontaire), décelables en analysant les flux vidéo. Ces signes lui permettent d'appréhender la relation du patient à sa douleur avec davantage d'objectivité. Certaines crispations sont lisibles à la surface du corps (plissures aux coins des lèvres, froncement des sourcils), et sont autant de témoignages pour le soignant, des difficultés de son patient, comme illustré dans la Figure 1.1-B. Les défis à relever par ces systèmes résident essentiellement dans le fait de percevoir ce que le patient cherche parfois à dissimuler. Dans ce cas, il est nécessaire de proposer des systèmes permettant de détecter de subtiles informations, comme les micro-mouvements, souvent involontaires, mais qui sont porteuses d'informations pour le médecin.

- **La communication** : les nouvelles technologies de communication (NTIC) et notamment le Web 2.0 permettent de mettre en relation les personnes entre elles et de les faire échanger ou collaborer (i.e. Facebook, Twitter). Cependant, les NTIC tendent à développer un lien virtuel qui se fait au détriment du lien réel, ce qui appauvrit la communication et aboutit à une certaine forme de rupture "sociale". La communication n'est pas uniquement verbale. La communication non-verbale, représentée par les gestes, les mimiques, les comportements, la posture, transmettent des éléments de communication qui humanise l'échange. Lors d'échanges par caméras interposées, notamment lorsqu'il y a plus de deux interlocuteurs (visioconférence, e-learning), les infor-

mations non-verbales sont plus difficilement interprétables et appauvrissent l'échange social. Les systèmes d'analyse vidéo permettent de renforcer cet échange en apportant des renseignements complémentaires sur le comportement des interlocuteurs, comme illustré dans la Figure 1.1-C. Grâce à cela, il est possible d'obtenir un retour sur l'état affectif d'une personne ou sur l'attention qu'il porte à la discussion. La difficulté de l'analyse dans un contexte de communication est principalement liée aux occultations du visage, notamment par les mouvements de la main et aux variations de pose du visage, qui cachent partiellement des parties du visage.

FIGURE 1.1 – Exemple d'applications où l'analyse du comportement humain est utilisée : sécurité - reconnaissance des visages et reconnaissance de genre (A), santé - analyse de la douleur (B), communication - analyse de l'état affectif (C).

L'analyse des états émotionnels est actuellement un domaine de recherche en plein développement. L'émergence de travaux sur la sécurité, l'interaction et sur la communication, inscrit l'étude des états émotionnels dans une perspective pluridisciplinaire. De nombreuses disciplines comme la psychologie, l'anthropologie, les sciences cognitives, la médecine ou encore l'éducation sont concernées. Cela nécessite de disposer de méthodologies adéquates, transposables et transdisciplinaires pour l'analyse et l'interprétation des comportements humains. Des travaux de recherche en psychologie ont démontré que les expressions faciales jouent un rôle prépondérant dans la conversation humaine [12], et transmettent plus d'informations que le message exprimé. Mehrabian [81] remarque qu'au cours d'une discussion "face à face", l'impact du message transmis représente à lui seulement 7%, alors que les signaux conversationnels et l'expression faciale du locuteur contribuent respectivement à 38 % et 55 % de l'impact global du message exprimé. Par conséquent, l'expression faciale est une source riche en information et est

définie comme une modalité essentielle de la communication humaine. L'ensemble de ces constats montre donc l'importance de l'analyse des expressions faciales pour caractériser l'état émotionnel chez l'homme.

Dans cette thèse, nous explorons de nouvelles approches permettant de répondre à différents problèmes liés à l'analyse des expressions faciales dans une séquence d'images, à destination d'un système d'acquisition où l'interaction est naturelle (l'intensité des mouvements faciaux peut varier) et où le visage est frontal à la caméra. Dans la section suivante, nous mettons en avant les problématiques que nous abordons dans ces travaux.

1.2 Problématique

Les systèmes d'analyse d'expressions faciales proposés dans l'état de l'art obtiennent de très bonnes performances dans des conditions où l'environnement est contrôlé (fond de la scène uniforme, lumière homogène) et où les expressions faciales sont maîtrisées (pose fixe, visage face à la caméra, expression de forte intensité), comme illustré sur la Figure 1.2-A. Cependant, ces données ne reflètent pas les conditions rencontrées dans une situation d'interaction naturelle (caméra de surveillance, visioconférence) où la personne est libre de ses mouvements. Dans ce contexte, l'environnement (changement lumineux, intérieur/extérieur), ainsi que l'interaction (variations de pose, larges déplacements dans la scène, occultation de visage, intensité des expressions variable) sont peu contraints, comme illustré sur la Figure 1.2-B et 1.2-C. Les problèmes liés aux conditions d'acquisition renforcent la difficulté d'analyse des expressions faciales et mettent en évidence plusieurs verrous scientifiques, encore non résolus. Dans la suite, nous illustrons plus spécifiquement les défis soulevés par la variation de l'intensité des expressions faciales et sur l'impact de la présence de variations de pose et de larges déplacements sur l'analyse faciale.

Souvent adapté pour analyser les mouvements faciaux de faibles et fortes intensités, il est difficile de concevoir une solution unique permettant d'analyser un large panel d'intensités, en s'appuyant sur la même méthodologie. Or, en situation d'interaction naturelle, les intensités des mouvements faciaux peuvent varier et nécessitent que l'on puisse traiter à la fois les petites et grandes intensités. La difficulté réside dans le fait que les ca-

ractéristiques de mouvement sont très différentes entre les expressions de faibles et fortes intensités, ce qui demande en général de passer par des techniques de prétraitements adaptés au mouvement analysé. Dans la majorité des cas, les prétraitements appliqués aux visages permettent de mieux caractériser un mouvement spécifique au détriment d'un autre mouvement. Ceci implique qu'il n'existe pas de solution unique prenant en compte à la fois les petites et les grandes variations de mouvement.

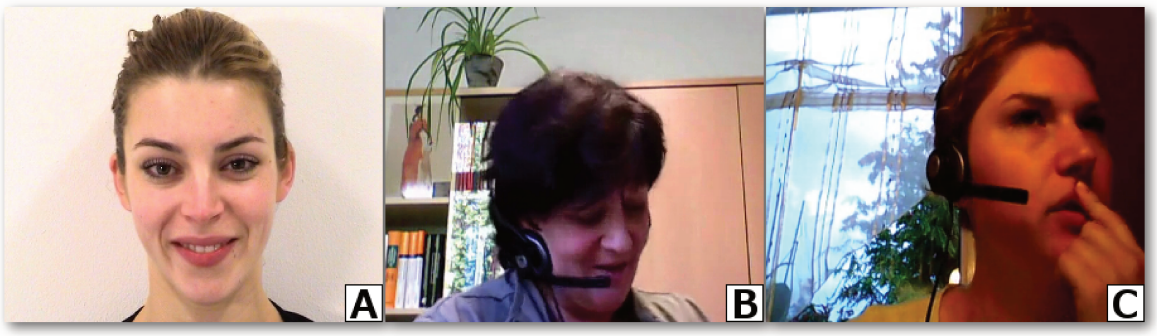


FIGURE 1.2 – Illustration des environnements de captation : (A) contrôlé (ADFES) et (B,C) non contrôlés (AViD).

Deux catégories d'approches se distinguent pour analyser les expressions faciales : les approches basées sur l'étude des caractéristiques du visage et les approches basées sur l'étude du mouvement. L'extraction du mouvement consiste à analyser les déformations apparentes du visage, causées par le mouvement relatif entre la caméra et la scène. Ambadar et al. [4] montrent l'importance de l'analyse de la dynamique faciale pour reconnaître les expressions, car cela permet d'identifier plus subtilement les déformations physiques du visage. Bien que l'analyse du mouvement permette d'améliorer les performances des méthodes proposées dans la littérature, plusieurs difficultés nécessitent d'être résolues.

Malgré le fait que les méthodes s'appuyant sur le mouvement sont plus performantes dans un contexte contrôlé, ce n'est pas le cas en situation d'interaction naturelle. Ici, la présence de variations de pose et de larges déplacements induisent des discontinuités de mouvement qui renforcent la difficulté de l'analyse. En effet, le bruit induit par le mouvement de la tête à un impact négatif sur le mouvement extrait au sein du visage. Des solutions de normalisation sont employées pour transformer le visage afin de simuler un contexte maîtrisé (frontal à la caméra). Bien que ces solutions permettent de réduire

la distance entre deux visages extraits d'images successives, la normalisation induit des déformations néfastes sur la géométrie du visage. Celles-ci résultent de l'apparition de mouvements incohérents sur le visage. Souvent utilisées dans le domaine de la reconnaissance faciale, les méthodes de normalisation ne garantissent pas la conservation des expressions faciales, et sont donc mal adaptées lorsque l'on s'intéresse à l'analyse des expressions faciales en situation d'interaction libre.

1.3 Objectifs

Dans ce mémoire de thèse, nous explorons et proposons des solutions permettant de répondre aux deux problèmes relevés précédemment :

- Comment analyser les faibles et fortes intensités de mouvement, en utilisant une méthode unique ?
- Comment adapter les méthodes s'appuyant sur le mouvement pour analyser les expressions faciales en présence de variations de pose et de larges déplacements de la tête ?

La prise en compte des variations d'intensité, nécessite d'élaborer un nouveau descripteur de mouvement prenant en compte les spécificités propres au visage pour les expressions faciales (élasticité de la peau, texture lisse). Il est important d'identifier les propriétés qui caractérisent un mouvement facial afin de construire un modèle adapté aux propriétés du visage (élasticité de la peau, contraintes musculaires). De ce fait, une étude de la propagation du mouvement facial doit être réalisée au préalable. Pour identifier les caractéristiques d'un mouvement facial, il est primordial de travailler sur des données où les conditions d'acquisition sont contrôlées (lumière homogène, pose fixe). En effet, cela garantit que le mouvement extrait au sein du visage est non bruité (aucun mouvement parasite provenant de la scène), et qu'il provient uniquement des expressions faciales. Les performances du descripteur à extraire des mouvements faciaux d'intensité variable, seront évaluées sur des bases de données composées de macro et de micro expressions, permettant ainsi de prendre en compte une large palette d'intensités.

L'analyse des expressions faciales en présence de variations de pose et de larges déplacements est souvent traitée à l'aide de méthodes de normalisation du visage. Un nombre

important de bases de données permettent de quantifier les performances des méthodes de normalisation du visage en présence de variations de pose. Ces bases sont généralement utilisées pour répondre à des problèmes d'identification de personnes. Ces mêmes solutions sont proposées pour analyser les expressions faciales. Or, en présence d'expressions faciales, il est important de pouvoir quantifier l'impact de la normalisation du visage, sur les performances de reconnaissance des expressions. Cependant, il n'existe pas à l'heure actuelle, de données permettant de répondre à ce besoin. De ce fait, il est judicieux de proposer une base de données permettant à la fois de quantifier la qualité de la normalisation du visage (minimiser les déformations faciales) et de pouvoir quantifier l'impact de la normalisation sur les expressions faciales (en termes de performances de reconnaissance). Il est d'autant plus important d'améliorer la précision de ces méthodes lorsque l'on désire analyser le mouvement facial, où la moindre déformation du visage a un impact considérable sur le mouvement extrait. Disposer d'un cadre pour une étude comparative des différentes méthodes de normalisation est essentiel pour identifier la meilleure solution adaptée à l'analyse des expressions faciales.

1.4 Plan

Après cette introduction qui fait ressortir notre approche globale et nos objectifs, ce mémoire de thèse est organisé comme suit.

Du fait que l'analyse des expressions faciales découle de l'analyse faciale, nous commençons ce mémoire en offrant un aperçu global du processus générique d'analyse faciale. Le Chapitre 2 permet d'introduire le processus d'analyse faciale, de manière générale, en passant en revue un ensemble des travaux représentatifs proposés dans la littérature. Nous présentons les différentes étapes du processus. Plus spécifiquement, nous nous focalisons sur les méthodes de détection du visage, sur l'extraction de caractéristiques visuelles, ainsi que sur les challenges de l'analyse faciale induits par l'accquisition dans des conditions non contrôlées.

Après avoir exposé les fondements de l'analyse faciale, le Chapitre 3 traite de manière exhaustive l'adaptation de ces systèmes pour la caractérisation des expressions faciales. Tout d'abord, nous présentons les modalités pour caractériser l'état affectif (représenté

sous forme de catégories ou de dimensions) à partir de l'expression faciale. Puis nous détaillons plus particulièrement les difficultés rencontrées dans ce domaine de recherche, notamment en présence de variations de pose, de larges déplacements, d'occultations et de variations d'intensité des mouvements faciaux. Nous discutons de l'évolution des systèmes, et de la manière dont sont construits les bases d'apprentissage, pour répondre aux différents défis préalablement cités.

Afin de répondre à un besoin de caractériser finement le mouvement facial en présence d'expressions faciales d'intensité variable, nous proposons dans le Chapitre 4 un descripteur de mouvement innovant appelé LMP (Local Motion Patterns). En s'appuyant sur les propriétés du visage (contrainte d'élasticité de la peau et contrainte musculaire), ce descripteur permet de filtrer tout mouvement qui n'est pas directement lié aux expressions faciales. Tout d'abord, nous étudions les caractéristiques d'un mouvement sur un objet non-rigide, en prenant comme référence le mouvement facial. Cette analyse permet de poser des hypothèses pour dissocier un mouvement cohérent d'un mouvement résiduel. Dans un deuxième temps, nous opérationnalisons l'utilisation de ces contraintes locales de magnitude et d'orientation, afin de conserver uniquement le mouvement cohérent sur le visage.

Dans le Chapitre 5, nous analysons les performances de notre descripteur sur des bases d'apprentissage composées d'expressions faciales de faibles et fortes intensités. Nous sélectionnons des bases où les données sont acquises dans différents contextes (lumière naturelle et infrarouge, présence de petites variations de pose) afin de montrer la robustesse de notre descripteur à s'appliquer sur des données plus naturelles. Nous analysons les régions les plus pertinentes pour analyser les expressions faciales en se basant sur le mouvement, ainsi que le paramétrage optimal de notre descripteur, pour analyser les visages.

Dans le Chapitre 6, nous nous intéressons à l'adaptation des approches de reconnaissance d'expressions faciales en présence de variations de pose et de larges déplacements. Plus particulièrement, nous passons en revue les différentes solutions et les bases d'apprentissage proposées dans la littérature. Puis, nous présentons un nouveau protocole d'acquisition SNaP-2DFe (Simultaneous Natural and Posed 2D Facial expression), per-

mettant de quantifier l'impact de la normalisation sur les expressions faciales ainsi que la quantification des performances de la normalisation sur le visage. Nous terminons par une étude comparative des différentes solutions de normalisation sur les données extraites par le biais de notre protocole.

Nous concluons ce mémoire de thèse avec le Chapitre 7, dans lequel nous passons en revue l'ensemble des contributions effectuées, et nous évoquons les perspectives à moyen et long terme ouvertes par ce travail.

Chapitre 2

La caractérisation d'un visage

« *Le visage d'une mère*
est pour l'enfant
son premier livre d'images. »

Christian Bobin

Sommaire

2.1 Introduction

L'analyse faciale est un domaine de recherche en plein essor puisqu'elle concerne de nombreux domaines d'application tels que par exemple la sécurité (biométrie, surveillance), la robotique (interaction homme-machine) ou les télécommunications (l'analyse affective). Ces systèmes ont le potentiel d'être déployés facilement et sont considérés comme peu invasifs, en comparaison à d'autres systèmes biométriques (empreintes digitales, reconnaissance de l'iris...), et leur usage se démocratise avec l'évolution des outils de communication (Internet, mobile).

Les systèmes permettant de caractériser les visages suivent généralement des processus très similaires. La Figure 2.1 représente les différentes étapes d'un processus générique d'analyse faciale. La première étape consiste à détecter le visage dans une image. Une fois le visage détecté, des informations telles que la texture et la géométrie faciale en sont extraites. Une étape de normalisation est parfois nécessaire pour réduire la différence des visages intra- et inter-individus, en présence notamment de variations de pose, de luminosité et d'occultations (ex : lunettes). Une fois l'extraction des informations faciales terminée, les visages sont labellisés par un système de classification en fonction de l'analyse souhaitée.

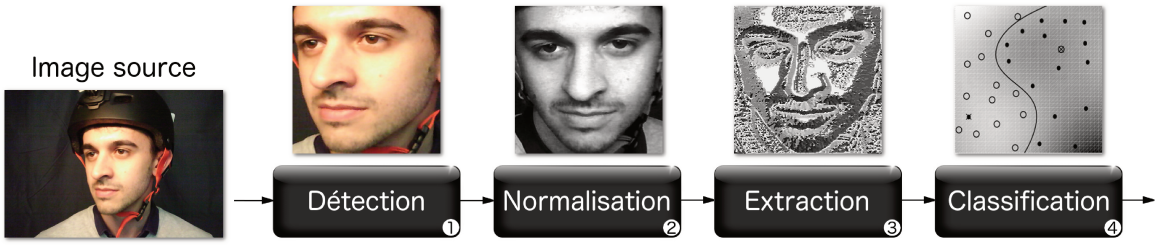


FIGURE 2.1 – Processus de caractérisation de visages (SNaP-2DFe).

Dans la suite de ce chapitre, nous discutons en détail des différentes étapes du processus d'analyse faciale. Pour chacune de ces étapes, nous présentons les méthodes les plus représentatives de la littérature. Dans ce mémoire, nous concentrons notre étude sur les processus de pré-traitement et sur l'extraction de caractéristiques, c'est pourquoi nous ne nous attardons pas sur l'étape de classification. Nous finissons ce chapitre par une syn-

thèse des différents systèmes d'analyse faciale existants et nous abordons dans le chapitre suivant, l'adaptation de ces systèmes pour caractériser les expressions faciales.

2.2 Détection de visages

Avant de procéder à l'analyse d'un visage fixe ou en mouvement, il convient de le détecter ou de le suivre afin d'en extraire des informations pertinentes. La détection de visage dans une image numérique consiste à mettre en évidence des zones de cette image jugées "intéressantes" pour l'analyse, c'est-à-dire présentant des propriétés locales qui caractérisent un visage. De ce fait, la qualité de l'algorithme utilisé pour détecter les zones d'intérêt conditionne souvent la qualité du résultat de la chaîne de traitement entière que l'on souhaite appliquer à une image. Selon l'algorithme utilisé, les zones définissant les visages sont modélisées sous la forme de points, de courbes continues, ou encore de régions connexes, qui constituent le résultat de la détection. Elles peuvent être définies manuellement ou estimées automatiquement à l'aide de fenêtres de détection (masques) ou par détection de points d'intérêts faciaux à l'aide d'algorithmes d'alignement facial. Dans cette section, nous présentons les approches les plus représentatives pour détecter les visages dans une image, organisées selon la modalité utilisée pour la détection : l'utilisation des pseudo-Haar, des points d'intérêts et les modèles d'alignement facial.

2.2.1 Caractéristiques pseudo-Haar

La méthode proposée par Viola et Jones [106], est souvent utilisée pour détecter les visages dans une image. Cette méthode consiste à balayer l'image avec des fenêtres de détection. Cette méthode considère des fenêtres de détection (ou masques) délimitant des zones rectangulaires adjacentes (rectangles dont la moitié des pixels sont blancs et l'autre moitié sont noirs), comme illustré dans la Figure 2.2-A. Les intensités de pixels de ces rectangles sont additionnées, formant des sommes dont la différence constitue une caractéristique. La caractéristique associée correspondant à la somme de pixels délimités par la zone sombre, soustraite à la somme des pixels délimités par la zone claire. Cette caractéristique encode les variations des pixels à une position donnée dans la fenêtre de détection. La présence de contours et les changements de texture sont ainsi traduits numériquement par les valeurs des caractéristiques pseudo-Haar.

L'image est balayée avec différents masques (Figure 2.2-B) afin de caractériser la texture. L'analyse peut se faire à différents niveaux d'échelle (analyse en cascade) avec des masques ayant des tailles de plus en plus petit. Cela dépend principalement du niveau d'analyse souhaité (taille des visages à détecter) mais également de la contrainte du temps de calcul allouée à la détection du visage. L'ensemble des masques permet ainsi de caractériser un visage et les différentes parties qui le composent (les yeux, le nez et la bouche), comme illustré dans la Figure 2.2-C.

FIGURE 2.2 – Processus de détection de visage par caractéristiques pseudo-Haar [106].

L'avantage déterminant des caractéristiques pseudo-Haar est la rapidité de leur calcul car elles peuvent être calculées à l'aide d'images intégrales. C'est une représentation sous la forme d'une image, de même taille que l'image d'origine, qui en chacun de ses points contient la somme des pixels situés au-dessus et à gauche de ce point. Plus formellement, l'image intégrale i' est définie à partir de l'image i par :

$$i'(x,y) = \sum_{x' \leqslant x, y' \leqslant y} i(x',y'). \tag{2.1}$$

Grâce à cette représentation, la somme des valeurs dans une zone rectangulaire peut être calculée en seulement 4 accès à l'image intégrale (6 accès pour deux zones rectangulaires contiguës), et donc en temps constant quelle que soit la taille de la zone. Cette somme peut se calculer par récurrence, comme :

$$\begin{aligned} s(x,y) &= s(x,y-1) + i(x,y) \\ i'(x,y) &= i'(x-1,y) + s(x,y). \end{aligned} \tag{2.2}$$

où $s(x,y)$ est la somme cumulée de la ligne x jusqu'à la colonne y. L'image intégrale peut donc se calculer avec un seul parcours de l'image d'origine. Une fois l'image intégrale calculée, la somme des pixels à l'intérieur de n'importe quel rectangle ABCD peut être évaluée en seulement 4 accès :

$$\sum_{x_A < x' \leqslant x_C, y_A < y' \leqslant y_C} i(x',y') = i'(A) + i'(C) - i'(B) - i'(D). \tag{2.3}$$

où A,B,C et D sont les points délimitant la région rectangulaire, x_A, x_C représentent les positions en abscisse des points A et C, et y_A, y_C représentent les positions en ordonnée des points A et C.

Un des problèmes majeurs de la méthode proposée par Viola et Jones est qu'il n'existe pas de méthode optimale pour choisir les différents paramètres régissant l'algorithme. Le nombre de couches d'analyse, leur ordre ou les taux de détection et de fausses détections (faux positifs) pour chaque couche doivent être choisis par essais et erreurs.

Un autre reproche fait à cette méthode, concerne la perte d'informations au passage d'une couche à l'autre de l'analyse en cascade. La perte est due à l'effet couperet des décisions d'acceptation ou de rejet prises à chaque niveau de l'analyse. Certains chercheurs proposent la solution de garder l'information contenue dans la somme pondérée des classifieurs [14, 104]. D'autres [112, 11], proposent de supprimer le concept d'analyse en cascade, en formant un seul classifieur, ce qui permet de réduire le nombre de détections et par la même occasion, les fausses détections (faux positifs).

L'utilisation de l'algorithme de Viola et Jones s'avère parfois mal adaptée, notamment dû à la grande variabilité d'apparence des visages dans des conditions non contraintes.

Dans ces conditions, la forte variabilité intra-classe des visages ne permet pas de détecter facilement les visages, et nécessite des solutions plus robustes.

2.2.2 Caractéristiques par points d'intérêts

Les algorithmes de détection de points d'intérêt se focalisent en général sur des points particuliers des contours, sélectionnés selon un critère précis. Les coins (corners) sont les points de l'image où le contour change brutalement de direction. Il s'agit de points particulièrement stables, et donc susceptibles d'être suivis efficacement entre deux images proches. Il s'agit d'identifier une caractéristique au sein de l'image qui a une forte probabilité à se répéter dans le temps.

La plupart des techniques de détection de points d'intérêt sont basées sur une analyse locale de l'image au deuxième ordre. La différence entre les différentes techniques réside dans l'opérateur de dérivation utilisé. Nous pouvons par exemple citer les méthodes basées sur l'analyse des DoG [73] (Difference of Gaussians), des LoG [9] (Laplacian of Gaussian) ou des DoH[68] (Difference of Hessians).

En général, les points d'intérêt servent à caractériser le barycentre de régions d'intérêt autour de ces mêmes points. En effet, lorsque les structures recherchées dans une image ne correspondent pas à des points saillants ; par exemple, lorsque l'image a subi un lissage important ou lorsque les contours sont épais et progressifs (sourcils, lèvres), l'information contenue dans le voisinage d'un point est susceptible d'améliorer la qualité de détection.

L'analyse des régions d'intérêt repose principalement sur des méthodes multi-échelles [16] fondées sur l'étude des détecteurs de points d'intérêt cités précédemment (DoG, LoG, etc). Ceci permet d'obtenir des régions soit circulaires soit elliptiques, selon le niveau de raffinement voulu. Ces méthodes sont souvent intégrées à des algorithmes tels que SIFT [73] ou SURF [8], qui incluent un descripteur de région d'intérêt en plus d'un détecteur.

Bien que l'information locale autour d'un point d'intérêt permette de renforcer la détection et le suivi de ce point, les régions dépourvues de textures restent difficiles à caractériser. C'est notamment le cas sur le visage, où la texture de la peau est majoritairement lisse. La Figure 2.3 illustre la correspondance de deux points d'intérêt sur deux visages

similaires. L'un est situé dans une zone à fort gradient (rouge) et l'autre sur une zone à faible gradient (vert). On remarque que le point rouge conserve sa position, par rapport au point vert, qui a tendance à être instable. Ceci est dû au fait que les régions adjacentes autour du point vert, ont des caractéristiques très similaires à la région d'origine.

L'usage des points d'intérêt pour caractériser un visage permette de facilement identifier les régions autour des yeux, des sourcils et de la bouche. Cependant, ces solutions sont difficilement exploitables lorsque l'on désire caractériser des éléments peu texturés comme le nez ou les joues, où les points ont tendance à devenir rapidement instables. Cette imprécision peut induire de fausses détections du visage, ainsi augmenter le taux de faux positif et de réduire significativement les performances des systèmes d'analyse faciale.

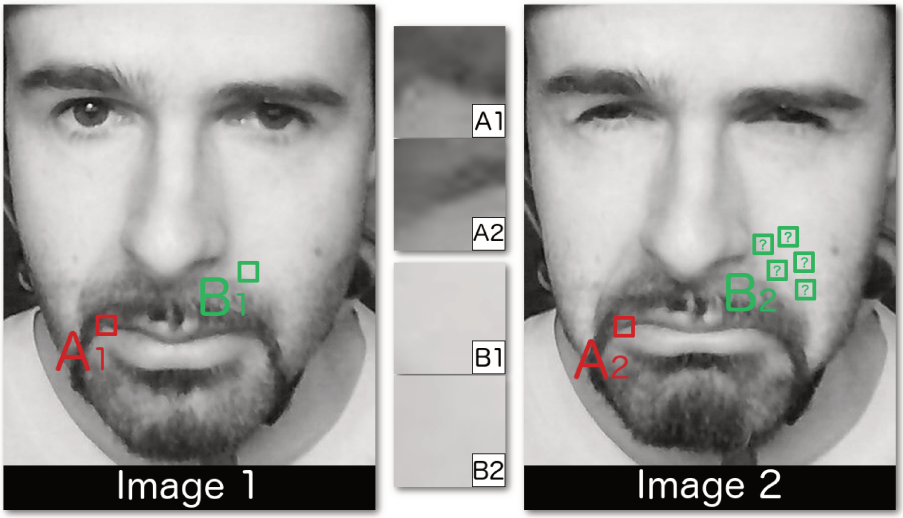


FIGURE 2.3 – Exemple de location de points d'intérêt sur un visage. Problème d'appariement du point B entre l'image 1 et 2, dû à la caractérisation de la texture lisse de la peau (SNaP-2DFe).

2.2.3 Caractéristiques par alignement facial

Afin d'améliorer la précision de la détection des points d'intérêt sur un visage, des nouvelles méthodes ont été proposées. Ces nouvelles méthodes s'appuient sur un modèle de distribution de points (en anglais : Point Distribution Model, PDM) [21], pour garantir une certaine cohérence dans la position des points. Ces solutions réduisent considérablement l'instabilité des points d'intérêts sur un visage, car chaque point dépend de

l'ensemble des points appartenant au modèle. Dans le domaine de l'alignement facial, les points d'intérêts faciaux sont communément appelés "landmarks".

Un modèle s = $(X_1^T, ..., X_N^T)^T$ est construit à partir de collections de distributions de landmarks annotées manuellement, où T représente la forme du visage et $X_i = (x_i, y_i)$ est la localisation 2D du i-ème landmark. De manière générale, les modèles contiennent un nombre prédéfini de landmarks, souvent compris entre 5 et 68. Un exemple de modèle de forme comportant 68 points est illustré dans la Figure 2.4.

Actuellement, on distingue deux approches principales dans les algorithmes d'alignement facial : les approches génératives qui se basent sur des modèles pré-construits de forme et/ou d'apparence pour estimer la forme du visage; et les approches discriminatives qui analysent directement la texture de l'image pour estimer la forme du visage. Dans la suite de cette section, nous détaillons ces deux catégories d'approches.

Caractéristiques par approche générative

En 1995, Cootes et Taylor [22] proposent un modèle statistique déformable (en anglais : Active Shape Model ou ASM) pour détecter et caractériser des visages. L'ASM analyse la texture (intensité des pixels) du visage sur des petites régions correspondant à l'emplacement des landmarks qui le composent. L'algorithme tend à minimiser la distance entre le modèle de forme synthétique et la forme du visage analysé. Pour chaque image X, le modèle estime sa forme globale B en fonction de trois paramètres : la position (T_x, T_y), l'orientation θ et l'échelle s. À chaque itération, l'algorithme examine le voisinage en chaque landmark i du modèle, pour trouver la meilleure correspondance entre la région (x_i, y_i) et la région (x'_i, y'_i). À chaque nouvelle correspondance, le modèle se met à jour $B(T_x, T_y, s, \theta)$, jusqu'à converger vers le modèle d'entrainement le plus similaire au visage analysé.

L'inconvénient des ASM, est le fait qu'ils reposent principalement sur la géométrie (la texture étant seulement analysée autour des points du modèle). L'optimisation du modèle de points se fait uniquement à partir des informations extraites autour des landmarks, comme illustré dans la Figure 2.4. De ce fait, les ASM ne prennent pas en compte l'intégrité des informations contenues dans le visage.

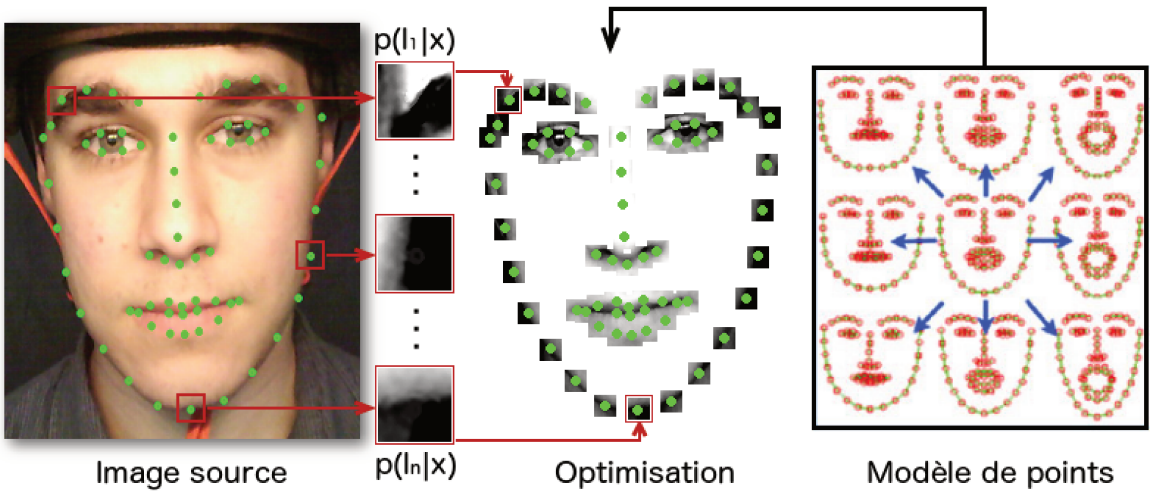


FIGURE 2.4 – Détection des landmarks en appliquant un modèle statistique déformable (ASM).

En 1998, Edwards et al. [20] proposent un nouveau modèle basé sur l'apparence (en anglais : Active Appearance Model, AAM). Un AAM repose entièrement sur un modèle d'apparence, qui caractérise à la fois la texture et la géométrie. L'algorithme analyse l'ensemble de la région, correspondant à la forme convexe autour de la distribution des landmarks. Contrairement aux ASM, un AAM utilise un modèle synthétique, construit à partir d'une base d'apprentissage, pour estimer sa position à chaque itération. Les données d'apprentissage correspondent à des distributions de landmarks labellisés. À chaque itération, l'algorithme tend à minimiser la différence entre son modèle synthétique et le visage analysé. Si la géométrie du modèle synthétique est représentée par un vecteur x et la texture par un vecteur g, alors l'AAM met à jour son modèle synthétique C par rapport à la géométrie et à la texture, en respectant l'équation suivante :

$$\begin{aligned} x &= \overline{x} + \mathrm{Q}_x \mathrm{C} \\ g &= \overline{g} + \mathrm{Q}_g \mathrm{C} \end{aligned} \tag{2.4}$$

où $\overline{x}$ est la forme moyenne, $\overline{g}$ la texture moyenne et Q_x,Q_g sont les matrices correspondant à la variation du modèle par rapport aux données d'apprentissages.

Les méthodes de suivi reposant sur des modèles de type AAM et ASM sont connues pour leur manque de robustesse lorsque les conditions d'acquisition des images diffèrent

de celles de la base d'apprentissage (éclairage, occultation, etc.). Seuls des visages présentant une apparence suffisamment proche de celle des visages appartenant à la base peuvent être suivis avec une bonne précision.

Des travaux sont proposés pour adapter les AAM dans des conditions naturelles (variation d'illumination, variations de pose). Tzimiropoulos et al. [105] montrent qu'en adaptant les données d'apprentissage aux conditions naturelles, les AAM arrivent à mieux estimer la forme du visage dans ces conditions. D'autres travaux [78] proposent d'employer des descripteurs plus robustes comme les SIFT [74], SURF [8] et HOG [24] pour caractériser les pixels, ce qui augmente la précision du modèle, notamment en présence de variation de luminosité. Malgré cela, ces solutions sont difficilement adaptables dans des conditions naturelles, car elles nécessitent un modèle d'apprentissage prenant en compte un nombre infini de possibilités, ce qui rend très difficile l'optimisation d'un tel système.

Caractéristiques par approche discriminative

Contrairement aux méthodes génératives, les méthodes discriminatives reposent sur un apprentissage par regression appliqué directement sur l'image. Quant au modèle de distribution de points, il permet simplement de garantir une cohérence dans le positionnement des points. Parmi les méthodes discriminatives, deux méthodes se distinguent : les approches basées sur l'apprentissage par régression vectorielle et les méthodes de deep learning.

Le processus d'analyse des méthodes basées sur l'apprentissage par régression vectorielle est très similaire aux approches génératives. À partir d'un modèle initial de distribution de points, l'ensemble des landmarks est successivement mis à jour en utilisant un ensemble de N régressions en cascade [29]. À l'analyse d'un visage, l'entrée du vecteur de regression R_t à l'itération t appartient à (I,S_{t-1}), où I est une image et S_{t-1} est la forme estimée à l'itération précédente. La forme initiale S_0 correspond à la forme moyenne de la base d'apprentissage. La régression extrait les caractéristiques en fonction du modèle de forme actuel, et met à jour son modèle par l'équation suivante :

$$S_t = S_{t-1} + R_t(\Phi_t(I, S_{t-1})). \tag{2.5}$$

où $\Phi_t(I, S_{t-1})$ représente les caractéristiques de la forme précédentes. La régression en cascade permet d'affiner la forme au fur et à mesure des itérations, jusqu'à converger vers un modèle de forme similaire au visage analysé. Un exemple d'affinement du modèle de forme est illustré dans la Figure 2.5, où l'algorithme part d'une forme initiale S_0, pour ensuite converger de manière itérative sur le visage.

La différence majeure avec les approches génératives réside dans le fait que le modèle de forme peut s'adapter automatiquement au visage, sans que le modèle du visage appartienne à la base d'apprentissage.

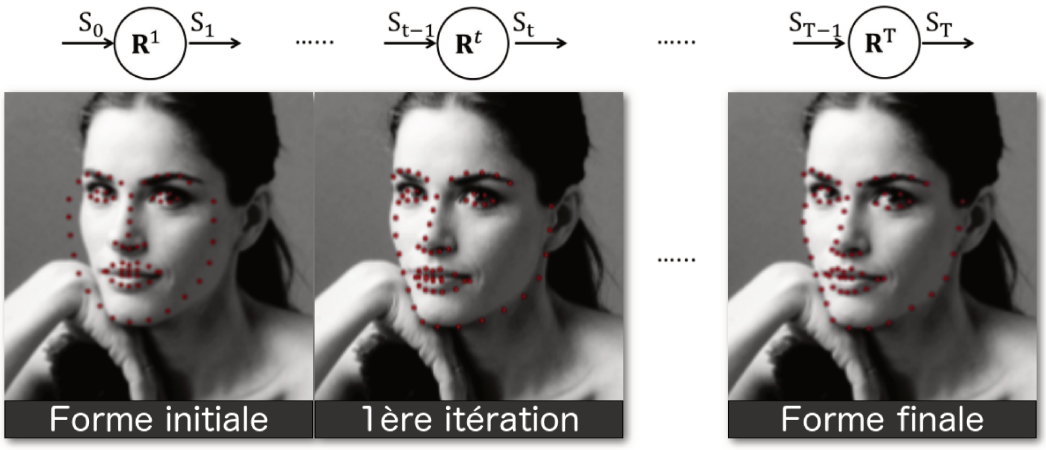


FIGURE 2.5 – Affinement du modèle du visage par méthode de regression en cascade.

Les méthodes basées sur le deep learning consistent à entraîner un réseau de neurones pour analyser un visage en vue de localiser les points caractéristiques. Désignés par l'acronyme CNN (en anglais : Convolutional Neural Network), ils comportent deux parties bien distinctes.

La première partie d'un CNN est la partie convolutive à proprement parler. Elle fonctionne comme un extracteur de caractéristiques des images. Une image est passée à travers une succession de filtres, ou noyaux de convolution, créant de nouvelles images ap-

pelées cartes de convolutions. Certains filtres intermédiaires réduisent la résolution de l'image par une opération de maximum local. Au final, les cartes de convolutions sont mises à plat et concaténées en un vecteur de caractéristiques, appelé code CNN. Un exemple d'architecture de CNN appliqué à la détection des visages (inspiré du fonctionnement des premières couches de la vision humaine [61]), est illustré dans la Figure 2.6.

Le code CNN en sortie de la partie convolutive (C1) est ensuite branché en entrée d'une deuxième partie, constituée de neurones entièrement connectés. Le rôle de cette partie est de combiner les caractéristiques du code CNN pour caractériser l'image. La sortie est une dernière couche comportant un neurone par classe. Les valeurs numériques obtenues sont généralement normalisées entre 0 et 1, de somme 1, pour produire une distribution de probabilité sur les classes. Dans le cas de la détection des landmarks, la classe obtenue correspond à un modèle de forme.

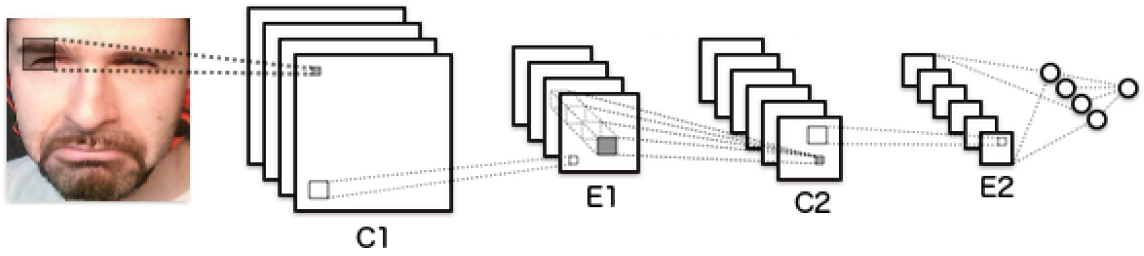


FIGURE 2.6 – Exemple d'architecture de CNN appliquée à la détection du visage, où C1, C2 représentent les convolutions et E1, E2 représentent les sous-échantillonnages.

La difficulté de ces approches repose essentiellement dans l'élaboration de la structure du réseau. Il s'agit d'abord de fixer l'architecture du réseau, c'est-à-dire le nombre de couches, leurs tailles et les opérations matricielles qui les connectent. L'entraînement consiste alors à optimiser les coefficients du réseau pour minimiser l'erreur de classification en sortie. La robustesse des CNN réside principalement dans la richesse de ces données d'apprentissage, où chaque cas (conditions d'acquisition, pose, expression) doit être représenté afin de converger vers une bonne estimation des landmarks. Ceci nécessite de prendre en considération un grand nombre très variées de données annotées. Cet entraînement peut prendre plusieurs semaines pour les meilleurs CNN, avec de nombreux GPU travaillant sur des centaines de milliers d'images annotées.

2.2.4 Synthèse

Nous avons vu dans cette section différentes approches pour détecter et caractériser un visage. Un visage peut être caractérisé sous forme de zones prédéfinies, sous forme de points d'intérêt, ou de modèles statistiques d'alignement facial. Le choix de la méthode dépend de ce que l'on désire caractériser au sein du visage.

La majorité des approches tend à utiliser des méthodes d'alignement facial pour caractériser localement le visage. Les algorithmes récents d'alignement facial basé sur un apprentissage profond, permettent de caractériser précisément un visage tout en satisfaisant la contrainte du temps réel. L'utilisation de ces solutions permet de renseigner à la fois des informations sur la géométrie, l'orientation et les mouvements faciaux.

Toutefois, ces méthodes semblent rencontrer des difficultés pour modéliser à la fois de larges déformations faciales (exemple : expressions faciales) et les variations de pose du visage. Ceci est dû au fait que les bases d'apprentissage destinées à l'entrainement de tels algorithmes sont essentiellement composées de visages frontaux ou légèrement inclinés, et où les visages sont généralement neutres.

La majorité des systèmes d'analyse faciale sont conçus pour analyser les visages dans de bonnes conditions d'acquisition (visage frontal à la caméra, illumination homogène). En supposant que le visage est analysé dans de bonnes conditions, cela assure que les caractéristiques visuelles sont parfaitement exploitables. Cependant, dans un contexte naturel, où les conditions d'acquisition changent, les visages ne sont pas directement exploitables dans la plupart des systèmes dans la littérature. Dans la section suivante, nous discutons des méthodes permettant d'améliorer la robustesse des systèmes dans un contexte naturel.

2.3 Normalisation du visage

La différence d'apparence d'un même visage capturé dans deux conditions d'acquisition distinctes pose de nombreux problèmes dans le domaine de l'analyse faciale. Cette différence est due, généralement, à des facteurs d'environnement comme les conditions

d'éclairage et de la position des capteurs par rapport au visage lors de l'acquisition. Cette variation peut aussi être due aux modifications liées aux expressions, à l'âge ou au genre. Dans ces conditions, les caractéristiques extraites sur le visage ne sont plus semblables, comme illustré dans la Figure 2.7. L'utilisation de ces caractéristiques peut alors avoir des répercussions sur des tâches sous-jacentes, telle que la reconnaissance des expressions faciales.

FIGURE 2.7 – Exemple d'un même visage extrait dans différentes conditions. A) expression, B) variation de pose, C) occultation par un élement structurel et D) changement d'éclairage.

Des méthodes de normalisation du visage sont employées afin de ramener dans un cadre commun de caractérisation plusieurs visages et d'en faciliter l'analyse. Plusieurs catégories de normalisation sont proposées dans la littérature. Nous pouvons distinguer trois catégories majeures de normalisation : la normalisation géométrique, la normalisation photométrique et la normalisation inter-individu. Dans la suite de cette section, nous discutons plus spécifiquement des problèmes liés aux variations de pose et aux occultations du visage. Nous décidons de ne pas nous attarder sur les problèmes d'illumination car ceux-ci ne sont pas considérés dans ce mémoire.

2.3.1 Normalisation géométrique

La normalisation géométrique consiste à appliquer une transformation géométrique sur un visage dans une image, afin de l'amener dans une configuration frontale. Cela permet d'obtenir les invariances du visage en translation, rotation, et changements d'échelle. Cette technique vise à garantir une superposition parfaite des différents composants du visage (yeux, nez, bouche), permettant ainsi de caractériser de manière robuste deux visages ayant une pose différente entre deux images. La Figure 2.8 illustre un processus

de normalisation de visages appliqué à un système d'identification faciale [25]. Le visage est tout d'abord détecté, puis une normalisation de la pose (Figure 2.8-1) et une mise à l'échelle (Figure 2.8-2) sont appliquées afin d'obtenir un visage adapté aux conditions d'analyse souhaitées (Figure 2.8-3).

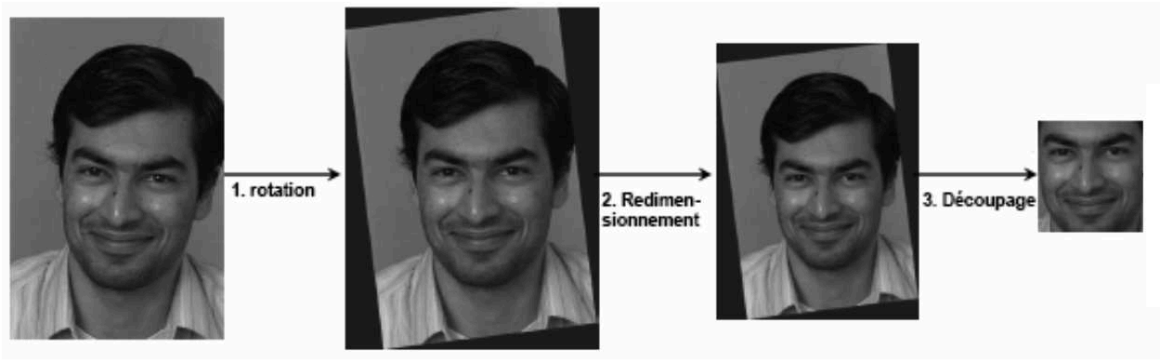


FIGURE 2.8 – Processus de normalisation d'un visage [25].

La normalisation 2D du visage consiste à estimer la meilleure transformation à appliquer aux visages, pour corriger les différences en termes de translation, rotation et de changements d'échelle. Cette transformation est généralement modélisée sous forme de multiplication matricielle (transformation linéaire), appliquée sur un visage. Basée sur une transformation affine, elle est représentée par une matrice M de dimension 2*3, calculée comme suit :

$$A = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}_{2*2} \quad B = \begin{bmatrix} b_{00} \\ b_{10} \end{bmatrix}_{2*1} \quad M = \begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & b_{00} \\ a_{10} & a_{11} & b_{10} \end{bmatrix}_{2*3} \tag{2.6}$$

où A représente la matrice de rotation et B la matrice de translation. La plupart des systèmes normalisent le visage en appliquant une transformation affine par rapport au centre des yeux car ces points ont tendance à rester stables indépendamment des expressions faciales [53, 92, 58]. Cependant, dû à la forte dépendance de ces deux points, d'importantes erreurs d'appariement peuvent apparaître, dans le cas où l'un des deux yeux vient à disparaitre (lunettes, pose extrême, ombre), ou lorsque leur localisation est erronée.

Des extensions de ces travaux utilisent un nombre plus important de points d'intérêt afin de garantir une meilleure stabilité en cas de mauvaise détection locale. Ces approches s'appuient sur la position des landmarks calculés par des méthodes de détection du visage (discutées dans la section 2.2). Certaines approches [53, 92] utilisent les points d'intérêt uniquement sur la région centrale du visage. D'autres [27] prennent aussi en compte les contours du visage afin d'obtenir des informations complémentaires sur la morphologie du visage. Bien que ces solutions soient plus robustes en présence de variations de pose et d'occultations du visage (lunettes, ombre), elles induisent d'importantes déformations faciales en présence de pose extrême du visage, et restent assujetties à la bonne détection des landmarks.

La qualité de la normalisation du visage dépend fortement de la position des landmarks. Or, un système d'acquisition d'images ne fournit que la projection des scènes observées sur un plan en deux dimensions. L'image ne permet donc d'exploiter que l'information résultant de la projection sur le plan 2D. Dans ce cas, le visage peut être fortement occulté, ce qui provoque de mauvaises estimations dans la position des landmarks. Comme illustré dans la Figure 2.9, plus le visage tourne, moins la détection précise des landmarks est assurée.

