Construction non supervisée d'un modèle expressif spécifique à la personne

Raphaël Weber

ANNÉE 2017

CentraleSupélec

UNIVERSITE BRETAGNE LOIRE

THÈSE / CENTRALESUPÉLEC

sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE CENTRALESUPÉLEC

Mention : Signal, Image, Vision

École doctorale 601 « Mathématiques et Sciences et Technologies de l'Information et de la Communication (MATHSTIC) »

présentée par

Raphaël Weber

Préparée à l'UMR 6164 - IETR (Équipe FAST)
Institut d'Électronique et de Télécommunications de Rennes

Construction non supervisée d'un modèle expressif spécifique à la personne

Thèse soutenue le 23 novembre 2017

devant le jury composé de :

Alice CAPLIER
Prof. à Grenoble INP / *rapporteur*

Sylvie GIBET
Prof. à Université Bretagne Sud / *rapporteur*

Catherine PELACHAUD
DR CNRS, Université Pierre et Marie Curie / *examinatrice*

Nicolas SABOURET
Prof. à Université Paris-Sud / *examinateur*

Renaud SÉGUIER
Prof. à CentraleSupélec / *directeur de thèse*

Catherine SOLADIÉ
Prof. associée à CentraleSupélec / *encadrante*

Remerciements

Je tiens tout d'abord à remercier Alice Caplier, Professeur à Grenoble INP (GIPSA-Lab), et Sylvie GIBET, Professeur à Université Bretagne Sud (IRISA) d'avoir accepté d'être rapporteurs de cette thèse. Je remercie également Catherine Pelachaud, directrice de recherche au CNRS (Université Pierre et Marie Curie, ISIR), et Nicolas Sabouret, Professeur à Université Paris Sud (LIMSI-CNRS), d'avoir accepté de juger mon travail en tant que membres du jury.

Je souhaite remercier vivement Renaud Séguier pour la direction de cette thèse, ainsi que Catherine Soladié pour son encadrement et pour tout le temps qu'elle m'a accordé afin que je me familiarise avec le sujet de cette thèse. Je les remercie tout particulièrement pour la liberté qu'ils m'ont laissé pour mener mes travaux et la confiance qu'ils m'ont accordée. Je remercie également Pierre-Yves Richard, professeur à CentraleSupelec, campus de Rennes, qui m'a mis sur la piste de cette thèse alors que les candidatures étaient en cours.

Je tiens à remercier chaleureusement l'ensemble des membres de l'équipe FAST et de l'équipe SCEE, avec une mention particulière pour Salah-Eddine et Vincent Barrielle. Mes remerciements vont aussi à toutes les personnes qui ont accepté de faire partie de la base de données dont j'ai fait l'acquisition.

Une pensée toute particulière s'adresse à Vincent Gouldieff, qui a effectué en même temps que moi sa thèse à CentraleSupelec, campus de Rennes. Un colocataire de choc.

Ces trois années de thèse ont aussi été l'occasion pour moi de me plonger plus profondément dans la musique Reggae et la culture *Sound System* associée et de progressivement en devenir un activiste, une sorte de thèse en parallèle. *Big up* à tous les activistes avec qui j'ai collaboré.

Enfin, mes remerciements vont à ma famille et mes amis pour leur soutien, ainsi qu'à Anna pour son soutien continu et tous les moments que l'on a partagés ensemble.

Résumé

L'analyse automatique des expressions faciales connaît un intérêt grandissant ces dernières décennies du fait du large champ d'applications qu'elle peut servir. Un des domaines d'applications visé est le médical avec notamment l'analyse automatique des variations de comportement pour le maintien à domicile des personnes âgées.

Cette thèse propose un modèle expressif continu spécifique à la personne construit de manière non supervisée (*i.e.* sans connaissance *a priori* sur la morphologie du sujet) pour répondre à ce besoin d'analyse automatique. Dans notre cadre applicatif, le système est amené à analyser un seul sujet, c'est pourquoi le modèle est spécifique à la personne, ce qui permet une analyse plus fine qu'un modèle générique. La construction non supervisée du modèle permet de s'affranchir de toute étape de calibration qui peut être laborieuse et désagréable pour le sujet. L'aspect continu du modèle permet d'analyser des expressions non prototypiques que l'on retrouve dans la vie courante. De plus, notre système doit être capable d'analyser les expressions faciales dans un environnement non contraint en termes de pose de la tête de parole.

Les travaux réalisés se basent sur des travaux existants sur la représentation invariante des expressions faciales. Le modèle sur lequel nous nous basons nécessite l'acquisition du visage neutre, il est donc construit de manière supervisée. De plus, il est construit sur des expressions de base synthétisées à partir du visage neutre, il ne rend donc pas compte des expressions faciales réelles du sujet. Nous proposons dans cette thèse de rendre la construction de ce modèle non supervisée en détectant automatiquement le visage neutre et d'adapter le modèle automatiquement aux expressions faciales réelles du sujet de manière non supervisée. L'idée de l'adaptation est de détecter, à la fois globalement et localement, les expressions de base réelles du sujet afin de remplacer les expressions de base synthétisées du modèle puis de les affiner, tout en maintenant un ensemble de contraintes. Nous proposons deux variantes de la méthode d'adaptation : séquentielle et sur une fenêtre temporelle. Pour pouvoir mener à bien la détection automatique du visage neutre et l'adaptation, nous avons développé notre méthode de

reconnaissance locale et globale d'expressions faciales. Pour pouvoir mener l'analyse dans un environnement non contraint en terme de pose de la tête, cette méthode est robuste à la pose.

Nous avons testé notre système d'adaptation sur des expressions posées, des expressions spontanées dans un environnement contraint et des expressions spontanées dans un environnement non contraint. Les résultats montrent l'efficacité de l'adaptation et l'importance de l'étape de vérification des contraintes lors des tests dans un environnement non contraint.

Mots clefs Analyse des expressions faciales, Reconnaissance d'expressions faciales, Pose de la tête, Base de données d'expressions faciales, Variété, Non supervisé, Adaptation

Table des matières

Introduction

Sommaire

L'analyse des émotions fait partie intégrante de notre vie quotidienne dans les interactions sociales. En effet, les émotions peuvent être vues comme les interfaces de notre organisme avec le monde extérieur. Le processus émotionnel consiste alors en une réponse de notre organisme à des stimuli extérieurs [1]. Analyser les émotions de notre interlocuteur lors d'une interaction nous permet donc de nous renseigner sur ses réactions et ses intentions. Pour ce faire, nous analysons diverses informations vecteurs de l'état émotionnel : la voix (intonation, contenu du message, ...), l'expression faciale ou le mouvement du corps (pose de la tête, gestuelle, ...). Parmi ces modalités, les expressions faciales ont fait l'objet d'un grand nombre d'études sur le plan psychologique, que ce soit sur leurs origines et leur signification ou sur l'interprétation que nous en faisons. Ces études montrent que les expressions faciales sont d'une importance cruciale lorsque nous cherchons à analyser l'état émotionnel de notre interlocuteur. Il a notamment été montré que 55% de l'impact d'une communication est dû au langage corporel et aux expressions faciales [2]. Les travaux d'Ekman ont mis en évidence l'existence de 6 émotions de base, chacune se manifestant par une expression faciale particulière, appelée « expression prototypique » [3]. Ces travaux ont cherché à montrer que ces expressions prototypiques sont reconnues universellement, *i.e.* indépendamment de l'origine culturelle.

Dès la Grèce antique, nous retrouvons cette idée qu'une expression faciale bien précise correspond à une émotion particulière avec les masques de théâtre (voir figure 1). Ces masques sont utilisés pour différencier des archétypes de personnages. Sur certains d'entre eux, nous retrouvons une expression faciale associée à l'émotion qu'est censé sus-

FIGURE 1 – Masques de théâtre de la tragédie et de la comédie, image reprise de [4].

citer le personnage. Par exemple, les masques de la tragédie inspirent la terreur et les masques de la comédie accentuent le ridicule.

Lorsque la communauté scientifique en traitement du signal et en vision par ordinateur a commencé à s'intéresser à l'analyse automatique des émotions, elle s'est naturellement tourné vers l'analyse automatique des expressions faciales, en se basant notamment sur les travaux d'Ekman sur les expressions prototypiques [3]. Les travaux de cette thèse s'inscrivent dans ce cadre. Dans cette introduction, nous présentons dans un premier temps le contexte et les motivations. Puis nous précisons à quelle problématique répondent les travaux de cette thèse. Enfin, nous présentons l'organisation du présent manuscrit de cette thèse.

Contexte et motivations

Contexte scientifique

L'analyse automatique des expressions faciales peut s'inscrire dans une analyse unimodale, *i.e.* seule l'image ou la vidéo du visage est utilisée comme donnée d'entrée, ou multimodale, *i.e.* l'image ou la vidéo du visage est combinée à d'autres modalités.

Historiquement, deux problèmes sont largement étudiés dans un premier temps : la reconnaissance d'expressions faciales et la reconnaissance d'unités d'action [5]. Il s'agit dans les deux cas d'un problème de classification des expressions faciales. Dans le premier cas, les classes sont des expressions globales (sur tout le visage), correspondant souvent aux expressions prototypiques d'Ekman [3]. Dans le second cas, les classes sont des expressions locales (ne concernant qu'une zone du visage), correspondant aux unités d'action du système FACS [6]. Le système FACS permet de décrire les déformations faciales selon les muscles faciaux qui sont activés, chaque unité d'action correspondant

à l'activation d'un muscle. Par la suite, le problème d'estimation des dimensions émotionnelles est également considéré. Contrairement aux problèmes de reconnaissance, il s'agit ici d'un problème de régression où il faut estimer la valeur d'une ou plusieurs dimensions continues permettant de décrire l'état émotionnel. Pour résoudre ces problèmes, différentes approches sont considérées et nous pouvons les distinguer selon que la modélisation des expressions faciales est générique, *i.e.* indépendante de la personne, ou spécifique à la personne, *i.e.* le système est paramétré pour un seul et unique sujet et ne peut être appliqué que sur celui-ci.

Les premiers systèmes sont développés sur des bases de données d'expressions faciales posées dans un environnement contraint, typiquement le laboratoire. Cependant, il existe des différences entre les expressions posées et les expressions spontanées [7, 8, 9, 10]. Dans les années 2000, de plus en plus de bases de données d'expressions spontanées sont apparues pour pouvoir développer des systèmes capables d'analyser de telles expressions. Plus récemment, l'analyse des expressions « in-the-wild », *i.e.* dans des conditions réelles, est identifiée comme la nouvelle problématique sur laquelle se pencher pour pouvoir proposer des systèmes véritablement robustes.

Contexte applicatif

L'intérêt grandissant que connaît l'analyse des expressions faciales depuis les 3 dernières décennies va de paire avec les applications qu'elle peut servir. Ces applications touchent autant le génie logiciel, l'éducation, le « marketing », l'interaction homme-machine que le médical [11, 12].

En ce qui concerne le médical, les applications potentielles consistent majoritairement en l'analyse de variations de comportement caractéristiques de certaines pathologies ou conditions médicales, telle la dépression ou l'anxiété [12]. Cela concerne également le maintien à domicile de personnes âgées. En effet, l'allongement de la durée de vie implique un vieillissement de la société et il peut être souhaitable de prolonger l'indépendance des personnes âgées grâce au maintien à domicile. Une des applications qui nous intéresse à terme est l'analyse des variations de comportement pour pouvoir lever une alarme lorsque quelque chose d'anormal arrive à la personne âgée.

Enoncé du problème

Dans le cadre applicatif présenté ci-dessus, notre système doit répondre aux besoins suivants :

- Analyse d'un seul sujet,

- Aucune étape de calibration nécessaire, cette étape peut en effet être laborieuse et désagréable pour le sujet,
- Analyse d'expressions non prototypiques car ce sont celles que l'on rencontre le plus souvent dans la vie réelle [13],
- Analyse dans un environnement non contraint en termes de pose de la tête et de parole.

Étant donné que notre système est amené à analyser un seul sujet, nous choisissons une modélisation spécifique à personne, ce qui permet d'avoir une analyse plus fine. Pour répondre au besoin d'absence de calibration, nous devons construire notre modèle de manière non supervisée, *i.e.* sans connaissance *a priori* de la morphologie du sujet. Pour pouvoir analyser des expressions non prototypiques, nous proposons une modélisation continue de l'ensemble des déformations faciales du sujet. Notre système est donc un modèle continu des expressions faciales, spécifique à la personne et construit de manière non supervisée. Nous testerons la robustesse de notre système sur une base de données d'expressions spontanées dans un environnement non contraint.

Organisation de la thèse

Dans le chapitre 1, nous dressons un état de l'art de l'analyse des expressions faciales. Nous nous intéressons d'abord à la théorie de l'émotion et aux modèles proposés par les psychologues. Nous nous penchons ensuite sur l'analyse des expressions faciales à proprement parler : nous présentons les différentes problématiques existantes, puis nous dressons un état de l'art des bases de données d'expressions faciales, nous décrivons le système classique d'analyse des expressions faciales, nous dressons un état de l'art de la reconnaissance d'expressions faciales robuste à la pose et enfin nous dressons un état de l'art des modèles expressifs spécifiques à la personne.

