Reconnaissance des émotions par traitement d'images

Sonia Gharsalli

Centre Val de Loire UNIVERSITÉ Spirit of the Loire Valley

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

ÉCOLE DOCTORALE MATHEMATIQUE, INFORMATIQUE, PHISIQUE THEORIQUE ET INGENIERIE DES SYSTEMES

LABORATOIRE PRISME

THÈSE présentée par :

Sonia GHARSALLI

soutenue le : **12 juillet 2016**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline/ Spécialité : Sciences et techniques industrielles

Reconnaissance des émotions par traitement d'images

THÈSE dirigée par :

Bruno EMILE	Maître de conférences-HDR, Université d'Orléans, Laboratoire PRISME
Hélène Laurent	Maître de conférences-HDR, INSA Centre Val de Loire, Laboratoire PRISME

RAPPORTEURS :

Christophe ROSENBERGER	Professeur des Universités, ENSICAEN, Laboratoire GREYC
Alain PRUSKI	Professeur des Universités, Université de Loraine, Laboratoire LCOMS

JURY

Christophe LEGER	Professeur des Universités, Polythech Orléans, Laboratoire PRISME, Président du Jury
Christophe ROSENBERGER	Professeur des Universités, ENSICAEN, Laboratoire GREYC
Alain PRUSKI	Professeur des Universités, Université de Loraine, Laboratoire LCOMS
Yannick BENEZETH	Maître de conférences-HDR, Université de Dijon, Laboratoire Le2i
Bruno EMILE	Maître de conférences-HDR, Université d'Orléans, Laboratoire PRISME
Hélène Laurent	Maître de conférences-HDR, INSA Centre Val de Loire, Laboratoire PRISME

Résumé

La reconnaissance des émotions est l'un des domaines scientifiques les plus complexes. Ces dernières années, de plus en plus d'applications tentent de l'automatiser. Ces applications innovantes concernent plusieurs domaines comme l'aide aux enfants autistes [1], les jeux vidéo [2], l'interaction homme-machine.
Les émotions sont véhiculées par plusieurs canaux. Nous traitons dans notre recherche les expressions émotionnelles faciales en s'intéressant spécifiquement aux six émotions de base [3] à savoir la joie, la colère, la peur, le dégoût, la tristesse et la surprise. Une étude comparative de deux méthodes de reconnaissance des émotions l'une basée sur les descripteurs géométriques et l'autre basée sur les descripteurs d'apparence est effectuée sur la base CK+ [4], base d'émotions simulées, et la base FEEDTUM [5], base d'émotions spontanées. Différentes contraintes telles que le changement de résolution, le nombre limité d'images labélisées dans les bases d'émotions, la reconnaissance de nouveaux sujets non inclus dans la base d'apprentissage sont également prises en compte. Une évaluation de différents schémas de fusion est ensuite réalisée lorsque de nouveaux cas, non inclus dans l'ensemble d'apprentissage, sont considérés. Les résultats obtenus sont prometteurs pour les émotions simulées (ils dépassent 86%), mais restent insuffisant pour les émotions spontanées. Nous avons appliqué également une étude sur des zones locales du visage, ce qui nous a permis de développer des méthodes hybrides par zone. Ces dernières améliorent les taux de reconnaissance des émotions spontanées. Finalement, nous avons développé une méthode de sélection des descripteurs d'apparence basée sur le taux d'importance que nous avons comparée avec d'autres méthodes de sélection. La méthode de sélection proposée permet d'améliorer le taux de reconnaissance par rapport aux résultats obtenus par deux méthodes reprises de la littérature.

Remerciments

Je souhaite remercier en premier lieu Christophe Rosenberger, professeur des universités à ENSI de Caen et Alain Pruski, professeur des universités à l'université de Lorraine d'avoir accepté d'être les rapporteurs de cette thèse, ainsi qu'à Christophe Leger, professeur des universités à Polytech Orléans et Yannick Benezeth, maître de conférences à Dijon pour en avoir examiné les travaux.
Je tiens à remercier tout particulièrement mes directeurs et encadrants de thèse pour leur appui scientifique et les conseils qu'ils m'ont donné tout au long de cette thèse. D'abord à Hélène Laurent pour son aide et ses corrections méticuleuses ; ensuite à Bruno Emile pour son aide et son bon sens ; enfin à Xavier Desquesnes pour ses conseils.
Je tiens à remercier vivement Laure Spina pour sa gentillesse, sa rapidité et son efficacité dans la gestion de tous nos papiers administratives.
Je remercie également mes amis pour le bon temps que nous avons passé ensemble et leurs encouragements.
Je ne peux terminer ces remerciements sans une pensée pour ma famille et mon fiancé, qui m'ont beaucoup soutenue et encouragée.

Sonia Gharsalli

Table des matières

Chapitre 1

Introduction

Les émotions sont indispensables à notre vie. Elles permettent d'améliorer la communication entre les individus, d'assurer une meilleure compréhension du message véhiculé et de s'adapter à une situation donnée. Les émotions jouent un rôle primordial pour la prise de décision. Elles influencent également les comportements et façonnent la personnalité.
Ces différents rôles intéressent beaucoup d'applications qui tentent d'automatiser la reconnaissance des émotions. Ces applications sont liées à de nombreux domaines. Nous citons à titre d'exemples :

- Marketing : des applications pour mesurer la satisfaction des clients, prévoir les produits qui les intéressent.
- Médecine : aide à la détection de certaines maladies psychologiques, aide à l'apprentissage des émotions pour les enfants autistes.
- Sécurité : détection du stress.
- Interaction Homme-Machine : robot d'accompagnement, voiture intelligente.
- Éducation : apprentissage à distance.

La reconnaissance automatique des émotions est un réel défi. D'une part, les représentations des émotions proposées par les théoriciens sont variées. Elles sont soit considérées comme des dimensions (représentation continue), soit considérées comme des groupes catégoriques avec des limites bien définies (représentation discrète). Ce manque d'une représentation standard rend l'automatisation de la reconnaissance des émotions plus difficile. D'autre part, les émotions sont multi-componentielles. Elles se manifestent sur différents canaux (verbal, vocal et non

verbal). D'après les travaux de Mehrabian [6], connu pour ses recherches sur la communication, la communication des émotions est à 7% verbale, à 38% vocale (tonalité de la voix) et 55% non verbale (expressions faciales, expressions corporelles).

Dans cette thèse, nous traitons le cas de la reconnaissance des émotions par traitement d'images et considérons uniquement le visage qui est le vecteur de reconnaissance le plus utilisé dans la littérature. Cependant, le visage représente, en plus des émotions, l'intention, l'état cognitif et l'expression des paroles. Une description bien définie de chaque émotion devient alors indispensable pour une bonne reconnaissance. L'extraction de descripteurs significatifs caractérisant les émotions est une étape très importante dans la reconnaissance des émotions. Le choix de la méthode d'extraction des descripteurs, le type de descripteurs et la sélection des meilleurs descripteurs sont ainsi des critères primordiaux dans un système de reconnaissance des émotions et sont directement liés à l'amélioration du taux de reconnaissance des émotions.

Contraintes

Les expressions faciales émotionnelles se basent principalement sur le mouvement des muscles. Ces mouvements sont détectables par le changement de l'aspect visuel qui se résume à des changements de la forme des caractéristiques faciales et des déformations de la peau pour former des rides et des sillons. Cependant, l'aspect visuel est très lié aux sujets. La différence inter-sujets présente ainsi l'une des difficultés majeures de la reconnaissance des émotions. L'intensité des émotions est variable selon la situation à laquelle le sujet est confronté. L'expression de l'émotion par le même sujet est alors différente selon la situation. La différence intra-sujet présente également une contrainte pour la reconnaissance des émotions.
Plusieurs contraintes, liées aux environnement spécifiques à chaque application, existent également, parmi lesquelles nous citons le changement de résolution, la variation d'illumination, les occultations. Souvent, les travaux de la littérature [7], [8], [9] traitant ces contraintes utilisent les bases d'émotions simulées. A notre connaissance aucune de ces contraintes n'a été testée dans le cadre de la reconnaissance des émotions spontanées et nous n'avons pas trouvé de base d'émotions spontanées les comportant.
La labellisation des bases de données est très importante dans l'apprentissage et

l'évaluation des méthodes de reconnaissance des émotions. Dans le cas d'émotions simulées, le label de la séquence est souvent attribué selon le label de la dernière image. Ceci est également vrai pour les actions unitaires constituant les codes des expressions faciales. Dans le cas des bases d'images spontanées, les labels appliqués sont ceux des séquences stimulantes de l'émotion. Le choix des bases d'apprentissage est également très important pour que le système soit opérationnel dans les conditions réelles. Les émotions spontanées sont souvent moins prononcées que les émotions simulées. Elles sont plus proches de la réalité et présentent plus de difficultés pour la reconnaissance automatique.
Dans le cadre de notre travail, nous traitons des émotions simulées et des émotions spontanées, ainsi que des contraintes souvent présentes dans les applications de reconnaissance des émotions à savoir le changement de résolution, la faible taille des ensembles d'apprentissage, l'exécution en temps réel et les variations inter-sujets.

Contribution

Le but de cette thèse est de trouver un ensemble de descripteurs représentant les changements faciaux des émotions. L'ensemble de descripteurs doit être robuste face à des contraintes présentes dans les applications réelles et indépendant des sujets d'apprentissage. Les méthodes proposées doivent être applicables dans le cas des émotions simulées et des émotions spontanées.

Les contributions effectuées dans cette thèse sont :

a) Étude comparative des descripteurs géométriques et des descripteurs d'apparence sous différentes contraintes (variation de résolution, taille de l'ensemble d'apprentissage, temps d'exécution) et évaluation de différents schémas de fusion sur les émotions simulées et les émotions spontanées [10].

b) Étude de la reconnaissance des émotions sur des zones du visage.

c) Méthode de sélection des descripteurs les plus pertinents en terme de reconnaissance des émotions. Le choix du nombre de descripteurs est obtenu à partir d'une étude des taux de reconnaissance des Séparateurs à Vaste Marge (SVM) et des taux d'erreur des forêts aléatoires [11].

Organisation de la thèse

La thèse est organisée en trois chapitres qui sont présentés comme suit :

- Le chapitre II expose tout d'abord les grandes théories des émotions et les différentes représentations proposées par les psychologues. Ensuite, les méthodes automatiques de reconnaissance des émotions sont présentées. Nous clôturons par une présentation des bases de données et des différents critères qu'elles doivent vérifier pour se rapprocher des conditions réelles.
- Le chapitre III porte sur l'évaluation subjective d'une base d'émotions spontanées (la base FEEDTUM [5]), suivie d'une présentation de deux méthodes de reconnaissance d'émotions de la littérature utilisant deux types de descripteurs différents. Une évaluation des deux méthodes sur les émotions simulées et les émotions spontanées est ensuite effectuée, suivi par une comparaison des deux méthodes sous différentes contraintes (changement de résolution, diminution du nombre d'échantillons d'apprentissage, temps d'exécution).
- Le chapitre IV présente une étude comparative entre plusieurs schémas de fusion et les méthodes présentées dans le second chapitre après quelques modifications. Les comparaisons sont effectuées pour des émotions simulées et des émotions spontanées. Une évaluation des meilleures méthodes de fusion est effectuée en prenant en considération la contrainte inter-sujets. Une évaluation de trois méthodes de sélection est enfin effectuée afin de trouver l'ensemble de descripteurs le plus représentatif des émotions.

Chapitre 2

Etat de l'art

2.1 Introduction

Les émotions sont très présentes dans notre vie quotidienne. Elles se manifestent par des réactions physiologiques (palpitations, chaleur, accélération du pouls...), des réactions motrices (expressions gestuelles, expressions faciales...) et des changements de la voix. La complexité des émotions a suscité l'intérêt de nombreux chercheurs. Différentes représentations des émotions ont alors été proposées permettant ainsi de produire un ensemble de théories. L'automatisation de la reconnaissance des émotions peut être alors abordée de diverses manières. Nous nous sommes intéressés principalement à la reconnaissance des expressions faciales émotionnelles. Dans ce cadre la reconnaissance se base sur différentes caractéristiques visuelles de l'expression.

Ce chapitre est dédié dans un premier temps à l'exposition de la définition des émotions, suivie des grandes théories qui ont vu le jour dans ce domaine. Dans un second temps, nous présentons les méthodes d'extraction des caractéristiques faciales et les méthodes de sélection. Enfin, quelques bases de données utilisées pour la reconnaissance automatique des émotions sont présentées.

2.2 Définition des émotions

L'émotion est un phénomène complexe et multicomponentiel [12]. Les théoriciens de l'émotion ont étudié plusieurs de ses composantes au fil des années, permettant

ainsi de définir différents aspects de l'émotion :

- Un aspect expressif constitué des expressions faciales, des intonations vocales et des expressions corporelles (gestes).
- Un aspect physiologique comme les changements de la fréquence respiratoire, de la température corporelle, du flux sanguin, la production de sueur...
- Un aspect cognitif étroitement lié à la compréhension par l'individu de la scène et sa propre évaluation du stimulus extérieur.
- Un aspect subjectif lié au ressenti par les individus.

Une grande variété de définitions des émotions existe, suivant l'angle selon lequel nous les considérons. Ces différences ont motivé Kleinginna et Kleinginna [13] à proposer une définition qui englobe les différentes facettes des émotions. L'émotion selon Kleinginna et al [13] "*est une interaction complexe entre des facteurs objectifs et subjectifs, présentée par un système neural/hormonal qui donne naissance à des expériences affectives telles que le sentiment d'excitation, de plaisir/déplaisir. Elle génère un processus cognitif comme les effets perceptuels émotionnels, le jugement et le processus de labellisation. Elle active un vaste ensemble de régulation pour les conditions d'excitation et entraine un comportement souvent, mais pas toujours, expressif, dirigé vers un but et adaptatif*" [13].

2.3 Les théories et représentations des émotions

Plusieurs études se sont intéressées aux émotions, ce qui a permis de proposer une variété de théories. Nous abordons dans cette section les théories les plus connues à savoir la théorie de l'universalité des émotions qui suppose que l'émotion est universelle, dotée de fonctions adaptatives et de fonctions communicatives, et une théorie physiologique liée à l'acheminement temporel des changements corporels et l'état subjectif (l'émotion ressentie).

2.3.1 Théorie de l'universalité

Dans la continuité de sa théorie de l'évolution, Darwin présente dans son livre *"L'expression des émotions chez l'homme et les animaux"* [14] le rôle adaptatif des émotions permettant la survie des espèces et postule également que les émotions sont imprimées dans le système nerveux humain, rendant ainsi les émotions universelles. Darwin classe les changements des expressions faciales en sept groupes

pour lesquels les déformations du visage ont été décrites d'une façon détaillée. Plusieurs chercheurs contemporains ont essayé de prouver la théorie de l'universalité pour un ensemble restreint d'émotions appelé les émotions de base.
Les émotions de base, encore appelées émotions discrètes, émotions primaires ou émotions fondamentales, sont souvent représentées comme des émotions innées et indépendantes des cultures. D'après Ekman [15] les émotions de base ont des caractéristiques uniques et leurs réactions sont préprogrammées.
Les émotions jouent également un rôle communicatif. Un échange non-verbal peut passer par le biais des expressions émotionnelles permettant de régulariser le comportement du récepteur ou d'informer sur l'état émotionnel de l'individu. Le caractère universel de l'émotion communiquée facilite sa compréhension.

2.3.2 Théorie physiologique

Le séquencement des changements corporels et l'état subjectif ont constitué l'objet d'une contestation qui a opposé la théorie *périphéraliste* de James à la théorie *centraliste* de Cannon.
La théorie défendue par William James [16] et soutenue par Carl Lange [17] propose que les réponses du système nerveux périphérique (sueurs, augmentation de la fréquence respiratoire, changement de la chaleur corporelle, larmes...) sont à la base de l'expérience émotionnelle, autrement dit les réponses corporelles ou physiologiques face à une situation externe sont la source de la sensation de l'émotion. Ainsi, les émotions sont liées à un pattern [1] périphérique spécifique qui déclenche l'émotion lors de son activation. La théorie de James a été supportée par des recherches plus récentes telles que les travaux de Levenson et al [18]. Les expériences menées ont montré que la configuration volontaire d'une expression faciale émotionnelle engendre son ressenti émotionnel chez des sujets spécialistes des émotions (acteurs et scientifiques) et chez des sujets non spécialistes.
A l'opposé, Cannon [19] propose la théorie de l'émotion thalamique (centrale), confirmée par Philip Bard [20]. Il avance que le système nerveux central est à l'origine de l'expérience émotionnelle déclenchant par la suite les changements corporels et physiologiques. Les expériences menées par Cannon ont mis en évidence que l'état émotionnel persiste même en rendant les réponses du système nerveux périphérique impossible.

1. Modèle spécifique à un comportement

2.3.3 Représentation des émotions

Les émotions sont représentées généralement soit de façon discrète (catégorielle) soit de façon continue (dimensionnelle). Ces deux représentations ont fait l'objet d'une grande controverse opposant deux courants de la psychologie.

Représentation discrète

La représentation discrète catégorise certaines émotions dans des groupes prédéfinis ayant des caractéristiques distinctes les unes des autres. En général, ces émotions appelées émotions de base ou encore émotions fondamentales sont considérées par la plupart des théoriciens comme universelles et innées. En plus de ces deux critères Ekman définit sept autres critères pour caractériser les émotions fondamentales [3], tels que la présence de la même émotion chez un autre primate ou la présence de caractéristiques physiologiques qui la distinguent des autres émotions...
Les chercheurs qui adoptent le concept des émotions de base présentent un ensemble restreint d'émotions, mais leurs recherches divergent lorsqu'il s'agit de les identifier. Les travaux ci-dessous présentent les émotions de base selon chaque auteur :

- Izard [21] : colère, surprise, dégoût, joie, peur, tristesse, mépris, intérêt, culpabilité, honte.
- Ekman [3] : colère, peur, tristesse, joie, dégoût, surprise.
- Tomkins [22] : colère, dégoût, joie, peur, surprise, mépris, honte, intérêt, anxiété.

La représentation discrète est très présente dans les systèmes de reconnaissance automatique, notamment l'ensemble des émotions proposé par Ekman. Plusieurs bases de données préparées pour la reconnaissance des émotions catégorisent les expressions faciales avec les six émotions de bases proposées par Ekman. C'est le cas des bases Cohn-Kanade (et de sa version étendue CK+), FEEDTUM, MMI... Des représentations des expressions faciales définissent des prototypes spécifiques aux six émotions de base comme le système FACS et le standard MPEG-4. Plus de détails sur ce point seront fournis ultérieurement dans ce chapitre.

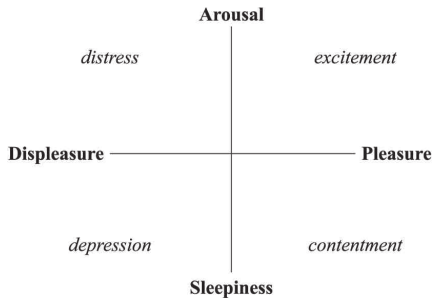


FIGURE 2.1 – Modèle circulaire de Russell [23]

Représentation continue

La représentation dimensionnelle présente les émotions dans un espace continu basé sur deux ou plusieurs dimensions. Contrairement à la représentation catégorielle, les émotions sont définies en se basant seulement sur les dimensions. Russell [23] définit les états affectifs sur un modèle bidimensionnel, représenté par un cercle basé sur un axe de plaisir (plaisir/peine) et un axe quantifiant la force du ressenti. La figure 2.1 illustre le modèle de Russell avec huit états affectifs. Toutes les émotions peuvent ainsi être définies sur ce modèle ; l'excitation par exemple est un état associant un taux de plaisir élevé à un taux d'activation élevé. D'après Russell, c'est un état qui se trouve à un angle de 45^o en considérant que le plaisir est l'angle 0^o et l'état d'éveil est l'angle 90^o. D'autre part, Cowie et al [24] proposent un outil "Feeltrace" pour l'annotation des émotions sur un modèle bidimensionnel très proche de celui de Russell, qui se base sur l'activation et l'évaluation (positif-négatif).

Une autre manière de représenter les émotions a vu le jour avec le modèle de Plutchik [25]. Ce dernier présente un modèle qu'on peut qualifier d'hybride, puisqu'il définit huit émotions de base tout en gardant en partie la notion de dimension. Huit émotions de base à savoir colère, dégoût, peur, joie, tristesse, surprise, acceptation et anticipation sont représentées sur un cercle d'émotions de manière à avoir quatre ensembles d'émotions opposées sans pour autant les définir par des dimensions. Les émotions secondaires sont également présentées dans son modèle. Elles sont définies par des combinaisons des émotions de base adjacentes. Une dimension d'intensité détermine également l'intensité des émotions de base, par exemple l'extase résulte de l'augmentation de l'intensité de la joie et la sérénité résulte de la diminution de l'intensité de la joie. La figure 2.2 illustre le modèle de Plutchik.

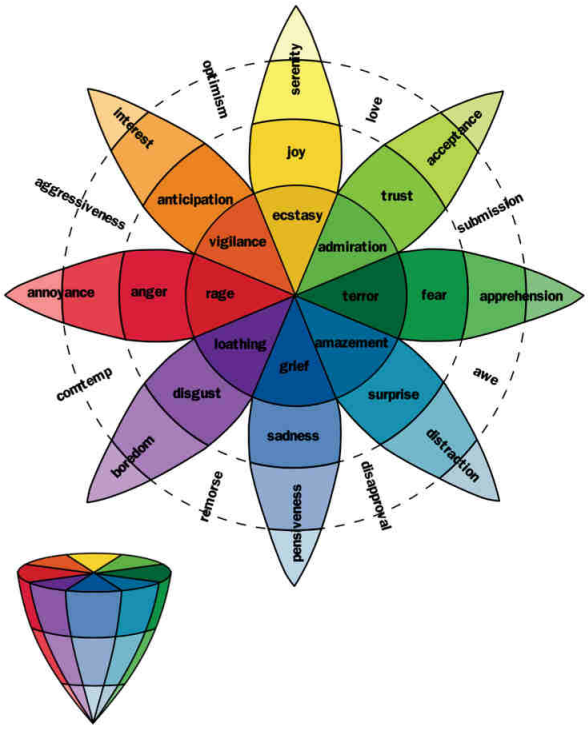


FIGURE 2.2 – Modèle de Plutchik [26]

La représentation continue a également servi à représenter des émotions pour des applications de reconnaissance automatique. Soladié et al [27] synthétisent huit expressions pour chaque sujet à partir des déformations acquises lors de l'apprentissage. La reconnaissance de l'expression faciale est effectuée par sa projection sur l'espace des huit expressions. La base utilisée pour l'apprentissage et le test du système est la base AVEC du chalenge AVEC2012, qui lui présente les émotions sur quatre dimensions. Karpouzis et al [28] représentent les émotions dans les cercles d'émotions de Whissell et Plutchik. Ils modélisent également des émotions intermédiaires aux émotions de base.

Des bases dédiées à la reconnaissance automatique des émotions labellisent les images en considérant la représentation dimensionnelle. Nous citons à titre d'exemple la base SEMAINE représentant cinq dimensions à savoir *valence* (positif/négatif), *Activation* (évaluation de la dynamique de la personne), *Power* (force et contrôle de la situation), *Anticipation/Expectation* (anticipation/attente), *intensity* (intensité de l'état exprimé).

2.3.4 Conclusion

Dans cette section, nous avons cité les théories les plus connues dans les travaux de recherche sur les émotions. La théorie de l'universalité a facilité la représentation des émotions, notamment la représentation discrète. D'autre part, les théories physiologiques ont soulevé la question du séquencement entre le ressenti et les changements corporels. Ces différents points de vues mettent l'accent sur les expressions simulées et leurs rôles. Levenson et al [18] montrent que la simulation d'une expression engendre le ressenti de son émotion. Ceci peut être également considéré lors de la construction des bases de données des émotions.
Les grands courants de recherche sur les émotions et les différents modèles définis pour les représenter ont beaucoup inspiré les travaux sur la reconnaissance automatique des émotions que nous présentons dans la section suivante.

2.4 La reconnaissance automatique des émotions

Composées de différents aspects, les émotions sont définies dans les systèmes de reconnaissance automatique des émotions à travers leurs aspects mesurables à savoir l'aspect physiologique et l'aspect expressif.
L'aspect physiologique des émotions se manifeste par les changements corporels accompagnant le changement de l'état subjectif comme le changement de la température, de la fréquence cardiaque, de la fréquence respiratoire. La détection de ces aspects nécessite des mesures avec des capteurs liés aux sujets, ce qui s'avère être intrusif. Notre travail s'inscrit dans le cadre d'une reconnaissance des émotions sans l'utilisation d'outils intrusifs.
D'autre part, l'aspect expressif est composé de plusieurs canaux à savoir l'expression vocale, l'expression via des gestes et l'expression faciale. Dans le cadre de cette thèse, nous ne prenons en considération que la reconnaissance des émotions à partir des expressions faciales. L'expression faciale est en effet un élément clé dans l'étude et la compréhension des émotions [29]. Plusieurs chercheurs ont privilégié le visage pour l'étude des états émotionnels. Darwin [14] s'est particulièrement intéressé à l'étude des changements des expressions engendrées par les émotions. Tomkins, Izard et Ekman ont par ailleurs été parmi les premiers psychologues à s'intéresser aux expressions faciales notamment les expressions émotionnelles.
Les expressions faciales remplissent plusieurs fonctions [30]. Elles présentent :

- des expressions spécifiques à l'état mental telles que la réflexion, la conviction, l'émotion ressentie...,
- la communication non verbale (clins d'œil, grimaces pour passer un message),
- la communication verbale (mouvement de la bouche),
- des états physiologiques (fatigue, peine...).

Cependant, l'une des difficultés majeures dans la reconnaissance d'une expression faciale, et notamment d'une expression émotionnelle, est de s'affranchir des différences inter-personnes. Ainsi, il faut prendre en considération l'existence de critères liés à l'identité, et par conséquent spécifiques à la personne, et de critères spécifiques à l'expression émotionnelle.
Des solutions pour rendre les informations de l'expression indépendantes de l'identité existent. Le standard MPEG-4 présente par exemple des unités définies sur le visage neutre (FAPU). Ces paramètres normalisent les paramètres des actions faciales (FAPs) calculés sur l'expression, ainsi la dépendance à l'identité est éliminée. Différentes approches ont été utilisées dans la littérature pour gérer la dépendance à l'identité des sujets. Abdat et al [31] normalisent les distances caractéristiques calculées dans l'image de l'expression par les distances calculées dans l'image neutre. Soladié et al [32] présentent un modèle spécifique à chaque sujet : à partir de son visage neutre, les différentes expressions faciales sont synthétisées.

2.5 Représentation des expressions faciales émotionnelles

Les représentations faciales relient les mouvements des muscles faciaux aux émotions exprimées. Elles représentent une sorte de dictionnaire utile à la reconnaissance des émotions. Dans cette section nous exposons trois catégories de représentations des expressions faciales émotionnelles.

2.5.1 Représentation par le système FACS

Dans le cadre de la représentation discrète des expressions faciales, Ekman a proposé un système purement objectif pour le codage des expressions faciales en quantifiant le mouvement des muscles. Ce système, appelé FACS (*Facial Action Coding*

System), permet de s'affranchir la labellisation subjective des expressions souvent biaisée par l'influence culturelle de l'observateur. Le système FACS a prouvé sa fiabilité dans la reconnaissance des expressions faciales en liaison avec plusieurs thèmes de psychologie comme la détection des mensonges [33], la mesure des émotions spontanées pour des patients ayant des lésions cérébrales [34], la relation entre expressions faciales et lésions coronariennes [35], la relation entre les expressions faciales et le comportement chez l'enfant [36]...
Développé par Ekman et Friesen [37], le système FACS est composé de 44 actions unitaires. Chaque action unitaire (AU) correspond à un mouvement des muscles faciaux. L'intensité de la contraction des muscles est également codée sur cinq niveaux. Les émotions de base définies par Ekman ont été décrites par des prototypes spécifiques à chaque expression d'émotion. A titre d'exemple, la tristesse est représentée par AU1+AU4+AU15+AU44. La figure 2.3 présente la description de la tristesse par l'un des prototypes formés des actions unitaires.

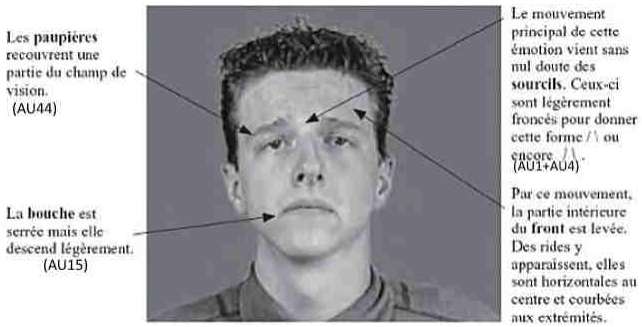


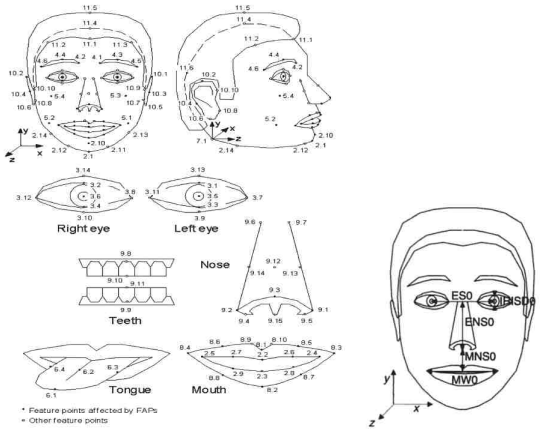
FIGURE 2.3 – Expression prototypique de la tristesse [38]

Ce système de codage est largement utilisé dans les travaux de reconnaissance automatique des expressions et des émotions. Il permet à la fois de décortiquer l'expression faciale en ses fines actions musculaires et ses différentes phases (onset, apex et offset) [39]. Plusieurs travaux ont étudié la reconnaissance des actions unitaires du système FACS afin d'avoir une reconnaissance des expressions faciales, notamment des expressions émotionnelles, de façon plus objective. Bartlett et al [39] ont utilisé une méthode hybride, qui combine la détection des caractéristiques, les flux optiques et l'analyse en composante principales (ACP) pour la reconnaissance des sept premières actions unitaires de AU1 à AU7. Quelques années plus tard, Bartlett et al [40] développent une méthode plus évoluée pour la reconnaissance de 20 AUs représentant des expressions spontanées. Les textures du visage sont extraites par 72 filtres de Gabor et elles sont classées par les Sépa-

rateurs à Vaste Marge et Adaboost. Les résultats obtenus sont assez prometteurs. Le taux de reconnaissance est de 93%. Tian et al [41] détectent les caractéristiques permanentes du visage (les yeux, les lèvres, les sourcils...) et les caractéristiques temporaires dues aux changements faciaux pour la reconnaissance de 16 AUs à partir d'une séquence d'images.

La détection des phases temporelles a été également le sujet de travaux de reconnaissance automatique des AUs. Valstar et Pantic [42] proposent une approche qui permet la reconnaissance de 15 AUs et de leurs phases temporelles (onset, apex et offset). Pour capturer la dynamique des actions unitaires (AU1, AU2, AU4, AU5, AU10, AU12, AU14, AU20), Yang et al [43] se basent sur les descripteurs de Haar dynamiques ainsi que la technique d'Adaboost. Une façon intéressante pour calculer la progression temporelle des AUs, indépendamment de l'état neutre, est proposée par Khademi et al [44]. Une fusion des caractéristiques d'apparence extraites par les ondelettes de Gabor et des caractéristiques géométriques extraites par le mouvement des points est appliquée par une analyse canonique à noyau de la corrélation (*Kernel Canonical Correlation Analysis (KCCA)*). En utilisant la différence entre l'intensité de l'AU dans l'image courante et les images aux voisinages, la progression temporelle est détectée. Appliquée sur trois bases d'images à savoir CK+, Bosphorus et DISFA, cette méthode a montré son efficacité dans la reconnaissance de 15 AUs et leurs évolutions temporelles. Les scores F1 obtenus pour la base CK+, la base Bosphorus et la base DISFA sont respectivement 0.79, 0.68 et 0.8. Zhang et al [45] combinent, en se basant sur les réseaux dynamiques bayésiens, des caractéristiques d'apparence et des caractéristiques géométriques pour la détection dans un premier lieu de plusieurs AUs, qui sont à leur tour combinées pour la reconnaissance des émotions spontanées. Les caractéristiques visuelles appliquées dans l'image précédente sont introduites dans les réseaux bayésiens pour améliorer la robustesse de la reconnaissance des émotions spontanées.

Le système FACS a cependant été critiqué, Essa et al [46] estiment que la diversité des expressions a été ignorée. Ils ajoutent que FACS ne détecte pas les variations subtiles des changements faciaux, les AUs étant localisées dans des zones prédéfinies.



(a) (b)

FIGURE 2.4 – Modèle du visage par le standard MPEG-4. (a) Représentation des points caractéristiques FAs. (b) Modèle des paramètres unitaires FAPUs

2.5.2 Représentation par le standard MPEG-4

Un deuxième codage très utilisé pour la description des expressions faciales est défini par les paramètres du standard de compression MPEG-4 [47]. Utilisé principalement pour l'animation et la synthèse du visage, MPEG-4 définit un ensemble de paramètres (FAPs : *Facial Animation Parameters*) nécessaire au codage de la déformation d'un visage à l'état neutre [48]. Les valeurs des paramètres d'animation FAPs sont proportionnelles à l'amplitude de la déformation faciale. Ces paramètres sont représentés par un total de 84 points caractéristiques (FPs : *Feature Points*) définis sur un visage neutre (voir la figure 2.4 (a)). Le modèle ainsi défini doit être applicable sur des visages de différentes tailles. Pour ce faire, les paramètres d'animation faciale FAPs sont définis en fonction des unités de paramètres d'animation faciale (FAPU : *Face Animation Parameter Units*). Ces dernières sont les rapports des distances entre les caractéristiques faciales clés d'un visage à l'état neutre. La figure 2.4 (b) illustre ces distances. Comme le codage FACS, le codage par FAPs est très lié aux changements des muscles faciaux. Il définit également six expressions faciales émotionnelles qui coïncident avec les émotions de base définies par le système FACS.

Plusieurs recherches dans le domaine de la reconnaissance automatique des émotions s'appuient sur le codage MPEG-4 pour la représentation des changements du visage. Pardãs et al [49] extraient d'abord les FAPs pour la représentation des mouvements faciaux sur une séquence vidéo, puis les analysent avec un HMM semi

continu. Les résultats obtenus sur des séquences extraites de la base Cohn-Kanade montrant les expressions simulées des six émotions de base, donnent un taux de 84% pour des sujets non inclus dans l'apprentissage. Cependant, le test sur des séquences plus longues donne des taux de reconnaissance plus faibles.
En se basant sur l'analyse des variations des FAPs, Ioannou et al [50] présentent des règles définies pour les émotions discrètes [51] et des règles définies pour l'espace affectif continu. Des techniques d'extraction de caractéristiques faciales sont mises en place. Chaque caractéristique extraite est accompagnée d'un taux de confiance qui se base sur des règles anthropométriques. Les neurones flous sont créés en se basant sur les règles définies par les variations des FAPs. Une étape d'adaptation est proposée pour intégrer de nouveaux sujets en ajoutant dans la base la définition des règles initialisées pour les sujets en question.

2.5.3 Représentation par catégorie

La représentation par catégorie d'émotion est adoptée par plusieurs travaux. Une certaine liberté est donnée au système pour apprendre au fur et à mesure, à partir des données présentées, les déformations liées à chaque type d'émotions. Certes cette technique est liée aux données, mais elle est plus efficace puisque la détection et la reconnaissance des unités des représentations FACS et FAPs entrainent une incertitude dans la reconnaissance de l'émotion.
Plusieurs méthodes utilisent ce mode de représentation. Les descripteurs sont d'abord extraits dans des vecteurs caractéristiques. L'étape d'apprentissage se charge alors de construire un ou plusieurs modèles spécifiant le comportement des expressions émotionnelles. Lors du test le problème se réduit à une classification d'un vecteur de descripteurs dans des groupes d'émotions. Dans ce cadre, Bartlett et al [52] présentent un système assez robuste qui se base sur Adaboost et les Séparateurs à Vaste Marge pour réduire et classer des vecteurs descripteurs de Gabor. Ce système a su faire face aux changements inter-sujets et a été intégré dans trois applications différentes, tel le *CU animator* qui consiste à reproduire les mimiques des sujets. Le système a été également intégré dans le robot domestique *Aibo robot* et dans le robot *ATR's RoboVie*. La comparaison entre deux systèmes de reconnaissance des expressions émotionnelles proposés par Kotsia et al [53] montre que le système utilisant la représentation FACS a des taux de reconnaissance inférieurs au système qui utilise directement un apprentissage des

caractéristiques (représentation par catégorie).
Nous adoptons ce mode de représentation dans la suite de nos travaux. Les déformations de chaque classe d'émotion sont modélisées à partir de l'étape d'apprentissage, sans passer par une représentation prédéfinie.

Les systèmes de reconnaissance des expressions, notamment les expressions émotionnelles, sont constitués de trois étapes principales [54] :

1. Acquisition du visage : cette étape peut être réalisée soit par une détection du visage soit par une estimation de la pause de tête.
2. Extraction des informations faciales liées à l'expression : ces informations sont liées soit à l'apparence de l'expression, soit à la géométrie des déformations. Dans le cadre d'une séquence d'images, l'information liée à la dynamique de l'expression est également utile.
3. Reconnaissance de l'expression : cette étape est effectuée par une classification dans le cas d'une représentation discrète des émotions ou par une régression dans le cas d'une représentation continue des émotions.

Dans le cadre de notre thèse, nous nous focalisons sur l'extraction des descripteurs et leurs rôles dans la reconnaissance des émotions.

2.6 Extraction des descripteurs

Les descripteurs faciaux sont d'une importance majeure dans la reconnaissance automatique des émotions. Ils décrivent les changements faciaux à la base de l'expression des émotions. Il existe deux classes principales dans la littérature à savoir les descripteurs géométriques et les descripteurs d'apparence. Nous nous intéressons dans cette section à la présentation des approches les plus utilisées dans les deux classes.
Un descripteur doit remplir plusieurs critères pour réaliser une description de l'expression émotionnelle :

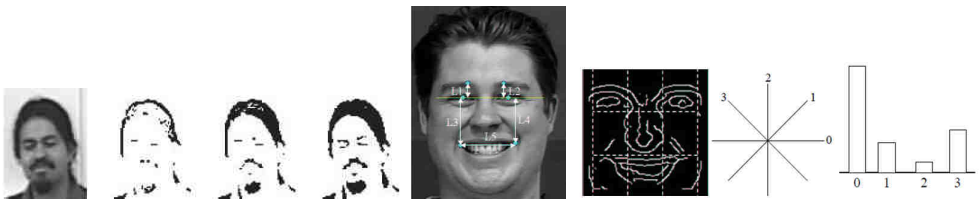
- décrire les informations nécessaires et suffisantes pour la reconnaissance des émotions,
- être robuste face aux changements de luminance,
- détecter des caractéristiques dans le cadre de faibles résolutions.

2.6.1 Descripteurs géométriques

L'expression faciale engendre des déformations géométriques sur le visage. Ces déformations se manifestent au niveau des changements de la forme des caractéristiques permanentes comme la bouche, le nez, les sourcils et par l'apparition de caractéristiques transitoires comme les rides d'expression. La représentation des déformations peut également être décrite par des points caractéristiques ou par des modèles géométriques qui détectent le mouvement des muscles faciaux.