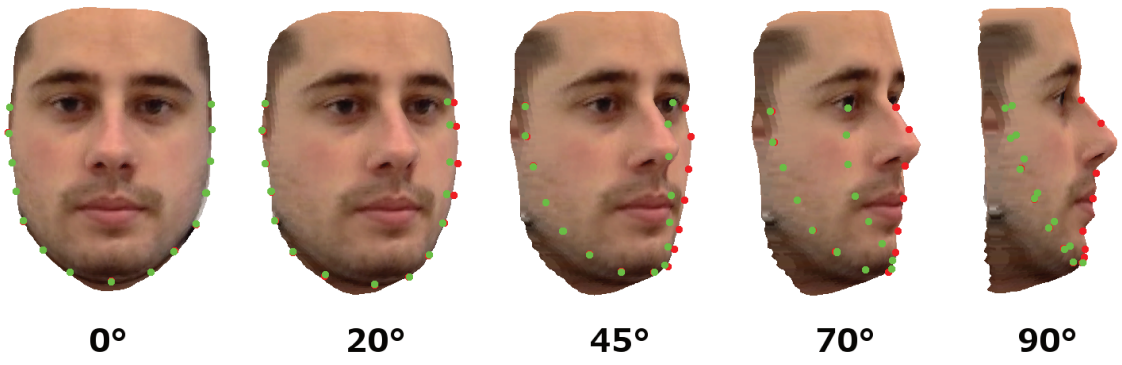


FIGURE 2.9 – Erreur de correspondance des points faciaux du visage 2D (rouge) et 3D (vert) sous plusieurs angles.

Afin d'être plus robuste aux problèmes de pose, de nouvelles approches [127, 46] proposent de reconstruire le visage 3D à partir de l'information 2D. La transformation géométrique est appliquée sur un modèle de visage déformable 3D afin d'aligner correcte-

ment le visage dans l'espace 2D. Le modèle 3D garantit la conservation de la morphologie du visage (sa forme globale et ses contours). Un système d'alignement facial 3D est illustré dans la Figure 2.10.

Dans ces méthodes, la première étape consiste à estimer les landmarks sur le visage 2D à l'aide des méthodes vues précédemment. Ensuite, la pose 2D est estimée à partir de la distribution des landmarks grâce à des algorithmes d'estimation de pose [118, 114], ce qui permet d'estimer la pose du modèle 3D en fonction du visage 2D. Puis, la texture du visage 2D est plaquée sur le modèle 3D en appliquant un maillage 3D sur le visage. Enfin, le modèle 3D est orienté dans le plan de l'image, afin d'obtenir un visage frontal.

En présence de variations de pose, certaines régions du modèle 3D peuvent ne pas être fortement texturées et nécessitent d'être reconstruites. Pour combler la perte d'information, des méthodes utilisent des techniques de remplissage de texture [49]. Ces méthodes sont efficaces sur des petites régions peu texturées (texture lisse), mais deviennent vite problématiques en présence d'occultations importantes ou sur les régions très texturées. D'autres [40] estiment les informations manquantes en se basant sur la symétrie du visage. Ces méthodes sont plus robustes en présence de fortes variations de pose. Cependant, elles supposent que la partie visible du visage n'est pas bruitée (ex : main sur le visage). Les méthodes récentes [50] utilisent l'information temporelle pour reconstruire les régions non texturées. Cette technique implique que le visage soit correctement appris sur une séquence d'images afin de pouvoir ensuite reconstruire les parties cachées.

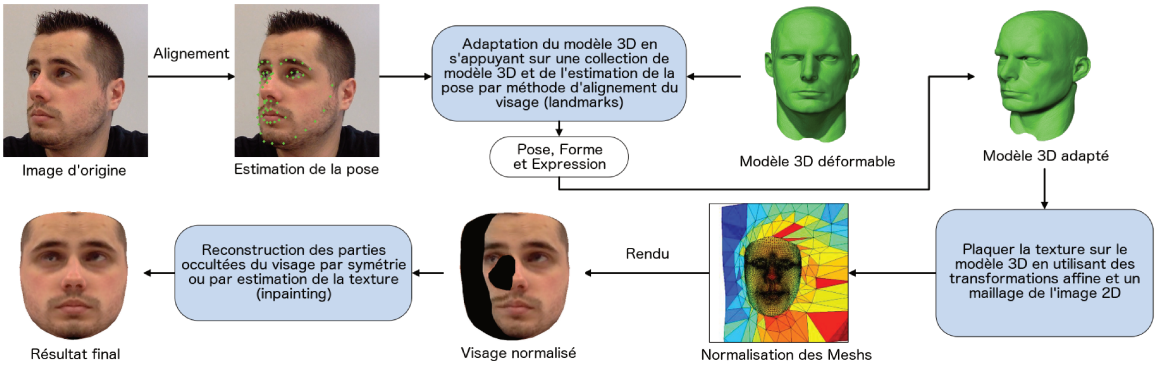


FIGURE 2.10 – Alignement du visage en plaquant la texture sur un modèle 3D et en reconstruisant les parties occultées.

Bien que les méthodes de normalisation 3D sont plus robustes en présence de rotations hors plan du visage et aux déformations faciales, elles ne sont pas infaillibles. La précision des approches 3D repose essentiellement sur la localisation des landmarks, où l'estimation de la pose 3D et le plaquage de la texture sont déterminants pour assurer une reconstruction parfaite entre le visage 2D et 3D. Les conditions d'illumination ou les occultations du visage (lunette, cheveux), peuvent venir bruiter la localisation des landmarks et impacter la précision du recalage 3D du visage. Il est important de noter que ces méthodes s'appuient sur des modèles déformables de visage 3D, qui demandent un apprentissage relativement long et un nombre de données volumineux pour s'adapter à un large panel de situations (position, morphologie, expression, genre, ...).

Toutes les approches de normalisation géométrique s'appuient sur l'hypothèse forte que les landmarks sont parfaitement détectés. Cela garantit une bonne correction de la pose. Ceci induit que les landmarks soient invariants aux variations de pose mais également aux changements lumineux et aux occultations du visage.

2.3.2 Présence d'occultations

La correction des occultations reste un défi majeur dans le domaine de la vision. Nous distinguons deux catégories d'occultation du visage : les occultations structurelles qui sont induites par des éléments directement placés sur le visage (lunette, maquillage, barbe), et les occultations extérieures provoquées par des éléments de la scène qui surgissent devant le visage (main, objet, personne), comme illustré sur la première ligne de la Figure 2.11.

La présence de composants structurels tels que la barbe, la moustache, ou bien les lunettes modifie les caractéristiques faciales telles que la forme ou la couleur. Ces occultations ont longtemps constituées des défis sérieux dans la détection des visages. Avec l'évolution des méthodes d'alignement de landmarks, principalement basées sur la géométrie faciale et sur des modèles 3D, il est possible de détecter les visages malgré la présence d'occultations [15]. L'inconvénient de ces approches est qu'elles ne garantissent pas une localisation exacte des landmarks au niveau des régions occultées, ce qui peut induire des déformations faciales et des erreurs d'appariement lors de la normalisation géométrique.

Bien que la localisation des landmarks soit de plus en plus précise sur les visages occultés, il n'en reste pas moins, que la caractérisation des visages occultés reste un problème difficile. En effet, certains composants structurels peuvent cacher complètement une partie du visage et de ce fait, rendre indisponible des informations importantes pour l'analyse faciale, comme les yeux. Dans certains cas, l'information perçue est suffisante afin de caractériser un visage, comme l'expression, le genre ou l'identification. Dans le cas où les informations perçues sur le visage permettent d'identifier la personne, certaines approches s'appuient sur des collections d'images de cette même personne afin de reconstruire les régions occultées. Généralement, ces solutions obtiennent de bonnes performances sur des célébrités, où les images sont abondantes (par le biais des réseaux sociaux et des médias).

Il n'est pas anodin que le visage subisse d'autres occultations, comme celles produites par les mains aidant à la construction du discours verbal. Ces occultations peuvent nuire aux systèmes d'analyse faciale. Pour détecter ces régions occultées sur un visage, plusieurs auteurs s'accordent à utiliser l'information temporelle. Smith et al. [**?**] utilisent un mélange de Gaussiennes (MoG). Les régions sur le point de devenir occultées sont caractérisées par une forte concentration de mouvements sur un court laps de temps. Mahmoud et al. [76] construisent des vecteurs d'occultation en fusionnant des informations spatiales (HOG, points d'intérêt) et spatio-temporelles (STIP). Dans l'hypothèse où l'occultation est ponctuelle, le contenu des séquences précédentes peut être exploité pour reconstruire les régions occultées [50].

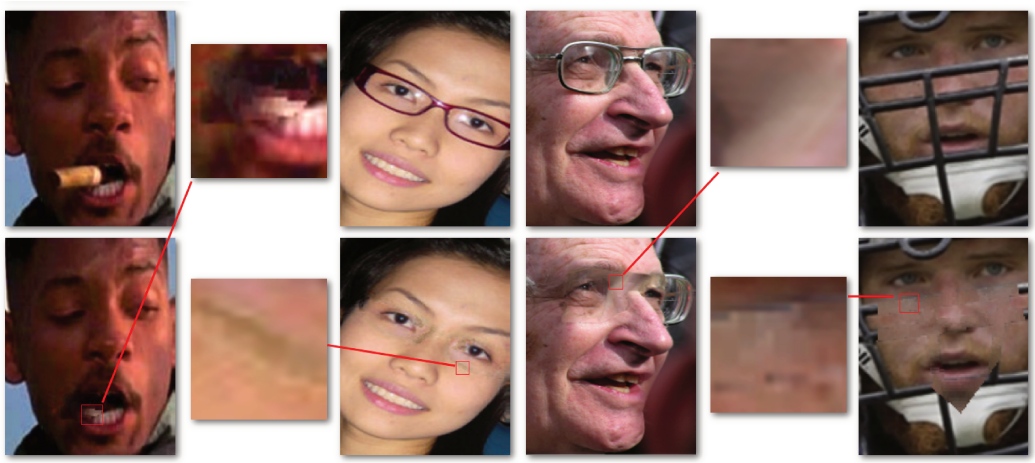


FIGURE 2.11 – Reconstruction des parties occultées de visage par inpainting [35].

Récemment, Jia et al. [121] ont proposé une solution permettant de reconstruire les parties du visage occultées à l'aide d'un auto-encodeur. L'approche consiste à détecter les régions occultées et à les reconstruire le plus fidèlement en calculant une image moyenne de la région cachée, en se basant sur une grande collection d'images. Farrugia et al. [35] proposent de reconstruire les parties occultées du visage en utilisant une méthode de remplissage (en anglais : inpainting). L'inpainting repose sur le principe du patch-match, qui consiste à estimer les valeurs des pixels manquants en fonction des pixels voisinant la région occultée, tout en conservant une cohérence de la texture entre des pixels connexes. Bien que les méthodes de reconstruction permettent de faciliter l'analyse faciale, les régions reconstruites ne garantissent pas l'authenticité par rapport à l'image originale (changement d'expression, de couleur, de forme), comme illustré dans la Figure 2.11. Il est également important de considérer que le temps de calcul de ces méthodes est relativement long et dépend notamment de la dimension des régions à reconstruire.

2.3.3 Synthèse

Avec un besoin grandissant de pouvoir analyser les visages dans un contexte naturel, la normalisation du visage est devenue une étape importante dans le processus d'analyse faciale. Elle garantit l'analyse parfaite des visages, en dépit des conditions d'acquisitions naturelles (variation de pose, occultation, changement de luminosité, ...). Les progrès réalisés dans ce domaine ont permis d'améliorer significativement les performances des systèmes face aux défis relevant de l'analyse faciale dans des contextes peu contraints.

Malgré les progrès réalisés, l'analyse faciale reste encore un challenge ouvert dans le domaine de la vision. Les propriétés du visage (texture lisse, élasticité de la peau, ...) font de celui-ci, un élément complexe à caractériser. Combiné aux problèmes d'acquisitions (pose, illumination, occultation), cela augmente considérablement la difficulté de l'analyse. Souvent traitées individuellement, les différentes problématiques liées à la normalisation peuvent apparaître simultanément et rendent généralement les méthodes de normalisation inefficaces en condition d'acquisition naturelle.

2.4 Extraction des caractéristiques visuelles

Considérant que les visages sont normalisés (pose et illumination homogènes, reconstruction des régions occultées), les conditions d'analyse sont excellentes pour pouvoir extraire des caractéristiques faciales. La caractérisation d'un visage consiste à extraire des informations contenues au sein du visage. Le visage peut être caractérisé soit :

- globalement : une seule région d'intérêt est analysée, et correspond au visage dans son intégralité ;
- localement : le visage est divisé en plusieurs régions d'intérêt, chacune caractérisée localement par des vecteurs de caractéristiques. Les vecteurs obtenus sont fusionnés pour construire le vecteur de caractéristiques associé au visage.

Les caractéristiques visuelles extraites au sein d'un visage permettent de renseigner diverses informations, telles que l'apparence (texture), la forme (géométrie) et la dynamique (mouvement) du visage. Ces informations ont leurs propres caractéristiques qui requièrent des traitements spécifiques. Dans la suite de cette section, nous discutons des méthodes employées pour caractériser chacune de ces informations.

2.4.1 Caractérisation de l'apparence faciale

Les méthodes pour caractériser la texture au sein d'un visage, consistent à appliquer des transformations mathématiques calculées sur l'ensemble des pixels de l'image pour ensuite y extraire un vecteur caractéristique global.

L'une des méthodes les plus utilisées pour caractériser la texture faciale, consiste à appliquer un descripteur de motif binaire local (en anglais : Local Binary Pattern, LBP) [1]. Le principe général est de comparer le niveau de luminance d'un pixel avec les niveaux de ses voisins. Cela rend donc compte d'une information relative à des motifs réguliers dans l'image, autrement dit une texture. Chaque pixel est représenté par un code binaire de 8 bits. Pour chaque pixel de l'image, la valeur du pixel en niveaux de gris est comparé à celle de ses voisins en respectant une 8 connectivité. La valeur du pixel de chaque voisin est égale à 0 si elle est plus petite que la valeur du pixel central, sinon elle est égale à 1. Selon l'échelle du voisinage utilisé, certaines régions d'intérêt telles des coins ou des

bords peuvent être détectées par ce descripteur. Pour réduire la dimension du vecteur, le visage est divisé en plusieurs régions en appliquant une grille de N*M blocs. Un vecteur est calculé au sein de chaque région, puis l'ensemble de ces vecteurs sont concaténés afin de créer un vecteur global permettant de caractériser le visage. La Figure 2.12 illustre le processus de caractérisation d'un visage à l'aide des LBP.

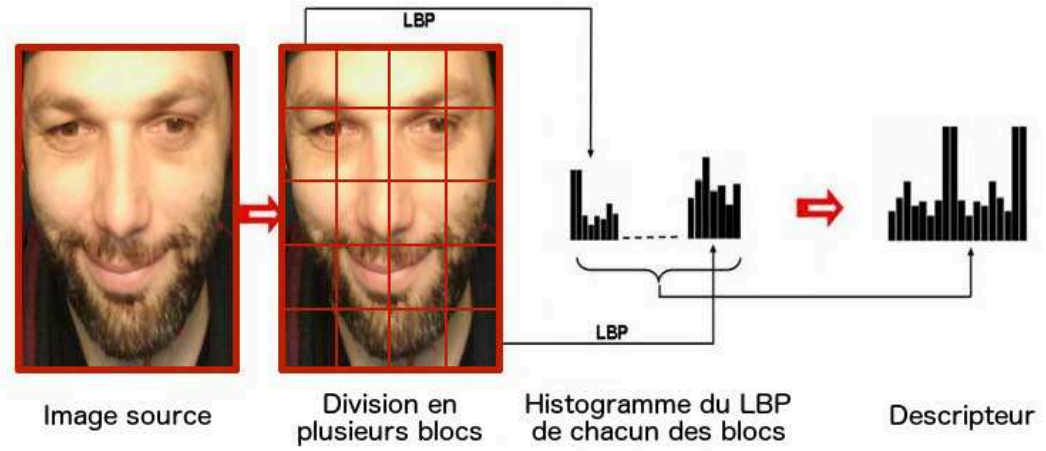


FIGURE 2.12 – Extraction des caractéristiques faciales à l'aide des descripteurs LBP (SNaP-2DFe).

Plusieurs variantes de ce processus d'analyse ont été proposées en y appliquant d'autres descripteurs comme les LGBP [100] (en anglais : local gabor binary pattern). Le descripteur LGBP consiste à appliquer des LBP sur des visages sur lesquels un filtre de gabor a été préalablement appliqué. Ce descripteur permet de prendre en considération la géométrie du visage en s'appuyant sur l'analyse des contours au sein du visage.

L'inconvénient majeur de ces méthodes, repose essentiellement sur la dimension du vecteur (proportionnel à la taille de l'image et aux informations extraites), ce qui augmente considérablement la complexité du calcul lors de l'étape de classification. Des méthodes sont proposées dans la littérature afin de diminuer la dimension du vecteur, et ainsi réduire la complexité du calcul. L'une de ces méthodes consiste à réduire la taille du visage à analyser, ce qui tend à faire disparaitre certaines informations contenues dans l'image. Danisman et al. [25] montrent qu'en réduisant la taille d'un visage, seules les informations au niveau des régions saillantes du visage (sourcil, yeux, bouche) sont conservées. Une autre méthode consiste à calculer un vecteur de caractéristiques sur un sous-ensemble de régions d'intérêt [39]. Parmi les caractéristiques locales couramment

calculées, on retrouve des motifs préalablement utilisés globalement, tels que des histogrammes de couleur ou des vecteurs rendant compte de l'orientation des gradients des niveaux de gris. Une autre méthode employée dans la littérature consiste à appliquer une analyse en composantes principales (ACP) qui permet de transformer des variables corrélées en nouvelles variables décorrélées les unes des autres [26]. L'ACP permet de réduire le nombre de variables du descripteur et de rendre l'information moins redondante.

Récemment, des méthodes de deep learning basées sur des réseaux de neurones convolutionnels (CNN) ont été proposées pour caractériser la texture au sein d'un visage [88, 72]. Ces méthodes intègrent à la fois dans le processus d'analyse, l'extraction et la classification des données. Leur fonctionnement consiste en un empilage multicouche de perceptrons, dont le but est de prétraiter de petites quantités d'informations. Chaque région du visage est traitée individuellement par un neurone artificiel qui effectue une opération de filtrage classique en associant un poids à chaque pixel de la région faciale. L'ensemble de ces neurones forme un noyau de convolution, qui analyse une caractéristique spécifique de l'image d'entrée. Pour analyser plusieurs caractéristiques, différentes couches de noyaux de convolution sont employées successivement, où chacune d'elles se spécialise dans une caractéristique spécifique.

Comparé aux autres méthodes plus traditionnelles, les réseaux de neurones convolutionnels utilisent relativement peu de pré-traitement. Cependant, cela nécessite d'avoir des bases de données exhaustives prenantes en compte l'ensemble des variations que peuvent subir les visages (occultations, variations de pose, illuminations). Pour obtenir suffisamment de données, une augmentation artificielle des données est employée sur les visages afin de simuler des variations de pose , des occultations et des changements lumineux à partir des données initiales des bases d'apprentissage. Les techniques généralement utilisées consiste à appliquer des filtres, des changements d'échelles et des rotations sur les visages.

2.4.2 Caractérisation de la géométrie faciale

Bien que la texture soit majoritairement utilisée pour caractériser les visages, d'autres méthodes s'intéressent à caractériser la forme du visage. La robustesse des méthodes d'alignement faciale facilite l'émergence des approches géométriques. Grâce à la loca-

lisation des landmarks, certaines approches [37] proposent de construire un vecteur caractéristique global en calculant la distance entre les différents points du modèle facial, comme illustré dans la Figure 2.13-A et 2.13-B. D'autres méthodes [39] caractérisent la forme des différents contours du visage en s'affranchissant de la localisation des landmarks, en utilisant des algorithmes de détection de contours (voir Figure 2.13-C). Les méthodes géométriques ont l'avantage d'être invariantes aux translations, rotations et aux changements d'échelles et sont donc moins sensibles aux déformations faciales induites par les méthodes de normalisation. L'inconvénient principal des méthodes géométriques est qu'elles ne permettent pas de distinguer facilement deux morphologies faciales relativement proches.

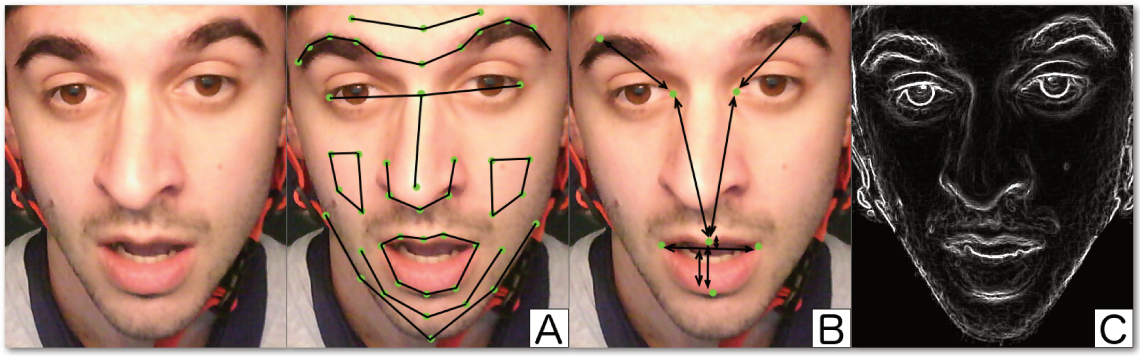


FIGURE 2.13 – Représentation géométrique du visage à l'aide du modèle facial A) et B) ou par analyse des contours C) (SNaP-2DFe).

Les caractéristiques géométriques sont employées dans certaines situations pour délimiter les régions d'intérêt au sein du visage, afin de pouvoir y calculer des caractéristiques d'apparence. Han et al. [38] appliquent une grille de N*M régions sur le visage, où chaque région est caractérisée par des LBP. À l'instar des approches reposant uniquement sur la texture, ils utilisent les landmarks pour adapter la géométrie des régions en fonction de la morphologie du visage. Les angles et les distances de chaque région sont ensuite regroupés dans un unique vecteur, caractérisant la géométrie faciale. Enfin, les deux caractéristiques (texture et géométrie) sont fusionnées pour caractériser le visage. La fusion de ces deux caractéristiques présente souvent un gain des performances car elles renseignent sur des caractéristiques complémentaires.

2.4.3 Caractérisation de la dynamique faciale

La caractérisation de la dynamique faciale consiste à analyser les déformations apparentes sur le visage dans le temps. Les travaux de Ambadar et al. [4] et de Bassili [7] montrent l'importance de la prise en compte de la dynamique faciale pour caractériser les visages car cela permet d'extraire des informations subtiles, parfois non perçues par les approches d'apparences classiques.

Au vu des bonnes performances obtenues avec les approches caractérisant l'apparence faciale, la majorité des approches proposées ont été etendues afin de caractériser la dynamique faciale. Zhao et al. [124] proposent une extension des LBP, nommé LBP-TOP (en anglais : Local Binary Patterns from Three Orthogonal Planes). Les LBP-TOP consistent à appliquer les LBP sur les 3 plans de l'image, plus précisément sur le plan XY (image d'origine), sur le plan X_t (l'évolution des pixels sur l'axe X) et sur le plan Y_t (l'évolution des pixels sur l'axe Y). Un vecteur moyen est ensuite calculé en concaténant le vecteur de ces 3 plans. Les LBP-TOP permettent de combiner l'information de mouvement et d'apparence au sein du même descripteur. La Figure 2.14 illustre le processus d'analyse faciale à l'aide des LBP-TOP. D'autres méthodes ont été étendues en suivant le même procédé comme les LGBP-TOP [3] et les LQP-TOP [3].

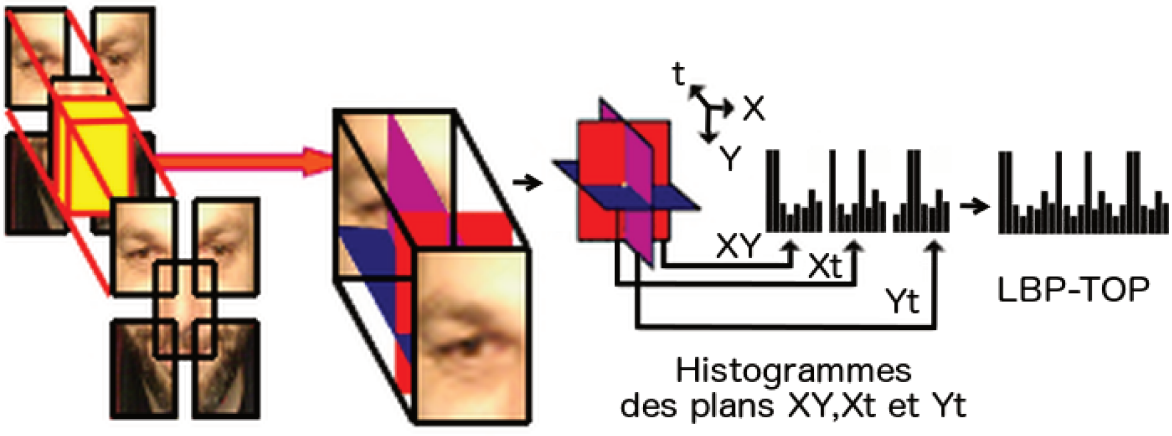


FIGURE 2.14 – Extraction des caractéristiques faciales à l'aide des descripteurs LBP-TOP.

Il es est de même pour les approches neuronales, où des réseaux de neurones récurrents (RNN) ont été adaptés pour analyser des données d'entrée de tailles variables [42].

Ils conviennent en particulier pour l'analyse de séries temporelles. Les techniques d'entraînement du réseau sont les mêmes que pour les réseaux classiques (rétropropagation du gradient), néanmoins les réseaux de neurones récurrents se heurtent au problème de disparition du gradient pour apprendre à mémoriser des évènements passés. Des architectures particulières répondent à ce dernier problème. Les réseaux les plus utilisés pour répondre au problème de disparition de gradient sont les réseaux récurrents à mémoire court et long terme (en anglais : Long short-term memory, LSTM) [13, 57]. Dans un LSTM chaque unité de traitement est liée non seulement à un état caché mais également à un état mémoire. Le passage entre deux états se fait par transfert à gain constant et égal à 1. De cette façon, les erreurs se propagent sans phénomène de disparition de gradient.

Parmi les approches caractérisant la dynamique faciale, les méthodes de flux optique sont très populaires car elles sont mieux adaptées pour caractériser un mouvement. Le flux optique caractérise le mouvement apparent des objets, surfaces et contours d'une scène visuelle, causé par le mouvement relatif entre un observateur (l'œil ou une caméra) et la scène. Le flux optique est représenté par un vecteur de mouvement qui décrit une transformation d'une image en deux dimensions vers une autre. Les niveaux de granularité sont différents selon l'algorithme : toute l'image peut être liée au vecteur comme c'est le cas pour l'estimation de mouvement global (échantillonnage dense) [66], ou juste des parties spécifiques de l'image (échantillonnage local), tels que des formes (sourcil, bouche) [96]. La Figure 2.15 représente le mouvement extrait à partir d'un flux optique dense en présence d'une expression faciale. Bien que l'expression faciale soit très subtile, le flux optique arrive tout de même à percevoir le mouvement induit par l'expression.

L'utilisation des approches dynamiques est particulièrement délicate lorsqu'elle est appliquée à l'analyse faciale en conditions de captation peu contrainte. La difficulté réside dans la spécificité des mouvements associés au visage. Ces mouvements se caractérisent entre autres par des fortes variations de pose et de larges déplacements. Dans ce contexte, les solutions actuelles issues de la littérature, qui s'appuient sur des caractéristiques stables de la fonction de luminance et sur l'hypothèse d'un mouvement rigide, s'avèrent en général mal adaptées. En effet, le mouvement est un problème mal posé en vidéo puisqu'il décrit un contexte en trois dimensions alors que les images sont une projection de scènes 3D dans un plan en 2D.

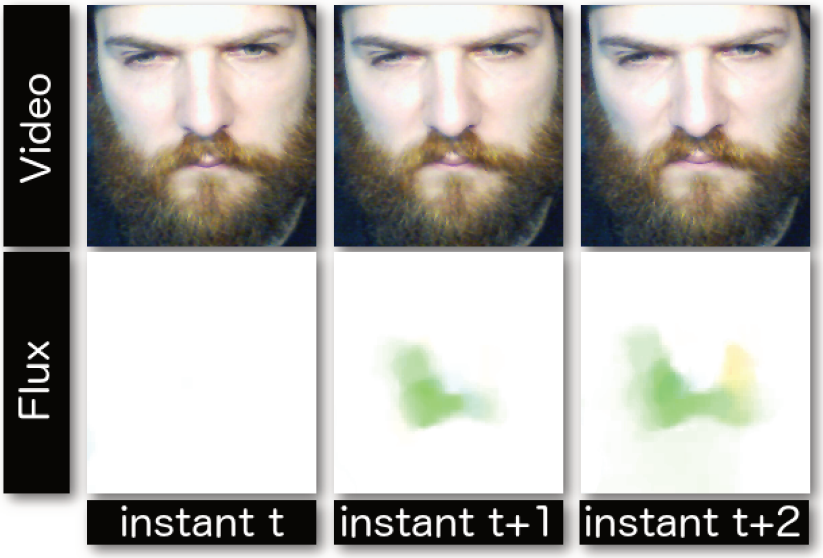


FIGURE 2.15 – Caractérisation du mouvement facial par flux optique. La direction du mouvement est associée à la couleur et l'amplitude du mouvement est associée à l'intensité des pixels (SNaP-2DFe).

2.4.4 Modèles de segmentation faciale

Toutes les approches citées précédemment sont associées à un modèle de segmentation faciale afin de caractériser les visages. Un modèle de segmentation faciale consiste à diviser le visage en fonction des différentes régions que l'on souhaite caractériser. Le modèle de segmentation dépend généralement du niveau de granularité de l'analyse (ex : nombre de région) et des caractéristiques souhaitées. La Figure 2.16 illustre différents modèles de segmentation faciale utilisés dans la littérature pour caractériser un visage.

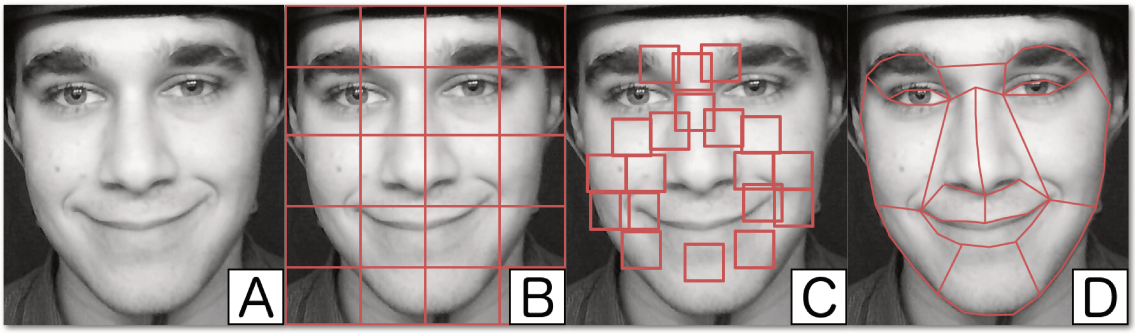


FIGURE 2.16 – Différents modèles de segmentation faciale pour caractériser un visage. (A) visage entier, (B) grille, (C) régions saliantes et (D) maillage (SNaP-2DFe).

Assumant que la tête reste fixe sur l'ensemble des visages caractérisés, le modèle de segmentation par grille est le plus utilisé [33]. Ce modèle a l'avantage d'être rapide à calculer et le nombre de régions est facilement paramétrable. Cependant, ce modèle est très sensible aux mouvements de la tête, car dans ce cas les régions faciales ne correspondent plus entre deux images successives. C'est-à-dire que les pixels d'une région N à l'instant t ne reflèteront pas les mêmes éléments du visage à l'instant $t+1$. Cela provoque de faux appariements entre les pixels et donc l'apparition de faux mouvements, ou des différences en termes de texture. Dans la majorité des cas, les composants du visage comme les yeux, le nez et la bouche sont utilisés pour normaliser le visage, comme vue dans la section 2.3.

L'inconvénient majeur des modèles de segmentation par grille, est qu'ils ne prennent pas en compte les déformations faciales induites par les mouvements à l'intérieur du visage (déplacement des lèvres, des sourcils, ...). C'est-à-dire qu'une région faciale ne s'adapte pas automatiquement à la géométrie faciale. Dans ce cas, une telle modélisation est difficilement adaptable, car la géométrie faciale d'un nombre important d'individus peut varier significativement. Plusieurs solutions permettent de résoudre ce problème en s'appuyant directement sur la géométrie faciale. Happy et al. [39] utilisent les composants faciaux tels que les yeux et les coins des lèvres pour estimer la position de plusieurs régions saillantes, qui s'adaptent dynamiquement en fonction de la géométrie faciale. Cette technique garantie de conserver les mêmes pixels malgré le fait que le coin des lèvres bouge.

D'autres approches utilisent les landmarks pour définir des régions faciales, ce qui augmente la robustesse du modèle de segmentation en présence de déformations faciales. Sadeghi et al. [94] améliorent le modèle de segmentation par grille en utilisant les landmarks. Grâce à cela, ils obtiennent une grille dynamiquement déformable en fonction de la géométrie faciale. Jiang et al. [52] définissent un maillage du visage en utilisant un ASM, ce qui rend leur méthode plus robuste aux changements de pose.

Les modèles de segmentation du visage basés sur des grilles, des régions saillantes ou des maillages ont été utilisés avec différentes approches de caractérisation du visage. Cependant, malgré l'utilisation de ces différents modèles, il est difficile d'identifier un modèle universellement adapté à l'ensemble des processus d'analyse. Dans la section sui-

vante, nous faisons une synthèse des différentes approches et des modèles de segmentation permettant de caractériser les visages.

2.4.5 Synthèse

Durant ces dernières années, plusieurs méthodes ont été proposées pour extraire des caractéristiques faciales sur des images 2D. Les méthodes basées sur l'apparence faciale ont prouvé leur efficacité et ont été beaucoup utilisées dans la littérature. Bien que la géométrie faciale permette de renseigner de nombreuses informations sur le visage, elles sont moins utilisées que les méthodes basées sur l'apparence. Généralement, les méthodes géométriques sont utilisées pour adapter les méthodes d'apparence à la morphologie faciale. Des récents travaux s'intéressent à la caractérisation de la dynamique faciale. Ces méthodes s'appuient sur des données contextuelles, en se basant sur une petite séquence d'images successives. Bien que ces méthodes aient prouvé leur efficacité dans des conditions contrôlées (pose fixe, lumière homogène), elles sont généralement mal adaptées en conditions d'interaction naturelle.

L'extraction de caractéristiques faciales peut être appliquée globalement ou localement sur le visage. Bien que l'analyse globale soit utilisée dans de nombreuses tâches sous-jacentes à l'analyse faciale, la dimension du vecteur caractéristique est généralement trop importante et rend la classification difficile. La caractérisation du visage repose principalement sur une analyse locale. Souvent utilisé dans la littérature, le modèle de segmentation reposant sur une grille a été progressivement remplacé par un modèle de maillage facial. Ces modèles s'appuyant sur la position des landmarks, ils permettent d'adapter les régions à la géométrie faciale, tout en étant plus robuste aux variations de pose et aux mouvements faciaux (ex : expressions).

2.5 Conclusion

Ce chapitre a présenté un état de l'art portant sur l'analyse faciale. Nous avons vu qu'un processus d'analyse faciale se compose de plusieurs étapes qui sont : la détection, la normalisation et l'extraction de caractéristiques. Chacune de ces étapes a fait l'objet de nombreuses recherches, permettant ainsi d'apporter bon nombre d'informations, et d'adapter le choix des méthodes employées en fonction de l'analyse souhaitée.

L'adaptation du processus d'analyse faciale repose essentiellement sur la spécificité des données analysées. L'un des premiers critères à prendre en considération est de regarder dans quelles conditions les données analysées sont extraites. La majorité des systèmes d'analyse faciale proposés dans la littérature permettent d'obtenir de bonnes performances dans des conditions contrôlées (lumière homogène, pose du visage fixe et frontal par rapport à la caméra). Malheureusement, ces conditions d'acquisition ne reflètent pas les conditions naturelles où la présence de variations de pose, d'occultations et de variations lumineuses viennent renforcer la difficulté de l'analyse. Dans ces conditions, les systèmes cités précédemment ne parviennent pas à analyser correctement le visage.

Afin d'adapter les systèmes aux conditions d'acquisition naturelles, des méthodes de normalisation du visage ont été proposées. La normalisation du visage permet de corriger les différentes invariances entre des visages pour les amener dans des conditions d'analyse idéales. Indépendamment des méthodes de normalisation, d'autres étapes du processus d'analyse faciale ont évolué afin d'améliorer la robustesse des méthodes proposées, notamment concernant la détection du visage, où la présence de variations de pose nécessite de pouvoir détecter correctement un visage malgré une occultation partielle. Les méthodes d'extraction de caractéristiques visuelles ont également évolué afin d'améliorer leur robustesse.

Bien que la normalisation du visage permette d'améliorer significativement les performances des systèmes d'analyse faciale, la normalisation induit parfois des changements néfastes au sein du visage. Cela est principalement dû au fait que le visage est un élément difficilement analysable. Cette difficulté réside notamment par le fait qu'il soit non-rigide, que sa texture est principalement lisse, et que sa morphologie varie entre deux individus. Ajouté à cela, les problèmes d'occultations liés aux variations de pose et aux problèmes d'acquisition, l'analyse faciale est une tâche complexe.

Il est important d'adapter les méthodes de caractérisation du visage en fonction de ce que l'on désire analyser, mais également en fonction des normalisations appliquées en amont de l'analyse. En effet, certaines méthodes de caractérisation sont plus sensibles que d'autres à divers changements (luminosité, morphologie, texture). Il est alors néces-

saire de trouver le meilleur compromis entre la correction des invariances par les méthodes de normalisation et la prise en compte des invariances dans les méthodes de caractérisation.

Dans le chapitre suivant, nous présentons les systèmes d'analyse faciale adaptés pour l'analyse des expressions faciales afin de caractériser l'état affectif d'un individu. Nous discutons dans un premier temps, de la représentation des états affectifs en exploitant les informations extraites du visage. Puis, nous présentons pour chacune des étapes du processus d'analyse faciale, quelles sont les solutions retenues pour construire un système adapté aux expressions faciales. Ensuite, nous discutons sur la conception des bases d'apprentissage et de leur évolution afin de s'adapter aux différentes conditions d'acquisition. Enfin, nous présentons notre positionnement par rapport aux différents systèmes actuels.

Chapitre 3

L'étude des expressions faciales

« *Pour exprimer son âme,*
on n'a que son visage. »

Jean Cocteau

Sommaire

3.1 Introduction

La manifestation de nos émotions passe par plusieurs composantes, notamment cognitive (évaluation et traitement de l'information perçue), physiologique (nos émotions suscitent parfois une augmentation de notre fréquence cardiaque, de notre rythme respiratoire, de notre transpiration, etc) et expressive (expression externe de l'émotion : expressions faciales, posture, gestuelle, timbre de voix, etc). Ainsi, là où un neurologue par exemple, sera attaché à des notions de facteurs somatiques ou d'activation cérébrale, un sociologue, aura une vision bien plus globale, et déterminera des valeurs liées à des paramètres sociaux donnés pour étudier par exemple un mouvement de panique de masse ou encore l'anxiété face à une situation donnée. Il n'en reste pas moins que les différentes approches conceptuelles apportent des points de vue complémentaires, qui permettent de mieux cibler la difficulté du problème.

Ne nécessitant aucun capteur intrusif, l'analyse des expressions faciales attire l'attention de nombreuses recherches. Outre le fait des défis liés aux différences inter-individu comme le genre, la culture, d'autres problématiques sont à prendre en compte lorsque l'on s'intéresse aux expressions faciales. Ekman [30] montre qu'une expression faciale est le résultat d'une activation des muscles faciaux, où l'intensité de ces changements est intimement corrélée à l'intention d'une personne à communiquer son état affectif. D'une manière générale, les processus d'analyse des expressions faciales sont par nature très complexes, où les expressions faciales sont difficiles à caractériser et dépendent du contexte dans lequel elles sont traitées. La difficulté d'analyse est renforcée par la présence de variations de pose, et de larges déplacements liés aux conditions d'acquisitions naturelles. Bien que les problèmes soient divers et variés, la plupart des approches de la littérature s'appuient sur des processus standard d'analyse faciale pour caractériser les expressions, en y apportant progressivement des améliorations afin d'être plus robustes, notamment aux changements d'intensité et aux problèmes de pose.

Les systèmes permettant d'analyser les expressions faciales s'appuient généralement sur des classifieurs et nécessitent donc des bases d'apprentissages. Ces bases d'apprentissages sont constituées d'un ensemble d'images et/ou de vidéos destinées à fournir une collection de données exhaustive pour caractériser les expressions faciales. Chaque base

d'apprentissage est élaborée pour répondre à une problématique particulière (présence d'occultations, de variations lumineuses, ...) en respectant un protocole d'acquisition particulier. En complément des données visuelles, les bases d'apprentissages peuvent contenir d'autres informations, telles que des points caractéristiques sur le visage, des informations sur les expressions faciales (durée, fréquence, intensité). Ces informations favorisent l'émergence de nouveaux algorithmes nécessitant une vérité-terrain plus approfondie.

Dans la suite de ce chapitre, nous discutons de la manière dont les états affectifs sont caractérisés au regard des expressions faciales. Puis, nous présentons deux défis majeurs de l'analyse des expressions faciales, qui consistent d'une part à rendre l'analyse invariante aux changements d'intensité des expressions et d'une autre part, d'être robuste aux changements de pose et de larges déplacements du visage. Nous détaillons comment les processus d'analyse faciale s'adaptent pour prendre en compte ces défis. Enfin, nous parlons des différentes bases d'apprentissage, de leur conception et de leur capacité à refléter les défis posés par une analyse des expressions faciales en contexte d'usage naturel.

3.2 Modélisation de l'état affectif

Plusieurs modèles ont été présentés dans la littérature pour décrire formellement les émotions. Ces modèles sont construits en fonction des différentes composantes qui permettent de manifester nos émotions (cognitive, physiologique et expressive). Deux modélisations se démarquent dans la littérature, la modélisation catégorielle, basée sur un ensemble d'émotions dites "basiques", universelles, non réductible et innées; et la modélisation dimensionnelle, basée sur le principe que les émotions résultent d'un nombre fixé de concepts, et propose donc de les représenter dans un espace multidimensionnel. Dans la suite de ce chapitre, nous expliquons la différence entre ces deux représentations.

3.2.1 Modélisation catégorielle

Dans les années 1970, Paul Ekman met en évidence les mouvements automatiques, universels et innés des émotions de base (haine, dégoût, peur, joie, tristesse, surprise). Pour décrire ces changements, le système d'action du visage de Paul Ekman [30] (FACS) a été largement utilisé pour étiqueter manuellement les mouvements faciaux. Dans le

système FACS, les contractions ou décontractions du visage sont décomposées en unités d'action (en anglais : action unit, AU). Le système FACS repose sur la description de 46 AUs identifiées par un numéro dans la nomenclature FACS. Par exemple, l'AU 1 correspond au mouvement de lever intérieur des sourcils. Chaque AU peut correspondre à la contraction ou à la détente d'un ou plusieurs muscles, qui se traduit par un mouvement d'une partie donnée du visage. La Figure 3.1 illustre l'enclenchement de différents AUs sur un visage. Une expression faciale correspond donc, à la mise en jeu de plusieurs unités d'action. Par exemple, dans le cas d'un visage apeuré (3.1-A), les AUs mises en jeu seraient : AU 1 (lever les sourcils intérieurs) + AU 2 (lever les sourcils extérieurs) + AU 23 (tension des lèvres) + AU 26 (abaissement de la mâchoire inférieure).

FIGURE 3.1 – Modélisation des expressions faciales basée sur le système FACS. A) expression de peur, B) expression de dégoût (SNaP-2DFe).

Dans le cas d'une modélisation catégorielle, les expressions faciales sont labélisées en fonction des émotions universelles (haine, dégoût, peur, joie, tristesse, surprise) ou en fonction de l'enclenchement des AUs. L'ensemble de ces valeurs de sortie est fini ($Y = \{1,\ldots,I\}$), on parle alors d'un problème de classification, qui revient à attribuer une étiquette à chaque entrée. La fonction de prédiction est alors appelée un classifieur.

Plusieurs processus d'analyse faciale s'appuient sur une méthode d'apprentissage par modèle catégoriel pour classifier les expressions en fonction des émotions universelles (6+N, où N est l'expression neutre). Happy et al. [39] définissent des régions saillantes sur le visage, permettant de mieux discriminer les expressions faciales. Chaque région

est alors représentée par un descripteur LBP. Puis, l'ensemble des descripteurs est concaténé pour caractériser l'expression. Zhao et al. [124] proposent d'appliquer une segmentation faciale à l'aide d'une grille où chaque bloc est caractérisé par des LBP-TOP, prenant ainsi en considération l'information temporelle. Liao et al. [66] définissent des patrons de mouvement faciaux pour chaque expression en utilisant les flux optiques. Une séquence de mouvement est alors associée à un patron correspondant à l'une des six expressions universelles.

D'autres préfèrent représenter une expression faciale par l'activation des muscles faciaux, en se basant sur le système FACS. Jiang et al. [53] proposent d'appliquer une segmentation faciale par grille pour décomposer le visage en plusieurs blocs, où chaque bloc est caractérisé par un LQP-TOP. Les différentes AUs sont alors associées à un ou plusieurs blocs. L'ensemble des descripteurs correspondant à chaque AU sont alors concaténés pour caractériser une expression. Zhu et al. [128] proposent de construire un automate retraçant l'activation des AUs. Les différentes règles appliquées permettent de conserver une certaine cohérence dans l'ordre d'activation et d'enlever les séquences qui ne satisfont pas un mouvement facial physiquement réalisable (i.e. l'ordre d'apparition des AUs est non conforme au système FACS). Jaiswal et al. [47] proposent de segmenter le visage en 27 régions, à l'aide d'un maillage basé sur les muscles faciaux. Chaque zone délimitée par le maillage est alors représentée par un LBP. Enfin, ils proposent d'utiliser un réseau de neurones pour estimer une cohérence dans l'ordre d'activation des AUs. Ce système tend à améliorer l'utilisation du système FACS, de manière automatique et évolutive en fonction des données alimentant le réseau.

Selon Ekman, les expressions du visage ne seraient pas déterminées par la culture. De nature biologique, ces émotions de base paraîtraient identiques pour tous, de façon indépendante. Le rôle de l'imitation dans l'apprentissage de ces expressions ne serait que secondaire, ne servant qu'à renforcer le système. La colère, le mépris, la peine seraient ainsi mondialement reconnaissables. Cependant, ces théories de l'universalité des émotions faciales ne font pas l'unanimité du monde scientifique et trouvent des contradicteurs. Une étude de Jack et al. [45] les remet en question à l'aide de données opposées qui démontrent le concept de différences culturelles dans le jugement des ressentis. Les signaux ne seraient donc pas universels et doivent être interprétés en fonction du contexte.

Que ce soit pour analyser les expressions universelles ou les AUs, les systèmes effectuent souvent leur apprentissage sur une base de données contenant une grande quantité de visages expressifs. Cela permet aux classifieurs d'apprendre les déformations du visage liées à l'expression, et cela pour tous les types d'identité et pour toutes les intensités. Généralement, ces systèmes se focalisent sur la reconnaissance d'un petit nombre d'émotions, mais ne sont pas bien adaptés pour traiter les émotions complexes comme une émotion qui évolue graduellement vers une autre émotion ou pour différencier les différentes nuances pour une même émotion.

3.2.2 Modélisation dimensionnelle

L'approche dimensionnelle est basée sur le principe que les émotions résultent d'un nombre fixé de concepts, et propose donc de les représenter dans un espace multidimensionnel [93]. Par exemple, les dimensions peuvent correspondre à un axe de plaisir et de déplaisir (valence), d'éveil ou d'ennuie (arousal), de nervosité, de puissance, de maîtrise de soi et bien d'autres au besoin du modèle. Ce modèle a rencontré beaucoup de succès car il permet de représenter une infinité d'émotions. Il permet également de représenter des émotions nuancées.

De nos jours, la plupart des chercheurs qui s'inscrivent dans cette approche s'accordent sur les deux premières dimensions : la valence et l'activation (ou arousal). La valence permet de distinguer les émotions positives, agréables, comme la joie, des émotions négatives, désagréables, comme la colère. L'activation représente le niveau d'excitation corporelle, qui transparaît par nombre de réactions physiologiques, comme l'accélération du cœur, la transpiration.