Dans le chapitre 2, nous proposons deux outils qui sont utilisés dans notre système. Le premier outil est le modèle spécifique à la personne construit de manière faiblement supervisée basé sur [14]. Le second outil est un système de reconnaissance d'expressions faciales robuste à la pose que nous avons développé.

Dans le chapitre 3, nous présentons notre système dans sa globalité. Dans un premier temps, nous détectons automatiquement le visage neutre du sujet pour pouvoir ensuite synthétiser des expressions. Le modèle spécifique à la personne est initialisé avec le visage neutre et les expressions synthétisées. Ensuite, le modèle s'adapte aux expressions réelles du sujet de manière non supervisée. Nous présentons deux variantes de notre méthode d'adaptation non supervisée : séquentielle et sur une fenêtre temporelle.

Dans le chapitre 4, nous présentons les résultats de notre méthode d'adaptation non supervisée. Dans un premier temps, nous présentons le protocole expérimental que nous suivons. Ensuite, nous présentons les résultats relatifs à l'initialisation du modèle spécifique à la personne. Enfin, nous présentons les résultats de notre méthode d'adaptation non supervisée pour les deux variantes. Pour chacune des variantes, nous testons notre méthode sur des expressions posées dans un environnement contraint, sur des expressions spontanées dans un environnement contraint et sur des expressions spontanées dans un environnement non contraint.

Chapitre 1

État de l'art

Sommaire

Dans ce chapitre, nous présentons l'état de l'art sur l'analyse automatique des expressions faciales.

Dans la première section, nous nous intéressons à la théorie de l'émotion comme sujet de recherche en psychologie. En effet, l'analyse automatique des expressions faciales découle de la volonté de la communauté de vision par ordinateur et de traitement du signal d'analyser automatiquement les émotions. Les chercheurs se sont donc basés sur les modèles de représentation des émotions développés par les psychologues pour ensuite les transposer à un problème d'apprentissage machine.

La seconde section porte sur l'analyse des expressions faciales. Nous présentons d'abord les différentes problématiques liées à l'analyse des expressions faciales. Puis nous dressons un état de l'art des bases de données d'expressions faciales, essentielles pour développer et tester un système d'analyse des expressions faciales. Nous nous intéressons ensuite à l'analyse des expressions faciales à proprement parler et nous portons

une attention particulière à la reconnaissance d'expressions faciales robuste à la pose et aux modèles expressifs spécifiques à la personne. Nous concluons par les contributions de cette thèse.

1.1 Théorie de l'émotion

Parmi les premiers travaux de recherche sur les émotions, l'ouvrage de Darwin intitulé *The expression of the emotion in man and animals* publié en 1872 [15] a eu un impact notable. Il y reprend sa théorie évolutionniste en considérant que l'humain a hérité son expressivité émotionnelle des systèmes comportementaux des autres espèces animales. Depuis lors, la recherche sur les émotions connaît un certain engouement.

Il n'existe pas de définition unique et arrêtée de l'émotion, cela reste une question ouverte pour les psychologues. Parmi les théories qui ont eu le plus d'influence, on peut citer Scherer et son modèle de processus par composantes (« Component Process Model ») [1]. Les émotions sont définies comme étant les interfaces de l'organisme avec le monde extérieur. Le processus émotionnel est alors vu comme une réponse à des stimuli extérieurs, se décomposant en 3 étapes : analyse des stimuli par l'organisme, mise en place d'actions aux niveaux physiologique et psychologique en réponse aux stimuli, et communication par l'organisme de l'état de l'individu à son environnement [1].

Dans cette section, nous commençons par décrire brièvement les principaux modes de représentation des émotions proposés par les psychologues (sous-section 1.1.1). Ensuite, nous énumérons les différentes modalités des émotions dans la sous-section 1.1.2, dont font partie les expressions faciales.

1.1.1 Modes de représentation des émotions

Plusieurs modes de représentation des émotions ont été proposés par les psychologues ces 50 dernières années. Nous présentons brièvement ici les trois modes de représentation les plus populaires : la représentation catégorielle, la représentation dimensionnelle et la représentation basée sur l'évaluation. Par la suite, la communauté de vision par ordinateur et de traitement du signal a utilisé certaines de ces représentations pour analyser automatiquement les émotions. Le lecteur peut se référer à [16] pour avoir une revue plus complète des différentes théories de l'émotion.

FIGURE 1.1 – Les émotions de base d'Ekman [3] : colère, dégoût, peur, joie, tristesse, surprise. Images extraites de la base de données MUG [22].

Représentation catégorielle

La représentation catégorielle des émotions consiste à associer un mot de la vie courante à chaque émotion identifiable. Elle repose sur l'hypothèse de l'universalité des émotions [3, 17, 18] selon laquelle il existe un ensemble d'émotions de base qui sont communes à toutes les cultures. Chaque émotion de base est identifiée par une même expression faciale et est interprétée de la même manière quelle que soit la culture.

Une méta-analyse a été menée sur les études de l'universalité des émotions [19]. Elle montre que le taux de reconnaissance des émotions, au sein d'un même groupe culturel et entre des groupes culturels différents, est meilleur que le hasard. Ceci étant dit, les auteurs ont mis en évidence un avantage pour la reconnaissance au sein d'un même groupe culturel, ce qui vient nuancer le propos sur l'universalité des émotions.

La catégorisation la plus populaire à ce jour résulte des travaux d'Ekman sur l'universalité des émotions [3]. Il a identifié 6 émotions de base [3, 20] : la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Les expressions faciales associées à ces émotions de base sont appellées expressions prototypiques et sont illustrées dans la figure 1.1. Le mépris a ensuite été ajouté à cette liste d'émotions de base universelles [21].

Par la suite sont apparues d'autres catégorisations incluant des émotions secondaires, c'est-à-dire résultant des interactions sociales. Par exemple, dans [23] une taxonomie est développée à partir d'une analyse linguistique des termes émotionnels de la langue anglaise. Elle résulte en une catégorisation de 412 émotions arrangées en 24 groupes, incluant par exemple l'ennui, l'intérêt ou la frustration. Malgré son aspect exhaustif, cette catégorisation est moins utilisée que les 6 émotions de base d'Ekman [3] pour l'analyse automatique des expressions faciales. Cela est notamment dû au fait que moins d'études ont porté sur l'universalité de ces émotions secondaires. De plus, il est plus complexe d'obtenir de bons taux de reconnaissance sur 412 classes ou 42 classes que sur 6 classes et l'acquisition d'une base de données avec autant d'émotions s'avère compliquée.

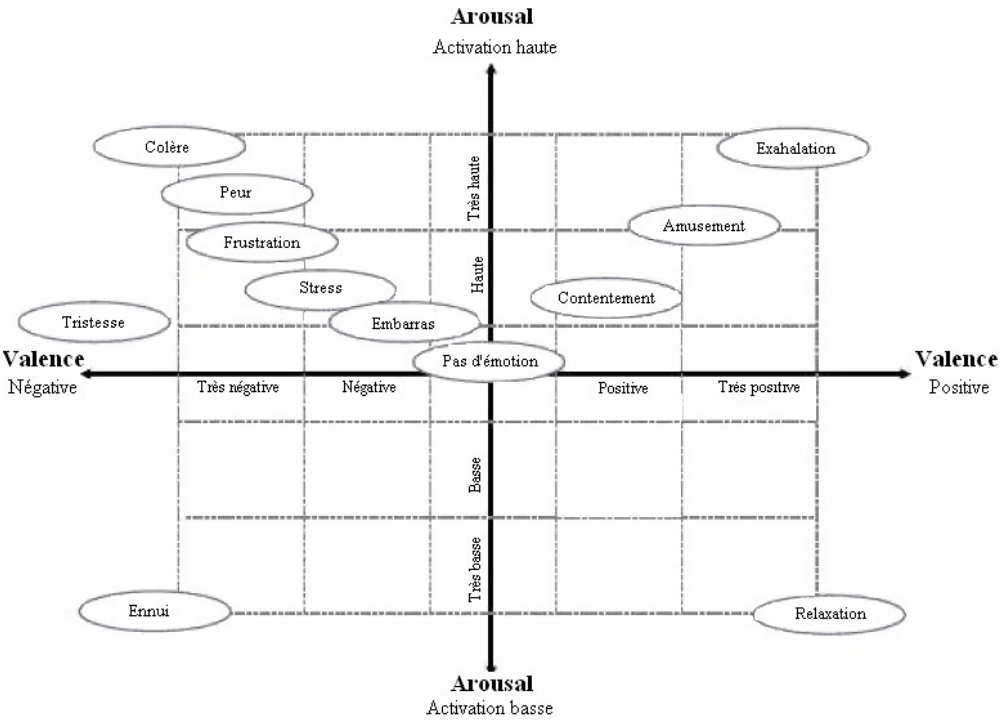


FIGURE 1.2 – Exemple d'émotions représentées à l'aide des dimensions valence et arousal (figure reprise de [29]).

Représentation dimensionnelle

Une autre approche très populaire parmi les psychologues est la représentation dimensionnelle. L'émotion est alors définie de manière continue selon plusieurs dimensions émotionnelles, aux antipodes de la représentation catégorielle.

Dans [24], il est proposé de définir l'état émotionnel à partir de trois dimensions indépendantes et bipolaires que sont la valence (plaisir - mécontentement ou positif - négatif), l'arousal (degré d'activation physiologique, actif - passif) et le pouvoir (dominance - soumission). Des études ont suivi sur la représentation à deux dimensions en utilisant uniquement la valence et l'arousal [25, 26, 27]. Le but de cette représentation bidimensionnelle est de rendre compte des similarités et des différences des émotions ressenties [28]. La figure 1.2 illustre quelques exemples d'émotions représentées selon ces deux dimensions.

Plus récemment, la représentation bidimensionnelle valence/arousal a été étendue à une représentation à 4 dimensions [28] : la valence, l'arousal, le pouvoir et l'imprévisibilité. D'autres combinaisons de dimensions ont été étudiées pour la représentation des

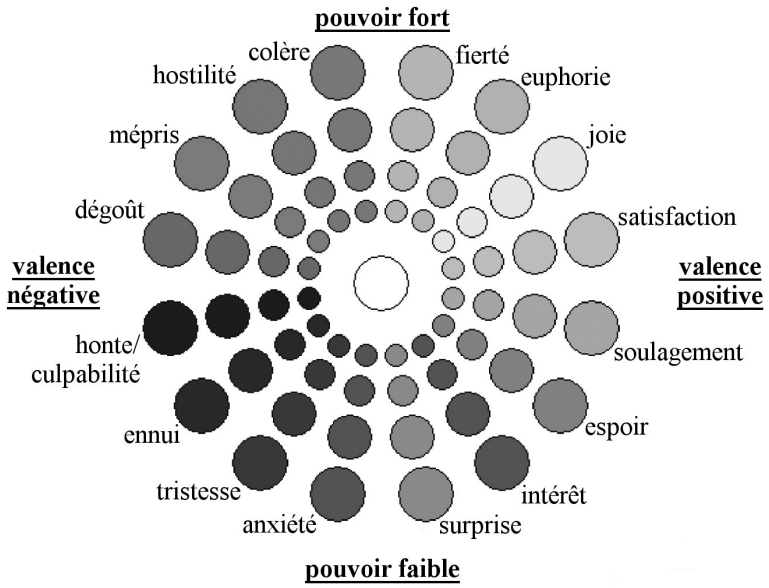


FIGURE 1.3 – Représentation des émotions avec la roue émotionnelle de Genève (Geneva Emotion Wheel) [30] selon les dimensions valence-pouvoir.

émotions. Par exemple, la roue émotionnelle de Genève (« Geneva Emotion Wheel ») [30] repose sur un représentation bidimensionnelle valence-pouvoir (voir figure 1.3).

Représentation basée sur l'évaluation

La théorie de l'évaluation postule le fait que les émotions sont le résultat de notre évaluation d'événements [31] et tente de donner une explication au fait que les personnes ne réagissent pas de la même manière à un même événement. Suite à un événement, le sujet évalue la situation selon plusieurs composantes internes et externes, puis cette évaluation cause une réponse émotionnelle. La particularité de cette théorie est de considérer que même en l'absence de stimulation physiologique les émotions peuvent résulter de notre interprétation d'événements. Plusieurs modèles reposant sur cette théorie ont été proposés, les plus populaires étant le modèle structurel et le modèle de processus.

Dans le cas du modèle structurel [32], trois aspects sont pris en compte pour évaluer l'émotion : l'aspect relationnel (relation entre le sujet et l'environnement), l'aspect motivationnel (évaluation de la pertinence de l'événement vis-à-vis des objectifs personnels du sujet) et l'aspect cognitif (évaluation de l'événement). Le modèle structurel suggère que différentes émotions sont suscitées lorsque l'événement est évalué différemment selon ces trois aspects.