Les formes des caractéristiques du visage

La détection de la forme, de la position et de l'état des caractéristiques du visage nécessitent des informations à priori. Cet ensemble d'informations doit être stable pour différentes conditions, comme le changement de résolution [55]. Dans le cadre de la reconnaissance de l'état émotionnel d'une personne dans une vidéo à basse résolution, comme les vidéos de réunions à distance, Tian et al [55] développent des techniques pour la détection des positions et des formes des caractéristiques permanentes du visage. Les yeux et les sourcils sont détectés par un seuillage itératif (voir figure 2.5 (a)). La position de la bouche est détectée par la projection de ses contours. En utilisant des informations à priori, Tian et al [55] précisent six points pour localiser ces caractéristiques et calculer les distances les séparant (voir figure 2.5 (b)). La forme de la bouche est également détectée par le calcul d'un histogramme présentant les quatre orientations illustrées dans la figure 2.5 (c). Appliquée sur la base PETS2003 qui contient des vidéos de faible résolution, cette méthode a montré certaines faiblesses lorsqu'il s'agit de très faibles résolutions entrainant des confusions entre les émotions et l'état neutre.



(a) (b) (c)

FIGURE 2.5 – Techniques utilisées dans [55]. (a) Seuillage itératif pour la détection des yeux et des sourcils, (b) Distances définies par les positions des caractéristiques permanentes du visage, (c) Quantification de la forme de la bouche par le calcul de l'histogramme des directions des contours

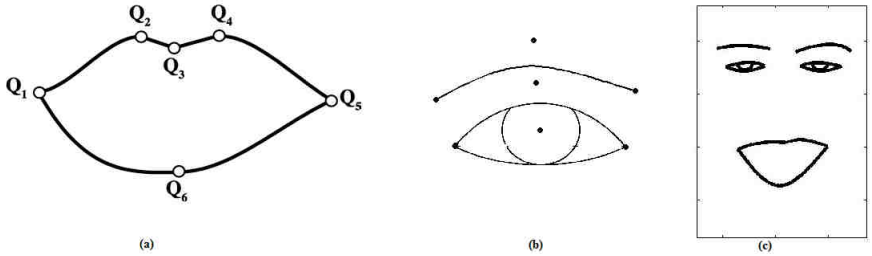


FIGURE 2.6 – Modèles paramétriques pour la représentation des caractéristiques du visage. (a) Modèle paramétrique de la bouche, (b) Modèles paramétriques des yeux et des sourcils, (c) Squelette du visage

La segmentation s'affiche comme une solution pour l'extraction des caractéristiques permanentes du visage. De multiples méthodes sont utilisées pour réaliser ce but. Khandait et al [56] couplent à la fois les traitements morphologiques du visage et la détection des contours des caractéristiques par l'algorithme SUSAN dans les zones locales à chaque caractéristique. Une fois la segmentation terminée, les changements des formes des caractéristiques sont alors détectés par la variation de leurs dimensions. Leurs positions sont calculées par les variations des distances entre les caractéristiques.

Des solutions plus performantes pour la segmentation et la détection des états des caractéristiques permanentes sont données par les modèles paramétriques [57]. Hammal et al [57] proposent des modèles paramétriques pour former un squelette du visage. La bouche est modélisée par quatre courbes cubiques (voir figure 2.6 (a)), les contours supérieurs des yeux et les sourcils sont modélisés par des courbes de Bézier contrôlées par trois points, le contour du bas des yeux est défini par un contour parabolique contrôlé également par trois points (figure 2.6 (b)). En se basant sur ce squelette formé par les modèles paramétriques (voir figure 2.6 (c)), Hammal et al [58] définissent une représentation de quatre expressions à savoir l'expression neutre, la joie, le dégoût et la surprise, qui ressemble fortement à la représentation donnée par le standard MPEG-4. La théorie des croyances est alors adoptée pour classer les changements faciaux en une des quatre émotions ou une union de deux émotions. Le cas de doute entre les émotions étant considéré, il permet l'ajout de caractéristiques transitoires (les rides qui apparaissent au-dessus du nez) dans le système.

Caractéristiques géométriques basées sur les points caractéristiques

Différentes méthodes sont basées sur l'extraction d'informations telles que l'état et la forme des caractéristiques faciales, ce qui les rend difficiles à mettre en œuvre et coûteuses en temps de calcul. De plus la ressemblance des états et des formes des caractéristiques peut engendrer une confusion entre certaines émotions comme par exemple la joie et le dégoût [58]. Dans ce qui suit, nous présentons des techniques, plus simples que les techniques présentées précédemment, basées sur le mouvement d'un ensemble de points caractéristiques pour le codage de l'expression.
Les points caractéristiques décrivent les expressions soit par leurs déplacements, soit par la variation des distances entre eux. Le déplacement des points caractéristiques nécessite deux étapes à savoir l'extraction de leurs positions et leur suivi. Pantic et al [59] proposent un modèle géométrique basé sur 32 points caractéristiques définis sur la face frontale et le profil (voir figure 2.7). Les points sont extraits après application d'une méthode hybride de contours sur les caractéristiques permanentes du visage. Les mouvements des muscles sont ensuite codés par le déplacement des points par rapport à l'image neutre. Des règles sur le déplacement des points sont définies pour la reconnaissance des AUs et des expressions faciales. Le même modèle géométrique a été utilisé dans [60] avec seulement 12 points pour la reconnaissance de sourires posés et de sourires spontanés. Le déplacement des points est fusionné avec la pose de la tête ainsi que le mouvement des épaules.

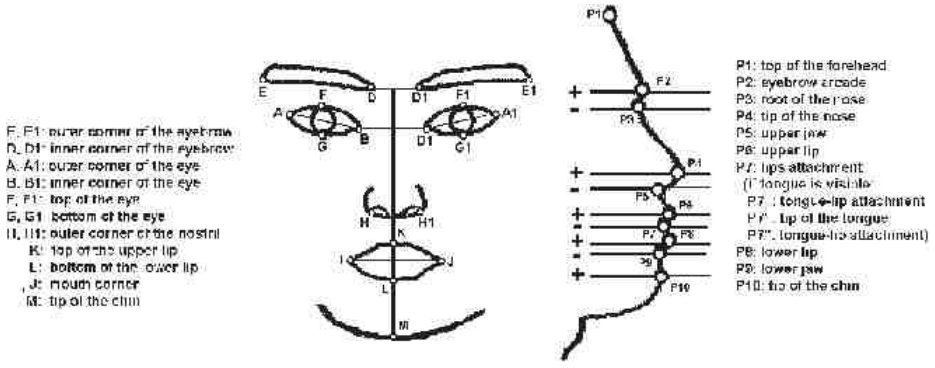


FIGURE 2.7 – Modèle géométrique des points caractéristiques proposé dans [59]

Bien d'autres méthodes sont également basées sur le déplacement des points dans le cadre d'applications temps réel. Cid et al [61] présentent une méthode basée sur l'approche bayésienne, dans le cadre de la reconnaissance et l'imitation des expressions émotionnelles (joie, colère, peur, tristesse) et de la pose de tête d'un

interlocuteur lors de son interaction avec un robot. Les caractéristiques géométriques sont extraites par l'algorithme *Good feature to track* puis suivies par l'algorithme Lukas-Kanade. Le déplacement de ces points sert de codage pour les AUs qui forment l'expression faciale. La projection de cet ensemble de points sur une surface cylindrique modélise la pose de la tête. Le système dépasse un taux de reconnaissance de 84% pour les quatre émotions.

Les méthodes basées sur la variation des distances représentent soit les distances entre tous les points caractéristiques [62], soit les distances entre les couples de points correspondant aux mouvements des muscles faciaux [31].

Ghimire et al [62] extraient un ensemble de points caractéristiques par le graphe de regroupement élastique (*elastic bunch graph*) de l'image neutre et appliquent un suivi par les jets définis par les ondelettes de Gabor dans chaque image. Les distances calculées lient tous les points caractéristiques. Adaboost est ensuite utilisé pour la sélection des distances les plus discriminantes (voir figure 2.8). Deux méthodes ont été testées pour la classification des six émotions de base dans la base Cohn-Kanade. Les Séparateurs à Vaste Marge calculés sur les distances sélectionnées par Adaboost dépassent 97% de taux de reconnaissance. La seconde technique de classification se base sur la similitude du vecteur caractéristique et du prototype de chaque émotion à l'aide d'Adaboost multi-classe et de la déformation temporelle dynamique (*Dynamic time warping DTW*). Le taux de reconnaissance obtenu atteint 95%.

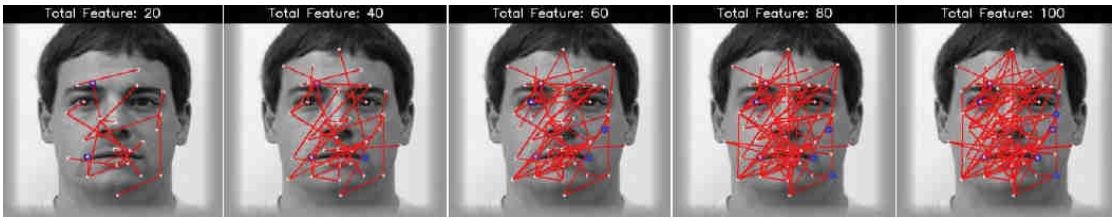


FIGURE 2.8 – Distances caractéristiques après sélection par Adaboost [62]

Abdat et al [31] détectent 38 points divisés en points fixes et points mobiles. Les points fixes sont localisés sur le contour du visage et les zones invariantes par l'expression, alors que les points mobiles sont localisés sur les zones variables liées à l'expression notamment les caractéristiques du visage (sourcils, yeux, bouche, nez). Les distances entre les points mobiles et les points fixes sont calculées et normalisées par les distances de l'image neutre. Chaque distance représente le mouvement d'un muscle. Cette méthode atteint des taux de reconnaissance élevés, 95.8% pour la base Cohn-Kanade et 97.5% pour la base FEEDTUM. Elle est

présentée d'une façon plus détaillée dans le chapitre suivant.

Caractéristiques géométriques basées sur les modèles

Des modèles pour la détection des déformations des muscles et des changements des états des caractéristiques sont utilisés pour la reconnaissance des expressions émotionnelles. La définition et l'initialisation de ces modèles se fait généralement sur un visage neutre. Ils codent souvent le mouvement des muscles soit par le déplacement des nœuds placés sur les caractéristiques du visage [53], soit en le transformant en une information 3D, qui renseigne sur l'intensité du mouvement [63] [64].

Pour quantifier le mouvement facial, Kotsia et al [53] utilisent les nœuds de la grille de Candide, initialisés manuellement sur l'image neutre. Les nœuds sélectionnés sont suivis au cours de la séquence par la méthode Kanade-Lucas-Tomasi (KLT). Le déplacement géométrique de ces nœuds est ensuite calculé comme la différence entre les nœuds présents dans la première fenêtre et les nœuds détectés dans la fenêtre de l'expression faciale maximale. Les Séparateurs à Vaste Marges multi-classes sont enfin utilisés pour la classification des déplacements des nœuds. Un taux dépassant 99% est obtenu pour la reconnaissance des six émotions fondamentales dans la base Cohn-Kanade.

Sebe et al [63] proposent la déformation du volume de Bézier (*piecewise Bézier volume deformation (PBVD)*) pour la détection et le suivi des caractéristiques du visage. Ce modèle définit 16 surfaces qui s'adaptent au visage par des points de contrôle. La déformation de chaque caractéristique faciale est détectée et leurs mouvements sont mesurés entre images consécutives. Ces mouvements sont transformés dans un espace 3D permettant de donner des mouvements unitaires (MUs) renseignant sur l'activation et la direction des muscles faciaux participant à l'expression ainsi que l'intensité de leurs mouvements (voir figure 2.9). Sebe et al testent leur méthode de reconnaissance des émotions sur une base d'émotions spontanées contenant les émotions joie, surprise, dégoût et l'expression neutre et sur une base d'expressions posées (Cohn-Kanade). Des taux de reconnaissance intéressants sont obtenus, dépassant 91% pour les deux bases.

Benli et al [64] utilisent un modèle facial, qui comporte 612 nœuds et 1128 polygones. Ce modèle inclut 18 muscles faciaux représentés linéairement (voir la figure 2.10). En se basant sur ces muscles, les régions d'intérêt sont définies, permet-

FIGURE 2.9 – Modèle de déformation des volumes de Bézier [63]

tant ainsi l'initialisation des points caractéristiques dans l'image neutre. Le suivi de ces derniers est réalisé par les flux optiques et leur déplacement au cours de la séquence permet l'estimation de la pose de tête ainsi que le déplacement des sommets du polygone. L'estimation de la déformation du visage est également résolue par un système basé sur les contraintes de l'anatomie faciale, permettant ainsi la mesure du niveau d'activation des muscles. Une sélection de l'ensemble des muscles les plus descriptifs est appliquée. Le résultat obtenu pour ce système de reconnaissance sur une base d'expressions posées ne dépasse pas 77.6%.

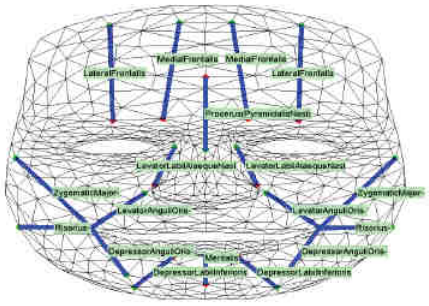
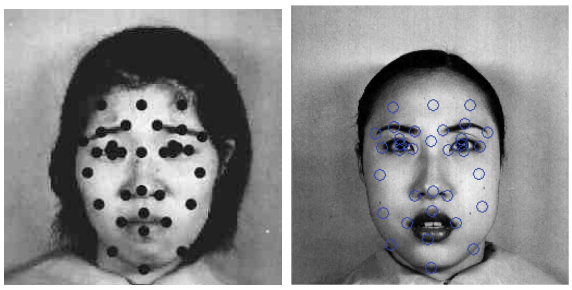


FIGURE 2.10 – Modèle de détection des déformations faciales [64]

Les méthodes géométriques présentent le mouvement des muscles soit par la détection de la forme des caractéristiques, soit par le mouvement des points caractéristiques, soit par des modèles qui détectent directement les mouvements des muscles. L'efficacité de ces méthodes devient cependant critique lorsqu'il s'agit d'un faible mouvement des muscles. La section suivante présente un autre aspect de l'expression, qui se base sur la description de son apparence.

2.6.2 Descripteurs d'apparence

Les expressions faciales notamment les expressions émotionnelles engendrent des changements transitoires sur le visage impactant la texture autour de la bouche, les rides sur le front, les côtés des yeux et bien d'autres zones qui dépendent en général de l'élasticité de la peau et de l'expression. En effet, suivant l'expression, différentes zones du visage sont sollicitées. D'après Raouzaiou et al [51], la tristesse est seulement décrite grâce aux mouvements des paupières qui se traduisent dans le standard MPEG4 par (F19, F20, F21, F22) et aux mouvements des sourcils (F31, F32, F33, F34, F35, F36), tandis que le dégoût est décrit par le mouvement de la mâchoire (F3) (qui ne traduit pas forcement l'ouverture de la bouche), le mouvement de plusieurs points de la lèvre (F4, F5, F8, F9,F55, F56, F57, F58, F59), le mouvement de la paupière (F19, F20, F21, F22) et le mouvement des sourcils (F33, F34).
Se pose alors la question du choix du niveau de traitement : "Faut-il détecter les textures locales ou extraire la texture globale du visage ?". Plusieurs méthodes d'extraction de texture ont été utilisées dans la littérature pour la reconnaissance des émotions. Parmi celles-ci nous pouvons citer les filtres de Gabor [65] [66] [67] [68], les filtres de Gabor logarithmiques [69] [70], la méthode des motifs binaires locaux (*local binary patterns (LBP)*) [9] [71] [70], l'analyse par composantes principales [72], les flux optiques [39] [73], la méthode de transformation de caractéristiques visuelles invariantes à l'échelle (*scale-invariant feature transform (SIFT)*) [74] [75]. La grande majorité de ces méthodes est utilisée pour les deux niveaux de traitement, localement autour de points caractéristiques et sur la totalité du visage. Ces deux niveaux de traitement ont leurs avantages et leurs inconvénients. Les méthodes d'extraction des informations locales au niveau des points sont plus intéressantes en terme de temps de calcul et en terme de mémoire utilisée puisque leurs vecteurs descripteurs sont moins volumineux que ceux des méthodes globales. Cependant, la robustesse de ces méthodes reste fortement liée à la précision de la position des points, qui peut être soit manuelle [65] [66], soit automatique [75]. Les méthodes globales extraient quant à elles toutes les textures du visage, permettant ainsi de ne pas négliger des zones indispensables à la reconnaissance de l'expression. Elles nécessitent cependant toujours une phase de prétraitement pour normaliser le visage afin d'éliminer les éventuelles erreurs causées par la pose de la tête et la différence de taille entre les visages.



(a) (b)

FIGURE 2.11 – Extraction des textures par filtres de Gabor appliqués localement sur des points choisis manuellement. (a) Localisation des points dans [65] (b) Localisation des points dans [66]

L'une des méthodes les plus utilisées pour l'extraction des textures est la méthode de Gabor. Elle a prouvé son efficacité dans plusieurs domaines comme la reconnaissance du visage [76] [77], la reconnaissance des empreintes [78], la détection de points caractéristiques [79]. Sa robustesse dans ces domaines a favorisé son application dans la reconnaissance des émotions. Dans [65], Zheng et al localisent manuellement 34 points caractéristiques sur différentes zones du visage comme le montre la figure 2.11 (a). Les filtres de Gabor appliqués sur chaque point donnent ainsi 1020 descripteurs qui constituent un vecteur de graphe étiqueté. En outre, un vecteur des taux sémantiques de chaque image est extrait à partir d'évaluations de psychologues. Une analyse de la corrélation canonique à noyau est ensuite effectuée sur les deux vecteurs. Lors de la reconnaissance des émotions, l'estimation du vecteur sémantique est alors calculée en se basant sur la corrélation canonique à noyau. L'expression est enfin labellisée selon l'émotion ayant le taux maximal dans le vecteur sémantique. Le résultat obtenu sur la base JAFEE, constituée seulement de sujets féminins simulant les six émotions de base (cette base est présentée dans la section 2.8), dépasse 98%, tandis que le résultat obtenu pour la base d'Ekman (une base d'émotions posées) atteint seulement le taux de 81%. Une autre approche proposée par Bashyal et al [66] applique 18 filtres de Gabor sur un ensemble de 34 points. Les points sont localisés par une interface utilisateurs (voir figure 2.11 (b)). Une analyse en composantes principales est ensuite appliquée pour réduire la taille des descripteurs obtenus. Une quantification vectorielle appliquée pour la classification des émotions donne un taux de reconnaissance de 90.2% pour la base JAFEE. Les points extraits dans les deux méthodes sont positionnés sur les yeux,

sur la bouche, sur les sourcils et sur le contour du visage. Ce faible nombre de points suffit-il vraiment à coder toute l'information pertinente pour la reconnaissance des émotions notamment les émotions spontanées ?
Dans le cadre de l'amélioration de la reconnaissance des émotions spontanées, Grimm et al [80] proposent une approche traitant localement une région centrée sur les yeux et une région centrée sur la bouche. Ils appliquent un banc de 18 filtres de Gabor sur chaque région, suivi d'une analyse des composantes principales. La reconnaissance est effectuée dans les deux cas de représentation des émotions avec un réseau de neurones. Pour la représentation discrète, le taux de reconnaissance obtenu avec la combinaison des deux vecteurs locaux est de 68.2% et pour la représentation continue le taux de reconnaissance est de 75%. Les auteurs proposent alors de procéder à une description plus fine des régions, qui s'effectue en deux couches. Dans un premier temps, la reconnaissance de l'état de la composante est effectuée, par exemple l'état de la bouche *"Ouverte"* ou *"Fermée"*. Dans un second temps, une description plus fine est obtenue, par exemple pour un état de bouche *"Ouverte"*, la recherche se porte sur *"Sourire"* ou *"Pas sourire"* (voir la figure 2.12). Les taux de reconnaissance des états de la première couche et de la deuxième couche de description sont tous supérieurs à 81%. Les auteurs proposent cette méthode comme la solution pour une bonne reconnaissance des émotions spontanées.

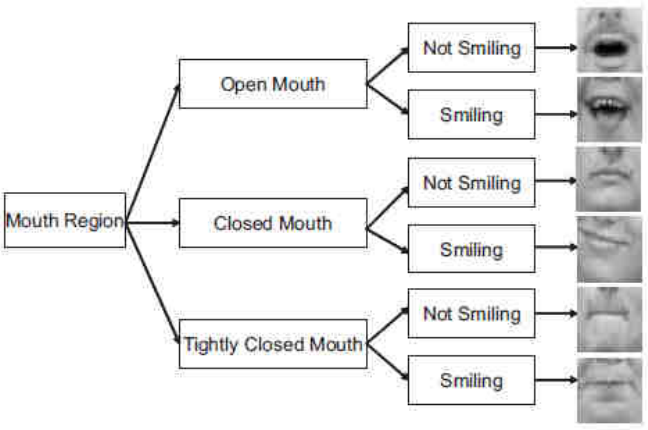


FIGURE 2.12 – Exemple de la description à deux couches de Grimm et al [80]

Les filtres de Gabor sont également utilisés pour l'extraction des textures sur tout le visage. Buciu et al [68] proposent une approche de reconnaissance des expressions émotionnelles dans le cas d'une occultation. La méthode applique 12 filtres de Gabor, constitués de quatre orientations et trois fréquences, pour l'extraction

des caractéristiques de tout le visage. Les résultats obtenus sur la base Cohn-Kanade sans occultations dépassent 94%, ils atteignent un taux de 91.5% pour une occultation des yeux et dépassent 87% pour une occultation de la bouche. Ainsi l'absence d'information dans certaines zones du visage influence le résultat de reconnaissance, d'où l'importance de l'extraction de texture sur tout le visage. Dans le cadre de la reconnaissance des actions unitaires pour les émotions spontanées, Bartlett et al [40] utilisent 72 filtres de Gabor pour extraire les informations de la totalité du visage, formant ainsi un vecteur de taille 663552. Les séparateurs à vaste marge et Adaboost sont testés pour la reconnaissance des actions unitaires ainsi que leurs intensités. Adaboost atteint un taux de reconnaissance des actions unitaires spontanées de 93% pour une reconnaissance image par image dans la base RU-FACS. Ce taux dépasse légèrement le taux obtenu par les séparateurs à vaste marge.
La grande taille des vecteurs caractéristiques générés par l'application des filtres de Gabor sur la totalité du visage est l'un des inconvénients majeurs de cette méthode. Optimiser le choix des caractéristiques extraites a suscité l'intérêt de plusieurs chercheurs. Deng et al [67] étudient la sélection d'un sous ensemble de filtres à partir de deux bancs de filtres : 3 fréquences $\times$ 8 orientations et 4 fréquences $\times$ 8 orientations. Une étape de réduction des descripteurs extraits par les sous-ensembles de filtres est effectuée par une analyse des composantes principales, suivie d'une analyse linéaire discriminante. Deng et al ont montré qu'un sous ensemble des filtres de Gabor suffit pour la reconnaissance des émotions.

Les motifs binaires locaux sont également appliqués pour l'extraction des textures du visage dans les approches de reconnaissance des expressions émotionnelles. Ils sont constitués d'un ensemble d'histogrammes décrivant la distribution des motifs locaux à savoir les contours, les points et les régions sans texture. Les motifs binaires locaux ont montré une robustesse pour la reconnaissance des expressions dans des images à faibles résolutions. Shan et al [9] utilisent les motifs binaires locaux pour extraire l'apparence du visage. Adaboost a été utilisé pour sélectionner les descripteurs les plus discriminants. Cette approche a montré sa robustesse face aux changements de résolutions dans la base PETS. Liao et al [81] ont quant à eux comparé les motifs binaires et les filtres de Gabor sur la base JAFEE. Ils montrent que les motifs binaires locaux se sont avérés plus performants que les filtres de Gabor dans le cas de visages à faibles résolutions.

Les Gabor logarithmiques sont également des méthodes très performantes dans l'extraction de l'information de texture. Elles ont des composantes DC nulles ce qui améliore la représentation des contrastes des crêtes et la représentation des contours [70]. Lajevardi et al [70] ont comparé l'utilisation de Gabor logarithmique et l'utilisation des motifs binaires locaux. Les résultats montrent que les deux méthodes d'extraction donnent des taux de reconnaissance comparables, même dans le cas d'image de faibles résolutions. Une deuxième étude utilisant l'extraction de Gabor logarithmique, menée par Lajevardi et al [69], montre que la fusion de descripteurs extraits de zones locales (bouche et yeux) et des descripteurs extraits de la totalité du visage améliore la reconnaissance des émotions de 5% par rapport à la reconnaissance par les descripteurs du visage seulement.

L'analyse en composantes principales est une technique pour décorréler les variables afin d'avoir une nouvelle base de composantes principales orthogonales les unes aux autres. Ces composantes sont les combinaisons linéaires des variables de départ. Elles représentent les principales déformations de la base [82]. Cette technique a été utilisée pour l'obtention de paramètres de forme des yeux dans [72]. Kapoor et al [72] détectent les déformations de la forme des yeux et des sourcils dans le haut du visage en projetant l'image dans la base de composantes principales. Le taux de reconnaissance obtenu pour les actions unitaires individuelles du haut du visage dans une base comportant différentes poses de tête, des mouvements du visage et des occultations, est de seulement 69.3%. De plus, Kapoor et al ont montré que la détection des déformations par composantes principales est très liée aux sujets de la base.

Les ondelettes de Haar sont aussi considérées parmi les meilleurs moyens pour détecter les textures du visage. Utilisées dans la détection d'objet notamment la détection des visages dans la méthode de Viola et Jones [83], elles sont appliquées pour la reconnaissance des expressions dans [84]. Whitehill et al [84] extraient les descripteurs de Haar de trois régions autour de la bouche, autour de chaque œil et sur le front. Adaboost est ensuite appliqué pour la sélection des descripteurs les plus performants. La comparaison entre la reconnaissance des actions unitaires avec les descripteurs de Haar sélectionnés par Adaboost et les filtres de Gabor classés par les séparateurs à vaste marge montre que la méthode basée sur les descripteurs de Haar est légèrement plus performante que les filtres de Gabor appliqués localement. L'étude menée montre également que l'extraction par les

descripteurs de Haar est moins couteuse en temps que l'extraction par les filtres de Gabor.

Toutes les méthodes précédemment citées ont montré leur efficacité et nous pouvons les substituer les unes aux autres. Elles sont toutes utilisées pour l'extraction des caractéristiques spatiales. Pour certaines d'entre elles, une extension existe pour les adapter à une extraction spatio-temporelle.

Méthode spatio-temporelle

Les méthodes traitant le coté dynamique des émotions sont nombreuses. L'aspect dynamique est intégré soit à l'aide d'une approche statique qui intègre le coté dynamique par un modèle traitant les sorties de chaque image tel que le modèle de Markov caché [85] [86] [60] ou les réseaux bayésiens dynamiques [87] [88] [45], soit à l'aide d'une méthode qui extrait l'aspect temporel en même temps que les aspects spatiaux, ainsi les caractéristiques extraites sont en 3 dimensions.

C'est ce dernier cas auquel nous nous intéressons dans cette sous-section, puisque le premier groupe fait intervenir les méthodes de classification que nous ne traitons pas dans ce chapitre.
Wu et al [89] utilisent une méthode basée sur l'extraction des caractéristiques spatio-temporelles par des filtres de Gabor à 3 dimensions. Ces derniers sont formés par la combinaison de la sortie des filtres de Gabor spatiaux traitant la vidéo image par image et le filtre temporel 1D. Le banc des filtres spatio-temporels est composé de 15 fréquences spatio-temporelles et de 16 orientations. L'extraction des caractéristiques spatio-temporelles de la vidéo améliore la reconnaissance des expressions surtout lorsqu'il s'agit d'une expression à faible intensité. En effet, une augmentation de 7% du taux de reconnaissance des expressions faciales dans la phase *onset* (début de l'expression) est obtenue avec les filtres de Gabor spatio-temporels par rapport à la reconnaissance par les filtres de Gabor spatiaux.
Une extension de la méthode des motifs binaires locaux à une troisième dimension est également possible pour permettre le codage de la dynamique des émotions ainsi que le proposent Zhao et al [90]. Les motifs binaires sont extraits à partir de trois plans orthogonaux comme le montre la figure 2.13 et des motifs de différentes résolutions. Adaboost est ensuite appliqué pour sélectionner les caractéristiques les plus discriminantes. Les Séparateurs à Vaste Marge sont enfin appliqués donnant

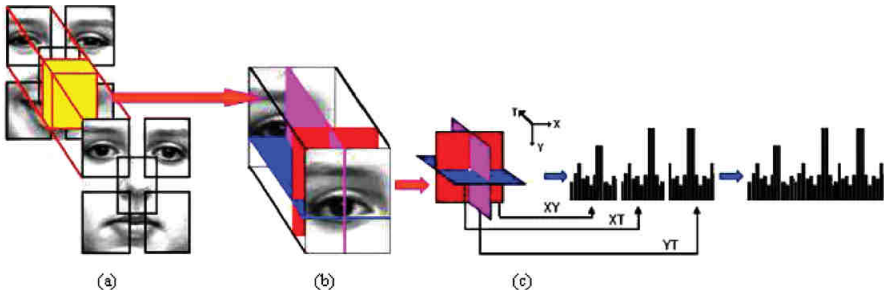


FIGURE 2.13 – Motifs binaires locaux spatio-temporels [90]. (a) bloque de volume (b) descripteurs des motifs binaires à partir de trois plans orthogonaux (c) concaténation des descripteurs en un bloque contenant l'apparence et le mouvement

ainsi des résultats intéressants, qui dépassent 93% dans la base Cohn-Kanade.

Yang et al [43] proposent une méthode de reconnaissance des actions unitaires et des expressions faciales basée sur les descripteurs de Haar dynamiques, combinaison des descripteurs de Haar extraits image par image de toute la séquence vidéo, ayant la même position et le même motif. Les descripteurs sont ensuite codés de façon binaire dans des dictionnaires pour chaque émotion et pour chaque action unitaire. Un apprentissage par Adaboost est enfin appliqué pour chaque émotion en utilisant les dictionnaires comme de faibles classifieurs. Comparée sur la base Cohn-Kanade avec les descripteurs de Haar statiques, cette méthode montre une amélioration des taux de reconnaissance des six émotions et de la reconnaissance des actions unitaires.

Les méthodes d'apparence spatiales ou bien spatio-temporelles décrivent les changements de la peau, liés à l'expression faciale. Cependant, la peau est très différente d'une personne à une autre et son aspect peut grandement varier, ne serait-ce qu'en fonction de l'âge du sujet considéré. L'apparence peut ainsi s'avérer insuffisante à la description de l'expression. La section suivante est consacrée aux propositions qui ont été faites pour combiner descripteurs d'apparence et descripteurs géométriques.

2.6.3 Descripteurs hybrides

Les expressions faciales sont définies à la fois par un changement de la forme des caractéristiques faciales et par une variation de la texture, autour de ces caractéristiques et dans d'autres zones du visage. La forme et la texture du visage sont

ainsi fortement liées et la présence des deux informations est indispensable pour une bonne reconnaissance des expressions faciales. Les méthodes représentant l'apparence seulement sont souvent critiquées par un manque de représentation des mouvements des caractéristiques du visage [91]. Kotsia et al [92] expliquent que l'utilisation de l'information d'apparence seule peut entrainer une confusion dans la reconnaissance de la colère et la peur. Cependant, l'absence d'informations qui décrivent l'apparence et la texture de l'expression peut amener à négliger de faibles mouvements des muscles faciaux. Les méthodes représentant seulement la forme et les mouvements géométriques négligent quant à elles des informations telles que les rides transitoires qui peuvent être des caractéristiques indispensables à la différentiation entre les émotions.

Une solution pour une description plus efficace des changements faciaux liés à l'expression est la présentation à la fois de l'information d'apparence et de l'information de forme. Fasel et al [30] indiquent que la combinaison de l'aspect géométrique et de l'aspect d'apparence peut être très intéressante notamment lorsqu'ils n'ont pas les mêmes erreurs. La combinaison de la description géométrique de l'expression et de la description d'apparence peut être effectuée soit directement avec des modèles présentant les deux informations comme le modèle actif d'apparence (*Active Appearance Model (AAM)*), soit avec une combinaison de descripteurs extraits par une méthode géométrique et de descripteurs extraits par une méthode d'apparence.

Le modèle actif d'apparence est un modèle performant, qui considère à la fois les contraintes de forme et les contraintes d'apparence dans l'alignement et l'extraction des caractéristiques faciales. Plusieurs représentations sont dérivées de ce modèle à savoir la similarité normalisée de forme, la similarité normalisée d'apparence et la représentation canonique d'apparence [93] [91]. La similarité normalisée de forme représente le visage et ses caractéristiques par 74 points. La similarité normalisée d'apparence représente l'apparence de chaque image détectée par un masque modélisé sur la forme de base. La représentation canonique est une représentation où toutes les variations de forme dues à l'expression sont éliminées de l'apparence. La contribution de chacune de ces trois représentations, la contribution de la fusion entre la représentation de forme et la représentation canonique d'apparence sont étudiées dans [93] dans le cadre de la reconnaissance des expressions faciales de la peine. Ashraf et al [93] ont montré que la fusion de la

FIGURE 2.14 – Schéma de fusion en amont

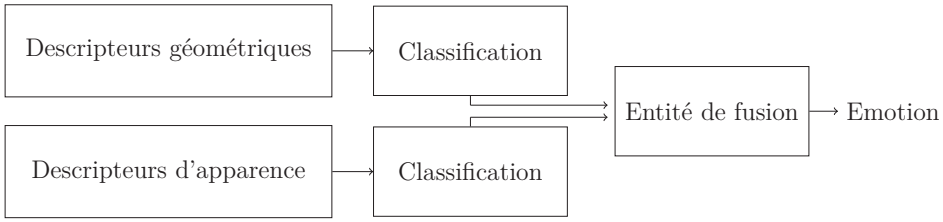


FIGURE 2.15 – Schéma de fusion en aval

forme de l'expression par le biais de la représentation de similarité de forme et de l'apparence par le biais de la représentation canonique d'apparence améliore la reconnaissance de la peine, permettant de dépasser un taux de reconnaissance de 81%. L'avantage de cette combinaison dans un même vecteur est confirmée par Lucey et al [91] dans le cadre de la reconnaissance des actions unitaires. La performance de la fusion de ces deux représentations est également utilisée pour la reconnaissance des six émotions de base dans [4].

Une deuxième alternative est appliquée pour décrire à la fois la forme et l'apparence de l'expression. Elle consiste à extraire chaque information de façon individuelle par une méthode dédiée pour ce but. Une fusion est ensuite appliquée. En général deux schémas de fusion sont utilisés, à savoir un schéma en amont et un schéma en aval.

- Le schéma en amont combine les descripteurs de différents types d'information avant le passage à l'étape de classification. Dans ce cas un prétraitement est appliqué aux deux informations pour pouvoir les fusionner dans un même vecteur. Ce dernier est ensuite utilisé comme donnée d'entrée par la méthode de classification (voir la figure 2.14).
- Le schéma en aval combine les descripteurs après l'étape de classification. Pour chaque type de descripteurs extraits une classification est appliquée. Les décisions issues de l'étape de classification sont ensuite combinées dans une entité de fusion. La figure 2.15 présente le schéma en aval.

Dans le cadre de la fusion entre les descripteurs géométriques et les descripteurs

d'apparence, le schéma en amont a été préféré dans plusieurs travaux de la littérature [94] [95] [74] [96]. Zhang et al [94] codent l'aspect géométrique par 43 distances entre des points détectés par le modèle actif de forme. Les distances décrivent la représentation des paramètres d'animation faciale (FAP) du standard MPEG-4. D'autre part, plusieurs méthodes ont été utilisées pour l'extraction des descripteurs d'apparence à savoir les filtres de Gabor, la méthode de transformation de caractéristiques visuelles invariantes à l'échelle (SIFT) et la méthode des motifs binaires locaux (LBP). Ces méthodes ont été appliquées localement dans des fenêtres autour de points caractéristiques. Les descripteurs d'apparence extraits par chaque méthode sont d'abord sélectionnés puis des fusions en amont avec les descripteurs géométriques sont appliquées. Testée sur la base NVIE et la base FEEDTUM, la fusion entre les descripteurs géométriques et les descripteurs d'apparence extraits par la méthode SIFT et sélectionnés par la méthode de redondance minimale et pertinence maximale (mRMR) permet d'obtenir de meilleurs taux de reconnaissance des émotions que les méthodes de fusion combinant les descripteurs géométriques et les descripteurs extraits par les filtres de Gabor et les descripteurs extraits par LBP. Dans [74] les mêmes méthodes sont utilisées pour l'extraction des caractéristiques. Après la fusion, une régression est appliquée à la place de la classification pour une représentation dimensionnelle des émotions. La fusion de la méthode LBP avec les descripteurs géométriques testée sur la base NVIE permet une meilleure représentation des émotions que la fusion des descripteurs géométriques avec les descripteurs de Gabor et la fusion des descripteurs géométriques avec les descripteurs extraits par la méthode SIFT.
Chen et al [95] codent l'aspect géométrique de l'expression par le calcul du déplacement de 21 points. L'aspect d'apparence est extrait par la différence de gradient normalisé entre l'image neutre et l'expression. Ce calcul est effectué localement autour des 21 points. Une fusion en amont est ensuite appliquée, permettant d'avoir un taux de reconnaissance de 95% pour des expressions d'émotions posées.
Zhengyou et al [96] combinent les coordonnées de 34 points caractéristiques détectés manuellement et les coefficients obtenus par l'application de 18 filtres de Gabor. La fusion est appliquée au niveau de perceptrons à deux couches. La première couche réduit de façon non linéaire les descripteurs de chaque type d'information. La deuxième calcule les taux de reconnaissance pour chaque expression émotionnelle. Testée sur la base JAFFE, le taux de reconnaissance atteint 90%.

La fusion en aval combine quant à elle les décisions [97] ou des paramètres qui

résultent de la classification [92]. Song et al [97] différencient deux types d'expressions à savoir les expressions symétriques et les expressions non symétriques en se basant sur les paramètres d'animation faciales (FAPs) codés par 27 points. Ils détectent également les déformations de la peau en calculant les rapports d'intensité dans 8 patchs. Un SVM est appliqué pour détecter la présence ou l'absence de déformations de la peau liées à l'expression dans chacun des patchs. Ces deux décisions issues de l'aspect géométrique et de l'aspect d'apparence sont combinées dans un seul vecteur formant ainsi de nouveaux paramètres pour une meilleure description des expressions. La reconnaissance des émotions est ensuite effectuée en fonction du type de l'expression symétrique ou asymétrique. Dans le cas où l'expression est détectée comme symétrique, sept SVMs sont utilisés pour classer l'expression parmi l'une des six émotions de base ou l'expression neutre. Dans le cas où l'expression est détectée en tant qu'expression asymétrique, six SVMs classifient l'expression parmi l'une des six émotions de base, puisque l'expression neutre est toujours symétrique. Les taux de reconnaissance de cette méthode sur la base Cohn-Kanade, la base JAFFE et une base préparée par les auteurs dépassent 88% même en présence de variation de luminance. Kotsia et al [92] combinent les distances issues de la classification des informations géométriques et des informations d'apparence. La grille de Candide est appliquée afin de suivre les déformations des muscles pendant l'expression. Les distances entre les nœuds de la grille sont classées par SVM parmi l'une des sept classes (les six émotions de base et l'expression neutre). D'un autre côté, la déformation de la texture est extraite par la factorisation matricielle discriminante non négative (Discriminant Non-negative Matrix Factorization (DNMF)). La distance entre le vecteur texture de la séquence vidéo et la classe la plus proche est combinée avec la distance calculée par le SVM pour la plus proche classe du vecteur de déformation de Candide. La fusion est ensuite classée par un réseau de neurones à fonction radiale. Un taux de reconnaissance de 92.3% est obtenu pour les six expressions émotionnelles de base et l'expression neutre dans la base Cohn-Kanade.