La modélisation dimensionnelle est souvent abordée comme un problème de prédiction, où l'interaction est analysée dans le temps pour déterminer l'évolution des dimensions affectives. Dans ce cas, la modélisation dimensionnelle est estimée par un problème de régression, où chaque dimension correspond à une valeur dans un ensemble continu de réels ($Y \subset \mathbb{R}$). La modélisation dimensionnelle n'étant pas abordable par un modèle linéaire, il est possible d'effectuer une régression approchée par des algorithmes itératifs, on parle alors de régression non linéaire.

Dans une modélisation dimensionnelle, il est plus difficile de définir des classes car il en existe une infinité. Contrairement à la modélisation catégorielle, à partir d'exemples d'entrainement annotés suivant ces classes, le calcul de l'hyperplan est effectué parmi une infinité d'hyperplans possibles.

Afin de réduire la complexité de la classification, la modélisation dimensionnelle peut être toutefois abordée comme un problème de classification [83]. Dans ce cas, une expression peut être représentée par un couple de valeur valence/arousal, ainsi la joie se caractérise par une valence positive et un arousal positif. En contrepartie, la richesse du modèle dimensionnel est perdue, où une infinité de classes sont réduites à un ensemble fini de classes. La Figure 2.2 est une représentation graphique d'un modèle dimensionnel, où chaque émotion est caractérisée par un niveau de valence (en abscisse) et par un niveau d'arousal (en ordonnée).

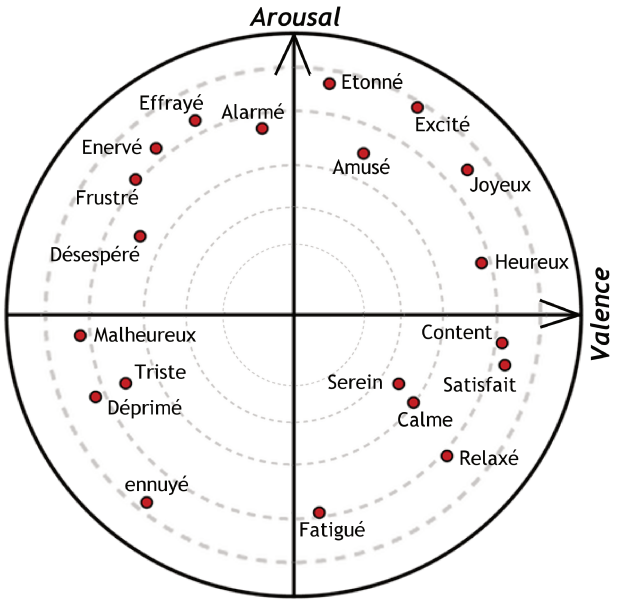


FIGURE 3.2 – Modelisation des états émotionnels par deux dimensions : valence/arousal.

Nicolaou et al. [84] constatent que les dimensions d'arousal et de valence sont corrélées. Leurs résultats suggèrent que les corrélations offrent une plus grande précision de la prédiction continue des états affectifs. Les indices visuels tels que les expressions faciales semblent obtenir de meilleurs résultats pour la prédiction de valence que pour l'arousal. Li et al. [63] proposent une nouvelle architecture de CNN adaptée à l'analyse faciale

et aux dimensions affectives. Leur architecture est basée sur un système d'analyse forward/backward permettant de prendre en compte l'évolution des dimensions affectives dans une séquence d'images. Pour réduire la complexité de l'analyse, ils décomposent les séquences en plusieurs sous-séquences, en conservant un chevauchement entre les sous-séquences pour respecter la continuité des informations. Kollias et al. [59] utilisent un réseau de neurones récurrent pour estimer les dimensions affectives de valence et d'arousal dans une séquence d'images. L'utilisation d'un tel réseau leur permet de s'adapter à des séquences en entrée de tailles variables et d'apprendre à mémoriser des évènements contextuels (informations extraites des images précédentes). Les résultats obtenus montrent que les réseaux de neurones semblent bien adaptés pour caractériser les dimensions affectives.

3.2.3 Synthèse

Deux modélisations sont principalement utilisées dans la littérature pour analyser les états affectifs à partir des expressions faciales : la modélisation catégorielle et la modélisation dimensionnelle. Quelle que soit la modélisation choisie, les expressions faciales se caractérisent généralement par manifestation observable d'un ensemble d'activations des muscles faciaux. L'ensemble de ces observations est ensuite traité pour identifier une expression faciale.

La différence majeure entre les deux modélisations réside essentiellement dans le modèle de prédiction utilisé. Dans le cas de la modélisation catégorielle, les observations permettent une représentation grossière des états affectifs, généralement associées aux six expressions universelles. La modélisation dimensionnelle consiste à représenter les expressions faciales en fonction de dimensions affectives. Dans ce cas, les observations sont traitées afin de définir les valeurs de ces dimensions. L'ensemble de ces dimensions sont alors mises en correspondance afin de caractériser l'état affectif observé par l'expression faciale analysée.

Le modèle catégoriel est plutôt utilisé en informatique et le modèle dimensionnel en psychologie. Sur un plan général, la modélisation catégorielle représente en partie seulement la modélisation dimensionnelle. elle condense l'information par le regroupement d'individus présentant les mêmes caractéristiques. Le sujet appartient à une catégorie

ou classe spécifique, dans une catégorisation limitée. Alors que la modélisation dimensionnelle permet de prendre en considération un nombre infini de classes définies par plusieurs dimensions telles que la valence et l'arousal.

Quel que soit le modèle employé, l'analyse des expressions repose essentiellement sur des observations visuelles extraites du visage. Dans la section suivante, nous présentons les différents défis s'inscrivant dans le processus de reconnaissance des expressions faciales.

3.3 Les défis de la reconnaissance des expressions faciales

La majorité des systèmes d'analyse faciale sont conçus pour analyser les expressions faciales où les conditions d'acquisition sont maîtrisées (pose du visage fixe et frontale à la caméra, luminosité constante) et les expressions faciales sont volontaires (forte intensité du mouvement facial). Assumant que le visage ne subit aucun changement autre que celui induit par les expressions, cela assure que les caractéristiques visuelles sont parfaitement exploitables. Cependant, ces conditions ne reflètent pas les conditions d'acquisition naturelles, où la personne est libre de ces mouvements (expressions spontanées, mouvement de la tête) et où des occultations (ombre, objet) font leur apparition.

Dans des conditions d'acquisition naturelles, la majorité des systèmes actuels ne parviennent pas à caractériser correctement les expressions faciales. En effet, en présence de variations de pose et de larges déplacements, le visage a tendance à subir des déformations (occultation partielle du visage, effet de floue). Dans ces conditions, les caractéristiques extraites sur le visage ne sont plus garanties car elles peuvent subir des déformations, ce qui réduit la qualité des données et peut avoir des répercussions sur des tâches sous-jacentes, telle que la reconnaissance des expressions faciales.

En complément des mouvements du visage, les intensités des mouvements faciaux sont amenées à varier et nécessitent que l'on puisse traiter à la fois les petites et grandes intensités. La difficulté réside dans le fait que les caractéristiques de mouvement sont très différentes entre les expressions de faibles et fortes intensités, ce qui demande en général de passer par des techniques de prétraitements adaptés au mouvement analysé.

La difficulté de l'analyse est proportionnellement renforcée en fonction des mouvements de la tête. Plus le mouvement de la tête sera important, plus les mouvements faciaux de petites intensités liés aux expressions seront englobés avec le mouvement global de la tête.

Dans la suite de cette section, nous illustrons les deux défis mentionnés, qui sont les problèmes liés aux variations d'intensité des expressions et les problèmes liés à la variation de mouvement du visage.

3.3.1 La variation de l'intensité des expressions

Les expressions faciales sont composées d'un ensemble de mouvement de faible et forte intensité. Plus spécifiquement, Ekman [30] divise les expressions faciales en deux catégories : les macro et les micro expressions. Les macro expressions sont caractérisées par des expressions faciales volontaires, impliquant une majorité des muscles faciaux. Les micro expressions sont des expressions faciales involontaires, souvent associées à des émotions que l'on désire cacher.

La différence entre les macro et micro expressions réside essentiellement dans l'intensité du mouvement facial et de la durée des expressions faciales. La Figure 3.3 illustre une séquence d'images représentant une macro et une micro expression. La durée d'activation des expressions faciales est considérée comme la mesure la plus discriminante pour dissocier les macro et micro expressions. Les macro expressions apparaissent généralement lorsque l'on souhaite communiquer nos émotions. Elles durent en moyenne entre 0.5 à 4 secondes et elles sont clairement visibles sur le visage [30]. Les micro expressions sont plus rapides et subtiles, au point d'être difficilement perceptibles par l'oeil humain. Des travaux montrent que la durée totale d'une micro expression est comprise entre 170 et 500 millisecondes [117]. Plus spécifiquement, Yan et al. [117] montrent que la phase d'activation de l'expression (onset) est un bon indicateur pour caractériser une micro expression, qui au minimum 65 millisecondes et au maximum dure 260 millisecondes.

Les micro expressions ne sont pas caractérisées uniquement par leur courte durée d'activation mais aussi par leur faible intensité [89, 116]. La plupart des micro expressions n'atteignent jamais la plus petite intensité (niveau A) du système FACS proposé

par Ekman [30]. Les intensités des micro expressions sont vraiment faibles et se caractérisent par des micro mouvements des muscles faciaux et par conséquent de très faibles changements de texture. Ces spécificités requièrent des protocoles adaptés et des caméras hautes fréquences pour pouvoir être perceptibles. Les nouvelles bases d'apprentissage pour l'analyse des micro expressions [65] capturent les expressions en utilisant des caméras hautes fréquences (100 - 200 fps). Cela permet d'acquérir matériellement des déformations faciales observables. Contrairement aux micro expressions, 25 à 30 fps sont largement suffisants pour détecter les macro expressions.

Concernant les macro expressions, la propagation du mouvement couvre généralement une majeure partie du visage contrairement aux micro expressions, où seule de petites régions spécifiques sont activées. Porter et al. [89] montrent que l'activation individuelle d'une micro expression ne peut se situer simultanément dans la partie haute et basse du visage. Le mouvement étant moins intense, la propagation de mouvement a tendance à être discontinue dans l'ensemble du visage. Des méthodes de magnification du mouvement sont parfois employées pour amplifier l'intensité du mouvement en augmentant artificiellement la fréquence d'une séquence d'images. La deuxième ligne de la Figure 3.3 illustre les séquences d'images de la macro et de la micro expression magnifiée avec un facteur de 10. Bien que ces méthodes permettent de visualiser la déformation faciale liée à l'expression, elle a tendance à déformer la géométrie faciale, notamment en présence de macro expression.

FIGURE 3.3 – Différence visuelle de l'intensite de mouvement d'une macro (CK+) et micro (SMIC) expression.

Dans la section suivante, nous discutons de l'impact des variations de mouvement du visage pour la reconnaissance des expressions faciales.

3.3.2 La variation de mouvement du visage

En situation d'interaction naturelle (liberté de mouvement), il arrive souvent que le visage soit soumis à des rotations 3D (rotations hors plan). Souvent, les systèmes proposés pour analyser les expressions faciales ne sont pas invariants aux transformations géométriques 3D. En effet, la majorité des systèmes s'appuient sur des modèles de segmentation par grille pour caractériser les expressions faciales [124, 120]. Bien que ces systèmes obtiennent de bonnes performances dans des conditions statiques (voir Tableau 3.2), ils ne sont pas adaptés aux mouvements de translations, rotations et changements d'échelle. La Figure 3.4 illustre l'utilisation d'une modélisation faciale par grille pour l'analyse d'expressions faciales. Le visage est divisé en plusieurs régions, où dans chacune d'elles des vecteurs caractéristiques y sont extraits. Puis, les vecteurs sont concaténés dans un unique vecteur qui caractérise l'expression faciale. En présence de variations de pose, on observe un mauvais alignement du visage, qui implique une mauvaise correspondance des composants du visage en chaque région. Ce problème d'alignement entraine une mauvaise estimation de l'expression, due au fait que la distance calculée entre les vecteurs extraits des deux visages, est trop importante.

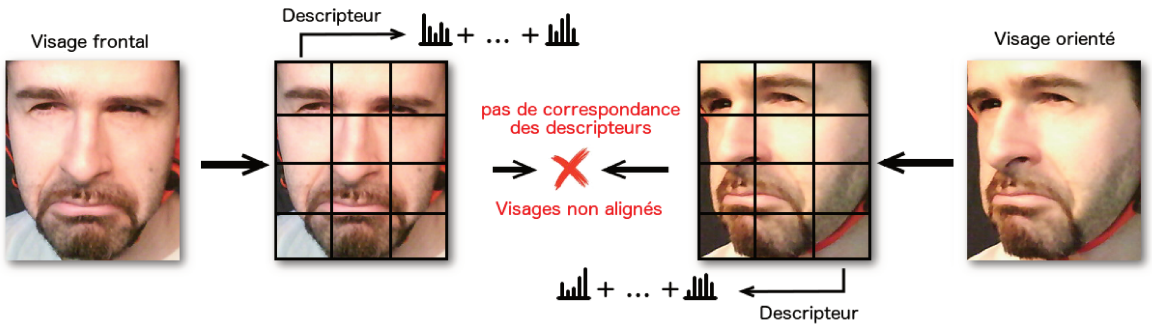


FIGURE 3.4 – Exemple de mauvais alignement facial en présence de variation de pose (SNaP-2DFe).

Les larges déplacements correspondent à des mouvements de forte intensité entre deux images successives. Dans ce contexte, les descripteurs de mouvement comme les flux optiques ou les LBP-TOP sont mal adaptés. Cela s'explique par le fait que les mouve-

ments induits par les expressions faciales vont être confondus avec le mouvement de la tête, comme illustré dans la Figure 3.5. La deuxième ligne de la Figure 3.5 représente le flux optique extrait sur le visage statique (sans mouvement de la tête). La troisième ligne de la Figure 3.5 correspond à la même séquence avec cette fois, la présence d'un mouvement de la tête. On remarque dans ce cas, que le mouvement de la tête est plus intense que celui de l'expression faciale. Cela a pour conséquence d'englober le mouvement induit par les expressions, ce qui ne permet plus de caractériser ce dernier mouvement indépendamment. L'impact du mouvement de la tête est directement proportionnel à l'intensité du déplacement. Plus le déplacement est important, plus il est difficile d'extraire le mouvement lié aux expressions faciales.

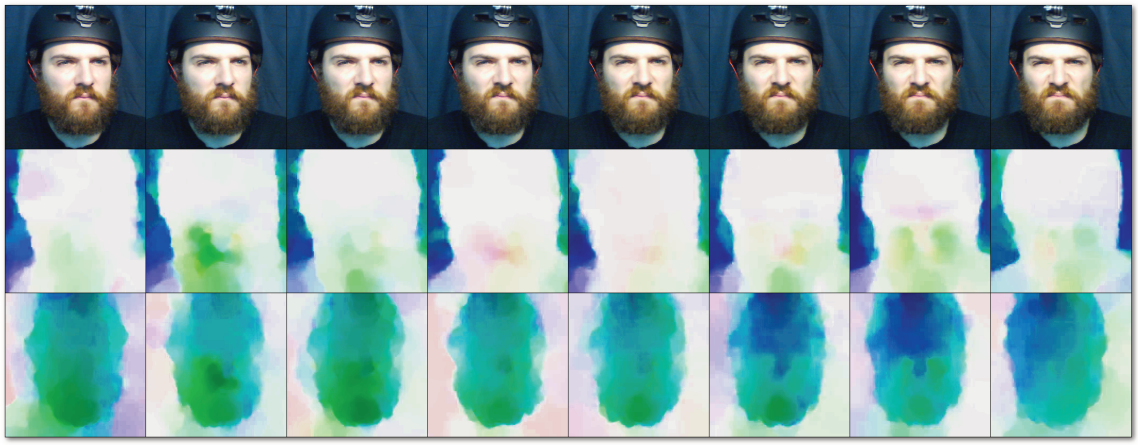


FIGURE 3.5 – Perte des mouvement induits par les expressions faciales dû à la présence de mouvements de la tête (SNaP-2DFe).

En complément des problèmes cités précèdemment, les larges déplacements produisent un effet de flou sur les visages. Le bruit généré induit des déformations au sein du visage, ce qui se traduit par une détérioration de la qualité de l'image [41]. Plus spécifiquement, au niveau des contours au sein du visage. Les larges déplacements se caractérisent par un déplacement de la tête qui peut-être corrigé avec les mêmes méthodes de normalisation employées pour corriger la pose. Cependant, étant donné que les contours ont été détectés de manière moins précise, cela ne garantit plus la localisation exacte des landmarks, comme l'illustre la Figure 3.6. La mauvaise estimation des landmarks impacte directement sur la précision de la détection du visage.

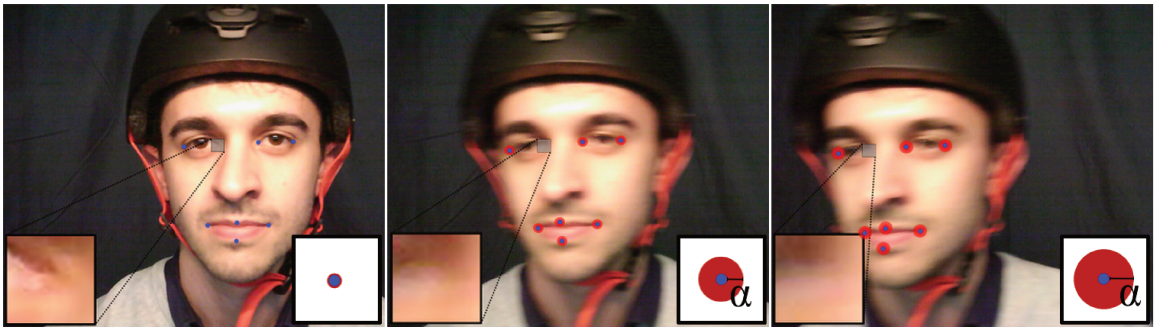


FIGURE 3.6 – Mauvaise estimation de la localisation d'un landmark due à l'effet de flou occasionné par un large déplacement. α correspond au niveau d'incertitude de la localisation du landmark, qui croît progressivement en fonction de la qualité de l'image (SNaP-2DFe).

3.3.3 Synthèse

Les approches de l'état de l'art obtiennent de bonnes performances sur des collections d'images contrôlées. Généralement, elles contiennent des adaptations pour augmenter leur robustesse dans des conditions proches de l'interaction naturelle (liberté du mouvement de la tête, large palette d'intensité des expressions).

Afin de pouvoir améliorer la robustesse des systèmes d'analyse d'expressions faciales en situation d'interaction naturelle, il est nécessaire de pouvoir travailler sur des données contenant des visages avec des poses variées et en présence de larges déplacements. Dans la section suivante, nous discutons de la représentativité de ces défis dans les bases d'apprentissages.

3.4 Les bases d'apprentissage

Les bases d'apprentissage jouent un rôle important pour la reconnaissance des expressions faciales. Construites sur un ensemble d'images et/ou de vidéos, elles mettent à disposition de nombreuses données permettant d'entraîner et d'améliorer la robustesse des algorithmes. Souvent élaborées pour répondre à un besoin, à un contexte d'acquisition spécifique (vidéo-surveillance, visio-conférence, ...), les bases de données tendent à adapter les algorithmes aux contraintes induites par le développement de nouvelles technologies. Ces contraintes sont principalement liées aux nouvelles méthodes de captation

(caméra thermique, caméra de profondeur, stéréovision, support mobile) ou aux conditions d'analyses (variations de pose, de luminosité, occultations), qui permettent de fournir des données proches des conditions d'acquisition naturelle.

Dans la suite de cette section, nous présentons un historique de l'évolution des bases d'apprentissage, où nous expliquons les différents défis auxquels elles ont permis de répondre au cours de ces dernières années. Puis, nous passons en revue les principales bases d'apprentissage, en discutant des caractéristiques qui les différencient les unes des autres. Nous finissons par une synthèse permettant d'identifier plus facilement les bases d'apprentissage à utiliser en fonction d'un besoin spécifique.

3.4.1 L'évolution des données

Les bases d'apprentissage ont été initialement élaborées pour caractériser les expressions faciales dans des conditions d'acquisition contrôlées (luminosité homogène, pose fixe et frontale à la caméra, expression exagérée). Ces bases, telle que CK+ [75], garantissent que les informations extraites sont non bruitées et permettent ainsi d'analyser les performances des algorithmes dans d'excellentes conditions. À ce jour, de nouvelles bases d'apprentissage, conçues sur le même protocole, sont proposées pour fournir de nouvelles données, en s'appuyant sur de nouveaux capteurs (haute résolution, fréquence d'acquisition, information 3D, domaine infrarouge).

Bien que ces bases d'apprentissage sont majoritairement utilisées dans la communauté, elles ne reflètent pas les conditions réelles d'acquisition, où l'interaction est naturelle (variations de pose, de luminosité, présence d'occultations, ...), et où les expressions sont spontanées. Suite à cela, les bases d'apprentissages se sont succédées, fournissant des scénarios de plus en plus élaborés, permettant ainsi d'obtenir des données de plus en plus proches d'un contexte naturel d'interaction, comme illustré dans la Figure 3.7.

Des bases telles que DISFA [79], se focalisent essentiellement sur l'analyse des expressions spontanées (vitesse et intensité des expressions variables), dans un contexte d'acquisition contrôlé. Ces données permettent d'analyser les performances des algorithmes, pour la caractérisation des déformations subtiles du visage. D'autres bases comme GEMEP [5], mettent en évidence des problèmes liés aux mouvements du visage durant l'in-

teraction sociale. Conçu sur des données où les expressions sont actées, exagérées, l'objectif est d'analyser la robustesse des algorithmes en présence d'occultations partielles du visage, dues à des variations de poses et de larges déplacements.

L'élaboration des nouvelles bases d'apprentissage, telle que RECOLA [91], s'appuie sur l'évolution des précédentes bases pour fournir des données de plus en plus réalistes. Bien que ces bases soient de plus en plus élaborées, l'analyse de leurs données devient de plus en plus complexe. Cette complexité concerne la caractérisation des expressions faciales mais aussi la conception de scénarios d'acquisition. En effet, les scénarios doivent garantir que l'interaction est naturelle et que les données extraites correspondent bien au besoin ciblé. La richesse de ces informations augmente considérablement le temps de conception, la difficulté d'annotation et le volume des données.

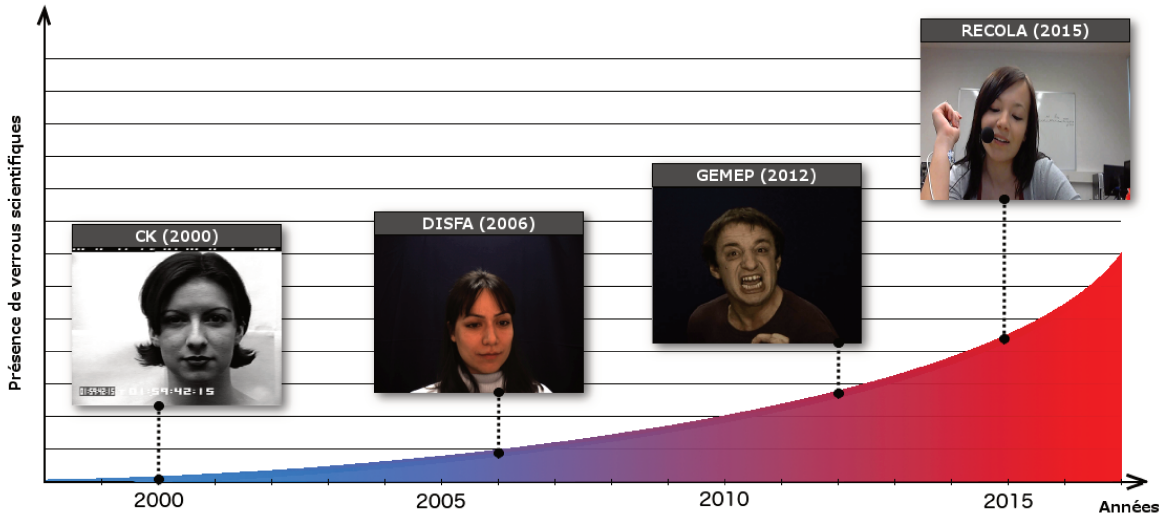


FIGURE 3.7 – Evolution des bases d'apprentissage en fonction de la difficulté d'analyse.

L'ensemble des bases de données permet de concevoir des algorithmes de plus en plus robustes aux situations d'acquisition naturelle. Bien que les récentes bases mettent en évidence les différentes problématiques rencontrées dans un cas d'usage réel, les bases telles que CK+, DISFA et GEMEP restent majoritairement utilisées dans la communauté. En effet, ces bases permettent d'analyser les performances des algorithmes sur des problématiques bien spécifiques, et d'augmenter progressivement la robustesse de ces derniers. L'idée étant, qu'un algorithme ayant de bonnes performances sur l'ensemble des

bases d'apprentissage, devrait bien se généraliser en situation naturelle, lorsque l'ensemble des problèmes seront réglés lors des phases de pré-traitement.

Chaque base de données met à disposition un ensemble d'annotations permettant de renforcer la robustesse des algorithmes, en respectant un protocole d'acquisition spécifique. Bien qu'il existe un grand nombre de bases conçues sur des scénarios similaires, certaines bases sont plus utilisées que d'autres. Dans la section suivante, nous discutons des critères de comparaison entre ces différentes bases d'apprentissage.

3.4.2 Comparatif des bases d'apprentissage

La conception d'une base d'apprentissage est une tâche complexe, qui demande une grande réflexion dans son élaboration. La création d'une base d'apprentissage destinée à l'analyse des expressions faciales repose sur trois critères principaux : la définition du système d'acquisition, la définition des verrous scientifiques et la caractérisation des états affectifs.

- La définition du système d'acquisition permet de déterminer le type des données (image, vidéo, son, réaction électrodermale), le type et le nombre de capteurs (système mono et multi modale) et les conditions d'acquisition (in-door, out-door, contexte social). L'ensemble de ces informations permet de fournir un large panel de données, pour s'adapter à un grand nombre de scénarios d'usage.

- La définition des problèmes/défis liés à l'acquisition est un critère important car cela permet de définir les différentes situations auxquelles le processus de caractérisation d'expressions faciales doit pouvoir s'adapter (occultation du visage, ombre). Certaines bases d'apprentissage mettent en évidence plusieurs problèmes rencontrés dans des situations d'acquisition réelles, où les expressions sont spontanées et la pose est libre. Des bases telles que RECOLA, SEMAINE [80] répondent à ces conditions en proposant d'étudier les problèmes de variations de pose (VP), de larges déplacements (LD), et de variations lumineuses. D'autres bases d'apprentissage fournissent des données permettant de traiter l'invariance à l'intensité des expressions. On distingue principalement deux catégories d'expression : les micro (faible intensité) et les macro (forte intensité) expressions. Les bases telles que SMIC et CASME II s'appuient sur les mêmes

principes de captation que les bases d'aprentissage classiques telle que CK+, mais se distinguent des autres bases en fournissant des données permettant d'analyser des expressions subtiles.

- La caractérisation des émotions permet d'identifier la manière dont les états affectifs sont annotés. Suite aux travaux d'Ekman [30], certaines émotions de base sont universellement reconnues (i.e. joie, peur, tristesse, surprise, colère, dégoût) (6+N). Les bases d'apprentissage reposant sur ce système d'annotation, associent chaque séquence à une expression. En complément des expressions faciales, certaines bases telle que MMI [87] contiennent une annotation continue sur l'enclenchement des muscles faciaux (AUs), basée sur le système d'action faciale (FACS) proposé par Ekman. Une alternative à cette représentation est l'utilisation de dimensions émotionnelles : "agréable ou non agréable" (Valence), "réveil ou soumission" (Arousal). Un scénario d'usage est élaboré en fonction de la manière dont les expressions sont annotées.

TABLEAU 3.1 – Bases d'apprentissage pour la caractérisation des expressions faciales.

	Système d'acquisition		Caractérisation des défis proposés				Caractérisation des émotions		
	Sequence	Capteur	Position	VP	LD	Lumière	Scenario	Emotion	AUs
CK+	10-60 img/seq	1 caméra	Fixe	-	✻	✻	Reproduction	6+N	✗
BU-4DFe	100 img/seq	2 caméras	Fixe	-	✻	✻	Reproduction	6+N	✗
Oulu-CASIA	20-50 img/seq	3 caméras + IR	Fixe	-	✻	✻✻✻	Reproduction	6+N	✗
SMIC	10-50 img/seq	2 caméras + IR	Fixe	-	-	✻✻	Reproduction	3	✗
CASME II	50-200 mig/seq	1 caméras	Fixe	-	-	✻	Reproduction	5	✓
Multi-Pie	20 img/seq	15 caméras	Fixe	-	✻	✻✻✻	Reproduction	6+N	✗
MMI	40-520 img/seq	1 caméra	Fixe	✻	✻	✻	Reproduction	6+N	✓
DISFA	4844 img/seq	2 caméras	Fixe	✻	✻✻	✻	Regarde des vidéo	-	✓
GEMEP	100 img/seq	1 caméra	Libre	✻✻	✻✻✻	✻	Reproduction	6+N	✓
SEMAINE	2K-25K img/seq	1 caméra	Libre	✻✻	✻✻✻	✻	Communication	A/V	✗
RECOLA	7500 img/seq	1 cam + ECG/RED	Libre	✻✻✻	✻✻✻	✻✻	Communication	A/V	✗

Les bases de données couramment utilisées dans la littérature sont recensées dans le Tableau 3.1. Le tableau est structuré en trois colonnes, représentant chacune l'un des cri-

tères cités précédemment. Pour chaque défi abordé, un indicateur compris entre une et trois étoiles (❋) permet de quantifier la difficulté contenue dans les données. Outre les différences proposées au niveau des défis, il est important de constater les différences entre les scénarios élaborés pour éliciter des expressions faciales et l'impact de ces derniers sur les caractéristiques globales des bases d'apprentissage.

La majorité des bases d'apprentissage repose sur un scénario de type 'reproduction" qui consiste à simuler des patrons de mouvement fidèles à des expressions simples telle que la joie, la peur, la tristesse... Dans ce cas, chaque séquence vidéo caractérise une expression, d'où le fait que la durée de séquence est courte. La base DISFA propose un autre scénario, où le sujet regarde un ensemble de collections de vidéos, ayant pour objectif de provoquer une expression particulière. Malheureusement, les vidéos sélectionnées n'ont pas toujours eu l'effet escompté sur les différents sujets, où certaines personnes ont ressenti des émotions différentes à la vue de certaines vidéos, ce qui a eu un impact sur l'équilibre des annotations. Les récentes bases comme RECOLA et SEMAINE, s'appuient sur un scénario de communication, où deux interlocuteurs (homme/homme, homme/agent virtuel) échangent sur un thème précis. Contrairement aux autres scénarios, ce scénario permet d'enregistrer des données dans un contexte d'interaction naturelle et où les séquences sont plus longues. Il est important de constater un changement au niveau de la caractérisation des émotions, où l'utilisation de dimensions émotionnelles semble plus adaptée aux données analysées.

Bien que les bases d'apprentissage tendent à fournir des données proches d'un système d'acquisition naturelle, peu de bases d'apprentissage prennent en compte toutes les contraintes liées à ces systèmes. De telles bases d'apprentissage demandent un nombre conséquent de données annotées, dans des conditions parfois difficilement analysables. Dans ces conditions, l'annotation manuelle est obligatoire, ce qui prend beaucoup de temps à concevoir.

3.4.3 Synthèse

Dans cette section, nous avons présenté les différentes bases d'apprentissage, en passant en revue leurs évolutions et leurs conceptions. L'objectif des bases d'apprentissage étant de fournir des données permettant de caractériser les expressions dans un cas d'usage

spécifique et en présence de défis, un nombre important de bases d'apprentissage ont été proposées. Souvent élaborées dans un contexte laboratoire, où les données sont non bruitées (illumination et scène homogènes, pose fixe et frontale), les bases d'apprentissage ont progressivement évolué, mettant à disposition de nouveaux défis, afin de rendre les algorithmes plus robustes dans des contextes d'interactions naturelles.

Bien qu'il existe un grand nombre de bases d'apprentissage, chaque base se distingue des autres bases, que cela concerne le système d'acquisition, les défis proposés ou la caractérisation des émotions. Dans le domaine des expressions faciales, la conception d'une base d'apprentissage repose essentiellement sur son scénario. Le scénario a pour objectif de définir le contexte social dans lequel les expressions sont capturées. Cela a une répercussion importante sur les données (taille des séquences, type de capteur, difficultés des défis) et sur la caractérisation des émotions (expressions universelles, dimension affective).

Il est important de noter qu'il n'existe pas de base d'apprentissage universelle. C'est le regroupement de toutes ces données qui tendent à améliorer les algorithmes vers des cas d'usage réels (vidéo-surveillance, visio-conférence,...). Le choix de la base d'apprentissage dépend principalement des objectifs fixés. Si l'objectif est de pouvoir quantifier les performances des algorithmes à caractériser une expression, alors le choix se portera sur des bases d'apprentissage où les données sont "saines". Si l'objectif est de pouvoir analyser la robustesse d'un algorithme face à un défi en particulier, comme la variation de pose ou d'illumination, alors il faut travailler sur des bases d'apprentissage fournissant ce type de défi. L'objectif final étant d'aboutir à la conception d'un algorithme capable de reproduire les capacités d'analyse et d'interprétation de la vision humaine en s'appuyant sur des technologies emphatiques.

Dans la section suivante, nous présentons comment les approches proposées dans la littérature utilisent les bases d'apprentissage afin d'améliorer leur robustesse en présence de variations de pose et de larges déplacements.

3.5 Invariance à l'intensité des expressions faciales

Dans cette section nous présentons les approches les plus significatives de la littérature pour la reconnaissance des macro et micro expressions. Nous commençons par discuter de différentes techniques permettant de caractériser les macro et les micro expressions. Puis, nous faisons une synthèse globale des approches couramment utilisées pour analyser les expressions faciales afin de mettre en exergue les besoins pour construire une approche la plus adaptée pour caractériser à la fois les macro et les micro mouvements faciaux.

3.5.1 Macro expression

Les macro expressions sont caractérisées par d'importants mouvements des muscles faciaux. Au vu des différentes approches permettant d'analyser les déformations faciales, on distingue plusieurs techniques. Ces techniques s'appuient soit sur l'analyse de la texture, soit sur la géométrie, ou bien sur les deux.

Des approches statiques s'appuyant sur l'analyse de la texture ont été proposées comme les LBP [86] et les HOG [56]. Ces techniques obtiennent de bonnes performances pour caractériser les macro expressions. Récemment, plusieurs approches basées sur les réseaux de neurones convolutionnels (CNN) ont prouvé leur efficacité [72, 28, 82]. Comme pour les autres méthodes statiques, l'apprentissage consiste à associer une expression à un ensemble de caractéristiques spatiales du visage. Ces caractéristiques sont extraites lorsque l'apex de l'expression est atteint (i.e. correspond à l'instant où l'intensité de l'expression est la plus élevée durant une séquence). En s'appuyant uniquement sur l'information spatiale, les approches statiques LBP, HOG et CNN n'exploitent pas la dynamicité des expressions faciales dans le processus de reconnaissance, ce qui limite leurs performances.

Des expériences psychologiques proposées par Bassili [7] ont prouvé que les expressions faciales sont mieux reconnues lorsque l'expression faciale est analysée dynamiquement, ce qui revient à analyser une séquence d'images. Une extension dynamique des LBP, appelée Local Binary Pattern on Three Orthogonal Plans (LBP-TOP), a été proposée par Zhao et al. [124]. D'autres approches basées sur la même méthode ont été proposées, comme les LGBP-TOP et les HOG-TOP. Les approches dynamiques permettent d'obtenir

de meilleurs résultats que leurs homologues statiques, ce qui montre bien l'importance de considérer l'information temporelle dans l'analyse des macro expressions. Au vu des résultats obtenus par les approches dynamiques récentes, il y a une émergence des approches basées sur les méthodes de flux optique [67, 36]. Les méthodes de flux optique denses ont l'avantage de détecter des mouvements subtils sur le visage, ce qui permet de discriminer plus facilement deux expressions similaires. Dans les récentes approches de deep learning [13, 122], un réseau de neurones récurrent (RNN) est combiné avec un CNN pour encoder l'information temporelle. Ces architectures permettent d'améliorer significativement les performances des CNN conventionnels.

La géométrie faciale est souvent utilisée pour analyser les macro expressions. Principalement basées sur la position des landmarks, les approches géométriques visent à analyser la forme du modèle de points caractérisant le visage. Les landmarks sont utilisés de différentes manières afin d'extraire la forme du visage et le mouvement des muscles faciaux [77]. De nombreux travaux associent des patrons de forme directement construits à partir des landmarks pour chaque expression [37, 54]. Ces patrons peuvent être analysés dynamiquement pour prendre en considération l'aspect temporel. Ghimire et al. [37] montrent que l'information temporelle permet de réduire significativement le taux de faux positifs entre deux expressions proches. En effet, dans ce cas, le modèle géométrique seul ne suffit pas, notamment pour les expressions de peur et de surprise, où les déformations faciales résultant de l'expression sont relativement identiques. Saeed et Al. [95] montrent qu'il n'est pas nécessaire d'utiliser l'ensemble des landmarks afin d'analyser les expressions faciales. Seul huit landmarks situés dans des régions saillantes du visage (sourcil, yeux, nez, bouche) permettent d'obtenir des performances similaires aux approches utilisant les modèles de 68 points.

Des approches hybrides sont proposées dans la littérature. Ces approches consistent à combiner des informations de texture et de géométrie pour caractériser une expression faciale. Comme le suggère Kotsia [60], la combinaison des caractéristiques de la géométrie et de l'apparence faciale sont complémentaires pour reconnaître une expression faciale. Han et al. [38] utilisent un AAM pour créer des régions déformables dans le temps au sein du visage. Chaque région est alors représentée par une forme géométrique dans laquelle ils caractérisent les pixels de ces régions par des LBP. Le descripteur permet d'obtenir de

meilleures performances que les deux descripteurs seuls. Jaiswal et al. [48] combinent des informations dynamiques d'apparence et de forme dans un réseau de neurones. Ils montrent que la combinaison des deux informations permet au réseau de converger plus facilement. Plusieurs approches de deep learning [54, 122] combinent les informations géométriques du visage afin de renforcer l'apprentissage basé sur l'apparence.

L'ensemble de ces approches a été évalué sur des collections d'images où les expressions faciales sont actées et obtiennent de très bonnes performances ($\simeq$ 90-95%). L'information temporelle permet d'améliorer significativement les performances des approches, car elle permet de mieux distinguer des expressions similaires, par exemple la surprise et la peur, où les deux expressions sont caractérisées par les mêmes AUs mais où l'activation des muscles faciaux est différente (ordre, intensité).

Dans la section suivante, nous discutons des méthodes proposées pour caractériser les micro expressions. Plus spécifiquement, nous vérifions si l'analyse des micro expressions s'appuie sur les mêmes méthodes que celles utilisées pour caractériser les macro expressions.

3.5.2 Micro expression

Au vu des performances obtenues dans l'analyse des macro expressions, de nombreux travaux ont appliqué les mêmes approches que pour analyser les micro expressions. Liu et al. [71] appliquent directement les LBP et les HOG pour caractériser des micro expressions. Les résultats obtenus montrent que les approches optimisées pour la macro expression ne semblent pas adaptées ($\simeq$ 50-53%). En effet, les caractéristiques des micro expressions (faible intensité de mouvement) sont très subtiles pour pouvoir appliquer directement des méthodes conçues pour analyser des mouvements de forte intensité. Les mouvements faciaux de faible intensité sont difficilement perceptibles par ces approches car le mouvement extrait se confond entre l'expression et le bruit (discontinuité de mouvement, mouvement de la tête, bruit du capteur, variation lumineuse). La même constat est observé concernant les approches de deep learning [88].

Comme pour les macro expressions, Liu et al. [64] montrent que les micro expressions sont plus facilement caractérisables lorsque l'information temporelle est prise en

compte. De ce fait, plusieurs travaux ont adapté les LBP-TOP pour l'analyse des micro expressions. Liong et al. [69] étendent les LBP-TOP en utilisant l'information de contrainte optique comme fonction de pondération pour trouver de plus petits mouvements. Wang et al. [109] proposent une représentation plus compacte et plus légère en réduisant au minimum la redondance des informations contenues dans le descripteur. Huang et al. [43] étendent les LBP-TOP en appliquant une projection intégrale améliorée pour combiner à la fois l'information de texture et de géométrie.

Bien que la majorité des approches adaptées pour la micro expression sont à base de LBP-TOP, d'autres approches alternatives ont été proposé. Huang et al. [44] proposent une extension dynamique des LQP appelée STCLQP (spatio-temporal completed local quantized pattern). La raison pour laquelle cette méthode obtient de bonnes performances sur la micro expression et qu'elle permet d'extraire conjointement l'information sur l'orientation et sur la magnitude (intensité du mouvement). Li et al. [64] utilisent une méthode d'interpolation temporelle et de magnification du mouvement pour extraire les mouvements de faible intensité. Ils prouvent qu'au-delà d'une certaine longueur d'interpolation pour une séquence, l'interpolation n'améliore pas les performances car la micro expression n'est plus perceptible. Dans ce cas, une interpolation sur 10 frames est largement suffisante. Récemment, Liu et al. [71] ont proposé un descripteur basé sur le flux optique appelé MDMO (Main Directional Mean Optical-flow), pour conserver uniquement le mouvement pertinent à l'expression faciale (non bruité). Ils montrent que l'information de magnitude est plus discriminante que l'orientation lorsque l'on caractérise des micro expressions. Leur méthode obtient de meilleures performances que l'ensemble des autres approches proposées ($\simeq$ 68%). De récentes approches deep learning ont été proposées pour reconnaître les micro expressions [13, 57, 88]. Cependant, les performances observées restent basses au vu des performances obtenues par les autres approches.

En présence de micro expressions, les approches prenant en compte l'information temporelle obtiennent de meilleures performances. Notamment, l'utilisation des méthodes de flux optique semble bien adaptée car ces descripteurs permettent de détecter des déformations faciales plus subtiles. Les récents travaux sur le flux optique, montrent également que la magnitude (i.e. intensité du mouvement) est une information importante à prendre en considération lorsque l'on souhaite caractériser les micro expressions.

Dans la section suivante, nous faisons une synthèse des différentes approches utilisées pour analyser les macro et des micro expressions, afin d'identifier une méthode permettant de caractériser les deux types d'expressions simultanément.

3.5.3 Synthèse

Les caractéristiques de mouvement des macro et des micro expressions font qu'il est nécessaire d'adapter les systèmes d'analyse en fonction du type d'expression recherchée. Un tableau de synthèse des récents systèmes d'analyse de macro et micro expression est donné dans le Tableau 3.2.

Le Tableau 3.2 présente les différentes catégories des approches basées sur l'apparence, la géométrie et le mouvement (approches dynamiques) pour l'analyse des macro et micro expressions. L'objectif principal de cette table est de présenter les différences en matière de performances entre les macro et micro expressions, quand une même approche et un même modèle de segmentation faciale sont utilisés. Les résultats entre les macro et les micro expressions ne sont pas directement comparables, dû au fait que les bases de données utilisées entre les deux types d'expressions ne sont pas les mêmes. Cependant, nous les présentons ensemble afin de visualiser les tendances et identifier quelle approche semblent être la mieux adaptée à chaque type d'expression. Pour garantir une comparaison équitable, tous les systèmes cités dans la colonne correspondant aux macro expressions du Tableau 3.2 utilisent le classifieur SVM avec une validation croisée (10-fold), et tous les systèmes cités pour la micro expression utilisent une validation par LOSO (leave-one-subject-out).

Au vu des résultats du Tableau 3.2, les approches statiques, représentées ici par les LBP [101] et les HOG [56], obtiennent de faibles performances pour l'analyse des micro expressions [64]. La différence s'explique par le fait que ces méthodes ne sont pas adaptées pour détecter des mouvements subtils [64]. On remarque que les LBP-TOP obtiennent de meilleures performances que les LBP, aussi bien sur les macro [124] et les micro expressions [43], ce qui souligne l'importance de prendre en compte l'information temporelle. Cela s'explique car les méthodes temporelles permettent de mieux caractériser le mouvement, notamment en identifiant de subtiles déformations. Depuis, de nombreux chercheurs se sont focalisés sur les LBP-TOP, pour caractériser les micro expressions.

TABLEAU 3.2 – Synthèse des récents systèmes de caractérisation des macro et micro expressions (* données augmentées / deep learning)

Basé sur	Macro expression (CK+)		Micro expression (CASME II)	
Apparence	LBP [101] Grille	90.05%	LBP [64] Grille	55.87%
	PHOG [56] Régions saillantes	95.30%	HIGO [64] Grille	67.21% *magnifié*
	CNN [72] Globale	* 96.76%	CNN [88] Globale	* 47.30%
Géométrie	Gabor Jet [37] Landmarks	95.17%	/	/
	DTGN [54] Landmarks	* 92.35%	/	/
Mouvement	LBP-TOP [124] Grille	**96.26%**	DiSTLBP-IIP [43] Grille	64.78%
	Optical flow [2] Maillage	93.17%	MDMO [71] Maillage	**67.37%**
	CNN + AUs + LSTM [13] Globale	*** 98.62%**	CNN + LSTM [57] Globale	* 60.98%

Les approches géométriques, représentées par l'approche proposée par Ghimire et al. [37], obtiennent de bonnes performances pour caractériser des macro expressions. De manière générale, les algorithmes géométriques reposent sur la précision des landmarks et le suivi de ces points pour détecter les mouvements faciaux. Cependant, ces approches parviennent uniquement à détecter des mouvements faciaux d'une certaine intensité et ne sont pas adaptées aux micro expressions. Ceci implique qu'il n'existe pas à notre connaissance, de travaux s'appuyant uniquement sur la géométrie faciale, pour caractériser les micro expressions.

Concernant les approches basées sur la texture, les approches dynamiques obtiennent les meilleures performances dans l'analyse des expressions de faibles intensités. Plus particulièrement, les approches à base de flux optique semblent les mieux adaptées pour caractériser les micro expressions [71]. De plus, il est important de constater que le flux optique permet d'obtenir des résultats compétitifs concernant l'analyse des macro expressions [2]. Cependant, l'utilisation du flux optique a longuement été critiquée dû au fait que ce descripteur soit très sensible aux discontinuités de mouvement et aux changements lumineux. Les récents algorithmes de flux optique [90] sont de plus en plus ro-

bustes aux bruits. La majorité de ces algorithmes se basent sur des filtres complexes et sur des méthodes de lissage du mouvement afin de réduire les discontinuités. Ces algorithmes améliorent nettement la qualité du flux optique, cependant, les méthodes de lissage appliquées tendent parfois à induire de faux mouvements.

D'autres techniques consistent à amplifier artificiellement le mouvement. Ces techniques sont de plus en plus utilisées pour détecter les micro mouvements et ont prouvé leur efficacité pour analyser des micro expressions [64]. Li et al. [64] constatent un gain de près de 10% (de 57.09% à 67.21%) sur le taux de reconnaissance en amplifiant la fréquence des vidéos de leur base d'apprentissage. L'inconvénient majeur de ces techniques se révèle en présence de mouvements intenses. L'amplification du mouvement induit de fortes déformations faciales, ce qui réduit considérablement leurs performances en présence de macro expressions.

Concernant les approches de deep learning, nous observons un important contraste entre les performances obtenues pour les macro et les micro expressions. Les approches de deep learning obtiennent de très bonnes performances pour analyser les macro expressions et sont majoritairement supérieures aux autres approches de la littérature. Cependant, ces mêmes méthodes ne parviennent pas à obtenir de bonnes performances pour analyser les micro expressions. De plus, il est important de noter que les approches de deep learning nécessitent d'augmenter artificiellement les données initiales des bases d'apprentissage (rotation, flou, symétrie) afin d'obtenir suffisamment de données pour s'entraîner. De ce fait, il est difficile d'affirmer que ces approches soient plus performantes que les autres approches proposées dans la littérature, en sachant que les données utilisées ne sont pas identiques.

Le Tableau 3.2 montre que même si des tendances communes peuvent être trouvées pour caractériser les macro et les micro expressions, il n'existe pas un système unique d'analyse faciale permettant de caractériser les deux types d'expressions simultanément. Pour rappel, une expression faciale se caractérise par un ensemble de macro et micro mouvements. De ce fait, on peut émettre l'hypothèse qu'actuellement, les méthodes proposées ne garantissent pas d'extraire entièrement l'information contenue dans les expressions faciales.

Afin de distinguer correctement les performances de la caractérisation des macro et des micro expressions, il est important que seules les déformations induites par les expressions faciales soient observables. De ce fait, aucunes autres déformations, notamment par des mouvements de la tête, ne doivent apparaitre. Dans la section suivante, nous étudions l'impact des mouvements de la tête dans la caractérisation des expressions faciales, et des approches employées pour les réduire. Plus spécifiquement, nous nous intéressons aux variations de pose (rotations hors plan) et aux larges déplacements (déplacements rapides).