Plusieurs modèles de processus ont été proposés comme une extension du modèle structurel. Alors que le modèle structurel met l'accent sur ce qui est évalué par le sujet, les modèles de processus analysent la manière dont le sujet évalue les événements à l'aide de plusieurs composantes. Les principales composantes des modèles de processus sont les stimuli perceptuels, le traitement des informations par association et le raisonnement. Dans le modèle bi-processus de l'évaluation [33], le traitement par association et le raisonnement fonctionnent en parallèle suite aux stimuli perceptuels, ce qui permet une évaluation plus complexe de l'émotion. Dans le modèle par vérification séquentielle multi-niveaux [34], l'évaluation se fait sur trois niveaux (inné, appris et délibéré) avec des contraintes séquentielles à chaque niveau qui impliquent un séquencement spécifique de l'évaluation.

1.1.2 Modalités des émotions

Plusieurs modalités ont été identifiées comme étant porteuses d'informations sur l'état émotionnel : les signaux physiologiques, les signaux audio (la voix) et les signaux visuels (les expressions faciales, les mouvement du corps, l'orientation de la tête, la direction du regard).

Il a été établi par Mehrabian [2] que 55% de l'impact d'une communication est dû au langage corporel et aux expressions faciales, contre 38% pour la partie vocale (intonation, ...) et 7% pour la partie verbale (sens du message). Les expressions faciales jouent donc un rôle crucial dans l'analyse des émotions et du comportement.

Ceci étant dit, la seule modalité des expressions faciales ne suffit pas forcément pour distinguer toutes les émotions les unes des autres. Une étude a montré l'importance du contexte pour reconnaître les émotions négatives à partir de l'expression faciale, ajoutant qu'il est possible qu'une émotion ne laisse aucun indice visuel [35].

1.2 Analyse des expressions faciales

Au début des années 90, l'analyse automatique des expressions faciales en est encore à ses balbutiements [36]. Par la suite, un effort significatif a porté sur le développement d'algorithmes de détection du visage et de suivi de points caractéristiques du visage. Ces outils ont permis à la recherche sur l'analyse automatique d'expressions faciales de prendre son envol et d'être en plein essor, en témoignent les nombreux états de l'art sur le sujet [5, 37, 38, 39, 40, 12, 41].

La figure 1.4 récapitule les sujets qui sont traités dans cette section. Nous présentons d'abord dans la sous-section 1.2.1 les différentes problématiques liées à l'analyse des

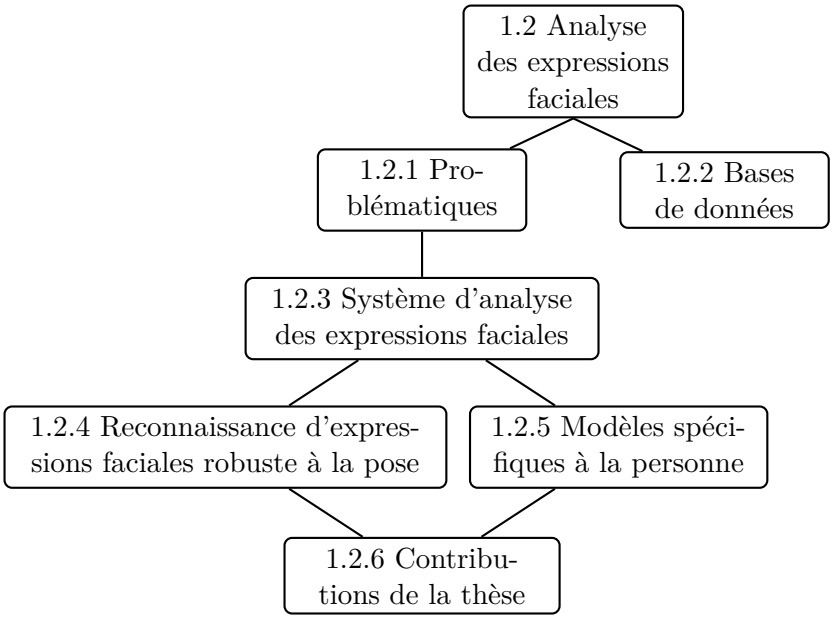


FIGURE 1.4 – Vue globale de la section 1.2 sur l'état de l'art de l'analyse des expressions faciales.

expressions faciales. Pour pouvoir traiter ces problématiques, le choix d'une base de données adéquate est essentiel. C'est pourquoi dans la sous-section 1.2.2 nous dressons un état de l'art des bases de données d'expressions faciales. Dans la sous-section 1.2.3, nous décrivons le schéma classique d'un système d'analyse des expressions faciales et proposons une classification des différentes méthodes existantes. Si d'excellents résultats ont été obtenus sur des données acquises dans l'environnement contrôlé du laboratoire, il reste un certain nombre de défis à relever lorsqu'on analyse des données proches du contexte de la vie réelle, notamment en ce qui concerne la robustesse à la pose de la tête et l'analyse d'expressions non prototypiques. Dans la sous-section 1.2.4 nous présentons les méthodes de reconnaissance d'expressions faciales robustes à la pose. En ce qui concerne l'analyse d'expressions non prototypiques, nous proposons dans cette thèse de construire un modèle continu et spécifique à la personne de manière non supervisée, *i.e.* sans information *a priori* sur le sujet. L'aspect continu du modèle permet d'analyser une grande variété d'expressions tandis que l'aspect spécifique à la personne permet une analyse plus fine des expressions du sujet. Nous présentons donc dans la sous-section 1.2.5 les méthodes de modélisation spécifique à la personne. Enfin, nous terminons cette section par les contributions de la thèse (sous-section 1.2.6).

1.2.1 Problématiques liées à l'analyse des expressions faciales

Reconnaissance d'expressions faciales

La reconnaissance d'expressions faciales consiste en la classification d'une expression faciale en y associant un label. Cette classification repose sur la représentation catégorielle des émotions (voir sous-section 1.1.1). Les labels utilisés pour la classification sont très souvent les six émotions de base mises en évidence par Ekman [3] (colère, dégoût, peur, joie, tristesse et surprise), même si certains travaux ajoutent des émotions secondaires dans la classification [42, 43]. L'expression faciale est directement interprétée comme une émotion, cette approche est donc basée sur le message (ou jugement) [37].

De nombreux chercheurs se sont penchés sur la reconnaissance d'expressions faciales et les méthodes développées atteignent désormais des taux de reconnaissance compris entre 90% et 100% pour les émotions de base dans un environnement contrôlé [44, 45].

Cette manière de caractériser les expressions faciales présente toutefois quelques limites. En effet, elle repose sur le fait que l'on associe une expression faciale à un état émotionnel, or ce n'est pas toujours le cas [46]. De plus, une étude a montré que les émotions non comprises dans les six émotions de base d'Ekman [3] se produisent assez fréquemment dans la vie courante alors que certaines des émotions de base se produisent rarement dans la vie courante [13]. Cette même étude montre que les émotions de base qui se produisent le plus fréquemment dans la vie courante sont la joie et la colère. Plus récemment, une autre étude a montré que des émotions non basiques, telles que l'ennui ou la frustration, se produisent cinq fois plus souvent que les émotions de base lors d'une interaction humain-machine [47]. Un classifieur entraîné sur les six émotions de base ne pourra donc pas fournir une analyse pertinente sur des données de la vie réelle.

Reconnaissance d'unités d'action

Une autre manière de caractériser les expressions faciales est l'utilisation du système FACS [6]. Ce système propose de caractériser une expression faciale par l'activation ou non des muscles du visage, ce qui se traduit par l'activation d'unités d'action (« action units » - AUs). Le système FACS est composé de 44 unités d'action. Une expression faciale est donc caractérisée par une combinaison d'unités d'action et leur intensité respective. L'annotation manuelle de ces unités d'action est une tâche ardue réservée à des experts ayant suivi une formation.

La reconnaissance d'unités d'action consiste en une classification des déformations du visage selon les unités d'action. Les systèmes développés se concentrent sur un sous-ensemble des unités d'action, souvent entre une dizaine et une quinzaine [48, 49, 50]. Cela

s'explique notamment par le fait que les bases de données annotées en unités d'action ne le sont que pour ces sous-ensembles. L'expression faciale est décrite par les déformations faciales observées et non pas par l'émotion qui peut y être associée, cette approche est donc basée sur le signal [37] et non sur le message.

Estimation continue de dimensions émotionnelles

La communauté scientifique s'est aussi penchée sur l'estimation automatique de dimensions émotionnelles, telles que arousal et valence. Contrairement à la reconnaissance d'expressions faciales et d'unités d'action qui sont des problèmes de classification, le but est ici d'estimer la valeur d'une grandeur continue, il s'agit donc de résoudre un problème de régression.

L'estimation de dimensions émotionnelles connaît un intérêt grandissant, comme en témoigne l'organisation du « challenge » AVEC qui depuis 2012 s'intéresse à la représentation dimensionnelle des émotions de manière continue. Dans un premier temps, le « challenge » proposait d'estimer ces dimensions à partir des modalités audio et vidéo à l'aide d'une régression [51]. Des modalités supplémentaires apparaissent à partir de l'édition de 2015 avec plusieurs signaux physiologiques [52]. Les résultats de la « baseline » des éditions de 2015 [52] et 2016 [53] montrent que la régression est plus performante en fusionnant les modalités plutôt que d'en sélectionner une seule, ce qui laisse entendre une complémentarité entre les modalités pour rendre compte des dimensions émotionnelle.

Défis de la vie réelle

Les premiers travaux sur l'analyse automatique des expressions faciales se sont basés sur des données acquises dans l'environnement contrôlé du laboratoire : le sujet reproduit une consigne précise de manière délibérée, le visage est de face et la luminosité est plus ou moins uniforme. L'analyse de données acquises dans un contexte réaliste apporte un certain nombre de défis à relever [38, 12] :

- Variation de pose de la tête,
- Variation d'illumination,
- Occlusions,
- Mouvements de la bouche non expressifs dûs à la parole,
- Expressions subtiles (de faible intensité) ou mélangées (non connues du système).

L'analyse et la reconnaissance d'expressions faciales dans les conditions de la vie réelle (ou conditions « in-the-wild ») est une problématique assez récente. Depuis 2013, le « challenge » EmotiW [54] propose aux chercheurs du monde entier de relever les

défis du contexte « in-the-wild » sur une base commune avec des données audio-visuelles dans le cadre de la reconnaissance d'expressions faciales. Un nouveau « challenge » « in-the-wild » a fait son apparition plus récemment, répondant au nom de « EmotioNet Challenge » et se décomposant en deux sous-parties : la reconnaissance d'unités d'action et la reconnaissance d'expressions faciales [55].

1.2.2 Bases de données d'expressions faciales

Les bases de données sont incontournables pour le développement de systèmes capables d'analyser automatiquement les expressions faciales. Avant d'élaborer une base de données, il est important de définir le contexte applicatif. Si les données ne sont pas pertinentes vis-à-vis du contexte applicatif, le système entraîné sur ces données n'aura aucune chance d'être performant en conditions réelles. De plus, si l'on veut être en mesure de généraliser au mieux les résultats du système, il faut se doter de données les plus diverses possibles au regard du contexte applicatif.

L'analyse automatique des expressions faciales a pris son envol dans les années 90 suite aux progrès réalisés en détection automatique du visage ou en suivi de points caractéristiques du visage. Des bases de données d'expressions faciales ont alors commencé à être accessibles à la communauté scientifique [56, 57, 58]. Il est à noter que certaines de ces bases de données sont toujours utilisées de nos jours comme banc d'essai pour pouvoir se comparer à d'autres méthodes. Ces premières bases de données sont constituées d'expressions posées, c'est-à-dire que le sujet effectue les expressions de manière délibérée en reproduisant une consigne.

Des études ont montré qu'il existe des différences entre les expressions posées et les expressions spontanées, ces dernières étant les expressions qu'une personne va faire naturellement dans la vie courante. Par exemple, Ekman a montré que certains muscles ne peuvent être sollicités délibérément lors des expressions de la colère, de la peur et de la tristesse [7]. Par la suite, d'autres études se sont concentrées sur le sourire [59, 8, 9] et les mouvements des sourcils [10]. Il est montré que les différences se trouvent dans l'intensité et la dynamique de l'expression. Un système entraîné sur des expressions posées verra donc ses performances amoindries lorsqu'il sera testé sur des expressions spontanées. Il y a donc un réel besoin de bases de données d'expressions spontanées. Elles commencent à apparaître quelques années après les premières bases de données d'expressions posées [60, 61] et depuis de nouvelles bases sont rendues disponibles quasiment tous les ans.

A notre connaissance, l'état de l'art des bases de données d'expressions faciales n'a pas été reporté de manière très complète dans les articles récents : sont passées en revue une quinzaine de bases données dans [62], une vingtaine dans [63, 64] et une trentaine

dans [65]. Dans le cadre de cette thèse, nous avons fait l'acquisition de notre propre base de données d'expressions faciales et un travail d'état de l'art a été effectué, nous dressons donc dans cette sous-section un état de l'art sur ce sujet. Nous décrivons chaque étape de l'élaboration d'une base de données d'expressions faciales en prenant en compte une cinquantaine de bases de données (voir Annexe A). Nous regroupons les différentes caractéristiques des bases de données dans six catégories : population, modalités, matériel d'acquisition, conditions expérimentales, protocole expérimental et annotations.