2.7 Sélection des descripteurs

Les descripteurs extraits du visage sont souvent volumineux, nécessitent un grand temps de calcul et beaucoup de mémoire. Pour résoudre ces problèmes, une étape de sélection devient alors indispensable dans le système de reconnaissance des

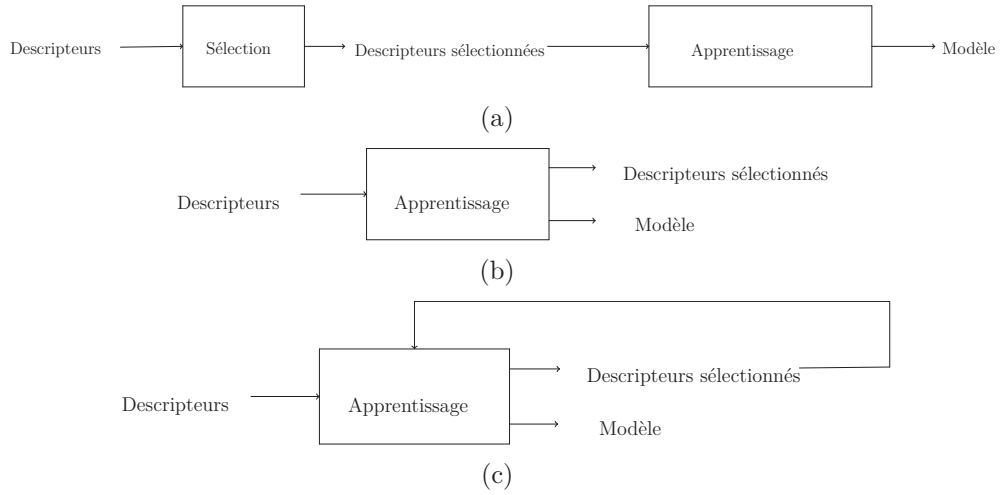


FIGURE 2.16 – Schémas des trois classes de sélection. (a) approches de type "*filter*", (b) approches de type "*embedded*", (c) approches de type "*wrapper*"

émotions. Les approches de sélection cherchent à extraire de manière optimale un sous-ensemble de descripteurs, qui obéissent à certains critères dont la définition change selon l'approche utilisée. La sélection des descripteurs réduit les bruits dus aux conditions d'acquisition et aux mouvements de la tête dans les séquences d'émotions naturelles et spontanées [98]. L'étape de sélection réduit également la redondance des descripteurs et améliore dans certains cas la classification.

Les méthodes de sélection sont souvent classées en trois groupes :

- Approches dites "*filter*" : ces techniques sont appliquées avant l'étape de classification. Leur critère de sélection est complètement indépendant des résultats de classification (voir figure 2.16 (a)).
- Approches dites "*embedded*" : ces techniques sélectionnent les descripteurs durant l'étape d'apprentissage (voir figure 2.16 (b)). Elles combinent ainsi l'estimation du modèle et le choix des descripteurs les plus performants [99].
- Approches dites "*wrapper*" : ces approches sélectionnent les descripteurs de façon itérative en évaluant leur pertinence par la méthode de classification [99] (voir figure 2.16 (c)).

Les méthodes de reconnaissance des émotions basées sur les approches "*filter*" sont nombreuses. Elles présentent une étape de sélection indépendante, qui prend en entrée tous les descripteurs extraits du visage et donne en sortie seulement les descripteurs choisis. Les critères de sélection des descripteurs dépendent de la méthode choisie.

L'étude de la corrélation des descripteurs est un critère important pour le choix

de ceux-ci. La technique de sélection des caractéristiques par corrélation (*Correlation Feature Selection (CFS)*), adoptée dans [74] choisit les descripteurs les plus corrélés avec la vérité terrain et les moins corrélés avec les autres classes. L'analyse par composantes principales [66] [100] exploite également la notion de corrélation. Elle transforme les descripteurs de départ en composantes principales non corrélées. Ainsi, le choix d'un sous-ensemble de composantes, ayant assez de variance pour représenter les données, réduit le nombre de descripteurs de départ. Pour substituer la variance dans le choix des composantes principales retenues, Soyel et al [101] utilisent le critère de Fisher. Une sélection itérative des composantes optimales est alors appliquée. L'analyse des composantes principales 2D [102] est également utilisée pour la réduction des descripteurs. Elle réduit les images filtrées par les filtres de Gabor en opérant directement sur les images sans les convertir en vecteurs.
Enfin l'information mutuelle est un critère utilisé dans la sélection des descripteurs. Contrairement à la sélection basée sur la corrélation, la sélection basée sur l'information mutuelle est indépendante des relations de linéarité qui peuvent exister entre les descripteurs [103]. L'information mutuelle mesure les dépendances entre les descripteurs et les classes. Elle est utilisée dans [69] pour la sélection des descripteurs extraits par les filtres de Gabor logarithmique sur les régions hybrides locales et globales. Lajevardi et al [103] montrent que la sélection par information mutuelle donne de meilleurs résultats que la sélection par l'analyse des composantes principales et la sélection par filtres optimaux.

Les méthodes de type "*embedded* " sélectionnent en général les descripteurs les plus importants au fur et à mesure de l'apprentissage. Ces méthodes ne sont pas très utilisées dans les systèmes de reconnaissance des émotions. Cependant, celle qui revient le plus souvent est la technique Adaboost. Adaboost est une méthode performante permettant de constituer à partir d'un ensemble de faibles classifieurs un classifieur fort. Souvent dans le cas de sélection les arbres de décision utilisée comme faibles classifieurs dans Adaboost sont constitués d'un seul niveau (*Decision stump*). Ces faibles classifieurs présentent les descripteurs choisis par Adaboost. Shan et al [9] et Bartlett et al [52] appliquent Adaboost pour réduire respectivement les descripteurs issus de motifs binaires locaux et les réponses de filtres de Gabor.

Les méthodes de type "*wrapper*" sélectionnent de façon itérative les descripteurs

après leur évaluation par la méthode de classification. Les algorithmes génétiques sont souvent utilisés dans le but de sélectionner les caractéristiques. Dans la reconnaissance des expressions émotionnelles, une extension des algorithmes génétiques, appelée *programmation génétique* est utilisée. Yu et Bhanu [104] présentent une approche de reconnaissance des expressions utilisant celle-ci. Une extraction des caractéristiques est d'abord effectuée par les ondelettes de Gabor. Ces caractéristiques considérées comme primitives alimentent la programmation génétique pour constituer des vecteurs composés. Ces derniers sont évalués par les séparateurs à vaste marge. L'ensemble des vecteurs composés retenus sont fixés après un ajustement itératif de la fonction de fitness qui se base sur les résultats des séparateurs à vaste marge. Testé sur la base JAFEE, la méthode atteint un taux de reconnaissance de 80.9%.
Li et al [105] proposent une technique de sélection de type "*wrapper*" pour choisir l'ensemble de descripteurs saillants parmi les descripteurs d'apparence du visage. L'ensemble de descripteurs choisi est d'abord initialisé par le descripteur ayant le taux maximal de séparation entre les classes. Ce taux est calculé pour tous les descripteurs. Les descripteurs intégrés de façon itérative dans l'ensemble choisi sont les descripteurs qui maximisent le taux de reconnaissance et minimisent le score d'information mutuelle.

2.8 Bases de données d'expressions faciales

L'évaluation des méthodes de reconnaissance d'expressions faciales et de reconnaissance émotionnelle nécessite le recours à une ou plusieurs bases de données. De façon à obtenir l'évaluation la plus objective possible, les bases de données doivent être générales et indépendantes des restrictions ou des hypothèses liées à un domaine précis. Ainsi, les performances des méthodes peuvent être référencées et comparées plus facilement.
Deux types de bases existent à savoir les bases d'émotions simulées et les bases d'émotions spontanées. Dans ce chapitre nous présentons les bases de données les plus utilisées dans la littérature. Puis, nous étudions les contraintes fréquemment rencontrées dans les applications de reconnaissance d'émotions et évaluons leur prise en compte dans les bases.

2.8.1 Cohn-Kanade

La base de données d'expressions faciales Cohn-Kanade est destinée aux travaux d'analyse et de synthèse automatiques d'images, ainsi qu'aux études de perception comme la reconnaissance des expressions faciales [106]. La première version de cette base de données est constituée d'expressions simulées par des acteurs. Des experts ayant décrit et montré visuellement l'action unitaire ou la combinaison d'actions unitaires à réaliser, les sujets les reproduisent devant la caméra. La partie de la base mise en accès libre contient 97 sujets, âgés de 18 à 30 ans. Ces sujets sont constitués de 65% de femmes, 15% d'afro-américains et 3% d'asiatiques et de latinos. Les séquences d'images présentées commencent par l'expression neutre et se terminent par le maximum de l'expression. Par ailleurs, la dernière image de la séquence est toujours codée par des experts. Parmi ces images, 17% sont codées par deux experts. L'accord de jugement des deux est calculé par le coefficient Kappa [2]. Cette base comporte des expressions impliquant une ou plusieurs actions musculaires, parmi elles les six émotions universelles (joie, colère, peur, surprise, tristesse, dégout). Ces émotions sont labellisées par des prototypes du système FACS [107].

La deuxième version de cette base CK+ décrite dans [4] comporte deux types d'expressions, posées et spontanées. Les expressions posées présentent une continuation de la première version, le nombre de vidéos augmentant de 22% et le nombre de sujets de 27%. Les séquences d'images sont organisées de la même manière que les séquences de la première version ; elles commencent par l'image neutre et se terminent par une dernière image présentant le pic de l'expression demandée. La figure 2.17 présente un sujet appartenant aux deux versions de la base Cohn-Kanade simulant les six émotions.
La labellisation des émotions posées est également effectuée par le codage FACS. Celle-ci se fait en trois étapes :

1. Les expressions des émotions sont d'abord comparées avec les prototypes des émotions prédéfinies par le FACS et leurs variations majeures. Par exemple l'action unitaire (AU9) présente le dégoût et la combinaison des actions unitaires (AU9 + AU17) est une deuxième variante du dégoût. Si

2. Le coefficient Kappa permet de mesurer l'accord entre des jugements au delà de l'effet du hasard

l'expression de l'émotion est compatible avec le prototype, elle obtient le label de cette émotion.

2. Dans le cas où la séquence comporte une AU non incluse dans le prototype ou l'une de ses variantes, l'adéquation de ces actions unitaires avec l'émotion en question est testée. Par exemple AU4 n'est présente que dans les émotions négatives, ainsi sa présence dans une séquence de surprise n'est pas validée.
3. Une vérification de toute la séquence est effectuée afin d'avoir un jugement perceptuel de la cohérence de l'expression avec l'émotion demandée. Cette étape est indispensable puisque les codages FACS ne présentent que le pic de l'émotion.

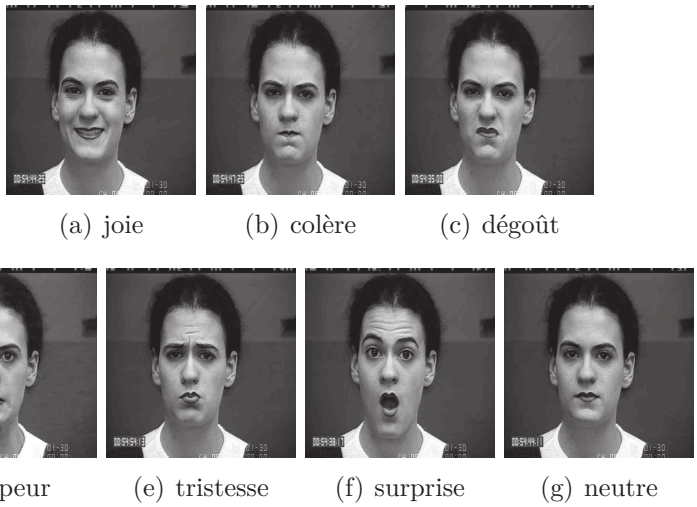
Dans la version CK+, les sourires spontanés acquis à l'insu des sujets juste avant l'enregistrement des expressions posées sont également proposés. Différentes intensités de sourires existent et leur insertion dans la base obéit à des conditions strictes comme :

- avoir une expression neutre au début de la séquence.
- ne pas présenter d'occultations ou d'artéfacts causés par le mouvement de la caméra dans la séquence.

Les sourires spontanés sont labellisés par la prise en compte des jugements de plusieurs participants (des groupes de 10 à 17 personnes) qui visualisent les séquences une par une et attribuent un type aux sourires, à savoir amusé, embarrassé, nerveux, poli et autres. L'intensité des sourires est également quantifiée sur une échelle de 1 à 7 ainsi que le niveau de confiance du sujet dans son choix du type de sourire. Le sourire est affecté d'un label si le niveau de confiance de celui-ci indique au moins 50% et moins de 25% pour les autres labels. Plus d'informations sur les sourires spontanés sont disponibles dans [108].

2.8.2 MMI

La base MMI est une base d'expressions faciales, continuellement en expansion. Elle comporte cinq parties : la première est constituée de 1767 séquences d'images de 20 participants de différentes origines (américains, asiatiques, européens). Les sujets effectuent 31 actions unitaires et des états affectifs tels que la joie et l'ennui. Ces actions sont répétées deux fois afin d'augmenter la variabilité de la base. Cette



(a) joie (b) colère (c) dégoût

(d) peur (e) tristesse (f) surprise (g) neutre

FIGURE 2.17 – Exemple extrait de la base CK+

première partie comporte des vues frontales et des vues de profil obtenues à l'aide d'un miroir positionné à un angle de 45 degrés de la vue frontale du sujet et en respectant toujours la position de la caméra [109].
La deuxième partie de la base est constituée de 238 séquences de 28 sujets qui simulent les six émotions de base. Tous les sujets sont enregistrés deux fois notamment ceux portant des lunettes, enregistrés une fois avec leurs lunettes et une deuxième fois sans les lunettes. Dans cette partie de la base, le profil est enregistré par une caméra placée sur le côté des sujets afin d'avoir une meilleure résolution que dans le cas d'images reflétées par un miroir.
Dans la troisième partie de la base, 5 sujets expriment toutes les actions unitaires et les six émotions de base permettant d'obtenir un total de 484 images.
Dans ces trois parties, un ensemble de vidéos est enregistré dans des conditions naturelles d'illumination et avec un arrière-plan complexe comportant plusieurs visages de profil comme le montre la figure 2.18. D'autre part, un deuxième ensemble est enregistré dans des conditions de luminosité contrôlée et avec un arrière plan simple (voir la figure 2.19).
La labellisation de ces parties est effectuée par deux spécialistes du codage FACS travaillant de façon indépendante. Dans le cas où les deux codes proposés sont différents, les codeurs se mettent ensuite d'accord sur le codage final à retenir. Un codage image par image par des actions unitaires temporelles est également appliqué sur un ensemble des vidéos. Plus d'informations sont disponibles dans [110].

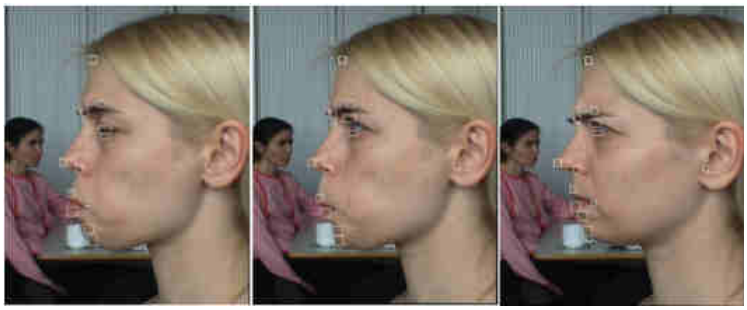


FIGURE 2.18 – Exemple d'image de la base MMI où l'arrière plan est complexe et la lumière naturelle [110]

FIGURE 2.19 – Exemple d'image de la base MMI où l'arrière plan est simple avec des lumières de studio (vue frontale et vue de profil) [110]

D'autres parties ont été ajoutées à la base de données MMI. Les dernières versions sont les parties IV et V décrites dans [109]. Les émotions exprimées dans ces deux parties sont des émotions spontanées à savoir joie, dégoût et surprise. Celles-ci sont en effet plus faciles à obtenir de façon naturelle que les émotions colère, tristesse et peur. Les participants sont invités à visualiser des vidéos et des images permettant de stimuler leurs émotions. Pour stimuler la joie, des vidéos de dessins animés sont par exemple visualisées, tandis que des vidéos et des images de maladies et de chirurgie sont projetées pour stimuler le dégoût.
Dans la partie IV de la base de données, les expérimentateurs sont présents dans la même pièce que les participants ce qui peut influencer ou inhiber les émotions de ces derniers. En revanche, dans la partie V de la base, les participants visualisent les vidéos sans la présence de l'équipe d'expérimentateurs. Ainsi, les émotions sont exprimées plus naturellement.
Les sujets participant à ces deux expériences sont au nombre de 25 (16 pour la partie IV et 9 pour la partie V). Ils sont âgés de 20 à 32 ans, représentent les deux sexes (12 femmes et 13 hommes) et sont de différentes origines (européenne, asiatique et sud-américaine).

2.8.3 FEEDTUM

La base de données FEEDTUM est réalisée dans le cadre du projet européen FG-NET (réseau de reconnaissance des visages et des gestes). Elle présente les six émotions de base (joie, colère, dégoût, peur, tristesse et surprise) et l'expression neutre. Le système d'enregistrement est constitué de 18 ordinateurs équipés de caméras (voir la figure 2.20). Il permet à la fois une projection des vidéos sur les écrans et un enregistrement des émotions stimulées.
Cette base comporte 18 sujets. Chaque sujet visualise trois vidéos pour chacune des émotions. Les réactions des sujets sont alors enregistrées et labellisées suivant la vidéo stimulante. Par exemple pour une vidéo censée stimuler la joie, les expressions faciales du sujet sont enregistrées et labellisées comme des expressions de joie. Ces expressions sont considérées comme spontanées.
Les séquences d'images présentes dans la base de données sont acquises par une

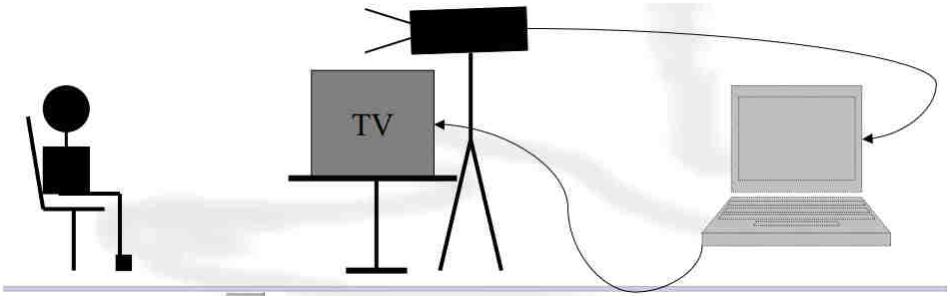
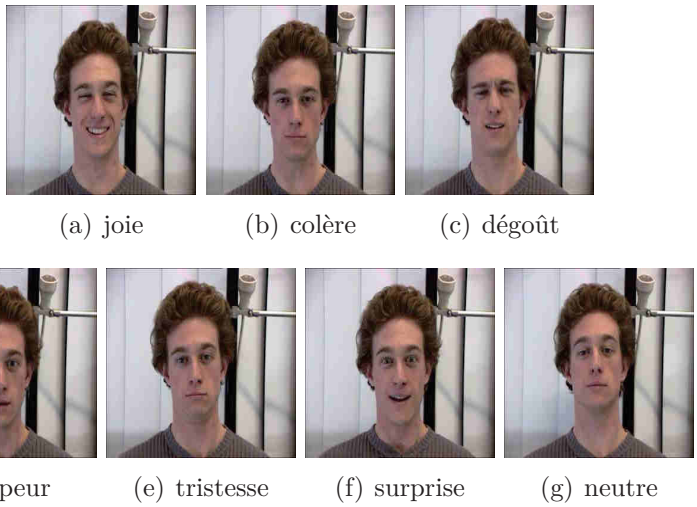


FIGURE 2.20 – Schéma du protocole utilisé pour l'enregistrement de la base FEEDTUM (schéma extrait de [5])

caméra Sony XC-999P. Elles ont une résolution de 640 × 480, de profondeur de 24 bits. Les images sont cependant disponibles dans la base en 8 bits et de résolution 320 × 240 pixels [5]. La figure 2.21 présente un exemple extrait de la base FEEDTUM.

2.8.4 JAFFE

La base de données JAFFE (*Japanese Female Facial Expression Database*) comporte 219 images statiques. Ces images concernent dix femmes japonaises qui simulent les six émotions universelles et l'expression neutre. La labellisation de la base est faite par 60 autres femmes japonaises qui quantifient l'émotion présente

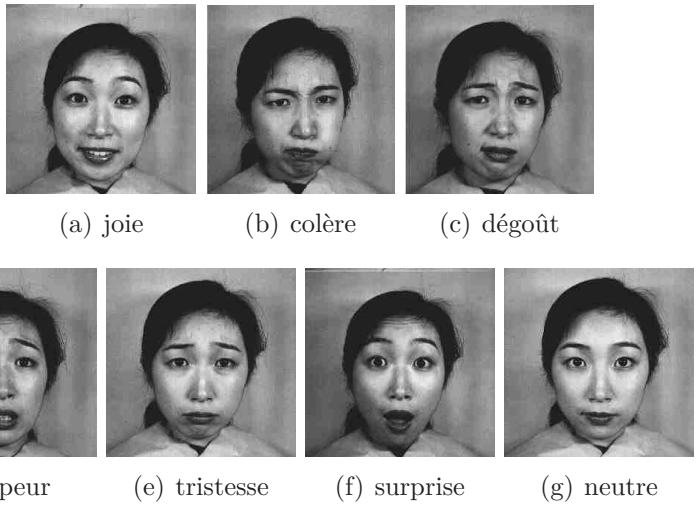


(a) joie (b) colère (c) dégoût

(d) peur (e) tristesse (f) surprise (g) neutre

FIGURE 2.21 – Exemple extrait de la base FEEDTUM

dans chaque image sur une échelle de 5 points pour les six émotions [111]. La figure 2.22 présente un exemple extrait de cette base.



(a) joie (b) colère (c) dégoût

(d) peur (e) tristesse (f) surprise (g) neutre

FIGURE 2.22 – Exemple extrait de la base JAFFE [112]

2.8.5 SEMAINE

Cette base comporte des vidéos avec des conversations entre des participants et des avatars. Ces derniers présentent différents états afin de stimuler ce même état chez les participants. Par exemple l'avatar *Poppy* est toujours heureuse, tandis que l'avatar *Spike* est toujours en colère et les mots et les répliques utilisés dans la conversation stimulent également ces émotions chez les participants. Cette base

comporte 150 participants, âgés de 22 à 60 ans, donnant un total de 959 séquences de conversation [113]. Les avatars sont guidés par des opérateurs suivant trois scénarios de différentes complexités. Le plus simple consiste en une conversation du participant avec un opérateur se trouvant dans une autre pièce jouant le rôle de l'avatar. Le deuxième scénario est semi automatique, l'opérateur choisi les répliques de l'avatar au fur et à mesure de la conversation, suivant un script prédéfini. Le troisième scénario est complètement automatique, l'avatar formule les répliques suivant la situation et l'expression faciale qu'il doit adopter. Le système utilisé pour l'enregistrement est présenté dans la figure 2.23. La labellisation de la base est effectuée par un ensemble de 6 à 8 évaluateurs pour chaque séquence. L'annotation des émotions se base sur une représentation continue sous différentes dimensions à savoir *arousal*, *valence*, *power*, *expectancy* et *intensity*. La figure 2.24 présente un exemple extrait de la base SEMAINE.

FIGURE 2.23 – Le système utilisé lors de l'enregistrement de la base SEMAINE. A gauche le participant interagit avec l'avatar, à droite l'opérateur guide l'avatar [113]

FIGURE 2.24 – Exemple extrait de la base SEMAINE où l'opérateur est présenté dans les deux images de gauche et où les trois images de droite présentent le participant [113]

2.8.6 Autres bases d'expressions faciales utilisées dans des travaux de la littérature

Nous présentons dans le tableau 2.1 un ensemble de bases utilisées pour évaluer les systèmes de reconnaissance des émotions. Elles sont majoritairement simulées.

Bases de données	Types des expressions émotionnelles	Détails de la base
Hammal-Caplier [114]	simulées	• 21 sujets. • Joie, dégoût, surprise. • Les séquences commencent par l'expression neutre et se terminent par un retour à l'état neutre.
Dailey-Cottrell [115]	simulées	• 16 sujets. • Joie, colère, peur, dégoût, tristesse et surprise. • Pour chaque émotion deux images existent l'expression et l'image neutre.
PIE [116]	simulées	• 68 sujets. • Expression neutre, sourire. • Différentes poses de têtes et différentes conditions d'illumination.
SALD [117]	spontanées	• 20 sujets. • Labellisation suivant les dimensions *Valence*, *Arousal*. • Les émotions sont simulées lors de conversation avec des agents artificiels.
Yale [118]	simulées	• 15 sujets. • Plusieurs poses de tête. • Expression neutre, tristesse, surprise, somnolence. • Changements d'illumination.
DaFEx [119]	simulées	• 8 sujets.

Bases de données	Types des expressions émotionnelles	Détails de la base
		• Joie, colère, dégoût, peur, tristesse, surprise et l'expression neutre.
HUT [120]	simulées	• 2 sujets. • Joie, colère, dégoût, peur, tristesse et surprise.
PETS 2003 [121]	simulées	• Expression neutre, joie et colère. • Les séquences d'images sont présentées à basse résolution, approximativement de 69×93.
NVIE [122]	simulées et spontanées	• 100 sujets enregistrés simultanément par des caméra d'image visible et des caméras infra-rouge thermique. • Les expressions posées sont enregistrées avec et sans lunettes.
DISFA [123]	spontanées	• 27 sujets de différentes origines. • Codage des séquences image par image par le système FACS en utilisant les 5 niveaux d'intensité.

TABLE 2.1 – Bases d'expressions faciales utilisées dans la littérature

2.9 Critères des bases de données

Une base de données générique pour la comparaison des approches de reconnaissance d'expressions faciales et émotionnelles doit vérifier certains critères afin de se rapprocher des contraintes réelles des applications. Elle doit également fournir des vérités terrain fiables pour permettre l'évaluation de la robustesse des applications de reconnaissance d'émotions. Nous détaillons ces critères dans les points suivants :

2.9.1 Choix des sujets

Le choix des sujets dans une base de données peut avoir une influence importante sur les résultats de la reconnaissance. En effet, les expressions sont différentes se-

lon l'âge des participants. Un bébé n'exprime pas une émotion de la même façon qu'une grande personne. Il a un visage très lisse, qui ne contient pas de texture, contrairement au visage d'une personne âgée contenant des rides. Ainsi, le niveau de texture présent dans un visage varie selon l'âge [54]. De plus, les morphologies des visages changent d'une personne à une autre et les origines interviennent également dans la forme et l'apparence des visages. Les asiatiques ont par exemple une forme d'yeux spécifique. Les expériences de P. Ekman [124] prouvent également que la culture intervient dans l'expression des émotions. En effet, Ekman montre à travers une expérience menée sur des américains et des japonais que ces derniers sont plus réservés lors de l'expression de leurs émotions devant des personnes étrangères. Pour ces raisons, les bases de données doivent contenir des personnes d'âges et d'origines variés. Dans toutes les bases que nous avons décrites précédemment il n'existe pas de sujets très jeunes (bébés) et de sujets très âgés (dépassant les 70 ans). Cependant, la majorité des bases choisissent des sujets d'origines variées.

2.9.2 Environnement

Beaucoup de contraintes liées à la scène dans laquelle se trouve le sujet influent sur la reconnaissance des expressions faciales. Nous citons ici l'arrière-plan, la lumière ambiante et les caméras utilisées. Dans la plupart des bases de données, l'arrière-plan est statique et uniforme, comme le montre la figure 2.19 de la base MMI et la figure 2.22 de la base JAFFE. Cependant, les approches faisant leurs apprentissages sur ce type d'images ne sont pas robustes face à des séquences d'images de scènes réelles où les arrière-plans sont non uniformes et dynamiques. Malgré l'intégration d'arrière-plans complexes pour quelques séquences de la base MMI (voir la figure 2.18), la présence d'arrière-plans dynamiques, dus aux mouvements de la caméra ou aux mouvements des sujets, restent toujours un critère important qui n'est que peu inclus dans les bases de reconnaissance des émotions.
La variation de lumière ambiante est également un critère qui impacte les résultats de reconnaissance. Dans la plupart des bases, la lumière est contrôlée à l'intérieur des laboratoires. Cependant, des séquences à l'extérieur avec une lumière naturelle sont aussi indispensables dans le cas de certaines applications. La variation de lumière dans une scène réelle peut projeter une ombre sur le visage, déformant ainsi les caractéristiques de celui-ci [54] ce qui entrainera plusieurs erreurs.

Le choix des caméras a également plusieurs impacts sur l'analyse des expressions. En effet, les propriétés des caméras comme le nombre d'images par seconde et la résolution des images sont très importantes dans la construction d'une base puisqu'elles influencent directement les résultats d'apprentissage des méthodes. Une base de données doit comporter, en plus des images hautes résolutions, des images de basses résolutions, l'extraction des informations réalisable sur des images hautes résolutions pouvant devenir impossible pour des images basses résolutions. Ainsi, les méthodes faisant leurs apprentissages sur des bases contenant les deux types de résolutions sont plus robustes dans les conditions réelles. La majorité des bases existantes comportent seulement des images hautes résolutions. Parmi les rares comportant des images basses résolutions, on peut citer la base PETS 2003 [121] avec des résolutions d'environ 69×93 pixels. L'échantillonnage des vidéos peut également avoir un impact sur les performances. D'après Kanade et al [107], les méthodes utilisant les flux optiques supposent toujours que le mouvement d'une image à une autre est très faible. Ainsi, le choix de l'échantillonnage de la vidéo doit être précis. La base MMI utilise par exemple un échantillonnage de 24 images par seconde.

2.9.3 Pose de la tête

Dans les scènes réelles, les expressions sont souvent stimulées par un évènement ou par une autre personne. Ainsi, dans plusieurs cas les émotions sont exprimées dans une vue inclinée, voir une vue de profil du visage. Tian et al constatent également que les expressions sont souvent accompagnées par un mouvement de la tête. En effet, dans une étude sur le message social et émotionnel du sourire, Kraut et al [125] mentionnent que le sourire se produit dans une interaction avec autrui en tournant la tête vers la personne en question. Dans ces cas de figure, la reconnaissance des émotions doit intégrer des images comportant différentes inclinaisons de la tête et différents angles par rapport à la vue frontale de la tête. Une base de données doit ainsi contenir des images de différentes vues du visage. Quelques efforts ont été accomplis dans ce domaine, comme la base MMI qui a intégré des images de vues frontales du visage et de vues de profil. Les bases de données intégrant ce genre d'informations restent cependant peu nombreuses et la majorité des bases d'images ne présentent que la vue frontale ou proche frontale du visage.

2.9.4 Expression spontanée et posée

Les expressions posées et spontanées sont générées par différentes zones du cerveau. Ainsi l'activation des muscles et la dynamique des expressions diffèrent selon le type d'émotions. Les expressions spontanées sont plus synchronisées, plus symétriques et plus cohérentes que les expressions forcées. D'après [40], les expressions spontanées peuvent correspondre à des expressions innées, tandis que les expressions forcées correspondent à un comportement socialement appris. Une base de données doit contenir les deux types d'expressions permettant ainsi une meilleure reconnaissance des émotions par les approches de reconnaissance automatique. Cependant, plusieurs difficultés apparaissent lorsqu'il s'agit de collecter les émotions spontanées. En effet, la stimulation des émotions telles que la colère, la tristesse et la peur est très difficile dans un environnement de laboratoire. Les bases intégrant les émotions spontanées ont adopté différents protocoles pour les stimuler. Sebe et al [63] présentent ainsi des vidéos pour stimuler les émotions des sujets qui restent seuls dans une salle contenant des caméras cachées afin de donner la liberté aux sujets de s'exprimer et préserver l'authenticité de l'expression. Sebe et al [63] n'ont détecté que la joie, le dégoût et la surprise. Par une technique de visualisation de vidéos semblable à celle utilisée dans [63], les six émotions sont stimulées et enregistrées dans la base FEEDTUM. Certaines émotions sont à peine visibles et restent très proches de l'expression neutre. Une étude subjective de cette base est présentée dans le chapitre suivant. Une autre technique est utilisée dans la base CK+ pour collecter les émotions spontanées. Les sourires spontanés des participants sont enregistrés avant et après leur participation aux expériences de simulations forcées.

2.9.5 Labellisation

La labellisation des émotions est une tâche d'un intérêt majeur dans la construction des bases de données. Deux approches de labellisation sont possibles [30]. La première utilise un ensemble de labels d'émotions prédéfinies, généralement les émotions universelles et une labellisation par jugement, qui se base sur l'avis de plusieurs participants experts ou non experts. Cette technique est utilisée dans la base FEEDTUM. Dans la base présentée par Sebe et al [63], les sujets sont interrogés sur les émotions qu'ils ont éprouvées. La deuxième approche se base

sur des prototypes prédéfinis pour chaque émotion. Ces prototypes se composent d'un ensemble de mouvements musculaires. Plusieurs bases de données ont utilisé cette approche parmi lesquelles nous citons la base MMI, qui a utilisé les actions unitaires du système FACS. La base CK+ est labellisée également par le système FACS. Deux experts effectuent la labellisation de façon indépendantes. Puis, le taux d'accord entre les deux est calculé pour un pourcentage des données.

2.9.6 Occultation

L'occultation présente une contrainte très fréquente dans les applications de reconnaissance des émotions. Cependant, les bases présentant ce genre de problème sont très rares. Certains travaux portant sur la reconnaissance des expressions faciales dans le cas d'occultations rencontrent des difficultés dans l'évaluation de leurs méthodes. Quelques idées ont été mises en place comme la construction de bases contenant de telles situations à partir de bases existantes. Dans Kotsia et al [7], les bases JAFFE et Cohn-Kanade ont ainsi été utilisées pour construire des bases prenant en compte cette contrainte.

Conclusion

Dans ce chapitre nous avons présenté les théories et les représentations les plus connues des émotions. Les approches d'extraction des caractéristiques, classées en trois catégories (géométriques, d'apparence et hybrides), ainsi que les méthodes utilisant la sélection de descripteurs ont été introduites. Finalement, les bases de données fréquemment utilisées ont été décrites, ainsi que les contraintes présentes dans les applications de reconnaissance d'émotions. A l'issue de cette étude, nous avons choisi deux bases pour la suite de notre travail. Au vu de l'importance des émotions posées dans le comportement social et de l'importance des émotions spontanées dans la vie quotidienne, nous avons sélectionné une base pour chaque type d'émotions. Pour les émotions posées, la base CK+, qui comporte des participants d'origines différentes et d'âges différents a été retenue. Concernant les émotions spontanées, notre choix s'est porté sur la base FEEDTUM, puisqu'elle présente les six émotions universelles, contrairement à d'autres bases qui ne présentent que quelques émotions.

Dans le chapitre suivant nous présenterons d'abord une étude subjective de la base des émotions spontanées FEEDTUM. Deux méthodes pour l'extraction de descripteurs d'expressions faciales sont par la suite présentées et une comparaison de ces descripteurs, sous différents critères, est enfin effectuée.

Chapitre 3

Méthodes d'extraction des descripteurs

Introduction

Les émotions spontanées et les émotions simulées sont différentes en terme de schéma d'activation, de dynamique et d'intensité d'expression. Les émotions spontanées sont en particulier plus difficiles à reconnaitre parce qu'elles ont des phases temporelles plus courtes (onset, apex et offset) et de faibles intensités. Si la reconnaissance automatique des émotions spontanées constitue en conséquence un challenge intéressant, on peut légitimement se demander dans quelle mesure un humain est lui même capable de distinguer ces différents cas de figures dans le cadre des bases d'images mises à disposition. Pour quantifier la reconnaissance des émotions spontanées par des humains, nous avons donc effectué une étude subjective sur la base FEEDTUM.

Une bonne reconnaissance des émotions spontanées et des émotions simulées passe d'abord par une fidèle description des changements faciaux déclenchés par l'expression émotionnelle. Ces changements apparaissent au niveau de la texture (la peau, les rides, les fossettes...) et au niveau de la forme (changement géométrique des caractéristiques). Dans ce chapitre, nous étudions l'importance des descripteurs d'apparence (texture) et des descripteurs géométriques dans la reconnaissance des émotions simulées et des émotions spontanées. Nous fixons alors la technique permettant de fournir l'entrée (détection du visage) et la technique traitant les sorties (la méthode de classification). Deux méthodes d'extraction des descripteurs sont

présentées. La première extrait les descripteurs d'apparence par un banc de 40 filtres de Gabor de différentes orientations et de différentes fréquences. La seconde extrait les descripteurs géométriques par une méthode qui code les mouvements faciaux en se basant sur un ensemble de points. Les améliorations proposées dans les techniques de détection de ces points augmentent le taux de reconnaissance des émotions, qu'elles soient simulées ou spontanées. Nous comparons ensuite les deux méthodes suivant différents critères qui interviennent dans plusieurs applications de reconnaissance des émotions à savoir l'influence du changement de résolution des images, la variation de l'ensemble d'apprentissage et le temps d'exécution de la méthode.

3.1 Etude subjective de la base FEEDTUM

La reconnaissance des émotions spontanées est une tâche difficile. Pour mesurer cette difficulté nous avons réalisé une étude subjective de la reconnaissance des émotions d'une base d'expressions spontanées. Nous avons choisi la base FEEDTUM puisqu'elle comporte les six émotions universelles. Neuf personnes non spécialistes du domaine d'étude ont participé à la reconnaissance des émotions de chacune des séquences vidéos de la base. Nous leur avons demandé d'indiquer un seul label pour chaque séquence et d'évaluer la difficulté du test.
La table 3.1 présente les taux de reconnaissance des émotions de la base FEEDTUM calculés à partir des avis des neuf participants. La moyenne des taux de reconnaissance de toutes les émotions est de 61.5%. Ce faible taux de reconnaissance montre la difficulté de la reconnaissance des émotions spontanées. En effet, privés de toute information sur le contexte de la scène permettant la compréhension des causes stimulant l'émotion, les sujets ont plus tendance à confondre les émotions notamment la colère, la peur et la tristesse. Cependant, la joie atteint un taux de reconnaissance de 86%. Ainsi, indépendamment du contexte, la joie est plus compréhensible que les autres émotions. La faible intensité des expressions et la difficulté de stimuler certaines émotions sont également parmi les difficultés majeures pour la reconnaissance des émotions spontanées et peuvent être la cause de confusion entre les émotions. La figure 3.1 présente des exemples d'expressions des émotions de faible intensité portant à confusion. Les expressions de la colère, du dégoût et de la tristesse se ressemblent fortement, pouvant ainsi être considérées comme l'expression d'une seule émotion.