3.6 Invariance aux déplacements du visage

Dans cette section, nous concentrons notre analyse sur les problèmes liés aux mouvements du visage. Plus spécifiquement, nous présentons les méthodes proposées dans littérature pour réduire les larges déplacements (LD) et les variations de pose (VP).

3.6.1 Variations de pose (VP) et Larges déplacements (LD)

Afin d'obtenir une invariance aux transformations géométriques, une étape de normalisation géométrique est généralement appliquée dans les récents systèmes [66, 97]. La normalisation géométrique a pour objectif de trouver la meilleure transformation (ou déformation) qui réduit la distance entre deux visages. Cette transformation consiste à enlever le mouvement de la tête en amenant le visage dans un plan frontal, ce qui permet de garantir un alignement parfait des visages durant tout le processus d'analyse.

Indépendamment des variations de pose, d'autres défis comme la présence de larges déplacements de la tête viennent renforcer la difficulté de l'analyse des expressions faciales. Les méthodes pour corriger les larges déplacements sont généralement conçues pour corriger le mouvement de la caméra dans la scène. Ce mouvement peut être linéaire, en forme d'arc ou en rotation. La correction du mouvement de la caméra filtre le mouvement induit par le capteur pour conserver uniquement le mouvement propre à la scène. Bien que ces méthodes aient prouvé leur efficacité dans de nombreux domaines d'application, elles ne sont pas bien adaptées pour filtrer le mouvement global en présence de mouvements générés par les expressions faciales. Cela s'explique par le fait que ces méthodes fonctionnent mieux pour extraire le mouvement sur des objets rigides. Or, les

mouvements induits par les expressions faciales sont par nature complexes, dû au fait de l'élasticité de la peau, et contiennent de nombreuses variations de mouvement difficilement traitées par ces méthodes.

Les méthodes de normalisation géométriques 2D sont majoritairement employées dans la littérature pour corriger les invariances géométriques (translations, rotations, changements d'échelle) afin d'analyser les expressions faciales [23, 85]. De nombreux travaux utilisent une normalisation affine rigide qui consiste à trouver le meilleur alignement entre les landmarks de deux visages [39, 124]. L'inconvénient de cette approche est qu'elle ne prend pas en compte les déformations faciales induites par les expressions, ce qui peut engendrer d'importantes erreurs d'alignement. Les méthodes de normalisation affine non-rigide permettent de résoudre ce problème, en enlevant la contrainte géométrique du visage initial. Bien que l'alignement soit plus précis, cela peut induire de fortes déformations de la texture au sein du visage (effet d'étirement), notamment entre deux landmarks qui ont subi un important changement de position. Pour pallier ce problème, la normalisation affine peut-être appliquée par morceaux, c'est-à-dire localement à différentes régions du visage [98]. Cette méthode permet de réduire les déformations induites par les méthodes globales. En contrepartie, la normalisation affine par morceaux à tendance à supprimer les mouvements faciaux et de ce fait, induire des pertes d'information lorsque l'on s'intéresse à l'analyse des expressions faciales.

Bien que les systèmes d'analyse d'expressions faciales montrent que la normalisation 2D apporte un gain significatif des performances en présence du mouvement de la tête, les résultats obtenus restent encore relativement faibles en présence de rotations hors plan du visage. En effet, un système d'acquisition d'images ne fournit que la projection du visage observé sur un plan en deux dimensions. L'image ne permet donc d'exploiter que l'information résultant de la projection sur le plan 2D. En présence de rotation hors plan, la normalisation induit de fortes déformations faciales, ce qui se traduit généralement par un étirement de la texture au niveau de la région occultée. Comme illustré dans la Figure 3.8, plus la région occultée du visage est grande, plus la déformation faciale est importante.

FIGURE 3.8 – Déformation faciale provoquée par les méthodes de normalisation 2D (SNaP-2DFe).

Pour corriger les problèmes liés aux méthodes de normalisation 2D, les récents travaux de la littérature tendent à converger vers l'utilisation de solutions 3D. Comme l'illustre la Figure 3.9, l'utilisation des méthodes 3D consiste à modifier artificiellement les poses des visages apparaissant dans le plan 2D à l'aide d'un avatar 3D. Cette solution a l'avantage d'être plus robuste en présence de rotations hors plan et permet de conserver la géométrie du visage, ce qui garantit de préserver les expressions faciales. Zhu et al. [127] proposent de normaliser les visages en plaquant la texture du visage 2D sur un modèle facial 3D déformable. Le visage 2D est tout d'abord détecté et segmenté sous forme de maillage grâce aux landmarks. Le modèle 3D est ensuite orienté afin d'amener le visage dans une configuration frontale. Enfin, les régions non texturées correspondant aux régions occultées sont reconstruites par symétrie. Jeni et al. [50] proposent une approche permettant de reconstruire le visage 3D sur une séquence de visage. L'avantage de cette approche est que les parties occultées sont reconstruites à partir des informations contextuelles.

FIGURE 3.9 – Normalisation de visage par méthode de normalisation 3D (SNaP-2DFe).

Bien que ces méthodes permettent de corriger significativement les problèmes de pose, ces méthodes deviennent difficilement applicables en présence de rotations extrêmes (au-delà de 45°) [46]. Cela s'explique par le fait que l'information contenue dans la région faciale occultée est inconnue et qu'il est complexe de reconstruire les données perdues, qui contiennent à la fois des informations géométriques et de texture. L'inconvénient majeur des méthodes de normalisation 3D est que le nombre de données nécessaires à l'apprentissage est très volumineux, dû au fait que le modèle 3D doit être capable de s'adapter à la fois aux différentes variations de pose mais également aux déformations induites par les expressions faciales. Il faut également prendre en compte que ces méthodes demandent un temps de calcul conséquent pour transposer le visage 2D en 3D, indépendamment des occultations faciales à corriger.

3.6.2 Synthèse

La présence de variations de pose et de larges déplacements de la tête sont deux principaux défis pour la caractérisation des expressions faciales. En effet, de nombreuses contraintes comme l'occultation du visage, de mauvaises détections du visage viennent renforcer la difficulté de l'analyse et impactent fortement la précision des résultats. Les récents travaux tendent à converger vers l'utilisation de méthodes 3D afin de pallier ces contraintes.

Les méthodes expliquées précédemment ont été appliquées sur différentes bases d'apprentissage. Les résultats obtenus pas ces méthodes sont illustrés dans la Figure 3.10. Au vu des résultats obtenus sur les bases CK+ et MMI, nous pouvons voir que les méthodes donnent de bonnes performances. Cela est dû au fait que la présence de variations de pose et de larges déplacements est faible et que l'intensité des expressions faciales est importante (expressions actées).

Comme illustré dans la Figure 3.10, les méthodes de normalisation du visage basées sur des modèles 3D et sur des modèles d'alignements faciale 2D sont majoritairement utilisées pour reconnaître les expressions faciales en situation d'interaction naturelle. En dépit du fait que les approches proposées obtiennent de meilleures performances en utilisant ces méthodes de normalisation, les performances des systèmes restent toujours plus faible en comparaison de celles obtenues dans un contexte contrôlé. Plus les don-

nées fournies par les bases d'apprentissage sont complexes (proche des conditions d'acquisition naturelle), plus les performances diminues. C'est notamment le cas sur les bases telles que GEMEP, DISFA et SEMAINE, où les résultats obtenus sont relativement faibles car les données sont relativement complexes à analyser.

Afin de mieux visualiser les performances en fonction des méthodes de normalisation appliquées sur les différentes bases d'apprentissage, nous avons inclus deux graphiques en bas de la Figure 3.10. La Figure 3.10-A regroupe les méthodes évaluées en utilisant un taux de reconnaissance ou une moyenne du taux de reconnaissance. Quant à La Figure 3.10-B, elle regroupe les méthodes évaluées par une approche de validation croisée par personne. Chaque couleur du graphique est associée à un duo comprenant la méthode de normalisation appliquée et la base d'apprentissage utilisée.

References	Features	Registration	Databases		Performances
Zhao *et al.* [124]	LBP-TOP	Eyes	CK+		ar:95.2%
Happy *et al.* [39]	Salient LBP Patches	Eyes	CK+		ar:94.14%
Allaert *et al.* [2]	Optical flow & Geometry	Eyes	CK+		cr:95.34%
Koelstra *et al.* [58]	FFDs	Eyes	MMI		cr:94.3%
Jiang *et al.* [52]	LPQ-TOP	Shape	MMI		cr:94.7%
Rivera *et al.* [92]	DNG	Shape	MMI		ar:97.6%
Jiang *et al.* [53]	LPQ	N/A	GEMEP		cr:66%
Yang *et al.* [120]	LBP,LPQ	Shape	GEMEP		ar:84%
Sandbach *et al.* [97]	LBP	Eyes	DISFA		cc:0.342
Cruz *et al.* [23]	LPQ	Shape	SEMAINE		ar:55%
Nicolle *et al.* [85]	Appearance & Geometry	Shape	SEMAINE		cc:0.46
Chen *et al.* [18]	3D Facial Shape	3D Model	SEMAINE		cc:0.51

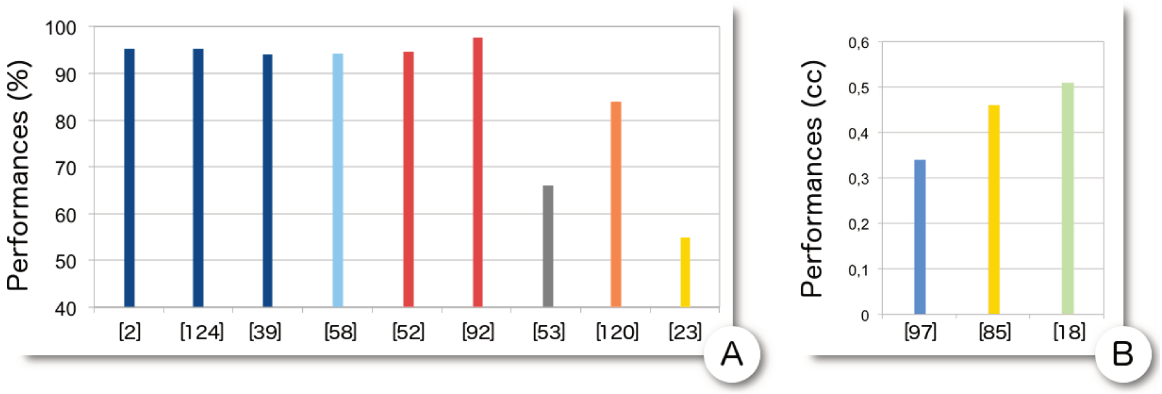


FIGURE 3.10 – Synthèse des résultats obtenus sur plusieurs bases d'apprentissage en utilisant différentes méthodes de normalisation.

Il est important de souligner que les méthodes de normalisation du visage sont employées principalement dans les systèmes de reconnaissance faciale où la présence de certains artefacts (déformation faciale, texture approximée) ou la perte de certaines informations est peu importante par rapport aux systèmes de reconnaissance d'expressions faciales. Or, lorsque l'on cherche à caractériser des expressions, la présence des déformations faciales provoque l'apparition de faux mouvements qui ne peuvent être supprimés sans impacter les mouvements pertinents liés aux expressions [19]. La normalisation a tendance à modifier la géométrie faciale et de ce fait change l'expression. Il est donc important de considérer les limitations des méthodes de normalisation lorsque l'on analyse les expressions faciales.

3.7 Conclusion

Ce chapitre présente un état de l'art portant sur l'analyse des expressions faciales. De cette synthèse, nous pouvons en conclure qu'il existe encore de nombreux défis qui ne sont que partiellement résolus. Dans ce chapitre, nous portons notre attention sur deux défis majeurs. Le premier défi est de pouvoir analyser des expressions faciales dans des conditions d'acquisition non contrôlées (variations de pose et larges déplacements). Le second est de pouvoir analyser des expressions avec des intensités très variables, où l'on distingue d'un côté les macro expressions (mouvements intenses) et de l'autre, les micro expressions (mouvements subtils).

Les conditions d'acquisitions soulèvent généralement les principaux défis dans le domaine de l'analyse des expressions faciales. Dans des conditions contrôlées, l'analyse est idéale du fait qu'aucune variation lumineuse, d'occultations ou de changements de pose modifient l'apparence du visage. Les systèmes actuels de la littérature obtiennent des résultats très satisfaisants dans ces conditions ($\simeq$ 95%). Les systèmes proposés sont généralement construits en utilisant des détecteurs de visages peu complexes, comme celui de Viola et Jones. Que les systèmes soient basés sur l'apparence, la géométrie ou bien les deux, la prise en compte de l'information temporelle permet d'améliorer significativement leur robustesse. Cela est principalement dû au fait que l'analyse s'affine au cours du temps grâce aux informations contextuelles. Actuellement, les systèmes basés sur des approches dynamiques denses telles que les LBP-TOP et les flux optiques obtiennent les

meilleures performances. En effet, ces méthodes ont l'avantage de détecter des mouvements subtils sur le visage, ce qui permet de discriminer plus facilement deux mouvements similaires. Dans des conditions non contrôlées, les variations de pose et la présence de larges déplacements renforcent la difficulté de l'analyse. Les systèmes proposés sont mal adaptés à ces conditions et ne parviennent pas à obtenir des résultats satisfaisants (variants de 25% à 70% en fonction des données). Bien que les approches dynamiques aient prouvé leur efficacité dans des conditions d'acquisition contrôlées, dans des conditions non contrôlées, leur efficacité dépendra fortement des données et de la robustesse des méthodes de normalisation.

Le deuxième défi majeur abordé dans ce chapitre, est la variation de l'intensité des expressions faciales. Les expressions faciales sont à la fois composées de macro et de micro expressions [30], l'une et l'autre portant des informations complémentaires. Par comparaison, les micro expressions sont des expressions faciales involontaires, souvent associées à des émotions que l'on désire cacher. Les macro expressions sont caractérisées par des expressions faciales volontaires, impliquant une majorité des muscles faciaux. Bien que les approches dynamiques obtiennent de meilleures performances pour caractériser à la fois les macro et les micro expressions, il est difficile de trouver une approche unique permettant d'analyser les différentes variations d'intensité de manière concomitante. En effet, les caractéristiques de mouvement des macro et des micro expressions font qu'il est nécessaire d'adapter les systèmes d'analyse en fonction du type d'expression. Généralement, il est plus difficile d'analyser les micro expressions où l'intensité du mouvement est très limitée, car la frontière entre l'information extraite par les descripteurs et le bruit résiduel causé par l'environnement de captation n'est pas évidente.

Il est important de noter que les approches dynamiques sont très sensibles au bruit, notamment à celui induit même par les légères variations de pose et les larges déplacements de la tête. Dans ce cas, il faut s'assurer que la caractérisation du visage ne subisse aucune déformation autre que celles induites par les expressions. Souvent adaptées pour la reconnaissance faciale, les méthodes de normalisation sont généralement mal adaptées pour analyser des expressions faciales, où il est important de pouvoir conserver avec exactitude les informations de texture et de géométrie faciale. Bien que les méthodes de normalisation 3D semblent être mieux adaptées pour normaliser les visages tout en

conservant les expressions, ces méthodes ne sont pas toujours parfaitement adaptées. En effet, les modèles 3D employés dans le processus d'apprentissage peuvent être soit en quantité absolument limitée (combinaison de pose et d'expressions faciales) ou trop complexe et volumineux à collecter en nombre suffisant.

Dans le chapitre suivant, nous proposons une approche commune pour l'analyse des macro et micro expressions en s'appuyant sur un filtrage local du mouvement dense au sein du visage.

Chapitre 4

Caractérisation du mouvement facial

« *Je sais calculer le mouvement*
des corps pesants,
mais pas la folie des foules. »

Isaac Newton

Sommaire

4.1 Introduction

Les descripteurs basés sur la caractérisation du mouvement dense (i.e. calcul du mouvement en chaque pixel de l'image) ont prouvé leur efficacité dans l'analyse des expressions faciales, et semblent mieux adaptés pour caractériser la dynamique de celles-ci. Bien que de nombreux processus d'analyse d'expressions faciales ont été proposés dans la littérature, il est difficile de trouver un processus aussi bien adapté pour caractériser des mouvements faciaux de faible et forte intensité. Cette difficulté est directement liée aux caractéristiques de mouvement de ces expressions. En présence de macro expressions, les déformations faciales induites par les muscles du visage sont facilement perceptibles. Cependant, en présence de micro expressions, où les intensités de mouvement sont très faibles, une attention particulière doit être apportée pour encoder les subtiles déformations.

Lors de l'acquisition des visages, l'apparition de bruit (i.e. discontinuités de mouvement) provenant de différents facteurs (illumination, bruit de capteur, occultations) renforce la difficulté de l'analyse des mouvements faciaux. En complément du bruit d'acquisition, l'analyse du visage est délicate car certaines déformations faciales occasionnent l'apparition ou la disparition de rides, ce qui provoquent des discontinuités de mouvement. Ceci exige d'adapter le processus de caractérisation du mouvement facial afin de renforcer la distinction entre le bruit et le mouvement induit par les expressions. Les discontinuités de mouvement complexifient l'analyse des mouvements et réduisent considérablement la performance du processus de reconnaissance.

Les récentes approches proposées dans la littérature exploitent directement l'information extraite par les approches de flux optique dense afin de caractériser les expressions faciales. Cependant, ces approches ne prennent pas en considération les spécificités des approches de flux optiques employées, qui ne sont généralement pas adaptées pour l'analyse des expressions faciales.

Pour faire face aux discontinuités, plusieurs approches de flux optique dense (i.e DeepFlow [111]) ont été proposées. Cependant, ces approches sont basées sur des algorithmes génériques de lissage et de filtrage qui ne tiennent pas compte des spécificités du mouve-

ment facial et requièrent un temps de calcul relativement long. De plus, les algorithmes utilisées pour traiter les discontinuités ont tendance à supprimer les informations pertinentes (micro mouvements, rides, ...) liées aux expressions faciales et ne conviennent donc pas à ces applications. D'autres approches, comme celle proposée par Farnebäck [34], permettent de calculer le mouvement dense rapidement, sans réduire le bruit induit par les caractéristiques faciales. Bien que ces approches ne filtrent pas le bruit, elles garantissent que les données extraites n'ont subit aucunes modifications et que l'information initiale liée aux expressions faciales est conservée. Les approches ne filtrant pas le bruit, constituent une base plus fiable pour caractériser les expressions faciales.

La Figure 4.1 représente le calcul du flux optique dense entre les deux images 4.1-A et 4.1-B à l'aide des approches de Farnebäck 4.1-C et de DeepFlow 4.1-D. Bien que visuellement l'approche DeepFlow permettent de corriger les discontinuités de mouvement sur le visage, on constate que certaines régions dénuées de mouvement ont été estimées en lissant le mouvement (i.e le front). Au niveau de la joue gauche, nous pouvons voir que le lissage a enlevé le mouvement induit par l'apparition de la ride (que l'on visualise sur l'autre flux). Dans ce cas, rien ne garanti que les informations estimées sont relatives aux expressions faciales. De ce fait, il nous semble plus intéressant de construire un modèle basé sur un flux optique dense non filtré, comme celui proposé par Farnebäck et d'y ajouter certaines contraites afin de s'abstraire des discontinuités de mouvement, et de conserver uniquement l'information induite par les expressions faciales.

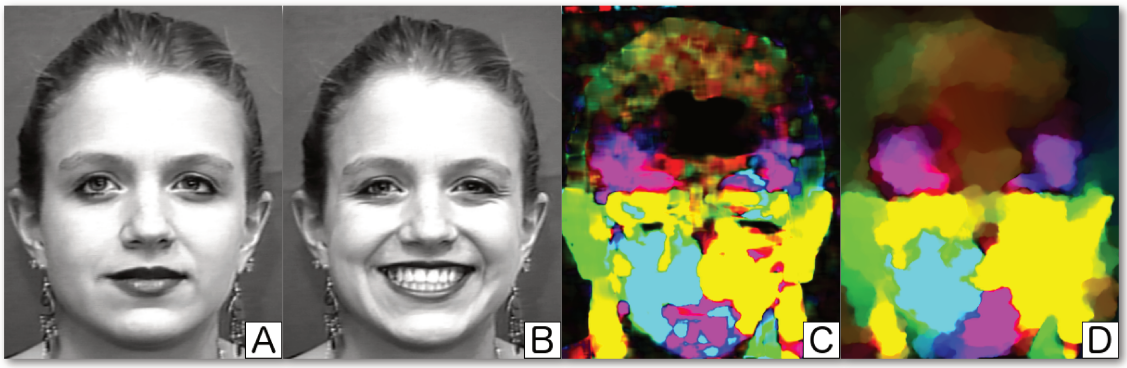


FIGURE 4.1 – Application des méthodes de flux optique dense de Farnebäck (C) et de DeepFlow (D) pour calculer le mouvement entre les images A et B (CK+).

Inspiré par les récentes utilisations des approches basées sur le mouvement pour caractériser les macro et les micro expressions, nous explorons l'utilisation des contraintes de magnitude et d'orientation afin d'extraire uniquement le mouvement induit par les expressions faciales. Dans ce chapitre, nous proposons un descripteur innovant appelé LMP (Local Motion Patterns), permettant de filtrer et de caractériser le mouvement des expressions faciales en nous affranchissant des discontinuités de mouvement. Nous nous appuyons sur les spécificités du mouvement facial afin de renforcer les caractéristiques des mouvements induits par les muscles faciaux, et d'extraire les directions principales du mouvement liées aux expressions faciales en s'affranchissant des discontinuités de mouvement.

Dans la suite de ce chapitre, nous discutons des spécificités des mouvements faciaux. Plus spécifiquement, nous analysons les caractéristiques du mouvement en présence d'expressions faciales, calculé à l'aide d'une approche de mouvement dense rapide (i.e. Farnebäck [34]). Puis, nous décrivons le processus permettant de dissocier plus distinctement les directions principales associées au mouvement facial. Enfin, nous concluons en présentant les apports de notre filtre de caractéristiques du mouvement pour caractériser un mouvement facial.

4.2 Caractéristiques du mouvement facial

L'extraction du mouvement facial demande avant tout d'étudier les spécificités liées au visage. Il est donc important de s'intéresser aux propriétés déformables des tissus biologiques tels que la peau et les muscles faciaux.

Les propriétés élastiques du visage impliquent que la déformation faciale induite par les expressions se caractérise par une force de traction (étirement) ou de compression résultant de la contraction des muscles faciaux. Les forces exercées par les muscles faciaux en présence d'expressions faciales induisent un déséquilibre thermo-dynamique dans les tissus biologiques du visage. Afin de retrouver un équilibre thermo-dynamique, la peau subit des déformations. Selon les lois physiques s'intéressant à la déformation des tissus biologiques, les hypothèses suivantes peuvent être énoncées pour caractériser un mouvement facial :

(Hypothèse A) Si l'on s'intéresse à une petite région faciale subissant de petites déformations, alors on peut dire que la déformation est linéaire et réversible quelle que soit la sollicitation. On peut donc supposer qu'il existe une cohérence locale de la distribution du mouvement en termes de magnitude et de direction.

(Hypothèse B) Les déformations appliquées à un solide sont contraintes à s'appliquer en fonction de la forme et de la matière du solide. Il est alors possible de considérer que le mouvement d'une petite région induit par un muscle facial est contraint à suivre une direction principale directement liée à l'activation musculaire.

(Hypothèse C) Pour de petites déformations, l'allongement d'un corps élastique est proportionnel à la force appliquée. Dans ce cas, la propagation du mouvement au sein du visage est directement proportionnelle à l'intensité d'une contraction musculaire.

(Hypothèse D) L'équilibre thermo-dynamique des tissus biologique permet de considérer que la magnitude et la direction de la déformation se propagent de manière continue dans le voisinage d'un muscle facial.

Dans la suite de cette section, nous vérifions de manière empirique si les différentes hypothèses énoncées sont vérifiées lorsque l'on caractérise un mouvement facial. Pour chacune de ces hypothèses, nous analysons plusieurs distributions de mouvements extraites de différentes régions du visage. Le mouvement facial est calculé à l'aide de la méthode de flux optique dense proposée par Farnebäck [34]. Contrairement aux méthodes récentes de flux optique dense, cette méthode n'inclut aucun pré-traitements (i.e. lissages) qui peuvent modifier l'information extraite. De plus, cette solution a la particularité de s'exécuter très rapidement. Les analyses sont réalisées en présence de déformations faciales induites d'une part par des macro expressions, et d'autre part par des micro expressions.

4.2.1 Contrainte locale de magnitude et de direction

Dans cette section, nous nous intéressons à la cohérence locale de la distribution du mouvement en termes de magnitude et de direction. Plus spécifiquement, nous analysons la corrélation entre la direction et la magnitude du mouvement en présence d'une

expression faciale. Selon l'hypothèse A, une déformation faciale est linéaire, ce qui implique qu'il existe une cohérence locale de la distribution du mouvement en termes de magnitude et de direction. Dans le cas contraire, la cohérence du mouvement local n'est pas garantie.

Afin de vérifier l'hypothèse A, nous analysons la déformation faciale induite par les muscles faciaux en présence d'un sourire (élévation du coin des lèvres). La Figure 4.2 illustre la distribution locale du mouvement extraite de différentes régions du visage. Les distributions sont calculées à partir de l'algorithme de flux optique dense proposé par Färneback [34]. Dans cette analyse, la dimension des régions (paramètre λ) correspond à 3 pourcents de la taille du visage analysé, ce qui correspond ici, à des régions de dimension 15*15 pixels, que ce soit pour la macro ou la micro expression. La distribution de la direction du mouvement est analysée sur plusieurs couches de magnitudes. Les couches de magnitudes varient de 0 à 10, avec un pas de 0.2 entre chaque couche (50 couches de magnitudes). Nous avons décidé de ne pas considérer les magnitudes allant au-delà de 10 au vu de la taille moyenne des visages contenus dans les bases de données analysées. Cet intervalle a montré son efficacité pour caractériser équitablement les macro (cadence d'enregistrement de 25 img/s) et les micro (cadence d'enregistrement de 100-200 img/s) expressions. Au même titre que les différents paramètres du descripteur proposé dans la suite de la section, cette valeur peut être modulée en fonction de la taille du visage et/ou de la cadence d'enregistrement de la vidéo. Actuellement, l'amplitude de 10 pixels caractérisant la magnitude maximale correspond à 2% de la diagonale du visage moyen. Dans la Figure 4.2, plus la magnitude est élevée, plus la courbe associée est rouge. La courbe bleue représente la distribution totale du mouvement sans filtrage de magnitude. L'abscisse représente la direction, divisée en 36 bins (10° par bin). L'ordonnée représente le nombre d'occurrences des pixels au sein de la distribution (%).

La première colonne de la Figure 4.2 représente la distribution du mouvement localisée au niveau du coin des lèvres. En présence d'une macro expression (première ligne de la Figure 4.2), nous remarquons une succession importante de couches de magnitudes convergeant dans une direction commune. Nous faisons le même constat en présence d'une micro expression (deuxième ligne de la Figure 4.2), à la seule différence du nombre

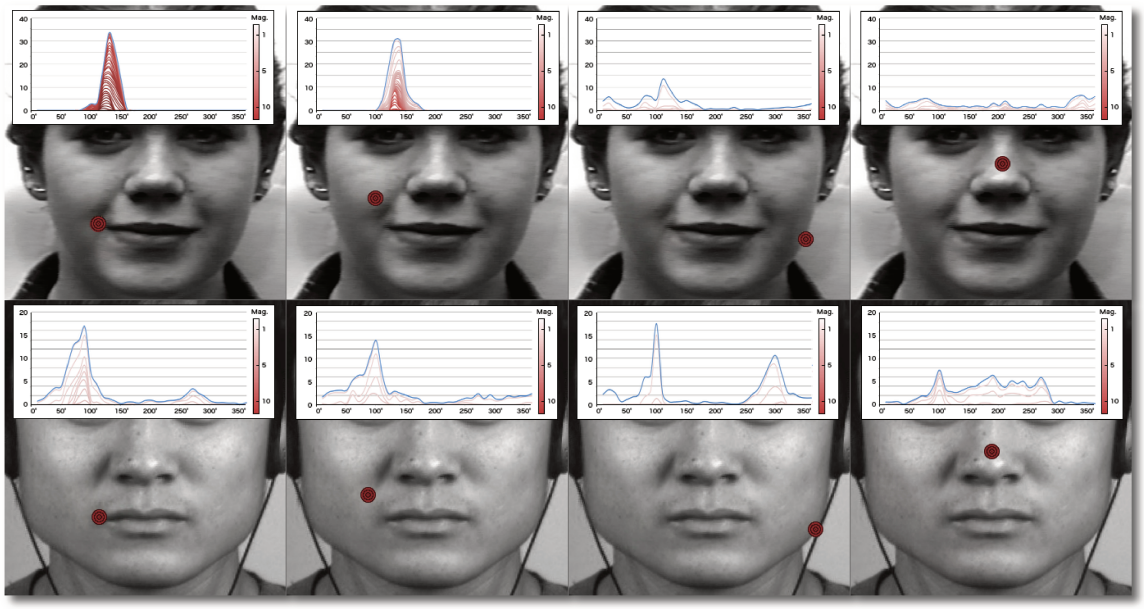


FIGURE 4.2 – Analyse de la distribution du mouvement sur différentes échelles de magnitude, en présence d'une macro (première ligne - CK+) et d'une micro (deuxième ligne - CASME II) expression (sourire).

de couches de magnitudes. Cela s'explique par le fait qu'une micro expression est moins intense qu'une macro expression. Au vu de ces deux distributions, nous pouvons identifier que la direction principale du mouvement a tendance à rester identique au travers des différentes couches de magnitudes.

La deuxième colonne de la Figure 4.2 représente la distribution du mouvement dans une région proche du coin des lèvres. Nous remarquons toujours le même comportement, aussi bien en présence de macro et de micro expressions. La distribution étant plus éloignée de l'épicentre du mouvement, la magnitude est moins élevée. Ceci est également illustré par le fait que le nombre de couches de magnitudes est moins important. Nous remarquons dans ce cas, que l'écart entre les couches successives a tendance à augmenter, ce qui montre que la linéarité du mouvement local tend à s'atténuer.

Les deux dernières colonnes de la Figure 4.2 représentent des distributions du mouvement extraites dans des régions non affectées par le mouvement du coin des lèvres. Nous remarquons dans ce cas, que la distribution du mouvement a tendance à être plus hétérogène. Ceci implique que la direction principale du mouvement ne se distingue plus autant que précédemment. Parfois, la direction principale se distingue très clairement

dans la distribution, c'est le cas de la colonne trois, correspondant au micro expression. Dans ce cas, il est important de noter que le nombre de couche de magnitude est faible et que les différentes couches de magnitudes convergeant dans une même direction sont très espacées. Ce qui signifie que la linéarité du mouvement local n'est plus garantie.

Au vu des différentes distributions, nous pouvons en conclure qu'une déformation faciale induite par un muscle se caractérise par un mouvement linéaire, où la direction et la magnitude sont corrélées. Dans ce cas, la direction principale se caractérise par une progression continue de l'orientation sur plusieurs niveaux de magnitude, ce qui permet de valider l'hypothèse A. La cohérence du mouvement local peut s'évaluer en fonction de la convergence successive des différentes couches de magnitudes dans une même direction. Plus le nombre de couches de magnitudes est important et l'écart entre ces couches est réduit, plus le mouvement à de chance d'être cohérent localement.

La corrélation entre la direction et la magnitude est un critère à prendre en compte pour caractériser un mouvement facial. Cependant, ce critère seul ne suffit pas toujours à garantir que la direction principale reflète un mouvement cohérent. C'est notamment le cas, lorsque la distribution du mouvement est hétérogène et où le mouvement facial a tendance à se confondre avec le bruit (discontinuité de mouvement, bruit d'acquisition). Dans ce cas, il est important de vérifier un deuxième critère de cohérence s'appuyant sur les contraintes liées à la forme et à la matière d'un solide déformable (hypothèse B). Dans la section suivante, nous proposons une analyse permettant de vérifier ce critère.

4.2.2 Contrainte locale de la distribution du mouvement

Dans cette section, nous nous intéressons à la répartition de la direction du mouvement facial au sein d'une petite région. Plus spécifiquement, nous vérifions que l'hypothèse B stipulant que la déformation locale d'un solide est directement liée à sa forme et à sa matière, s'applique bien au visage.

Dans le cas d'une déformation faciale, nous devons identifier comment se manifeste la déformation de la peau en fonction du mouvement induit par un muscle facial. Pour cela, nous reprenons les mêmes régions faciales que précédemment, mais ici nous analysons uniquement la direction de la distribution totale des pixels au sein d'une région

(15*15 pixels), sans prendre en compte la magnitude. La Figure 4.3 illustre l'ensemble des distributions de mouvements induits par une macro (première ligne) et une micro (deuxième ligne) expression. Les distributions sont calculées à partir de l'algorithme de flux optique dense proposé par Färneback [34]. Concernant les distributions, l'abscisse représente la direction, divisée en 36 bins (10° par bin), et l'ordonnée représente le nombre d'occurrences des pixels au sein de la distribution (en pourcentage).

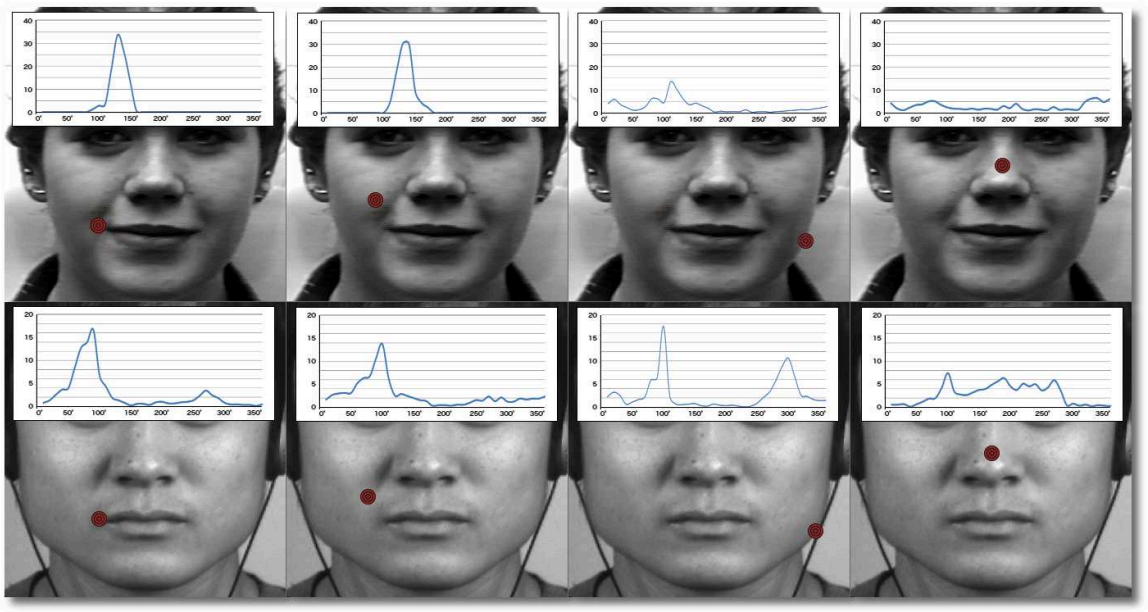


FIGURE 4.3 – Analyse de la répartition de la direction du mouvement facial au sein de la distribution, en présence d'une macro (première ligne - CK+) et d'une micro (deuxième ligne - CASME II) expression (sourire).

Les deux premières colonnes de la Figure 4.3 représentent les distributions du mouvement extraites dans des régions proches de l'épicentre du mouvement. Nous remarquons que la distribution du mouvement autour de la direction principale a tendance à s'étaler en diminuant progressivement sur les directions voisines. Le constat est le même que ce soit en présence d'une macro (première ligne) ou d'une micro (deuxième ligne) expression. Ceci montre que la déformation locale converge progressivement vers une direction principale.

Concernant les deux dernières colonnes de la Figure 4.3, où aucun mouvement induit par l'expression n'est observable, nous distinguons plusieurs catégories de distributions :

- Dans certaines régions, la distribution s'étend sur l'ensemble des directions avec une très faible intensité. Dans ce cas, l'information contenue dans la distribution n'est pas suffisamment importante pour caractériser une déformation faciale. Les petites variations observées sont généralement dues aux bruits d'acquisition.

- Dans d'autres cas, un large étalement en termes de distribution significative peut être observé, comme dans la dernière colonne de la deuxième ligne. Il est alors important de vérifier si le mouvement représenté par la distribution est pertinent par rapport aux spécificités des déformations faciales. Si l'on considère que l'on analyse des petites régions, il y a de forte chance que la direction principale ne s'étale pas au-delà d'un certain seuil ($\in$ [0°-60°]).

- Parfois, une direction principale se caractérise par une forte concentration dans une unique direction. C'est notamment le cas dans la troisième colonne de la deuxième ligne. Cette forte concentration est généralement observée en présence de discontinuités de mouvement, principalement au niveau des contours. Cela s'explique par le fait que les vecteurs de mouvement associés aux pixels des contours ont tendance à être mal estimés à cause des occultations. Dans ce cas, il n'y a pas de continuité du mouvement autour de la direction principale, ce qui se traduit par une très forte variation entre deux directions différentes au sein de la distribution.

Au vu de ces distributions, nous remarquons que le mouvement au sein d'une petite région est contraint à couvrir un intervalle de directions assez restreint, lié à la force appliquée, tout en conservant une certaine cohérence dans la convergence des directions. Il est alors important de s'assurer que les directions principales restent cohérentes en termes d'intensité, de variation et de recouvrement, pour se conformer aux contraintes physiques caractérisant une déformation faciale.

Jusqu'ici, nous avons discuté de plusieurs critères permettant de s'assurer de la cohérence du mouvement au sein d'une petite région. Cependant, la force appliquée en présence d'une expression faciale implique généralement que la déformation se propage au-delà de la région contenant l'épicentre du mouvement. Dans la section suivante, nous analysons comment se propage le mouvement induit par une déformation faciale.

4.2.3 Contrainte de propagation du mouvement

Dans cette section, nous nous intéressons à la propagation du mouvement facial au sein d'une petite région (15*15 pixels). Plus spécifiquement, nous vérifions les hypothèses C et D liées aux caractéristiques élastiques de la peau du visage. L'objectif est de vérifier si la direction principale du mouvement au sein de différentes régions voisines reste cohérente malgré la distance qui les sépare du coin des lèvres.

La Figure 4.4 illustre les distributions extraites de trois régions situées à des distances différentes du coin des lèvres. Les distributions sont calculées à partir de l'algorithme de flux optique dense proposé par Färneback [34]. La dernière colonne regroupe les trois distributions pour comparer facilement le chevauchement entre les directions principales. Concernant les distributions, l'abscisse représente la direction, divisée en 36 bins (10° par bin), et l'ordonnée représente le nombre d'occurrences des pixels au sein de la distribution (en pourcentage).

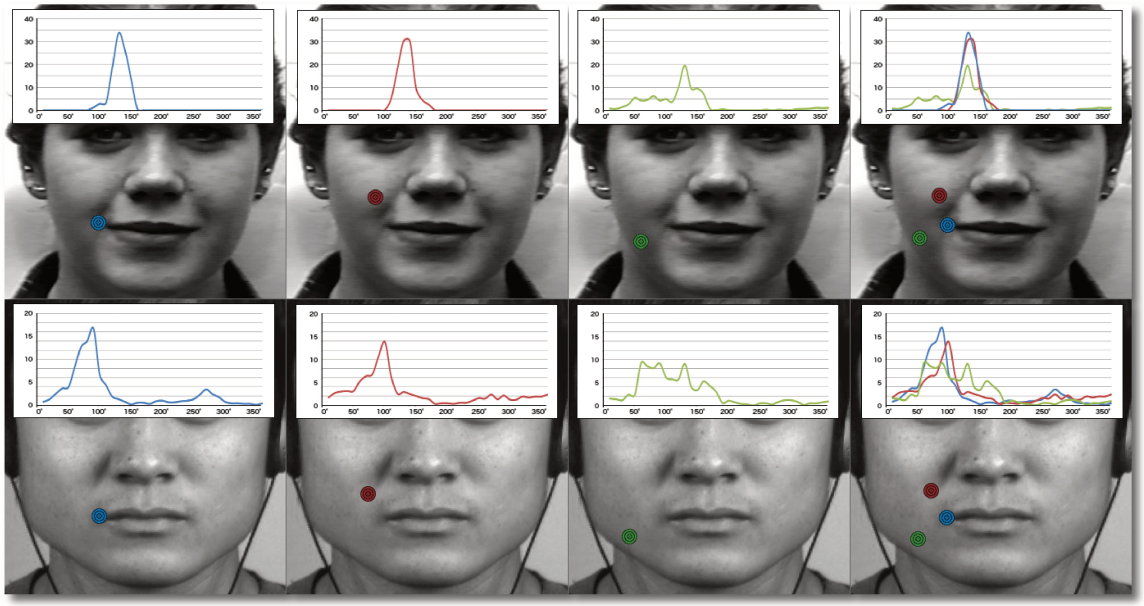


FIGURE 4.4 – Analyse de la propagation du mouvement autour du coin des lèvres, en présence d'une macro (première ligne - CK+) et d'une micro (deuxième ligne - CASME II) expression (sourire).

Au vu des différentes distributions concernant la macro expression (Figure 4.4, pre-

mière ligne), nous remarquons une correspondance forte entre les directions principales malgré la distance qui sépare les régions du coin des lèvres. Cela montre que la déformation se propage de manière continue dans le voisinage d'un muscle facial en présence d'un mouvement de forte intensité. Cependant, plus la distance avec l'épicentre du mouvement est grande, plus la direction principale diverge vers d'autres bins.

Concernant la micro expression (Figure 4.4, deuxième ligne), nous remarquons le même comportement. En revanche l'intensité musculaire étant plus faible, la propagation du mouvement est moins importante. Ceci est illustré dans la distribution de la troisième colonne, où la direction principale du mouvement induit par le coin des lèvres ne se distingue plus aussi clairement.

Au vu de ces résultats, nous pouvons conclure que la propagation du mouvement au sein du visage semble liée aux propriétés élastiques de la peau, et est directement proportionnelle à l'intensité d'une contraction musculaire. Ceci confirme que les hypothèses C et D s'appliquent dans le cadre des mouvements faciaux. La déformation faciale implique donc qu'il existe une cohérence de magnitude et de direction du mouvement entre deux régions voisines.

Dans la section suivante, nous faisons une synthèse des différentes hypothèses liées aux spécificités du visage. De ces hypothèses, nous proposons un processus de caractérisation du mouvement adapté aux déformations de solide tel que le visage.

4.2.4 Synthèse

Au vu des hypothèses liées au mouvement facial, la caractérisation du mouvement facial se fait à deux niveaux d'analyse différents. La validation des hypothèses A et B consistent à analyser la véracité d'un mouvement au sein d'une petite région, en s'assurant qu'il existe une cohérence locale du mouvement en fonction de la force appliquée. Quant aux hypothèses C et D, elles permettent de valider la cohérence du mouvement en s'assurant que celui-ci se répercute dans le voisinage de l'épicentre. Bien que les expérimentations présentées dans cette section se sont uniquement focalisées sur le comportement du mouvement autour du muscle facial de la bouche, il est important de noter que le même comportement s'applique aux autres muscles du visage. Nous considérons que les direc-

tions principales du mouvement facial doivent satisfaire l'ensemble de ces contraintes afin de dissocier plus distinctement le mouvement facial du bruit.

En partant de ces différentes hypothèses, nous proposons un processus de caractérisation du mouvement facial permettant d'extraire uniquement les directions principales, en s'abstrayant du bruit (bruit d'acquisition, discontinuité du mouvement). Une illustration des différentes étapes du processus de filtrage et de caractérisation du mouvement est présentée dans la Figure 4.5. Chaque région est analysée localement afin de vérifier si la distribution du mouvement local est cohérente avec la déformation faciale observée (Figure 4.5-1 (hypothèse A) et Figure 4.5-2 (hypothèse B)). Dans le cas où le mouvement est cohérent localement, l'analyse est appliquée à une région connexe pour s'assurer que le mouvement se propage correctement dans son voisinage (Figure 4.5-3 (hypothèses C et D)). Si l'une des étapes du processus échoue, alors le mouvement analysé n'est pas associé à une véritable déformation faciale.

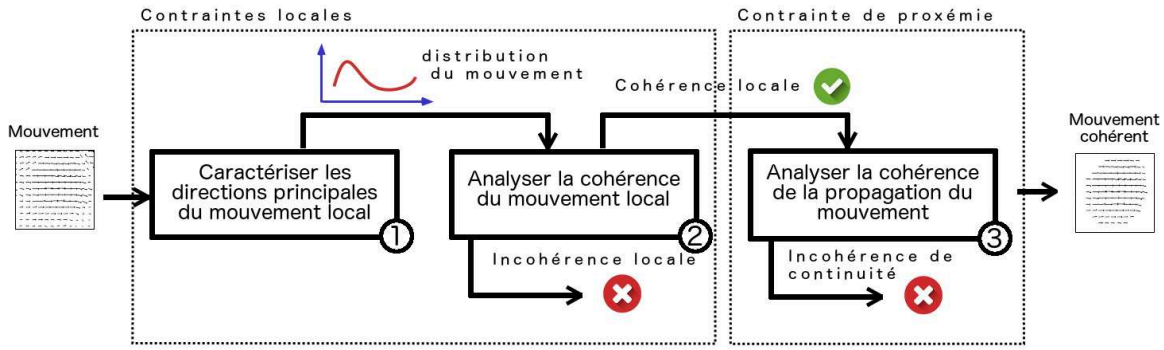


FIGURE 4.5 – Schéma du processus de filtrage pour la caractérisation du mouvement.

Dans la suite, nous détaillons le processus de filtrage et de caractérisation du mouvement facial appelé LMP (Local Motion Patterns) en nous basant sur le processus défini dans la Figure 4.5.

4.3 LMP

Les caractéristiques faciales (élasticité et réflectance de la peau, texture lisse) induisent des inconsistances et du bruit dans le processus d'extraction des mouvements faciaux.

Nous nous appuyons sur les spécificités du mouvement facial afin de renforcer les caractéristiques des mouvements induits par les muscles faciaux. Au vu des caractéristiques faciales, nous considérons qu'un mouvement naturel implique une certaine cohérence locale (pas de discontinuité) et doit se propager de manière continue autour des régions voisines.

Pour prendre en compte ces hypothèses de contraintes du mouvement, et extraire le mouvement cohérent d'une région faciale spécifique, nous proposons un nouveau descripteur appelé LMP (Local Motion Pattern). Un LMP filtre le bruit et caractérise les directions principales du mouvement au sein d'une région, en vérifiant localement la contrainte de propagation du mouvement dans une région. Chaque région définie par rapport à son centre C(x,y), appelée LMR (Local Motion Region) est caractérisée par un histogramme de flux optique $H_{LMR_{x,y}}$. Nous définissons deux types de LMR :

- La région centrale (qui suppose la direction principale que prendra le mouvement au sein du LMP) est appelée CMR (Central Motion Region). Cette région permet de déterminer la nature du mouvement à l'épicentre du LMP afin d'anticiper la propagation sur les régions voisines.
- Les régions voisines situées autour de l'épicentre sont appelées NMR (Neighboring Motion Region). Elles permettent d'analyser la propagation du mouvement et de quantifier la cohérence du mouvement par rapport à la direction principale définie au sein du CMR.

Une représentation du LMP est donnée dans la Figure 4.6. Huit NMRs sont générées autour du CMR. Toutes les régions sont placées à une distance Δ par rapport au CMR. La distance définie par Δ représente le niveau de chevauchement entre deux régions voisines. La variable λ définit la dimension d'une région, où chaque région recouvre $\lambda * \lambda$ pixels. Enfin, la variable β définit le nombre de propagations directes à partir de l'épicentre pour vérifier la cohérence de la propagation du mouvement dans les régions voisines. L'impact de ces différents paramètres sur la qualité du filtrage du mouvement sera détaillé lors de l'évaluation du descripteur dans la section 5.5.2.

Dans la suite de cette section, nous présentons en détail la méthodologie permettant d'extraire le mouvement pertinent sur un visage, en s'appuyant sur les caractéristiques

du visage, et les contraintes de mouvements induites par les muscles faciaux.