Population

La forme et la texture du visage varient avec le sexe, l'âge et le groupe ethnique : cela correspond aux différences individuelles. Par exemple, l'ouverture moyenne des yeux diffère de façon prononcée entre les Asiatiques et les Caucasiens. Si l'on veut développer une méthode robuste à ces différences individuelles, il est essentiel que la base de données contienne le plus large panel de groupes ethniques et une bonne distribution de l'âge et du sexe des sujets, c'est-à-dire la plus grande variabilité interpersonnelle possible.

La plupart des bases de données ont une population de moins de 50 sujets. Certaines ont une population d'une centaine de sujets, telles que CK [58], BU-3D FE [66], BU-4D FE [67], Bosphorus [68], RU-FACS [69] et BINED [70], voire plus de 300 sujets comme dans Multi-PIE [71]. Les bases de données composées d'extraits de programmes télévisés ou de films ont une population allant d'une centaine de sujets (Belfast Naturalistic [72], VAM [73]) à 330 sujets (AFEW [74]). Une autre approche intéressante pour avoir une large population est le « crowdsourcing » où les sujets sont recrutés via une campagne en ligne pour être filmés chez eux avec leur « webcam ». Nous n'avons recensé que deux bases de données utilisant cette technique : AM-FED [75] et Vinereactor [76].

La répartition hommes/femmes est souvent comprise entre 40/60 et 60/40, quelques exceptions existent cependant. La plus notable est la base JAFFE [56] qui est composée uniquement de femmes. Les bases CK [58], Belfast Naturalistic (HUMAINE) [72], UT Dallas [61] et CAS(ME)2 [77] sont composées en majorité de femmes (plus de 70%). Les bases Multi-PIE [71], NVIE [78] et ICT-3DRFE [79] sont quant à elles composées en majorité d'hommes (plus de 70%).

Concernant la répartition des âges, l'information n'est pas toujours disponible dans les articles associés aux bases de données. Nous pouvons toutefois faire ressortir deux tendances parmi les bases de données fournissant cette information : celles ne contenant que des sujets jeunes (entre 18 et 30 ans) et celles contenant une grande variété d'âges (entre 18 et 60 ans). Les exceptions notables sont Radboud Faces [80] et AFEW [74] qui contiennent aussi des sujets enfants.

L'information sur les groupes ethniques n'est pas non plus toujours présente dans les articles associés aux bases de données. Il ressort que la plupart des bases de données sont composées de plusieurs groupes ethniques avec une majorité de sujets caucasiens. Il existe aussi quelques bases de données ne contenant que des sujets asiatiques : JAFFE [56], NVIE [78], CAS(ME)2 [77] et CHEAVD [81]. Nous avons aussi relevé trois bases de données ne contenant que des sujets caucasiens : D3DFACS [82], MPI [65] et DynEmo [83]. Enfin, la base BAUM-1 [84] ne contient que des sujets Turques.

Modalités

Les toutes premières bases de données d'expressions faciales ne proposent que la modalité visuelle avec des images du visage (JAFFE [56], KDEF [57]) ou des vidéos du visage (University Of Maryland [85], CK [58]).

Par la suite, la fusion de modalités a connu un intérêt grandissant, en particulier l'association de la vidéo avec l'audio ou des signaux physiologiques. Cela a commencé en 2000 avec la base Belfast Naturalistic [72] constituée d'extraits de programmes télévisés, les signaux disponibles sont donc la vidéo et l'audio. A la même période, les signaux physiologiques ont été utilisés pour étudier le sourire, ce qui a fait l'objet de la création d'une base de données combinant de la vidéo et des signaux électromyographique faciaux (EMG) [60].

Dans les années 2000, l'analyse bi-modale audio/vidéo connaît un certain engouement [39] et les bases de données de ce type commencent à émerger (projet HUMAINE [86], GEMEP [87], IEMOCAP [88], VAM [73]). Il continue à en apparaître de nos jours : SEMAINE [89], MPI [65], AFEW [74], B3D(AC)2 [90], CAM3D [91], AViD-Corpus [92], BAUM-1 [84], CHEAVD [81] et GFT [93].

Nous avons relevé un nombre moins important de bases de données contenant de la vidéo et des signaux physiologiques (Enterface [94], DEAP [95], et BioVid Emo [96]), ce qui n'est pas surprenant étant donné que l'acquisition de signaux physiologiques demande un matériel très spécifique et est intrusive pour le sujet. Certaines bases de données sont allées jusqu'à combiner la vidéo, l'audio et les signaux physiologiques (MAHNOB-HCI [97] et RECOLA [98]), devenant alors des outils très précieux pour l'analyse multimodale des émotions.

En parallèle, des études ont porté sur l'analyse des émotions en combinant les expressions faciales et les mouvements du corps [99, 42]. Les bases de données incluant ces deux types de données utilisent 2 caméras ou plus. Dans le cas de GEMEP [87], FABO [100], BINED [70] et DynEmo [83], le sujet est seul ; dans le cas d'EmoTABOO [101] et IEMOCAP [88], deux sujets se font face et interagissent.

Une autre modalité visuelle qui connaît un intérêt de la part de la communauté scientifique est la modélisation 3D du visage. La première base de données à la proposer est BU-3D FE [66], où les modèles 3D sont statiques. Très vite, la même équipe de recherche élabore une nouvelle base de données, BU-4D FE [67], avec des modèles 3D dynamiques, c'est-à-dire une vidéo 3D du visage. Les bases Bosphorus [68] et ICT-3DRFE [79] sont restées sur des modèles 3D statiques. La base PICS - Stirling ESRC 3D Face Database [102] propose des modèles 3D statiques en plus de vidéos 2D d'expressions faciales. Les bases de données de vidéos 3D continuent à apparaître récemment : CAM3D [91], D3DFACS [82], B3D(AC)2 [90] et BP4D-Spontaneous [103].

Une dernière modalité qui a été explorée pour l'analyse des expressions faciales est l'infrarouge : seule la base NVIE [78] propose des vidéos d'expressions faciales à la fois dans le visible et l'infrarouge.

Matériel d'acquisition

Nous nous intéressons ici au matériel d'acquisition de l'image et de la vidéo.

Pour l'acquisition 3D, plusieurs techniques sont employées. Dans BU-3D FE [66], 6 caméras sont utilisées pour construire le modèle 3D. La même équipe de recherche a utilisé ensuite deux caméras stéréo et une caméra pour la texture dans les bases BU-4D FE [67] et BP4D-Spontaneous [103]. Dans la base CAM3D [91] l'information 3D est contenue dans une carte de profondeur obtenue avec une Microsoft Kinect.

Pour l'acquisition 2D, les caractéristiques à prendre en compte sont la résolution et le taux d'images par seconde. La résolution est souvent dans les alentours de 720x576. Dans certaines bases, telles Oulu-Casia [104], VAM [73], AM-FED [75] et Vinereactor [76], la résolution de la caméra est plus faible, dans les alentours de 320x240. Au contraire, on trouve une haute résolution de l'ordre de 1024x768 dans les bases FABO [100], BINED [70] et DISFA [105]. Le taux d'images par seconde est souvent compris entre 25 et 30, une caractéristique que l'on retrouve dans la plupart des caméras grand public. Quelques bases de données, telles FABO [100], MUG [22], DISFA [105] et AM-FED [75], sont composées de vidéos avec un taux d'images par seconde faible, compris entre 14 et 20. Pour des études sur la dynamique des expressions, il est préférable d'avoir un taux d'images par seconde élévé. Les bases de données MPI [65], D3DFACS [82], MAHNOB-HCI [97] et IEMOCAP [88] proposent des vidéos acquises avec un taux d'images par seconde de 50, 60, 60 et 120 respectivement.

Certaines bases de données font l'acquisition du visage simultanément depuis plusieurs points de vue, ce qui permet d'étudier les expressions faciales avec une variation de la pose de la tête. Deux caméras sont utilisées dans les bases MMI [106] (face et pro-

fil), ADFES [107] et BAUM-1 [84] (face et 45° pour les deux). D'autres bases utilisent un nombre impair de caméras pour avoir la pose de la tête des deux côtés : 3 caméras sont utilisées (face et +/-23°) dans la base MPI [65] ; 5 caméras (0°, +/-45°, +/-90°) sont utilisées dans KDEF [57] et Radboud Faces [80] ; 9 caméras (-90° :22,5° :90°) sont utilisées dans UT-Dallas [61]. La base de données Multi-PIE [71] propose quant à elle 15 vues différentes du visage : 13 au niveau du visage avec un pas de 15° et 2 en hauteur. Bien que cette base soit destinée aux problèmes de reconnaissance de visage ou d'identification, elle a été utilisée pour la reconnaissance d'expressions faciales robuste à la pose ou à l'illumination. Elle ne contient pourtant que 2 expressions prototypiques parmi ses 6 expressions (neutre, dégoût, surprise, sourire, yeux plissés et cri). Autrement, les bases de données 3D sont privilégiées pour analyser les expressions faciales avec une variation de la pose de la tête.

Pour les bases de données combinant les expressions faciales et les mouvements du corps, plusieurs caméras sont aussi nécessaires. Dans les bases FABO [100] et EmoTABOO [101], 2 caméras sont utilisées pour avoir le visage de face et le haut du corps. Dans la base GEMEP [87], 3 caméras sont utilisées pour avoir le visage de face et le corps en entier de face et de profil. Dans la base RU-FACS [69], 4 caméras sont utilisées pour avoir le visage de face, le corps avec deux angles de vue et le corps en contre-plongée. Dans la base MAHNOB-HCI [97], 5 caméras sont dédiées à l'acquisition du visage avec plusieurs angles de vue et 1 caméra est dédiée à l'acquisition du haut du corps. Dans la base IEMOCAP [88], une technique de capture de mouvement est mise en œuvre à l'aide de 8 caméras, ce qui en fait une base unique.

Conditions expérimentales

La grande majorité des bases de données a été acquise dans un laboratoire. Dans la plupart des cas un fond uni est installé derrière le sujet et ainsi la détection du visage dans l'image est facilitée, mais il existe quelques bases de données où l'arrière-plan est simplement l'environnement du laboratoire. Des cas particuliers sont à noter cependant. Les bases de données Belfast Naturalistic [72], VAM [73] et CHEAVD [81] consistent en des extraits de programmes télévisés, l'arrière-plan est donc varié. De même, la base EmoTV [108] contient des extraits d'interviews télévisées tournées en extérieur. Dans le même esprit, les bases AFEW [74] et SFEW [109] contiennent des extraits de films et fournissent donc une grande variabilité d'arrière-plans. Enfin, les bases de données AM-FED [75] et Vinereactor [76] ont été construites grâce au « crowdsourcing », ce qui veut dire que les sujets ont été filmés chez eux grâce à leur « webcam », elles offrent donc aussi une grande variabilité d'arrière-plans.

L'illumination du sujet joue aussi un rôle important dans les conditions expérimentales. En effet, si l'illumination est uniforme et constante, il est impossible de pouvoir tester la robustesse d'une méthode face à ce paramètre. Dans des applications de la vie réelle, il est évident qu'il faut faire face à une grande variabilité d'illumination. On retrouve trois grandes catégories par rapport à cette caractéristique : illumination uniforme, illumination ambiante et illumination variée. La dernière catégorie s'applique notamment aux bases AM-FED [75], Vinereactor [76], AFEW [74] et SFEW [109]. Dans les bases Oulu-CASIA [104] et NVIE [78], les expressions faciales sont acquises sous 3 conditions d'illumination différentes. Dans la base Multi-PIE [71], 19 conditions d'illuminations sont disponibles. La base ICT-3DRFE [79] propose quant à elle un modèle 3D du visage sur lequel on peut faire varier l'illumination librement, obtenu à l'aide d'un « light stage » avec 156 LEDs.

Un autre aspect à prendre en compte pour tester la robustesse d'une méthode d'analyse d'expressions faciales est la présence ou non d'occlusions du visage. Une grande partie des articles associés aux bases de données ne fournissent pas d'information à ce propos, ou alors il est précisé qu'aucune occlusion n'est autorisée lors de l'acquisition. La base Bosphorus [68] propose quant à elle des acquisitions avec des occlusions au niveau des yeux et de la bouche, avec le port de lunettes et avec des cheveux sur le visage. Dans le cas des bases Smile Database [60], Enterface [94] et RECOLA [98], il y a une occlusion sur tous les sujets due aux matériel d'acquisition des signaux physiologiques dans les deux premières bases et au micro dans la dernière base. La base RECOLA [98] contient d'autres occlusions avec notamment le passage de la main du sujet sur le visage ou des cheveux sur le visage. Le même type d'occlusions naturelles se trouvent dans la base CAM3D [91].

Protocole expérimental

Comme nous l'avons dit en introduction de cette sous-section, une attention toute particulière est portée sur le réalisme selon que les bases de données contiennent des expressions posées ou spontanées. C'est d'ailleurs très souvent l'argument mis en avant lors de la présentation de la base de données. Nous allons donc ici décrire les différents protocoles expérimentaux pour obtenir les expressions faciales en séparant les expressions posées des expressions spontanées. Nous terminons par les bases de données « in-the-wild » qui fournissent des données permettant de répondre aux défis de la vie réelle.