Une évaluation subjective a été menée par Wallhoff et al [86] sur la base FEEDTUM. Les vidéos de la base ont été visualisées par vingt sujets. La moyenne des taux de reconnaissance des émotions est 61%. Elle est très proche du taux de reconnaissance que nous avons obtenu. Wallhoff et al [86] proposent également une méthode de reconnaissance automatique des émotions. Les vecteurs de mouvement, constitués du centre de masse du mouvement, sa direction et sa variance dans les deux dimensions, sont d'abord extraits de cinq zones du visage. Une sélection des descripteurs est ensuite effectuée. Enfin, une classification par SVM des séquences d'images est appliquée. Le taux de reconnaissance des émotions obtenu pour les séquences de la base FEEDTUM est égal à 61.6%. Les résultats de la reconnaissance des émotions spontanées présentés dans la base FEEDTUM obtenus par la reconnaissance automatique et par les évaluations subjectives montrent la difficulté de la tâche entreprise. Nous avons effectué une évaluation du niveau de difficulté de la reconnaissance des émotions spontanées chez les sujets. La table 3.2 montre que 40% des participants trouvent cet exercice difficile et 60% le trouvent d'une difficulté moyenne, mais aucun participant n'a indiqué avoir trouvé la reconnaissance des émotions facile.

(a) Colère (b) Dégoût (c) Tristesse

FIGURE 3.1 – Exemple d'expression pouvant prêter à confusion

%	Joie	Colère	Peur	Dégoût	Tristesse	Surprise
Taux de reconnaissance des émotions	86	42.3	42.6	73.2	57.6	67.6

TABLE 3.1 – Taux de reconnaissance des émotions spontanées de la base FEEDTUM dans le cadre de notre étude subjective

%	Facile	Moyen	Difficile
Avis des participants	0	60	40

TABLE 3.2 – Evaluation de la difficulté du test de reconnaissance des émotions par les participants à l'étude subjective mise en place

3.2 Présentation des étapes de l'approche de reconnaissance des émotions

Dans notre travail, nous nous plaçons dans le cadre de la représentation discrète des émotions, présentée dans la section 2.3.3 du chapitre précédent. En général les approches de reconnaissance des émotions dans ce cadre sont constituées principalement de trois étapes, présentées dans la figure 3.2. Malgré l'importance de chacune de ces étapes dans la reconnaissance des émotions, nous nous intéressons particulièrement à la deuxième, puisqu'elle permet la description des émotions. Nous explorons alors deux types d'approches s'intéressant respectivement aux modifications d'apparence et aux déformations géométriques introduites par l'expression.

Pour extraire les descripteurs d'apparence du visage nous utilisons des filtres de Gabor. Cette méthode prend en entrée le visage comportant l'expression et fournit en sortie un vecteur de descripteurs d'apparence de l'expression. La méthode d'extraction est décrite dans la section 3.5.1.

La méthode d'extraction des descripteurs géométriques prend en entrée le visage neutre pour l'extraction des points caractéristiques et considère le visage courant pour effectuer le suivi de ces points. En sortie, elle retourne un vecteur décrivant la variation des distances qui code les mouvements des muscles faciaux. La description de cette méthode est présentée dans la section 3.5.2.

Afin d'alimenter cette étape centrale dans la reconnaissance des émotions, la détection du visage est d'abord effectuée pour assurer les entrées des méthodes d'extraction des caractéristiques faciales (voir la section 3.3). Les sorties de ces méthodes sont ensuite classées dans la troisième étape. Nous utilisons, dans notre cas, la méthode des Séparateurs à Vaste Marge (voir la section 3.4).

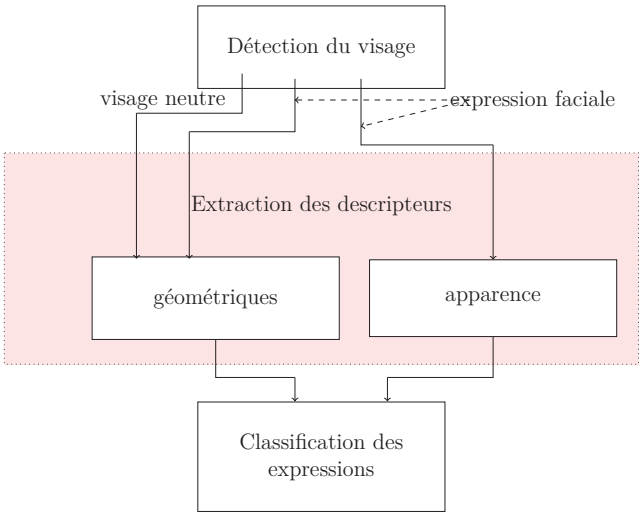
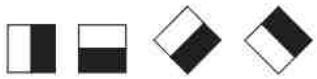


FIGURE 3.2 – Etapes de la méthode de reconnaissance des émotions

3.3 Détection du visage

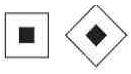
La détection du visage est une étape indispensable dans les systèmes de reconnaissance des émotions. Elle constitue la première étape de notre système permettant de définir une région d'intérêt pour l'extraction des caractéristiques.
Pour réaliser cette première étape, nous appliquons une méthode fréquemment utilisée dans la littérature pour la détection du visage et applicable en temps réel. Implémentée dans la bibliothèque OpenCV [126], cette méthode a été proposée par P. Viola et M. Jones [83]. Elle a ensuite été améliorée par R. Lienhart et al [127]. Elle se base principalement sur l'extraction des caractéristiques par les descripteurs de Haar et la classification par une cascade de classifieurs dopés.
Les descripteurs de Haar sont des fonctions permettant le calcul du contraste entre deux régions contigües représentées par des régions en blanc et des régions en noir. Ils facilitent l'extraction de plusieurs informations comme les contours, les lignes et les centres (voir la figure 3.3). Pour ce faire, la différence entre la somme des pixels sous les zones en blanc et la somme des pixels sous les zones en noir est calculée. Viola et al [83] ont développé une représentation des images appelée *"image intégrale"* pour accélérer le calcul de ces régions.

La méthode de classification est constituée d'une cascade de classifieurs dopés. Le principe de cascade se base sur la présence de plusieurs étages de classifieurs pour une estimation finale. Autrement dit, un résultat positif obtenu par un classifieur déclenche son traitement par le classifieur suivant et ainsi de suite ; par contre un



(a) Descripteurs de contour

(b) Descripteurs de ligne



(c) Descripteurs de centre

FIGURE 3.3 – Des prototypes des descripteurs de Haar

résultat négatif obtenu par un classifieur est directement rejeté (voir la figure 3.4) [83]. La cascade ainsi utilisée améliore la détection et augmente la rapidité de la méthode, puisqu'elle élimine les descripteurs qui n'appartiennent pas au visage dès les premiers étages de classifieurs.
Un classifieur dopé est une combinaison linéaire de plusieurs classifieurs. Viola et al [83] appliquent une variante de Adaboost comme classifieur dopé, de sorte à sélectionner un nombre limité de descripteurs dans chaque étage de la cascade. Un équilibre entre le taux de bonne détection du visage et le taux de faux positif est trouvé afin de permettre la détection de la totalité du visage.

La méthode de Viola et Jones comporte cependant des limites. Un visage dans la position de profil n'est pas détectable. En revanche, cette limite n'influe pas sur notre système puisque nous ne traitons pas les occultations et supposons que les sujets sont face à la caméra.
L'application de la détection du visage sur la base CK+ et sur la base FEEDTUM donne des taux de détection de 100%.

3.4 Méthode de classification

La classification représente la troisième étape de l'approche de reconnaissance des émotions. Elle classe les descripteurs extraits du visage dans les six classes d'émotions à savoir joie, colère, peur, dégoût, tristesse, surprise et l'expression neutre. Nous utilisons dans cette étape la méthode des Séparateurs à Vaste Marge (SVM). Rapide et performante, elle a prouvé son efficacité dans plusieurs travaux [128], [53], [64]. Elle est appliquée principalement dans le cadre de grandes dimensions,

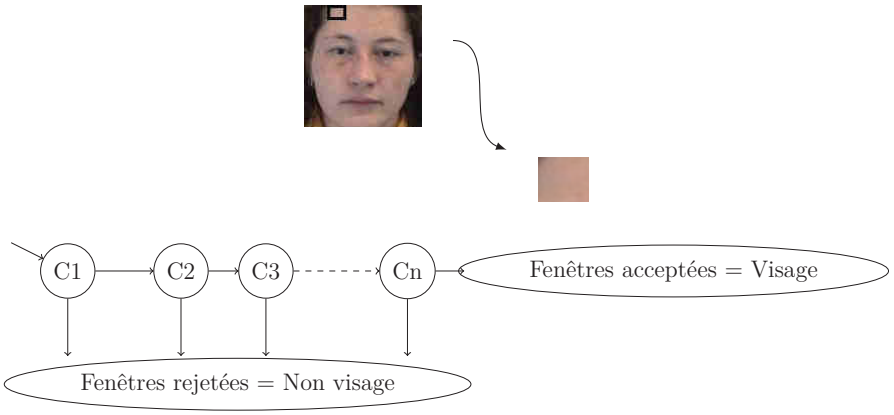


FIGURE 3.4 – Schéma de la cascade de classifieurs

ce qui s'applique parfaitement à notre problématique.

3.4.1 Séparateurs à Vaste Marge (SVM)

Les séparateurs à vaste marge présentent une méthode d'apprentissage et de classification considérant principalement deux classes. Elle est fondée sur deux principes :

- Des marges maximales permettant d'agrandir la distance entre les vecteurs supports et la séparatrice des deux ensembles.
- La projection des variables dans un espace plus grand où la séparation linéaire des variables est possible.

L'espace vectoriel de projection est généré par une famille de fonction ϕ_k transformant les vecteurs d'entrées dans un espace de dimension plus grande voir une dimension infinie. Le produit scalaire des vecteurs dans le nouvel espace est très couteux en terme de calcul. Pour remédier à ce problème un noyau est défini vérifiant :

$$K(x, y) = \sum_{k \in \mathbb{N}} \phi_k(x) \phi_k(y) \tag{3.1}$$

La classification des vecteurs d'entrées se base sur le signe de la fonction de décision $f(x)$ qui dépend désormais du noyau K (voir l'équation 3.2)

$$f(x) = \sum_{i=1}^{M} y_i \alpha_i K(x, x_i) + b \tag{3.2}$$

où α_i et b sont les coefficients déterminant la marge entre la frontière de décision ($f(x) = 0$) et les couples $\{x_i, y_i\}_{i=1...M}$ tels que $x_i \in \mathbb{R}$ et $y_i \in \{-1, 1\}$ présentent

l'ensemble d'apprentissage de taille M. Le cas le plus simple de la fonction noyau est le cas d'une fonction linéaire. Dans ce cas, la classification est appliquée par un hyperplan séparant les deux classes dans le même espace de vecteur d'entrée. Le noyau peut être également polynomial ou gaussien.
Pour la suite de ce chapitre, nous avons choisi un SVM linéaire. En effet, le SVM linéaire nécessite moins de paramètres à fixer pour les deux méthodes d'extraction présentées dans la section suivante. Une comparaison entre les SVMs linéaires et les SVMs gaussiens est présentée dans le chapitre 4. Le codage de cette méthode est basée sur la bibliothèque LibSVM [129].

3.4.2 Validation Croisée

La validation croisée est une technique d'estimation de fiabilité très utilisée en raison de sa simplicité de mise en œuvre et de sa précision. Elle est souvent appliquée pour une meilleure utilisation d'un ensemble de données limité (indépendamment de la dimension des vecteurs de données). Elle consiste à diviser les données en N ensembles égaux contenant le même nombre d'échantillons de chaque classe. L'apprentissage du SVM est appliqué sur $N-1$ ensembles et l'ensemble restant est consacré au test. Ce traitement est répété N fois, à chaque fois un nouvel ensemble de test est choisi et l'ancien est intégré dans l'ensemble d'apprentissage. Le taux de validation croisée est donné par la moyenne des N taux obtenus en considérant chaque test.

3.5 Méthodes d'extraction des caractéristiques

Dans cette section, nous présentons deux méthodes d'extraction des caractéristiques. La première extrait les textures et les apparences du visage. Elle est basée sur les filtres de Gabor. La deuxième méthode extrait les mouvements des muscles du visage.

3.5.1 Méthode d'apparence

Principe

Les filtres de Gabor sont parmi les techniques les plus robustes d'extraction des textures pour différentes échelles et suivant différentes orientations. Ils semblent alors tout à fait adaptés à l'extraction des changements d'apparence engendrés par les expressions faciales. Ils sont appliqués soit au niveau local [130], soit sur la totalité du visage [131], [40]. Nous appelons, dans la suite, la méthode basée sur les filtres de Gabor "*méthode d'apparence*".

Un filtre de Gabor est composé d'une sinusoïde complexe limitée par une gaussienne. L'équation de Gabor s'écrit alors comme suit :

$$G(x, y) = s(x, y)g(x, y) \tag{3.3}$$

où $s(x, y)$ est une sinusoïde complexe et $g(x, y)$ est l'enveloppe gaussienne.
La sinusoïde est composée d'une partie réelle (Re) et d'une partie imaginaire (Im), qui s'expriment de la façon suivante :

$$Re(s(x, y)) = cos(k\,x\ +\ k\,y)$$
$$Im(s(x, y)) = sin(k\,x\ +\ k\,y)$$

La gaussienne g de largeur de bande $\frac{\sigma}{k}$ s'écrit $g(x, y) = \frac{\|k\|^2}{\sigma^2} \exp(-\frac{(x^2+y^2)\|k\|^2}{2\sigma^2})$.
Nous supposons que la gaussienne et la sinusoïde ont la même direction (le même angle ϕ_μ). Cette approximation donne un résultat semblable au cortex visuel [132]. Ainsi la caractéristique de la sinusoïde est exprimée par $k = [k_\nu \cos\phi_\mu, k_\nu \sin\phi_\mu]^t$ avec $k_\nu = 2^{-\frac{\nu+2}{2}}\pi$ où $\nu \in \{0, ..., N-1\}$ et N est le nombre de fréquences, l'angle $\phi_\mu = \mu\,\frac{\pi}{M}$ où $\mu \in \{0, ..., M-1\}$ et M est le nombre d'orientations.
La fonction de Gabor s'écrit alors dans le domaine spatial de la façon suivante [7] :

$$G_k(z) = \frac{||k||^2}{\sigma^2} \exp\,(-\frac{||k||^2||z||^2}{2\sigma^2})(\exp\,(ik^t z) - \exp\,(\frac{\sigma^2}{2})), \tag{3.4}$$

où $z = (x, y)$ présente le pixel, $\sigma = 2\pi$ [7]. La soustraction du deuxième terme de la sinusoïde permet d'éliminer les réponses DC du filtre [132] obtenues par la

partie réelle du filtre (cosinus). La présence des réponses DC peut détériorer la détection des hautes fréquences et perturber la détection des basses fréquences. Les résultats de l'amplitude de Gabor peuvent en être affectés.
Le filtre de Gabor doit respecter certaines règles pour donner des résultats fiables dans le domaine discret. La plus évidente selon Kamarainen et al [133] est la fréquence de Nyquist. En effet, la réponse du filtre doit être négligeable en dépassant la fréquence de Nyquist. Un pourcentage considérable de l'enveloppe du filtre doit ainsi respecter la limite de Nyquist dans le domaine fréquentiel. Dans le domaine spatial, la limite de Nyquist est exprimée par la taille minimale n_{min} du rectangle englobant le filtre de Gabor, approximé par la condition [133] :

$$\int\int_{-\frac{n_{min}}{2}}^{\frac{n_{min}}{2}} |G(x,y)|\, dxdy > p_s \int\int_{-\infty}^{\infty} |G(x,y)|\, dxdy \tag{3.5}$$

où p_s est le pourcentage de l'enveloppe du filtre à l'intérieur d'une fenêtre de taille $n_{min} \times n_{min}$ où $n_{min} \leq min(M,N)$ pour une image de taille $M \times N$ [133]. Nous prenons cette contrainte en considération dans la construction du banc de filtres.

Dans les sections suivantes, nous détaillons la normalisation du visage et sa convolution avec un banc de filtre.

Normalisation des visages

Cette étape est indispensable pour le filtrage du visage. Elle permet de mettre tous les visages à la même échelle en se basant sur la position des yeux. En d'autres termes, à l'issue de cette étape, les visages ont le même nombre de pixels et ces derniers réfèrent aux mêmes zones géométriques. Ainsi les pixels situés sur le front ou appartenant aux yeux pour un sujet, le sont pour tous les autres sujets. La figure 3.5 présente des exemples de la normalisation de plusieurs visages en se basant sur la position des yeux, puisque la distance entre les deux yeux est fixe.

FIGURE 3.5 – Normalisation des visages en se basant sur la position des yeux

Pour normaliser les visages nous procédons d'abord à une détection de la position de chaque œil par la même approche utilisée précédemment pour la détection du visage [126]. Le centre $C(C_x, C_y)$ entre les deux yeux est ensuite calculé, ainsi que l'angle de rotation ω entre l'axe des yeux et l'axe horizontal. Nous calculons l'échelle s selon laquelle nous diminuons la taille de l'image. Finalement, nous appliquons une transformation affine au visage en utilisant la matrice M : $M = \begin{pmatrix} a & b & -a\,C_x - b\,C_y + \alpha\,l \\ -b & a & b\,C_x - a\,C_y + \beta\,h \end{pmatrix}$ avec $a = s\,cos\omega$ et $b = s\,sin\omega$, les valeurs de α et β sont choisies de façon empirique (pour une normalisation du visage 60×80, $\alpha = 0.45$ et $\beta = 0.40$).
Dans la suite, nous normalisons les visages à 60×80 pixels.

Banc de filtres

Une fois le visage normalisé, nous le convoluons avec un banc de filtres de Gabor. Le banc de filtres est obtenu suite à la variation des fréquences et des orientations dans le filtre principal (voir l'équation 3.4). La figure 3.6 illustre la variation des fréquences et des orientations d'un banc de filtres de Gabor. L'utilisation de plusieurs orientations permet de détecter les contours et les textures dans différentes directions, tandis que le changement des fréquences permet de détecter différents détails du visage de tailles variées. La figure 3.7 présente les résultats de la convolution du visage avec différents filtres.

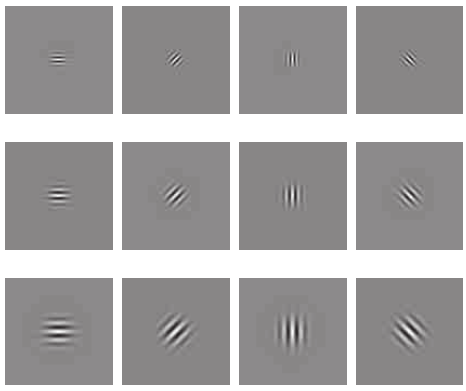


FIGURE 3.6 – Banc de filtres de Gabor (Partie réelle)

Plusieurs types de bancs ont été utilisés dans la littérature de reconnaissance des expressions. Kotsia et al. [7] ont utilisé 12 filtres de Gabor pour des images de

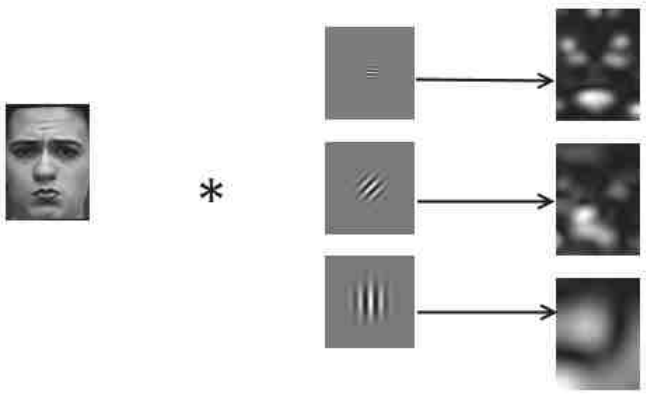


FIGURE 3.7 – Amplitude des convolutions du visage avec trois filtres de différentes orientations et de différentes fréquences

taille 60×80 pixels. Un nombre plus grand de filtres est utilisé par Littelwort et al. [131] à savoir 40 filtres constitués de 8 orientations et 5 fréquences sur un visage normalisé à 48×48 pixels. Bartlett et al. [40] ont également appliqué un filtrage de Gabor pour la reconnaissance des expressions spontanées, 72 filtres sont appliqués sur des images de 96×96.

Dans le cadre de notre travail, nous comparons deux bancs de filtres. Le premier contient 12 filtres (4 orientations et 3 fréquences) et le second contient 40 filtres (8 orientations et 5 fréquences). Nous normalisons tout d'abord le visage à 60×80, puis nous le filtrons par un banc de filtre. Le filtrage est appliqué dans le domaine spatial. Les résultats sont ensuite scannés ligne par ligne et concaténés pour former le vecteur final. Dans le cas d'un banc de 12 filtres, nous obtenons 57600 descripteurs pour représenter un visage, tandis que ce dernier sera représenté par 192000 descripteurs pour un banc de 40 filtres. Nous appliquons finalement une validation croisée. Les taux de reconnaissance des émotions simulées (base de données CK+) et des émotions spontanées (base de données FEEDTUM) obtenus sont présentés dans la table 3.3.

%	Base CK+	Base FEEDTUM
Banc 12 filtres	90.3	84.2
Banc 40 filtres	92.5	84.7

TABLE 3.3 – Taux de reconnaissance avec deux bancs de filtres

Nous remarquons que l'extraction des textures du visage par un banc de 40 filtres améliore le taux de reconnaissance de plus de 2% dans la base CK+ et de 0.5% dans la base FEEDTUM. Le choix d'un banc de filtres plus grand améliore donc

la reconnaissance des émotions. Nous adoptons dans la suite de notre travail un banc de 40 filtres de Gabor.

3.5.2 Méthode Géométrique

Les méthodes géométriques extraient les informations sur les formes, les positions et les mouvements des caractéristiques du visage à savoir les yeux, les sourcils, le nez et la bouche. Généralement, des points sont utilisés pour coder ces informations. Le nombre de points extraits diffère d'une méthode à une autre selon les caractéristiques faciales à coder. Les déformations sont codées soit en utilisant directement les coordonnées des points, soit en utilisant les distances entre les points. Dans les deux cas, la localisation des points peut engendrer des erreurs.
Nous utilisons la méthode de Abdat et al. [31] basée sur des mesures anthropométriques du visage. Ces mesures collectées à partir de plusieurs individus mettent en évidence certaines informations communes à tous les sujets. Des règles sont ainsi définies pour localiser les points caractéristiques permettant en même temps de considérer la différence des morphologies faciales inter-sujets.
La localisation de ces points est basée sur le calcul de la position des axes suivants :

- Axe horizontal des yeux.
- Axe horizontal de la bouche.
- Axe de symétrie du visage.

Localisation des axes

Deux hypothèses sont indispensables à la localisation des axes à savoir :

- Le visage doit être face à la caméra (face frontale).
- La localisation des axes s'effectue sur un visage neutre.

Axe des yeux

La localisation de l'axe des yeux est basée sur la projection des gradients. Un calcul du gradient horizontal est d'abord appliqué par la différence des colonnes sur tout le visage. La projection horizontale de ces gradients permet de déterminer l'axe des yeux, qui présente le gradient maximal. En effet, le passage de la couleur de peau

au blanc de l'œil puis à l'iris produit plusieurs gradients. Leurs sommes présentent un pic lors de la projection. La figure 3.8 illustre la projection des gradients et montre le pic correspondant à l'axe des yeux.

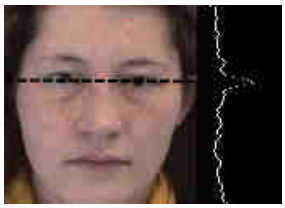


FIGURE 3.8 – Projection horizontale des gradients pour la détection de l'axe des yeux

Axe de la bouche

Une région d'intérêt (ROI) est d'abord définie autour de la bouche, puis transformée vers l'espace couleur TSV (Teinte, Saturation, Valeur). Une segmentation de la bouche est appliquée en utilisant les seuils suivants :

- $120 < T < 200$
- $S > 70$

L'axe de la bouche est la droite horizontale passant par le centre $C\,(C_x, C_y)$ de la bouche segmentée.

$$C_x = \frac{M_{10}}{M_{00}},\ C_y = \frac{M_{01}}{M_{00}}$$

avec

$$M_{00} = \sum_{x,y \in ROI} ROI\,(x,y)$$
$$M_{10} = \sum_{x,y \in ROI} xROI\,(x,y)$$
$$M_{01} = \sum_{x,y \in ROI} yROI\,(x,y)$$

M_{00} présente le moment de premier ordre et M_{10} et M_{01} sont les moments de second ordre. La segmentation de la bouche est présentée dans la figure 3.9.

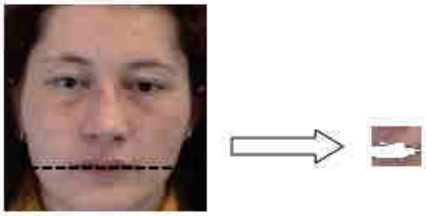


FIGURE 3.9 – Segmentation de la bouche par la couleur et la saturation

Axe de symétrie

Entre les deux axes précédents, une région d'intérêt est localisée. Ensuite, les niveaux de gris de cette région sont projetés verticalement. L'axe de symétrie est présenté par le pic de l'histogramme des niveaux de gris projetés verticalement comme le montre la figure 3.10.

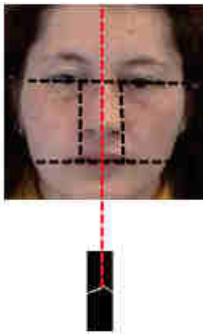


FIGURE 3.10 – Projection verticale des niveaux de gris de la région du nez limitée par l'axe des yeux et l'axe de la bouche. Le pic de l'histogramme des niveaux de gris présente la position de l'axe de symétrie

Positionnement des points

En se basant sur trois critères, à savoir l'axe de symétrie, la distance entre l'axe de la bouche et l'axe des yeux et le centre du visage, les coordonnées de 38 points sont définies. La figure 3.11 illustre les positions de ces points.
Deux types de points sont distingués :

- des points invariants par rapport à l'expression sont localisés dans les zones les plus stables du visage. Représentés dans la figure 3.11 par des points en rouge, ils sont détectés sur le contour du visage, les coins extérieurs des yeux et le haut du nez.

- des points mobiles sont localisés au niveau des coins inférieurs du nez, du contour de la bouche, sur les paupières et sur les sourcils. Ils sont représentés par des points verts dans la figure 3.11.

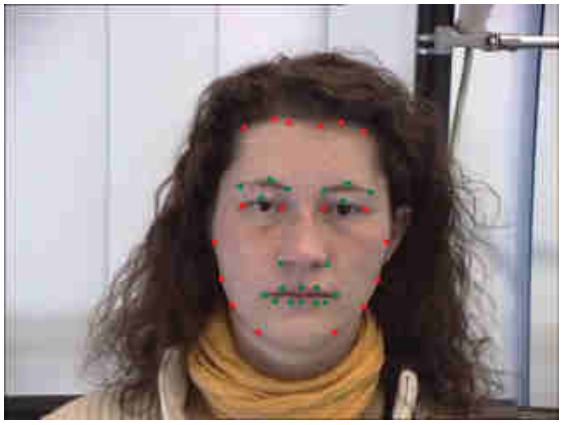


FIGURE 3.11 – Détection des points invariants (en rouge) et des points mobiles (en vert) par rapport à l'expression. Tous les points sont définis sur le visage neutre

Précision de la position des points

Abdat et al. [31] ont amélioré la précision de la localisation des points par l'application de la méthode Shi-Thomasi [134] au voisinage de chaque point. Cette méthode est connue pour son efficacité dans la détection des points d'intérêt dans une image. Elle permet la localisation du point ayant la plus grande variation d'intensité pouvant correspondre à un point physique dans la réalité.
Une fenêtre est définie autour de chaque point dans laquelle la méthode Shi-Thomasi [134] est appliquée pour ajuster la précision des points. Plusieurs tailles de fenêtre ont été testées ; des fenêtres 5×5, 8×8 et 15×15. D'après [31], les fenêtres de taille 5×5 sont très petites donc les points restent les mêmes. Les fenêtres de taille 15×15 sont très larges et peuvent englober d'autres points proches comme les points au niveau de la bouche. L'application de la méthode dans une fenêtre de taille 8×8 améliore la précision de la localisation de certains points (le résultat est présenté dans la figure 3.12).

Calcul des distances entre les points

Une fois la précision des points améliorée par la méthode Shi-Thomasi, les distances entre les points invariants et les points mobiles sont calculées. Sept distances sont

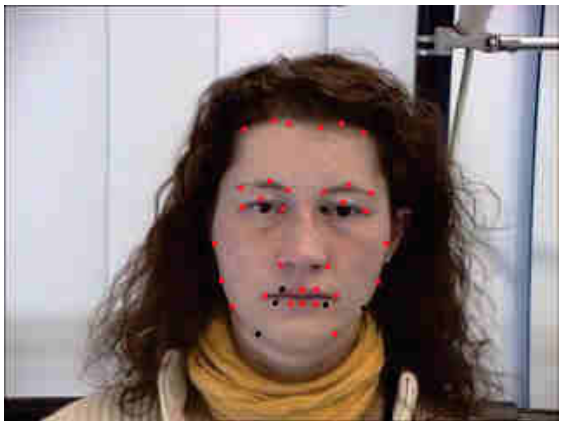


FIGURE 3.12 – Précision des points par la méthode Shi-Thomasi [134] sur des fenêtres de taille 8×8. Les points rouges sont redéfinis par la méthode Shi-Thomasi et les points noirs sont invariants par rapport à l'application de la méthode

utilisées pour coder les mouvements des sourcils, deux décrivent le mouvement des paupières et deux décrivent les déformations du nez. La bouche est codée par 10 distances. La figure 3.13 illustre toutes ces distances. L'expression est codée par les distances calculées dans l'image courante (D_i) et normalisées par les distances dans l'image neutre D_{0i}.

$$\Delta D = \left(\frac{D_1}{D_{01}}, \ldots, \frac{D_i}{D_{0i}}, \frac{D_{21}}{D_{021}} \right)$$

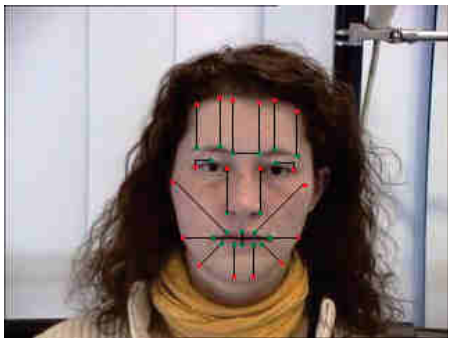


FIGURE 3.13 – Distances entre les points caractéristiques

La méthode géométrique proposée dans [31] permet principalement une bonne localisation des points indépendamment des sujets et des changements de luminosité puisque le positionnement des points est lié seulement à un calcul de coordonnées. Peut-on en dire autant pour la localisation des axes ? En effet, les axes sont localisés par différentes techniques qui peuvent avoir leurs propres limites. Dans la

section suivante, nous nous sommes intéressés à la quantification de la précision du positionnement des axes.

Quantification de la précision de localisation des axes

La bonne localisation des axes est une étape indispensable à la localisation des points. Une erreur dans cette étape engendre des erreurs dans la localisation de plusieurs points caractéristiques. Afin de pouvoir quantifier de potentielles erreurs de localisation des axes, nous avons construit un ensemble de vérités terrain réalisées à partir de la base CK+ et de la base FEEDTUM.
Pour chaque image neutre, nous plaçons manuellement six points de repère sur le visage. Comme le montre l'image (a) de la figure 3.14, deux points sont positionnés sur l'iris des yeux pour localiser l'axe des yeux ; deux points positionnés sur les coins de la bouche permettent le tracé de l'axe de la bouche ; deux points sont placés verticalement, un sur le nez et l'autre sur le front pour le tracé de l'axe de symétrie. Un tracé automatique des axes idéaux à partir des points repères est ensuite effectué. L'image (b) de la figure 3.14 illustre ce processus et l'obtention des trois axes idéaux.

(a) (b)

FIGURE 3.14 – Vérité terrain pour un sujet de la base FEEDTUM. Les points sont placés manuellement et servent de référence pour l'obtention des axes de la vérité terrain.

En se basant sur les axes de la vérité terrain, nous quantifions le taux d'erreur de positionnement des axes de la méthode présentée dans [31] ; nous fixons de façon empirique une plage autour de chaque axe de vérité terrain, si l'axe calculé par la méthode [31] est en dehors de cette plage alors il est considéré comme erroné. La quantification du taux de déplacement de ces axes par rapport à la position idéale montre que les techniques utilisées pour la localisation présentent un taux

d'erreur élevé, notamment pour l'axe de symétrie et l'axe des yeux (voir la table 3.4). En effet, la détection de l'axe de symétrie est basée sur la projection verticale des niveaux de gris, ceci suppose que l'axe de symétrie reflète plus de lumière. Par conséquent, cette hypothèse est très liée à la position de la source de lumière. L'axe des yeux présente également un taux élevé de détections erronées. Nous pensons que la façon de calculer les gradients en est la cause. Nous proposons, dans la suite, des améliorations pour la détection des trois axes.

Axes du visage	Base CK+	Base FEEDTUM
Axe de symétrie	14.5%	7.7%
Axe des yeux	11.3%	5%
Axe de la bouche	0.6%	1.5%

TABLE 3.4 – Quantification des erreurs de positionnement des axes du visage calculés par [31]

Amélioration de la détection des axes

Nous proposons quelques modifications pour améliorer la détection de chaque axe. Nous avons mentionné précédemment que l'axe de symétrie est très lié à la source de lumière et pour être plus indépendant des conditions d'enregistrement des sujets de la base, nous proposons de choisir la médiane de la boite englobante du visage. Ainsi, nous gardons l'hypothèse d'un visage face à la caméra cité précédemment dans la sous-section 3.5.2. La projection horizontale des gradients est retenue pour le calcul de l'axe des yeux. Cependant, nous modifions la façon de les calculer. A la place de la différence entre les colonnes utilisée dans [31], nous utilisons un calcul par le filtre de Sobel. En effet, nous pouvons remarquer dans la figure 3.15 que les contours verticaux détectés par un filtrage de Sobel sont plus visibles que ceux détectés par le calcul de la différence des colonnes de l'image. Le filtrage par Sobel met en évidence les gradients présentant le passage de la peau au blanc des yeux, puis du blanc des yeux à l'iris. Pour détecter l'axe de la bouche, nous appliquons le filtre de Sobel sur la même région d'intérêt définie dans [31] autour de la bouche afin de détecter les gradients verticaux. Ensuite, nous projetons horizontalement la somme des gradients. La somme maximale présente l'axe de la bouche.

Pour procéder à une quantification des erreurs de localisation des axes dans la version que nous proposons, nous appliquons la même procédure que celle décrite précédemment et comparons la position des axes obtenus après modifications par

(a) Résultat de la différence de colonne (b) Résultat d'un filtrage par Sobel

FIGURE 3.15 – Résultat de la détection de gradient par (a) la différence de colonne (b) le filtrage par Sobel

rapport à la plage fixée autour des axes idéaux. La table 3.5 montre que le taux d'erreur de localisation a diminué dans les deux bases. En effet, l'amélioration de la localisation de l'axe de symétrie permet d'annuler le taux d'erreur pour cet axe dans les deux bases. Une nette diminution des erreurs de positionnement de l'axe des yeux est également à noter, se traduisant par une baisse de 6.9% dans la base CK+ et de 3.8% dans la base FEEDTUM. La modification du calcul de l'axe de la bouche améliore sa détection dans la base FEEDTUM. Pour cette base, le taux d'erreur de localisation est même réduit à 0%. Cependant dans la base CK+, le taux d'erreur de l'axe de la bouche augmente à 2%.

Axes du visage	Base CK+	Base FEEDTUM
Axe de symétrie	0%	0%
Axe des yeux	4.4%	1.2%
Axe de la bouche	2%	0%

TABLE 3.5 – Quantification des erreurs de positionnement des axes du visage après les modifications proposées

Taux de reconnaissance en se basant sur	Base CK+	Base FEEDTUM
Les axes de [31]	86.8	45.1
Les axes modifiés	90	48.7

TABLE 3.6 – Taux de reconnaissance des émotions par la méthode géométrique en utilisant les axes basés sur les techniques utilisées dans [31] et les axes modifiés.

Nous avons également calculé le taux de reconnaissance des émotions suite à l'amélioration des axes (voir la table 3.6). Les changements apportés à la localisation des trois axes ont amélioré le taux de reconnaissance dans les deux bases. En effet,

le taux de reconnaissance des émotions simulées a augmenté de 3.2%, tandis que le taux de reconnaissance des émotions spontanées s'est amélioré de 3.6%.

3.6 Comparaison entre les descripteurs d'apparence et les descripteurs géométriques

Les expressions faciales provoquent des changements au niveau de l'apparence du visage et au niveau des formes de ses caractéristiques. Nous distinguons ainsi deux types de descripteurs à savoir les descripteurs d'apparence extraits par des filtres de Gabor et les descripteurs issus de la méthode géométrique qui traduisent le changement de la forme des caractéristiques faciales.
Dans cette section, nous comparons les taux de reconnaissance des émotions décrites par chaque type de descripteurs, d'abord pour les émotions simulées de la base CK+, ensuite pour les émotions spontanées de la base FEEDTUM. La robustesse de ces descripteurs face à des contraintes, comme la variation des résolutions, la variation de la taille des ensembles d'apprentissage et le temps d'exécution, est également étudiée.
Dans la suite de ce chapitre, nous utilisons la technique de validation croisée pour le calcul des taux de reconnaissance. Les matrices de confusion présentent les taux de reconnaissance de chaque émotion et les taux de fausse détection. Les lignes des matrices correspondent aux labels des émotions à reconnaitre et les colonnes correspondent aux émotions estimées.

3.6.1 Emotions simulées (base CK+)

Les émotions simulées correspondent à des expressions socialement apprises. Elles sont très utiles lors de l'interaction sociale et font partie de la communication non verbale. Déclenchées par un mécanisme différent des émotions spontanées, les émotions simulées présentent un sujet d'étude aussi intéressant que les émotions spontanées. Nous comparons dans cette sous-section les taux de reconnaissance obtenus par la méthode d'apparence et par la méthode géométrique dans le cas des émotions simulées.
Les tables 3.7 et 3.8 sont respectivement les matrices de confusion de la méthode de Gabor et de la méthode géométrique calculées sur la base CK+. Nous remar-

quons que la moyenne du taux de reconnaissance de la méthode de Gabor est supérieure de 2.5% de celle de la méthode géométrique. Cependant, les taux de reconnaissance de la peur, de la surprise et de l'expression neutre calculés avec la méthode géométrique sont plus élevés respectivement de 3.4%, 1.7% et 16% que ceux calculés avec la méthode de Gabor. Nous constatons ainsi que chaque méthode est avantageuse pour un ensemble d'émotions.
Nous notons également que les deux méthodes atteignent des taux de reconnaissance supérieurs à 90% pour la joie et la surprise et que le taux de la reconnaissance de la colère est de 90% pour la méthode de Gabor et de 89.5% pour la méthode géométrique. Bien que toutes les expressions dans cette base d'images soient simulées, ces trois émotions semblent être plus exagérées que d'autres.
Nous constatons une large différence entre les taux de reconnaissance calculés par la méthode de Gabor et les taux de reconnaissance calculés par la méthode géométrique pour quelques émotions. Nous citons à titre d'exemple la tristesse qui atteint un taux de 100% avec la méthode de Gabor et un taux de 83.1% avec la méthode géométrique et inversement l'expression neutre atteint un taux de 100% avec la méthode géométrique et un taux de 84% avec la méthode de Gabor. Nous pouvons alors conclure que la tristesse est mieux définie par la déformation de la texture faciale et l'expression neutre est bien définie par les distances entre les points caractéristiques.