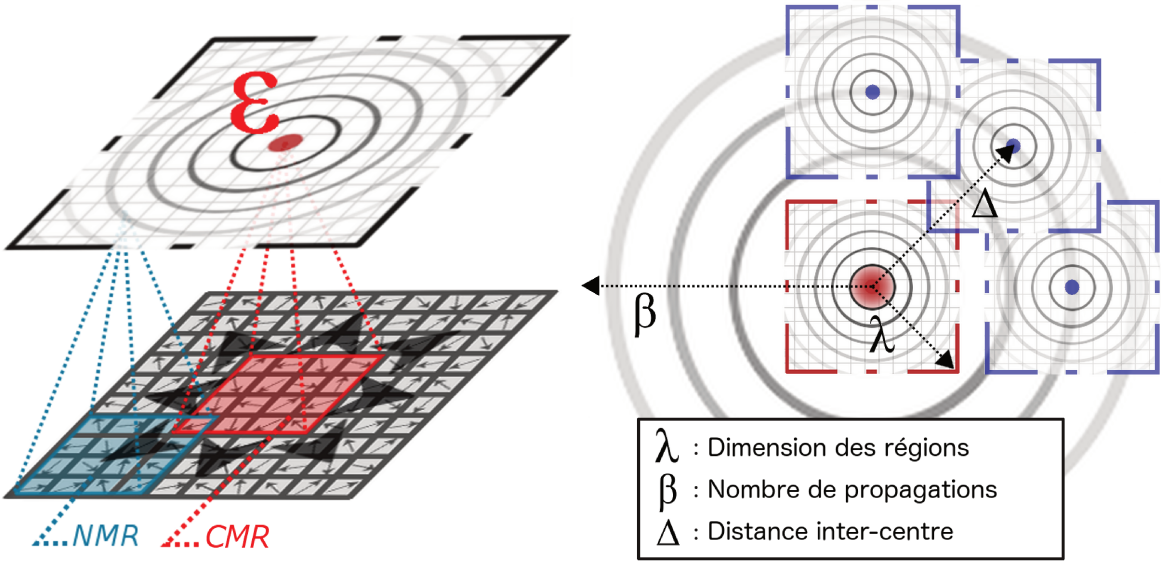


FIGURE 4.6 – Vue générale de la structure d'un LMP.

4.3.1 Cohérence locale du mouvement

Afin de filtrer le mouvement facial, nous commençons par analyser la distribution du mouvement au sein d'une CMR. L'objectif est de conserver uniquement le mouvement pertinent contenu dans l'information retournée par le flux optique. Pour cela, nous proposons d'extraire uniquement le mouvement des directions principales en s'abstrayant du bruit (bruit d'acquisition, discontinuité de mouvement). La première étape du processus consiste à calculer la distribution du mouvement d'une petite région sur plusieurs couches de magnitudes. Puis, pour chaque direction, nous calculons le nombre d'occurrences d'une même direction à différents niveaux de magnitudes. Enfin, nous pondérons la distribution du mouvement en fonction de ces différents niveaux de magnitudes pour chaque direction. Cela nous permet d'obtenir une nouvelle distribution contenant uniquement les directions principales du mouvement, en privilégiant la cohérence forte en termes de magnitude. Chacune de ces étapes est détaillée dans la suite de cette section.

[Etape 1] : Comme toutes régions au sein du LMP, une CMR est une LMR et est définie par une localisation (x,y). Une LMR se caractérise par un histogramme de direction

$H_{LMR_{x,y}}$, de taille B bins (répartition en B classes de direction). La distribution du mouvement de la LMR est divisée en q histogrammes correspondant à différentes couches de magnitudes. Ce découpage permet d'analyser plus finement le mouvement local. L'analyse multi-couches de la magnitude du mouvement permet d'identifier les mouvements de faibles et fortes intensités qui se manifestent sur plusieurs couches successives. Chaque couche de magnitude de la LMR est définie comme suit :

$$MH_{LMR_{x,y}}(n,m) = \{(bin_i, mag_i) \in H_{LMR_{x,y}} \mid mag_i \in [n,m]\}. \tag{4.1}$$

où n et m représentent les intervalles de magnitudes et $i = 1,2,...,Bin$ est l'index des bins (classes) de direction. La distribution du mouvement par couches de magnitude ($MH_{LMR_{x,y}}$), d'une macro et d'une micro expression est illustrée dans la Figure 4.7. Dans le cas présent, nous avons fait varier le paramètre n de 0 à 10, par pas de 0.2. Quant au paramètre m, celui-ci est fixé à 10 afin de garantir un recouvrement des différentes couches de magnitudes. Les couches successives de magnitudes permettent facilement de distinguer les directions principales.

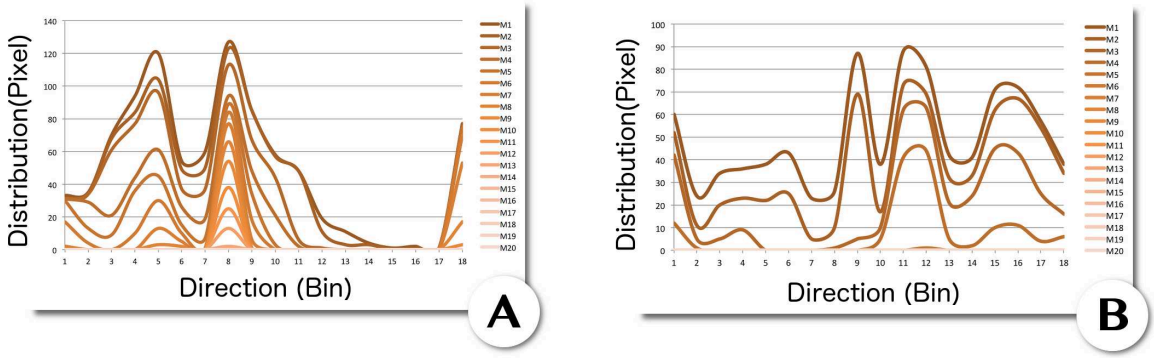


FIGURE 4.7 – Représentation de la distribution du mouvement par couche de magnitude d'une macro (A) et d'une micro (B) expression.

[Etape 2] : Chaque $MH_{LMR_{x,y}}$ est ensuite normalisé. Les directions ayant une magnitude inférieure à 10% sont filtrées (mises à 0). Chaque distribution est ensuite divisée en trois intervalles $P_1 \in (0\%, 33\%]$, $P_2 \in]33\%, 66\%]$ et $P_3 \in]66\%, 100\%]$, représentés par trois histo-

grammes cumulés $\mathrm{ML}_{\mathrm{LMR}_{x,y}}(m1, m2)$, calculés comme suit :

$$\mathrm{ML}_{\mathrm{LMR}_{x,y}}(m1, m2) = \{(bin_i, Card\{(n, m) \mid \exists(bin_i, mag_i) \in \{\mathrm{MH}_{\mathrm{LMR}_{x,y}}(n, m) \mid m1 < mag_i < m2\})\}. \tag{4.2}$$

Les différents intervalles permettent ainsi de dissocier les directions ayant une représentativité faible, moyenne et forte dans l'ensemble des niveaux de magnitude. Grâce à cela, nous pouvons appliquer un facteur de pondération aux différents niveaux de co-occurrence de direction à travers les magnitudes, ce qui permet de privilégier certaines directions plutôt que d'autres. La Figure 4.8 représente la division des $\mathrm{MH}_{\mathrm{LMR}_{x,y}}$ par les trois intervalles P_1, P_2 et P_3. Chaque $\mathrm{ML}_{\mathrm{LMR}_{x,y}}$ correspond à une ligne du tableau et le nombre inscrit dans chaque cellule représente le nombre d'occurrences de chaque magnitude pour chaque direction.

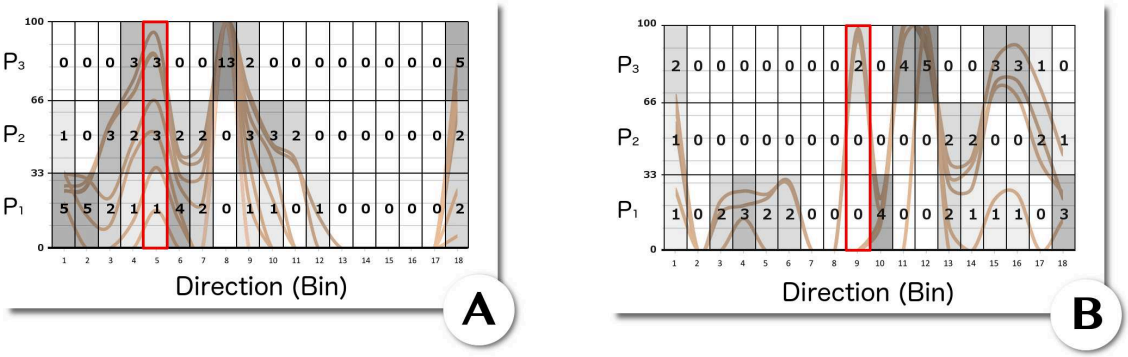


FIGURE 4.8 – Division des $\mathrm{MH}_{\mathrm{LMR}_{x,y}}$ en fonction de trois intervalles de magnitude. (A) macro et (B) micro-expression.

[Etape 3] : L'histogramme de magnitude et de direction $\mathrm{DMH}_{\mathrm{LMR}_{x,y}}$ est calculé en appliquant différents facteurs de pondération ω_1, ω_2 et ω_3, à chaque histogramme cumulé $\mathrm{ML}_{\mathrm{LMR}_{x,y}}(m1, m2)$. $\mathrm{DMH}_{\mathrm{LMR}_{x,y}}$ est défini par l'équation suivante :

$$\begin{aligned} \mathrm{DMH}_{\mathrm{LMR}_{x,y}} = \mathrm{ML}_{\mathrm{LMR}_{x,y}}(m1, m2) * \omega_1 + \mathrm{ML}_{\mathrm{LMR}_{x,y}}(m2, m3) * \omega_2 \\ + \mathrm{ML}_{\mathrm{LMR}_{x,y}}(m3, m4) * \omega_3. \end{aligned} \tag{4.3}$$

Cette équation permet de renforcer la cohérence locale de la magnitude en chaque direction. Nous avons appliqué un facteur de 10 entre chaque couche (ω1 = 1, ω2 = 10 et ω3 = 100). Les valeurs associées aux trois facteurs peuvent varier en fonction de ce que l'on désire quantifier dans l'analyse du mouvement local. Dans notre cas, plus la co-occurrence d'une direction à travers les magnitudes est élevée, plus le poids associé est grand. Ceci permet de donner plus de poids à un mouvement qui a une magnitude constante entre plusieurs couches. Une représentation des histogrammes de directions et de magnitudes $\mathrm{DMH}_{\mathrm{LMR}_{x,y}}$ des deux exemples précédents est illustrée à la Figure 4.9.

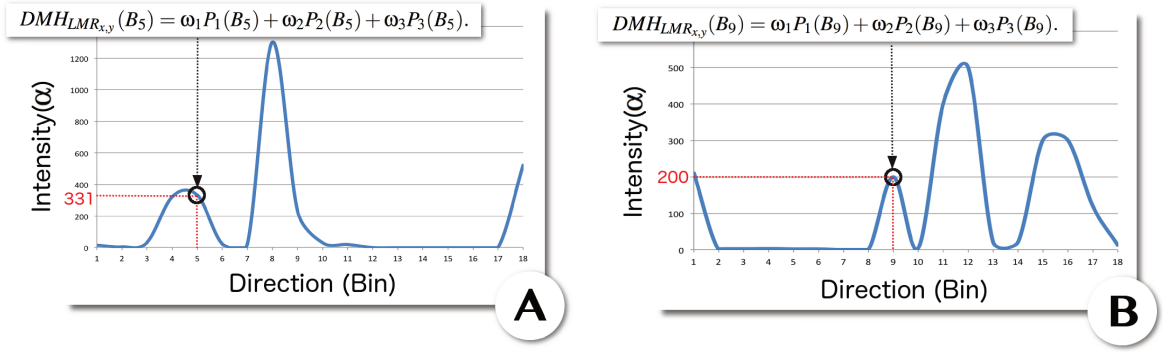


FIGURE 4.9 – Calcul des histogrammes de direction et de magnitude. (A) macro et (B) micro-expression.

À ce stade du processus, nous obtenons une nouvelle distribution du mouvement, où seule les directions principales apparaissent. À présent, nous devons vérifier si ces directions sont cohérentes par rapport aux spécificités liées aux déformations faciales. Dans la section suivante, nous présentons la seconde étape du processus de filtrage de caractéristiques du mouvement au sein du LMP.

4.3.2 Cohérence de la distribution locale

Afin de vérifier les hypothèses du mouvement facial, l'histogramme de magnitudes et de directions du LMR ($\mathrm{DMH}_{\mathrm{LMR}_{x,y}}$) doit contenir au moins une direction principale cohé-

rente par rapport aux spécificités des déformations faciales. Pour cela, nous vérifions que la distribution du LMR ($DMH_{LMR_{x,y}}$) contient au moins une direction principale vérifiant les trois critères suivants :

- L'intensité du mouvement est suffisament élevée.
- Le mouvement principal ne se propage pas dans toutes les directions.
- Le mouvement converge progressivement vers la direction principale.

[Etape 1] : Pour identifier les directions principales au sein d'une distribution, nous appliquons un seuillage α sur la valeur de la co-occurrence d'une direction à travers les niveaux de magnitudes. Le seuil α dépend principalement des poids associés dans l'équation 4.3. Si aucune direction n'est suffisamment intense, cela signifie que le mouvement local ne permet pas de se distinguer suffisamment du bruit dans la région de l'épicentre du LMP. Dans ce cas, nous considérons que le CMR représentant le LMP ne contient pas d'informations importantes, ce qui implique que l'analyse au sein du LMP ne sera pas approfondie.

[Etape 2] : Dans le cas où au moins une direction principale est trouvée, il faut s'assurer que les directions retenues au sein du $DMH_{LMR_{x,y}}$ représentent des mouvements cohérents pour un visage. En effet, un mouvement peut-être cohérent en termes de magnitude, mais il faut également s'assurer qu'il soit cohérent en termes de direction. Plus spécifiquement, la direction principale du mouvement ne peut couvrir un large intervalle de direction au sein d'une petite région du visage. Cette étape permet de conserver un cadre cohérent et sélectif pour l'analyse de la propagation du mouvement.

Pour s'assurer de la cohérence du mouvement en termes de direction, nous analysons la densité des k directions principales du $DMH_{LMR_{x,y}}$. Chaque direction principale doit satisfaire plusieurs critères. Le premier critère vérifie que chaque direction principale s'étend sur un angle de diffusion limité. En effet, si nous analysons une petite région faciale, un mouvement cohérent s'étend rarement sur un large éventail de bins de directions et l'étalement du mouvement est progressif autour de la direction principale. Si

aucune direction principale ne valide ce critère, alors le mouvement au sein du CMR est incohérent, et l'analyse du LMP est arrêtée. Malgré la cohérence du mouvement en termes de magnitude, une direction recouvrant un angle assez large à une forte chance de correspondre à un faux mouvement, et peut entraîner des erreurs lors des propagations au sein du LMP. Le critère est défini par les équations suivantes. La première équation permet de calculer l'étalement d'une direction principale quant à la seconde, elle permet de contrôler si ces intervalles respectent la limite tolérée.

$$\begin{aligned} \mathrm{C}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}) = \{\mathrm{E} = [a..b] \mid \forall_i \in [a..b] \mid \mathrm{DMH}_{\mathrm{LMR}_{x,y}}(i) > \alpha \\ \wedge \nexists j \in \{a-1, b+1\} \mid \mathrm{DMH}_{\mathrm{LMR}_{x,y}}(j) > \alpha\}. \end{aligned} \tag{4.4}$$

$$\mathrm{C}^{'}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}, s) = \{\mathrm{E} \in \mathrm{C}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}) \mid card(\mathrm{E}) < s\}. \tag{4.5}$$

où a et b représentent les limites de l'étalement d'une direction principale et α, le seuil de magnitude. Pour chaque direction principale, nous conservons uniquement les mouvements s'étalant au maximum sur s bins consécutifs.

[Etape 3] : Pour renforcer le fait que le mouvement s'étale progressivement autour d'une direction, il est important de vérifier que la transition entre les différents bins successifs est progressive. Une tolérance Φ est acceptée entre les différents bins. Le critère est validé par l'équation suivante :

$$\begin{aligned} \mathrm{C}^{''}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}) = \{\mathrm{E} = [a..b] \in \mathrm{C}^{'}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}, s) \mid \\ \forall_{i,j} \in \mathrm{E}, ||i-j|| \leq 1 \mid ||\mathrm{DMH}_{\mathrm{LMR}_{x,y}}(i) - \mathrm{DMH}_{\mathrm{LMR}_{x,y}}(j)|| < \Phi\}. \end{aligned} \tag{4.6}$$

Pour finir, l'histogramme filtré de directions et de magnitudes $\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}$ correspond aux k directions principales extraites du $\mathrm{DMH}_{\mathrm{LMR}_{x,y}}$, satisfaisant chacun les critères cités précédemment. L'histogramme du $\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}$ est décrit par l'équation suivante :

$$\mathrm{FDMH}_{\mathrm{LMR}_{x,y}} = \{(b_i, m_i) \in \mathrm{DMH}_{\mathrm{LMR}_{x,y}} \mid \exists \mathrm{E} = [a..b] \in \mathrm{C}^{''}(\mathrm{DMH}_{\mathrm{LMR}_{x,y}}) \wedge (b_i \in \mathrm{E})\}. \quad (4.7)$$

Bien que le mouvement au sein du CMR soit cohérent, la conformité du mouvement au sein du LMP n'est pas encore vérifiée. En effet, bien que l'épicentre du mouvement permette d'identifier de potentielles directions principales, il faut s'assurer que ces derniers se propagent correctement dans leur voisinage. L'analyse de la propagation du mouvement au sein du LMP est détaillée dans la section suivante.

4.3.3 Cohérence dans la propagation du mouvement

Lorsque la distribution du mouvement d'un LMP est cohérente au sein de son CMR (épicentre du mouvement), il faut vérifier que le mouvement calculé se propage correctement dans son voisinage. En effet, l'élasticité de la peau assure qu'un mouvement facial se propage sur le visage de manière progressive et continue. Dans ce cas, la cohérence du mouvement en termes d'intensité et de direction au sein des régions voisines doit être respectée afin de garantir qu'il ne s'agit pas d'un bruit. Le fait que le visage est un élément non-rigide implique que la direction et la magnitude du mouvement peuvent subir des déformations graduelles au cours de la propagation.

Avant de vérifier la cohérence de la propagation du mouvement, il est important de s'assurer de la cohérence local du mouvement au sein de chaque NMR. Pour rappel, chaque NMR est représenté par un LMR, ce qui signifie que chaque NMR est caractérisé par un histogramme filtré de directions et de magnitudes $\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}$. Comme pour le CMR, le mouvement local au sein de chaque NMR doit être cohérent en termes de magnitude et de directions.

En plus de s'assurer de la cohérence de la distribution au sein d'un NMR, il est important de vérifier qu'il existe une corrélation en termes de directions principales du NMR par rapport aux directions principales caractérisant le CMR. Cette étape du processus vérifie l'hypothèse C qui stipule que le mouvement se propage en fonction de la contrainte d'élasticité de la peau. Pour cela, le coefficiant de Bhattacharyya [10] est employée pour déterminer la proximité relative de deux distributions de mouvement. Le coefficient de

Bhattacharyya est calculée comme suit :

$$C'''(\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}, \mathrm{FDMH}'_{\mathrm{LMR}_{x,y}}) = \sum_{i=1}^{B} \sqrt{\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}(i)\mathrm{FDMH}'_{\mathrm{LMR}_{x,y}}(i)}. \quad (4.8)$$

où $\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}$ et $\mathrm{FDMH}'_{\mathrm{LMR}_{x,y}}$ représentent les deux distributions locales et B le nombre de bins. La cohérence du mouvement entre les deux régions adjacentes est assurée si le coefficient est supérieur à un seuil fixé ρ.

Une représentation de l'analyse de la propagation du mouvement au sein d'un LMP est représentée dans la Figure 4.10. La Figure montre la première itération de l'analyse du mouvement autour du CMR, c'est-à-dire la propagation du mouvement du CMR aux huit NMR adjacentes. La cohérence d'une NMR est mesurée en fonction de plusieurs critères :

- La distribution du mouvement au sein de la NMR en termes de magnitude.
- La distribution du mouvement au sein de la NMR en termes de direction.
- L'indice de similarité entre les distributions du mouvement de deux régions connexes.

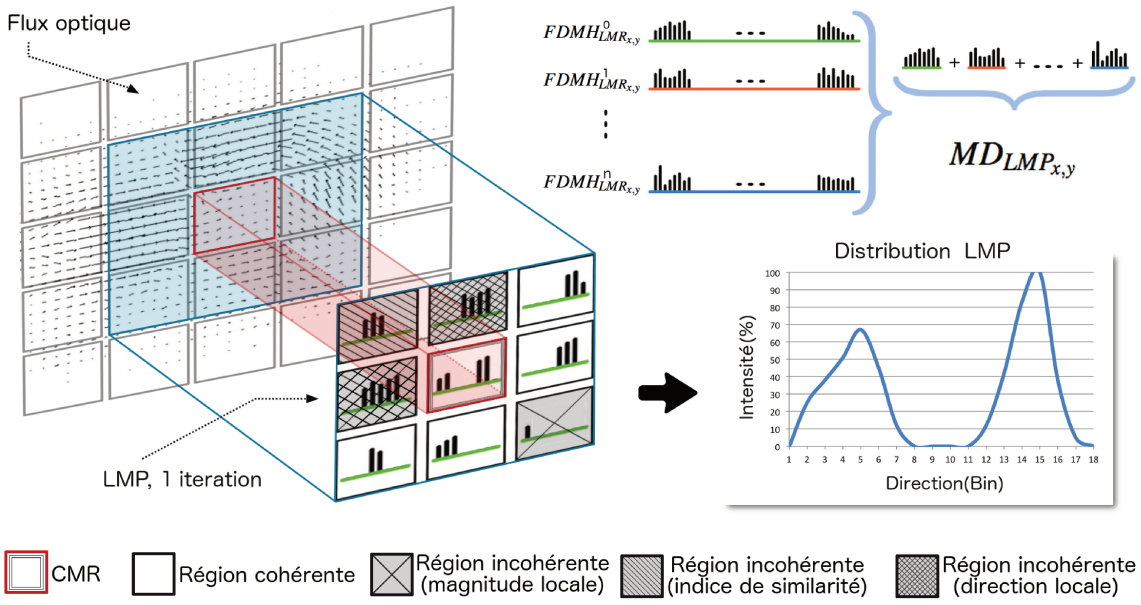


FIGURE 4.10 – Calcule de la distribution du mouvement du LMP au sein du voisinage directe du CMR (première itération).

Si ces trois critères sont vérifiés, une carte de cohérence binaire est mise à jour afin de garder la trace des régions cohérentes et d'éviter de les traiter à nouveau au cours des itérations suivantes. Dans le cas où la distribution du mouvement d'une NMR est incohérente, celle-ci peut-être réévaluée et vérifiée en fonction d'une autre région adjacente lors des prochaines itérations de l'analyse. C'est notamment le cas en présence de discontinuité du mouvement associée à l'apparition ou la disparition d'une ride, où dans ce cas, le mouvement peut être indirectement lié à son voisinage en contournant l'élément perturbateur.

Ensuite, de manière récursive, pour chaque NMR vérifiant les trois critères de cohérence, l'analyse de la propagation du mouvement est réitérée à son voisinage. L'analyse récursive est réitérée tant qu'il existe une cohérence de mouvement avec au moins une région adjacente, ou que le nombre d'itérations souhaitées est atteint (le nombre de progations est défini par le paramètre β du LMP).

Si aucune cohérence de mouvement entre le CMR et les huit NMRs adjacentes n'a été retrouvée, alors cela signifie qu'il n'existe pas de mouvement vérifiant les hypothèses C et D d'un mouvement facial dans la région caractérisant le LMP.

Dans le cas où plusieurs régions adjacentes sont cohérentes, alors les distributions de mouvement ($\mathrm{FDMH}_{\mathrm{LMR}_{x,y}}$) des NMRs ayant une correspondance directe ou indirecte avec le mouvement local du CMR sont cumulées pour caractériser le mouvement au sein du LMP. La distribution du mouvement du LMP est calculée comme suit :

$$\mathrm{MD}_{\mathrm{LMP}_{x,y}} = \{\sum_{i=0}^{n} \mathrm{FDMH}_{\mathrm{LMR}_{x,y}} \mid \mathrm{FDMH}_{\mathrm{LMR}_{x,y}} \in \mathrm{LMP}_{x,y}\}. \tag{4.9}$$

où n représente le nombre de régions où la distribution du mouvement local est cohérente (la CMR et l'ensemble des NMRs cohérentes). Dans le cas d'une analyse 8 connexes (huit régions voisines), le nombre maximum que peut atteindre n peut être calculé par l'équation suivante :

$$Max(n) = 1 + 8 * \frac{\beta(\beta+1)}{2} \text{ si } \beta \geq 1 \text{ , 1 sinon.} \tag{4.10}$$

où β représente le nombre d'itérations. La distribution du mouvement du LMP ($MD_{LMP_{x,y}}$) se caractérise par un histogramme de B bins, où la valeur de chaque bin correspond au cumul des différents histogrammes ($FDMH_{LMR_{x,y}}$) de chaque région où le mouvement local est cohérent. À cette étape du processus, nous pouvons calculer le mouvement cohérent d'une région faciale spécifique définie par un LMP.

Une application de notre filtrage de caractéristiques du mouvement est illustré dans la Figure 4.11. Un LMP est placé sur un visage afin de caractériser le mouvement facial cohérent. Le LMP est caractérisé par un histogramme de directions (décomposé en 12 bins dans l'exemple) représentant la distribution du mouvement qui recouvre la région faciale ayant une cohérence avec l'épicentre représenté par la région verte. Le LMP est directement caractérisé à partir du flux optique calculé par la méthode de Farnebäck [34] (deuxième ligne). Bien que le flux optique soit fortement bruité tout au long de la séquence, le LMP permet de conserver uniquement le mouvement cohérent (ligne 3).

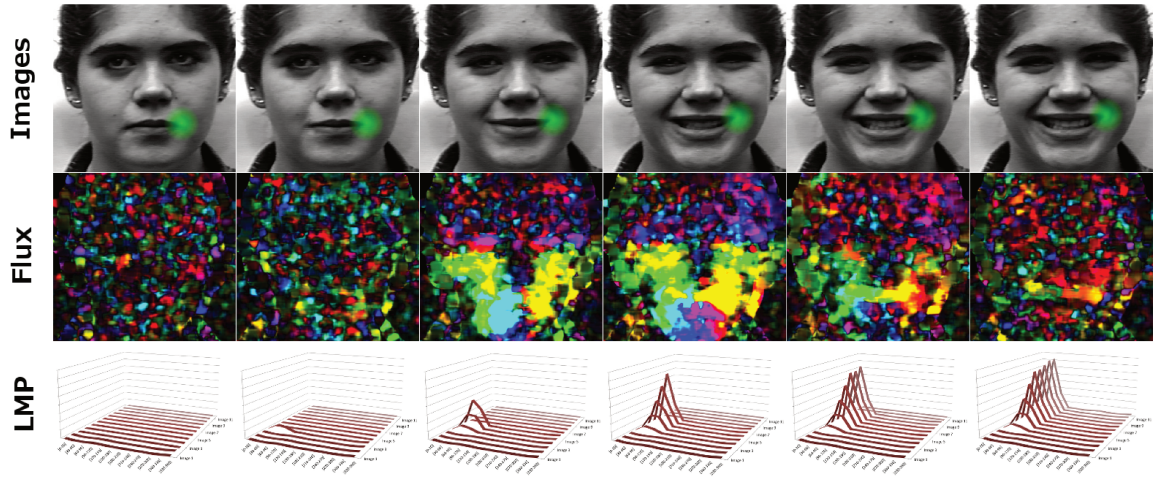


FIGURE 4.11 – Application d'un LMP pour caractériser le mouvement facial cohérent autour du coin des lèvres (région verte) (CK+).

4.4 Conclusion

Dans ce chapitre, nous avons proposé un nouveau descripteur s'appuyant sur les caractéristiques faciales afin de filtrer le bruit d'acquisition et les discontinuités de mouvement occasionnées par différents facteurs (illumination, bruit de capteur, occultations). Le filtrage proposé permet de conserver uniquement les informations liées aux expressions faciales d'intensité variable, en s'appuyant sur le mouvement calculé par une méthode de flux optique dense rapide.

Afin d'identifier et de comprendre les spécificités du mouvement facial, nous avons préalablement analysé le comportement du mouvement facial en analysant localement plusieurs régions du visage. Pour analyser le comportement du mouvement facial, nous avons utilisé une approche de mouvement dense rapide(i.e Färneback [34]) afin d'extraire le mouvement facial. La distribution locale de ces régions a permis d'identifier plusieurs hypothèses concernant le comportement naturel d'un mouvement facial. Les caractéristiques faciales telles que l'élasticité de la peau et les contraintes musculaires impliquent que le mouvement se propage de manière continue au sein du visage, tout en conservant une certaine cohérence de direction et de magnitude. Basée sur ces hypothèses, nous avons proposé un nouveau filtre de caractéristiques du mouvement appelé LMP (Local Motion Pattern), permettant de conserver uniquement le mouvement validant les spécificités liées au visage.

Le filtrage de caractéristiques du mouvement proposé au sein du LMP vérifie trois hypothèses pour s'assurer que le mouvement local est représentatif d'un mouvement facial. Les différentes étapes du processus du filtrage permettent de vérifier qu'il existe a) une convergence entre les différentes couches de magnitude dans une même direction, b) que la distribution locale du mouvement caractérise une ou plusieurs directions principales, et c) qu'il existe une cohérence dans la propagation du mouvement. Ces trois critères s'appuient directement des hypothèses posées en fonction des caractéristiques liées au mouvement facial. Cela assure que le mouvement extrait au sein du LMP est conforme à un mouvement cohérent par rapport aux spécificités du visage.

Dans notre approche, au lieu de calculer explicitement le mouvement sur l'ensemble

du visage, le mouvement est calculé dans des régions spécifiques afin de conserver uniquement les mouvements induits par les expressions faciales. Ces régions sont définies en relation avec le système FACS et permettent d'analyser directement les mouvements cohérents induits par les muscles faciaux. Nous avons ensuite exploité les hypothèses de cohérence du mouvement et amélioré la distinction entre l'information liée au mouvement et le bruit présent dans les données.

Au vu des différentes étapes du processus de filtrage, nous avons vu que le LMP permet, à la fois, de caractériser un mouvement facial avec une magnitude de faible ou de forte intensité. Chaque critère d'évaluation du processus de filtrage peut être configurable indépendamment des autres critères, ce qui rend le LMP entièrement adaptable. Pour vérifier les performances de notre descripteur, nous appliquons, dans le chapitre suivant, notre LMP pour la reconnaissance des macro et des micro expressions.

Chapitre 5

Analyse des micro et macro expressions

« *Le visage contient souvent*
deux messages :
ce que le menteur veut montrer
et ce qu'il cherche à dissimuler »

Paul Ekman

Sommaire

5.1 Introduction

Dans ce chapitre, nous concentrons notre attention sur la reconnaissance des expressions en présence d'une grande diversité des amplitudes de mouvements faciaux. Pour cela, nous portons un réel intérêt à l'étude des macro et micro expressions qui permettent de fournir un panel exhaustif de mouvements faciaux. Pour rappel, les macro expressions sont définies par des mouvements volontaires, caractérisées par des mouvements de fortes intensités d'une durée comprise entre 0.5 et 4 secondes. Quant aux micro expressions, elles sont souvent involontaires et durent une fraction de seconde, en moyenne entre 170ms et 500ms [117]. Bien que ces mouvements sont très rapides et généralement non perceptibles pour l'oeil humain, les micro expressions apportent de précieux renseignements sur l'état affectif d'une personne.

Une large variété d'approches ont été proposées pour la reconnaissance des macro expressions dans des séquences vidéo. Cependant, ces solutions ne conviennent pas en présence des expressions de faible intensité telles que les micro expressions [117]. Bien que les approches récentes tendent vers des techniques communes pour analyser les macro et micro expressions, il n'existe pas une solution unifiée permettant de traiter de manière optimale les deux simultanément. Cela est principalement dû aux différentes caractéristiques de changement d'intensité de mouvement et/ou de la texture.

Dans le chapitre précédent, nous avons présenté un nouveau descripteur, nommé LMP, qui filtre et caractérise le mouvement cohérent dans une région du visage en termes de direction principale. Dans ce descripteur, la direction et l'intensité du mouvement sont conjointement analysées pour obtenir localement un modèle cohérent du mouvement facial. À présent, nous proposons d'employer ce descripteur dans un processus d'analyse d'expressions faciales en unifiant la reconnaissance des macro et des micro expressions. Dans la suite de ce chapitre, nous commençons par étudier les différentes régions faciales qui permettent de caractériser les macro et micro expressions. Ensuite, nous proposons un nouveau modèle de segmentation faciale en fonction des régions d'intérêt sélectionnées. Puis, nous évaluons notre processus d'analyse sur des bases d'apprentissage composées de macro et de micro expressions.

5.2 Définition d'un modèle de segmentation facial

Dans cette section, nous nous intéressons à identifier les régions faciales où il est intéressant d'extraire les patrons de mouvements cohérents nécessaires à une caractérisation optimale des macro et micro expressions. Pour cela, nous étudions deux bases de données composées de macro expressions (CK+) et de micro expressions (CASME II [115]). La base de données CK+ contient plusieurs séquences d'images d'expressions actées dans des conditions contrôlées. CK+ couvre six expressions universelles : joie, peur, tristesse, colère, dégoût et surprise. Quant à la base de données CASME II, elle contient des séquences d'images de micro expressions dans des conditions contrôlées, où les expressions sont : la joie, la surprise, le dégoût et la répression.

Pour identifier les régions du visage où il y a une forte probabilité de mouvement, nous proposons d'analyser la répartition du mouvement facial caractérisant différentes macro et micro expressions. La Figure 5.1 illustre le processus d'extraction de cartes de mouvements correspondant à l'expression joie de la base CK+. Nous commençons pour chacune des deux bases de données, par aligner l'ensemble des visages de chaque séquence en fonction de la position des yeux, puis nous calculons le flux optique sur toute les images de chaque séquence. Chaque image d'une séquence est segmentée à l'aide d'une grille de 20*30 blocs (Figure 5.1-1). Cela garantit un découpage fin du visage en regions recouvrant chacune environ 0.16% de la boîte englobante du visage. La segmentation appliquée permet d'analyser de manière dense, le mouvement facial en présence d'une expression, tout en s'assurant d'analyser les régions non caractérisées par des landmarks (i.e joue, front, ...). Un LMP est extrait dans chacun de ces blocs pour caractériser le mouvement cohérent au sein du visage. La granularité fine de l'analyse du mouvement facial est assurée en analysant le mouvement au sein des LMP, sur des petites régions avec une propagation maximale de 6 voisins. La dimension des régions (λ) est variable en fonction de la taille des visages afin de garantir une homogénéité du processus sur l'ensemble des séquences (en moyenne, les régions ont une dimension de 15*15 pixels pour des visages ayant en moyenne 300*400 pixels). Une région appartenant à plusieurs LMP est prise en compte une seule fois dans le processus afin d'éviter que l'information ne soit sur-représentée. Suite à cela, une carte binaire est construite à partir du nouveau flux optique filtré grâce au LMP. Toutes les cartes binaires correspondant à une même expression, provenant de

toutes les séquences d'une base de données, sont fusionnées pour construire une carte de chaleur. Ces cartes de chaleur représentent la quantification de la présence de mouvements dans les régions respectives pour chaque expression faciale (Figure 5.1-3). Les cartes de chaleur des six expressions faciales de la base CK+ sont illustrées dans la première ligne de la Figure 5.2. La même stratégie a été employée pour caractériser le mouvement lors de la production des micro expressions dans la base de données CASME II. Les cartes de chaleurs obtenues sont illustrées dans la deuxième ligne de la Figure 5.2.

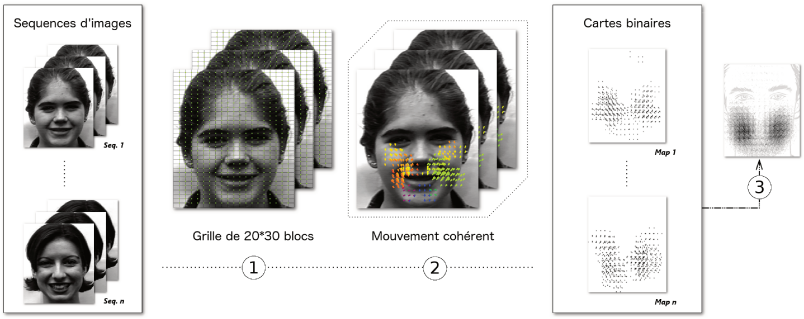


FIGURE 5.1 – Carte du mouvement facial de l'expression joie de la base CK+.

Les cartes de chaleur concernant les macro expressions indiquent que le mouvement a tendance à se localiser au niveau des yeux, du front, des joues et de la bouche. Certaines expressions font intervenir les mêmes régions faciales, mais nous pouvons tout de même distinguer des nuances en termes d'intensité, de direction et de densité de mouvement. Par exemple, l'expression de colère et de tristesse ont des cartes de chaleur assez similaires. Cependant, lorsqu'une personne est en colère, les mouvements associés à l'expression faciale ont tendance à converger vers le centre du visage (la bouche s'élève et les sourcils s'abaissent) tandis que chez une personne triste, les mouvements ont tendance à diverger. Si nous comparons les cartes de chaleur de macro et de micro expressions, nous pouvons voir que les épicentres de mouvement se situent dans les mêmes régions. Mais à la différence des macro expressions, les micro expressions génèrent une propagation et une densité du mouvement très réduites.

À ce stade, les régions principales de mouvements faciaux pour la reconnaissance des expressions sont précisément identifiées. À présent, nous pouvons construire un modèle facial permettant de positionner de manière adéquate les épicentres des LMPs afin de

caractériser de manière précise les mouvements locaux. Inspiré des travaux de Jiang et al. [53], nous utilisons un modèle d'alignement facial pour définir des régions faciales s'adaptant dynamiquement à la morphologie du visage. Pour cela, nous appliquons la méthode proposée par Kazemi et al. [55] qui permet d'extraire les points caractéristiques du visage (landmarks). Cependant, les modèles de landmarks fournissent uniquement des points autours des yeux, du nez et de la bouche, où il existe un fort contrast. Ces points d'intérêts sont généralement utilisés dû à leur forte répétabilité. Cependant, certaines parties du visage sont très peu texturées, comme les joues ou le front. Or, comme l'illustre la Figure 5.2, ces régions fournissent des informations importantes pour la caractérisation des expressions faciales. De ce fait, dans notre modèle facial, nous nous appuyons sur la position des landmarks et sur la morphologie du visage pour calculer de nouveaux points caractéristiques du visage dans des régions peu texturées, comme le front et les joues. Certains landmarks fournis par la méthode de landmarks utilisée ne sont pas pris en compte dans notre modèle. Ces points concernent la région intérieure de la bouche, où il n'est pas évident de caractériser fidèlement le mouvement. Cela est dû à la présence de fortes discontinuités de mouvement quand la bouche s'ouvre ou se ferme (apparition et disparition des pixels). Pour l'analyse des expressions faciales, nous supposons qu'il est plus fiable d'analyser le mouvement sur les contours extérieurs de la bouche.

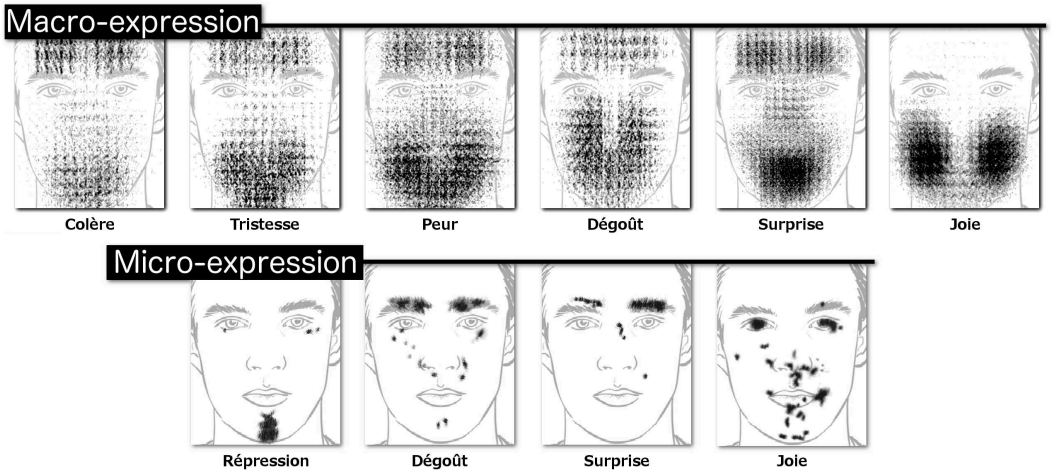


FIGURE 5.2 – Cartes de chaleur du mouvement correspondant aux macro expressions (ligne 1 - CK+) et aux micro expressions (ligne 2 - CASME II).

Le modèle de segmentation faciale proposé est représenté dans la Figure 5.3. La segmentation des différentes régions faciales est élaborée en fonction des différentes cartes de chaleur du mouvement représentant les macro et les micro expressions, extraites depuis les bases de données CK+ et CASME II. De plus, chaque région correspond à l'un des différents muscles faciaux (AUs) afin de couvrir l'ensemble des mouvements liés aux expressions. La segmentation des régions repose uniquement sur la localisation des landmarks. Ainsi l'estimation des régions se fait dynamiquement et les régions s'adaptent en présence de mouvements de la tête (translation, rotation et changement d'échelle). Par exemple, la localisation du point caractéristique Q est estimée en fonction de la localisation des landmarks P10 et P55. Pour définir les régions situées au niveau du front, nous définissons plusieurs points caractéristiques (A,B,...,F) où la distance entre les sourcils et ces points est estimée en fonction de la longueur du nez, correspondant au quart de la distance entre les points P27 et P33. Afin de dissocier le mouvement facial autour des coins des lèvres et le mouvement au niveau de la mâchoire, nous définissons deux couples de régions partageant une partie commune du visage. Les régions en question sont les régions 19,18 et les régions 22, 23.

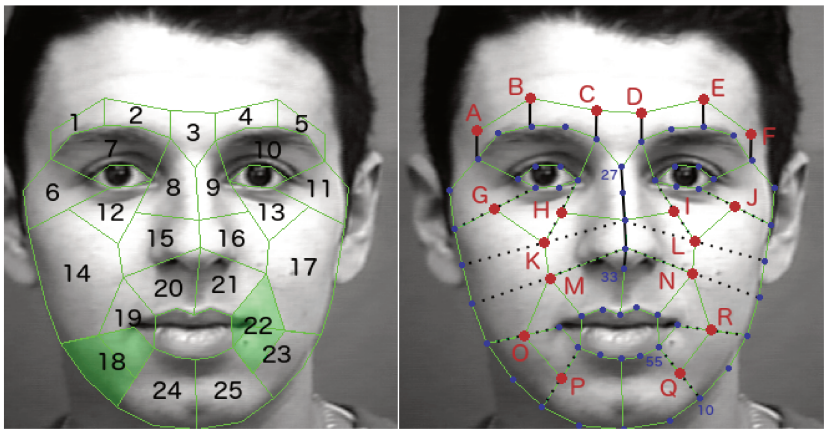


FIGURE 5.3 – Modèle de segmentation facial (CK+).

Dans la section suivante, nous expliquons comment le vecteur de caractéristiques est construit afin d'encoder la relation entre les régions faciales et les macro et micro expressions.

5.3 Construction du vecteur de caractéristiques

Inspiré des approches hybrides, nous proposons de combiner l'information du mouvement à l'information géométrique extraite à partir des 25 régions faciales du modèle de segmentation appliqué au visage. Chaque région est alors représentée par la concaténation de deux vecteurs : le vecteur de mouvements fournit par la concaténation des LMPs, et un vecteur géométrique encodant les relations géométriques entre les landmarks. Dans la suite de cette section, nous détaillons la construction de ces deux vecteurs.

5.3.1 Vecteur de caractéristiques de mouvement

Le vecteur de caractéristiques est construit à l'aide des 25 ROIs du modèle de segmentation faciale proposé précédemment. Dans chaque image f_t, nous calculons la distribution du mouvement filtrée $\mathrm{MD}_{\mathrm{LMP}_{x,y}}$ au sein des différentes ROIs R_t^k, en y appliquant un LMP, où t correspond à l'index de l'image et $k = 1, 2, ..., 25$ représente l'index des ROIs. À chaque image, la distribution du mouvement d'une région R_t^k est cumulée dans un histogramme η^k, où cet histogramme représente la distribution du mouvement facial local à la ROI k pour une séquence d'images. L'histogramme η^k est calculé comme suit :

$$\eta^k = \sum_{t=1}^{time} \mathrm{R}_t^k. \tag{5.1}$$

où R_t^k correspond à la région k de l'image f_t d'une séquence de $time$ images.

Les 25 histogrammes η^k correspondant aux différentes ROIs du modèle facial sont concaténés dans un vecteur GMD, où GMD = $(\eta^1, \eta^2, ..., \eta^{25})$ représente le vecteur de caractéristiques d'une expression faciale de la séquence d'images analysée. La dimension du vecteur GMD est égale au nombre de ROIs du modèle facial multiplié par le nombre de bins retenus dans les histogrammes de distribution du mouvement. Le processus de construction du vecteur GMD est illustré dans la Figure 5.4, où toutes les distributions du mouvement $\mathrm{MD}_{\mathrm{LMP}_{x,y}}$ correspondant aux régions R_t^1, R_t^2, ..., R_t^{25}, et où $t \in [1, time]$, sont cumulées dans le vecteur η^1, η^2, ..., η^{25} respectivement. Puis, chaque vecteur η^k est concaténé dans le vecteur de caractéristiques GMD.

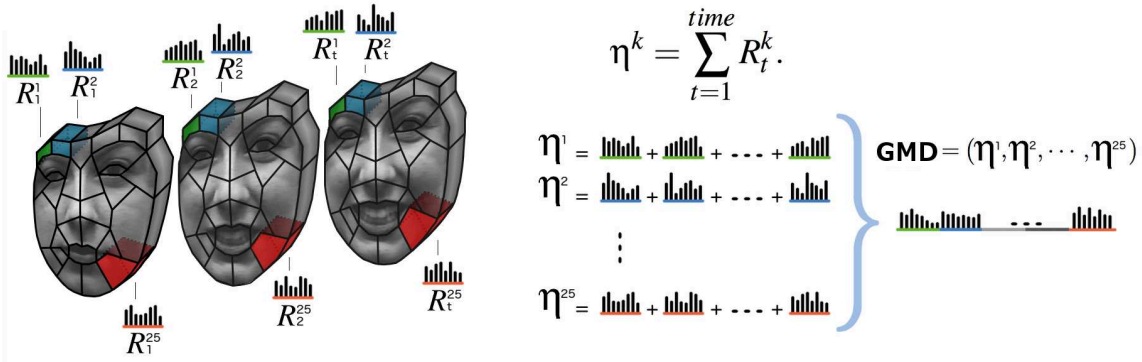


FIGURE 5.4 – Construction du vecteur de caractéristiques de mouvement.

Dans la section suivante, nous détaillons comment le vecteur de caractéristiques géométriques est construit à partir du modèle de segmentation faciale proposé.

5.3.2 Vecteur de caractéristiques géométriques

Le vecteur caractérisant la géométrie du visage consiste à extraire les informations relatives à la géométrie des 25 régions du visage. Le vecteur de caractéristiques d'une région correspond donc à la concaténation de la longueur des différents côtés qui la compose et des angles entre chaque côté. Chaque région est représentée par un vecteur de caractéristiques nommé Gr représensant sa géométrie.

Les 25 vecteurs correspondants aux différentes ROIs du modèle facial sont concaténés dans un vecteur GRD, où GRD = $(Gr_1, Gr_2, ..., Gr_{25})$ représente le vecteur de caractéristiques géométriques d'une expression faciale. La dimension du vecteur GRD dépend directement du nombre de côtés et d'angles composant l'ensemble des régions du modèle facial considéré. Dans notre cas, la dimension est égale à 318. Le processus de construction du vecteur GRD est illustré dans la Figure 5.5.

Dans la section suivante, nous présentons comment le vecteur de caractéristiques de mouvement et de géométrie sont réunis afin de construire un vecteur hybride regroupant les informations de ces deux vecteurs.