Expressions posées On trouve majoritairement deux protocoles pour obtenir des expressions posées : reproduction libre et reproduction d'une consigne. Dans le premier

cas, le sujet est uniquement informé de l'émotion à reproduire et doit le faire de manière expressive sans aucune consigne supplémentaire. Les bases suivantes sont concernées : University Of Maryland [85], JAFFE [56], ICT-3DRFE [79], FABO [100] et BU-3D FE [66]. Pour les deux dernières, le sujet peut recevoir une consigne d'un expert si cela est nécessaire. Dans le second cas (reproduction d'une consigne), soit le sujet a été préalablement entraîné à reproduire les expressions voulues, soit il est en présence d'un expert qui lui donne une consigne lors de l'acquisition. Les bases suivantes sont concernées : KDEF [57], CK [58], MMI [106], Multi-PIE [71], BU-4D FE [67], Bosphorus [68], Oulu-CASIA [104], Radboud Faces [80], MUG [22] et NVIE [78]. Il existe un troisième protocole pour obtenir des expressions posées : le portrait d'émotion. Dans ce cas, le sujet est amené à improviser sur un scénario riche émotionnellement. On retrouve ce protocole pour la première fois dans la base GEMEP [87], puis dans les bases MPI [65] et BAUM-1 [84]. La base IEMOCAP [88] reprend aussi ce protocole mais l'étend à un dialogue entre deux sujets. Dans la base B3D(AC)2 [90], le sujet visionne une vidéo censée induire une émotion spécifique et doit ensuite lire une phrase en utilisant l'intonation émotionnelle qu'il a perçue en visionnant la vidéo.

Les expressions réalisées dans ces bases de données contiennent toujours les expressions prototypiques d'Ekman correspondant aux émotions de base [3]. Les bases suivantes ne contiennent que les 6 émotions de base (colère, dégoût, peur, joie, tristesse et surprise) : University Of Maryland [85], BU-4D FE [67], Oulu-CASIA [104], Radboud Faces [80] (expression supplémentaire avec le mépris) et NVIE [78]. Pour certaines bases de données, il y a une acquisition supplémentaire pour le neutre : JAFFE [56], KDEF [57], BU-3D FE [66], MUG [22] et ICT-3DRFE [79]. D'autres bases de données ont cherché à diversifier le corpus expressif et émotionnel en ajoutant des émotions secondaires (par exemple l'ennui, l'embarras ou l'anxiété) : GEMEP [87], FABO [100], ADFES [107], BAUM-1 [84] (sans la tristesse). Enfin, nous avons relevé quatre bases de données se basant sur le système FACS pour définir les expressions à reproduire : CK [58], MMI [106], Bosphorus [68] et D3DFACS [82]. Les expressions sont alors définies par des combinaisons d'unités d'action, les expressions prototypiques d'Ekman [3] pouvant être obtenues avec certaines combinaisons.

Expressions spontanées Comme nous l'avons dit dans l'introduction de cette sous-section, la communauté scientifique a vite mis en évidence les limites des systèmes entraînés sur des expressions posées lorsqu'il s'agit de les tester dans des conditions réelles sur des expressions spontanées. On peut distinguer deux manières d'obtenir des expressions spontanées. D'un côté, on peut utiliser une méthode d'induction émotionnelle pour

induire au sujet un certain état émotionnel (par exemple avec le visionnage d'une vidéo censée induire une émotion spécifique). De l'autre, on peut chercher des données naturelles dans le sens où le protocole ne cherche pas à induire une émotion spécifique (par exemple avec une interaction humain-humain ou humain-machine).

La mise en place de méthodes d'induction émotionnelle n'est pas sans difficulté [70]. Il n'est pas possible de savoir objectivement quelle émotion est ressentie par le sujet et comment elle est perçue par un tiers, et de savoir à quel point l'expression faciale observée reflète l'émotion ressentie. Plus les expressions sont spontanées et naturelles, moins elles sont faciles à capturer, moins on dispose d'information sur l'état émotionnel du sujet et moins le protocole expérimental est reproductible. A contrario, l'acquisition d'expressions posées permet de garder un contrôle parfait sur le protocole expérimental et sa reproductibilité mais ne nous procure aucune information sur le véritable état émotionnel du sujet. L'idée est alors de trouver un compromis en exerçant un contrôle sur le protocole expérimental grâce à des tâches relativement standardisées qui permettent de collecter des informations sur l'état émotionnel du sujet tout en laissant le sujet réagir naturellement au contexte [70]. Il existe deux types de méthodes d'induction émotionnelle : les tâches passives et les tâches actives.

Les tâches passives restent les plus utilisées et consistent en le visionnage de vidéos ou d'images inductrices. Les vidéos et les images sont sélectionnées pour leur capacité à faire ressentir au sujet une émotion clairement identifiée. Pour les images, la base « International Affective Picture System » (IAPS) [110] est populaire parmi les psychologues pour l'étude des émotions et propose une grande variété d'émotions pouvant être induites par les images. Cette base a été sélectionnée comme inducteur émotionnel dans les bases Enterface [94] et BAUM-1 [84]. Concernant les vidéos inductrices, des extraits de programmes télévisés ou de films sont souvent utilisés. Les bases de données utilisant des images et vidéos inductrices sont les suivantes : Smile Database [60], UT-Dallas [61], MMI+ [111], MUG [22], NVIE [78], MANOB-HCI [97], DEAP [95], DISFA [105], CAS(ME)2 [77], BioVid Emo [96].

Une autre méthode d'induction émotionnelle est d'amener le sujet à exécuter des tâches actives. Contrairement au visionnage de vidéos ou d'images, le sujet est ici directement impliqué. Cette technique a été popularisée par la base de données BINED [70] qui combine des tâches actives et passives. Un exemple de tâche active consiste à faire passer un anneau le long d'un tube sans le toucher alors que le système est conçu tel qu'il est impossible de ne pas toucher l'anneau. Cette tâche est censée induire la frustration. Ce type de méthodes d'induction émotionnelle a été repris par quelques bases de données par la suite : UNBC-McMaster Shoulder Pain Expression Archive [112], DynEmo [83],

AVEC 2013 AViD-Corpus [92] et BP4D-Spontaneous [103]. Le cas de UNBC-McMaster Shoulder Pain Expression Archive [112] est particulier car il s'agit de faire faire une série d'exercices de mouvement de l'épaule à un patient dans un hôpital. C'est la seule base spécifiquement dédiée aux expressions de douleurs, on retrouve aussi des expressions de douleur dans la base BP4D-Spontaneous [103].

Les tâches actives et passives ont pour but d'induire des émotions spécifiques chez le sujet. Les six émotions de base d'Ekman [3] que l'on retrouve dans les bases d'expressions posées restent populaires et sont présentes dans une grande majorité de bases de données. Les bases de données d'expressions induites spontanées ne contenant que des émotions de base sont les suivantes : Enterface [94] (uniquement joie et tristesse), MMI+ [111] (uniquement dégoût, joie et surprise), MUG [22], NVIE [78] et CAS(ME)2 [77] (uniquement joie, colère, dégoût). Nous avons aussi relevé un nombre important de bases de données combinant les émotions de base d'Ekman [3] et des émotions secondaires : UT-Dallas [61], BINED [70], IEMOCAP [88] (neutre, joie, colère, tristesse et frustration), B3D(AC)2 [90] (sans le dégoût), CAM3D [91] (uniquement colère et surprise comme émotions de base), MAHNOB-HCI [97] (sans la colère et la surprise), DynEmo [83] (uniquement dégoût et peur comme émotions de base), BP4D-Spontaneous [103], BioVid Emo [96] (sans la joie et la surprise) et BAUM-1 [84]. Nous avons relevé une exception dans la description des émotions à induire. La base DEAP [95] fait figure d'exception car les émotions induites sont décrites par des dimensions émotionnelles au lieu des catégories émotionnelles. Le sujet doit visionner des clips musicaux censés induire un panel d'émotions balayant les quatre quadrants de l'espace bidimensionnel arousal-valence.

Il est à noter que certaines de ces bases de données sont acquises en deux temps avec des expressions posées et des expressions spontanées, ce qui peut être intéressant pour étudier les différences entre les deux. Les bases concernées sont IEMOCAP [88], MUG [22], NVIE [78], PICS - Stirling ESRC 3D Face Database [102], DISFA+ [113] et BAUM-1 [84].

Il existe d'autres méthodes d'induction émotionnelle qui ne visent pas forcément à induire une émotion spécifique. Elles reposent sur une interaction entre deux protagonistes. Il peut s'agir d'une interaction humain-humain où l'un des deux sujets est au courant de la procédure et cherche à diriger l'interaction pour la rendre riche émotionnellement (RU-FACS [69], EmoTABOO [101]), ou d'une interaction où les sujets sont amenés à interagir naturellement dans un contexte précis (RECOLA [98] et GFT [93]). Dans la base RECOLA [98], deux sujets interagissent par ordinateur interposé, alors que dans la base GFT [93], 3 sujets interagissent directement autour d'une table. L'interaction humain-machine a aussi été explorée dans les bases de données SAL [114] et SEMAINE

[89]. Le sujet interagit alors avec un agent virtuel commandé à distance par l'expérimentateur dans une configuration « wizard of oz ». L'expérimentateur peut choisir parmi 4 personnalités pour l'agent virtuel et ainsi orienter la teneur émotionnelle de l'interaction.

Expressions « in-the-wild » Dans un récent article d'état de l'art [12], l'analyse des expressions faciales dans un environnement « in-the-wild », donc non contraint et dans des conditions réelles, est identifié comme l'un des prochains défis majeurs que la communauté doit relever. Il existe 3 méthodes pour obtenir des expressions « in-the-wild » : corpus de vidéos/d'images d'expressions spontanées, corpus de vidéos/d'images d'expressions posées et « crowd sourcing ».

Les premières bases de données commençant à répondre à ce critère datent des années 2000 : Belfast Naturalistic [72], EmoTV [108] et VAM [73]. Elles sont composées d'extraits de programmes télévisés qui contiennent des expressions spontanées résultant d'une interaction humain-humain. Cependant, elles ne sont pas enregistrées dans des conditions que l'on pourrait qualifier de réelles. De plus, le nombre de sujets est de 48 pour VAM [73] et d'une centaine pour les deux autres ; il y a donc un manque de variabilité inter-personnelle. Récemment, la base de données Aff-Wild [115] reprend l'idée d'extraits de vidéos d'expressions spontanées, mais cette fois elles sont extraites de Youtube et offrent une plus grande variabilité de conditions expérimentales. Les vidéos sélectionnées montrent une personne qui affiche des expressions spontanées en regardant une vidéo, en pratiquant une activité, ou en réagissant à une blague ou une surprise. La base contient aussi des images recueillies sur Google Image.

La base de données AFEW [74] propose un corpus d'extraits de films. La sélection est effectuée automatiquement parmi 54 films en analysant les sous-titres pour sourds et malentendants, qui contiennent, entres autres, des informations sur le contexte émotionnel des acteurs et de la scène. Ensuite un annotateur va rentrer les informations à propos de chaque extrait. La base SFEW [109] est une version statique de la base AFEW [74] contenant des images extraites des vidéos de cette dernière. L'avantage de ces bases est de proposer une forte variabilité inter-personnelle avec une population de 330 sujets allant de 1 à 70 ans et une forte variabilité dans les conditions expérimentales. Cependant, les émotions disponibles restent les 6 émotions de base d'Ekman [3] et les expressions sont posées. Plus récemment, la base CHEAVD [81] combine un corpus de vidéos d'expressions posées avec des extraits de films et de séries télévisées et un corpus d'expressions spontanées avec des extraits de programmes télévisés.

Le « crowd sourcing » est utilisé pour construire une base de données offrant une grande variabilité inter-personnelle et de conditions expérimentales tout en étant com-

posée d'expressions spontanées. Le principe est de recruter les sujets par internet pour une étude et de les filmer directement chez eux via leur « webcam ». A notre connaissance, deux bases de données sont construites de cette manière : AM-FED [75] et Vinereactor [76]. Dans les deux cas, le sujet visionne une vidéo inductrice et sa réaction est enregistrée. Dans AM-FED [75], seul le sourire est induit. Dans Vinereactor [76], les émotions induites ne semblent être en relation qu'avec l'amusement puisque le sujet remplit par la suite un questionnaire pour noter à quel point la vidéo inductrice l'a amusé.

Annotations

Les annotations sont des méta-données donnant des informations bas-niveau (caractéristiques faciales, unités d'action du système FACS [6]) ou haut-niveau (labels émotionnels, dimensions émotionnelles). Cette étape est cruciale car selon l'application que l'on vise, certaines annotations serviront de vérité terrain et rendront donc la base de données attractive au reste de la communauté scientifique.

Caractéristiques faciales Pour décrire le visage, la plupart des annotations bas-niveau disponibles sont les points caractéristiques du visage extraits automatiquement (BU-3D FE [66], IEMOCAP [88] (marqueurs capture de mouvement), CK+ [116], MUG [22], DISFA [105], AM-FED [75], Vinereactor [76], GFT [93]). Dans la base NVIE [78], des points caractéristiques sont donnés dans le domaine du visible et de l'infrarouge. Les bases de données contenant des acquisitions 3D mettent à disposition les maillages 3D. Dans la base BP4D-Spontaneous [103], les annotations sont à la fois en 2D et en 3D. Les bases AFEW [74] et SFEW [109] ne proposent pas de description du visage avec des points caractéristiques mais incluent la pose de la tête dans les annotations. La pose de la tête est aussi inclue dans les annotations de la base GFT [93] en plus des points caractéristiques. La base AVEC 2013 AViD-Corpus [92], quant à elle, fournit la position du visage et des yeux ainsi que les caractéristiques dynamiques d'apparence LPQ-TOP basé sur les caractéristiques « Local Phase Quantisation » (LPQ). C'est la seule base de données fournissant des caractéristiques basées sur la texture et non la forme.