%	Joie	Colère	Peur	Dégout	Tristesse	Surprise	Neutre
Joie	**100**	0	0	0	0	0	0
Colère	0	**90**	0	0	0	0	10
Peur	4	0	**85.7**	2.2	4	0	4
Dégout	0	0	4	**96**	0	0	0
Tristesse	0	0	0	0	**100**	0	0
Surprise	0	0	0	0	4	**92**	4
Neutre	0	6	0	2	8	0	**84**

TABLE 3.7 – Matrice de confusion de la méthode de Gabor calculée sur la base CK+

3.6.2 Emotions spontanées (la base FEEDTUM)

Les émotions spontanées sont des émotions innées. Elles sont caractérisées par leur faible intensité par rapport aux expressions simulées. L'évaluation subjective

%	Joie	Colère	Peur	Dégout	Tristesse	Surprise	Neutre
Joie	**93.7**	0	0	6.2	0	0	0
Colère	0	**89.5**	0	0	6.2	0	4.2
Peur	10.8	0	**89.1**	0	0	0	0
Dégout	4.2	4.2	2	**81.3**	6.2	0	2
Tristesse	0	2	4.2	0	**83.1**	0	10.6
Surprise	2	0	2	2.2	0	**93.7**	0
Neutre	0	0	0	0	0	0	**100**

TABLE 3.8 – Matrice de confusion de la méthode géométrique calculée sur la base CK+

a montré que ce point constitue l'inconvénient majeur de la reconnaissance des émotions spontanées.

Dans cette sous-section, nous avons choisi dans la base FEEDTUM les séquences d'images les plus représentatives comportant une seule émotion du début jusqu'à la fin. Nous comparons dans ce qui suit les taux de reconnaissance des émotions calculés par la méthode de Gabor et par la méthode géométrique.

Les tables 3.9 et 3.10 présentent les matrices de confusion respectives de la méthode de Gabor et de la méthode géométrique sur la base FEEDTUM. Le taux de reconnaissance moyen de la méthode de Gabor est de 84.7%, tandis que celui de la méthode géométrique est de 48.7%. Nous remarquons que les deux méthodes ont des taux de reconnaissance inférieurs à ceux obtenus pour les émotions simulées. Cette diminution est due aux faibles intensités des expressions spontanées. Nous notons également que le taux de reconnaissance moyen de la méthode de Gabor est supérieur au taux de reconnaissance moyen de la méthode géométrique de 36%. La méthode de Gabor est ainsi plus adaptée pour la reconnaissance des émotions spontanées. Nous constatons donc, que malgré la faible intensité, les déformations des textures faciales permettent de détecter les changements dus aux expressions. Nous remarquons également que la méthode géométrique atteint des taux de reconnaissance supérieurs à 74% pour trois émotions spontanées à savoir la joie, la surprise et l'expression neutre. Dans le cas de la base CK+, ces trois émotions avaient également les meilleurs taux de reconnaissance par la méthode géométrique par rapport aux autres émotions. Elles semblent ainsi bien définies par les distances caractéristiques.

%	Joie	Colère	Peur	Dégout	Tristesse	Surprise	Neutre
Joie	**98**	2	0	0	0	0	0
Colère	0	**85.1**	2.2	4.2	8.4	0	0
Peur	4.4	0	**76**	2	2.2	8.8	6.4
Dégout	4	6.2	0	**83.7**	4	0	2
Tristesse	0	10	2	0	**73.3**	0	14.6
Surprise	0	0	10.6	0	0	**89.3**	0
Neutre	0	2	0	0	10.2	0	**87.7**

TABLE 3.9 – Matrice de confusion de la méthode de Gabor calculée sur la base FEEDTUM

%	Joie	Colère	Peur	Dégout	Tristesse	Surprise	Neutre
Joie	**79.1**	4.2	0	4.2	0	8.2	4.2
Colère	8	**56.6**	0	19.1	8	0	8.2
Peur	8.8	25.5	**25.7**	0	4.2	33.3	22.2
Dégout	26	34.6	4	**12.6**	10.2	8	4.4
Tristesse	12.8	39.5	8	2	**16.6**	2	18.8
Surprise	10.4	2.2	10.8	0	2	**74.4**	0
Neutre	0	6	0	0	18	0	**76**

TABLE 3.10 – Matrice de confusion de la méthode géométrique calculée sur la base FEEDTUM

3.6.3 Variation des résolutions

Les systèmes de reconnaissance des émotions sont utilisés dans différentes applications. Ces dernières ne sont pas toujours basées sur des vidéos de bonne qualité. En effet, les vidéos telles que les vidéo-surveillances et les vidéos de conférences à distance ont de faibles résolutions, ce qui rend la reconnaissance des expressions faciales d'autant plus difficile. La compression des images avec perte dégrade également la qualité des images. Ainsi, la prise en compte de la dégradation de la qualité de l'image lors de la conception d'un système de reconnaissance des émotions est indispensable.

Dans cette section, nous étudions l'impact du changement de résolution des visages sur les taux de reconnaissance calculés par la méthode de Gabor et la méthode géométrique, sur la base CK+. La figure 3.16 montre les différentes résolutions testées. Dans le cas des visages de résolutions 120×160 et 60×80, les détails du visages sont visibles et détectables comme les contours des yeux et le contour de la bouche. Cependant, la détection de ces détails devient plus difficile pour des résolutions plus petites comme 30×40 et encore plus pour la résolution 15×20

où les contours deviennent flous.

La table 3.11 présente les taux de reconnaissance des deux méthodes présentées

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

FIGURE 3.16 – Changements de résolution du visage : (a) 120 × 160 pixels, (b) 60 × 80 pixels, (c) 30 × 40 pixels (d) 15 × 20 pixels

précédemment. Nous remarquons que la méthode de Gabor permet d'extraire les caractéristiques du visage même lorsque les résolutions médiocres sont considérées. Les taux de reconnaissance des émotions calculés par les descripteurs d'apparence diminuent cependant avec la résolution. En effet, d'un visage de résolution 120×160 à une résolution de 15 × 20, le taux de reconnaissance des émotions diminue de presque 20%. En revanche, le taux de reconnaissance de la méthode géométrique diminue de 46.5% pour un passage d'une résolution de 120 × 160 à une résolution de 30 × 40. Nous constatons également que la méthode géométrique n'arrive pas à extraire les caractéristiques du visage pour une résolution de 15 × 20, puisque les gradients deviennent indétectables.

La méthode de reconnaissance basée sur les descripteurs géométriques est ainsi moins fiable pour la reconnaissance des émotions à faibles résolutions. D'une part, le taux de reconnaissance se détériore très vite avec la dégradation de la résolution. D'autre part, les descripteurs ne sont même plus détectables pour les visages de très faible résolution (15 × 20 pixels). Sur ce point, nous rejoignons les conclusions de l'étude menée par Shan et al [9]. Leur expérience consiste à tester l'impact du changement de résolution dans la base Cohn-Kanade (première version de CK+). Ils comparent leur méthode avec d'autres méthodes présentées dans [55]. Les résultats montrent que les descripteurs d'apparence sont plus performants que les descripteurs géométriques dans la reconnaissance des émotions dans des images à faibles résolutions.

%	120 × 160	60 × 80	30 × 40	15 × 20
Descripteurs d'apparence	95	92.5	87.5	75.7
Descripteurs géométriques	83.5	83.5	37	_

TABLE 3.11 – Taux de reconnaissance des émotions après changements de résolution des visages

	268 images	200 images	136 images	68 images
Méthode d'apparence	92.5	87.3	78.3	65.2
Méthode géométrique	90	87	82.5	69.5

TABLE 3.12 – Taux de reconnaissance de la méthode de Gabor et de la méthode géométrique sur différents ensembles d'apprentissage

3.6.4 Variation de la taille des ensembles d'apprentissage

Dans certains cas le nombre d'images d'apprentissage est limité. De telles situations doivent être prises en considération dans les systèmes de reconnaissance automatique des émotions. Nous étudions dans ce paragraphe les effets du changement de la taille des ensembles d'apprentissage sur les taux de reconnaissance calculés par les deux méthodes décrites précédemment.

La table 3.12 présente les taux de reconnaissance calculés par la validation croisée pour la méthode d'apparence et pour la méthode géométrique sur différents ensembles d'apprentissage de la base CK+. Nous remarquons que les taux de reconnaissance des deux méthodes diminuent avec la diminution de la taille de l'ensemble d'apprentissage. Pour un ensemble de 268 images, le taux de reconnaissance de la méthode d'apparence dépasse celui de la méthode géométrique de 2.5%, mais il chute rapidement avec la diminution des ensembles d'apprentissage pour devenir inférieur de 4.3% au taux de reconnaissance de la méthode géométrique pour un ensemble d'apprentissage de 68 images.

Les deux méthodes ont besoin d'au moins 200 images pour avoir des résultats satisfaisants, qui dépassent 87%. Cependant, la méthode géométrique reste légèrement plus robuste pour des ensembles d'apprentissage au alentour de 100 images.

3.6.5 Temps d'exécution

Le temps d'exécution du système de reconnaissance des émotions est d'un grand intérêt. En effet, la rapidité de la méthode de reconnaissance des émotions est un critère important pour évaluer la qualité du système. Par exemple, dans une application d'agent virtuel, il faut que l'agent soit capable de connaitre les émotions de l'utilisateur pour une interaction en temps réel.
Nous évaluons et comparons dans cette section le temps mis par chaque méthode pour extraire les descripteurs de l'expression. Dans le cas de la méthode d'apparence, les filtres sont appliqués sur l'image comportant l'expression. En revanche, la méthode géométrique extrait les descripteurs du visage neutre, puis une étape de suivi est appliquée. La table 3.13 présente le temps d'extraction des caractéristiques d'apparence de l'expression, le temps d'extraction des caractéristiques géométriques du visage neutre et le temps de suivi de tous les points d'une image à une autre. Le temps mis pour l'extraction des caractéristiques géométriques d'une image dans une séquence vidéo dépend de la position de l'image par rapport à l'image neutre. Dans le cas où la reconnaissance est effectuée pour toutes les images de la séquence, alors l'extraction des descripteurs géométriques est nettement plus rapide que l'extraction des descripteurs de Gabor.

	Temps d'extraction (en s)
Descripteurs d'apparence	0.3
Descripteurs géométriques	0.014 (pour l'image neutre), 0.013 (suivi d'une image)

TABLE 3.13 – Temps d'extraction des descripteurs d'apparence et des descripteurs géométriques

	Temps de classification (en s)
Descripteurs d'apparence	0.18
Descripteurs géométriques	0.00012

TABLE 3.14 – Temps de classification d'une expression par les descripteurs d'apparence et les descripteurs géométriques

Nous comparons également le temps mis pour la classification des descripteurs d'apparence et la classification des descripteurs géométriques. D'après la table 3.14, la classification d'une expression décrite par les descripteurs géométriques est plus rapide que sa classification par des descripteurs d'apparence. Ce résultat

est dû à la grande différence de taille existant entre les vecteurs de descripteurs associés à chaque méthode.

Conclusion

Dans ce chapitre, nous avons d'abord présenté une étude subjective montrant les difficultés de la reconnaissance des émotions spontanées dans la base FEEDTUM. Nous avons ensuite présenté la structure générale des méthodes de reconnaissance des émotions constituée principalement de trois étapes. La première étape est la détection du visage. Elle se base sur les descripteurs de Haar. La dernière étape est la classification de l'expression dans l'une des sept classes (joie, colère, peur, dégoût, tristesse, surprise et l'expression neutre). Nous nous sommes intéressés de plus près à la deuxième étape qui décrit l'expression. Deux méthodes d'extraction de descripteurs faciaux ont alors été étudiées. La méthode d'apparence extrait les descripteurs du visage en se basant sur des filtres de Gabor de différentes tailles et de différentes orientations. Une deuxième méthode a été également appliquée pour l'extraction de l'aspect géométrique de l'expression. Cette dernière présentée dans [31] se base principalement sur le changement des distances entre des points caractéristiques. Ces points sont localisés par trois axes faciaux. Après la quantification des erreurs de localisation de ces axes, nous avons proposé des améliorations pour leur localisation ce qui nous a permis d'améliorer les taux de reconnaissances dans les deux bases considérées. Les deux types de descripteurs extraits ont été comparés sous différents critères présents dans les applications de reconnaissance des émotions. Nous avons remarqué que les descripteurs d'apparence extraits par la méthode de Gabor sont plus robustes pour les applications comportant des visages de faibles résolutions tandis que les descripteurs géométriques extraits par la méthode géométrique peuvent avoir des résultats satisfaisants pour un ensemble d'apprentissage plus restreint contenant approximativement 100 images. Les temps d'extraction et de classification de ces derniers sont également plus rapides que ceux des descripteurs d'apparence.

Dans le chapitre suivant nous fusionnons les descripteurs d'apparence et les descripteurs géométriques par différentes techniques et comparons les résultats obtenus sur la base des émotions simulées et sur la base des émotions spontanées. Une étude de la reconnaissance dans des zones locales est également effectuée, suivie d'une sélection des descripteurs d'apparence.

Chapitre 4

Différentes approches d'amélioration des descripteurs

4.1 Introduction

Dans le chapitre précédent nous avons comparé les descripteurs d'apparence et les descripteurs géométriques sous plusieurs contraintes, tout en gardant la même méthode pour la détection du visage et la même méthode pour la classification. Nous avons remarqué que chaque type de descripteurs peut être avantageux dans certains contextes. Dans ce chapitre nous nous intéressons à la méthode de classification utilisée afin de voir dans quelle mesure les choix effectués peuvent influer sur les taux de reconnaissance selon chaque type de descripteurs. De plus, l'expression d'une émotion engendre à la fois des changements dans la forme et l'apparence du visage. Nous étudions donc dans ce chapitre une fusion possible entre descripteurs géométriques et descripteurs d'apparence afin d'obtenir une représentation plus complète des émotions. Le choix de la méthode de fusion étant crucial dans ce cadre, nous appliquons plusieurs schémas de fusion, que nous comparons sur les deux bases d'étude et dans le cas d'ajout de nouveaux sujets non inclus dans l'ensemble d'apprentissage. Nous étudions ensuite l'importance des zones locales dans la reconnaissance des émotions et comparons des méthodes utilisant la fusion entre les zones pour améliorer la reconnaissance. Enfin, nous étudions trois méthodes de sélection des descripteurs. La comparaison de ces méthodes est à nouveau effectuée sur les deux bases.

Méthodes	SVM linéaire	SVM gaussien
Méthode d'apparence	92.5	87.8
Méthode géométrique	90	96.4

TABLE 4.1 – Taux de reconnaissance de la méthode d'apparence et la méthode géométrique pour un SVM linaire et de un SVM gaussien dans la base CK+

4.2 Noyau gaussien

Dans le chapitre précédent, nous nous sommes focalisés sur la comparaison entre les descripteurs d'apparence et les descripteurs géométriques. Pour assurer cette comparaison, l'étape de détection du visage et l'étape de classification avaient été fixées. Nous avons ainsi utilisé un SVM linéaire pour classer les deux types de descripteurs. Dans cette section, nous cherchons à améliorer la classification des descripteurs en modifiant le noyau du SVM par un noyau gaussien ($K(x, y) = \exp(-\gamma \, |x - y|^2)$). Les tables 4.1 et 4.2 présentent les taux de reconnaissance de la méthode géométrique et de la méthode d'apparence calculés par un SVM linéaire et un SVM gaussien respectivement dans la base CK+ et dans la base FEEDTUM. Nous avons utilisé la méthode de grille de recherche ("*grid search*") basée sur une validation croisée pour trouver les paramètres optimaux du noyau gaussien, notamment γ.

Nous remarquons que les taux de reconnaissance de la méthode d'apparence calculés par un SVM gaussien dans la base CK+ sont inférieurs aux taux de reconnaissance calculés par un SVM linéaire de 4.7%, tandis que pour la base spontanée le taux de reconnaissance reste stable comme le montre la table 4.2. Ainsi les descripteurs d'apparence sont mieux classés par une discrimination linéaire. En effet, le grand nombre de descripteurs rend le problème linéairement séparable. De plus, la faible valeur de γ (de l'ordre de 10^{-9}) montre que le noyau gaussien a une très faible amplitude confirmant ainsi l'hypothèse de linéarité. Par contre, le taux moyen de reconnaissance des émotions posées calculé par la méthode géométrique passe de 90% avec un noyau linéaire à 96.4% avec un noyau gaussien. Dans la base FEEDTUM, le taux de reconnaissance calculé avec la méthode géométrique pour un SVM gaussien dépasse de 16% le taux de reconnaissance calculé avec le SVM linéaire.

Dans la suite de ce chapitre, nous utilisons un SVM linéaire pour les descripteurs d'apparence et un SVM gaussien pour les descripteurs géométriques.

Méthodes	SVM linéaire	SVM gaussien
Méthode d'apparence	84.7	84.8
Méthode géométrique	48.7	65.4

TABLE 4.2 – Taux de reconnaissance de la méthode d'apparence et de la méthode géométrique pour un SVM linaire et un SVM gaussien dans la base FEEDTUM

4.3 Méthodes de fusion des descripteurs d'apparence et des descripteurs géométriques

L'expression faciale des émotions apportante un ensemble de changements au niveau de l'apparence et au niveau de la forme du visage, une représentation de ces expressions doit également comporter les deux informations. Dans cette section, nous proposons un ensemble de méthodes de fusion permettant de combiner à la fois les informations d'apparence et les informations de forme du visage.

Les méthodes de fusion sont classées principalement en deux types. La fusion en amont, appelée également fusion au niveau des descripteurs, combine les deux types de descripteurs dans un même vecteur qui est ensuite considéré comme entrée pour la méthode de classification. La fusion en aval, appelée aussi fusion au niveau des décisions, classe chaque type de descripteurs de façon individuelle. Les décisions sont ensuite fusionnées. Dans le cadre de notre travail, nous considérons seulement la fusion en aval, puisqu'elle apporte d'une part de la flexibilité dans le choix de la méthode de classification la plus adaptée à chaque type de descripteurs et que, d'autre part, la grande différence de taille entre les descripteurs d'apparence (192000 descripteurs) et les descripteurs géométriques (21 descripteurs) présente une difficulté lors de la fusion en amont. Différents schémas de fusion au niveau des décisions sont appliqués.

Les vecteurs de probabilités issus des classifications des descripteurs d'apparence par un SVM linéaire et des descripteurs géométriques par un SVM gaussien sont notés $P(\omega_k|X_i)$ où ω_k est la *kiéme* émotion, $k \in \{1, ..., n\}$ avec n le nombre d'émotions et $X_i \in \{X_A, X_G\}$ où X_A est le vecteur des descripteurs d'apparence et X_G est le vecteur des descripteurs géométriques. Ces vecteurs de probabilités sont ensuite transmis à l'entité de fusion comme le montre la figure 4.1. Cette entité est remplacée dans ce qui suit par différentes règles statistiques et par des méthodes de classification.

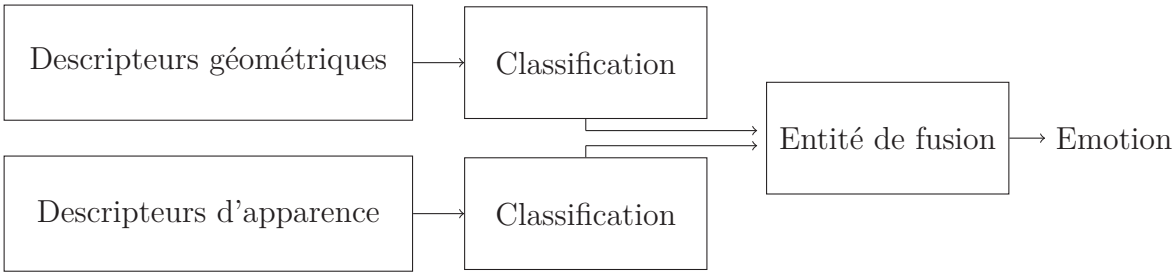


FIGURE 4.1 – Schéma des méthodes de fusion au niveau des décisions

4.3.1 Fusions basées sur des règles statistiques

Ce type de fusion permet de substituer l'entité de fusion par diverses règles statistiques comme la moyenne, le produit, le maximum, le vote majoritaire... Nous considérons les méthodes les plus adaptées à notre situation et qui ne nécessitent pas d'informations a priori.

Fusion basée sur la moyenne

Dans cette technique nous supposons que les deux vecteurs de décision ont la même probabilité a priori. La moyenne des deux vecteurs est d'abord calculée. Ensuite, l'émotion ayant la plus grande probabilité est choisie comme le montre l'équation ci-dessous où m présente le nombre de vecteurs et $X_i \in \{X_G, X_A\}$ présente le vecteur des descripteurs.

$$Z \rightarrow \omega_k$$
$$si\ \frac{1}{m}(\sum_{i=1}^{m} P(\omega_k|X_i)) = \max_k(\frac{1}{m}(\sum_{i=1}^{m} P(\omega_k|X_i))) \qquad |k = \{1, .., n\},$$

Un exemple de mise en œuvre de cette technique est présenté dans la figure 4.2.

Fusion basée sur le produit

En assumant l'indépendance des décisions prises au niveau de chaque classifieur, la densité de probabilité conjointe s'exprime par :

$$P(X_G, X_A|\omega_k) = P(X_G|\omega_k) \times P(X_A|\omega_k)$$

Nous définissons ainsi la fusion par le produit des vecteurs de décision. L'émotion sélectionnée est celle ayant le produit maximal :

$$Z \rightarrow \omega_k$$
$$si\ (\prod_{i=1}^{m} P(\omega_k|X_i)) = \max_k (\prod_{i=1}^{m} P(\omega_k|X_i)) \qquad |k = \{1, .., n\},$$

Un exemple de mise en œuvre est présenté dans la figure 4.2.

Fusion basée sur le maximum

L'émotion choisie est l'émotion ayant la plus grande probabilité dans les deux vecteurs de décision. L'expression suivante explique le choix de l'émotion :

$$Z \rightarrow \omega_k$$
$$si\ \max_i(P(\omega_k|X_i)) = \max_k(\max_i(P(\omega_k|X_i))) \qquad |k = \{1, .., n\},\ i = \{1, ., m\},$$

Un exemple de mise en œuvre est présenté dans la figure 4.2.

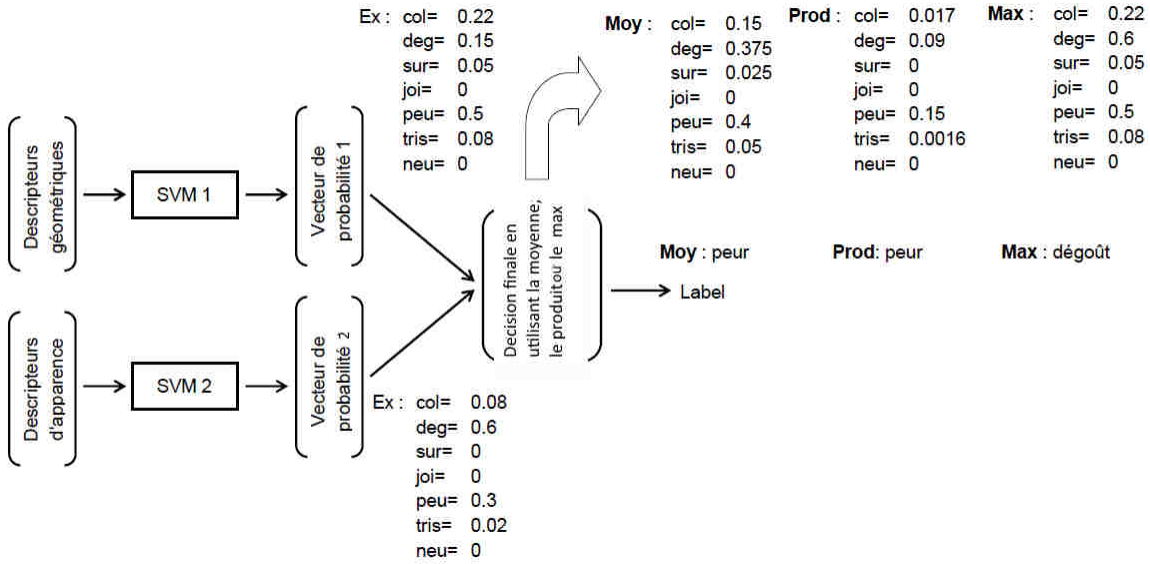


FIGURE 4.2 – Exemple de mise en œuvre des méthodes statistiques

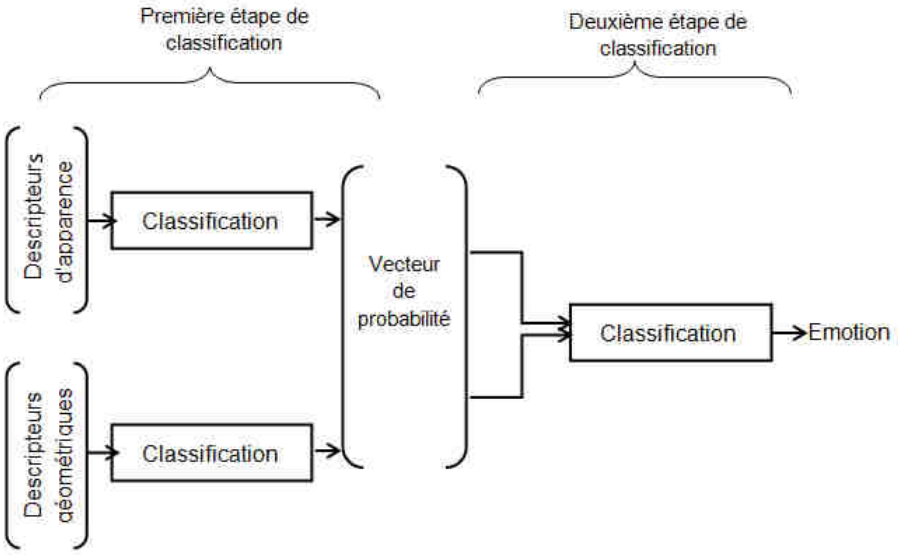


FIGURE 4.3 – Schéma de fusion basée sur une méthode de classification

4.3.2 Fusions basées sur les méthodes de classification

Cette technique est utilisée dans le domaine de l'analyse des multimédias [135], [136] pour fusionner plusieurs modalités de données. Les vecteurs de décisions $P(\omega_k|X_i)$ sont rassemblés dans un même vecteur qui est ensuite considéré comme l'entrée d'une nouvelle méthode de classification (voir figure 4.3). Différentes méthodes d'apprentissages sont utilisées pour la fusion : SVM linéaire, SVM gaussien, Bayésien Naïf (BN), les K les plus proches voisins (KPP).

Bayésien Naïf (BN)

Le Bayésien Naïf (BN) est une méthode de classification générative. Elle se base sur l'hypothèse d'indépendance des descripteurs pour une classe donnée, ce qui revient à :

$$\hat{\omega} = arg\, \max_{\omega} P(\omega) \prod_{i=1}^{n} P(X_i|\omega)$$

Les K plus proches voisins (KPP)

La méthode des K plus proches voisins (KPP) est une méthode d'apprentissage simple et robuste. Elle applique un vote sur les classes des vecteurs d'apprentissage

les plus proches. Le label de la classe présent sur la majorité des K vecteurs est alors assigné au vecteur candidat.

4.4 Evaluation des méthodes de fusion

4.4.1 Comparaison des méthodes de fusion dans la reconnaissance des émotions posées de la base CK+

Dans cette section, nous comparons les taux de reconnaissance de la méthode d'apparence et de la méthode géométrique ainsi que des méthodes de fusion sur la base des émotions posées.
D'après la table 4.3, la méthode d'apparence et la méthode géométrique ne se trompent pas sur les mêmes émotions. En effet, le taux de reconnaissance le plus faible obtenu par la méthode d'apparence correspond à l'expression neutre. Cependant, cette dernière est reconnue avec un taux de 100% par la méthode géométrique. D'autre part, le taux de reconnaissance le plus faible calculé avec la méthode géométrique est de 89.5%. Il est obtenu pour la tristesse, alors que le taux de reconnaissance de cette émotion par la méthode d'apparence est de 100%. La fusion dans ce cadre pourrait compenser les points faibles de chaque méthode. Nous remarquons que les méthodes basées sur des règles statistiques ont des taux de reconnaissance très similaires, néanmoins, la fusion basée sur le produit est légèrement inférieur à la fusion par la moyenne ou par le maximum. Les trois méthodes arrivent à reconnaitre avec un taux de 100% la joie, la tristesse, la surprise et l'expression neutre, permettant ainsi d'améliorer la reconnaissance des émotions pour lesquelles la méthode d'apparence ou la méthode géométrique ont des taux plus faibles comme l'expression neutre pour la méthode d'apparence et la tristesse pour la méthode géométrique. La fusion par des règles statistiques améliore le taux moyen de reconnaissance de 6.6% par rapport à la méthode d'apparence et de 2.7% par rapport à la méthode géométrique.
La fusion par classification améliore également les taux de reconnaissance des émotions par rapport à la méthode d'apparence et à la méthode géométrique, sauf dans le cas de la fusion basée sur la méthode BN. Si cette dernière permet de dépasser légèrement le taux de reconnaissance de la méthode d'apparence de 0.2%, ses performances restent inférieures à celles de la méthode géométrique. Toutefois, elle arrive à un taux de reconnaissance de 100% pour les émotions joie et tristesse.

Méthodes		TR	Joi	Col	Peu	Dég	Tri	Surp	Neu
Apparence		92.5	100	90	85.7	96	100	92	84
Géométrique		96.4	96	93.7	100	96	89.5	100	100
Règles statistiques	Moyenne	99.1	100	100	97.7	96	100	100	100
	Produit	98.8	100	98	97.7	96	100	100	100
	Max	99.1	100	100	97.7	96	100	100	100
Fusion par classification	Linéaire	97.9	100	96	97.7	98	100	100	94
	Gaussien	98.2	100	96	97.7	96	100	100	98
	BN	92.7	100	88	93.7	96	100	98	73.5
	KPP	97.6	100	96	97.7	96	100	100	94

TABLE 4.3 – Taux de reconnaissance des méthodes de fusion calculés par une validation croisée de 5 groupes sur la base de données CK+

Méthode	TR	Joi	Col	Peu	Dég	Tri	Surp	Neu
Chen et al [95]	95.0	97.5	92.5	90.0	96.0	93.5	96.5	_
Kotsia et al [137]	92.3	97.5	93.6	84.3	89.5	94.3	95.6	91.3

TABLE 4.4 – Taux de reconnaissance de méthodes de la littérature utilisant la fusion de descripteurs géométriques et de descripteurs d'apparence.

Nous remarquons également que la fusion par un SVM gaussien donne le meilleur taux de reconnaissance parmi les méthodes de fusion par classification. Cependant, celui-ci reste légèrement inférieur au taux de reconnaissance de la fusion par le produit.

Nous notons également que le dégoût reste l'émotion avec le taux de reconnaissance le plus faible pour toutes les méthodes de fusion basées sur des règles statistiques et pour la méthode de fusion basée sur le SVM gaussien. Par contre, la fusion par un SVM linéaire améliore la reconnaissance du dégoût.

4.4.2 Comparaison avec des méthodes de la littérature

Nous nous sommes intéressés à la comparaison des méthodes de fusion proposées ci-dessus avec des méthodes de la littérature utilisant la fusion de descripteurs géométriques et de descripteurs d'apparence. Chen et al [95] extraient 21 points caractéristiques sur le visage en se basant sur le modèle actif de forme (ASM :*Active Shape Model*). Les expressions sont ensuite codées par le déplacement de ces points par rapport à l'expression neutre. Les textures des expressions sont codées par la différence des gradients dans les deux orientations verticale et horizontale entre l'image de l'émotion et l'image neutre. Les textures sont extraites localement sur

des fenêtres centrées sur chaque point caractéristique. Une fusion au niveau des descripteurs est appliquée pour les descripteurs géométriques et les descripteurs d'apparence. Enfin, un SVM est utilisé pour la classification de 7 émotions à savoir la joie, la colère, la peur, le dégoût, la tristesse, la surprise et le mépris. Par contre l'expression neutre n'est pas considérée dans ce contexte. Kotsia et al [137] détectent l'information géométrique par la déformation du maillage de CANDIDE appliqué sur le visage. Les changements de texture sont extraits par l'algorithme DNMF, une extension de l'algorithme NMF (*Non Negative Matrix Factorization*), qui représente l'expression par une approximation linéaire. Une normalisation est ensuite appliquée aux deux types de descripteurs, suivie d'une fusion dans le même vecteur. Les expressions sont enfin classées par les réseaux de neurones.
La méthode de Chen et al [95] est calculée sur la base CK+, tandis que la deuxième méthode est évaluée sur la base Cohn-Kanade. Tous les sujets de la base Cohn-Kanade sont intégrés dans la base CK+, la base Cohn-Kanade constituant la première version de la base CK+ (une description des deux bases est présentée dans la section 2.8.1). D'après la table 4.4 présentant les taux de reconnaissance des deux méthodes décrites, les taux de reconnaissance des méthodes de fusion basées sur des règles statistiques et celles basées sur des méthodes de classification dépassent les taux de reconnaissance des méthodes présentées dans [95] et [137]. En effet, les taux de reconnaissance moyens des émotions des méthodes de fusion basées sur les règles statistiques sont supérieurs à 98% et les méthodes de fusion basées sur la classification par SVM dépassent le taux de 97%. Par contre, la fusion basée sur la classification par BN a un taux de reconnaissance inférieur à la méthode présentée dans [95], tout en restant légèrement supérieur au taux de reconnaissance de la méthode de fusion présentée dans [137]. Nous notons également que le taux de reconnaissance de la peur est le plus faible pour les deux méthodes de la littérature. Pour cette même émotion les méthodes de fusion que nous avons utilisées ont un taux de reconnaissance de 97.7%, sauf dans le cas de la fusion basée sur BN qui donne un taux de reconnaissance de 93.7%. Ce dernier reste malgré tout supérieur aux taux de reconnaissance de la peur atteints par les méthodes de la littérature. Le point commun entre toutes ces méthodes de fusion est le taux élevé de la reconnaissance de la joie. En effet, dans une base de données posées comme la base CK+ ou la base Cohn-Kanade, la joie est l'émotion la plus facile à simuler. Elle est représentée seulement par l'action unitaire 6 du système de codage facial (FACS) et les différentes intensités de l'action unitaire 12. Des exemples des actions

unitaires formant la joie sont présentés dans la figure 4.4. Nous notons également que la surprise a des taux de reconnaissance très élevés avec les méthodes de la littérature et nos méthodes de fusion, ce qui prouve que les émotions positives sont plus faciles à simuler dans les bases d'émotions posées et par la suite plus simples à reconnaître automatiquement.

(a) (b)

FIGURE 4.4 – Les actions faciales constituant la joie. L'image (a) représente l'action unitaire 6 (haussement des joues) et l'image (b) représente l'action unitaire 12 (étirement du coin des lèvres)

4.4.3 Comparaison des méthodes de fusion sur les émotions spontanées de la base FEEDTUM

Méthodes		TR	Joi	Col	Peu	Dég	Tri	Surp	Neu
Apparence		84.7	98	85.1	76	83.7	73.3	89.3	87.7
Géométrique		65.0	81.3	43.7	52.8	57.1	52.2	82.8	85.7
Règles statistiques	Moyenne	82.7	98	68.8	78.4	82	71.1	89.1	92
	Produit	81.8	98	66.6	82.6	77.7	71.1	86.8	90
	Max	81.2	98	70.6	74.2	71.7	71.1	89.3	93.7
Fusion par classification	Linéaire	85.2	98	87.3	71.7	83.7	79.3	89.3	87.7
	Gaussien	86.5	98	89.3	78.2	83.7	77.3	91.3	87.7
	BN	80.7	93.7	79.1	71.7	86	75.3	80.8	78.8
	KPP	86.2	98	91.3	78.2	79.7	79.3	89.3	87.7

TABLE 4.5 – Taux de reconnaissance des méthodes de fusions calculés par une validation croisée de 5 groupes sur la base de données FEEDTUM

Nous comparons dans la table 4.5 les taux de reconnaissance des émotions obtenus par la méthode d'apparence, la méthode géométrique et les méthodes de fusion sur la base FEEDTUM. Nous remarquons que le taux de reconnaissance de la méthode d'apparence dépasse le taux de reconnaissance de la méthode géométrique de 19.4%, ce qui représente une très grande différence. Comme nous l'avons déjà mentionné dans le chapitre précédent les émotions spontanées ne sont pas très marquées ce qui rend leurs détections plus difficiles avec les descripteurs géométriques.

Les taux de reconnaissance des méthodes de fusion par des règles statistiques sont supérieurs à celui obtenu par la méthode géométrique, cependant ils restent légèrement inférieurs au taux de reconnaissance de la méthode d'apparence. La fusion par moyenne améliore toutefois les taux de reconnaissance de la peur et de l'expression neutre par rapport à la méthode d'apparence et à la méthode géométrique.
Les méthodes de fusion basées sur la classification ont des taux de reconnaissance très intéressants. Les méthodes de fusion basées sur le SVM gaussien, le SVM linéaire et la méthode des K plus proches voisins (KPP) ont des taux de reconnaissance moyens supérieurs à 85%. Elles dépassent le taux de reconnaissance de la méthode géométrique d'au moins 20% et sont également supérieurs au taux de reconnaissance de la méthode d'apparence. La fusion par SVM gaussien et la fusion par KPP améliorent les taux de reconnaissance de la colère, de la peur, de la tristesse et de la surprise par rapport des deux méthodes de base. La méthode de fusion basée sur la classification par BN a quant à elle un taux de reconnaissance moyen inférieur au taux de reconnaissance de la méthode d'apparence. Elle améliore cependant les taux de reconnaissance du dégoût et de la tristesse par rapport à la méthode d'apparence et à la méthode géométrique.
Nous remarquons que globalement les émotions ayant les taux de reconnaissance les plus intéressants pour toutes les méthodes de fusion sont les émotions positives (joie et surprise) et l'expression neutre. Les méthodes de fusion par classification ont également des taux de reconnaissance élevés pour la colère.

4.5 Evaluation des méthodes hybrides face aux changements inter-sujets

Dans cette section, nous comparons les méthodes de fusion qui ont prouvé leur efficacité pour les émotions actées (la base CK+), ainsi que les émotions spontanées (la base FEEDTUM), à savoir la fusion par moyenne, la fusion par SVM gaussien et la fusion par KPP. Nous testons leur stabilité devant une généralisation de la reconnaissance d'émotions à de nouveaux sujets. Pour ce faire, une validation croisée des sujets est opérée. Cette dernière consiste à retirer toutes les images d'un sujet de l'ensemble d'apprentissage et à les utiliser ensuite comme échantillon de test. Cette opération est répétée pour tous les sujets.

4.5.1 Evaluation sur la base CK+

La table 4.6 présente les taux de reconnaissance de la méthode d'apparence, de la méthode géométrique et des trois méthodes de fusion citées précédemment sur la base CK+.
Le changement du noyau SVM d'un noyau linéaire à un noyau gaussien apporte une stabilité et une robustesse aux descripteurs géométriques qui atteignent un taux de reconnaissance assez élevé de 83.9%. En effet, les descripteurs géométriques sont basés sur la variation des distances entre des points caractéristiques. Ces derniers sont localisés indépendamment des sujets. A l'inverse, les descripteurs d'apparence sont très liés aux changements de textures des visages. La moyenne des taux de reconnaissance des émotions ne dépasse pas 74.2%. Cependant, les descripteurs d'apparence ont des taux de reconnaissance supérieurs à 93% pour les émotions positives (joie, surprise) et l'expression neutre.
Les méthodes de fusion ont des taux de reconnaissance intéressants, leurs moyennes dépassent le taux de 86%. La joie, la surprise et l'expression neutre sont reconnues avec un taux de 100% et le taux de reconnaissance de la colère est supérieur à 91% pour les trois méthodes de fusion. Les méthodes de fusion sont ainsi stables devant les changements inter-sujets dans le cas des émotions simulées.