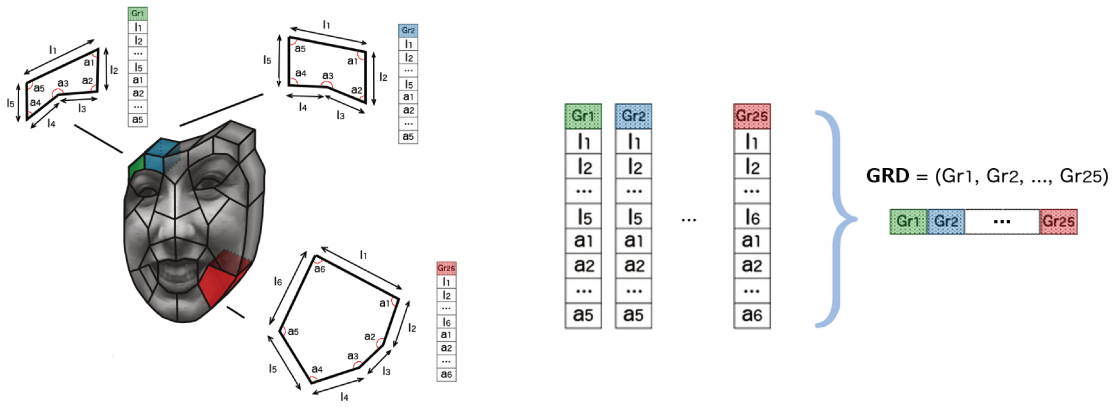


FIGURE 5.5 – Construction du vecteur de caractéristiques géométriques.

5.3.3 Fusion des vecteurs de caractéristiques

Plusieurs études présentent l'apport d'information lors de la fusion de données provenant de différents plans. La fusion des informations permet d'améliorer la prise de décision [96] dans plusieurs domaines d'activité, notamment dans l'analyse faciale, où l'apparence et la géométrie sont caractérisées de façon concomitante pour améliorer les performances d'analyse. Il existe deux types de fusion, la fusion dite de bas niveau, où la fusion se fait au niveau des descripteurs et la fusion dite de haut niveau, où la fusion se fait au niveau de la classification.

Dans notre approche, nous choisissons une fusion au niveau de l'espace des caractéristiques. Cela consiste à concaténer, avant l'étape d'apprentissage, toutes les caractéristiques en un seul vecteur. Ce vecteur est ensuite fourni en entrée d'un classifieur. Une fois les vecteurs fusionnés, le vecteur résultant est classifié pour prendre une décision. La fusion de caractéristiques a l'avantage de ne nécessiter qu'une seule phase d'apprentissage automatique pour l'ensemble des modalités. Celle-ci permet au classifieur d'apprendre des régularités dans un espace multimodal.

Ainsi, le vecteur de caractéristiques final est obtenu en appliquant une fusion au niveau des vecteurs de caractéristiques, entre le vecteur de mouvement GMD extrait à partir du LMP et le vecteur de géométrie GRD. Une représentation de la fusion de ces deux vecteurs de caractéristiques est illustrée dans la Figure 5.6.

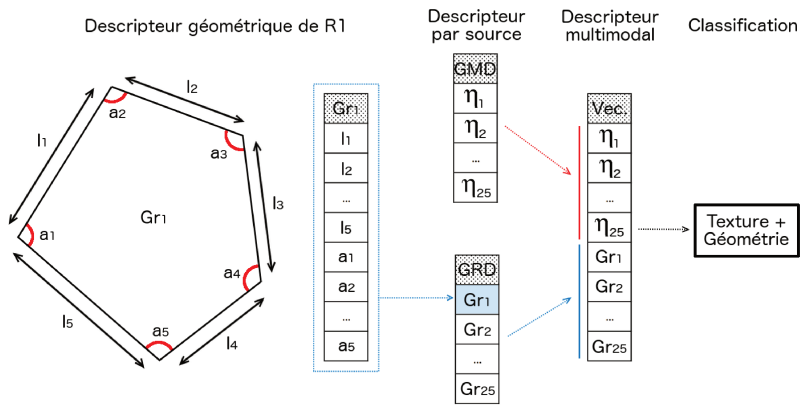


FIGURE 5.6 – Concatenation du vecteur de mouvement et du vecteur de géométrie.

Dans la section suivante, nous présentons notre processus générique de reconnaissance des expressions faciales qui regroupe l'ensemble des contributions proposées précédemment.

5.4 Processus générique de reconnaissance

Une représentation du processus de reconnaissance d'expressions faciales est illustré à la Figure 5.7. Avant de caractériser le mouvement facial, l'ensemble des visages de la séquence d'images est normalisé en utilisant une transformation géométrique 2D par rapport à la position du centre des yeux. Ce pré-traitement permet d'aligner les visages tout au long de la séquence d'images afin de réduire les invariances de poses.

Le mouvement facial est ensuite extrait avec la méthode de flux optique dense proposée par Farnebäck [34]. Nous avons choisi d'utiliser cette méthode car le mouvement calculé n'est pas affecté par un algorithme de lissage et que le calcul est relativement rapide. Une fois le mouvement calculé globalement sur le visage, le modèle de segmentation faciale illustré dans la Figure 5.3 est utilisé pour filtrer localement le mouvement au sein des 25 régions, en y appliquant des LMPs en leur centre respectif. Afin de prendre en considération la contrainte de dynamicité du mouvement dans le temps (voir hypothèse D), la distribution du mouvement de chaque région est cumulée dans le temps. Enfin, chaque vecteur cumulé correspondant aux 25 régions est concaténé dans un vecteur global de caractéristiques qui sera par la suite associé à une expression faciale. Concernant l'étape de classification, nous utilisons un processus de classification surpervisée à l'aide d'un SVM (norme euclidienne) avec un noyau RBF (utilisation de LibSVM [17]). Nous ap-

pliquons un protocole de validation croisée en considérant pour chaque dataset une décomposition en 10-folds. L'apprentissage est multi-classes, où chaque classe représente une expression. Le nombre de classe peut varier en fonction des bases de données ou des expérimentations réalisées dans les sections suivantes.

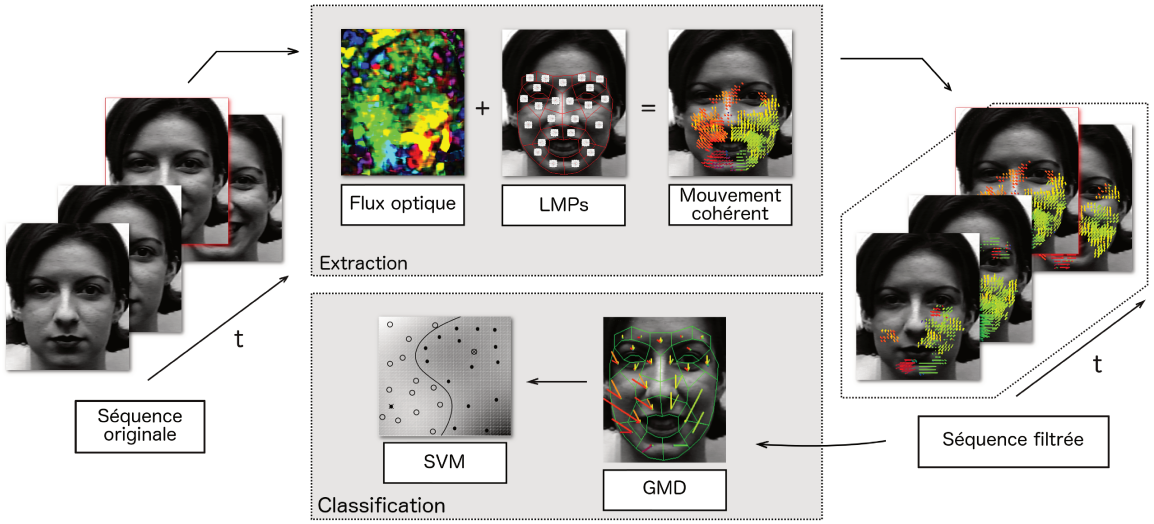


FIGURE 5.7 – Représentation du processus d'analyse pour la caractérisation des expressions (CK+).

Dans la section suivante, nous évaluons le processus de reconnaissance d'expressions faciales proposé. Nous nous intéressons particulièrement à la manière dont les propriétés d'élasticité de la peau et les contraintes musculaires prises en compte dans le filtrage et la caractérisation du mouvement sont adaptées à l'analyse des macro et des micro expressions.

5.5 Evaluation sur les micro et les macro expressions

Dans cette section, nous évaluons les performances de notre processus d'analyse sur plusieurs bases d'apprentissage composées d'expressions faciales de faible et forte intensités. Nous sélectionnons des bases couramment utilisées dans la littérature, afin de pouvoir positionner les performances de notre descripteur pour l'analyse des macro (CK+, Oulu-CASIA et MMI) et des micro (CASME II et SMIC) expressions. Les expérimentations et les différentes spécificités propres à chacune de ces bases tendent à couvrir plusieurs problématiques de l'analyse des expressions en condition naturelle : mouvements de tête, lumière naturelle ou infrarouge, intensité des expressions variable.

Nous commençons par présenter les différentes bases d'apprentissage que nous avons sélectionnées afin d'analyser les macro et les micro expressions. Puis, nous nous intéressons à la paramétrisation du descripteur de caractérisation du mouvement (dimension des régions, nombre de propagations, ...) pour estimer l'impact des différents paramètres sur les performances de la reconnaissance d'expressions faciales. Notamment, nous analysons l'impact des différents paramètres en fonction des caractéristiques liées aux macro et aux micro expressions. Ensuite, nous évaluons les performances de notre descripteur pour caractériser les macro et les micro expressions. Pour montrer la généricité de notre descripteur, nous sélectionnons des bases d'apprentissage qui proposent des données acquises dans différentes conditions : présence de petites variations de pose, changements lumineux. Enfin, nous proposons d'étendre notre analyse en mesurant la capacité de notre méthode à différencier une même expression à différents niveaux d'intensité.

5.5.1 Bases d'apprentissage

Pour évaluer les performances de notre descripteur, nous avons sélectionné plusieurs bases d'apprentissage disposant de données acquises dans différentes conditions mettant en évidance différentes problématiques comme la variation de pose et les changements lumineux. Deux bases sont utilisées pour caractériser les micro expressions (CASME II et SMIC) et trois pour de caractériser les macro expressions (CK+, Oulu-CASIA et MMI).

CASME II inclut 247 séquences d'images représentant des micro expressions, provenant de 26 participants. Les données sont réparties en cinq classes : joie (33 seqs.), dégoût (60 seqs.), surprise (25 seqs), répression (27 seqs.) et une classe "autres" (102 seqs.). La classe "autres" regroupe l'ensemble des données où les séquences d'images ne permettent pas de distinguer clairement l'une des quatre expressions. Les micro expressions sont enregistrées à l'aide d'une caméra haute fréquence (200 fps) dans un contexte d'acquisition contrôlé (pose fixe, lumière homogène).

SMIC est divisée en trois jeux de données : (1) le jeu de données HS correspond à des séquences acquises par une caméra haute fréquence (100 fps) et inclut 164 séquences provenant de 16 participants; (2) le jeu de données VIS contient des données acquises

par une caméra couleur standard de 25 fps ; (3) le jeu de données NIR contient des données acquises par une caméra infrarouge de 25 fps. Les données HS sont utilisées pour enregistrer l'ensemble des séquences, tandis que les données VIS et NIR contiennent uniquement les séquences correspondant aux huit derniers participants (77 séquences). Chaque jeu de données contient des micro expressions allant de l'état neutre à l'apex. Les séquences sont labelisées selon trois classes : positive, surprise et négative.

CK+ contient 593 séquences d'images représentant des expressions actées (volontaires) provenant de 123 participants de genre et d'ethnie différents. Les séquences couvrent sept expressions (colère, mépris, dégoût, peur, joie, tristesse et surprise). Au cours de chaque séquence d'images, les expressions commencent dans un état neutre et se terminent à l'instant où l'expression est à sa plus forte intensité (apex). Les données fournies dans cette base d'apprentissage permettent d'analyser les expressions faciales dans de parfaites conditions (pose fixe, lumière homogène, visage frontal). Le challenge apporté par cette base repose essentiellement dans le fait que les séquences temporelles d'activation des expressions sont d'une durée variable. La longueur moyenne des séquences est de 17.8 ± 7.42 images (les séquences allant de 4 à 66 images). Le fait que les séquences ne suivent pas le même patron d'activation temporelle, cela permet d'approcher les conditions d'interaction naturelle où chaque personne exprime ses émotions de manière primitive et non standardisée.

Oulu-CASIA inclut 480 séquences provenant de 80 participants, acquises dans trois conditions d'illumination différentes : normal, clair et sombre. Les données sont labélisées en fonction des six expressions universelles (joie, tristesse, colère, dégoût, surprise et peur). Chaque séquence commence par une expression neutre et se termine à l'apex de l'expression. Cette base d'apprentissage permet de tester la précision des systèmes en présence de différentes conditions d'illumination. Les séquences sont analysées en simultané par une caméra infrarouge permettant de fournir des données invariantes à la luminosité.

MMI contient 213 séquences provenant de 30 participants ayant comme instruction de reproduire les six expressions universelles (joie, tristesse, colère, dégoût, surprise et peur).

Comparée aux bases de données CK+ et Oulu-CASIA, cette base d'apprentissage fournie des données plus proches des conditions naturelles, où les participants sont libres de bouger la tête et leurs expressions sont plus spontanées. Ces données tendent à augmenter la robustesse des systèmes dans des conditions d'acquisition moins contrôlées.

Dans la section suivante, nous évaluons l'impact des paramètres impliqués dans la construction du LMP sur deux bases d'apprentissage : CK+ et CASME II. Plus spécifiquement, nous discutons du choix des différents paramètres et de leur homogénéité en fonction des caractéristiques liées aux macro et aux micro expressions. Nous choisissons volontairement ces bases car elles permettent de définir les paramètres optimaux de notre méthode sur les macro et les micro expressions, tout en s'affranchissant du bruit occasionné par les mouvements de tête et les changements lumineux.

5.5.2 Définition des paramètres optimaux

Dans cette section, nous passons en revue les différents paramètres permettant de configurer le processus du filtrage du mouvement. Nous structurons la suite de cette section en fonction des différentes hypothèses émises dans la section 4.2. Nous analysons l'impact de chaque paramètre en fonction de l'hypothèse qu'il permet de vérifier en appliquant un protocole de validation croisée en 10-fold à l'aide d'un SVM avec un noyau RBF. Nous rappelons les différentes hypothèses de cohérence :

(A) Il existe une cohérence locale de la distribution du mouvement en termes de magnitude et de direction (λ - la dimension d'une région locale au sein du LMP, B - le nombre de bins des histogrammes de mouvement).

(B) Le mouvement d'une petite région induit par un muscle facial est contraint à suivre une direction principale directement liée à l'activation musculaire (α - le seuil d'intensité accepté pour caractériser une direction principale, M - le seuil de l'étendue d'une direction principale accepté, V - le seuil de la variation de l'intensité entre deux bins successifs accepté pour caractériser une direction principale).

(C) La propagation du mouvement au sein du visage est proportionnelle à l'intensité d'une contraction musculaire (β - le nombre d'itérations de l'analyse de la propagation du mouvement).

(D) La magnitude et la direction de la déformation se propagent de manière continue

dans le voisinage d'un muscle facial (Δ - la distance de recouvrement entre deux régions voisines, ρ - le seuil de similarité entre deux distributions connexes).

L'hypothèse A vérifie qu'il existe une cohérence locale du mouvement en termes de magnitude et de direction. Deux paramètres sont à considérer au sein de notre LMP afin de caractériser le mouvement en fonction de ce critère. Les paramètres en question sont λ et B.

λ : ce paramètre représente la dimension d'une région locale au sein d'un LMP. Il est important que la dimension des régions s'adapte dynamiquement en fonction de la taille du visage analysé. Cela garantit de conserver une région faciale de taille similaire malgré la distance qui sépare le visage du capteur d'acquisition. Comme illustrée dans la Figure 5.8, la valeur idéale correspondant à λ est égale à 3 pourcents de la taille du visage analysé (distance estimée en fonction des landmarks), que ce soit pour la macro ou la micro expression. Une région doit avoir une taille adaptée pour pouvoir s'appuyer sur la cohérence de la distribution du mouvement facial tout en disposant de suffisamment d'informations pour renforcer le critère. Inversement, une trop grande région se caractérise par une distribution du mouvement éparse, et ne permet plus de vérifier l'hypothèse de déformation locale.

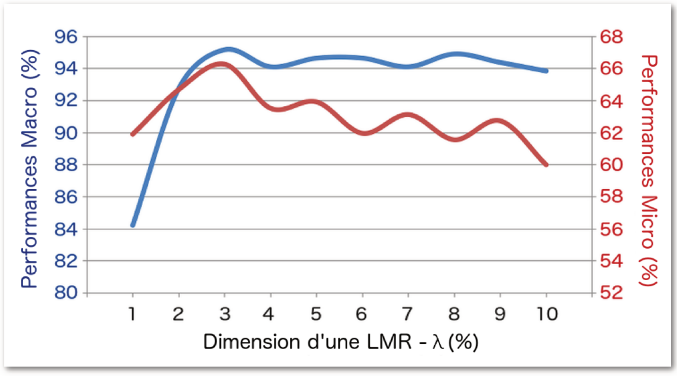


FIGURE 5.8 – Taux de reconnaissance en fonction de la dimension des LMR (λ) au sein du LMP.

B : ce paramètre représente le nombre de bins caractérisant la distribution locale au sein d'une région. Lorsque la dimension d'une région λ est petite, il est plus intéressant de prendre un petit nombre de bins. Ceci permet de distinguer plus facilement la direction principale. Comme illustré dans la Figure 5.9, les meilleures performances ont été obtenues en analysant la distribution du mouvement sur 9 ou 12 bins, ce qui correspond à une segmentation de 40° et de 30° respectivement. La prise en compte d'un grand nombre de bins fait que les distributions locales sont trop éparses, ce qui réduit la similarité entre deux distributions connexes. À l'inverse, un petit nombre de bins fait que la distribution n'est plus assez représentative pour dissocier deux mouvements relativement proches.

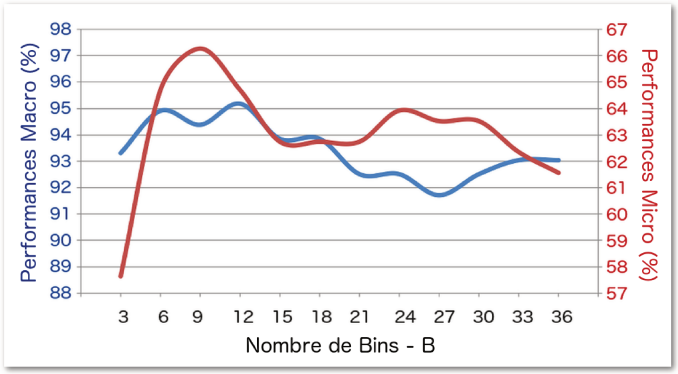


FIGURE 5.9 – Taux de reconnaissance en fonction du nombre de bins.

L'hypothèse B vérifie que le mouvement au sein d'une petite région induit par un muscle facial est contraint à suivre une direction principale directement liée à l'activation musculaire. Trois paramètres sont à considérer au sein de notre LMP pour définir si la distribution du mouvement local correspond aux critères souhaités. Nous appliquons différents seuils sur la distribution du mouvement afin de vérifier si le mouvement est cohérent. Les paramètres de seuillage en question sont α, M et V.

α : ce paramètre représente le seuil d'intensité accepté pour conserver une direction principale. L'intensité des directions au sein de la distribution du LMP est calculée à l'aide des différents poids associés aux couches de magnitudes. Au vu de la Figure 5.10, nous constatons que le fait d'appliquer un seuil minimum permet d'améliorer les performances. En effet, on prend uniquement en compte les directions ayant une représentativité suf-

fisamment importante pour caractériser une direction principale. En revanche, un seuil trop élevé tend à réduire les performances. Ceci prouve qu'il est important de conserver également les informations contenues dans les directions de plus faibles intensités pour caractériser un mouvement facial. Cela est d'autant plus important en présence de micro expressions, où un seuil trop élevé a pour conséquence d'enlever la majorité des directions principales de la distribution.

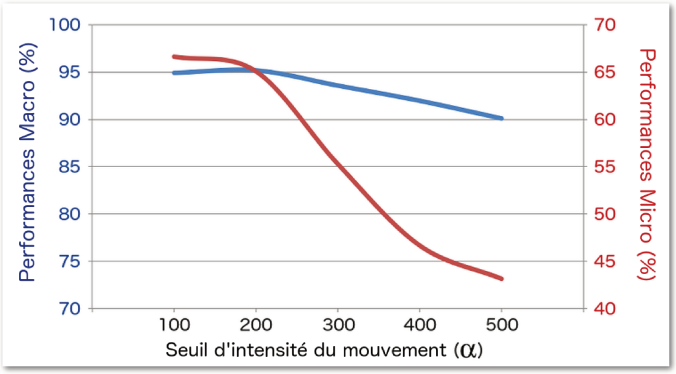


FIGURE 5.10 – Taux de reconnaissance en fonction du seuil appliqué à l'intensité.

M : ce paramètre représente le seuil de l'étendue du mouvement accepté pour caractériser une direction principale. Comme illustré dans la Figure 5.11, le seuillage appliqué à la densité de la direction augmente les performances jusqu'à une certaine valeur. En effet, un mouvement facial a tendance à s'étendre autour d'une direction dans un intervalle restreint. Ceci est directement lié aux contraintes de déformation de la peau et des muscles faciaux. Dans le cas contraire, un mouvement unidirectionnel ou un mouvement s'étendant sur plus de la moitié de la distribution est souvent caractéristique d'un bruit.

V : ce paramètre représente le seuil de la variation de l'intensité entre deux bins d'orientation successifs défini pour caractériser une direction principale. Pour vérifier l'influence de ce paramètre, nous diminuons progressivement la variation maximale tolérée (la pente) entre deux bins d'orientation successive. La Figure 5.12 illustre les taux de reconnaissance obtenus en présence de macro et de micro expressions. Au vu de ces résultats, nous pou-

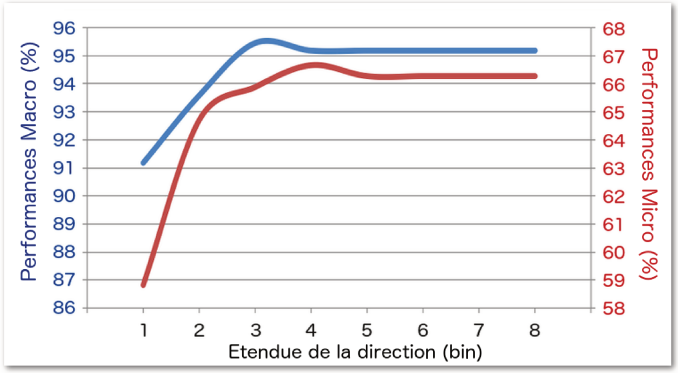


FIGURE 5.11 – Taux de reconnaissance en fonction du seuil de l'étendue du mouvement.

vons observer que l'influence de ce seuil n'a pas un impact très important sur les performances du LMP. Cependant, nous pouvons constater une légère augmentation des performances en appliquant un seuil de 95% pour les macro et micro expressions.

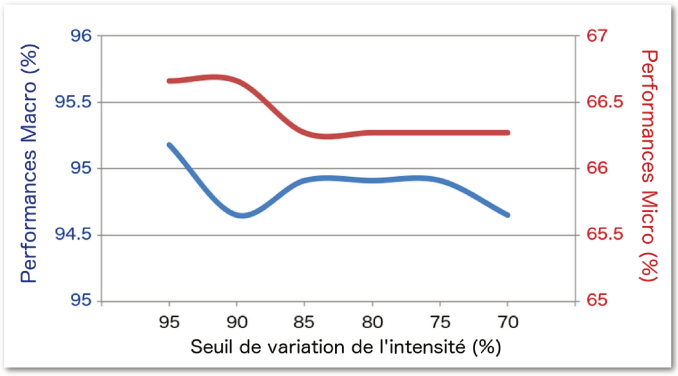


FIGURE 5.12 – Taux de reconnaissance en fonction du seuil de la variation de l'intensité entre deux bins d'orientation successive.

Les hypothèses C et D vérifient que le mouvement se propage sur le visage en présence d'une déformation faciale. Plus précisément, la propagation du mouvement au sein du visage est directement proportionnelle à l'intensité d'une contraction musculaire. La magnitude et la direction de la déformation se propagent de manière continue dans le voisinage d'un muscle facial. Trois paramètres sont à considérer au sein de notre LMP pour prendre en compte le critère de propagation du mouvement au sein du visage. Les différents paramètres en question sont Δ, ρ et β.

Δ, ρ : ces paramètres représentent le seuil de recouvrement entre deux LMR au sein d'un LMP (Δ) et le seuil de similarité entre les deux distributions de mouvement associées à ces régions (ρ). Afin de s'assurer de la continuité du mouvement (en termes de magnitude et de direction), il est important que les distributions de deux régions voisines se chevauchent. Comme illustré dans les tables de relation 5.13-1 et 5.13-2, le chevauchement entre deux régions permet d'augmenter significativement les performances. En effet, cela assure qu'il existe une cohérence du mouvement entre deux régions voisines et évite ainsi de considérer des mouvements n'ayant pas de cohérence avec le mouvement induit par un muscle facial. Nous faisons également varier le seuil de similarité, qui représente le coefficient de Bhattacharyya appliqué pour calculer la distance entre deux distributions connexes. Au vu des résultats obtenus, nous pouvons voir que les performances sont meilleures lorsque le recouvrement entre les distributions est compris entre 50% et 100%.

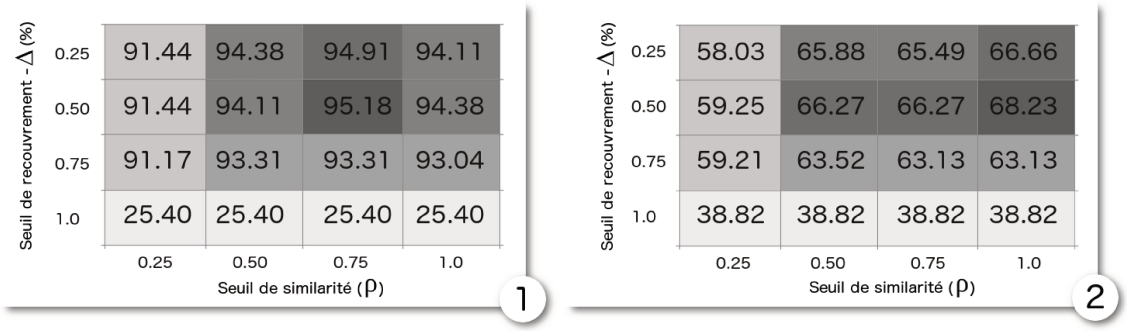


FIGURE 5.13 – Table de relation entre le seuil de recouvrement entre deux régions connexes et le seuil de similarité correspondant au coefficient de Bhattacharyya pour calculer l'indice de similarité de deux distributions voisines. Matrice 1 - macro expressions, Matrice 2 - micro expressions.

β : ce paramètre représente le nombre d'itérations de l'analyse de la propagation du mouvement autour de l'épicentre d'un LMP. Concernant le nombre d'itérations β, la Figure 5.14 permet de vérifier qu'il existe une contrainte sur la propagation du mouvement facial. Que ce soit pour l'analyse des macro ou des micro expressions, l'augmentation du nombre d'itérations permet d'augmenter les performances. La propagation du mouvement permet de mieux discriminer des mouvements relativement proches en cumu-

lant l'information pertinente, tout en réduisant le bruit. Il est important de noter qu'un nombre important d'itérations peut cependant réduire les performances. C'est notamment le cas en présence de macro expressions, où le mouvement a tendance à se répandre rapidement et à couvrir un large intervalle de direction.

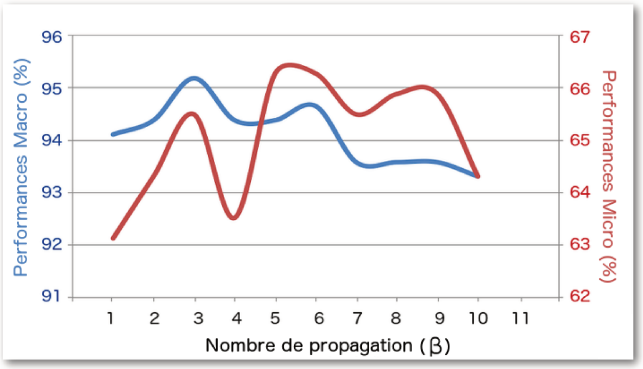


FIGURE 5.14 – Taux de reconnaissance en fonction du nombre d'itérations appliqué au sein du LMP.

L'analyse faite dans cette section tend à définir les meilleurs paramètres pour l'analyse des expressions faciales. Au vu de l'impact des différents paramètres de notre descripteur, il est important de noter que chaque paramètre a une influence spécifique sur les performances du système, en fonction des différentes hypothèses à considérer lorsque l'on caractérise un mouvement facial. De plus, il existe une certaine cohérence entre chaque paramètre du descripteur. Il est alors difficile d'identifier la meilleure paramétrisation en faisant varier un seul paramètre à la fois.

Jusqu'à présent, nous avons exploré les paramètres optimaux sur CK+ et CASME II afin de déterminer les maxima locaux correspondant à chaque paramètre. Connaissant les minima locaux des différents paramètres, nous déterminons la meilleure paramétrisation de manière empirique pour les différentes bases d'apprentissage utilisées dans nos expérimentations. Pour restreindre le nombre de configurations, nous faisons varier chaque paramètre en prenant pour chacun d'eux un intervalle de valeur voisinant leur maximum local. Ceci nous permet d'obtenir différentes configurations pour les macro et les micro-expressions. La Table 5.1 représente les différentes paramètrisations optimales obtenues sur les différentes bases d'apprentissage.

TABLEAU 5.1 – Paramètres optimaux pour les différentes bases d'apprentissage.

Bases		λ	Δ	ρ	$alpha$	M	V	β	bin
Micro	CASME II	4	0.5	0.75	100	4	5	6	9
	SMIC-HS	3	0.5	0.75	100	3	5	6	9
	SMIC-VIS	5	0.5	0.75	100	4	5	3	9
	SMIC-NIR	4	0.5	0.75	100	3	5	3	12
Macro	CK+	3	0.5	1	100	4	5	3	12
	MMI	3	0.5	1	100	4	5	6	12
	CASIA-VL	4	0.5	1	100	5	5	3	6
	CASIA-NI	5	0.5	0.75	100	5	5	6	9

Au vu des résultats de la Table 5.1, nous observons que les paramètres varient en fonction des différentes bases d'apprentissage. La légère variation observée pour certains paramètres s'explique notamment à cause des différentes conditions d'acquisition entre les différentes bases (distance de la caméra, résolution, taux de rafraichissement, ...). Il est intéressant de constater que certains paramètres (Δ, $alpha$ et V) restent identiques entre les différentes bases. Quant aux autres paramètres, il n'existe pas de grande différence entre ceux utilisés pour caractériser les macro et les micro expressions.

À présent que nous avons identifié les meilleures paramétrisations pour notre descripteur, dans la section suivante nous évaluons les performances du descripteur pour caractériser les micro expressions.

5.5.3 Reconnaissance des micro expressions

Dans cette section, nous évaluons les performances de notre descripteur à caractériser les micro expressions sur les bases CASME II et SMIC. Les résultats obtenus dans les sections suivantes sont obtenus en utilisant les paramètres inscrits dans la Table 5.1. Un protocole de validation croisée en 10-fold à l'aide d'un SVM avec un noyau RBF est employé pour classifier les expressions. Les informations géométriques ne sont pas prises en compte pour l'analyse des micro expressions, car les landmarks restent stables durant les séquences vidéo. Les landmarks ne permettent pas de renseigner de l'information complémentaire.

Experimentation sur CASME II

La Table 5.2 compare les performances de notre méthode par rapport aux méthodes les plus significatives de la littérature sur la base CASME II. Au cours de chaque séquence d'images, les expressions commencent dans un état neutre et se terminent dans un état d'offset. Pour notre évaluation, nous avons sélectionné uniquement la partie des séquences d'images correspondant à l'activation des expressions, c'est-à-dire de la phase neutre à l'apex. En effet, comme la méthode proposée s'appuie sur l'accumulation du mouvement au cours du temps, si nous conservons la phase d'offset de l'expression, des mouvements contraires vont se cumuler (onset vs offset). Dans la suite de l'évaluation, le mouvement facial est extrait en calculant le flux optique sur deux images successives.

TABLEAU 5.2 – Comparaison des performances sur CASME II *(* deep learning)*.

Méthode	Interpolat.	Magnifi.	Classes	Reco.(%)
Baseline [115]	✗	✗	5	63.41%
LBP-SIP [109]	✗	✗	5	67.21%
LSDF [108]	✓	✗	5	65.44%
TICS [107]	✓	✗	5	61.76%
STCLQP [44]	✓	✗	5	58.39%
STLBP-IIP [43]	✗	✗	5	62.75%
DiSTLBP-IPP [43]	✗	✗	5	64.78%
HIGO [64]	✓	✓	5	67.21%
MDMO [71]	✗	✗	4	67.37%
CNN + LSTM [57]	✗	✗	5	* 60.98%
CNN + AUs + LSTM [13]	✗	✗	5	* 59.47%
Local Motion Pattern (LMP)	✗	✗	**5**	**70.20%**

Au vu des résultats obtenus dans le Tableau 5.2, notre méthode obtient de meilleures performances par rapport aux approches proposées dans la littérature, incluant les méthodes traditionnelles et de deep learning. Etant donné que les données de la base CASME II sont acquises à l'aide d'une caméra haute fréquence (200 fps), certaines approches proposent d'interpoler les séquences vidéo [108, 64, 107, 44] afin de mieux percevoir les micro mouvements. En plus d'interpoler les séquences d'images, Li et al. [64] utilisent une

méthode de magnification, qui consiste à augmenter artificiellement la fréquence des vidéos, afin d'amplifier les micro mouvements. Cette technique convient bien à l'analyse des micro expressions. Cependant elle est mal adaptée en présence de macro expressions et de mouvements de la tête, car cela induit de fortes déformations faciales.

Bien que les méthodes de deep learning [57, 13] utilisent des données augmentées, leurs performances restent relativement basses comparées aux méthodes traditionnelles. Cela est probablement lié au fait qu'il n'existe pas de variation intra-classe suffisamment importante lorsque l'on caractérise les micro expressions.

Les résultats obtenus sur les données originales de la base CASME II montrent que notre méthode arrive à bien caractériser les micro expressions. Le descripteur proposé permet d'obtenir de meilleures performances que les approches récentes de la littérature, et cela sans amplifier le mouvement en magnifiant la fréquence ou en interpolant les séquences. Les résultats obtenus montrent que notre méthode donne de bonnes performances lorsque les conditions d'illumination sont homogènes. Dans la section suivante, nous évaluons la capacité de notre méthode à caractériser les micro expressions dans différentes conditions d'illuminations.

Expérimentation sur SMIC

La Table 5.3 compare les performances de notre méthode par rapport aux méthodes les plus représentatives de la littérature sur la base de données SMIC. Nous évaluons notre méthodes sur les trois jeux de données de la base : (HS) séquences enregistrées avec une caméra de 100 images/secondes, (VIS) séquences enregistrées par une caméra standard de 25 images/secondes et (NIR) séquences enregistrées par une caméra infrarouge de 25 images/secondes.

Au vu des résultats de la Table 5.3, notre méthode obtient de meilleures performances que les récentes approches de la littérature, incluant les méthodes traditionnelles et les méthodes de deep learning, sur l'ensemble des jeux de données, à l'exception du jeu de données SMIC-HS, pour lequel nous obtenons des résultats compétitifs. Comme dans l'expérimentation précédente, Li et al. [64] montrent qu'en amplifiant artificiellement la

TABLEAU 5.3 – Comparaison des performances sur SMIC *(* deep learning)*.

Méthode	Magnifi.	SMIC-HS	SMIC-VIS	SMIC-NIR
LBP-TOP [65]	✗	48.78%	52.11%	38.03%
Selective Deep features (CNN) [88]	✗	* 53.60%	* 56.30%	N/A
Facial Dynamics Map [113]	✗	54.88%	59.15%	57.75%
HIGO [64]	✗	65.24%	76.06%	59.15%
HIGO [64]	✓	68.29%	81.69%	67.61%
Local Motion Patterns (LMP)	✗	**67.68%**	**86.11%**	**80.56%**

fréquence des séquences, leur méthode obtient de meilleures performances. Cependant, cette technique n'est pas très adaptée à la reconnaissance des macro expressions.

Les résultats obtenus sur les différents jeux de données de la base d'apprentissage SMIC montrent que notre méthode obtient de bonnes performances à la fois en condition d'illumination naturelle (HS, VIS) et sur des données proches infrarouges (NIR). Il est intéressant de noter que notre méthode, s'appuyant sur le flux optique, semble mieux adaptée que les autres méthodes dynamiques sur des données infrarouges.

Au vu des deux expérimentations, nous avons montré que notre méthode permet de bien caractériser les micro expressions. Dans la section suivante, nous évaluons les capacités de notre méthode à caractériser les macro expressions.

5.5.4 Reconnaissance des macro expressions

Dans cette section, nous évaluons les performances de notre méthode à reconnaître les macro expressions en s'appuyant sur les bases CK+, Oulu-CASIA et MMI. Ces bases d'apprentissage permettent respectivement d'évaluer la capacité de notre méthode à être robuste dans différentes conditions lumineuses et en présence de petits mouvements de tête. Les résultats des expérimentations suivantes sont obtenus en utilisant les paramètres inscrits dans la Table 5.1. Un protocole de validation croisée en 10-fold à l'aide d'un SVM avec un noyau RBF est employé pour classifier les expressions. Pour chaque évaluation, nous reportons les résultats obtenus en exploitant uniquement l'information de mouvement et les résultats combinant l'information du mouvement et l'information géométrique.

Expérimentation sur CK+

La Table 5.4 compare les performances de notre descripteur par rapport aux approches les plus significatives de la littérature sur la base CK+. Bien que CK+ soit l'une des bases les plus utilisée dans la littérature, la comparaison avec les autres méthodes n'est pas simple, car les protocoles utilisés(nombre de séquences, nombre de classes) sont rarement identiques. Cela est principalement dû au fait de la présence de petites variations lumineuses ou de variations de pose présentes dans quelques séquences d'images, mais aussi à cause de l'absence de certaines annotations. De ce fait, nous utilisons les deux jeux de données les plus représentatifs pour évaluer les performances de notre méthode sur la base d'apprentissage CK+ :

- Le premier jeu de données contient 327 séquences d'images basées sur les annotations originales. Ce jeu de données contient sept classes associées aux six expressions universelles : colère (45), tristesse (28), joie (69), surprise (83), peur (25) et dégoût (59) auxquelles s'ajoute l'expression du mépris (18).
- Le second jeu de données contient 374 séquences d'images basées sur les séquences les plus représentatives. Ce jeu de données contient six classes correspondant aux expressions universelles : colère (37), tristesse (65), joie (95), surprise (83), peur (53) et dégoût (41).

Comparée aux méthodes proposées de flux optique [67, 103, 62] et aux autres méthodes traditionnelles [125, 32, 33, 70, 31], notre méthode, basée uniquement sur le flux optique, obtient des performances compétitives (96.94%). Ces résultats montrent que le fait de filtrer le mouvement facial en combinant la direction et la magnitude tout en s'appuyant sur les contraintes de la propagation du mouvement, permet d'améliorer la qualité du flux optique lorsqu'il est employé pour caractériser un mouvement facial.

Inspiré des approches hybrides, nous avons proposé de combiner l'information du mouvement à l'information géométrique du visage. Nous observons que la fusion de ces deux informations permet d'augmenter les performances du système (97.25%).

Les résultats des récentes méthodes de deep learning [72, 13, 122, 54] sont compa-

TABLEAU 5.4 – Comparaison des performances sur CK+ *(* deep learning)*.

Méthode	Mesure	Seq.	Classes	Reco.(%)
Spatial weight mask [67]	LOSO	442	6	92,50%
Optical flow and three MLPs [103]	train/test	415	5	93,27%
Sparse motion dictionary [62]	4-fold	N.A.	7	86,70%
Local Motion Patterns (LMP)	**10-fold**	**327**	**7**	**96.94%**
Local Motion Patterns (LMP)	**10-fold**	**374**	**6**	**96.26%**
LBP-TOP + Gabor [125]	LOO	309	6	95.80%
PHOG-TOP + Optical flow [32]	LOO	327	7	83.70%
MHI-OF + QLZM-MCF [33]	LOO	327	7	88.30%
Dis-ExpLet [70]	10-fold	327	7	95.10%
LBP-TOP [124] + manual alignment	10-fold	374	6	96.26%
Spatio-temporal RBM-based model [31]	10-fold	327	7	95.66%
CNN + Spatial features [72]	8-fold	327	7	* 86.67%
CNN + Spatial features [72]	8-fold	*augmentées*	7	* 96.76%
CNN + AUs + LTSM [13]	LOO	*augmentées*	7	* 98.62%
PHRNN-MCSNN [122]	10-fold	*augmentées*	7	* 98.50%
DTAGN (joint) [54]	10-fold	*augmentées*	7	* 97.25%
LMP + Géométrie	**10-fold**	**327**	**7**	**97.25%**
LMP + Géométrie	**10-fold**	**374**	**6**	**96.79%**

rables à notre méthode. Cependant, la comparaison des performances reste relativement délicate car les méthodes de deep learning nécessitent d'augmenter les données initiales en appliquant différentes opérations sur les données (rotation, symétrie). Alors que les méthodes traditionnelles comme notre méthode exploitent uniquement les données initiales de la base d'apprentissage. En utilisant uniquement les données initiales, Lopes et al. [72] montrent que leur méthode basée sur le deep learning obtient seulement 86.67%. Suite à l'augmentation synthétique des données, ils augmentent leurs performances de plus de 10% (96.76%). Au vu de la Table 5.4 les performances obtenues par notre méthode, sans augmenter les données, sont comparables aux récentes méthodes de deep learning, qui elles, se servent des données augmentées.

Nous avons montré que notre méthode obtient des performances compétitives sur la base CK+, où les expressions sont actées et où les conditions d'acquisition sont excellentes pour analyser les macro expressions (pose fixe, visage frontal, lumière homogène).

Dans la section suivante, nous évaluons la capacité de notre méthode à caractériser les macro expressions dans différentes conditions d'illumination.

Expérimentation sur Oulu-CASIA

La Table 5.5 compare les performances de notre méthode par rapport aux méthodes les plus représentatives de la littérature sur la base de données Oulu-CASIA, sur des données acquises par une caméra standard et par une caméra infrarouge. La majorité des méthodes s'évaluant sur Oulu-CASIA exploitent uniquement les données acquises par la caméra standard (VL). Certaines méthodes [51, 123] reportent leur performances obtenues sur les séquences acquises par la caméra proche infrarouge (NI).

TABLEAU 5.5 – Comparaison des performances sur Oulu-CASIA *(* deep learning)*.

Méthode	Mesure	Seq.	Classes	VL-Reco.(%)	NI-Reco.(%)
LBP-TOP [124]	10-fold	480	6	68.13%	-
CLM + 3D Landmarks [51]	LOSO	480	6	72.31%	71.59%
AdaLBP [123]	10-fold	480	6	73.54%	72.09%
LBP-TOP + Gabor [125]	10-fold	480	6	74.37%	-
Dis-ExpLet [70]	10-fold	480	6	79.00%	-
DTAGN (joint) [54]	10-fold	augm.	6	* 81.46%	-
PHRNN-MSCNN [122]	10-fold	augm.	6	* 86.25%	-
FN2EN [28]	10-fold	augm.	6	* 87.71%	-
Local Motion Patterns (LMP)	**10-fold**	**480**	**6**	**75.13%**	**81.88%**
LMP + Géométrie	**10-fold**	**480**	**6**	**84.58%**	**81.49%**

Au vu des résultats de la Table 5.5, notre méthode obtient de meilleures performances que les méthodes traditionnelles [124, 125, 70] et des résultats compétitifs par rapport aux récentes méthodes de deep learning [54, 122, 28]. Les performances obtenues sur les données infrarouges sont supérieures à celles obtenues par les autres méthodes (81.88%). Une fois de plus, la combinaison du mouvement et de l'information géométrique augmente significativement les performances (84.58%) sur les données acquises par la caméra standard (VL). Cependant, ce n'est pas le cas sur les données acquises par la caméra infrarouge. Cela est principalement dû au fait que la position des landmarks est moins précise sur ce type de données, ce qui détériore la qualité des informations géométriques.

Dans cette expérimentation, nous avons montré que notre méthode obtient de bonnes performances à la fois dans les domaines visible et proche infrarouge. Dans la section suivante, nous évaluons la capacité de notre méthode à caractériser les macro expressions en présence de petits mouvements de tête.

Expérimentation sur MMI

La Table 5.6 compare les performances de notre descripteur par rapport aux approches les plus significatives de la littérature sur la base MMI. Afin d'obtenir le même protocole de validation employé par les autres méthodes, nous utilisons les 205 séquences où le visage est frontal à la caméra. Pour chaque séquence, nous exploitons uniquement le segment allant de l'état neutre à l'apex de l'expression.

TABLEAU 5.6 – Comparaison des performances sur MMI *(* deep learning)*.

Méthode	Mesure	Seq.	Classes	Reco.(%)
LBP-TOP [124]	10-fold	205	6	59.51%
LBP-TOP + Gabor [125]	10-fold	203	6	71.92%
CSPL [126]	10-fold	205	6	73.53%
Dis-ExpLet [70]	10-fold	205	6	77.60%
DTAGN (joint) [54]	10-fold	augmentées	6	* 70.24%
Deep Neural Networks [82]	5-fold	augmentées	6	* 77.60%
PHRNN-MSCNN [122]	10-fold	augmentées	6	* 81.18%
Local Motion Patterns (LMP)	**10-fold**	**205**	**6**	**74.40%**
LMP + Géométrie	**10-fold**	**205**	**6**	**78.26%**

Comparé aux méthodes traditionnelles [124, 125, 126, 70], notre méthode, basée uniquement sur le flux optique, obtient des performances compétitives (74.40%). La combinaison des informations géométriques et de mouvement augmente significativement les performances et donne des résultats compétitifs (78.26%) par rapport aux méthodes de deep learning [122, 82, 54]. Cela est principalement dû à la présence de variations de pose dans les séquences, où le mouvement est plus sensible au bruit que la géométrie.

Nous avons montré, qu'en présence de petites variations de pose de la tête, notre mé-

thode obtient de bonnes performances sur la base d'apprentissage MMI. Dans la section suivante, nous faisons une synthèse des expérimentations réalisées sur les différentes bases d'apprentissage de micro et macro expressions afin de souligner les avantages de la solution unifiée apportée par notre méthode.

5.5.5 Synthèse des expérimentations sur les micro et macro expressions

La Table 5.7 rassemble les résultats obtenus par les méthodes les plus représentatives de la littérature sur les différentes bases d'apprentissage de micro et macro expressions.

TABLEAU 5.7 – Synthèse des performances sur les différentes bases d'apprentissage de micro et macro expressions *(* deep learning)*.

	Micro expression				Macro expression			
Méthode	CASME II	SMIC			CK+	CASIA		MMI
		HS	VIS	NIR		VL	NI	
LBP-TOP [124]	-	-	52.11%	-	96.26%	68.13%	-	59.51%
LBP-TOP + Gabor[125]	-	-	-	-	95.80%	74.37%	-	71.92%
Dis-ExpLet [70]	-	-	-	-	95.10%	79.00%	-	77.60%
HIGO + magnification [64]	67.21%	68.29%	81.69%	67.61%	-	-	-	-
* CNN + LSTM [13]	59.47%	-	-	-	98.62%	-	-	-
* PHRNN-MSCNN [122]	-	-	-	-	98.50%	86.25%	-	81.18%
Local Motion Pattern	**70.20%**	**67.68%**	**86.11%**	**80.56%**	**97.25%**	**84.58%**	**81.46%**	**78.26%**

Les méthodes de deep learning obtiennent à ce jour les meilleures performances sur les bases d'apprentissage de macro expressions. Cependant, ils n'arrivent pas à obtenir des performances similaires sur les micro expressions. Pour améliorer leurs performances, ces méthodes ont besoin de cas réels servant d'exemples pour leur apprentissage. Ces cas doivent être d'autant plus nombreux que le problème est complexe et que sa topologie est peu structurée. Cela fonctionne très bien dans le cadre des macro expressions, où l'augmentation artificielle des données d'apprentissage suffit à enrichir la connaissance du réseau. Cependant, en présence de micro expressions, il n'existe pas de variations intra-classes suffisamment importantes, ce qui tend à réduire les performances de ces systèmes.

Bien que notre méthode n'obtienne pas les meilleures performances sur l'ensemble des bases d'apprentissage, elle permet d'obtenir des résultats compétitifs aussi bien pour caractériser les micro et les macro expressions en s'appuyant sur le même système d'analyse (descripteur, modèle de segmentation). Au vu des différents résultats, notre méthode obtient les meilleures performances sur des données acquises par une caméra proche infrarouge, aussi bien sur les macro (81.46%) et les micro (80.56%) expressions. La combinaison de l'information de mouvement et l'information géométrique permet de renforcer la robustesse de notre méthode en présence de petites variations de pose du visage.