Unités d'action du système FACS Un autre type d'annotation bas-niveau que l'on retrouve assez souvent est le codage des expressions en unités d'action à l'aide du système FACS [6]. Les bases suivantes sont concernées : CK+ [116], MMI [106], D3DFACS [82], ICT-3DRFE [79], Smile Database [60], RU-FACS [69], MMI+ [111], DISFA [105], CAS(ME)2 [77], BP4D-Spontaneous [103], AM-FED [75], Aff-Wild [115], Vinereactor [76] et GFT [93]. Ces annotations sont essentielles pour pouvoir développer un système

de reconnaissance d'unités d'action car l'annotation prend beaucoup de temps et doit être réalisée par un expert.

Labels émotionnels L'annotation haut-niveau avec des labels émotionnels consiste à associer à chaque image ou vidéo un label correspondant à l'émotion affichée par le sujet. On trouve parfois une pondération à ce label, permettant d'apporter une information d'intensité. Ce type d'annotation est essentiel pour développer un système de reconnaissance d'expressions faciales. Dans certains cas, un seul label émotionnel est disponible (CK+ [116], BioVid Emo [96], BAUM-1 [84] avec pondération, AFEW [74], SFEW [109]), alors que dans d'autres cas au moins deux labels sont donnés, ce qui permet de décrire des émotions mélangées (JAFFE [56] avec pondération, EmoTABOO [101], BINED [70] avec pondération, SEMAINE [89] avec pondération, CAM3D [91], DynEmo [83], VAM [73], CHEAVD [81]). La base NVIE [78] donne deux labels : l'émotion réellement ressentie par le sujet et l'émotion supposément induite.

Dimensions émotionnelles Les vidéos sont annotées selon des dimensions émotionnelles en utilisant des outils tels que FEELTRACE [117] ou SAM [118]. Le couple arousal/valence est très souvent présent (GEMEP [87], RECOLA [98], AVEC 2013 AViD-Corpus [92], Belfast [72]), parfois augmenté d'une ou deux dimensions supplémentaires telles que l'intensité (NVIE [78], Radboud Faces [80], SAL [114]), le pouvoir (IEMOCAP [88], VAM [73], SEMAINE [89], MAHNOB-HCI [97], DEAP [95]) ou l'imprévisibilité (SEMAINE [89], MAHNOB-HCI [97]). La base UNBC-McMaster Shoulder Pain Expression Archive [112] reste un cas particulier puisqu'elle est annotée avec des dimensions spécifiques à la douleur.

Conclusion intermédiaire sur les bases de données

Nous venons de présenter dans cette section un état de l'art sur les bases de données d'expressions faciales. Pour développer et tester une méthode d'analyse d'expressions faciales, le choix de la base de données dépend des contraintes que l'on se fixe dans le contexte applicatif visé. Nous avons identifié six catégories pour regrouper les différentes caractéristiques des bases de données : population, modalités, matériel d'acquisition, conditions expérimentales, protocole expérimental et annotations.

Le choix de la population impacte la capacité de généralisation de la méthode d'analyse d'expressions faciales dans le cas où celle-ci est générique, *i.e.* indépendante de la personne. Il est important que la base de données contienne les variations inter-personnelles qui seront rencontrées dans le contexte applicatif pour que la méthode soit robuste à

ces variations. Les variations inter-personnelles concernent le sexe du sujet, son âge et son groupe ethnique. De manière générale, le nombre de sujets présents dans la base de données est une caractéristique importante à prendre en compte. En effet, plus il y a de données à tester, meilleure sera l'évaluation de la performance de la méthode.

Comme en témoignent certains articles sur l'état de l'art [39, 40], l'analyse multimodale de l'émotion connaît un intérêt grandissant, qui va de paire avec l'apparition des bases de données multimodales. Actuellement, les principales modalités disponibles sont l'image ou la vidéo 2D (expressions faciales et/ou mouvements du corps), l'image ou la vidéo 3D (expressions faciales), l'audio et les signaux physiologiques.

Le matériel d'acquisition a un impact si l'on cherche par exemple à étudier les expressions faciales sous plusieurs angles de vue. Il faut alors sélectionner une base de données faisant l'acquisition simultanée de l'expression faciales avec autant de caméras que d'angles de vue souhaités. Le même cas se présente lorsque l'on veut étudier conjointement les expressions faciales et les mouvements du corps. Selon le contexte applicatif, on peut souhaiter étudier les expressions faciales de manière très fine, ce qui implique une bonne illumination ainsi qu'un taux d'images par seconde et/ou une résolution importants, ou au contraire dans des conditions dégradées, ce qui implique une illumination variable ainsi qu'un taux d'images par seconde et/ou une résolution faibles.

Les conditions expérimentales prennent leur importance si l'on souhaite sortir du cadre déjà bien étudié où le sujet est face à la caméra sur un fond uni, sans bouger le corps ni la tête et sans parler.

Le protocole expérimental, quant à lui, joue un rôle important car c'est là que l'on détermine la nature des expressions faciales. Soit elles sont posées, *i.e.* réalisées de manière délibérée suivant une consigne, soit elles sont spontanées, *i.e.* non contrôlées. Le second cas est celui qui se rapproche le plus des conditions réelles, on trouve maintenant un grand nombre de bases de données d'expressions spontanées et il continue d'en apparaître de nos jours. Pour induire les expressions spontanées, deux familles de techniques sont utilisées jusqu'à présent : soit le sujet effectue des tâches actives ou passives censées induire une émotion spécifique, soit le sujet est en interaction avec un autre protagoniste, auquel cas aucune émotion spécifique n'est visée. Pour répondre aux problèmes du contexte de la vie réelle, les bases de données d'expressions « in-the-wild » proposent des corpus, pour la plupart audio-visuels, qui sont acquis dans des conditions réelles et non contrôlées ou utilisent le « crowd sourcing » pour avoir une large population dans des conditions expérimentales variées.

Les annotations fournies par la base de données sont aussi à prendre en compte. Selon que l'on veut développer un système de reconnaissance d'expressions faciales, de

reconnaissance d'unités d'actions ou d'estimation continue de dimensions émotionnelles (voir sous-section 1.2.1), certaines bases de données font gagner un temps précieux en ayant déjà annoté la vérité terrain correspondante. De plus, certaines bases de données contiennent aussi des caractéristiques faciales déjà extraites qui permettent de mettre rapidement en place un système d'analyse d'expressions faciales.

Le choix de la base de données est donc une étape cruciale pour développer et tester une méthode d'analyse d'expressions faciales car cela impacte directement les performances de la méthode dans son contexte applicatif. Dans notre cadre applicatif, le sujet affichera des expressions spontanées dans un environnement non contraint en termes de pose de la tête et de parole. La base RECOLA [98] répond à ces critères. Nous aurons aussi besoin de tester notre méthode d'adaptation du modèle spécifique à la personne sur des expressions posées pour pouvoir définir une vérité terrain et en tirer profit pour évaluer la performance. Nous choisissons alors la base MUG [22] car elle contient à la fois des expressions posées et des expressions spontanées pour les mêmes sujets, nous pourrons donc comparer notre méthode avec ces deux types d'expressions. Le modèle spécifique à la personne sur lequel nous nous basons est construit avec des expressions prototypiques et non prototypiques, ces dernières n'étant pas disponibles dans les bases de données que nous avons décrites. Nous avons donc fait l'acquisition de notre propre base de données avec ces expressions prototypiques et non prototypiques pour pouvoir faire des tests préliminaires sur la construction du modèle.

1.2.3 Système d'analyse des expressions faciales

Si l'on se réfère à la figure 1.4, nous venons de dresser dans la sous-section 1.2.2 l'état de l'art des bases de données qui sont nécessaires pour pouvoir développer et tester des méthodes répondant aux problématiques identifiées dans la sous-section 1.2.1. Nous revenons maintenant à l'analyse des expressions faciales en tant que telle et dans cette sous-section nous nous intéressons aux systèmes de reconnaissance d'expressions faciales et d'unités d'action. Le premier état de l'art dans le domaine date de 1992 [36]. La recherche en analyse automatique des expressions faciales n'en est alors qu'au commencement et peu d'études ont déjà été menées. Grâce aux avancées faites dans les années 90 en détection automatique du visage et en suivi automatique des points caractéristiques du visage, la recherche dans le domaine commence à réellement prendre son envol [5].

La communauté scientifique s'est d'abord penché sur les problèmes de reconnaissance d'expressions faciales et d'unités d'action, comme en témoignent les articles d'état de l'art du début des années 2000 [5, 37]. Les méthodes qui ont été développées pour répondre à

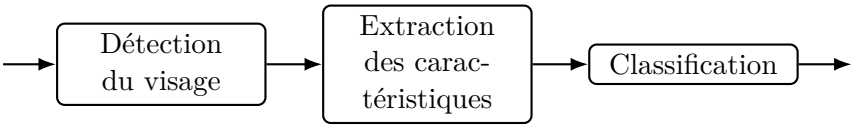


FIGURE 1.5 – Structure d'un système de reconnaissance d'expressions faciales et d'unités d'action.

ces deux problématiques suivent le même schéma d'apprentissage machine : extraction des caractéristiques après avoir détecté le visage puis classification (voir figure 1.5). Le lecteur peut se référer à [41] pour un récent état de l'art avec une évaluation des performances sur une base de données « in-the-wild ».

Comme nous l'avons dit dans l'introduction de la sous-section 1.2.2, les premiers systèmes sont développés sur des expressions posées dans l'environnement contrôlé d'un laboratoire. Des études ayant montré qu'il existe des différences entre les expressions posées et spontanées [8, 10, 9], les méthodes de reconnaissance d'expressions faciales et d'unités d'action ont été appliquées à des expressions spontanées [39]. La structure des systèmes développés reste identique à la figure 1.5.

Ces dernières années, un intérêt particulier porte sur l'apprentissage profond (« deep learning »). Le but des méthodes d'apprentissage profond est de trouver une modélisation haut-niveau des données à l'aide d'architectures basées sur les réseaux de neurones. Les étapes d'extraction des caractéristiques et de classification sont alors réalisées conjointement [119, 120]. Les réseaux de neurones peuvent aussi être utilisés uniquement pour l'extraction des caractéristiques et les caractéristiques ainsi obtenues sont ensuite utilisées pour l'apprentissage d'un classifieur [121]. La particularité de ces méthodes est de nécessiter une grande quantité de données pour l'apprentissage.

Détection du visage

Dans le cas où les données d'expressions faciales sont enregistrées dans un environnement contrôlé, la présence du visage dans les images est assurée et on peut connaître sa position a priori. Au contraire, dans des conditions proches de la vie réelle, la présence du visage n'est pas assurée pour toutes les images et sa position peut varier. Il faut alors pouvoir détecter le visage avant d'en extraire les caractéristiques. Dans [5], il est proposé de classer les méthodes de détection du visage selon deux approches : globale ou locale. Dans l'approche globale le visage est détecté comme un unique objet dans sa globalité, alors que dans l'approche locale le visage est détecté en deux temps. Dans un premier temps, des zones précises du visage sont détectées (yeux, nez, ...). Le visage est ensuite

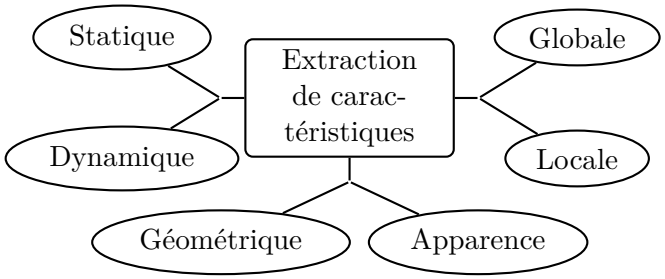


FIGURE 1.6 – Familles de méthodes d'extraction des caractéristiques faciales.

détecté grâce à la localisation de ces zones les unes par rapport aux autres. Le lecteur peut se référer à [122] pour un récent état de l'art sur la détection du visage.

La méthode proposée par Viola et Jones [123] reste l'une des plus populaires à ce jour pour la détection de visage. L'extraction des caractéristiques se fait à l'aide du descripteur de Haar. Une cascade de classifieurs est appliquée sur ces données pour détecter les points caractéristiques du visage.

Extraction des caractéristiques

Une fois que le visage est détecté, nous pouvons passer à l'étape cruciale qu'est l'extraction de caractéristiques. Avant de passer à cette étape, il peut être nécessaire d'effectuer un pré-traitement sur l'image si cela est requis par la méthode d'extraction des caractéristiques (par exemple une mise à l'échelle). Si l'on ne tient pas compte du bruit présent dans l'image, le visage contient deux informations : l'identité (ou morphologie) et l'expression. Le but de l'extraction de caractéristiques est donc d'extraire l'information d'expression tout en étant le plus invariant possible à l'identité. L'espace des images est transformé en un espace des caractéristiques où il devient possible de discriminer les expressions faciales les unes des autres. Nous pouvons différencier les familles de méthodes d'extraction de caractéristiques selon trois composantes (voir figure 1.6) :

- Statique vs. dynamique,
- Géométrique vs. apparence,
- Globale vs. locale.