4.5.2 Evaluation sur la base FEEDTUM

Dans cette section, nous généralisons le test sur de nouveaux sujets de la base des émotions spontanées (FEEDTUM). La table 4.7 présente les taux de reconnaissance de la méthode d'apparence, de la méthode géométrique et des méthodes de fusion. Nous remarquons que les taux de reconnaissance sont faibles pour toutes

Méthodes	TR	Joi	Col	Peu	Dég	Tri	Surp	Neu
Apparence	74.2	97.8	67.4	63.8	82.9	52.2	97.2	93.7
Géométrique	83.9	91.3	89.1	87.2	80.8	79.5	100	100
Fusion moyenne	86.6	100	95.6	89.3	82.9	79.5	100	100
Fusion par SVM gaussien	87.8	100	95.6	89.3	82.9	88.6	100	100
Fusion par KPP	86	100	91.3	89.3	80.8	81.8	100	100

TABLE 4.6 – Taux de reconnaissance de la méthode d'apparence, de la méthode géométrique et des méthodes de fusion calculés pour de nouveaux sujets de la base CK+

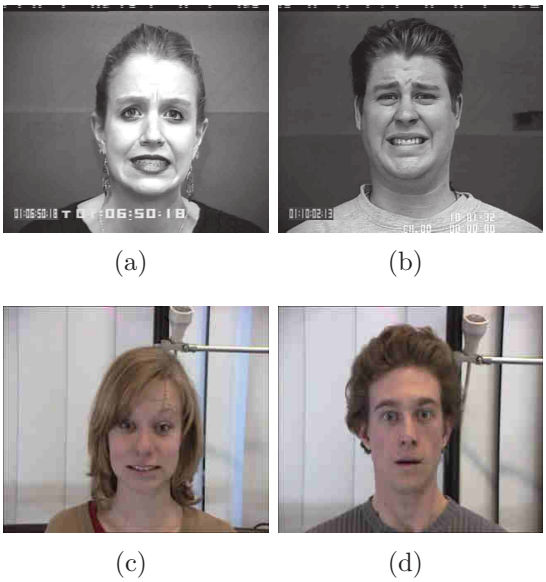
Méthodes	TR	Joi	Col	Peu	Dég	Tri	Surp	Neu
Apparence	40.7	72.9	38.1	50	40.9	34.1	66.6	36.1
Géométrique	46.1	70.2	26.2	50	36.3	39	69.2	85.1
Fusion moyenne	51.4	81	42.8	71	40.9	41.4	74.3	72.3
Fusion par SVM gaussien	43.1	72.9	35.7	55.2	50	34.1	71.7	38.2
Fusion par KPP	45.2	75.6	40.4	55.2	50	39	71.8	42.5

TABLE 4.7 – Taux de reconnaissance de la méthode d'apparence, de la méthode géométrique et des méthodes de fusion calculés pour de nouveaux sujets de la base FEEDTUM

les méthodes. En effet, si les émotions présentées dans la base actée suivent un ensemble de codes faciaux décrivant chaque émotion, les émotions spontanées sont en revanche subtiles et présentent les réactions des participants face aux vidéos visionnées. Les expressions des émotions spontanées sont ainsi très liées aux sujets. Les images (a) et (b) de la figure 4.5 montrent la ressemblance entre les expressions simulant la peur dans la base CK+. Ces expressions sont composées des mêmes actions unitaires. La difficulté de la reconnaissance des émotions de nouveaux sujets réside alors dans la différence de la texture de la peau de ce nouveau visage. Les images (c) et (d) présentent les expressions spontanées de la peur dans la base FEEDTUM. En plus de la différence de texture de la peau, la différence l'expression de la peur chez les deux sujets est visible, rendant ainsi la reconnaissance des émotions de nouveaux sujets encore plus difficile.
Les méthodes de fusion par classification ont des taux de reconnaissance légèrement inférieurs au taux de reconnaissance de la méthode géométrique. Par contre, la méthode de fusion basée sur la moyenne dépasse le taux de reconnaissance de la méthode d'apparence de 10.7% et le taux de reconnaissance de la méthode géométrique de 5.3%. Il reste cependant faible.

Conclusion

La comparaison des méthodes de fusion sur les deux bases montre que la fusion par la moyenne, par la classification SVM et par la classification KPP ont de meilleurs taux de reconnaissance que la méthode géométrique et la méthode d'apparence. Testées sur de nouveaux sujets des mêmes bases, ces méthodes restent robustes lorsqu'il s'agit d'expressions qui ressemblent aux expressions existantes dans l'ensemble d'apprentissage, comme c'est le cas dans la base des émotions simulées. En revanche, pour la base des émotions spontanées les taux de reconnaissance ne



(a) (b)

(c) (d)

FIGURE 4.5 – L'expression de la peur par différents sujets (a) et (b) appartiennent à la base CK+ et (c) et (d) appartiennent à la base FEEDTUM

dépassent pas 52%. En effet, le caractère spontané des émotions les rend plus difficiles à détecter. Ainsi, la même émotion peut être exprimée de façon différente d'une personne à une autre.

Les méthodes de reconnaissance des émotions que nous avons vues jusque là sont toutes appliquées sur la totalité du visage. Elles ont leurs points faibles et leurs points forts. Dans la section suivante, les études menées ont pour objectifs de mieux comprendre l'importance de diverses zones du visage dans la reconnaissance des émotions et les possibilités d'amélioration que peut apporter une fusion des informations locales contenues dans ces différentes zones.

4.6 Descripteurs locaux

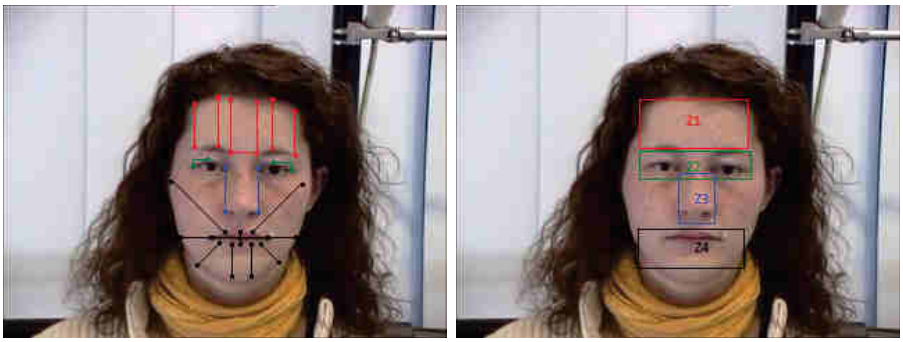
Dans cette section, nous nous intéressons à la reconnaissance des émotions par zones du visage. Nous cherchons tout d'abord à évaluer le taux de reconnaissance des émotions en considérant uniquement certaines parties du visage. Puis, nous évaluons la reconnaissance des émotions pour des méthodes fusionnant les parties du visage les plus représentatives des émotions.

4.6.1 Reconnaissance des émotions dans des zones locales du visage

Nous divisons le visage en quatre zones définies comme suit :

- La zone du front : le front est représenté dans la méthode géométrique par la variation des 7 distances présentées en rouge dans l'image (a) de la figure 4.6. Ces distances codent le mouvement des sourcils durant l'expression. Pour la méthode d'apparence, la zone du front considérée est la zone du visage au dessus des sourcils afin de capter les rides et le changement de la peau dû aux mouvements des sourcils. La fenêtre représentant le front varie avec le changement de l'expression. Elle est présentée dans l'image (b) de la figure 4.6 par le rectangle rouge ($Z1$).
- La zone des yeux : la méthode géométrique représente l'état des yeux par la variation de deux distances présentées en vert sur l'image (a) de la figure 4.6. Dans la méthode d'apparence, nous considérons les yeux et la zone autour des yeux afin de détecter les rides qui peuvent apparaitre durant l'expression. Cette zone est limitée par un rectangle vert ($Z2$) dans l'image (b) de la figure 4.6.
- La zone du nez : elle est définie dans la méthode géométrique par deux distances décrivant les déformations du nez durant l'expression. Dans la méthode d'apparence, nous considérons le rectangle englobant le nez. Cette zone est représentée par les lignes en bleu dans l'image (a) et par le rectangle bleu dans l'image (b) de la figure 4.6.
- La zone de la bouche : elle est définie dans la méthode géométrique par les 10 distances liant les points sur le contour de la bouche aux points se trouvant sur le contour du visage. Les distances sont représentées par les lignes noires dans l'image (a) de la figure 4.6. Pour la méthode d'apparence, nous considérons toute la zone située au dessous de l'axe horizontal passant par le point le plus haut de la lèvre supérieure, représentée par la zone $Z4$ dans l'image (b) de la figure 4.6. La fenêtre englobant la bouche varie avec le changement de l'expression.

La table 4.8 présente les moyennes des taux de reconnaissance des émotions calculées sur la base des émotions simulées (la base CK+) pour la méthode d'apparence et la méthode géométrique sur chacune des quatre zones présentées précédemment. Nous avons appliqué un SVM linéaire pour chacune des zones extraites par la mé-



(a) (b)

FIGURE 4.6 – Représentation des zones du visage : (a) pour la méthode géométrique (b) pour la méthode d'apparence

thode d'apparence et un SVM gaussien pour les descripteurs géométriques extraits de chaque zone. Les paramètres γ et C du SVM gaussien sont adaptés à chaque région.

Zones	Méthode d'apparence	Méthode géométrique
Zone du front	62.8	85.4
Zone des yeux	79.9	52.6
Zone du nez	79.1	63
Zone de la bouche	85.7	95.3

TABLE 4.8 – Taux de reconnaissance des émotions par zone du visage selon la méthode d'apparence et la méthode géométrique dans la base CK+

Nous remarquons dans le cas des émotions simulées que la zone de la bouche est la zone la plus significative en terme de reconnaissance des émotions. En effet en se basant seulement sur les dix distances codant les déformations de la bouche, la méthode géométrique atteint un taux de reconnaissance de 95.3%. La méthode d'apparence atteint quant à elle un taux de reconnaissance de 85.7% lorsque seul le rectangle englobant la bouche est considéré. Les taux de reconnaissance calculés sur la zone de la bouche par les deux méthodes sont proches des taux de reconnaissance obtenus sur la totalité du visage, notamment pour la méthode géométrique puisque son taux de reconnaissance sur la totalité du visage est supérieur de seulement 1% par rapport au taux de reconnaissance calculé sur la zone de la bouche. La zone du front est la seconde zone la plus représentative des émotions pour la méthode géométrique. Caractérisée par les sept distances codant les mouvements des sourcils, elle atteint un taux de reconnaissance de 85.4%. La comparaison entre la matrice de confusion de la méthode géométrique calculée sur la zone de

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	92	0	4	0	0	4	0
Colère	0	90	0	0	10	0	0
Peur	0	0	100	0	0	0	0
Dégoût	0	0	2	94	0	2	2
Tristesse	0	0	0	2.2	93.7	0	4
Surprise	0	0	0	0	0	100	0
Neutre	0	0	0	0	2.2	0	97.7

TABLE 4.9 – Matrice de confusion calculée par la méthode géométrique sur la zone de la bouche dans la base CK+

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	75	0	2	4.4	0	0	18.4
Colère	0	89.7	2	6.2	2	0	0
Peur	0	0	77.3	2	10.4	10	0
Dégoût	2	10.6	2	83.3	0	2	0
Tristesse	0	0	10.6	0	89.3	0	0
Surprise	0	0	4	0	4.4	91.5	0
Neutre	4	0	0	0	2	2	92

TABLE 4.10 – Matrice de confusion calculée par la méthode géométrique sur la zone du front dans la base CK+

la bouche (voir la table 4.9) et celle calculée sur la zone du front (voir la table 4.10) montre que la zone de la bouche a des taux de reconnaissance supérieurs aux taux de reconnaissance sur la zone du front pour toutes les émotions. Ainsi se pose la question de n'utiliser que la variation des distances présentant la bouche, sans lui ajouter les sept distances présentant les mouvements des sourcils. De plus, l'état des yeux n'apporte pas beaucoup d'informations pour la reconnaissance des émotions calculées par la méthode géométrique. En effet, le taux de reconnaissance des émotions calculé par la méthode géométrique sur la zone des yeux est égal à 52.6%. De même pour la zone du nez, la méthode géométrique atteint seulement le taux de 63%. En ce qui concerne la méthode d'apparence, les textures présentes dans la zone du front ne sont pas assez significatives pour toutes les émotions. Le taux de reconnaissance moyen de la méthode d'apparence calculé sur la zone $Z1$ ne dépasse pas 63%. Par contre, nous notons que les rides et les textures qui apparaissent au niveau des yeux et au niveau du nez sont importantes pour la reconnaissance des émotions simulées. En effet, les taux de reconnaissance des émotions calculés par la méthode d'apparence dans la zone des yeux et la zone du nez dépassent 79%. D'après les tables 4.11 et 4.12 présentant les matrices de

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	83.3	0	0	0	4	0	12.6
Colère	0	81.5	4	0	10.4	2	2
Peur	0	2	73.5	6.2	4	4	10.2
Dégoût	2	6.4	4	87.5	0	0	0
Tristesse	0	0	0	0	93.3	2.2	4.4
Surprise	0	0	2	2	2	87.7	6.2
Neutre	100	4.2	8.4	0	14.4	10.2	52.6

TABLE 4.11 – Matrice de confusion calculée par la méthode d'apparence sur la zone des yeux dans la base CK+

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	87.5	4	0	0	0	2	6.4
Colère	0	82	0	0	10	4	4
Peur	2	0	79.5	4.2	2	2	10
Dégoût	2	4	2	88	0	0	4
Tristesse	0	0	0	2	94	0	4
Surprise	0	2.2	0	0	2	83.1	12.6
Neutre	6.4	6	4.2	4	16.6	22.9	39.8

TABLE 4.12 – Matrice de confusion calculée par la méthode d'apparence sur la zone du nez dans la base CK+

confusion de la méthode d'apparence respectivement sur la zone des yeux et sur la zone du nez, les taux de reconnaissance des émotions joie, colère, dégoût, tristesse et surprise dépassent 81%. Cependant, les taux de reconnaissance de l'expression neutre restent assez faibles avec 52% pour une analyse menée sur la zone des yeux et 39.8% pour une analyse menée sur la zone du nez. Nous pouvons ainsi considérer une seule de ces deux zones pour la reconnaissance des émotions puisqu'elles fournissent des taux de reconnaissance assez proches pour les mêmes émotions. Cependant, la comparaison des taux de reconnaissance obtenus sur la zone des yeux (voir la table 4.11) et sur la zone de la bouche (voir la table 4.13) montre que la colère et le dégoût sont mieux détectés sur la zone des yeux et la joie, la peur et la surprise ont des taux plus élevés sur la zone de la bouche. Nous pensons alors que ces deux zones peuvent être complémentaires pour la reconnaissance des émotions avec la méthode d'apparence.

La table 4.14 présente les moyennes des taux de reconnaissance des émotions calculées par la méthode d'apparence et la méthode géométrique sur les quatre zones du visage dans la base FEEDTUM. Nous remarquons que le taux de reconnais-

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	100	0	0	0	0	0	0
Colère	0	73.1	2	6	0	0	18.9
Peur	0	0	93.8	4.2	2	0	0
Dégoût	0	4.2	2	85.8	4	0	4
Tristesse	0	0	0	0	93.3	0	6.7
Surprise	0	4	0	0	0	96	0
Neutre	0	10.4	2	16.6	12.4	0	58.4

TABLE 4.13 – Matrice de confusion calculée par la méthode d'apparence sur la zone de la bouche dans la base CK+

Zones	Méthode d'apparence	Méthode géométrique
Zone du front	50	62.8
Zone des yeux	68.7	31.5
Zone du nez	55.75	42.4
Zone de la bouche	68	77.5

TABLE 4.14 – Taux de reconnaissance des émotions par zone du visage selon la méthode d'apparence et la méthode géométrique dans la base FEEDTUM

sance des émotions calculé par la méthode géométrique sur la zone de la bouche dépasse de 12.5% le taux de reconnaissance des émotions calculé par la même méthode sur tout le visage. D'après les tables 4.15 et 4.16 présentant respectivement, pour la méthode géométrique, la matrice de confusion calculée sur tout le visage et la matrice de confusion calculée seulement sur la zone de la bouche, il n'y a pas de confusion entre la joie et la colère lorsqu'on considère seulement la région de la bouche et elle atteint 4.2% en considérant la totalité du visage. De plus, les taux de confusion entre la joie et la peur, la joie et le dégoût et la joie et la surprise sont inférieurs de 2% à ceux calculés sur la totalité du visage. Nous notons également que les taux de reconnaissance du dégoût et de l'expression neutre calculés sur la zone de la bouche dépassent respectivement de 14% et de 8.3% leurs homologues calculés sur la totalité du visage.

Pour la méthode d'apparence, les meilleurs taux de reconnaissance sont de 68.7% sur la zone des yeux et 68% sur la zone de la bouche (voir la table 4.14). L'utilisation d'une zone locale dans la reconnaissance des émotions pour la méthode d'apparence donne donc des taux de reconnaissance inférieurs à ceux obtenus avec la totalité du visage. La comparaison des matrices de confusion calculées avec la méthode d'apparence sur la zone des yeux (voir la table 4.17) et sur la zone de la bouche (voir la table 4.18) montrent que la zone des yeux a des taux de reconnais-

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	81.3	4.2	2	2.2	0	6	4.2
Colère	0	43.7	4.4	24.6	21.1	0	6
Peur	6.4	14.9	52.8	4.4	2.2	16.9	2.2
Dégoût	6	14.4	4	57.1	6.2	8	4.2
Tristesse	4.2	14.2	4.4	6.4	52.2	2	16.4
Surprise	2	2.2	12.9	0	0	82.8	0
Neutre	0	0	0	0	14.2	0	85.7

TABLE 4.15 – Matrice de confusion de la méthode géométrique calculée sur la totalité du visage dans la base FEEDTUM

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	94	0	2	0	0	4	0
Colère	0	66.6	4.2	10.2	8.4	0	10.4
Peur	4.4	4.4	67.8	10.6	2	6.2	4.4
Dégoût	4	8.2	0	71.1	2	4.2	10.4
Tristesse	0	10.2	6	6.4	66.6	0	10.4
Surprise	0	2.2	13.1	0	2.2	82.4	0
Neutre	0	0	0	0	6	0	94

TABLE 4.16 – Matrice de confusion de la méthode géométrique calculée sur la zone de la bouche dans la base FEEDTUM

sance plus élevés que ceux obtenus sur la zone de la bouche pour la colère et la peur, respectivement de 6.2% et 10.8%. Par contre, les taux reconnaissance de la joie et la surprise calculés sur la zone de la bouche sont plus élevés de 3.8% et de 10.2% que ceux obtenus sur la zone des yeux. Ces résultats confirment les résultats obtenus sur la base CK+ : ces deux zones se complètent pour la reconnaissance des émotions.

D'après les résultats de reconnaissance des émotions des deux bases sur les quatre

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	81.5	6	4.2	0	2.2	6	0
Colère	2.2	66.4	2	18.9	6.2	0	4.2
Peur	0	4.4	65.7	0	2	21.3	6.4
Dégoût	6.2	28.4	4	55.1	6.2	0	0
Tristesse	4.2	0	2	10.2	62.8	4	16.6
Surprise	2	0	12.6	2.2	2	68.4	12.6
Neutre	4.2	0	2	0	8	4.2	81.5

TABLE 4.17 – Matrice de confusion de la méthode d'apparence calculée sur la zone des yeux dans la base FEEDTUM

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	85.3	4.2	2	8.4	0	0	0
Colère	2	60.2	4	4.4	14.6	0	14.6
Peur	0	13.1	54.9	8.4	6.4	10.6	6.4
Dégoût	6	16	6.2	57.1	4.2	4.2	6.2
Tristesse	0	16.4	6	2	60.6	0	14.9
Surprise	0	0	10.7	10.6	0	78.6	0
Neutre	2	10	2.2	0	6.4	0	79.3

TABLE 4.18 – Matrice de confusion de la méthode d'apparence calculée sur la zone de la bouche dans la base FEEDTUM

zones, la région de la bouche représente la meilleure région pour la reconnaissance des émotions calculée par la méthode géométrique. Pour la méthode d'apparence, les taux de reconnaissance obtenus sur la région des yeux et la région de la bouche sont meilleurs que ceux obtenus sur les régions restantes. Nous avons remarqué également que ces deux régions sont complémentaires.

4.6.2 Méthodes de reconnaissance des émotions par régions

Après l'évaluation de la reconnaissance des émotions sur les zones du visage, nous testons deux approches de reconnaissance basées sur le principe de fusion entre la méthode géométrique et la méthode d'apparence, appliquées sur des zones du visage. Suivant les résultats obtenus précédemment, nous proposons deux méthodes de reconnaissance :

- La première présente une fusion en aval entre la méthode d'apparence et la méthode géométrique par un SVM gaussien. Les descripteurs d'apparence sont extraits à partir de la région des yeux et de la région de la bouche. La fusion entre ces descripteurs est réalisée par la moyenne. Les descripteurs géométriques quant à eux sont extraits de la région de la bouche et classés par un SVM gaussien. La figure 4.7 présente cette première méthode.
- La deuxième fusion est une fusion entre les résultats de la méthode d'apparence appliquée sur la totalité du visage et les résultats de la méthode géométrique appliquée sur la région de la bouche.

La table 4.19 présente les taux de reconnaissance des émotions des deux méthodes citées ci-dessus dans les deux bases (CK+ et FEEDTUM). Nous remarquons, dans

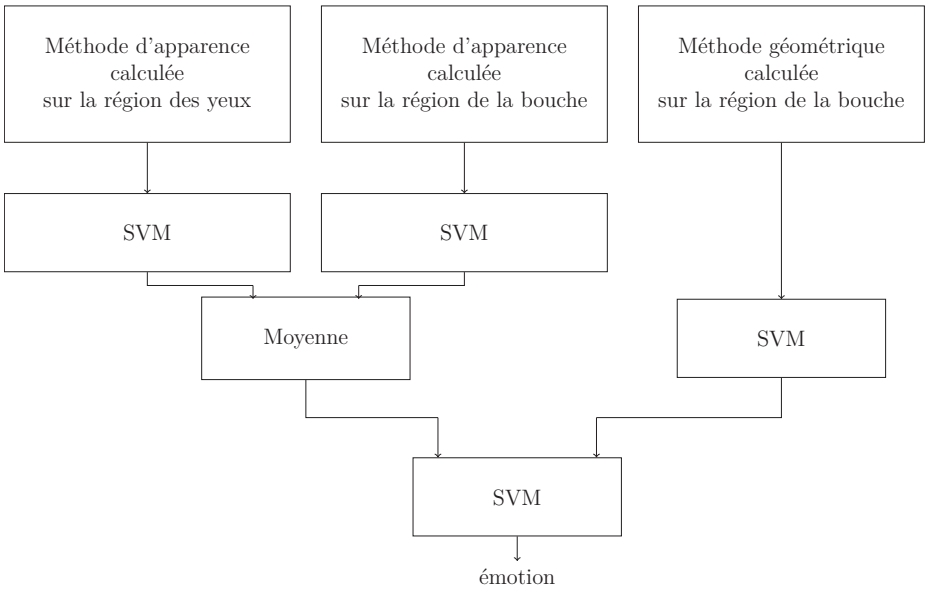


FIGURE 4.7 – Méthode de reconnaissance des émotions par zones du visage. Fusion entre la méthode géométrique appliquée sur la zone de la bouche et la méthode d'apparence appliquée sur la zone des yeux et la zone de la bouche

la base des émotions simulées (la base CK+), que la méthode utilisant les descripteurs géométriques de la bouche et les descripteurs d'apparence de la totalité du visage a un taux de reconnaissance supérieur de 0.6% au taux de reconnaissance de la méthode fusionnant seulement la géométrie de la bouche et l'apparence des régions de la bouche et des yeux. Les résultats obtenus par les deux méthodes utilisant les zones du visage sont malgré tout légèrement inférieurs à ceux obtenus par les méthodes de fusion appliquées sur la totalité du visage (voir la table 4.3). Dans la base des émotions spontanées (la base FEEDTUM), nous notons que la méthode fusionnant les résultats de l'apparence sur la totalité du visage et les résultats de la géométrie de la bouche a un taux de reconnaissance supérieur de 6.2% à celui de la méthode de fusion considérant seulement deux zones du visage pour la méthode d'apparence. Elle dépasse également les taux de reconnaissance des méthodes de fusion appliquées sur la totalité du visage présentées sur la table 4.5. D'après la matrice de confusion de la méthode fusionnant la géométrie de la bouche et l'apparence sur la totalité du visage dans la base FEEDTUM, présentée dans la table 4.20, nous remarquons que cette méthode améliore les taux de reconnaissance du dégoût, de la tristesse et de la surprise par rapport à ceux obtenus avec les méthodes de fusion présentées dans la table 4.5.

La méthode de fusion basée sur la géométrie de la bouche et l'apparence du visage

	Taux de reconnaissance dans la base CK+	Taux de reconnaissance dans la base FEEDTUM
Méthode géométrique calculée sur la région de la bouche et méthode d'apparence calculée sur les régions des yeux et de la bouche	97%	82%
Méthode géométrique calculée sur la région de la bouche et méthode d'apparence calculée sur la totalité du visage	97.6%	88.2%

TABLE 4.19 – Taux de reconnaissance des méthodes de reconnaissance des émotions par zones du visage sur les bases CK+ et FEEDTUM

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	98	0	0	2	0	0	0
Colère	0	85.1	0	8.4	6.4	0	0
Peur	2.2	4.4	78	2.2	2.2	6.6	4.2
Dégoût	2.2	6	0	90	2	0	0
Tristesse	0	4	2	0	83.5	0	10.4
Surprise	0	0	4.4	0	2.2	93.3	0
Neutre	0	0	0	0	10.2	0	89.8

TABLE 4.20 – Matrice de confusion de la méthode de fusion entre la géométrie de la bouche et l'apparence de la totalité du visage dans la base FEEDTUM

présente de bons résultats notamment pour les émotions spontanées. Cependant, la grande taille des descripteurs d'apparence reste un inconvénient pour une reconnaissance des émotions dans une séquence d'images. Nous pensons également que les descripteurs d'apparence extraits de la totalité du visage comportent de la redondance et du bruit. Une sélection de ces descripteurs pourrait permettre d'améliorer davantage les résultats de reconnaissance des émotions simulées et spontanées.

4.7 Sélection des descripteurs d'apparence

La grande taille des vecteurs descripteurs d'apparence entraine plusieurs inconvénients en termes de temps de calcul et la taille mémoire mobilisée. Les descripteurs

d'apparence peuvent également contenir des redondances et des informations qui apportent la confusion entre plusieurs émotions.
Nous procédons dans cette section à la sélection et la réduction du nombre de descripteurs de la méthode d'apparence par plusieurs approches. Nous appliquons tout d'abord, de façon naïve, des réductions de la taille des images après convolution par les filtres de Gabor. Une combinaison entre l'analyse des composantes principales et un seuillage des coefficients des descripteurs participant aux composantes choisies est ensuite réalisée. Enfin, une troisième méthode basée sur le calcul de l'importance des descripteurs par les forêts aléatoires est également appliquée pour la sélection des descripteurs les plus importants.

4.7.1 Réduction naïve

La méthode d'apparence commence par la normalisation du visage détecté à une taille de 60×80 en se basant sur la position des yeux. Une convolution avec 40 filtres est ensuite appliquée, suivie d'une étape de réduction pour diminuer le nombre de descripteurs. Cette dernière est basée sur la réduction de la taille des résultats de convolution. La figure 4.8 présente les étapes de la méthode d'apparence avec réduction des descripteurs.
Nous appelons cette technique *Naïve* parce que la réduction se fait sans l'utilisation d'un critère de choix entre descripteurs ; la réduction des images filtrées se base sur l'interpolation.
Dans le cas du changement de taille d'une image, l'interpolation cherche à estimer les valeurs des pixels à partir des pixels du voisinage. Plusieurs types d'interpolation existent, les plus utilisés sont :

- Interpolation du plus proche voisin : c'est une technique permettant d'estimer la valeur du point à interpoler à partir du pixel le plus proche. Ainsi, il n'y a pas de nouvelles valeurs dans l'image. Il y a cependant toujours le risque de perdre des informations importantes avec cette technique, puisque certains pixels pourront être solliciter plusieurs fois tandis que d'autres seront ignorés. Cette technique peut ainsi introduire de la redondance d'information.
- Interpolation bilinéaire : elle permet d'estimer la valeur du point à interpoler à partir des quatre pixels les plus proches.
- Interpolation bicubique : c'est une technique qui calcule le point à inter-

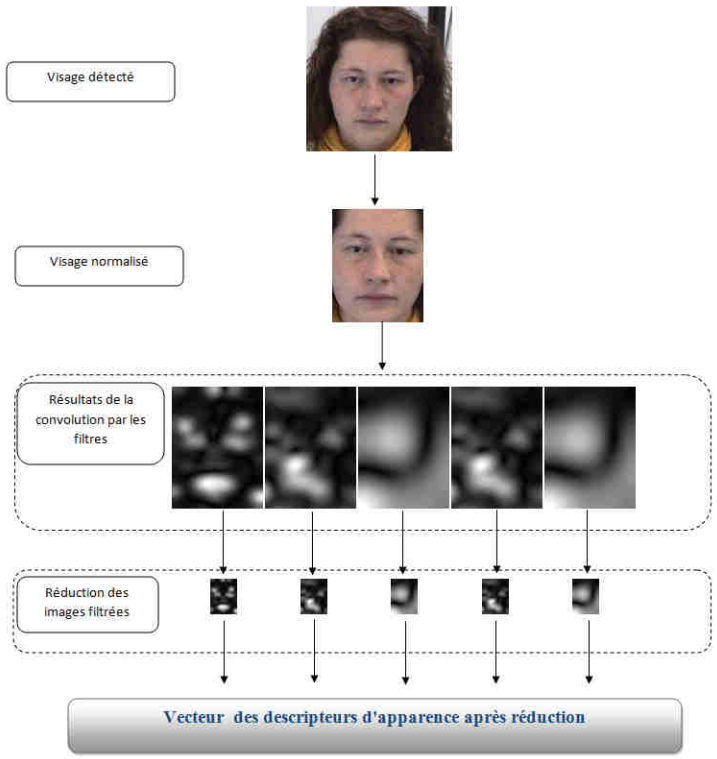


FIGURE 4.8 – Les étapes de la méthode d'extraction d'apparence du visage avec réduction des résultats de convolution

poler à partir des 16 points les plus proches, de façon à donner plus de pondération aux pixels les plus proches. L'interpolation bicubique peut être accomplie en utilisant soit des polynômes de Lagrange, des splines cubiques, ou l'algorithme de convolution cubique.

La table 4.21 présente l'influence du changement de la méthode d'interpolation sur le taux de reconnaissance après réduction de la taille des images filtrées à 15×20. Le vecteur des descripteurs d'apparence est alors réduit à 12000 (soit $15 \times 20 \times 40$) descripteurs.

%	CK+	FEEDTUM
Interpolation par le plus proche voisin	92.2	84.4
Interpolation bilinéaire	92.5	84.4
Interpolation bicubique	**92.5**	**84.7**
Sans interpolation	92.5	84.7

TABLE 4.21 – Influence de la technique d'interpolation utilisée dans la réduction du nombre de descripteurs sur le taux de reconnaissance des émotions

Nous remarquons, pour les deux bases de données considérées, que l'interpolation bicubique donne de meilleurs taux de reconnaissance que l'interpolation par le plus

proche voisin et l'interpolation bilinéaire. Ainsi, pour seulement 12000 descripteurs, nous gardons les mêmes taux de reconnaissance pour les émotions simulées et les émotions spontanées que ceux obtenus avant réduction des descripteurs.
D'autres tailles de réduction sont également testées à savoir 30×40 et 7×10 qui correspondent respectivement à des vecteurs d'apparence de 48000 et 2800 descripteurs. Les taux de reconnaissance des émotions posées et des émotions spontanées, obtenus pour diverses tailles des vecteurs d'apparence, sont présentés dans la table 4.22.

%	CK+	FEEDTUM
48000 descripteurs	92.5	84.7
12000 descripteurs	92.5	84.7
2800 descripteurs	93.1	84.1

TABLE 4.22 – Taux de reconnaissance de la méthode d'apparence après réduction des descripteurs par interpolation bicubique dans les bases CK+ et FEEDTUM

Dans le cas des émotions posées, la réduction du vecteur à 2800 descripteurs augmente le taux de reconnaissance de la méthode d'apparence de 0.6%. En revanche, la réduction des descripteurs à 48000 ou à 12000 descripteurs conduit au même taux de reconnaissance que celui qui avait été obtenu sans réduction. La reconnaissance des émotions posées peut donc être réaliser avec un nombre limité de descripteurs. La matrice de confusion de la méthode d'apparence calculée sur la base CK+, après une réduction des descripteurs de 192000 à 2800 descripteurs, montre une augmentation de 4% pour la reconnaissance de la colère (voir table 4.23). Les descripteurs éliminés diminuent ainsi la confusion entre la colère et l'expression neutre.

Dans le cas des émotions spontanées, la réduction à 48000 descripteurs et la réduction à 12000 descripteurs conservent le même taux de reconnaissance moyen que celui obtenu sans réduction, tandis que la réduction à 2800 descripteurs diminue le taux de reconnaissance de 0.6%. La matrice de confusion de la méthode d'apparence après réduction de 2800 descripteurs est présentée dans la table 4.24. Nous remarquons que le taux de reconnaissance de la colère augmente de 4% et que ses taux de confusion avec le dégoût et la tristesse diminuent par rapport à leurs équivalents avant la réduction (voir la table 3.9). D'autre part, les taux de reconnaissance de la peur, de la surprise et de l'expression neutre diminuent. La réduction des descripteurs à 2800 augmente la confusion entre la peur et l'expression neutre, de 6.4% avant la réduction, à 8.4%. La confusion entre la surprise et

%	Joi	Col	Peu	Dég	Tri	Surp	Neu
Joi	100	0	0	0	0	0	0
Col	0	**94**	0	0	0	0	6
Peu	4	0	85.7	2.2	4	0	4
Dég	0	0	4	96	0	0	0
Tri	0	0	0	0	100	0	0
Surp	0	0	0	0	4	92	4
Neu	0	6	0	2	8	0	84

TABLE 4.23 – Matrice de confusion de la méthode d'apparence après réduction des descripteurs à 2800 descripteurs dans la base CK+

%	Joi	Col	Peu	Dég	Tri	Surp	Neu
Joi	98	2	0	0	0	0	0
Col	0	**89.1**	2.2	2.2	6.4	0	0
Peu	4.4	0	***74***	2	2.2	8.8	8.4
Dég	4	6.2	0	83.7	4	0	2
Tri	0	10	2	0	73.3	0	14.6
Surp	0	0	14.6	0	0	***85.3***	0
Neu	0	0	0	0	14.4	0	***85.5***

TABLE 4.24 – Matrice de confusion de la méthode d'apparence après réduction des descripteurs à 2800 descripteurs dans la base FEEDTUM

la peur augmente également de 4% et la confusion de l'expression neutre avec la tristesse augmente de 4.2%.
Pour les émotions spontanées, la réduction des descripteurs à 2800 élimine donc des descripteurs qui caractérisent des émotions comme la surprise, la peur et l'expression neutre. La colère quant à elle semble ne nécessiter qu'un nombre réduit de descripteurs pour la définir.

La réduction Naïve est une méthode simple. Elle permet d'arriver à des taux importants de réduction des descripteurs tout en gardant le même taux de reconnaissance pour la méthode d'apparence. L'utilisation de l'interpolation bicubique permet ainsi de résumer les informations dans un vecteur plus petit, puisqu'elle pondère les valeurs des 16 pixels voisins.
Nous choisissons pour la suite de notre étude de conserver 12000 descripteurs afin de conserver des taux de reconnaissance équivalents aux cas sans sélection dans toutes les configurations.

4.7.2 Sélection par analyse des composantes principales seuillées

L'analyse des composantes principales (ACP) est une technique non supervisée permettant la transformation des variables dans un espace plus réduit. Ce dernier est représenté par des variables décorrélées appelées "composantes principales". Celles-ci correspondent aux vecteurs propres de la matrice de corrélation. Elles sont constituées des combinaisons linéaires des variables d'origine. Le sous-espace des composantes principales choisies doit retenir la variance nécessaire et suffisante à la représentation des données.

Dans cette section, nous utilisons l'ACP pour réduire les descripteurs d'apparence en utilisant un seuil sur les pondérations de ceux-ci au sein des composantes principales. Cette technique a été utilisée dans [138] pour sélectionner des descripteurs acoustiques pour la reconnaissance des émotions dans un signal vocal.

Une analyse par ACP est d'abord appliquée pour avoir un nouvel ensemble de variables décorrélées. Nous appliquons par la suite un seuil sur les pondérations des descripteurs constituant les composantes principales. Notre but dans cette section est de sélectionner les descripteurs qui ont plus de poids dans les composantes principales.

La table 4.25 montre un exemple de sélection des descripteurs en se basant sur les composantes principales sélectionnées au préalable $\{P_1, \ldots, P_m\}$. Pour un seuil S choisi, nous gardons les descripteurs $\{D_i\}$ ayant au moins une pondération (w_{ij}) au dessus du seuil. Dans l'exemple, nous supposons que $w_{11} > S$, $w_{1m} > S$ et $w_{32} > S$, les descripteurs sélectionnés seront alors D_1 et D_3.

	P_1	P_2	$\cdots$	P_m
D_1	w_{11}	w_{12}	$\cdots$	w_{1m}
D_2	w_{21}	w_{22}	$\cdots$	w_{2m}
D_3	w_{31}	w_{32}	$\cdots$	w_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
D_n	w_{n1}	w_{n2}	$\cdots$	w_{nm}

TABLE 4.25 – Un exemple pour l'application du seuil sur les pondérations des descripteurs constituant les composantes principales

Choix de la variance et du seuil des descripteurs

Les composantes principales construites par l'ACP sont ordonnées de façon décroissante en terme de représentation de variance, autrement dit var(P_1)>var(P_2)>...>var(P_{12000}). Une bonne représentation de la variance des données induit un grand nombre de composantes principales sélectionnées. Nous cherchons alors à trouver un compromis entre la représentation de la variance des données et le nombre de composantes principales prises en compte, un bon nombre de composantes étant défini par son amélioration de la reconnaissance des émotions. Chaque composante principale présente un sous-espace de descripteurs d'apparence. Le but de cette section est de trouver, sur un échantillon des données, le meilleur nombre de composantes principales et le meilleur seuil des descripteurs.
Différents taux de variances ont été testés pour la sélection des composantes principales à savoir 85%, 90%, 95% et 98%. Pour chaque ensemble de composantes choisies, une plage de seuils sur les pondérations des descripteurs d'apparence est également testée, de -0.02 à 0.04 avec un pas de 0.005.