Sachant que notre méthode permet de caractériser uniformément les macro et les micro expressions, nous évaluons, dans la section suivante, la capacité de notre méthode à dissocier une même expression sur plusieurs niveaux d'intensité.

5.6 Reconnaissance de plusieurs niveaux d'intensité

Lorsque l'on cherche à caractériser les expressions faciales dans un contexte d'acquisition naturelle, les intensités des mouvements faciaux peuvent être amenées à varier pour renforcer ou diminuer le degré d'une expression : une colère intense, un sourire subtil, etc. Nous évaluons la capacité de notre méthode à caractériser sur plusieurs niveaux d'intensité une même expression. Pour cela, nous réorganisons la base d'apprentissage CK+ afin de segmenter chaque expression en quatre niveaux d'intensité. Nous avons choisi d'utiliser la base d'apprentissage CK+ car les conditions d'acquisition sont parfaites (pose fixe, visage frontal, lumière homogène) ce qui garanti que les mouvements extraits correspondent uniquement aux expressions faciales.

Nous commençons par détailler comment les données de la base d'apprentissage CK+ sont segmentées afin de préparer cette évaluation. Ensuite, nous évaluons la capacité de notre méthode à reconnaitre une expression faciale en utilisant une fraction de l'information du mouvement. Cette expérimentation vise à analyser la capacité de notre méthode à anticiper la reconnaissance des expressions faciales. Puis, nous évaluons la capacité de notre méthode à distinguer une même expression sous plusieurs niveaux d'intensité. Enfin, nous faisons une synthèse sur les performances de notre méthode au vu des résultats obtenus.

5.6.1 Préparation des données

Afin de mesurer la capacité de notre méthode à caractériser des expressions faciales d'intensités variables, nous proposons une nouvelle décomposition des données de la base d'apprentissage CK+. Pour rappel, l'ensemble des séquences vidéo commencent dans un état neutre et se terminent à l'instant où l'expression est à sa plus forte intensité (apex). La Figure 5.15 illustre la décomposition temporelle de l'activation d'une expression faciale.

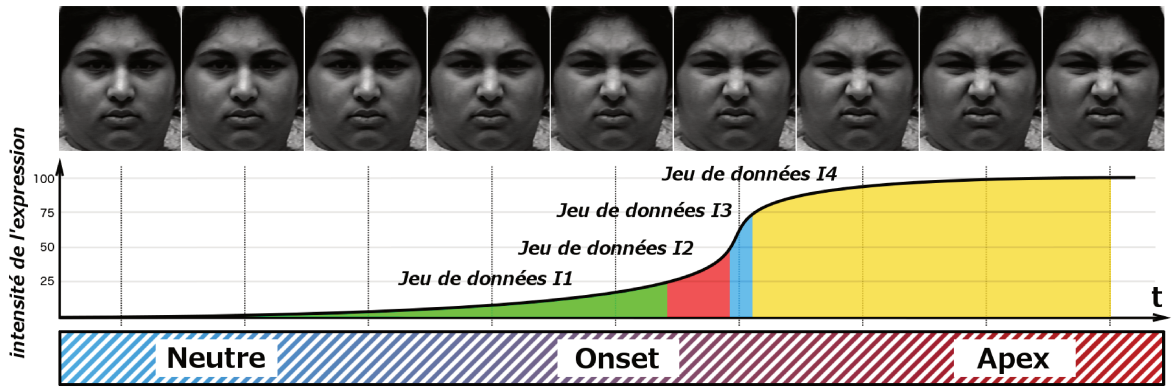


FIGURE 5.15 – Segmentation de l'activation temporelle des expressions de l'état neutre à l'apex (CK+).

Afin de simuler une même expression sur différents niveaux d'intensité, nous générons de nouvelles données à partir des séquences initiales en appliquant différentes segmentations temporelles. Etant donné que l'activation temporelle des expressions est continue (neutre vers apex), nous appliquons sur chaque séquence trois découpages correspondant aux différents niveaux d'intensité. Nous définissons trois points de coupe afin de décomposer une séquence en trois niveaux, correspondant à 25%, 50% et 75% de l'activation temporelle des expressions. Basé sur les 327 séquences couvrant 7 expressions faciales, nous générons quatre jeux de données en relation avec quatre niveaux d'intensité :

- Jeu de données I1 : sous-séquence contenant les images allant de l'image initiale (état neutre) à approximativement 25% de l'intensité totale de l'expression.
- Jeu de données I2 : sous-séquence contenant les images allant de l'image initiale (état

neutre) à approximativement 50% de l'intensité totale de l'expression.

- Jeu de données I3 : sous-séquence contenant les images allant de l'image initiale (état neutre) à approximativement 75% de l'intensité totale de l'expression.
- Jeu de données I4 : sous-séquence contenant les images allant de l'image initiale (état neutre) à approximativement 100% de l'intensité totale de l'expression.

Il est important de noter que la structure temporelle d'une séquence vidéo et la séquence d'activation d'une expression faciale sont deux choses différentes. En effet, atteindre 50% de la durée d'une séquence vidéo ne signifie pas que l'expression atteint 50% de son intensité. De ce fait, la construction des différents jeux de données I1 à I4 ne correspond pas à une segmentation naïve en relation avec la durée des séquences vidéo, mais en relation avec l'intensité du mouvement de l'expression entre l'état initial et son apex. Afin de définir les points de coupe pour une séquence, nous avons calculé l'intensité du mouvement cumulé lié aux expressions entre chaque image de la séquence en s'appuyant sur la variation des landmarks. La séparation d'une séquence en sous-séquence (I1, I2, I3 et I4) est appliquée en fonction d'un pourcentage défini par l'intensité totale du mouvement cumulé au sein de la séquence. Pour éviter de prendre en considération l'intensité induite par les petits mouvements de tête, nous appliquons un seuillage sur l'intensité qui permet de déduire si le mouvement analysé doit être pris en considération afin de définir un point de coupe dans la séquence. Ce seuil est déterminé en fonction de la distribution des différences entre les images successives des séquences, illustré dans la Figure 5.16. Comme la grande majorité des séquences ne comportent pas de mouvement de tête, les séquences comportant des mouvements de tête vont se retrouver dans la queue de la distribution.

La Figure 5.17 illustre la moyenne et l'écart type des découpages temporels des séquences appliquées pour chaque niveau d'intensité. L'axe des abscisses représente l'intensité du mouvement lié aux expressions (en pourcentage) et l'axe des ordonnées représente le pourcentage des données de la séquence. Par exemple, pour sélectionner une sous-sequence contenant 25% de l'intensité du mouvement liée aux expressions faciales (correspond au jeu de données I1), il faut appliquer un point de coupe à $31.46\% \pm 6.19\%$ de la durée totale des séquences. Au vu des écarts-types correspondant à chaque point

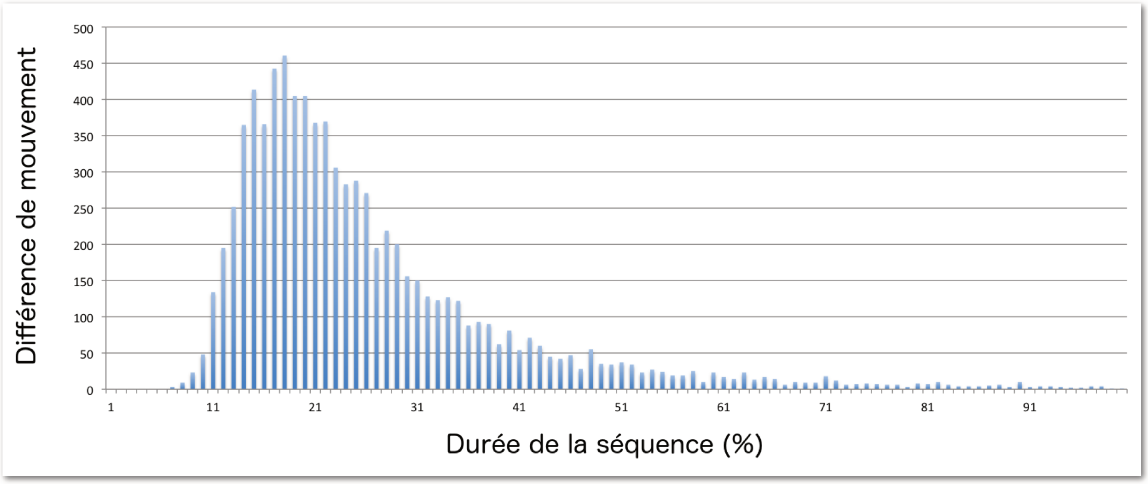
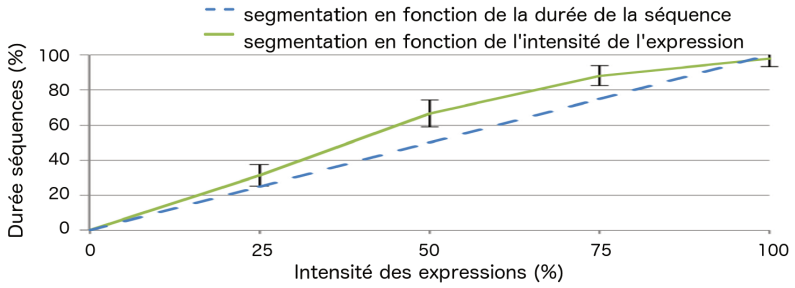


FIGURE 5.16 – Distribution des différences de mouvement entre les images successives des séquences.

de coupe, nous observons que l'activation temporelle des expressions peut varier fortement d'une séquence à une autre. Il est intéressant de constater que le point de coupe correspondant au jeu de données I4 (contienant 100% de l'intensité du mouvement lié à l'expression) ne se situe pas toujours à la fin de la séquence. Cela s'explique par le fait que l'apex de l'expression est parfois atteint avant la fin de la séquence. Au vu de la courbe représentant le pourcentage des données utilisées pour définir quatre jeux de données, les jeux I3 et I4 sont très similaires. Cela est dû au fait que l'activation des expressions faciales est généralement rapide pour passer de l'état neutre à 50% de l'expression, puis progresse légèrement pour atteindre l'apex.



	Jeu de données I1 (intensité 25%)	Jeu de données I2 (intensité 50%)	Jeu de données I3 (intensité 75%)	Jeu de données I4 (intensité 100%)
Moyenne	31.46%	66.61%	88.21%	97.88%
Ecart-type	±6.19%	±7.68%	±5.74%	±4.52%

FIGURE 5.17 – Segmentation des expressions par intensité du mouvement en fonction de la durée des séquences.

À partir de ces quatre jeux de données, nous évaluons par la suite la capacité de notre méthode à dissocier les expressions faciales selon plusieurs niveaux d'intensité. Pour cela, nous menons deux évaluations :

- La première mesure la capacité de notre méthode à reconnaitre les expressions faciales en utilisant une fraction de l'information du mouvement ;
- La seconde mesure la capacité de notre méthode à dissocier les expressions faciales sous différents niveaux d'intensité.

5.6.2 Analyse des expressions en utilisant une fraction du mouvement

Afin de mesurer la capacité de notre méthode à reconnaitre les expressions faciales en utilisant une fraction de l'information du mouvement, nous avons fait un apprentissage sur chaque jeu de données en utilisant un classifieur SVM et en appliquant une validation croisée en 10-folds. Les matrices de confusion obtenues pour les quatre jeux de données sont représentées dans la Figure 5.18. Au vu des matrices, plus la séquence d'activation des expressions est petite, plus les performances diminuent. Cependant, malgré le fait que le jeu de données I1 contient uniquement 25% de la séquence d'activation des expressions, notre méthode permet de reconnaitre les sept expressions faciales avec un taux de 87.16%. En utilisant 50% de la séquence d'activation (de l'état neutre jusqu'au milieu de l'onset), nous obtenons un taux de 95.11%. Avec 75% de la séquence d'activation des expressions, nous reconnaissons les sept expressions faciales à 96.69%. Les performances obtenues en utilisant 100% de la séquence d'activation sont très similaires avec un taux de 96.94%.

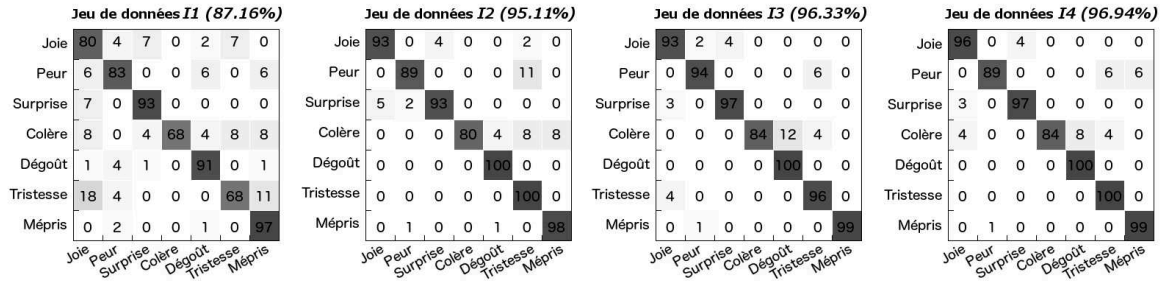


FIGURE 5.18 – Taux de reconnaissance (%) et matrices de confusion des différents jeux de données.

Cette expérimentation montre la capacité de notre méthode à reconnaître différentes expressions faciales en exploitant uniquement une fraction de l'information du mouvement. Les performances obtenues concernant les jeux de données I3 et I4 confirment la similarité de ces deux jeux de données observées précédemment dans la Figure 5.17. Dans la section suivante, nous évaluons la capacité de notre méthode à dissocier une expression faciale sous différents niveaux d'intensité.

5.6.3 Analyse des expressions sous différents niveaux d'intensité

Dans cette expérimentation, nous évaluons la capacité de notre méthode à dissocier une même expression sous différents niveaux d'intensité (forte, modérée, faible). Prenant en considération la grande similitude entre les jeux de données I3 et I4, nous choisissons de ne pas utiliser le jeu de données I3 dans la suite de cette expérimentation. Pour la suite, nous associons les différents jeux de données aux différents niveaux d'intensité. Nous considérons que le jeu I1, correspondant à 25% de la séquence d'activation, représente les expressions de faible intensité. Par analogie, le jeu de données I2 est associé aux expressions d'intensités modérées et le jeu de données I4 aux expressions de forte intensité.

Nous avons réalisé plusieurs regroupements en combinant les différents jeux de données afin d'évaluer la capacité de notre méthode à dissocier les sept expressions faciales sous différents niveaux d'intensité. Les résultats obtenus sur les différents regroupements sont représentés dans la Table 5.8. Pour les regroupements I1 vs. I2, I1 vs. I4 et I2 vs. I4, nous obtenons un jeu de données composé de 14 classes, où chaque expression est représentée par deux niveaux d'intensité en fonction du jeu de données associé. Par exemple, dans le regroupement I1 vs. I2, l'expression de joie est caractérisée par la classe "joie - intensité faible" et "joie - intensité modérée". Chaque colonne de la Table 5.8 représente les performances de notre méthode à dissocier une même expression sous différents niveaux d'intensité (apprentissage réalisé sur 2 ou 3 classes par expression). La dernière colonne représente la capacité de notre méthode à dissocier les sept expressions sous différents niveaux d'intensité (apprentissage réalisé sur 14 ou 21 classes en fonction du regroupement).

TABLEAU 5.8 – Taux de reconnaissance des expressions faciales sous différents niveaux d'intensité.

Regroupements		Joie	Peur	Surprise	Colère	Dégoût	Tristesse	Mépris	Tout
Jeu de donnée	Reco.(%)	83.33%	91.67%	90.68%	76.00%	94.93%	82.14%	86.75%	76.45%
I1, I2	# Classes	2	2	2	2	2	2	2	14
Jeu de donnée	Reco.(%)	96.67%	97.22%	96.61%	90%	99.28%	91.07%	95.18%	84.25%
I1, I4	# Classes	2	2	2	2	2	2	2	14
Jeu de donnée	Reco.(%)	67.78%	75.00%	72.03 %	66.00 %	81.88%	73.21 %	74.10 %	65.90%
I2, I4	# Classes	2	2	2	2	2	2	2	14
Jeu de donnée	Reco.(%)	68.15%	79.62%	73.45%	60.00%	84.54%	69.05%	71.89%	61.47%
I1, I2, I4	# Classes	3	3	3	3	3	3	3	21

Au vu des résultats obtenus dans la Table 5.8, nous montrons que notre méthode permet de dissocier une même expression sous plusieurs intensités lorsque la variation d'intensité du mouvement est suffisament importante (i.e. regroupement *I1 vs. I2* et *I1 vs. I4*). En ce qui concerne le regroupement *I2 vs. I4*, les taux de reconnaissance obtenus sont plus bas que pour les regroupements précédents. Cela s'explique par le fait que la différence entre ces deux niveaux d'intensité est moins nuancée. Cependant, les résultats obtenus pour les différentes expressions montrent que notre méthode arrive tout de même à dissocier ces nuances. Au vu des performances obtenues dans le regroupement *I1 vs. I2 vs. I4*, notre méthode arrive à dissocier de manière satisfaisante une même expression sur trois niveaux d'intensité.

Cette expérimentation montre la capacité de notre méthode à dissocier une même expression sous différents niveaux d'intensité. Dans la section suivante, nous faisons une synthèse des expérimentations réalisées sur l'évaluation des segments d'activation des expressions faciales.

5.6.4 Synthèse des expérimentations sur les segments d'activation

Ces deux évaluations mettent en évidence la robustesse de notre méthode à reconnaitre les expressions faciales malgré la prise en compte d'un nombre restreint d'informations. La première évaluation met en avant la capacité de notre méthode à identifier une expression faciale en exploitant une fraction de l'activation de l'expression. Au vu des performances obtenues, notre méthode atteint un taux de reconnaissance de 87.16% en utilisant uniquement 25% des séquences d'activation des 327 expressions analysées, réparties en sept classes. De ces résultats, nous pouvons supposer que notre méthode

puisse anticiper plus rapidement les expressions faciales. La deuxième évaluation montre la capacité de notre méthode à dissocier une même expression sous différents niveaux d'intensité. Ces résultats mettent en avant la robustesse de la méthode à nuancer plus subtilement les expressions.

5.7 Conclusion

Les contributions majeures de notre méthode s'articulent autour de trois axes :

- Premièrement, notre méthode repose sur un principe innovant permettant de mesurer les déformations faciales du visage en s'appuyant sur les propriétés physiques de l'élasticité de la peau. L'extraction spatio-temporelle du mouvement vérifie la conformité de la distribution du mouvement au sein d'une petite région du visage ainsi que la propagation de ce mouvement dans son voisinage local. Cela permet de conserver uniquement le mouvement lié aux expressions et enlever tout mouvement bruité dû aux conditions d'acquisitions (bruits de capteur, occultations, changements lumineux).
- Deuxièmement, notre méthode permet de caractériser de manière équivalente les micro et les macro expressions. En filtrant les mouvements discontinus au sein du visage, le mouvement lié aux expressions se distingue plus facilement du mouvement bruité, ce qui permet de caractériser à la fois les expressions subtiles et les expressions de forte intensité.
- Troisièmement, notre méthode tend à répondre à plusieurs défis scientifiques présents sur des données acquises en condition d'interaction naturelle. Nous obtenons de bonnes performances dans différentes configurations d'illumination (domaine visible et infrarouge), aussi bien pour les macro et les micro expressions. Malgré la présence de petits mouvements de tête, notre méthode arrive à caractériser les macro expressions. Enfin, notre méthode a la spécificité de pouvoir reconnaitre une expression faciale sous différents niveaux d'intensité, ce qui tend à permettre de reconnaitre les expressions plus rapidement, voir les anticiper.

Les expérimentations portées sur l'analyse des micro expressions montrent que la méthode proposée obtient de meilleures performances que les récentes méthodes de la littérature, à la fois sur la base CASME II (70.20%) et sur la base SMIC (86.11%). Sans augmentation artificielle des données initiales des bases d'apprentissage, comme le font

les méthodes de deep learning. Nous obtenons des résultats compétitifs pour la reconnaissance des macro expressions (97.25% sur CK+, 84.58% sur Oulu-CASIA et 78.26% sur MMI). De plus, en prenant uniquement en considération une fraction du mouvement lié aux expressions, notre méthode arrive à reconnaître les expressions faciales. Les performances obtenues en considérant 100%, 75%, 50% et 25% des séquences d'activation des expressions correspondent respectivement aux taux suivants : 96.94%, 96.69%, 95.11% et 87.16%.

Bien que dans l'état notre modèle permet de caractériser des micro et des macro-expressions, les visages analysés nécessitent d'être statiques et frontaux à la caméra. Dans ce cas, il est important de pouvoir adapter notre modèle pour apprendre à caractériser des expressions spontanées, où des problèmes de variations de pose et de larges déplacements viennent renforcer la difficulté de l'analyse. Pour palier à ce problème, des méthodes de normalisation géométrique du visage sont nécéssaires.

Dans l'objectif de pouvoir adapter notre descripteur à l'analyse des expressions faciales en présence de variations de pose et de larges déplacements, il est nécessaire de pouvoir identifier la méthode de normalisation géométrique la mieux adaptée. Dans le chapitre suivant, nous proposons un système d'acquisition innovant permettant de quantifier l'impact des méthodes de normalisation géométrique pour la caractérisation des expressions faciales.

Chapitre 6

Vers une adaptation aux problèmes de pose

« *Tout corps animé est*
un laboratoire de chimie :
Deus est philosophus
per quem. »

Voltaire

Sommaire

6.1 Introduction

La majorité des systèmes d'analyse d'expressions faciales sont conçues pour analyser les visages dans de bonnes conditions d'acquisition (visage fixe et frontal à la caméra). Cela assure que les caractéristiques visuelles sont parfaitement exploitables. Cependant, dans un contexte naturel, où les conditions d'acquisition changent, les visages ne sont pas directement exploitables sur la plupart des systèmes actuels. C'est principalement le cas en présence de variations de pose (déplacements 2D et 3D du visage par rapport à la caméra), où le visage a tendance à devenir partiellement occulté.

Au vu des récentes approches permettant de caractériser les expressions faciales, les approches dynamiques sont plus adaptées car elles permettent de détecter de subtils mouvements. Cependant, l'usage de ces approches nécessite que les textures dynamiques doivent être parfaitement segmentées localement et spatialement afin d'éviter d'induire des discontinuités de mouvement. Il est donc important de s'assurer que le visage est correctement aligné durant toute la séquence d'analyse en dépit de la présence de variations de pose ou de larges déplacements pour tirer tous les bénéfices de ces approches.

La solution utilisée pour adapter les approches dynamiques aux variations de pose et aux larges déplacements, est d'ajouter une étape de pré-traitement qui consiste à normaliser le visage afin d'amener le visage dans une configuration d'analyse idéale en corrigeant la transformation géométrique entre deux visages [124, 102]. Bien que ces solutions permettent d'améliorer les performances des systèmes, la normalisation induit des déformations néfastes sur la géométrie du visage, ce qui tend à modifier les expressions faciales [19]. La performance des approches dynamiques dépend directement de la qualité des méthodes de normalisation à réduire les invariances géométriques tout en conservant les déformations faciales induites par les expressions.

Pour répondre à ces défis, les récentes bases d'apprentissage [5, 80, 91] fournissent des données plus proches des conditions d'interaction naturelle, où les expressions sont spontanées (intensité du mouvement facial variable) et où les personnes sont libres de leurs mouvements. Bien que les données contenues dans ces bases d'apprentissage permettent de caractériser les expressions faciales en présence de variations de pose et de

larges déplacements, elles ne permettent pas de quantifier l'impact de la normalisation sur l'analyse des expressions faciales. Plus spécifiquement, la vérité-terrain du visage à reconstruire (visage frontal) n'est pas connue. Actuellement, les performances des méthodes de normalisation sont estimées à travers les résultats obtenus par le processus complet de l'analyse des expressions faciales : détection, normalisation, caractérisation, classification. De ce fait, il est difficile d'identifier si la perte des performances des systèmes d'analyse dans un contexte d'interaction naturelle est due à la normalisation ou à une autre étape du processus.

Afin de pouvoir quantifier la robustesse d'une méthode de normalisation et ainsi fournir une solution permettant d'améliorer ces méthodes, nous proposons un système d'acquisition innovant appelé SNaP-2DFe (Simultaneous Natural and Posed 2D Facial expression). Ce système permet de capturer simultanément un visage dans un plan fixe et dans un plan mobile (suit le déplacement de la tête). Grâce à cela, nous fournissons une connaissance du visage à reconstruire malgré les occultations induites par les rotations 3D (hors plan) de la tête. Le système proposé permet de répondre à deux objectifs :

- Connaissant le visage à reconstruire, nous pouvons quantifier la robustesse des méthodes de normalisation à reconstruire un visage en présence de variations de pose et de larges déplacements.
- Appliquée sur des visages expressifs où la pose n'est pas contrainte, nous pouvons quantifier la qualité des méthodes de normalisation à conserver les déformations géométriques liées aux expressions faciales. Ceci permettant d'identifier la méthode de normalisation la plus adaptée pour l'analyse des expressions faciales.

Dans la suite de ce chapitre, nous commençons par rappeler comment les problèmes de variations de pose et de larges déplacements sont illustrés à travers les récentes bases d'apprentissage. Puis, nous faisons une brève synthèse des récents systèmes d'analyse des expressions faciales employés sur ces bases d'apprentissage, en s'intéressant plus particulièrement aux méthodes de normalisation utilisées. Ensuite, nous détaillons notre système d'acquisition innovant pour capturer simultanément un visage dans un plan fixe et dans un plan mobile. Enfin, nous évaluons les méthodes de normalisation les plus représentatives de la littérature sur notre système d'acquisition. Nous concluons sur un positionnement par rapport aux différentes méthodes de normalisation employées de nos

jours pour caractériser les expressions faciales en présence de variations de pose et de larges déplacements.

6.2 Les défis posés par les bases d'apprentissage

La majorité des systèmes évaluent leurs performances sur des bases d'apprentissage où les expressions sont reproduites par des acteurs afin d'obtenir d'importantes déformations faciales [75, 87] et où les conditions sont contrôlées (pose fixe et frontale). Cependant, ces bases ne reflètent pas les données acquises dans des conditions d'interaction naturelle, où l'intensité des expressions est variable et des problèmes de variations de pose (VP) et de larges déplacements (LD) apparaissent. Pour cela, les bases d'apprentissage tendent à fournir des données de plus en plus proches des systèmes d'acquisition naturelle [6, 80, 91].

Les bases d'apprentissage les plus représentatives de la littérature sont illustrées dans la Figure 6.1. La Figure 6.1 montre que la complexité d'analyse des expressions faciales croît en fonction du type des expressions (actées ou spontanées) et des conditions d'acquisition. La présence de variations de pose et de larges déplacements au sein des données de chaque base est représentée par un indicateur variant de une à trois étoiles (✻).

	CK+	MMI	GEMEP	DISFA	SEMAINE	RECOLA
Variations de pose	-	✻	✻✻	✻	✻✻	✻✻✻
Larges déplacements	✻	✻	✻✻✻	✻✻	✻✻✻	✻✻✻

FIGURE 6.1 – Les bases d'apprentissage les plus représentatives pour l'analyse des expressions faciales.

Récemment, les méthodes de normalisation du visage basées sur des modèles déformables géométriques 2D et 3D sont généralement employées pour analyser les expressions faciales en présence de variations de pose et de larges déplacements [18]. En

dépit du fait que ces méthodes permettent d'augmenter les performances des systèmes d'analyse des expressions faciales en présence de variations de pose et de larges déplacements, les taux de reconnaissances restent relativement faibles par rapport aux taux obtenus dans des conditions contrôlées. La Figure 6.2 représente une synthèse de l'évolution des performances des méthodes de normalisation en fonction du taux de présence de variations de pose et de larges déplacements contenus dans les bases d'apprentissage. La synthèse proposée s'appuie sur les performances des systèmes de la Figure 3.11 du chapitre 3.

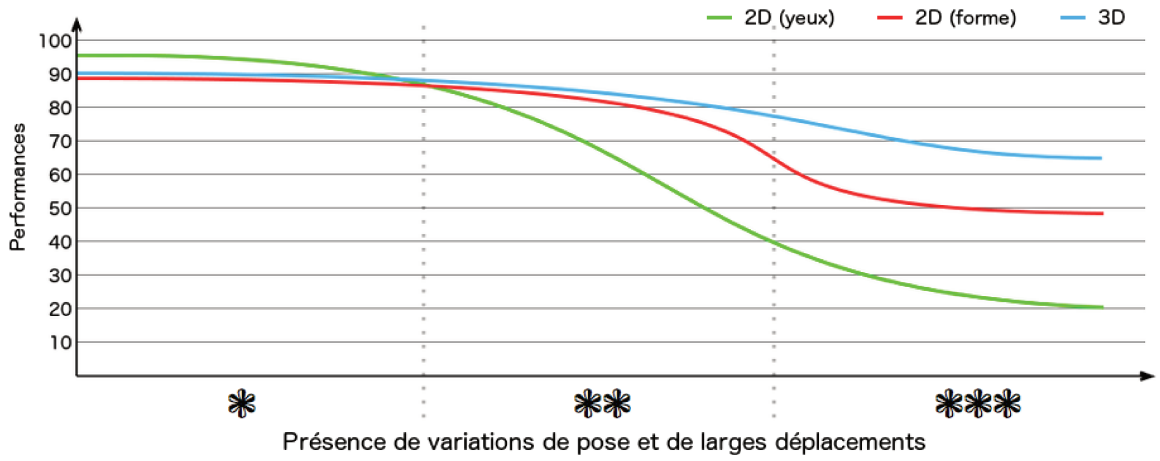


FIGURE 6.2 – Synthèse des performances des méthodes de normalisation en fonction du taux de présence de variations de pose et de larges déplacement dans les bases d'apprentissage.

Il est important de se questionner sur la capacité des méthodes de normalisation à corriger les transformations géométriques du visage tout en conservant les déformations induites par les expressions faciales. Afin de pouvoir quantifier la robustesse des méthodes de normalisation pour l'analyse des expressions faciales, nous proposons un système d'acquisition innovant qui permet de collecter simultanément des visages en présence et en absence du mouvement de la tête. Les détails de la conception de ce système sont expliquées dans la section suivante.

6.3 Système d'acquisition innovant (SNaP-2DFe)

Pour quantifier la robustesse des méthodes de normalisation dans la reconnaissance des expressions faciales, nous proposons un système d'acquisition innovant nommé Si-

multaneous Natural and Posed Facial expression (SNaP-2DFe). Chaque expression faciale est enregistrée simultanément à l'aide de deux caméras : l'une des caméras est fixée sur un casque, tandis que l'autre est placée sur un trépied devant la personne à une distance relativement proche. La caméra fixée sur le casque permet d'obtenir des données similaires à la base CK+ [75], où aucun mouvement de la tête n'apparaît. La caméra placée sur le trépied fournie des données similaires aux bases RECOLA [91] et SEMAINE [80], où des mouvements de tête apparaissent. Notre système d'acquisition permet ainsi de mesurer la qualité des méthodes de normalisation à corriger les mouvements liés au visage en se comparant aux données extraites simultanément sur la caméra fixée sur le casque.

Le casque est équipé de huit LEDs qui assure l'homogénéité de l'illumination du visage en dépit des mouvements de la tête. Il inclut également une carte "9DOF Razor IMU" développée par SparkFun, qui contient un gyroscope 3 axes, un accéléromètre, un magnétomètre et un micro contrôleur permettant de fusionner les données des différents capteurs. Une caméra est fixée sur le casque à cinquante centimètres du visage, maintenue par un rail en aluminium afin d'assurer la stabilité de la caméra. Un contrepoids est placé à l'arrière du casque pour améliorer le confort de l'utilisateur et pour garantir que le casque ne bouge pas lorsque l'utilisateur fait des mouvements de tête.

Chaque participant a comme instruction de porter un casque équipé d'une caméra (que nous nommerons Caméra 1 par la suite) et de s'asseoir devant un trépied situé à un mètre, où se trouve la caméra fixe (associée à la Caméra 2 par la suite). Chaque session est enregistrée à l'aide de deux caméras HD avec une résolution de 1280*720 pixels et d'un taux de rafraichissement de 30 img/s. Une session d'enregistrement en présence d'une expression de joie et d'un mouvement de tangage est illustrée dans la Figure 6.3. Le système d'acquisition est illustré dans la partie gauche de la Figure 6.3. La partie droite de la Figure 6.3 représente les images extraites simultanément des deux caméras. La première ligne représente les données extraites de la caméra fixée sur le casque et la deuxième ligne représente les données extraites de la caméra placée sur le trépied. Les courbes de la troisième ligne représentent les mouvements de la tête (lacet, tangage, roulis) extraits à l'aide d'un gyroscope placé sur le casque.

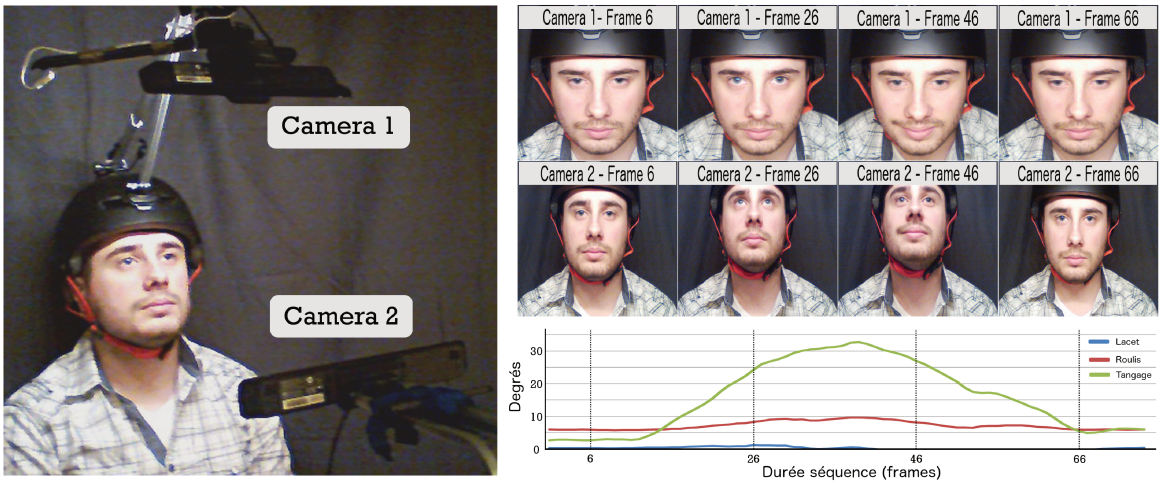


FIGURE 6.3 – Session d'acquisition d'une expression de joie en présence d'un mouvement de tangage (SNaP-2DFe).

Les données collectées par notre système sont composées de 840 séquences vidéo provenant de 10 sujets. Les sujets enregistrés ont signé un accord juridique donnant le droit d'exploiter les informations récoltées à des fins de recherche. Chaque vidéo correspond à une combinaison d'une expression faciale et d'un mouvement de tête. Dans chaque session, le sujet reproduit l'un de ces différents mouvements en suivant une animation projetée :

- une translation sur l'axe X (nommée T_x) qui simule un large déplacement.
- trois rotations (lacet, roulis et tangage) qui simulent des variations de pose.
- une session sans mouvement afin de récupérer uniquement l'expression faciale.
- un mouvement en diagonale combinant translation et rotation.

L'étendue des mouvements est relativement importante, avec des translations de plus de $150mm$ entre deux images successives et des rotations allant jusqu'à 40°. Pour chaque mouvement, le participant doit reproduire sept expressions faciales différentes (neutre, joie, tristesse, colère, peur, surprise et dégoût). Capturées simultanément par deux caméras, cela représente un total de 84 vidéos pour chaque participant (42 avec les mouvements de la tête et 42 sans).

Les participants ont pour instruction de produire une expression faciale à un moment précis dans la séquence vidéo et doivent conserver l'expression pendant une période donnée. Un témoin lumineux permet de donner les instructions au participant pendant la session d'enregistrement. Cela permet d'annoter les différentes étapes de l'évolution

d'activation des expressions faciales : neutre, onset, apex, offset et neutre. Bien que les expressions sont principalement actées (volontairement intensifiées), les participants ne sont pas des acteurs. Ainsi, les expressions capturées sont généralement spontanées, où leurs intensités sont variables et où différentes déformations faciales sont employées pour représenter une même expression.

Les données récoltées permettent ainsi d'évaluer les méthodes de normalisation en présence d'expressions faciales et de différents mouvements de tête, le tout dans des conditions d'acquisition contrôlées (lumière homogène, fond de la scène homogène). Dans la section suivante, nous évaluons les méthodes de normalisation les plus représentatives de la littérature en les appliquant directement sur les données collectées par notre système d'acquisition.

6.4 Evaluation des méthodes de normalisation du visage

Dans la suite de cette section, nous analysons la robustesse des méthodes de normalisation par le biais de deux expérimentations.

La première expérimentation consiste à vérifier si la normalisation appliquée au visage préserve la géométrie faciale. Pour cela, nous mesurons la capacité des méthodes de normalisation à reconstruire l'image frontale d'un visage. Nous évaluons les performances des méthodes de normalisation en calculant une distance de similarité entre le visage acquis par le casque (vérité-terrain), et le visage acquis par la caméra fixée sur le trépied après normalisation.

La deuxième expérimentation permet d'analyser la capacité des méthodes de normalisation à conserver les déformations induites par les expressions faciales. Dans cette expérimentation, nous construisons des modèles d'apprentissage d'expressions faciales distinctement sur les deux caméras. Puis nous évaluons les performances de reconnaissance de chacun de ces modèles. Cela permet d'identifier la perte de performances des méthodes de reconnaissance en présence de variations de poses et de larges déplacements. Nous construisons d'autres modèles en appliquant différentes méthodes de normalisation sur les séquences acquises par la caméra fixée sur le trépied afin d'identifier si

la normalisation permet de conserver les informations pertinentes du visage permettant de caractériser distinctement les expressions faciales.

6.4.1 Est-ce que la normalisation préserve la géométrie faciale ?

Pour cette évaluation, nous proposons d'étudier les méthodes de normalisation les plus représentatives de la littérature. Plus spécifiquement, nous considérons trois méthodes différentes : basée sur la position des yeux [39], basée sur la forme 2D du visage (landmarks - [38, 53]) et basée sur un modèle 3D déformable (3DMM - [127]). La Figure 6.4 permet de comparer visuellement les performances des méthodes de normalisation en présence de différents mouvements de tête provenant de la base SNaP-2DFe (c-à-d visage fixe, large déplacement sur X (T_x), roulis, lacet, tangage et diagonale).

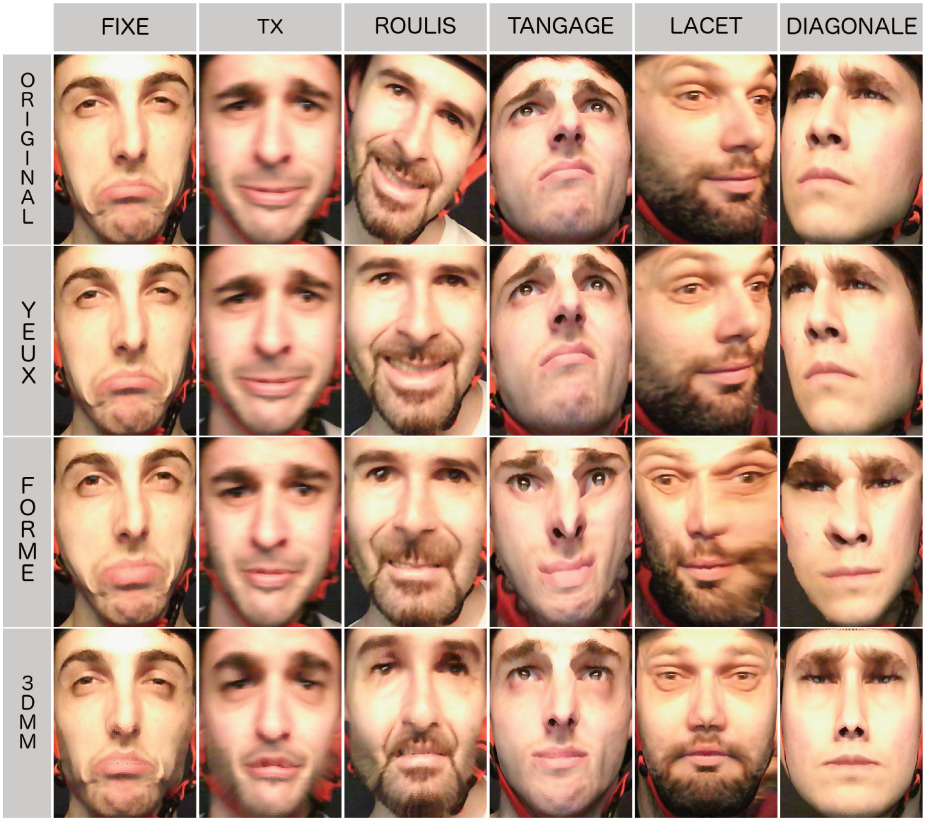


FIGURE 6.4 – Application de différentes méthodes de normalisation en présence de mouvements du visage (SNaP-2DFe).

Pour s'affranchir des mouvements de la tête dans le plan (T_x, roulis), la méthode de normalisation basée sur la position des yeux semble plus adaptée car cette méthode n'in-

duit aucune déformation sur la géométrie du visage. Cependant, en présence de fortes variations de poses hors plan (lacet, tangage, diagonale), la bonne localisation des yeux n'est plus garantie due aux occultations du visage, ce qui produit de fortes instabilités dans la normalisation. Dans ce contexte, les méthodes basées sur la forme 2D semblent plus robustes car elles s'appuient sur un nombre plus important de landmarks et sont donc moins assujetti aux occultations. Cependant, la normalisation basée sur la forme 2D du visage a tendance à induire des déformations au sein du visage (effet d'étirement de la texture au niveau des régions occultées) qui changent la géométrie faciale. La méthode de normalisation basée sur les modèles 3D montre une meilleure reconstruction du visage en présence de fortes variations de poses hors plan, bien que la géométrie du visage semble toutefois modifiée.

La géométrie du visage fournie de bonnes informations pour caractériser les déformations faciales, mais n'est pas adaptée pour caractériser les déformations subtiles, qui peuvent être observées par des changements au niveau de la texture. Afin d'évaluer la qualité des méthodes de normalisation, nous utilisons une mesure de similarité basée sur la texture, nommée SSIM (Structural Similarity Index Measure) [110]. SSIM mesure la similarité de structure entre deux images, plutôt qu'une différence pixel à pixel comme le fait un PSNR. SSIM compare les configurations locales des intensités de pixel qui ont été normalisées en fonction de la luminance et du contraste. La métrique SSIM est calculée sur plusieurs fenêtres d'une image. La mesure entre deux fenêtres x et y d'une dimension de $N * N$ est :

$$\mathrm{SSIM}(x,y) = l(x,y).c(x,y).s(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_x\sigma_y + c_2)(2cov_{xy} + c_3)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)(\sigma_x + \sigma_y + c_3)}. \tag{6.1}$$

où μ_x et μ_y représente la moyenne de x et y, σ_x^2 et σ_y^2 représente la variance de x et y, et cov_{xy} et la covariance de x et y. $c_1 = (k_1 \mathrm{L})^2$, $c_2 = (k_2 \mathrm{L})^2$, $c_3 = c_2/2$ permettent de stabiliser la division quand le dénominateur est très faible, où L représente la dynamique des valeurs des pixels, soit 255 pour des images codées sur 8 bits, $k_1 = 0.01$ et $k_2 = 0.03$.

La mesure de similarité SSIM est calculée pour chaque animation (mouvements) de

la base SNAP-2DFe après avoir appliqué différentes méthodes de normalisation, basées sur : un transformation affine rigide à partir de la position des yeux (Yeux), une transformation affine par morceaux déformables avec les moindres carrées [99] (Forme) et une transformation en plaquant la texture sur un modèle 3D déformable [127] (3DMM). Les résultats obtenus suite aux différentes normalisations sont reportés dans le Tableau 6.1.

Les performances des méthodes de normalisation basées sur les yeux et la forme 2D ont une moyenne très similaire (yeux - 55.16%, forme - 55.09%). La méthode basée sur les yeux est plus adaptée pour normaliser les visages en présence de mouvements dans le plan (fixe, T_x et roulis) tandis que la méthode basée sur la forme est plus adaptée en présence de variations hors plan (tangage, lacet et diagonale). Ces deux méthodes sont en général limitées car elles exploitent uniquement les informations visibles (2D). Bien qu'au vu des moyennes, les méthodes de normalisation basées sur les yeux et sur la forme aident à améliorer les performances, les écarts-types mettent en évidence d'importantes variations. En effet, les méthodes basées sur la forme et les yeux ne semblent pas être des bons candidats pour corriger les variations de pose. Quant à la méthode de normalisation basée sur un 3DMM, celle-ci donne les meilleures performances avec une moyenne de similarité de 61.33% et un écart-type relativement faible 4.38. Dans l'ensemble des cas, les méthodes de normalisation augmentent significativement les performances, cependant, elles n'assurent pas encore une similarité parfaite entre les visages acquis par les deux caméras.

Au vu des performances, les méthodes de normalisation ne parviennent pas à aligner parfaitement un visage sans induire de déformations. Suite à ce constat, il est important de vérifier que les déformations induites par les méthodes de normalisation ne détériorent pas les déformations induites par les expressions faciales. Dans la section suivante, nous analysons l'impact des méthodes de normalisation sur les performances des processus de reconnaissance des expressions faciales.

6.4.2 Est-ce que la normalisation préserve les expressions faciales ?

Dans cette évaluation, nous analysons l'impact de la normalisation sur l'analyse des expressions faciales en présence de variations de poses et de larges déplacements. Les

Normalisation		Fixe	Tx	Roulis	Tangage	Lacet	Diagonale	Total
Aucune	moyenne	53.30%	47.88%	48.68%	46.34%	46.54%	51.14%	48.98%
	écart-type	2.92	8.43	11.53	10.97	10.89	15.69	10.07
Yeux	moyenne	58.26%	55.06%	55.29%	54.57%	52.01%	55.79%	55.16%
	écart-type	3.46	6.31	5.76	13.50	8.99	14.37	8.73
Forme	moyenne	55.25%	53.49%	52.29%	57.40%	54.44%	57.67%	55.09%
	écart-type	6.02	6.95	8.60	17.89	10.25	17.35	11.17
3DMM	moyenne	**64.72%**	**61.81%**	**58.17%**	**60.52%**	**58.47%**	**64.29%**	**61.33%**
	écart-type	1.98	4.23	4.70	4.19	5.40	5.78	4.38

TABLEAU 6.1 – Indice de similarité par SSIM appliqué à différentes méthodes de normalisation en présence de variations de pose et de larges déplacements.

résultats présentés dans la suite de cette section ont été obtenus en utilisant un classifieur SVM [17] avec un noyau RBF (Radial Basis Function), en suivant un protocole de validation croisée (10-fold). Chaque expression est classifiée selon l'une des sept classes suivantes : neutre, joie, colère, surprise, tristesse, peur et dégoût. En supposant que les visages sont parfaitement alignés suite aux différentes normalisations, nous appliquons une segmentation du visage à l'aide d'une grille de dimension $5 * 5$ afin d'extraire les caractéristiques faciales.

Afin de caractériser les expressions faciales, nous sélectionnons trois descripteurs représentatifs de la littérature : les LBP et les LBP-TOP pour caractériser la texture et les HOOF pour caractériser le mouvement. Nous appliquons également les LMPs afin de comparer les performances de notre descripteur en présence de variations de pose du visage. Chaque descripteur est appliqué sur les données acquises par la caméra fixée sur le casque, où les conditions d'analyse sont excellentes (aucun mouvement de la tête). Les caractéristiques du visage sont stables durant toutes les séquences d'acquisition, ce qui ne demande aucune normalisation du visage. Les résultats obtenus par les différents descripteurs sont représentés dans le tableau 6.2. Les pourcentages sont obtenus en appliquant le descripteur sur l'ensemble des séquences de la base SNaP-2DFe (840 séquences, comprenant 7 expressions et 6 mouvements de tête).