La première distinction que l'on peut faire est selon la prise en compte ou non de l'information temporelle dans l'extraction de caractéristiques. Si elle est n'est pas prise en compte, on parle de méthode statique, sinon de méthode dynamique [124, 125, 126, 44].

La seconde distinction se fait sur la nature des données extraites. D'un côté, il y a les caractéristiques géométriques qui visent à extraire les informations de déformation faciale à partir de la détection et du suivi des points caractéristiques du visage [45, 44].

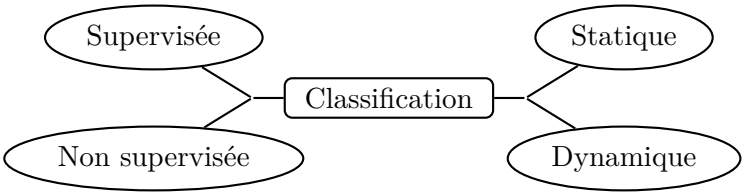


FIGURE 1.7 – Familles de méthodes de classification.

De l'autre côté, il y a les caractéristiques d'apparence qui se basent sur l'analyse de la texture et de ses variations dans l'image [124, 127]. Il existe aussi des méthodes hybrides alliant des caractéristiques géométriques et d'apparence [125, 126].

Enfin, nous faisons une troisième distinction selon que l'information extraite considère le visage en entier comme un tout (méthode globale) [45, 44] ou comme plusieurs zones indépendantes les unes des autres (méthode locale) [124, 127, 128]. Il existe aussi des méthodes hybrides alliant des caractéristiques globales et locales [126].

Classification

L'étape de classification cherche à déterminer à quelle classe appartiennent les caractéristiques extraites. Nous pouvons différencier les méthodes de classification selon deux composantes (voir figure 1.7) :

- Supervisée vs. non supervisée,
- Statique vs. dynamique.

Dans le cadre de la classification supervisée, les données d'apprentissage sont labellisées, c'est-à-dire que l'on sait à quelle classe appartient chacune de ces données. L'apprentissage du classifieur consiste à déterminer les frontières entre les différentes classes dans l'espace des caractéristiques en prenant en compte les labels des données d'apprentissage. Après l'apprentissage, le classifieur est capable d'associer un label à des données de test non labellisées. En classification non supervisée, les données d'apprentissage ne sont pas labellisées. Les caractéristiques extraites des données d'apprentissage sont partitionnées en sous-ensembles, idéalement sans intersections, en utilisant des métriques de similarité entre les données d'un même sous-ensemble.

De même que pour l'extraction de caractéristiques, la distinction entre classification statique et dynamique se fait selon que l'information temporelle est prise en compte ou non lors de l'apprentissage.

Parmi les classifieurs qui ont montré un grand intérêt pour la reconnaissance d'expressions faciales, on trouve la machine à vecteurs de support (« Support Vector Machine », SVM) [127, 129, 130, 45, 44], les plus proches voisins (« k-Nearest Neighbor »,

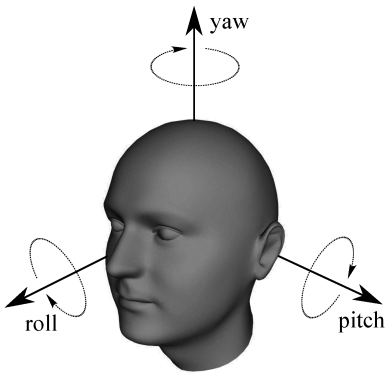


FIGURE 1.8 – Les angles de la pose de la tête (image extraite de [142]).

kNN) [131, 132, 133], les modèles de Markov [134, 135], la classification naïve de Bayes (« Naive Bayes », NB) [136, 137], la logique floue [138] et les réseaux de neurones [139, 140, 119, 141, 120], qui sont utilisés pour l'apprentissage profond (« deep learning »).

1.2.4 Reconnaissance d'expressions faciales robuste à la pose

La pose de la tête est définie par 3 angles : le « yaw », le « pitch » et le « roll » (voir figure 1.8). Lors de l'analyse des expressions faciales, la pose de la tête peut être assimilée à du bruit dont il faut s'abstraire car elle altère l'information expressive. Dans cette sous-section, nous nous intéressons aux méthodes de reconnaissance d'expressions faciales robuste à la pose. Cette problématique a été identifiée comme l'un des défis à relever pour pouvoir analyser les expressions faciales dans des conditions réelles [38, 12]. Nous pouvons distinguer six familles de méthodes pour répondre à cette problématique : les méthodes visant à extraire simultanément la pose et l'expression du visage, les méthodes effectuant une normalisation du visage par rapport à la pose avant l'extraction des caractéristiques, l'apprentissage de plusieurs classifieurs dépendant de la pose, l'apprentissage d'un unique classifieur avec plusieurs poses, l'apprentissage multi-vue et l'extraction de caractéristiques invariantes à la pose.

Extraction simultanée de la pose et de l'expression

Nous avons dit plus haut que le visage contenait à la fois l'information d'identité et d'expression. Si l'on se place dans un contexte où la pose de la tête varie, cette information vient alors s'ajouter aux deux précédentes. Si l'analyse des expressions faciales avec des variations de pose a été assez peu étudiée jusqu'à ces dix dernières années,

on trouve des travaux sur cette problématique dès la fin des années 90 [85]. L'idée est alors d'extraire simultanément l'information de pose et d'expression du visage. Dans [85], la méthode d'apparence dynamique et locale du flux optique est utilisée pour créer une modélisation paramétrique du mouvement dans l'image. Cette modélisation paramétrique locale permet de déterminer les mouvements rigides (pose de la tête) et non rigide (expression faciale). Les paramètres obtenus sont ensuite utilisés pour décrire et reconnaître les expressions faciales. L'estimation des mouvements repose sur une hypothèse de conservation d'illumination d'une image à l'autre, ce qui n'est pas toujours vérifié lorsqu'il y a des mouvements trop rapides. La robustesse à la pose a été testée qualitativement sur des extraits de films.

L'idée d'extraire simultanément l'information de pose et d'expression a été reprise par la suite mais n'a pas perduré dans le temps [143, 144]. La méthode de [143] se base sur des caractéristiques géométriques composées de 22 points caractéristiques. Le couplage entre la pose de la tête et l'expression faciale est modélisée par une fonction non linéaire. Les paramètres de pose et d'expression sont ensuite séparés à l'aide d'une décomposition en valeurs singulières normalisée (« Normalized Singular Value Decomposition », N-SVD). Pour augmenter la robustesse de l'estimation de ces paramètres, ces derniers doivent satisfaire certaines contraintes. La méthode a été testée sur des données synthétiques et sur une base de données maison. Dans [144], la méthode repose sur un modèle de variation d'intensité spécifique à la personne. La déformation faciale est décrite comme un changement dans l'intensité de plusieurs points fixés sur un modèle de forme rigide. Il faut alors apprendre le modèle d'intensité pour chaque personne avec une image frontale de chaque expression. Ensuite, les paramètres de pose et d'expression sont estimés simultanément en calculant, à l'aide d'un filtre particulaire, leur probabilité jointe sachant l'intensité observée sur les points caractéristiques du modèle. La méthode est testée sur une base maison avec les angles +/-20° pour le « yaw », +/-40° pour le « pitch » et +/-40° pour le « roll ».

Normalisation du visage par rapport à la pose avant extraction des caractéristiques

Une autre famille de méthodes qui a rapidement vu le jour propose de normaliser le visage par rapport à la pose, *i.e.* remettre le visage de face, avant l'extraction des caractéristiques dans le schéma classique de reconnaissance d'expressions faciales (voir figure 1.5).

Dans [145], la normalisation du visage par rapport à la pose se fait à l'aide d'une correspondance avec un modèle de tête cylindrique. La méthode proposée nécessite un

modèle initial de référence avec la pose de la tête correspondante. Pour les images suivantes, la pose de la tête est estimée et corrigée automatiquement. Cela n'a été appliqué qu'à la reconnaissance de clignement des yeux mais pas aux expressions faciales.

Ce type de méthode a été véritablement appliqué à l'analyse des expressions faciales dans [146]. Une fonction de correspondance entre les points caractéristiques 2D du visage non frontal et les points caractéristiques 2D du visage frontal est apprise à l'aide d'un modèle de régression par processus Gaussiens couplés (« Coupled Gaussian Process Regression », CGPR). La méthode se base sur la régression par processus Gaussiens (« Gaussian Process Regression », GPR) qui apprend la relation entre deux domaines, en l'occurrence entre les expressions avec de la pose et les expressions frontales. Une solution simpliste serait d'apprendre une unique GPR pour toutes les poses à la fois ou d'apprendre une GPR pour chaque pose indépendamment des autres. A la place, la CGPR reprend la solution consistant à apprendre une GPR pour chaque pose et propose ensuite d'apprendre la relation entre ces différentes GPRs. L'analyse discriminante de Fisher (« Linear Discriminative Analysis », LDA) est utilisée en premier lieu sur les points caractéristiques du visage pour réduire la dimensionalité. Dans cet espace réduit, la pose de la tête est estimée en calculant la probabilité qu'a l'expression d'appartenir à l'une des poses disponibles dans les données d'apprentissage. L'estimation de la pose est alors discrète. Une fois qu'elle est estimée, la CGPR est utilisée pour rendre le visage frontal. Si la pose estimée n'est pas disponible dans la base d'apprentissage, le visage frontalisé est calculé par une combinaison linéaire des visages frontalisés obtenus avec les différentes GPRs apprises pour chaque pose disponible dans les données d'apprentissage, la pondération de la combinaison linéaire est obtenue à partir des probabilités calculées lors de l'estimation de pose. La reconnaissance d'expressions s'effectue ensuite classiquement avec une SVM. La méthode a été testée sur les bases BU-3D FE [66] et Multi-PIE [71] (comme la plupart des méthodes que l'on mentionne dans la suite) avec des poses allant de -45° à +45° pour le « yaw » et de -30° à +30° pour le « pitch ».

Plus récemment, le « challenge » FERA propose dans sa troisième édition de s'attaquer à la reconnaissance d'unités d'action et l'estimation de leur intensité sous plusieurs points de vue [147]. La « baseline » s'appuie sur la normalisation du visage par rapport à la pose pour rendre ensuite le calcul des caractéristiques invariant à la pose. La normalisation s'effectue grâce aux paramètres du modèle de forme de la méthode de suivi des points caractéristiques du visage « Cascaded Continuous Regression » (CCR) [148].

Apprentissage de plusieurs classifieurs dépendant de la pose

Le principe de l'apprentissage de plusieurs classifieurs dépendant de la pose est de construire autant de systèmes de reconnaissance d'expressions faciales (ou d'unités d'action) que de poses disponibles pour l'apprentissage, puis de sélectionner le système à utiliser pour la reconnaissance selon la pose du visage à tester.

Nous avons relevé plusieurs méthodes reprenant ce schéma avec des types de caractéristiques différents. Dans [149], 5 poses différentes sont testées (0°, 30°, 45°, 60° et 90° pour le « yaw »). Pour l'extraction de caractéristiques, les distances entre chaque point caractéristique du visage neutre et du visage expressif sont calculées, puis normalisées pour avoir une moyenne nulle et une variance unitaire. Plusieurs classifieurs sont ensuite testés, la SVM donnant les meilleurs résultats. Les résultats montrent que le taux de reconnaissance n'est pas le plus élevé pour les images frontales mais pour les images avec un « yaw » de 45°. Dans [150], 5 poses sont aussi testées (0°, 30°, 45°, 60° et 90° pour le « yaw »). Des caractéristiques d'apparence sont utilisées : le visage est découpé en 64 blocs puis un histogramme de caractéristiques LBP (« Local Binary Pattern ») est calculé pour chaque bloc et les différents histogrammes obtenus sont concaténés. Dans [151], 13 poses sont testées, allant de -90° à +90° pour le « yaw » avec un pas de 15°. Des caractéristiques hybrides (géométrique et d'apparence) sont utilisées. Les points caractéristiques du visage sont obtenus avec des modèles actifs d'apparence (« Active Appearance Model », AAM) dépendant de la pose et les caractéristiques d'apparence sont calculées autour des points caractéristiques avec une transformée discrète de cosinus (« Discrete Cosine Transform », DCT). Pour réduire la dimension des caractéristiques hybrides, une méthode de sélection des caractéristiques est appliquée. Le classifieur choisi est la SVM.

Apprentissage d'un unique classifieur avec plusieurs poses

L'apprentissage d'un unique classifieur avec plusieurs poses prend le contre-pied des méthodes décrites ci-dessus. Au lieu d'apprendre autant de classifieurs que de poses disponibles dans les données d'apprentissage, un seul classifieur est entraîné et est rendu robuste à la pose en incluant plusieurs poses pour chaque expression dans l'apprentissage.