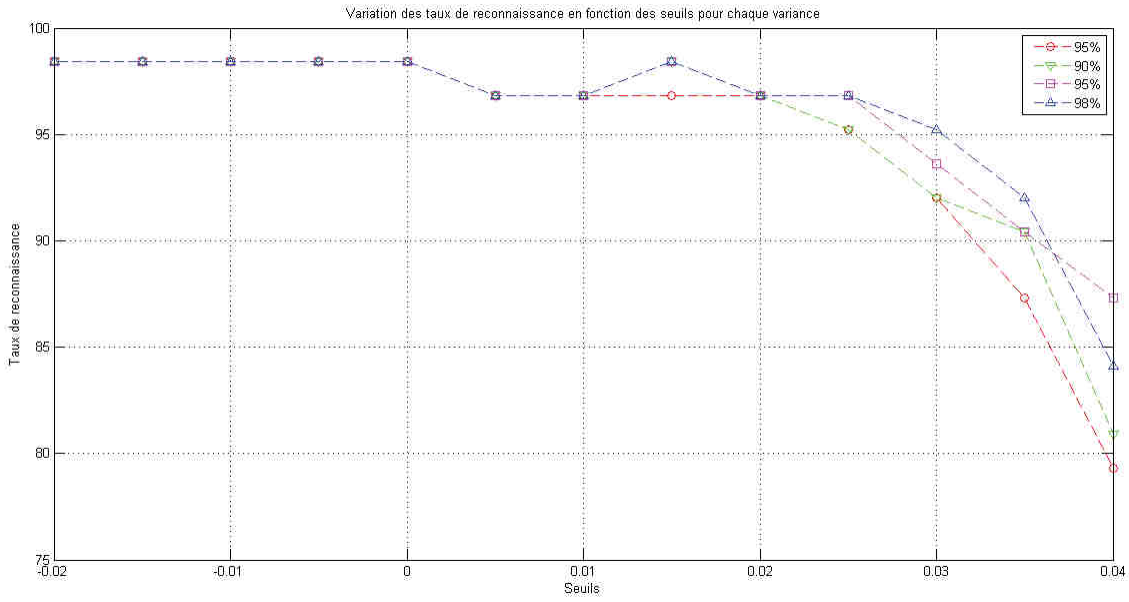


FIGURE 4.9 – Variation des taux de reconnaissance selon les seuils des descripteurs et le pourcentage de la variance gardé dans la base CK+

Dans le cas des émotions posées, la variation du taux de reconnaissance en fonction des seuils est représentée dans la figure 4.9 pour chaque pourcentage de la variance. Nous remarquons que les taux de reconnaissance pour les quatre pourcentages de la variance sont similaires jusqu'au seuil 0.01. Ainsi, le nombre de

descripteurs éliminés sur toutes les variances est minimal et n'affecte pas le taux de reconnaissance. Au delà de ce seuil, les taux de reconnaissance calculés en gardant 98% de la variance (courbe bleu), 95% de la variance (courbe rose) et 90% de la variance (courbe verte) ont des taux de reconnaissance plus élevés que les taux de reconnaissance obtenus en gardant 85% de la variance. Nous pensons alors que le nombre de descripteurs obtenus en appliquant un seuil de 85% à la variance n'est pas suffisant pour améliorer le taux de reconnaissance des émotions. Nous remarquons également que le seuil 0.015 donne le meilleur taux de reconnaissance en gardant une variance égale ou supérieure à 90%.
Dans la suite, nous étudions le comportement des taux de reconnaissance en fonction des descripteurs sélectionnés pour trois pourcentages de la variance à savoir 90%, 95% et 98%. Cette étude nous permet de choisir le seuil de la variance et le seuil des descripteurs en gardant un bon taux de reconnaissance et en réduisant le nombre de descripteurs. La figure 4.10 présente les taux de reconnaissance en fonction des ensembles de descripteurs pour les trois taux de variances cités précédemment. Les taux de reconnaissance présentés sont les taux obtenus à partir du seuil 0.01, celui-ci représentant la limite à partir de laquelle la réduction des descripteurs devient plus importante pour tous les pourcentages de la variance. Le pic du taux de reconnaissance détecté pour le seuil 0.015 dans la figure 4.9 est représenté dans la figure 4.10 pour chaque taux de pourcentage de variance. Les ensembles de descripteurs nécessaires à la détection de ce pic pour les pourcentages de variance 98%, 95% et 90% sont respectivement 10049, 10006 et 9578 descripteurs. La réduction dans ces trois ensembles de descripteurs reste faible. Les taux de reconnaissance atteignent 96.8% pour des ensembles de 5063 descripteurs pour 95% de la variance et un ensemble de 5678 descripteurs pour 98% de la variance. Cependant, 90% de la variance atteint ce même taux de reconnaissance pour 7295 descripteurs. Ainsi, le seuil de la variance le plus efficace est 95%, puisqu'il permet d'obtenir un bon taux de reconnaissance tout en réduisant l'ensemble des descripteurs. Nous testons dans la section suivante les seuils 0.015 et 0.025 pour la réduction des descripteurs, puisqu'ils donnent les meilleurs taux de reconnaissance.

Dans le cas des émotions spontanées, les variations du taux de reconnaissance en fonction des seuils sont représentées dans la figure 4.11 pour chaque pourcentage de la variance. Nous notons deux pics des taux de reconnaissance. Pour 95% et 98% de la variance représentés respectivement par la courbe rose et la courbe bleue, les

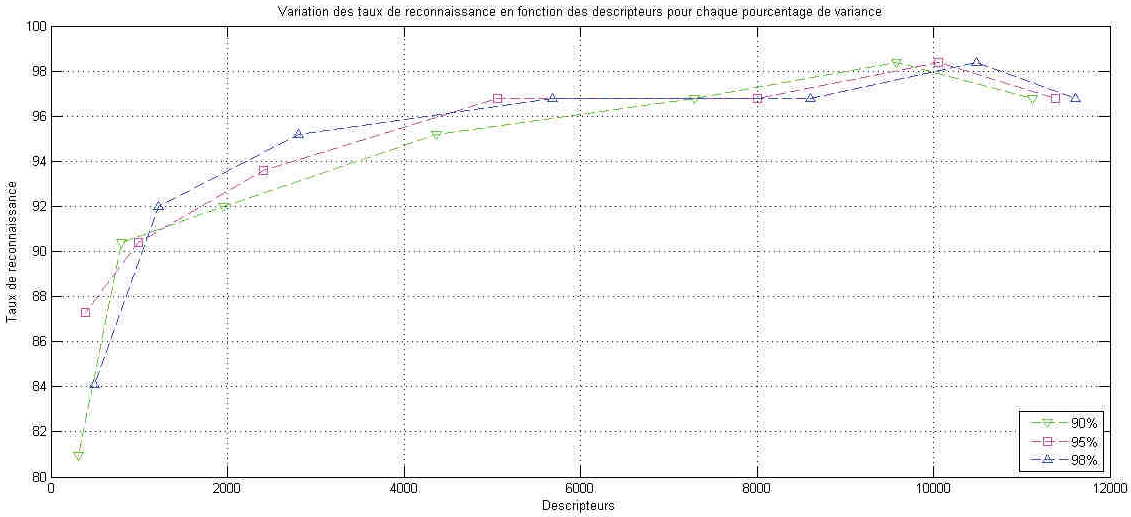


FIGURE 4.10 – Variation des taux de reconnaissance en fonction des descripteurs d'apparence sélectionnés pour 90% de la variance, 95% de la variance et 98% de la variance dans la base CK+

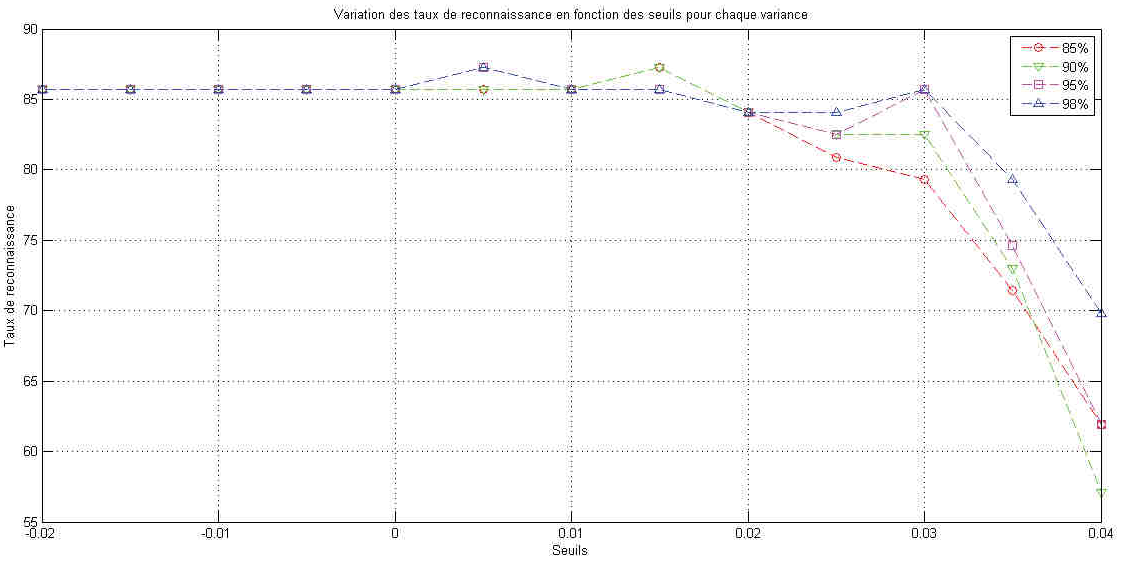


FIGURE 4.11 – Variation des taux de reconnaissance selon les seuils des descripteurs et le pourcentage de la variance gardé dans la base FEEDTUM

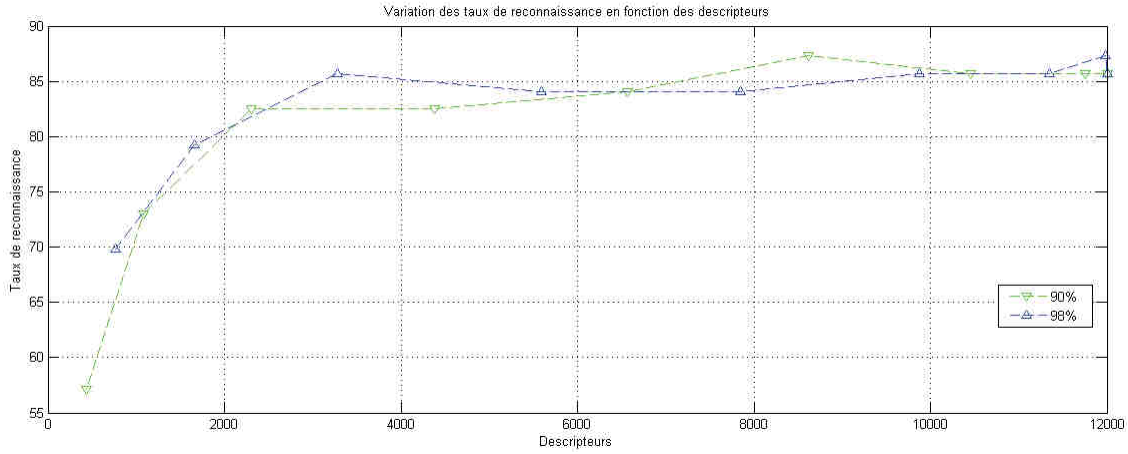


FIGURE 4.12 – Variation du taux de reconnaissance en fonction des descripteurs d'apparence sélectionnés pour 90% et 98% de la variance dans la base FEEDTUM

taux de reconnaissance atteignent 87% pour un seuil de 0.005. Pour 90% et 85% de la variance, représentés respectivement par la courbe verte et la courbe rouge, 87% de taux de reconnaissance sont atteints pour un seuil de 0.015. Ainsi, les pourcentages de variance forment deux groupes qui évoluent de façon différentes. Nous avons choisi un taux de variance de chaque groupe pour étudier l'évolution des taux de reconnaissance par rapport au nombre de descripteurs. Pour le premier groupe, c'est le cas à 98% de la variance qui a été retenu puisqu'il conduit aux meilleurs taux de reconnaissance pour les seuils supérieurs à 0.03. Dans le deuxième groupe, nous avons choisi le cas à 90% de la variance qui obtient de meilleurs taux de reconnaissance dès que le seuil dépasse 0.025.

La figure 4.12 présente la variation du taux de reconnaissance en fonction des descripteurs d'apparence sélectionnés par des seuils supérieurs à 0 pour 90% et 98% de la variance. Nous notons que la courbe à 98% de la variance atteint un taux de reconnaissance de 87% pour un ensemble de descripteurs de 11978 descripteurs. Son taux de reconnaissance se dégrade ensuite progressivement de 3%, puis remonte à 85% pour un nombre de descripteurs égal à 3277. La courbe verte correspondant à 90% de la variance atteint un taux de reconnaissance égal à 87% pour 8620 descripteurs et diminue rapidement à 82% pour un nombre de descripteurs égal à 4384.

Nous pensons alors que 98% de la variance reste assez robuste pour un ensemble de descripteurs égal à 3277 descripteurs et que 90% de la variance atteint un bon taux de reconnaissance mais pour un grand nombre de descripteurs. Nous choisissons pour le test 98% comme un seuil de la variance et un seuil de 0.03 pour les

%	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	100	0	0	0	0	0	0
Colère	0	90	0	0	0	0	10
Peur	4	0	85.7	2.2	4	0	4
Dégoût	0	0	4	96	0	0	0
Tristesse	0	0	0	0	100	0	0
Surprise	0	0	0	0	6	92	2
Neutre	0	4	0	2	8.2	2	**83.7**

TABLE 4.26 – Matrice de confusion de la méthode d'apparence après sélection des descripteurs avec la méthode ACP dans la base CK+ (95% de la variance - seuil descripteurs à 0.025)

descripteurs.

Sélection des descripteurs par ACP

La réduction des descripteurs d'apparence passe d'abord par la réduction naïve (présentée dans la section 4.7.1) puis par la sélection des descripteurs par l'ACP. Les taux de reconnaissance présentés dans cette section sont calculés par une validation croisée de cinq dossiers.
Pour la base CK+, nous avons choisi de laisser 95% la variance et nous avons testé deux seuils pour la sélection des descripteurs d'apparence, à savoir 0.015 et 0.025. Pour le seuil 0.015, nous obtenons un taux de reconnaissance égal à 92.2%, en revanche un taux de 92.48% est obtenu pour le seuil 0.025. La table 4.26 présente la matrice de confusion de la méthode d'apparence en utilisant la sélection des descripteurs par ACP avec le seuil 0.025.
La sélection des descripteurs d'apparence par ACP permet d'obtenir, avec seulement 5000 descripteurs, un taux de reconnaissance quasiment similaire à celui obtenu sans sélection. Une légère diminution de 0.3% du taux de reconnaissance de l'expression neutre peut être constatée.

Pour les émotions spontanées de la base FEEDTUM, nous avons choisi de garder 98% de la variance des données et le seuil des descripteurs est fixé à 0.03. La table 4.27 présente la matrice de confusion de la méthode d'apparence après sélection des descripteurs. Le taux de reconnaissance après sélection est égal à 82.7%. Il décroit de 2% par rapport au taux de reconnaissance avant la sélection. Nous remarquons une diminution de la reconnaissance de la joie, de la colère, du dégoût

%	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	94	0	0	6	0	0	0
Colère	0	83.1	2.3	6.2	4.2	0	4.2
Peur	4.4	2.2	76	2	2.3	11.1	2
Dégoût	8.5	10.2	0	75.3	4	0	2
Tristesse	0	6	2	0	79.6	0	12.4
Surprise	0	0	16.9	0	0	83.1	0
Neutre	0	2	0	0	10.4	0	87.6

TABLE 4.27 – Matrice de confusion de la méthode d'apparence après sélection des descripteurs avec la méthode ACP dans la base FEEDTUM (98% de la variance - seuil descripteurs à 0.03)

et de la surprise. Par contre, une augmentation de 6.3% dans la reconnaissance de la tristesse après la sélection par ACP peut être observée.

4.7.3 Sélection basée sur le calcul d'importance des descripteurs

Dans cette section, nous nous basons sur le taux d'importance des descripteurs d'apparence pour la sélection. Plusieurs méthodes ont été utilisées pour calculer les scores d'importance des variables afin de les hiérarchiser, telles que les SVMs [139] [140], les modèles linéaires généralisés (GLM) [141] et Adaboost [40]. Malgré leurs capacités à traiter des problèmes de grande densité de variables, leurs performances diminuent lorsque les variables sont corrélées. D'un autre côté la méthode des forêts aléatoires (FA) [142] a démontré sa performance dans le calcul des taux d'importance des variables et surtout sa stabilité en présence de bruit [99]. En effet, le choix aléatoire des variables pour la construction des arbres permet une sélection des variables non corrélées et assure une stabilité du calcul des scores d'importance des variables. C'est pour ces raisons que notre choix s'est porté sur les forêts aléatoires pour le calcul d'importance des descripteurs d'apparence.

Les forêts aléatoires

Les forêts aléatoires sont des ensembles d'arbres de décision binaires sur plusieurs échantillons tirés par *bootstrap* avec remise dans l'ensemble d'apprentissage N. Le *boostrapping* avec remise est une technique d'échantillonnage des données d'ap-

prentissage de façons aléatoires. La construction des nœuds des forêts aléatoires se base également sur un tirage aléatoire d'un ensemble de descripteurs, noté $mtry$, de la totalité des descripteurs p. L'aspect aléatoire dans le choix des descripteurs pour chaque nœud permet en plus de la décorrélation des nœuds, d'éviter un nombre de descripteurs plus grand que l'ensemble d'apprentissage ($p >> N$). Le paramètre $mtry$ est un paramètre d'un intérêt majeur dans la construction des forêts aléatoires, nous étudions dans la suite le choix de ce paramètre.
Comme nous l'avons déjà indiqué, l'apprentissage d'un arbre aléatoire nécessite un échantillon de l'ensemble d'apprentissage. La partie exclue de l'apprentissage (*Out-Of-Bag*) sert à estimer le taux d'erreur de l'arbre en question. Le calcul d'importance des descripteurs se base sur cette notion. Appelé calcul d'importance par permutation [142], il permute la valeur du descripteur (X_a^j) par une de ses valeurs tirées aléatoirement de l'ensemble *Out-Of-Bag* ($X_{a\_oob}^j$). Le taux d'importance d'un descripteur ($ID(X^j)$) est alors la moyenne de l'erreur calculée sur toutes les arbres de la forêt après permutation :

$$ID(X^j) = \frac{1}{nb\_arbres} \sum_a \left(|err\left(X_a^j\right) - err\left(X_{a\_oob}^j\right)| \right) \tag{4.1}$$

Sélection du paramètre $mtry$

Le paramètre $mtry$ défini précédemment comme le nombre des descripteurs participants dans l'apprentissage de chaque nœud, est un paramètre qui agit sur l'apprentissage de la forêt aléatoire. Dans cette section, nous étudions l'impact de sa variation sur la reconnaissance des émotions. Le paramètre $mtry$ varie de la valeur par défaut ($\sqrt{p}$) à la totalité des descripteurs (p). Dans ce dernier cas, la notion des forêts aléatoires devient un *Bagging*. Le paramètre $mtry$ varie alors dans l'ensemble $\left\{\sqrt{p}, \frac{p}{4}, \frac{p}{2}, \frac{3p}{4}, p\right\}$. La valeur du paramètre $mtry$ considérée est celle qui améliore le taux de de reconnaissance pour différents ensembles de descripteurs.
Les figures 4.13 et 4.14 présentent les variations des taux de reconnaissance en fonction des descripteurs sélectionnés pour chaque valeur de $mtry$, respectivement pour la base CK+ et la base FEEDTUM.

Dans la figure 4.13, nous remarquons que la courbe $mtry = \sqrt{p}$ (courbe en rose) est au dessous des autres courbes, traduisant ainsi des taux de reconnaissance inférieurs. Les courbes de $mtry = \frac{p}{2}$ (courbe verte) et $mtry = \frac{p}{4}$ (courbe bleue) ont

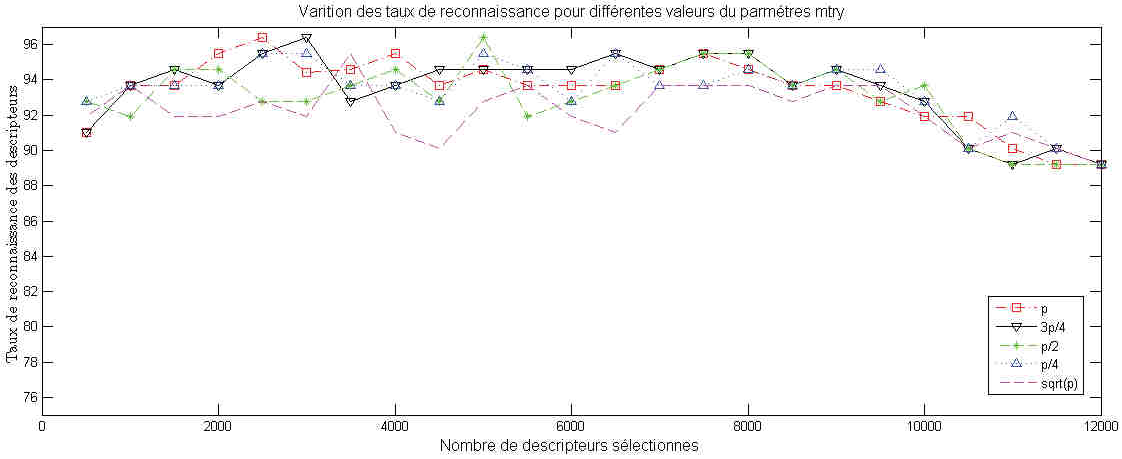


FIGURE 4.13 – Variation des taux de reconnaissance en fonction des descripteurs sélectionnés pour différentes valeurs de $mtry$ dans la base CK+

un pic au niveau de 5500 descripteurs, mais leurs taux de reconnaissance calculés pour la majorité des autres ensembles de descripteurs sont inférieurs aux taux de reconnaissance de $mtry = p$ (courbe rouge) et $mtry = \frac{3p}{4}$ (courbe noire). Pour la base CK+, nous choisissons le paramètre $mtry = \frac{3p}{4}$. Les taux de reconnaissance obtenus pour cette valeur du paramètre mtry sont supérieurs à ceux atteints pour le paramètre $mtry = p$ sur une grande plage de descripteurs allant de 4500 à 10000 descripteurs. De plus, le choix de $mtry = \frac{3p}{4}$ permet de garder la caractéristique du choix aléatoire des variables de chaque nœud, permettant ainsi de réduire le temps d'apprentissage des forêts aléatoires.

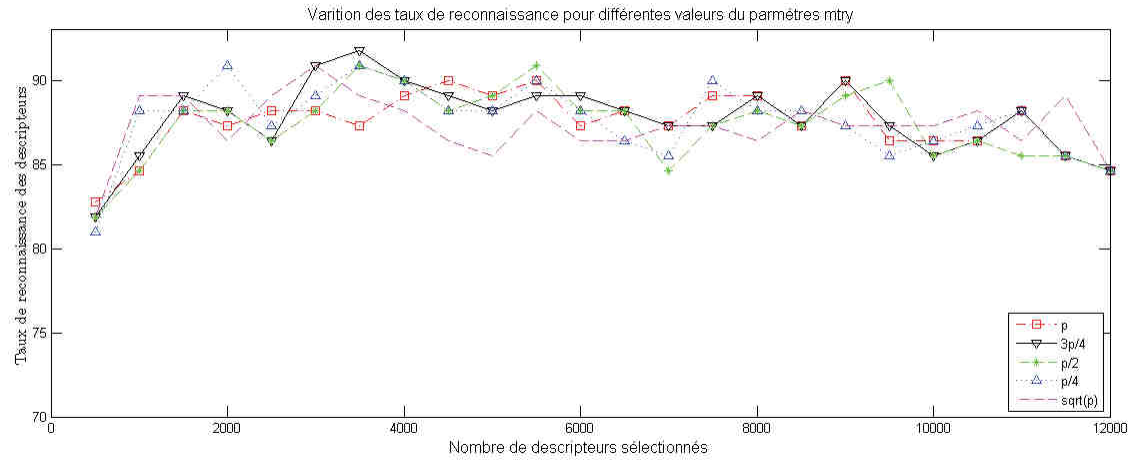


FIGURE 4.14 – Variation des taux de reconnaissance en fonction des descripteurs sélectionnés pour différentes valeurs de $mtry$ dans la base FEEDTUM

Dans la base FEEDTUM, nous notons que les courbes de $mtry = \sqrt{p}$ (courbe

rose) et $mtry = p$ (courbe rouge) ont des taux de reconnaissance inférieurs aux taux de reconnaissance des valeurs de $mtry = \frac{p}{4}$ (courbe bleue), $mtry = \frac{p}{2}$ (courbe verte) et $mtry = \frac{3p}{4}$ (courbe noire). La courbe de $mtry = \frac{p}{2}$ a deux pics aux niveaux de 5500 descripteurs et de 9500 descripteurs. D'un autre coté, la courbe de $mtry = \frac{p}{4}$ atteint d'importants taux de reconnaissance pour 2000 descripteurs et 7500 descripteurs. Ces deux courbes restent cependant inférieures à celle obtenues lors de l'utilisation de $mtry = \frac{3p}{4}$, sur une grande plage de descripteurs. Ainsi, notre choix se porte à nouveau sur $mtry = \frac{3p}{4}$ pour la base FEEDTUM.

Notre approche de sélection des descripteurs

Notre approche prend en considération deux méthodes d'apprentissage les SVMs et les forêt aléatoires (FAs). Elle est constituée principalement de trois étapes :

- Réduction itérative des descripteurs basée sur leurs calculs d'importance par les FAs.
- Sélection du nombre de descripteurs retenus en se basant sur les taux de reconnaissance par SVM et les taux d'erreur des FAs.
- Evaluation des performances de la reconnaissance des émotions.

Réduction itérative des descripteurs

Cette étape est présentée dans le schéma de la figure 4.15. Pour assurer la stabilité des résultats obtenus (les taux de reconnaissance des SVMs et les taux d'erreur des FAs), nous répétons cette étape sur plusieurs échantillons de données. Chaque échantillon est divisé en deux ensembles :

- Un ensemble d'apprentissage E_{app}.
- Un ensemble d'évaluation E_{eval}

Dans un premier temps, les forêts aléatoires sont construites sur E_{app} en utilisant le paramètre $mtry$ sélectionné précédemment. L'importance des descripteurs est ensuite calculée plusieurs fois (cinq fois) afin de renforcer la stabilité des taux d'importance des descripteurs. La figure 4.16 montre la variation du taux d'importance d'un ensemble de descripteurs. Nous remarquons qu'au cours des cinq répétitions, les descripteurs ayant des taux d'importance faibles sont peu variables. Par contre, les descripteurs ayant des taux d'importance élevés ont de larges plages de variabilité. Cette instabilité est causée par le bruit apporté par les descripteurs

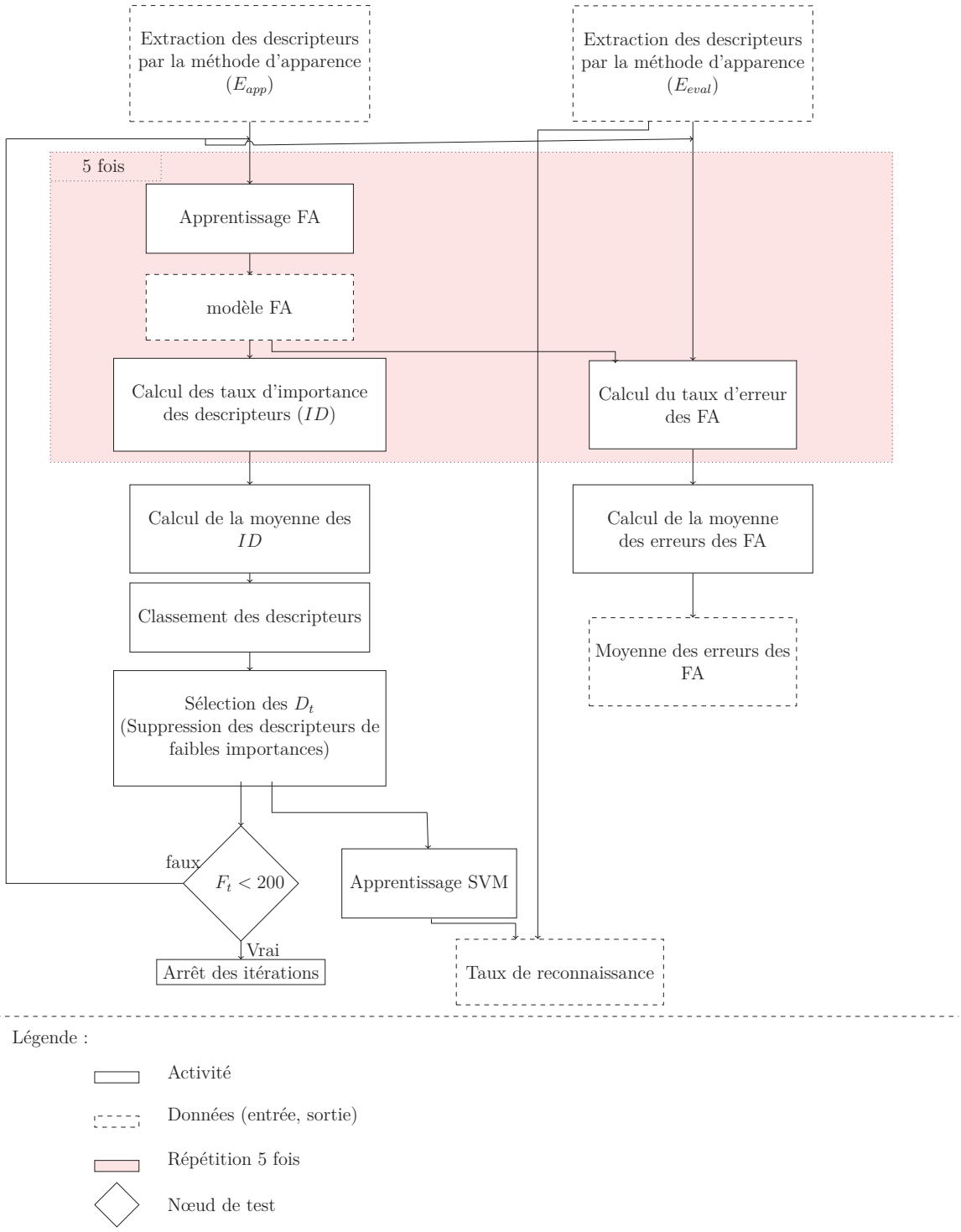


FIGURE 4.15 – Schéma de la réduction itérative des descripteurs

de faible importance. Ainsi, la moyenne des taux d'importance reste plus stable pour l'apprentissage.

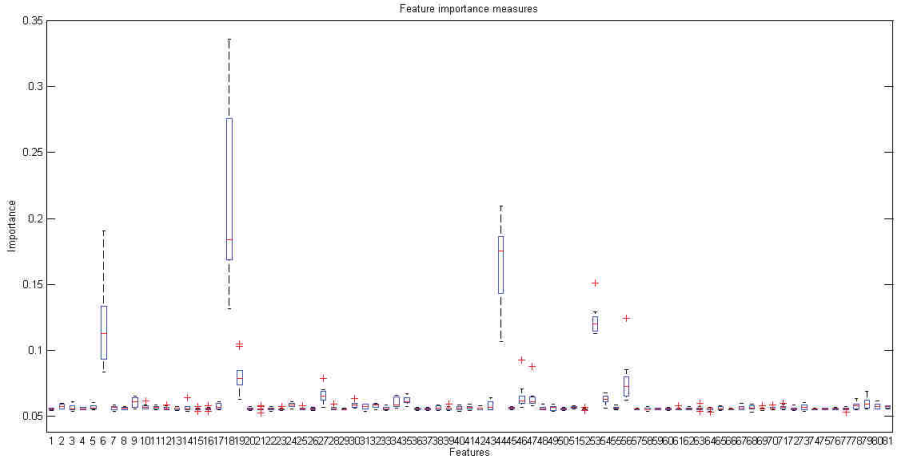


FIGURE 4.16 – Variation des taux d'importance d'un ensemble de descripteurs mesurés durant cinq répétitions

Nous classons ensuite les descripteurs selon l'ordre décroissant des moyennes de leurs taux d'importance (voir la figure 4.15) et éliminons l'ensemble des descripteurs formant un pourcentage Q du taux d'importance total. Le choix des descripteurs éliminés se fait parmi les descripteurs ayant les taux d'importance les plus faibles. Nous avons choisi dans ce cas une approche de sélection de type "*Backward*". Ce type de sélection commence par considérer la totalité des descripteurs et élimine les descripteurs les plus faibles. Puis, une nouvelle organisation des descripteurs restants est effectuée, jusqu'à satisfaction d'un critère d'arrêt. Cette approche est souvent utilisée pour la sélection dans des problèmes de grandes dimensions, contrairement à l'approche "*Forward*" qui commence par un ensemble vide de descripteurs et ajoute ensuite les descripteurs les plus importants jusqu'à la vérification du critère d'arrêt. Cette deuxième approche est plus favorisée pour les problèmes de faibles dimensions.
Nous avons testé l'élimination des descripteurs formant $Q = 40\%$, $Q = 30\%$ ou $Q = 20\%$ du taux d'importance total. Nous avons alors remarqué que pour 40%, le nombre de descripteurs éliminés est élevé, notamment après un certain nombre d'itérations l'ensemble de descripteurs restants devient très faible. D'un autre coté, pour 20% le nombre de descripteurs éliminés est faible et le nombre d'itérations augmente. Nous choisissons alors d'éliminer les descripteurs formant 30% du taux d'importance total, puisque ce taux nous permet d'éliminer un nombre raisonnable

de descripteurs à chaque itération.
Les descripteurs sélectionnés dans chaque itération (D_t) sont intégrés dans l'apprentissage des SVMs et le calcul des erreurs des forêts aléatoires (FA) sur l'ensemble E_{app}. Finalement, les taux de reconnaissance des émotions et le taux des erreurs des FA sont calculés pour chaque itération avec les D_t sélectionnés sur l'ensemble d'évaluation E_{eval}.
Cette première étape de réduction des descripteurs est réitérée jusqu'à avoir un nombre de descripteurs insuffisant à la classification des différentes classes d'émotions.

Sélection des descripteurs en se basant sur les taux de reconnaissance par SVM et les taux d'erreur des forêts aléatoires

Dans cette méthode de sélection, nous avons choisi la méthode des FAs pour le calcul des scores d'importance et nous avons gardé les SVMs pour le calcul des taux de reconnaissance des émotions. L'étude du choix du nombre optimal de descripteurs pour la sélection passe par l'étude des taux de reconnaissance de la méthode SVM et des taux d'erreur de la méthode FA. L'utilisation des deux méthodes permet d'assurer une meilleure robustesse pour le choix de l'ensemble de descripteurs. En effet, nous considérons un ensemble de descripteurs comme étant optimal s'il améliore le taux de reconnaissance des émotions calculé par SVM et minimise le taux d'erreur calculé par les FAs.
La première étape étant effectuée sur plusieurs échantillons de données, les moyennes des taux de reconnaissance et les moyennes des taux d'erreurs sont alors calculées pour chaque ensemble de descripteurs. Les figures 4.17 et 4.18 présentent les variations de ces moyennes sur les ensembles de descripteurs sélectionnés respectivement dans la base CK+ et la base FEEDTUM.
Dans la base CK+, nous remarquons que le taux de reconnaissance atteint un maximum de reconnaissance de 90% pour un ensemble de descripteurs égal à 6000 descripteurs. Puis, une diminution progressive de 2.5% est notée jusqu'à 2000 descripteurs. La diminution du taux de reconnaissance devient plus rapide en-deça de ce nombre de descripteurs (une perte de 3.4% peut être observée lors du passage de 2000 à 670 descripteurs). Le taux d'erreur calculé par les FAs reste stable à un taux de 20% pour des ensembles supérieurs à 2000 descripteurs. Une augmentation est ensuite notée pour des ensembles de descripteurs inférieurs à 2000.

Nous pensons que les ensembles inférieurs à 2000 descripteurs ne sont pas adaptés à la reconnaissance dans la base CK+, puisqu'ils diminuent à la fois le taux de reconnaissance des SVMs et augmentent le taux d'erreur des FAs. Nous testerons dans la suite deux ensembles de descripteurs. Un ensemble comportant 6000 descripteurs est d'abord testé, la courbe des taux de reconnaissance indiquant un pic pour cet ensemble. Nous testerons également un ensemble comportant 2000 descripteurs. En effet, le taux de reconnaissance pour cet ensemble reste élevé et le taux d'erreur reste stable.
Dans la base FEEDTUM, nous remarquons une augmentation progressive du taux de reconnaissance de 74.2% pour la totalité des descripteurs à 77.8% pour un ensemble de 6000 descripteurs. Le taux de reconnaissance reste ensuite stable pour les ensembles allant de 6000 à 4000 descripteurs. Puis, il commence à se dégrader jusqu'à arriver à un taux de 70% pour un ensemble de 677 descripteurs. Le taux d'erreur marque quant à lui une légère augmentation de 0.4% jusqu'à 6000 descripteurs. Puis il se stabilise dans la plage comprise entre 6000 et 4000 descripteurs. Enfin, le taux d'erreur augmente à nouveau de presque 10% pour 677 descripteurs. Dans la base FEEDTUM, la plage des ensembles de descripteurs comprenant de 4000 à 6000 descripteurs représente un bon choix, puisqu'ils ont à la fois les taux de reconnaissance les plus élevés et de faibles taux d'erreur. Nous choisissons de tester dans ce qui suit un ensemble de 4000 descripteurs.