La même expérimentation est faite sur les images acquises par la caméra fixée sur

le trépied, où les mouvements de tête sont présents. Les résultats sont représentés dans le Tableau 6.2. Pour chaque descripteur, nous calculons les performances de classification obtenues sur les données originales acquises par la caméra où aucune normalisation n'est appliquée, puis sur les trois méthodes de normalisation utilisées dans l'expérimentation précédente : yeux, forme et 3D. Les résultats obtenus montrent que le choix du descripteur et de la méthode de normalisation a un impact différent sur l'analyse des expressions faciales. On remarque également que les méthodes dynamiques (LBP-TOP, HOOF, LMP) donnent de meilleures performances, avec un résultat significatif pour les LMP, où la caractérisation dense du mouvement filtré parvient à mieux distinguer les expressions faciales. Ces résultats montrent que les méthodes dynamiques de la littérature parviennent à caractériser fidèlement les expressions faciales lorsque le visage est dans de bonnes conditions d'analyse (aucun mouvement de la tête).

	Caméra casque	Caméra trépied			
Descripteur	Données originales	Données originales	Yeux	Forme	3DMM
LBP	75.52%	30.55%	47.46%	47.76%	51.34%
LBP-TOP	78.34%	19.44%	49.12%	44.62%	46.93%
HOOF	83.21%	17.38%	50.01%	42.16%	48.73%
LMP	**87.36%**	**47.24%**	**60.72%**	**53.85%**	**58.05%**

TABLEAU 6.2 – Taux de reconnaissance des expressions faciales extraits à partir de plusieurs descripteurs en fonction des méthodes de normalisation.

Les résultats obtenus sur les données acquises par la caméra du casque montrent que les méthodes dynamiques comme les LBP-TOP, les HOOF et les LMP caractérisent mieux les expressions faciales que les LBP. Cependant, l'expérimentation montre une perte significative des performances sur les données acquises par la caméra posée sur le trépied. En effet, en présence de variations de poses et de larges déplacements, les méthodes dynamiques, à l'exception des LMP, obtiennent de moins bonnes performances que les LBP dû au fait qu'elles sont très sensibles aux mouvements de la tête. Dans ce contexte, il est important de s'assurer que le visage est parfaitement aligné afin de bénéficier pleinement des informations apportées par les approches dynamiques. Bien que les méthodes dynamique souffre en présence de mouvement du visage, le filtrage du mouvement au sein du

LMP permet toutefois de conserver l'information pertinente liée aux expressions, ce qui donne de meilleurs résultats en comparaison aux autres descripteurs.

L'utilisation des méthodes de normalisation améliore significativement les performances des systèmes de reconnaissance des expressions faciales en présence de variations de poses et de larges déplacements. Au vu des résultats dans le Tableau 6.2, la normalisation basée sur la position des yeux semble la mieux adaptée en termes d'efficacité et de robustesse pour conserver les expressions faciales. En dépit du gain obtenu par les autres méthodes de normalisation sur les LBP, les méthodes basées sur la forme 2D et le modèle déformable 3D paraissent moins adaptées pour les approches de caractérisation dynamique. Cela est principalement dû à la présence d'artéfacts qui apparaissent au cours du temps et qui engendrent du mouvement qui n'est pas induit par les expressions faciales.

Nous évaluons plus en détail la robustesse des méthodes de normalisation en fonction du type de mouvement de la tête. Afin d'étudier uniquement l'impact de la normalisation sur la reconnaissance d'expression, en s'affranchissant du biais de la temporalité, nous choisissons par la suite d'évaluer uniquement les performances fournies par les LBP. Les méthodes de normalisation sont évaluées sur les différentes animations de la base SNaP-2DFe. Les performances obtenues sur chaque animation sont représentées dans le Tableau 6.3.

Normalisation	Fixe	Tx	Roulis	Tangage	Lacet	Diagonale
Données originales	45.23%	38.09%	32.47%	37.71%	33.33%	14.26%
Yeux	**52.38%**	33.33%	**47.61%**	30.95%	40.47%	26.19%
Forme	47.61%	35.71%	42.85%	30.95%	38.02%	11.90%
3DMM	48.02%	**39.27%**	40.21%	**40.74%**	**41.08%**	**34.96%**

TABLEAU 6.3 – Taux de reconnaissance des expressions faciales par classe de mouvement de la tête.

La méthode de normalisation basée sur la position des yeux est la plus adaptée lorsque le visage est frontal (Fixe, Roulis). En effet, dans ces conditions, une simple rotation dans le plan suffit à aligner parfaitement les visages. Cette méthode garantie également de conserver la géométrie du visage. Cependant, elle n'est plus adaptée en présence de varia-

tions de poses et de larges déplacements. Grâce à la reconstruction des parties occultées du visage, les méthodes 3D obtiennent les meilleures performances en présence de rotations hors plan. Cependant, malgré les performances obtenues, les méthodes de normalisation 3D ne garantissent pas de reconnaitre parfaitement les expressions faciales. Cela est principalement dû à la difficulté de reconstruire les parties occultées du visage, où chaque information de texture et de géométrie est une source importante d'information pour caractériser une expression faciale.

Afin de pouvoir mesurer précisément l'impact des méthodes de normalisation sur la reconnaissance des expressions faciales, nous avons représenté dans la Figure 6.5 les courbes ROC associées à chacune des expressions de la base SNaP-2DFe, tout mouvement de la tête confondu. Ces courbes ont été obtenues en utilisant le descripteur LBP et une validation croisée en 10-folds sur l'ensemble des séquences de la base. La courbe bleue (cam1) représente les résultats obtenus en analysant les expressions faciales à partir de la caméra du casque et les courbes rouges (norm. 3D) et verte (norm. Forme) sont calculées à partir des données fournies par la caméra du trépied après normalisation. Nous avons sélectionné uniquement les méthodes de normalisation basées sur la forme 2D et le modèle 3D en fonction des performances obtenues dans la Tableau 6.2.

Au vu de la courbe moyenne, les performances obtenues à partir de la caméra du trépied, après normalisation, montrent des résultats plus bas que ceux obtenus directement à partir de la caméra du casque. Les différentes courbes révèlent que certaines expressions comme la colère et la surprise, souffrent moins que d'autres du processus de normalisation. Cela est probablement dû au fait que les méthodes de normalisation induisent moins de déformations au niveau de certaines régions du visage. Les courbes correspondant aux expressions de peur et de dégoût montrent que la normalisation basée sur le 3DMM est plus robuste que celle basée sur la forme. En ce qui concerne les expressions de joie, de tristesse et de neutre, elles sont plus impactées par la normalisation. Cela s'explique principalement par le fait, que les régions éloignées des landmarks, comme les joues, semblent plus affectées par les déformations induites par la normalisation. Dans le cas de l'expression neutre, la perte des performances met en avant le fait que les déformations induites par la normalisation du visage augmentent le taux de faux positifs dans la reconnaissance des expressions faciales.

Au vu de ces expérimentations, nous montrons que les méthodes de normalisation permettent d'augmenter significativement les performances des systèmes de reconnaissance des expressions faciales en présence de variations de pose et de larges déplacements. Cependant, les méthodes de normalisation actuelles ne permettent pas d'obtenir des performances similaires aux résultats obtenus dans des conditions contrôlées (pose fixe et frontale à la caméra). Nous avons également montré que le choix de la normalisation à appliquer dépend fortement du type de mouvement de la tête.

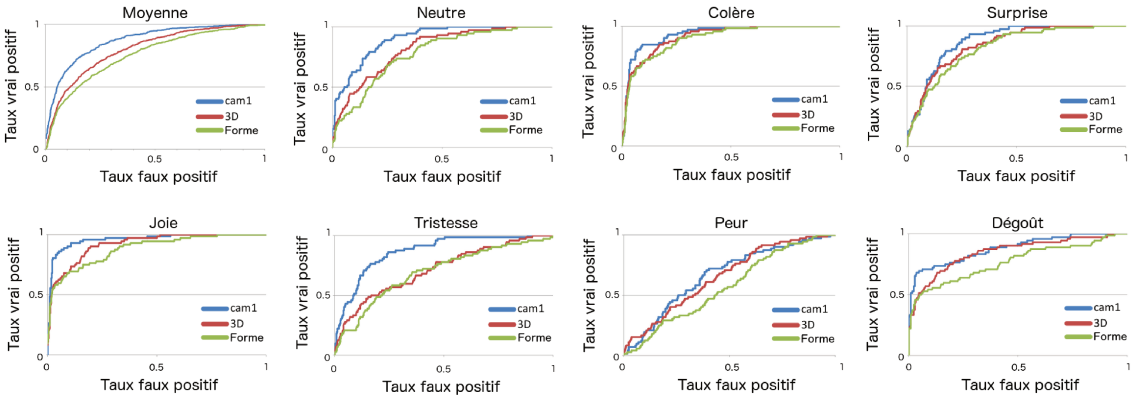


FIGURE 6.5 – Courbes ROC des expressions faciales calculées à partir du descripteur LBP.

Principalement utilisées dans les systèmes de reconnaissance faciale, les méthodes de normalisation ne sont pas parfaitement adaptées pour être utilisées dans les systèmes de reconnaissance des expressions faciales. En effet, dans ces systèmes, il est important de pouvoir conserver les déformations induites par les expressions toute en supprimant la déformation relative aux mouvements de la tête. Or les méthodes employées jusqu'à ce jour, ne semblent pas adaptées pour ce cas d'analyse. Dans la section suivante, nous discutons des perspectives d'évolution des méthodes de normalisation pour l'analyse des expressions faciales en présence de mouvement de la tête.

6.5 Conclusion

Quand les caractéristiques faciales (texture, géométrie, mouvement) sont utilisées pour caractériser une expression, la présence de déformations faciales induites par les variations de pose et de larges déplacements dégrade les performances des systèmes. Sup-

primer ces déformations est une tâche complexe et qui souvent à un impact négatif sur l'analyse des expressions faciales [19].

Nous proposons un système d'acquisition innovant permettant de quantifier la robustesse des systèmes de reconnaissance d'expressions faciales en présence de mouvements de la tête. Nous avons montré à travers plusieurs expérimentations que les méthodes de normalisation les plus représentatives de la littérature (basées sur les yeux, la forme 2D et les 3DMM) ont tendance à déformer le visage, ce qui induit des pertes dans les performances des systèmes. Bien que la normalisation 3D semble la plus adaptée parmi les trois catégories de méthodes analysées, les résultats montrent qu'en fonction du type de mouvement de la tête, il est préférable d'utiliser une méthode de normalisation spécifique. En présence de mouvement dans le plan (fixe, T_x, roulis), les méthodes basées sur la position des yeux à l'avantage de mieux préserver la géométrie faciale au sein du visage. Cependant, en présence de rotation hors plan (lacet, tangage, diagonale), les méthodes basées sur la 3D sont plus adaptées car elles sont plus robustes aux occultations du visage et permettent de reconstruire plus fidèlement les parties occultées du visage.

Les méthodes de normalisation basées sur des modèles 3D sont en perpétuelle progression afin de répondre aux challenges posés par l'acquisition de données en condition naturelle. Cependant, aucune solution ne garantit de conserver parfaitement les expressions faciales. En effet, l'usage des méthodes de normalisation ne semble pas adéquat pour aligner le visage sans induire de déformations au sein du visage. La perte des performances en présence de mouvement de la tête est particulièrement due au bruit induit par les méthodes de normalisation. Cela a une incidence directe sur les performances des systèmes utilisant des méthodes dynamiques. Bien que les méthodes dynamiques ont prouvé leur efficacité pour caractériser les expressions faciales dans un contexte contrôlé (pose fixe et frontale), elles ne supportent pas très bien l'association avec une méthode de normalisation, car le bruit généré par ces méthodes induit de fausse déformation qui ne sont pas liées aux expressions faciales.

De notre point de vue, il est envisageable d'explorer une autre solution pour normaliser un visage en s'appuyant sur l'information d'un flux optique dense. Une solution capable de séparer la source de mouvement liée au déplacement de la tête et la source de

mouvement liée aux expressions faciales. Récemment, Yang et al. [119] ont proposé une étude montrant que les méthodes de normalisation basées sur la géométrie et la texture ne sont pas parfaitement adaptées pour normaliser un visage en présence d'expressions faciales. Ils proposent ainsi une nouvelle méthode de normalisation basée sur le flux optique permettant de mieux conserver la géométrie du visage sans impacter les expressions faciales. Dans ce sens, la base de données proposée SNaP-2DFe peut-être employée à l'avenir pour approfondir les travaux dans cette direction. La caméra du casque permet de récupérer la vérité terrain du flux associé à l'expression faciale tandis que la caméra du trépied fournit des flux optique où le mouvement des expressions est confondu avec celui de la tête. Grâce à cela, il est possible de quantifier la qualité de la normalisation en calculant la distance entre le flux du casque et celui provenant de la soustraction du flux lié au mouvement de la tête au flux global.

Chapitre 7

Conclusion

« *Un bon acteur sait mettre de l'émotion dans l'action et de l'action dans l'émotion.* »

Charlie Chaplin

Sommaire

7.1 Résumé des contributions

Les systèmes d'analyse d'expressions faciales proposés dans l'état de l'art obtiennent de très bonnes performances dans des conditions où l'environnement est contrôlé (fond de la scène uniforme, lumière homogène) et où les expressions faciales sont maîtrisées (pose fixe, visage face à la caméra, expression de forte intensité). Cependant, ces données ne reflètent pas les conditions rencontrées dans une situation d'interaction naturelle (caméra de surveillance, visioconférence) où la personne est libre de ses mouvements.

Dans ce contexte, l'environnement (changements lumineux, intérieur/extérieur), ainsi que l'interaction (variations de pose, larges déplacements dans la scène, occultation de visage, intensité des expressions variable) sont peu contraints. L'ensemble de ces contraintes baisse significativement les performances des systèmes d'analyse des expressions faciales et mettent en évidence plusieurs verrous scientifiques, encore non résolus.

Nos travaux apportent plusieurs contributions permettant de répondre à différents problèmes liés à l'analyse des expressions faciales dans une séquence d'images, à destination d'un système d'acquisition où l'interaction est naturelle. Plus particulièrement, nous proposons des solutions innovantes aux problèmes de variations de l'intensité des mouvements faciaux et de variations de pose.

7.1.1 Variation de l'intensité du mouvement

En situation d'interaction naturelle, les intensités des mouvements faciaux ont tendance à varier d'une personne à une autre car elles n'expriment pas leur expression de la même manière (peu ou fortement expressif). De plus, comme le suggère Ekman [30], les expressions faciales sont composées d'un ensemble de mouvements de faible (micro expression) et forte (macro expression) intensités. L'association de ces deux mouvements permet de mieux nuancer les expressions.

La difficulté liée aux variations de l'intensité du mouvement réside dans le fait que les caractéristiques de mouvement sont très différentes entre les expressions de faibles et fortes intensités. Dans la majorité des cas, cela demande de passer par des techniques de prétraitements adaptés au mouvement analysé. En règle générale, les prétraitements ap-

pliqués aux visages permettent de mieux caractériser un mouvement spécifique au détriment d'un autre mouvement. Ceci implique qu'il n'existe pas de solution unique prenant en compte à la fois les petites et les grandes variations de mouvement.

Pour cela, nous proposons un descripteur innovant appelé LMP (Local Motion Patterns) s'appuyant sur les caractéristiques physiques déformables du visage afin de conserver uniquement les directions principales du mouvement facial. Le mouvement est calculé dans différentes régions du visage. Ces régions sont définies en relation avec le système FACS et permettent d'analyser directement les mouvements cohérents induits par les muscles faciaux. En s'appuyant sur les différents travaux autour de l'analyse des macro et micro expressions, nous avons utilisé une approche de mouvement dense (algorithme de Färneback) pour extraire le mouvement facial. Nous avons ensuite exploité les hypothèses de cohérence du mouvement et amélioré la distinction entre l'information liée au mouvement et le bruit présent dans les données.

Nous avons présenté l'efficacité de notre descripteur sur plusieurs bases d'apprentissage à la fois pour caractériser des macro et les micro expressions. Les résultats obtenus sur les bases d'apprentissage CASME2 (70.20%) et SMIC-VIS (86.11%) montrent que le descripteur proposé donne de meilleurs résultats que les méthodes récentes de l'état de l'art pour l'analyse des micro expressions. Sans augmentation artificielle des données initiales des bases d'apprentissage, comme le font les méthodes de deep learning, nous obtenons des résultats compétitifs pour la reconnaissance des macro expressions (97.25% sur CK+, 84.58% sur Oulu-CASIA et 78.26% sur MMI). Ces différentes bases montrent la robustesse de notre descripteur en présence de variations d'intensités et aux petites variations de pose sur des données acquises par des capteurs du domaine visible et proche infrarouge.

En prenant uniquement en considération une fraction du mouvement lié aux expressions, notre méthode arrive à reconnaître les expressions faciales. Les performances obtenues en considérant 100%, 75%, 50% and 25% des séquences d'activation des expressions correspondent respectivement aux taux suivants : 96.94%, 96.69%, 95.11% et 87.16%. Ces résultats mettent en avant le fait que notre méthode a la spécificité de pouvoir reconnaitre une expression faciale sous différents niveaux d'intensité, ce qui tend à pouvoir

reconnaitre les expressions plus rapidement, voir les anticiper.

En conclusion à cette problématique, nous avons proposé une solution innovante basée sur le mouvement, permettant de caractériser simultanément les micro et les macro expression en utilisant, pour les deux catégories d'intensité, le même système d'analyse (descripteur, modèle facial, configuration).

7.1.2 Variation de la pose

Les méthodes actuelles nécessitent généralement que le visage soit fixe et frontal à la caméra pour analyser les expressions. Cela garantit que le visage soit dans de parfaites conditions pour être analysé : aucune variation de pose (occultation) et aucun mouvement (détérioration des informations). Ceci explique pourquoi les systèmes actuels ne parviennent pas à obtenir de bonnes performances dans un contexte d'interaction naturelle où la personne est libre de ses mouvements.

Au vu de la littérature, nous avons vu que les méthodes s'appuyant sur l'information du mouvement sont plus performantes dans un contexte contrôlé. Or, ce n'est pas le cas en situation d'interaction naturelle. En effet, dans ces conditions, la présence de variations de pose et de larges déplacements induisent des discontinuités de mouvement qui renforcent la difficulté de l'analyse. Le bruit induit par le mouvement de la tête à un impact négatif sur le mouvement extrait au sein du visage. Il est alors difficile d'extraire uniquement le mouvement induit par les expressions faciales.

Des solutions de normalisation sont employées pour corriger la déformation géométrique d'un visage dans la scène afin de l'amener dans un cadre idéal d'analyse (frontal à la caméra). Bien que ces solutions permettent de réduire la distance entre deux visages extraits d'images successives, la normalisation induit des déformations néfastes sur la géométrie du visage. Celles-ci résultent de l'apparition de mouvements incohérents sur le visage.

Actuellement, il est difficile de pouvoir quantifier précisément la robustesse des méthodes de normalisation pour l'analyse des expressions faciales. Jusqu'à présent, ces méthodes sont employées directement dans des systèmes complets d'analyse des expres-

sions faciales. Les résultats obtenus en sorties de ces systèmes nous informent de manière informelle de l'apport de la normalisation sur les performances du système. Cependant, cela ne permet pas de distinguer avec précision les faiblesses des méthodes de normalisation et de ce fait, il est difficile d'améliorer efficacement leur robustesse en présence d'expressions faciales.

Dans cette thèse, nous proposons un système d'acquisition innovant appelé SNaP-2DFe (Simultaneous Natural and Posed 2D Facial expression) pour améliorer la robustesse des méthodes de normalisation pour l'analyse des expressions faciales. Ce système permet de capturer simultanément un visage dans un plan fixe et dans un plan mobile (suit le déplacement de la tête). Grâce à cela, nous fournissons une connaissance du visage à reconstruire malgré les occultations induites par les rotations 3D (hors plan) de la tête.

Nous montrons que les méthodes de normalisation employées dans les systèmes actuels ne sont pas parfaitement adaptées pour l'analyse des expressions faciales. Chacune des méthodes analysées (yeux, forme 2D, 3DMM) ont leurs avantages et leurs inconvénients en fonction du type de mouvement de la tête. Bien que ces méthodes ne permettent pas d'obtenir des performances similaires aux résultats obtenus dans des conditions maitrisées, les méthodes basées sur les modèles 3D semblent être les plus robustes.

Le système proposé permet de sensibiliser les chercheurs du domaine aux problèmes de variations de pose pour l'analyse des expressions faciales. Nous mettons principalement un accent sur le fait de choisir judicieusement la méthode la plus adaptée en fonction des mouvements de la tête observés. Enfin, nous espérons que notre système permettra, dans un avenir proche, d'améliorer la robustesse des méthodes de normalisation en présence d'expressions faciales.

7.2 Perspectives

Comme la majorité des approches de la littérature, nous avons principalement analysé des visages dans un contexte maitrisé afin de quantifier la robustesse de notre descripteur à caractériser les expressions faciales. Actuellement, le descripteur proposé ne

permet pas d'analyser, de manière fiable, les expressions faciales en présence de mouvement de la tête. C'est pourquoi, dans un avenir proche, nous souhaitons appliquer une méthode de normalisation en amont de l'extraction du mouvement lié aux expressions faciales afin de s'abstraire des mouvements de la tête. Cependant, nous avons observé que les méthodes de normalisation actuelles ne permettent pas de résoudre entièrement ce problème, car elles induisent des artéfacts qui viennent ajouter de faux mouvements. Afin de résoudre ce problème, nous souhaitons explorer une autre solution pour normaliser un visage en s'appuyant sur l'information d'un flux optique dense. Une solution qui soit capable de séparer la source de mouvement lié au déplacement de la tête et la source de mouvement lié aux expressions faciales. À l'inverse des autres méthodes de normalisation, nous pensons que normaliser et corriger le mouvement dans le domaine flux optique aura moins d'impact négatif que de reconstruire directement la géométrie du visage dans l'image. Les récents travaux de Yang et al. [119] tendent à confirmer cette hypothèse.

Plusieurs perspectives sont également envisageables afin d'améliorer la robustesse de notre descripteur. Nous souhaitons ajouter une contrainte spatio-temporelle supplémentaire afin de filtrer les discontinuités de mouvement. Actuellement, nous faisons l'hypothèse que les discontinuités temporelles ont tendance à disparaitre lorsque le vecteur global du mouvement est cumulé dans le temps. Or, il serait intéressant d'ajouter une contrainte où l'on vérifie que localement le mouvement reste cohérent entre plusieurs images successives comme le font les LBP-TOP.

Avec l'essor des méthodes de deep learning, nous souhaitons adapter notre descripteur dans un système d'apprentissage profond en s'inspirant des récentes architectures temporelles 3D. Dans ce cas, nous voudrions concevoir une architecture 3D où l'évolution des poids synaptiques entre les neurones seraient directement liés en fonction des différentes contraintes déformables du visage (élasticité de la peau). Nous pouvons également ajouter plusieurs modalités à l'architecture afin de prendre en compte la géométrie et la texture du visage pour améliorer la prise de décision finale en sortie du réseau.

Enfin, nous allons appliquer notre descripteur dans différents scénario d'usage. Nous pensons que notre descripteur peut être adapté pour l'analyse d'action (courir, marcher, danser) ou de gestes (interaction homme/machine), qui se base essentiellement sur le

mouvement. Le descripteur proposé peut également être utilisé pour améliorer le quotidien des personnes atteintes de handicap, notamment pour apprendre à lire sur les lèvres à l'aide des micro mouvements autour de la bouche, ou aider un médecin à mieux diagnostiquer les paralysies faciales.

7.3 Publications

7.3.1 Revue

Allaert, B., Mennesson, J., Bilasco, I. M., and Djeraba, C. Impact of the face registration techniques on facial expressions recognition, In Signal Processing : Image Communication, Volume 61, 2018, Pages 44-53, ISSN 0923-5965.

7.3.2 Chapitre de livre

Allaert, B., Mennesson, J., and Bilasco, I. M. EmoGame : Towards a Self-Rewarding Methodology for Capturing Children Faces in an Engaging Context. In International Workshop on Human Behavior Understanding (pp. 3-14), (2016, October). Springer International Publishing.

7.3.3 Conférences internationales

Mennesson, J., Allaert, B., Bilasco, I. M., Van Der Aa, N., Denis, A., and Cruz-Lara, S. Faces and thoughts : An empathic diary. In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on (Vol. 1, pp. 1-1), (2015, May). IEEE.

Allaert, B., Bilasco, I. M., Djeraba, C., Allaert, B., Mennesson, J., Bilasco, I. M., ... and Djeraba, C. Consistent Optical Flow Maps for Full and Micro Facial Expression Recognition. In VISIGRAPP (5 : VISAPP) (pp. 235-242), (2017, February).

7.3.4 Conférences nationales

Allaert, B., Bilasco, I. M., and Lablack, A. Vers une reconnaissance d'état affectif à base de mouvements du haut du corps et du visage. In Colloque National Compression et REprésentation des Signaux Audiovisuels (CORESA, (2014, November).

Allaert, B., Mennesson, J., Bilasco, I. M., and Djeraba, C. Etude de la dynamique du visage en situation d'interaction naturelle. In COmpression et REprésentation des Signaux Audiovisuels (CORESA), (2016, May).

Mennesson, J., Allaert, B., and Bilasco, I. M. Fast head turns detection in low quality videos using optical flow. In Reconnaissance des Formes et l'Intelligence Artificielle (RFIA), (2016, June).

Bibliographie

[1] AHONEN, T., HADID, A., AND PIETIKÄINEN, M. Face recognition with local binary patterns. *Computer vision-eccv 2004* (2004), 469–481. 31

[2] ALLAERT, B., BILASCO, I. M., AND DJERABA, C. Consistent optical flow maps for full and micro facial expression recognition. In *VISAPP* (2017), vol. 5, pp. 235–242. 67

[3] ALMAEV, T. R., AND VALSTAR, M. F. 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* (2013), IEEE, pp. 356–361. 35

[4] AMBADAR, Z., SCHOOLER, J., AND COHN, J. Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science 16*, 5 (2005), 403–410. 6, 35

[5] BÄNZIGER, T., MORTILLARO, M., AND SCHERER, K. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion 12*, 5 (2012), 1161. 56, 142

[6] BARTLETT, M., LITTLEWORT, G., FRANK, M., LAINSCSEK, C., FASEL, I., AND MOVELLAN, J. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia 1*, 6 (2006), 22–35. 144

[7] BASSILI, J. N. 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 (1979), 2049–58. 35, 62

[8] BAY, H., ESS, A., TUYTELAARS, T., AND VAN GOOL, L. Speeded-up robust features (surf). *Computer vision and image understanding 110*, 3 (2008), 346–359. 16, 20

[9] BERZINS, V. Accuracy of laplacian edge detectors. *Computer Vision, Graphics, and Image Processing 27*, 2 (1984), 195–210. 16

[10] BHATTACHARYYA, A. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Calcutta Math. Soc* (1943). 97

[11] BOURDEV, L., AND BRANDT, J. Robust object detection via soft cascade. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 2, IEEE, pp. 236–243. 15

[12] BOYLE, E. A., ANDERSON, A. H., AND NEWLANDS, A. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and*

speech 37, 1 (1994), 1–20. 4

[13] BREUER, R., AND KIMMEL, R. A deep learning perspective on the origin of facial expressions. In *Computer Vision and Pattern Recognition (CVPR) Honolulu - June 21-26* (2017). 36, 63, 65, 67, 124, 125, 127, 128, 131

[14] BRUBAKER, S. C., MULLIN, M. D., AND REHG, J. M. Towards optimal training of cascaded detectors. In *European Conference on Computer Vision* (2006), Springer, pp. 325–337. 15

[15] BURGOS-ARTIZZU, X. P., PERONA, P., AND DOLLÁR, P. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1513–1520. 28

[16] CARMINATI, L., BENOIS-PINEAU, J., AND JENNEWEIN, C. Knowledge-based supervised learning methods in a classical problem of video object tracking. In *Image Processing, 2006 IEEE International Conference on* (2006), IEEE, pp. 2385–2388. 16

[17] CHANG, C.-C., AND LIN, C.-J. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27. 112, 152

[18] CHEN, H., LI, J., ZHANG, F., LI, Y., AND WANG, H. 3d model-based continuous emotion recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1836–1845. 144

[19] CHEW, S., LUCEY, P., LUCEY, S., SARAGIH, J., COHN, J., MATTHEWS, I., AND SRIDHARAN, S. In the pursuit of effective affective computing : The relationship between features and registration. *Systems, Man, and Cybernetics, Part B : Cybernetics 42*, 4 (2012), 1006–1016. 74, 142, 157

[20] COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence 23*, 6 (2001), 681–685. 19

[21] COOTES, T. F., AND TAYLOR, C. J. Active shape models-'smart snakes'. In *BMVC* (1992), vol. 92, pp. 266–275. 17

[22] COOTES, T. F., TAYLOR, C. J., COOPER, D. H., AND GRAHAM, J. Active shape models-their training and application. *Computer vision and image understanding 61*, 1 (1995), 38–59. 18

[23] CRUZ, A., BHANU, B., AND YANG, S. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *International Conference on*

Affective Computing and Intelligent Interaction (ACII) (2011), Springer, pp. 341–350. 70

[24] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, pp. 886–893. 20

[25] DANISMAN, T., BILASCO, I. M., MARTINET, J., AND DJERABA, C. Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron. *Signal Processing 93*, 6 (2013), 1547–1556. 25, 32

[26] DENG, H.-B., JIN, L.-W., ZHEN, L.-X., HUANG, J.-C., ET AL. A new facial expression recognition method based on local gabor filter bank and pca plus lda. *International Journal of Information Technology 11*, 11 (2005), 86–96. 33

[27] DHALL, A., ASTHANA, A., GOECKE, R., AND GEDEON, T. Emotion recognition using phog and lpq features. In *Automatic Face and Gesture recognition (FG)* (2011), IEEE, pp. 878–883. 26

[28] DING, H., ZHOU, S. K., AND CHELLAPPA, R. Facenet2expnet : Regularizing a deep face recognition net for expression recognition. In *International Conference Automatic Face & Gesture Recognition (FG)* (2017), IEEE, pp. 118–126. 62, 129

[29] DOLLÁR, P., WELINDER, P., AND PERONA, P. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 1078–1085. 20

[30] EKMAN, P., AND ROSENBERG, E. *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997. 43, 44, 51, 52, 59, 75, 160

[31] ELAIWAT, S., BENNAMOUN, M., AND BOUSSAID, F. A spatio-temporal rbm-based model for facial expression recognition. *Pattern Recognition 49* (2016), 152–161. 127, 128

[32] FAN, X., AND TJAHJADI, T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition 48*, 11 (2015), 3407–3416. 127, 128

[33] FAN, X., AND TJAHJADI, T. A dynamic framework based on local zernike moment and motion history image for facial expression recognition. *Pattern Recognition 64* (2017), 399–406. 38, 127, 128

[34] FARNEBÄCK, G. Two-frame motion estimation based on polynomial expansion. In *Image Analysis : 13th Scandinavian Conference, SCIA 2003* (2003), Springer, pp. 363–370. 79, 80, 81, 82, 85, 87, 100, 101, 112

[35] FARRUGIA, R. A., AND GUILLEMOT, C. Model and dictionary guided face inpainting in the wild. In *Asian Conference on Computer Vision* (2016), Springer, pp. 62–78. 29, 30

[36] FORTUN, D., BOUTHEMY, P., AND KERVRANN, C. Optical flow modeling and computation : a survey. *Computer Vision and Image Understanding (CVIU) 134* (2015), 1–21. 63

[37] GHIMIRE, D., AND LEE, J. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors 13*, 6 (2013), 7714–7734. 34, 63, 67

[38] HAN, S., MENG, Z., LIU, P., AND TONG, Y. Facial grid transformation : A novel face registration approach for improving facial action unit recognition. In *ICIP* (2014), pp. 1415–1419. 34, 63, 149

[39] HAPPY, S., AND ROUTRAY, A. Automatic facial expression recognition using features of salient facial patches. *Affective Computing 6*, 1 (2015), 1–12. 32, 34, 38, 45, 70, 149

[40] HASSNER, T., HAREL, S., PAZ, E., AND ENBAR, R. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 4295–4304. 27

[41] HERRMANN, C., QU, C., WILLERSINN, D., AND BEYERER, J. Impact of resolution and image quality on video face analysis. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on* (2015), IEEE, pp. 1–6. 54

[42] HOCHREITER, S., BENGIO, Y., FRASCONI, P., SCHMIDHUBER, J., ET AL. Gradient flow in recurrent nets : the difficulty of learning long-term dependencies, 2001. 35

[43] HUANG, X., WANG, S., LIU, X., ZHAO, G., FENG, X., AND PIETIKAINEN, M. Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection. *CVPR* (2016). 65, 66, 67, 124

[44] HUANG, X., ZHAO, G., HONG, X., ZHENG, W., AND PIETIKÄINEN, M. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing 175* (2016), 564–578. 65, 124

[45] JACK, R. E., GARROD, O. G., YU, H., CALDARA, R., AND SCHYNS, P. G. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences 109*, 19 (2012), 7241–7244. 46

[46] JACKSON, A. S., BULAT, A., ARGYRIOU, V., AND TZIMIROPOULOS, G. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *arXiv preprint arXiv :1703.07834* (2017). 26, 72

[47] JAISWAL, S., MARTINEZ, B., AND VALSTAR, M. F. Learning to combine local models for facial action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (2015), vol. 6, IEEE, pp. 1–6. 46

[48] JAISWAL, S., AND VALSTAR, M. Deep learning the dynamic appearance and shape of facial action units. In *WACV* (2016), pp. 1–8. 64

[49] JAMPOUR, M., LI, C., YU, L.-F., ZHOU, K., LIN, S., AND BISCHOF, H. Face inpainting based on high-level facial attributes. *Computer Vision and Image Understanding* (2017). 27

[50] JENI, L. A., COHN, J. F., AND KANADE, T. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture recognition (FG)* (2015), vol. 1, IEEE, pp. 1–8. 27, 29, 71

[51] JENI, L. A., HASHIMOTO, H., AND KUBOTA, T. Robust facial expression recognition using near infrared cameras. *Journal of Advanced Computational Intelligence and Intelligent Informatics 16*, 2 (2012), 341–348. 129

[52] JIANG, B., MARTINEZ, B., VALSTAR, M. F., AND PANTIC, M. Decision level fusion of domain specific regions for facial action recognition. In *International Conference on Pattern Recognition (ICPR)* (2014), IEEE, pp. 1776–1781. 38

[53] JIANG, B., VALSTAR, M., MARTINEZ, B., AND PANTIC, M. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics 44*, 2 (2014), 161–174. 25, 26, 46, 107, 149

[54] JUNG, H., LEE, S., YIM, J., PARK, S., AND KIM, J. Joint fine-tuning in deep neural networks for facial expression recognition. In *International Conference on Computer Vision (ICCV)* (2015), pp. 2983–2991. 63, 64, 67, 127, 128, 129, 130

[55] KAZEMI, V., AND SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In *CVPR* (2014), pp. 1867–1874. 107

[56] KHAN, R. A., MEYER, A., KONIK, H., AND BOUAKAZ, S. Human vision inspired framework for facial expressions recognition. In *ICIP* (2012), pp. 2593–2596. 62, 66, 67

[57] KIM, D. H., BADDAR, W., JANG, J., AND RO, Y. M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *Transactions on Affective Computing - issue 99* (2017). 36, 65, 67, 124, 125

[58] KOELSTRA, S., PANTIC, M., AND PATRAS, I. A dynamic texture-based approach to recognition of facial actions and their temporal models. *Pattern Analysis and Machine Intelligence (PAMI) 32*, 11 (2010), 1940–1954. 25

[59] KOLLIAS, D., NICOLAOU, M. A., KOTSIA, I., ZHAO, G., AND ZAFEIRIOU, S. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on* (2017), IEEE, pp. 1972–1979. 49

[60] KOTSIA, I., ZAFEIRIOU, S., AND PITAS, I. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition 41*, 3 (2008), 833–851. 63

[61] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105. 22

[62] LEE, C.-S., AND CHELLAPPA, R. Sparse localized facial motion dictionary learning for facial expression recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE* (2014), pp. 3548–3552. 127, 128

[63] LI, J., CHEN, Y., XIAO, S., ZHAO, J., ROY, S., FENG, J., YAN, S., AND SIM, T. Estimation of affective level in the wild with multiple memory networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on* (2017), IEEE, pp. 1947–1954. 48

[64] LI, X., HONG, X., MOILANEN, A., HUANG, X., PFISTER, T., ZHAO, G., AND PIETIKÄINEN, M. Reading hidden emotions : spontaneous micro-expression spotting and recognition. In *CVPR* (2015), pp. 217–230. 64, 65, 66, 67, 68, 124, 125, 126, 131

[65] LI, X., PFISTER, T., HUANG, X., ZHAO, G., AND PIETIKÄINEN, M. A spontaneous micro-expression database : Inducement, collection and baseline. In *FG* (2013). 52, 126

[66] LIAO, C., CHUANG, H., DUAN, C., AND LAI, S. Learning spatial weighting via quadratic programming for facial expression analysis. In *Computer Vision and Pattern Recognition Workshops (CVPRW)* (2010), IEEE, pp. 86–93. 36, 46, 69

[67] LIAO, C.-T., CHUANG, H.-J., DUAN, C.-H., AND LAI, S.-H. Learning spatial weighting for facial expression analysis via constrained quadratic programming. *Pattern Recognition 46*, 11 (2013), 3103–3116. 63, 127, 128

[68] LINDEBERG, T., AND GÅRDING, J. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and vision computing 15*, 6 (1997), 415–434. 16

[69] LIONG, S.-T., SEE, J., PHAN, R. C.-W., LE NGO, A. C., OH, Y.-H., AND WONG, K. Subtle expression recognition using optical strain weighted features. In *ACCV* (2014), Springer, pp. 644–657. 65

[70] LIU, M., SHAN, S., WANG, R., AND CHEN, X. Learning expressionlets via universal manifold model for dynamic facial expression recognition. *Transactions on Image Processing 25*, 12 (2016), 5920–5932. 127, 128, 129, 130, 131

[71] LIU, Y.-J., ZHANG, J.-K., YAN, W.-J., WANG, S.-J., ZHAO, G., AND FU, X. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing 7*, 4 (2016), 299–310. 64, 65, 67, 124

[72] LOPES, A. T., DE AGUIAR, E., DE SOUZA, A. F., AND OLIVEIRA-SANTOS, T. Facial expression recognition with convolutional neural networks : Coping with few data and the training sample order. *Pattern Recognition 61* (2017), 610–628. 33, 62, 67, 127, 128

[73] LOWE, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (1999), vol. 2, Ieee, pp. 1150–1157. 16

[74] LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110. 20

[75] LUCEY, P., COHN, J., KANADE, T., SARAGIH, J., AMBADAR, Z., AND MATTHEWS, I. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshop (CVPR)* (2010), IEEE, pp. 94–101. 56, 144, 146

[76] MAHMOUD, M. M., BALTRUŠAITIS, T., AND ROBINSON, P. Automatic detection of naturalistic hand-over-face gesture descriptors. In *ICMI* (2014), ACM, pp. 319–326. 29

[77] MAJUMDER, A., BEHERA, L., AND SUBRAMANIAN, V. K. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition 47*, 3 (2014), 1282–1293. 63

[78] MATTHEWS, I., AND BAKER, S. Active appearance models revisited. *International journal of computer vision 60*, 2 (2004), 135–164. 20

[79] MAVADATI, S. M., MAHOOR, M. H., BARTLETT, K., TRINH, P., AND COHN, J. F. Disfa : A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing 4*, 2 (2013), 151–160. 56

[80] MCKEOWN, G., VALSTAR, M., COWIE, R., PANTIC, M., AND SCHRÖDER, M. The semaine database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing 3*, 1 (2012), 5–17. 58, 142, 144, 146

[81] MEHRABIAN, A. Communication without words. *Communication theory* (2008), 193–200. 4

[82] MOLLAHOSSEINI, A., CHAN, D., AND MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV)* (2016), pp. 1–10. 62, 130

[83] MOU, W., CELIKTUTAN, O., AND GUNES, H. Group-level arousal and valence recognition in static images : Face, body and context. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (2015), vol. 5, IEEE, pp. 1–6. 48

[84] NICOLAOU, M. A., ZAFEIRIOU, S., AND PANTIC, M. Correlated-spaces regression for learning continuous emotion dimensions. In *Proceedings of the 21st ACM international conference on Multimedia* (2013), ACM, pp. 773–776. 48

[85] NICOLLE, J., RAPP, V., BAILLY, K., PREVOST, L., AND CHETOUANI, M. Robust continuous prediction of human emotions using multiscale dynamic cues. In *International Conference on Multimodal Interaction (ICMI)* (2012), ACM, pp. 501–508. 70

[86] OJALA, T., PIETIKAINEN, M., AND MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence (PAMI) 24*, 7 (2002), 971–987. 62

[87] PANTIC, M., VALSTAR, M., RADEMAKER, R., AND MAAT, L. Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo (ICME)* (2005), IEEE, pp. 5–pp. 59, 144

[88] PATEL, D., HONG, X., AND ZHAO, G. Selective deep features for micro-expression recognition. In *International Conference on Pattern Recognition (ICPR)* (2016), IEEE, pp. 2258–2263. 33, 64, 65, 67, 126

[89] PORTER, S., AND TEN BRINKE, L. Reading between the lies : Identifying concealed and falsified emotions in universal facial expressions. *Psychological science 19*, 5 (2008), 508–514. 51, 52

[90] REVAUD, J., WEINZAEPFEL, P., HARCHAOUI, Z., AND SCHMID, C. Epicflow : Edge-preserving interpolation of correspondences for optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1164–1172. 67

[91] RINGEVAL, F., SONDEREGGER, A., SAUER, J., AND LALANNE, D. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (2013), IEEE, pp. 1–8. 57, 142, 144, 146

[92] RIVERA, A. R., AND CHAE, O. Spatiotemporal directional number transitional graph for dynamic texture recognition. *Pattern Analysis and Machine Intelligence (PAMI) 37*, 10 (2015), 2146–2152. 25, 26

[93] RUSSELL, J. A. A circumplex model of affect. *Journal of personality and social psychology 39*, 6 (1980), 1161. 47

[94] SADEGHI, H., RAIE, A.-A., AND MOHAMMADI, M.-R. Facial expression recognition using geometric normalization and appearance representation. In *Machine Vision and Image Processing (MVIP)* (2013), IEEE, pp. 159–163. 38

[95] SAEED, A., AL-HAMADI, A., AND NIESE, R. The effectiveness of using geometrical features for facial expression recognition. In *Cybernetics* (2013), pp. 122–127. 63

[96] SÁNCHEZ, A., RUIZ, J. V., MORENO, A. B., MONTEMAYOR, A. S., HERNÁNDEZ, J., AND PANTRIGO, J. J. Differential optical flow applied to automatic facial expression recognition. *Neurocomputing 74*, 8 (2011), 1272–1282. 36, 111

[97] SANDBACH, G., ZAFEIRIOU, S., AND PANTIC, M. Markov random field structures for facial action unit intensity estimation. In *International Conference on Computer Vision Workshops (ICCVW)* (2013), pp. 738–745. 69

[98] SAVRAN, A., AND SANKUR, B. Non-rigid registration based model-free 3d facial expression recognition. *Computer Vision and Image Understanding 162* (2017), 146–165. 70

[99] SCHAEFER, S., MCPHAIL, T., AND WARREN, J. Image deformation using moving least squares. In *ACM transactions on graphics (TOG)* (2006), vol. 25, ACM, pp. 533–540. 151

[100] SENECHAL, T., RAPP, V., SALAM, H., SEGUIER, R., BAILLY, K., AND PREVOST, L. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42*, 4 (2012), 993–1005. 32

[101] SHAN, C., GONG, S., AND MCOWAN, P. W. Facial expression recognition based on local binary patterns : A comprehensive study. *Image and Vision Computing 27*, 6 (2009), 803–816. 66, 67

[102] SIKKA, K., WU, T., SUSSKIND, J., AND BARTLETT, M. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision (ECCV)* (2012), Springer, pp. 250–259. 142

[103] SU, M., HSIEH, Y., AND HUANG, D.-Y. A simple approach to facial expression recognition. *Proceeding WSEAS (World Scientific and Engineering Academy and Society)* (2007). 127, 128

[104] SUN, J., REHG, J. M., AND BOBICK, A. Automatic cascade training with perturbation bias. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on* (2004), vol. 2, IEEE, pp. II–II. 15

[105] TZIMIROPOULOS, G., AND PANTIC, M. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 593–600. 20

[106] VIOLA, P., AND JONES, M. J. Robust real-time face detection. *International journal of computer vision 57*, 2 (2004), 137–154. 13, 14

[107] WANG, S., YAN, W.-J., LI, X., ZHAO, G., AND FU, X. Micro-expression recognition using dynamic textures on tensor independent color space. In *ICPR* (2014), pp. 4678–4683. 124

[108] WANG, S.-J., YAN, W.-J., ZHAO, G., FU, X., AND ZHOU, C.-G. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In *ECCV Workshop* (2014), pp. 325–338. 124

[109] WANG, Y., SEE, J., PHAN, R. C.-W., AND OH, Y.-H. Lbp with six intersection points : Reducing redundant information in lbp-top for micro-expression recognition. In *ACCV* (2014), pp. 525–537. 65, 124

[110] WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. Image quality assessment : from error visibility to structural similarity. *Image Processing 13*, 4 (2004), 600–612. 150

[111] WEINZAEPFEL, P., REVAUD, J., HARCHAOUI, Z., AND SCHMID, C. Deepflow : Large displacement optical flow with deep matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (2013), IEEE, pp. 1385–1392. 78

[112] XIAO, R., ZHU, L., AND ZHANG, H.-J. Boosting chain learning for object detection. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 709–715. 15

[113] XU, F., ZHANG, J., AND WANG, J. Z. Microexpression identification and categorization using a facial dynamics map. *Transactions on Affective Computing 8*, 2 (2017), 254–267. 126

[114] XU, X., AND KAKADIARIS, I. A. Joint head pose estimation and face alignment framework using global and local cnn features. In *Proc. 12th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC* (2017), vol. 2. 27

[115] YAN, W.-J., LI, X., WANG, S.-J., ZHAO, G., LIU, Y.-J., CHEN, Y.-H., AND FU, X. Casme ii : An improved spontaneous micro-expression database and the baseline evaluation. *PloS one 9*, 1 (2014), https ://doi.org/10.1371/journal.pone.0086041. 105, 124

[116] YAN, W.-J., WANG, S.-J., LIU, Y.-J., WU, Q., AND FU, X. For micro-expression recognition : Database and suggestions. *Neurocomputing 136* (2014), 82–87. 51

[117] YAN, W.-J., WU, Q., LIANG, J., CHEN, Y.-H., AND FU, X. How fast are the leaked facial expressions : The duration of micro-expressions. *Journal of Nonverbal Behavior 37*, 4 (2013), 217–230. 51, 104

[118] YAN, Y., RICCI, E., SUBRAMANIAN, R., LIU, G., LANZ, O., AND SEBE, N. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence 38*, 6 (2016), 1070–1083. 27

[119] YANG, S., AN, L., LEI, Y., LI, M., THAKOOR, N., BHANU, B., AND LIU, Y. A dense flow-based framework for real-time object registration under compound motion. *Pattern Recognition 63* (2017), 279–290. 158, 164

[120] YANG, S., AND BHANU, B. Facial expression recognition using emotion avatar image. In *Automatic Face and Gesture recognition (FG)* (2011), IEEE, pp. 866–871. 53

[121] ZHANG, J., KAN, M., SHAN, S., AND CHEN, X. Occlusion-free face alignment : deep regression networks coupled with de-corrupt autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 3428–3437. 30

[122] ZHANG, K., HUANG, Y., DU, Y., AND WANG, L. Facial expression recognition based on deep evolutional spatial-temporal networks. *Transactions on Image Processing 26*, 9 (2017), 4193–4203. 63, 64, 127, 128, 129, 130, 131

[123] ZHAO, G., HUANG, X., TAINI, M., LI, S. Z., AND PIETIKÄINEN, M. Facial expression recognition from near-infrared videos. *Image and Vision Computing 29*, 9 (2011), 607–619. 129

[124] ZHAO, G., AND PIETIKAINEN, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence (PAMI) 29*, 6 (2007), 915–928. 35, 46, 53, 62, 66, 67, 70, 128, 129, 130, 131, 142

[125] ZHAO, L., WANG, Z., AND ZHANG, G. Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram. *Mathematical Problems in Engineering* (2017). 127, 128, 129, 130, 131

[126] ZHONG, L., LIU, Q., YANG, P., LIU, B., HUANG, J., AND METAXAS, D. N. Learning active facial patches for expression analysis. In *CVPR* (2012), IEEE, pp. 2562–2569. 130

[127] ZHU, X., LEI, Z., YAN, J., YI, D., AND LI, S. High-fidelity pose and expression normalization for face recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 787–796. 26, 71, 149, 151

[128] ZHU, Y., WANG, S., YUE, L., AND JI, Q. Multiple-facial action unit recognition by shared feature learning and semantic relation modeling. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (2014), IEEE, pp. 1663–1668. 46