Evaluation des performances des descripteurs sélectionnés dans la reconnaissance des émotions

Les résultats obtenus dans cette section sont issus d'une validation croisée sur cinq dossiers. Une fois le taux d'importance des descripteurs calculé par les FAs, les ensembles de descripteurs fixés pour chaque base sont sélectionnés parmi les descripteurs ayant les scores les plus importants. Un apprentissage des SVMs avec l'ensemble des descripteurs choisi est ensuite effectué. Enfin, un test est appliqué sur l'ensemble de test.
Dans la base CK+, nous avons testé deux ensembles de descripteurs à savoir 2000 et 6000 descripteurs. Les tables 4.28 et 4.29 présentent respectivement les matrices de confusion après sélection de 2000 et de 6000 descripteurs. Les taux de reconnaissance des émotions après une sélection de 2000 descripteurs et une sélec-

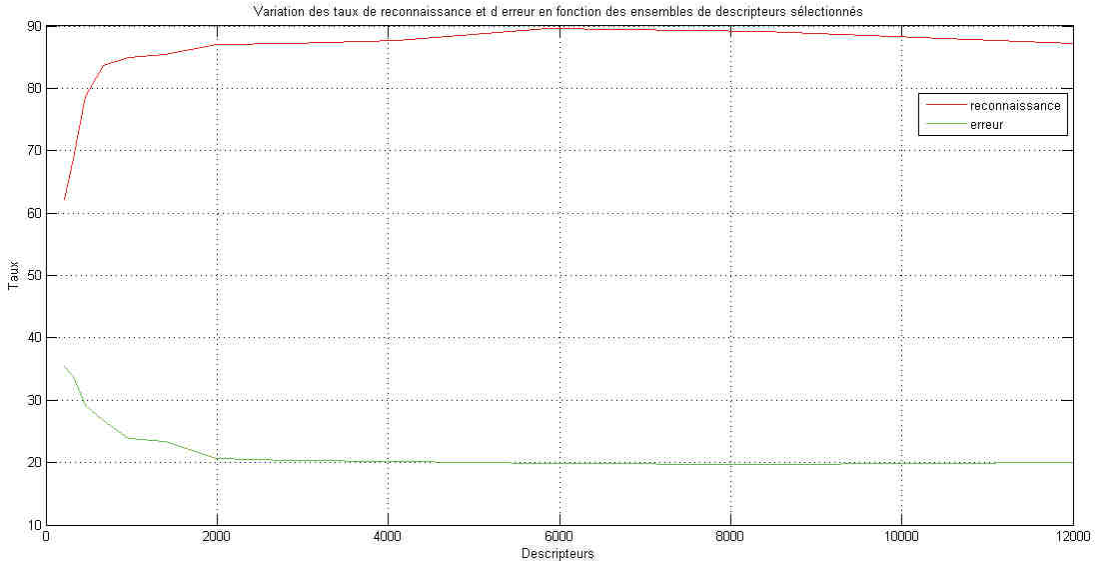


FIGURE 4.17 – Variation des taux de reconnaissance calculés par SVM et des taux d'erreur calculés par FA en fonction du nombre de descripteurs sélectionnés dans la base CK+

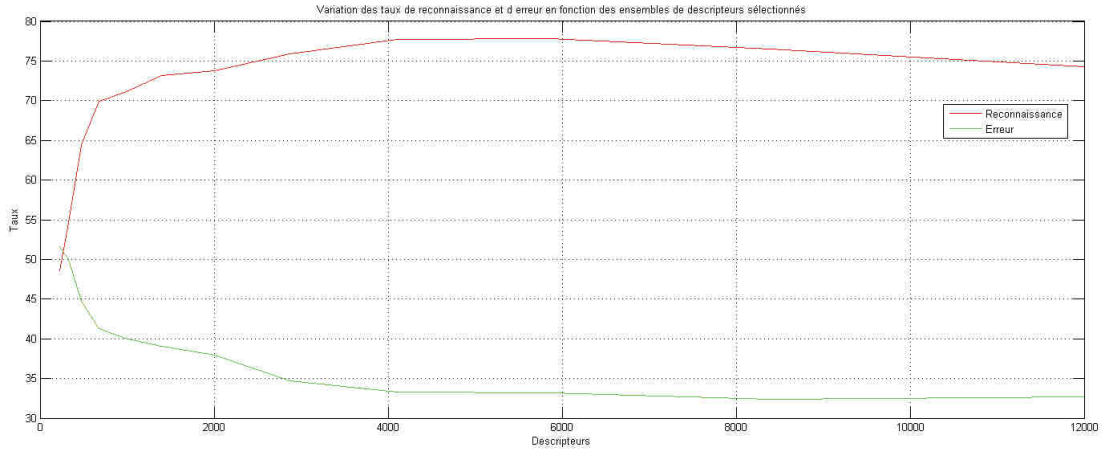


FIGURE 4.18 – Variation des taux de reconnaissance calculés par SVM et des taux d'erreur calculés par FA en fonction du nombre de descripteurs sélectionnés dans la base FEEDTUM

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	100	0	0	0	0	0	0
Colère	0	92	0	0	0	0	8
Peur	4	0	89.8	2.2	0	0	4
Dégoût	0	0	2	96	0	0	2
Tristesse	0	0	0	0	100	0	0
Surprise	0	0	0	0	2	98	0
Neutre	0	6.2	0	2	6	0	85.8

TABLE 4.28 – Matrice de confusion après sélection de 2000 descripteurs dans la base CK+

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	100	0	0	0	0	0	0
Colère	0	92	0	0	0	0	8
Peur	2	0	87.8	4.2	4	0	2
Dégoût	0	0	2	98	0	0	0
Tristesse	0	0	0	0	100	0	0
Surprise	0	0	0	0	4	96	0
Neutre	0	4	0	2	6	0	88

TABLE 4.29 – Matrice de confusion après sélection de 6000 descripteurs dans la base CK+

tion de 6000 descripteurs sont égaux à 94.5%. Notre méthode de sélection apporte ainsi une amélioration de 2% par rapport à la reconnaissance avant la sélection. Les matrices de confusion obtenues après sélection de 2000 ou de 4000 descripteurs montrent que les taux de reconnaissance de la colère, la peur, la surprise et l'expression neutre augmentent par rapport à leurs taux avant la sélection. Nous remarquons également que la peur et la surprise ont des taux de reconnaissance plus élevés avec une sélection de 2000 descripteurs qu'avec une sélection de 6000 descripteurs. Cependant, le dégoût et l'expression neutre sont mieux reconnus avec une sélection de 6000 descripteurs. Notre choix pour cette base de données se porte sur le plus petit ensemble obtenu à savoir la sélection à 2000 descripteurs.

Dans la base FEEDTUM, nous choisissons 4000 descripteurs pour représenter chaque émotion. La table 4.30 présente la matrice de confusion des émotions dans la base spontanée. Le taux de reconnaissance moyen dans la base FEEDTUM après une sélection de 4000 descripteurs est de 85.9%. Ce taux a augmenté de 1.2% par rapport au taux de reconnaissance obtenu sans sélection des descripteurs. La sélection dans ce cas permet d'améliorer la reconnaissance de la colère de 2% et de la tristesse de 6.3% par rapport à la reconnaissance sans sélection.

	Joie	Colère	Peur	Dégoût	Tristesse	Surprise	Neutre
Joie	98	2	0	0	0	0	0
Colère	0	87.1	0	4.5	6.2	0	2.2
Peur	4.4	0	76	2	0	8.7	8.9
Dégoût	4	6.2	0	83.8	4	0	2
Tristesse	0	6	2	0	79.6	0	12.4
Surprise	0	0	10.7	0	0	89.3	0
Neutre	0	4	0	0	8.4	0	87.6

TABLE 4.30 – Matrice de confusion après sélection de 4000 descripteurs dans la base FEEDTUM

Nous remarquons que les émotions spontanées nécessitent plus de descripteurs que les émotions simulées.

4.7.4 Comparaison entre les méthodes de sélection

Méthodes de sélection	Nombre de descripteurs	Taux de reconnaissance
Réduction Naïve	12000	92.5
ACP avec seuillage	5000	92.48
Approche proposée	2000	94.5

TABLE 4.31 – Taux de reconnaissance après les différentes approches de sélection sur la base CK+

Notre choix de taille des images filtrées dans la réduction naïve nous a permis de garder les mêmes taux de reconnaissance de toutes les émotions et de réduire les descripteurs de 192000 descripteurs à 12000 descripteurs. Nous avons remarqué qu'en dessous de ce nombre de descripteurs, les taux de reconnaissance sont modifiés. Nous avons ensuite appliqué cette méthode comme un pré-traitement pour faire la sélection des descripteurs. Nous avons testé une sélection par ACP en gardant les descripteurs ayant des pondérations supérieures à un seuil appliqué sur les composantes principales choisies. Dans un second temps, nous avons gardé les descripteurs ayant les meilleurs taux d'importance calculés par les FAs.

Dans la base des émotions simulées CK+ (voir la table 4.31), la sélection par une application d'un seuil sur les composantes principales permet une sélection de 5000 descripteurs. Le taux de reconnaissance obtenu après la sélection est légèrement inférieur au taux de reconnaissance avant sélection. Nous remarquons seulement une légère dégradation de 0.3% du taux de reconnaissance de l'expression neutre.

En revanche, la seconde méthode proposée permet de ne retenir que les 2000 descripteurs les plus importants. La moyenne du taux de reconnaissance augmente alors de 2%. Nous notons également une augmentation du taux de reconnaissance de la colère, la peur, la surprise et l'expression neutre par rapport à leurs taux de reconnaissance avant sélection.

Méthodes de sélection	Nombre de descripteurs	Taux de reconnaissance
Réduction Naïve	12000	84.7
ACP avec seuillage	3277	82.7
Approche proposée	4000	85.9

TABLE 4.32 – Taux de reconnaissance après les différentes approches de sélection sur la base FEEDTUM

Dans la base des émotions spontanées FEEDTUM, la sélection par utilisation d'un seuil sur les composantes principales des données diminue le taux de reconnaissance de la joie, de la colère, du dégoût, de la surprise et de l'expression neutre. En revanche, l'approche proposée, basée sur SVM et FA, permet de sélectionner 4000 descripteurs toute en améliorant le taux de reconnaissance moyen par 1.2%. Les taux de reconnaissance de la colère et de la tristesse sont également améliorés.

4.8 Conclusion

Ce chapitre propose plusieurs approches pour l'amélioration des taux de reconnaissance des émotions en se basant sur les méthodes présentées dans le chapitre 3. Dans un premier temps, nous avons amélioré les taux de reconnaissance de la méthode géométrique en substituant le SVM linéaire par un SVM gaussien. Nous avons proposé, dans un second temps, des méthodes de fusion permettant de représenter à la fois les changements au niveau de l'apparence et au niveau de la forme du visage. Deux types de fusion en aval sont appliqués à savoir des méthodes basées sur des règles statistiques et d'autres basées sur des méthodes de classification. Testées sur les deux bases, les méthodes de fusion ont des taux de reconnaissance supérieurs aux taux de reconnaissance de la méthode géométrique et de la méthode d'apparence. Trois méthodes de fusion sont alors testées avec des nouveaux sujets des deux bases. Pour les émotions simulées, les taux de reconnaissance obtenus dépassent 86%, alors que pour les émotions spontanées les taux de reconnaissance n'atteignent même pas 52%. Nous avons ensuite cherché les zones du visage les plus représentatives des émotions. Deux méthodes de reconnaissance

des émotions par région sont testées. La première fusionne la méthode géométrique appliquée sur la région de la bouche et la méthode d'apparence appliquée sur la région des yeux et sur la région de la bouche. La deuxième méthode fusionne les distances géométriques de la bouche et la méthode d'apparence appliquée sur tout le visage. Cette deuxième méthode améliore les taux de reconnaissance des émotions spontanées. Enfin les vecteurs descripteurs d'apparence extraits de la totalité du visage comportent de la redondance et du bruit. Une méthode de sélection basée sur les FAs et les SVMs est alors proposée. Elle permet d'augmenter le taux de reconnaissance dans la base CK+ de 2% et dans la base FEEDTUM de 1.2%.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

Notre travail étudie principalement l'amélioration des taux de reconnaissance des émotions simulées et des émotions spontanées. Plusieurs pistes ont été investiguées.

Une évaluation de deux méthodes de reconnaissance utilisant deux types de descripteurs différents à savoir les descripteurs géométriques et les descripteurs d'apparence est tout d'abord présentée. Les résultats obtenus montrent que la méthode basée sur les descripteurs géométriques est plus rapide et ne nécessite pas beaucoup d'images d'apprentissage tandis que la méthode d'apparence est plus robuste dans le cas d'images de basses résolutions.
Une étude de plusieurs schémas de fusion a ensuite été effectuée. Les taux de reconnaissance obtenus sur les deux bases montrent que les méthodes de fusion sont plus performantes que la méthode géométrique et la méthode d'apparence. Les méthodes de fusion basées sur la classification par SVM et par KPP et la fusion par la moyenne des probabilités a postériori ont montré leurs performances dans la reconnaissance des émotions simulées et des émotions spontanées. Leur évaluation sur des sujets non inclus dans l'ensemble de l'apprentissage a montré que les méthodes de fusion restent efficaces dans le cas des émotions simulées mais que, dans le cas des émotions spontanées, leurs taux de reconnaissance se détériorent.
Une étude de la reconnaissance des émotions sur des zones locales a été effectuée. Les zones choisies entourent les caractéristiques du visage à savoir le front, les yeux, le nez et la bouche. L'évaluation de la méthode d'apparence et de la méthode géométrique sur les zones locales a abouti à la construction d'une méthode

de fusion entre la géométrie de la bouche et l'apparence du visage. Cette dernière améliore les taux de reconnaissance des émotions spontanées.
Une troisième piste pour l'amélioration des taux de reconnaissance a enfin été testée. Une étude de plusieurs méthodes de sélection des descripteurs a été effectuée. La méthode de sélection basée sur les taux d'importance des descripteurs et sur l'étude de leur nombre optimal a permis une sélection de 2000 descripteurs pour les émotions simulées et de 4000 descripteurs pour les émotions spontanées. Une amélioration des taux de reconnaissance de la méthode d'apparence a été obtenue dans ce contexte.

5.2 Perspectives

Détection du visage

Dans le cadre de cette thèse, nous avons supposé une vue frontale du visage. La méthode de détection du visage [83] implémentée dans la bibliothèque OpenCV permet d'atteindre des taux de reconnaissance élevés. Cependant, dans le cas réel, les sujets ne sont pas toujours face à la caméra et certaines émotions sont parfois accompagnées d'un mouvement de tête comme le rire ou la peur. Envisager une détection du visage avec une certaine inclinaison améliorera la reconnaissance dans le cas réel. Deux cas peuvent alors être rencontrés. Le premier serait une inclinaison du visage où toutes les caractéristiques seraient encore visibles par la caméra. Une normalisation du visage et correction des déformations serait alors envisageable pour la méthode géométrique et la méthode d'apparence. Dans le deuxième cas où certaines zones du visage ne seraient pas visibles, un traitement des occultations devrait alors être pris en compte.

Reconnaissance par les zones locales

La reconnaissance des émotions par la fusion de la géométrie de la bouche et l'apparence du visage donne un taux de reconnaissance des émotions simulées supérieur à 97% et atteint un taux de 88.2% pour les émotions spontanées. Les zones locales utilisées pour l'application de la méthode d'apparence sont choisies sans étude préalable. Elles sont appliquées globalement autour des caractéristiques permanentes du visage en considérant leurs zones de déformation. Nous pensons

qu'une étude de patchs plus fins sur des points dynamiques pourrait améliorer les taux de reconnaissance par zone pour la méthode d'apparence. Une application de patchs dynamiques comme les patchs proposés dans [143] serait une éventuelle solution à ce problème.

Méthode de sélection

La sélection des descripteurs d'apparence apporte une amélioration du taux de reconnaissance global. Nous avons remarqué au cours de l'évaluation des méthodes de sélection que certaines émotions sont plus sensibles que d'autres à la réduction des descripteurs. Une sélection adaptée à chaque émotion améliorerait davantage la reconnaissance. En effet, l'importance des descripteurs change selon l'émotion. Dans ce cas une étude du nombre de descripteurs optimal pour chaque émotion pourrait être effectuée. Les descripteurs importants selon chaque émotion seraient alors sélectionnés. La reconnaissance serait enfin effectuée par plusieurs SVM de type un contre tous. La figure 5.1 présente le schéma alors envisagé pour la reconnaissance des émotions.

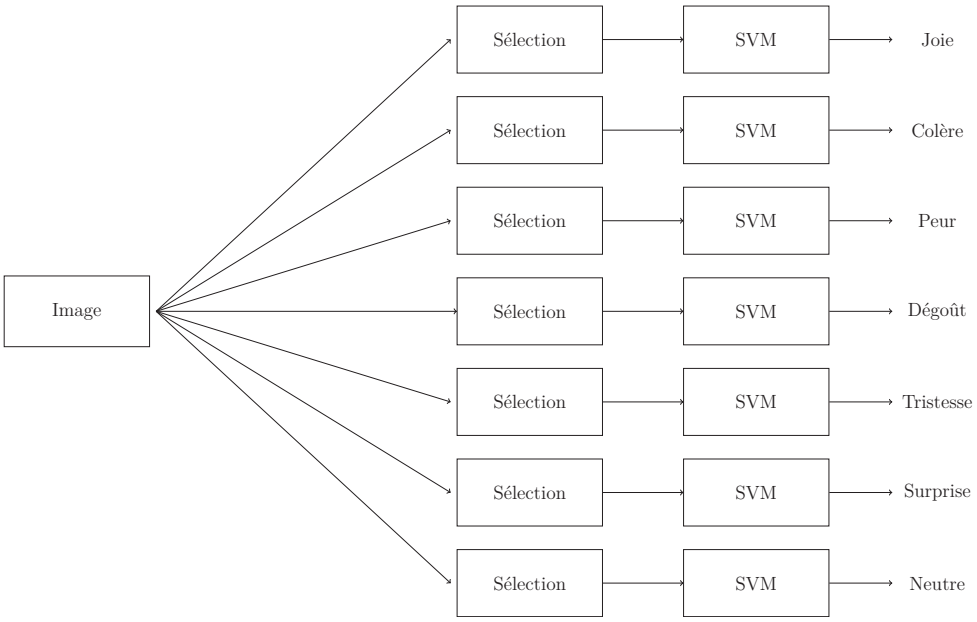


FIGURE 5.1 – Schéma de sélection par émotion

La dynamique des émotions

L'expression émotionnelle passe par trois phases dynamiques à savoir *onset* (le début de l'expression), *apex* (le maximum de l'expression) et *offset* (la fin de l'expression : retour à l'état neutre). La durée de l'expression est souvent liée à l'intensité de l'émotion. Les émotions spontanées étant souvent subtiles et souvent de plus courte durée nécessitent alors une intégration des informations temporelles qui modélisent leur aspect dynamique. Comme nous l'avons déjà mentionné dans le premier chapitre, deux types de méthodes traitant l'aspect dynamique des émotions existent. Nous proposons pour la suite une extraction de l'information temporelle en même temps que l'extraction de l'information spatiale. Nous suggérons pour cela les filtres de Gabor 3D [89].

Evaluation de la méthode de reconnaissance

Nous avons appliqué dans ce travail une évaluation basée principalement sur la validation croisée. Nous avons utilisé également la validation croisée par sujet, qui permet de tester des sujets de la base, mais non inclus dans l'ensemble d'apprentissage pour généraliser les résultats. D'autres techniques de test existent telles que la validation entre bases. Cette technique évalue le système dans des conditions très proches des conditions réelles puisque les sujets d'une autre base n'évoluent pas forcément dans le même environnement que la base d'apprentissage.

Bibliographie

[1] D. Deriso, J. Susskind, L. Krieger, and M. Bartlett. Emotion mirror : A novel intervention for autism based on real-time expression recognition. *Computer Vision ECCV 2012. Workshops and Demonstrations*, pages 2–5, 2012.

[2] A Mourao and J. Magalhaes. *ACM Multimedia*, pages 83–92.

[3] P. Ekman. An argument for basic emotions. *cognition and emotion*, pages 169–200, 1992.

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

[5] Frank Wallhoff. Facial expressions and emotion database, http ://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html, 2006.

[6] A. Mehrabian. Communication without words. *Psychology Today*, 2 :52–55, 1968.

[7] I. Kotsia, I. Buciu, and I. Pitas. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26 (7) :1052–1067, 2008.

[8] B. Jiang and K-b. Jia. Research of robust facial expression recognition under facial occlusion condition. In *Active Media Technology*, pages 92–100, 2011.

[9] C. Shan, S. Gong, and P. W. Mcowan. Facial expression recognition based on Local Binary Patterns : A comprehensive study. *Image and Vision Computing*, 27 :803–816, 2009.

[10] S. Gharsalli, H. Laurent, B. Emile, and X. Desquesnes. Various fusion schemes to recognize simulated and spontaneous emotions. In *the 10th International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 424–431, 2015.

[11] S. Gharsalli, B. Emile, H. Laurent, and X. Desquesnes. Feature selection for emotion recognition based on random forest. In *the 11th International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, 2016.

[12] A. Nugier. Histoire et grands courants de recherche sur les émotions. *Revu électronique de psychologie sociale*, pages 8–14, 2009.

[13] Kleinginna. P. R. and Kleinginna A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5, 1981.

[14] C. Darwin. *The expression of the emotions in man and animals*. 1872.

[15] P. Ekman. Are there basic emotions ? *Psychological Review*, 99 :550–553, 1992.

[16] J. William. What is emotion ? *Mind*, 9 :188–205, 1884.

[17] C. Lange. Ueber gemüthsbewegungen. 1887.

[18] R. Lebenson, Ekman. P., and W. V. Friesen. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, 27(4) :363–384, 1990.

[19] W. B. Cannon. The james-lange theory of emotion : A critical examination and an alternative theory. *American Journal of Psychology*, 39 :10–124, 1927.

[20] P. Bard. *A diencephalic mechanism for the expression of rage with special reference to the central nervous system*, 84 :490–513, 1928.

[21] C.E Izard. Human emotions. *Plenum Press, New York*, 1977.

[22] S. S. Tomkins. Affect theory. *Approaches to emotion*, pages 163–195, 1984.

[23] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 :1161–1178, 1980.

[24] R. Cowie, E. Douglas-cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder. “feeltrace” : An instrument for recording perceived emotion in real time. In *ISCA Workshop on Speech and Emotion*, pages 19–24, 2000.

[25] Plutchik R. Emotion a psycho-evolitionary synthesis. *Harper, New York*, 1980.

[26] R. Plutchik. The nature of emotions. *American Scientist*, 89, 2001.

[27] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier. A multimodal fuzzy inference system using a continuous facial expression representation

for emotion detection. *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 493–500, 2012.

[28] K. Karpouzis, N. Tsapatsoulis, and S. Kollias. Moving to continuous facial expression space using the mpeg-4 facial definition parameter (fdp) set. In *Proceedings of SPIE Electronic Imaging*, pages 443–450, 2000.

[29] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4) :384–392, 1993.

[30] I.R. Fasel, M.S. Bartlett, and J.R. Movellan. A comparison of gabor filter methods for automatic detection of facial landmarks. *Fifth International Conference on automatic face and gesture recognition*, pages 345–350, 2002.

[31] F. Abdat, C. Maaoui, and A. Pruski. Human-computer interaction using emotion recognition from facial expression. In *Computer Modeling and Simulation (EMS), 2011 Fifth UKSim European Symposium on*, pages 196–201, 2011.

[32] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In *Proceedings of International conference on Multimodal interaction*, pages 493–500, 2012.

[33] P. Ekman, W. V. Friesen, and M. O'Sullivan. Smiles when lying. *Journal of Personality and social psychology*, 54(3) :414–420, 1988.

[34] A. Mammucari, C. Caltagirone, P. Ekman, W. Friesen, G. Gainotti, L. Pizzamiglio, and P. Zoccolotti. Spontaneous facial expression of emotions in brain-damaged patients. *Cortex*, 24 :521–533, 1988.

[35] E. L. Rosenberg, P. Ekman, and J. A. Blumenthal. Facial expression and the affective component of cynical hostility in male coronary heart disease patients. *Health Psychology*, 17(4) :376–380, 1998.

[36] L. A. Camras. *Child Development*, volume 48(4), chapter Facial expressions used by children in conflict situations. Wiley on behalf of the Society for Research in Child Development, 1977.

[37] P. Ekman and W. Friesen. The facial action coding system : A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[38] P. Philippot. *Emotion et psychothérapie*. Mardaga, 2007.

[39] M. S. Bartlett, J. C. Hager, and P. Ekman. *Psychophysiology*, 1999.

[40] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 2006.

[41] Y. I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[42] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2006.

[43] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[44] M. Khademi and L-P. Morency. Relative facial action unit detection. In *Proceedings of the Winter conference on Applications in Computer Vision*, 2014.

[45] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5) :1–16, 2005.

[46] I.A. Essa. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :757–763, 1997.

[47] Standard mpeg-4 :. ISO/IEC 14496-2, Information Technology – Coding of Audio-Visual Objects, 2001.

[48] A. M. Tekalp and J. Ostermann. Face and 2-d mesh animation in mpeg-4. *Image Communication Journal*, pages 387–421, 1999.

[49] M. Pardas, A. Bonafonte, and J. L. Landabaso. Emotion recognition based on mpeg-4 facial animation parameters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 24–27, 2002.

[50] Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis a. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, and Stefanos D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4) :423–435, 2005.

[51] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, and S. Kollias. Parameterized facial expression synthesis based on mpeg-4. *EURASIP Journal on Applied Signal Processing*, page 1021–1038, 2002.

[52] M-S. Bartlett, L. Gwen, F. Ian, and R-M. Javier. Real time face detection and facial expression recognition : Development and applications to human computer interaction. *Computer Vision and Pattern Recognition Workshop*, 2003.

[53] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16 :172–187, 2007.

[54] Y. L. Tian, T.C. Kanade, and Jeffrey F., editors. *Facial expression analysis*. Springer-Verlag, 2004.

[55] Y.L. Tian, A. Brown, S. Hampapur, A. Pankanti, and B. Ruud. Real world real-time automatic recognition of facial expressions. In *IEEE workshop on Evaluation of Tracking and Surveillance*, 2003.

[56] S. P. Khandait, R. C. Thool, and P. D. Khandait. Automatic Facial Feature Extraction and Expression Recognition based on Neural Network. *International Journal of Advanced Computer Science and Applications*, 2(1) :113–118, 2011.

[57] Z. Hammal, N. Eveno, A. Caplier, and P Coulon. Parametric models for facial features segmentation. *Signal Processing*, 86(2) :399–413, 2005.

[58] Z Hammal, L Couvreur, A Caplier, and M Rombaut. Facial Expression Recognition Based on the Belief Theory : Comparison with Different Classifiers. In *International Conference on Image Analysis and Processing*, pages 743–752, 2005.

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

[60] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 38–45, 2007.

[61] F Cid, J A Prado, P Manzano, and P Bustos. Imitation system for humanoid robotics heads. *Journal of physical agents*, pages 23–30, 2013.

[62] D. Ghimire and J. Lee. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors (Basel, Switzerland)*, 13(6) :14–34, 2013.

[63] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, and T.S. Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25 :1856–1863, 2007.

[64] Kristin S. Benli and M. Taner Eskil. Extraction and Selection of Muscle Based Features for Facial Expression Recognition. *twenty-second International Conference on Pattern Recognition*, pages 1651–1656, 2014.

[65] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE transactions on neural networks*, 17(1) :233–238, 2006.

[66] S. Bashyal and Ganesh K. Venayagamoorthy. Recognition of facial expressions using Gabor wavelets and learning vector quantization. *Engineering Applications of Artificial Intelligence*, 21(7) :1056–1064, 2008.

[67] Hong-bo Deng, Lian-wen Jin, Li-xin Zhen, and Jian-cheng Huang. A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA. *International Journal of Information Technology*, pages 86–96, 2005.

[68] I. Buciu, I. Kotsia, and I. Pitas. Facial expression analysis under partial occlusion. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 5, pages 453–456, 2005.

[69] S.M. Lajevardi and Z. M. Hussain. Feature Extraction for Facial Expression Recognition based on Hybrid Face Regions. *Advances in Electrical and Computer Engineering*, 9(3) :63–67, 2009.

[70] S. M. Lajevardi and Z. M. Hussain. Facial Expression Recognition Using Log-Gabor Filters and Local Binary Pattern Operators. *International Conference on Communication, Computer and Power*, pages 349–353, 2009.

[71] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine*, 29 :1–14, 2007.

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

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

[74] L. Zhang, D. Tjondronegoro, and V. Chandran. Evaluation of Texture and Geometry for Dimensional Facial Expression Recognition. *International Conference on Digital Image Computing : Techniques and Applications*, pages 620–626, 2011.

[75] S. Berretti, B. Ben Amor, M. Daoudi, and A. Bimbo. 3D Facial Expression Recognition Using SIFT Descriptors of Automatically Detected Keypoints. *The Visual Computer*, 27(11) :1021–1036, 2011.

[76] K.B. Vinay and B.S. Shreyas. Face recognition using gabor wavelets. *Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 593–597, 2006.

[77] Yang Chen, Zhanping Fu, Yu Han, and Runsheng Wang. Independent component analysis of Gabor features for texture classification. *Optical Engineering*, 47(12), 2008.

[78] Chih-Jen Lee and Sheng-De Wang. Fingerprint feature extraction using Gabor filters. *Electronics Letters*, pages 288–290, 1999.

[79] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. *IEEE Conference of Systems, Man, and Cybernetics*, pages 1692–1698, 2005.

[80] Michael Grimm, D Dastidar, and K Kroschel. Recognizing emotions in spontaneous facial expressions. In *International Conference on Intelligent Systems And Computing (ISYC)*, 2006.

[81] C.S. Chung D.-Y. Yeung S. Liao, W. Fan. Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features. In *IEEE International Conference on Image Processing (ICIP)*, pages 665—-668, 2006.

[82] I. T. Jolliffe. Principal component analysis. *Springer-Verlag New York*, 487, 1986.

[83] P. Viola and M. Jones. Robust real-time object detection. *2nd international workshop on statistical and computational theories of vision - modeling, learning, computing, and sampling vancouver*, 2001.

[84] J. Whitehill and C.W. Omlin. Haar features for facs au recognition. In *Automatic Face and Gesture Recognition. 7th International Conference on*, pages 217–222, 2006.

[85] M. Pantic and I. Patras. Dynamics of Facial Expression : Recognition of Facial Actions and their Temporal Segments from Face Pro le Image Sequences. IEEE Transactions on. *IEEE transactions on Systems, Man, and Cybernetics - Part B : Cybernetics*, 36(2) :433–449, 2006.

[86] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll. Efficient recognition of authentic dynamic facial expressions on the feedtum database. In *IEEE International Conference on Multimedia and Expo*, pages 493–496, 2006.

[87] J. Hoey and J.J. Little. Value directed learning of gestures and facial displays. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[88] R.E. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *IEEE CVPR Workshop on Real-time Vision for Human–Computer Interaction*, 2004.

[89] T. Wu, M. S Bartlett, and J. R Movellan. Facial Expression Recognition Using Gabor Motion Energy Filters. In *Computer Vision and Pattern Recognition Workshops CVPRW 2010 IEEE Computer Society Conference*, volume 6, pages 42–47. Ieee, 2010.

[90] Guoying Zhao and Matti Pietikainen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, 30(12) :1117–1127, 2009.

[91] S. Lucey, Ahmed Bilal A., and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In K. Kurihara, editor, *Face Recognition Book*. Pro Literatur Verlag, Mammendorf, Germany, April 2007.

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

[93] Ahmed B. Ashraf, S. Lucey, Jeffrey F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B J. Theobald. The Painful Face - Pain Expression Recognition Using Active Appearance Models. *IEEE Transactions on Systems, Man, and Cybernetics*, 27(12) :9–14, 2007.

[94] L. Zhang, D. Tjondronegoro, and V. Chandran. Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition. *IEEE International Conference on Multimedia and Expo*, pages 1027–1032, 2012.

[95] J. Chen, D. Chen, Y. Gong, M. Yu, K. Zhang, and L. Wang. Facial expression recognition using geometric and appearance features. *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, pages 29–33, 2012.

[96] Zhengyou Z., M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *3td IEEE International Conference on Automatic Face and Gesture Recognition (FG 98)*, pages 454–459, 1998.

[97] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou. Image ratio features for facial expression Recognition Application. *IEEE Transactions on systems, man, and cybernetics-part B : Cybernetics*, 40(3) :779–788, 2010.

[98] Ligang Zhang. *Towards spontaneous facial expression recognition in real-world video.* PhD thesis, 2012.

[99] B. Ghattas and Ben Ishak A. Sélection de variables pour la classification binaire en grande dimension : comparaisons et application aux données de biopuces. *Journal de la Société Française de Statistique*, 2008.

[100] S-M. Lajevardi and Z-M. Hussain. Local feature extraction methods for facial expression recognition. *seventeenth European Signal Processing Conference(EUSIPCO)*, pages 60–64, 2009.

[101] H. Soyel and H. Demirel. Optimal feature selection for 3d facial expression recognition using coarse-tofine classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, 18(6) :1031—1040, 2010.

[102] L. Zhang and D. Tjondronegoro. Selecting , Optimizing and Fusing ‘ Salient ’ Gabor Features for Facial Expression Recognition. *Neural Information Processing (Lecture Notes in Computer Science)*, pages 724–732, 2009.

[103] Seyed Mehdi Lajevardi and Zahir M. Hussain. Feature selection for facial expression recognition based on mutual information. *fifth IEEE GCC Conference and Exhibition*, 2009.

[104] Y. Jiangang and B. Bir. Evolutionary feature synthesis for facial expression recognition. *Pattern Recognition Letters*, 27(11) :1289–1298, 2006.

[105] P. Li, P. Phung, A. Bouzerdoum, and Tivive. Feature selection for facial expression recognition. pages 35–39, 2010.

[106] Cohn-kanade au-coded expression database, http ://www.pitt.edu/ jeffcohn/ckandck+.htm.

[107] Takeo Kanade, Jeffrey F Cohn, and Tian Yingli. Comprehensive database for facial expression analysis. *the fourth IEEE International conference on Automatic Face and Gesture Recognition*, 5 :46, 2000.

[108] Z. Ambadar, Cohn J., and Reed L. All smiles are not created equal : Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, pages 17–34, 2009.

[109] M. F. Valstar and M. Pantic. Induced disgust , happiness and surprise" : an addition to the mmi facial expression database. *Workshop on Emotion*, pages 65–70, 2010.

[110] M. Pantic, M. Valstar, and L. Rademaker, R.and Maat. Web-based database for facial expression analysis. *IEEE International Conference on Multimedia and Expo*, pages 317 – 321, 2005.

[111] R. Gross. *Face Databases, Handbook of face recognition.* Springer-Verlag, 2005.

[112] M. Lyons, M. Kamachi, and J. Gyoba. The japanese female facial expression (jaffe) database.

[113] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3 :5–17, 2012.

[114] Hammal caplier database, http ://facialexpressions.free.fr/, 2003.

[115] Dailey M., Cottrell G. W., and Reilly J. California facial expressions (cafe). *unpublished digital images, University of California*, 2001.

[116] Pie database, http ://www.ri.cmu.edu/research_project_detail.html.

[117] I. Sneddon-C. Cox O. Lowry M. Mcrorie J. C. Martin L. Devillers S. Abrilian A. Batliner et al E. Douglas-Cowie, R. Cowie. The humaine database : Addressing the col- lection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, pages 488–500, 2007.

[118] Yale University. The yale face database b, http ://vision.ucsd.edu/ iskwak/extyaledatabase/extyaleb.html, 2001.

[119] A. Battocchi, F. Pianesi, and D. Goren-Bar. A first evaluation study of a database of kinetic facial expressions (dafex). In *Conference on Multimodal Interfaces ICMI*, pages 214–221, 2005.

[120] J. Katsyri, V. Klucharev, M. Frydrych, and M. Sams. Identification of synthetic and natural emotional facial expressions. In *Audio-Visual Speech Processing Converence AVSP*, 2003.

[121] Pets : Performance evaluation of tracking and surveillance.

[122] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7) :682–691, 2010.

[123] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa : A spontaneous facial action intensity database. In *IEEE Transactions on Affective Computing*, 2012.

[124] P. Ekman. Strong evidence for universals in facial expressions : A reply to russell's mistaken critique. *Psychological Bulletin*, 115 :268–287, 1994.

[125] R. Kraut and R. Johnson. Social and emotional messages of smiling : an ethological approach. *Journal of Personality and Social Psychology*, 37 :1539–1523, 1979.

[126] G. Bradski, T. Darrell, I. Essa, J. Malik, P. Perona, S. Sclaroff, and C. Tomasi. *http ://sourceforge.net/projects/opencvlibrary/*, 2006.

[127] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. *IEEE ICIP*, 1 :900–903, 2002.

[128] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi. A set of selected SIFT features for 3D facial expression recognition. *International Conference on Pattern Recognition*, pages 4125–4128, 2010.

[129] C.C. Chang and C.J. Lin. Libsvm : library for support vector machines. 2001.

[130] M. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*, 42 :28–43, 2012.

[131] G. Littlewort, I. Bartlett, M. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *IEEE Workshop Face Processing in Video*, 2004.

[132] J.R. Movellan. Tutorial on gabor filters. *MPLab Tutorials, UCSD MPLab, Tech*, 2005.

[133] Joni-Kristian Kamarainen, Ville Kyrki, and Heikki Kälviäinen. Robustness of gabor feature parameter selection. In *MVA*, pages 132–135, 2002.

[134] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[135] Cees G. and M. Snoek. Early versus late fusion in semantic video analysis. *ACM Multimedia*, pages 399–402, 2005.

[136] U. Niaz and B. Merialdo. Fusion methods for multi-modal indexing of web data. *fourteenth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 1–4, 2013.

[137] I. Kotsia, S. Zafeiriou, and I. Pitas. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, pages 833–851, 2008.

[138] Z-J. Chuang and C-H. Wu. Multi-modal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, 9 :45–62, 2004.

[139] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3) :389–422, 2002.

[140] A. Ishak and B. Ghattas. An efficient method for variable selection usingsvm-based criteria. *Pré-publication de l'Institut de Mathématiques de Luminy, Marseille, France.*, 2005.

[141] M. Y. Park and T. Hastie. L1 regularization path algorithm for generalized linear models. *Technical report, Stanford University*, 2006.

[142] L. Breiman. *Machine Learning*, 2001.

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

Sonia GHARSALLI

Reconnaissance des émotions par traitement d'images

Résumé :

La reconnaissance des émotions est l'un des domaines scientifiques les plus complexes. Ces dernières années, de plus en plus d'applications tentent de l'automatiser. Ces applications innovantes concernent plusieurs domaines comme l'aide aux enfants autistes, les jeux vidéo, l'interaction homme-machine.

Les émotions sont véhiculées par plusieurs canaux. Nous traitons dans notre recherche les expressions émotionnelles faciales en s'intéressant spécifiquement aux six émotions de base à savoir la joie, la colère, la peur, le dégoût, la tristesse et la surprise. Une étude comparative de deux méthodes de reconnaissance des émotions l'une basée sur les descripteurs géométriques et l'autre basée sur les descripteurs d'apparence est effectuée sur la base CK+, base d'émotions simulées, et la base FEEDTUM, base d'émotions spontanées. Différentes contraintes telles que le changement de résolution, le nombre limité d'images labélisées dans les bases d'émotions, la reconnaissance de nouveaux sujets non inclus dans la base d'apprentissage sont également prises en compte. Une évaluation de différents schémas de fusion est ensuite réalisée lorsque de nouveaux cas, non inclus dans l'ensemble d'apprentissage, sont considérés. Les résultats obtenus sont prometteurs pour les émotions simulées (ils dépassent 86%), mais restent insuffisant pour les émotions spontanées. Nous avons appliqué également une étude sur des zones locales du visage, ce qui nous a permis de développer des méthodes hybrides par zone. Ces dernières améliorent les taux de reconnaissance des émotions spontanées. Finalement, nous avons développé une méthode de sélection des descripteurs d'apparence basée sur le taux d'importance que nous avons comparée avec d'autres méthodes de sélection. La méthode de sélection proposée permet d'améliorer le taux de reconnaissance par rapport aux résultats obtenus par deux méthodes reprises de la littérature.

Mots clés : reconnaissance des émotions, émotions simulées, émotions spontanées, sélection des descripteurs, fusion des descripteurs.

Emotions recognition based on image processing

Summary :

Emotion recognition is one of the most complex scientific domains. In the last few years, various emotion recognition systems are developed. These innovative applications are applied in different domains such as autistic children, video games, human-machine interaction…

Different channels are used to express emotions. We focus on facial emotion recognition specially the six basic emotions namely happiness, anger, fear, disgust, sadness and surprise. A comparative study between geometric method and appearance method is performed on CK+ database as the posed emotion database, and FEEDTUM database as the spontaneous emotion database. We consider different constraints in this study such as different image resolutions, the low number of labelled images in learning step and new subjects. We evaluate afterward various fusion schemes on new subjects, not included in the training set. Good recognition rate is obtained for posed emotions (more than 86%), however it is still low for spontaneous emotions. Based on local feature study, we develop local features fusion methods. These ones increase spontaneous emotions recognition rates. A feature selection method is finally developed based on features importance scores. Compared with two methods, our developed approach increases the recognition rate.

Keywords : emotion recognition, posed emotions, spontaneous emotions, feature selection, feature fusion.

Laboratoire PRISME
IUT, 63 Avenue de Lattre de Tassigny, 18000 Bourges,
France