Dans [152], 35 poses sont utilisées, combinant 7 poses différentes pour le « yaw », allant de -45° à +45° avec un pas de 15°, et 5 poses différentes pour le « pitch », allant de -30° à +30° avec un pas de 15°. Des caractéristiques locales d'apparence SIFT sont extraites, puis une matrice de covariance par région (« Regional Covariance Matrix », RCM) est calculée. La RCM est utilisée à la base pour la représentation d'image et a été appliquée à la détection d'objets. Cela permet de capturer les propriétés statistiques des

données tout en étant invariant à la translation, le changement d'échelle et la rotation. Le classifieur est ensuite entraîné avec des vecteurs extraits de la matrice RCM.

Dans [153], les auteurs proposent une extension de leur méthode de reconnaissance d'expressions faciales basées sur les forêts aléatoires conditionnelles avec comparaison par paire (« Pairwise Conditional Random Forest », PCRF) [125] pour la rendre robuste à la pose. Les caractéristiques utilisées pour construire les arbres de la PCRF sont hybrides (géométrique et d'apparence), où les caractéristiques géométriques sont des distances et des angles entre certains points caractéristiques du visage. Ensuite la dynamique des expressions est prise en compte en nourrissant les arbres de la PCRF avec des paires d'expressions. Pour rendre la PCRF robuste à la pose, plusieurs poses sont générées pour l'apprentissage et le modèle est conditionné par rapport à la pose, ce qui donne la méthode « Multi-View Pairwise Conditional Random Forest » (MVPCRF). Le modèle est entraîné avec 15 poses, combinant 5 poses différentes pour le « yaw » (0°, +/-17.5° et +/-35°) et 3 poses différentes pour le « pitch » (0°, +/-25°).

Apprentissage multi-vues

Dans les quatre familles de méthodes décrites précédemment, la relation qui existe entre les différentes poses n'est pas explicitement modélisée [154]. L'apprentissage multi-vues propose d'apprendre une variété des expressions à partir de plusieurs vues de telle sorte que les différentes vues d'une même expression soient regroupées dans une même zone dans l'espace créé. On retrouve souvent l'apprentissage multi-vues dans les méthodes développées ces dernières années pour la reconnaissance d'expressions faciales robuste à la pose ainsi que pour la reconnaissance de visage robuste à la pose [155].

La méthode proposée par [156] est l'analyse multi-vues généralisée (« Generalized Multiview Analysis », GMA), qui est une extension de l'analyse canonique des corrélations (« Canonical Correlation Analysis », CCA). L'avantage de cette méthode est de pouvoir se généraliser à des classes inconnues. La robustesse à la pose est testée sur 5 poses allant de 0° à 75° avec un pas de 15° sur le « yaw ». Plus récemment, on trouve d'autres méthodes d'apprentissage multi-vues. Dans [154], les processus Gaussiens sont adaptés à la problématique multi-vues (« Discriminative Shared Gaussian Process Latent Variable Model », DS-GPLVM). La méthode a été testée sur la base Multi-PIE [71] avec des poses allant de -30° à +30° pour le « yaw » avec un pas de 15°, mais aussi sur les bases SFEW [74] et Labeled Face Parts in the Wild (LFPW) [157]. Tout comme SFEW, la base LPFW contient des données dans des conditions « in-the-wild », mais vise comme problématique la reconnaissance de visage et non l'analyse d'expressions faciales. Ainsi, la robustesse à la pose est testée sur une plus grande variété de poses.

Dans [158], la variété est apprise sur plusieurs types de caractéristiques (BoF, CNN, LBP-TOP, audio) disponibles pour plusieurs vues. Ce travail s'inscrit dans le cadre du « challenge » EmotiW 2016 [159], les tests se font donc sur des données « in-the-wild » avec une grande variété de poses.

Extraction de caractéristiques invariantes à la pose

Nous avons relevé une dernière famille de méthode pour l'analyse d'expressions faciales robuste à la pose qui consiste à extraire des caractéristiques invariantes par rapport à la pose. A notre connaissance, une seule étude a porté sur une telle approche [160]. La particularité de la méthode est d'apprendre automatiquement les caractéristiques invariantes à la pose à l'aide de deux réseaux de neurones. Premièrement, un PCANet effectue un apprentissage non supervisé des caractéristiques du visage de face. Ensuite, un réseaux de neurones convolutionnel (« Convolutional Neural Network », CNN) apprend de manière supervisée une fonction permettant de passer des images avec de la pose aux images de face. Cet apprentissage est réalisé avec 35 poses différentes combinant 7 poses différentes pour le « yaw », allant de -45° à +45° avec un pas de 15°, et 5 poses différentes pour le « pitch », allant de -30° à +30° avec un pas de 15°.

Conclusion intermédiaire sur la reconnaissance d'expressions faciales robuste à la pose

Dans cette sous-section, nous avons présenté les techniques existantes pour la reconnaissance d'expressions faciales robuste à la pose. Bien que cette problématique ait été étudiée dès la fin des années 90 [85], ce n'est que récemment qu'elle connaît un intérêt grandissant. Cela est en phase avec la volonté de la communauté scientifique de relever les défis techniques liés à l'analyse d'expressions faciales dans un contexte « in-the-wild » [12]. Nous avons relevé six familles de méthodes pour l'analyse d'expressions faciales robuste à la pose : extraction simultanée de la pose et de l'expression, normalisation du visage par rapport à la pose avant extraction des caractéristiques, apprentissage de plusieurs classifieurs dépendant de la pose, apprentissage d'un unique classifieur avec plusieurs poses, apprentissage multi-vues et extraction de caractéristiques invariantes à la pose. Les méthodes ont généralement été testées sur des poses allant de -45° à +45° pour le « yaw » et de -30° à +30° pour le « pitch ». L'apprentissage multi-vue est la famille de méthodes qui connaît le plus d'intérêt ces dernières années.

Dans notre cadre applicatif, le sujet est dans un environnement non contraint en termes de pose de la tête et de parole. Nous aurons besoin d'un système de reconnaissance

d'expressions faciales robuste à la pose. Nous choisissons l'apprentissage d'un unique classifieur avec plusieurs poses pour la simplicité de sa mise en œuvre.

1.2.5 Modèles spécifiques à la personne

Une grande majorité de la recherche sur l'analyse des expressions faciale se concentre sur une modélisation générique, *i.e.* indépendante de la personne. Les méthodes vont alors chercher à apprendre un unique classifieur (voir figure 1.5) qui va se généraliser le plus possible à des sujets inconnus de la base d'apprentissage. Comparativement, peu de travaux proposent un modèle spécifique à la personne. Pourtant, des études ont montré que les modèles spécifiques à la personne sont plus performants que les modèles indépendants de la personne pour la reconnaissance d'expressions faciales [137, 161]. De plus, les modèles spécifiques à la personne peuvent s'avérer intéressant dans un contexte « in-the-wild » pour pouvoir analyser plus finement des expressions subtiles ou mélangées qui ne sont pas incluses dans la base d'apprentissage. Dans cette sous-section, nous présentons quelques méthodes pour construire un modèle spécifique à la personne. Nous pouvons distinguer trois approches : l'apprentissage de variété, l'adaptation de domaine et l'apprentissage faiblement supervisé.

Apprentissage de variété

L'objectif de l'apprentissage de variété est de trouver une fonction permettant de projeter l'espace de grande dimension des caractéristiques faciales dans un sous-espace de faible dimension.

Dans [131], la variété est apprise sur des caractéristiques géométriques constituées de points caractéristiques du visage. Pour ce faire, deux méthodes sont étudiées : l'imbrication localement linéaire (« Locally Linear Embedding », LLE) et l'imbrication de Lipschitz (« Lipschitz Embedding », LE). Le visage neutre est au centre de la variété et chaque expression s'étend sur une direction spécifique en partant du neutre. Étant donné que le visage neutre et l'intensité des expressions varient selon les sujets, l'alignement des variétés entre les sujets est nécessaire afin de fournir un cadre unifié pour l'analyse des expressions faciales. Les expérimentations montrent que la LLE est adaptée pour la représentation visuelle des expressions faciales mais ne peut pas atteindre un bon taux de reconnaissance, contrairement à la LE. Dans [132], la variété est apprise avec les projections de préservation locale (« Locality Preserving Projections », LPP) à partir de données d'image brute et de caractéristiques d'apparence, en l'occurrence les LBP. Un algorithme d'alignement est également nécessaire pour fournir un cadre unifié

pour l'analyse des expressions faciales. L'inconvénient de l'apprentissage de variété dans [131, 132] est qu'un nombre important de données labellisées est nécessaire, ce qui n'est pas le cas dans un contexte « in-the-wild ».

Dans [133], la variété est apprise avec les caractéristiques géométriques différentielles AMM (« Differential AMM Features », DAFs), qui sont calculées par la différence entre les caractéristiques AAM de l'expression et du visage neutre. Ainsi, l'alignement entre les sujets est effectué avant d'apprendre la variété, contrairement à [131, 132]. Le visage neutre est détecté avec le modèle différentiel de densité de probabilité d'expression faciale (« Differential Facial Expression Probability Density Model », DFEPDM).

Dans [14], la variété est apprise avec les caractéristiques géométriques AAM. Les 8 expressions nécessaires pour l'apprentissage sont synthétisées à partir du visage neutre. La particularité de l'approche est de représenter ensuite une expression par sa position relative aux 8 expressions synthétisées dans la variété. Il en résulte un espace expressif indépendant de la personne. Il n'y a donc pas besoin d'alignement entre les sujets.

Adaptation de domaine

L'adaptation de domaine cherche à s'attaquer au problème de la quantité de données labellisées requises pour l'apprentissage. L'adaptation de domaine repose sur l'idée que la distribution de l'espace des caractéristiques varie selon les domaines (dans notre cas, les sujets), tandis que les labels (dans notre cas, l'expression faciale) sont identiques.

L'apprentissage par transfert est une approche populaire pour l'adaptation de domaine dans le cadre de l'adaptation de sujet. Cela vise à extraire les connaissances d'un ou plusieurs domaines sources (les expressions faciales labellisées des sujets connus) et de les transférer vers un domaine cible (les expressions faciales du sujet à modéliser). Dans [162], l'apprentissage par transfert inductif et par transfert transductif sont utilisés pour la reconnaissance de la douleur. Dans le premier cas, les expressions cibles sont labellisées, mais seulement une petite quantité est requise, tandis que dans le second cas, les expressions cibles ne sont pas labellisées, ce qui est plus probable dans un contexte « in-the-wild ». L'apprentissage par transfert transductif consiste à pondérer les données sources à l'aide des distributions marginales de la source et de la cible, puis les utiliser pour construire le modèle cible. Dans [48], le même problème est abordé avec une machine de transfert sélectif (« Selective Transfer Machine », STM). Un classifieur indépendant de la personne est personnalisé en réévaluant la pondération des données sources en fonction de leur proximité avec les données cibles. Ces méthodes ont l'avantage de ne nécessiter que des données cibles non labellisées, ce qui facilite la tâche de la modélisation spécifique à la personne.

Plus récemment, une autre approche d'adaptation de domaine basée sur le processus Gaussien (« Gaussian Process », GP) a été proposée dans [163]. Le processus Gaussien expert du domaine (« Gaussian Process Domain Expert », GPDE) est introduit pour construire le modèle cible. La méthode est appliquée à l'adaptation de sujet à partir de sources multiples ; une faible quantité de données labellisées du sujet cible est nécessaire. Contrairement à l'apprentissage par transfert, les paramètres du classifieur ne sont pas réévalués pour s'adapter au domaine cible. Dans un premier temps, des GPs sont entraînés séparément sur chaque sujet source et le sujet cible. Ensuite, l'adaptation au sujet cible se fait en conditionnant le GP cible sur les GPs sources. L'efficacité de la méthode d'adaptation de sujet est testée pour la reconnaissance des AU.

Apprentissage faiblement supervisé

Afin de ne pas être tributaire de l'annotation des données spécifiques à la personne, une méthode faiblement supervisée, appelée détection d'affect personnel avec annotation minimale (« Personnal Affect Detection With Minimal Annotation », PADMA), est proposée dans [164]. L'idée de base est qu'un changement dans l'état affectif du sujet implique un changement dans l'expression du visage. Le sujet visionne des vidéos induisant des émotions spécifiques puis sélectionne un label correspondant à l'émotion globale ressentie pour chaque vidéo. Les caractéristiques faciales sont ensuite extraites et regroupées en plusieurs classes (« clusters ») pour créer des labels expressifs. Les groupes similaires sont identifiés et fusionnés de manière adaptative afin d'extraire les principales déformations du visage. Enfin, la relation entre les principales déformations du visage et les labels est apprise avec l'apprentissage par instances multiples basé sur l'association (« Association-Based Multiple Instance Learning », AMIL).

Conclusion intermédiaire sur les modèles spécifiques à la personne

La modélisation spécifique à la personne est un sujet de recherche peu exploré comparé à la modélisation générique, *i.e.* indépendante de la personne. A notre connaissance, seules trois familles de méthodes existent pour la modélisation spécifique à la personne : apprentissage de variété, adaptation de domaine et apprentissage faiblement supervisé, cette dernière ne faisant l'objet que d'une seule étude à ce jour [164]. L'inconvénient des méthodes existantes d'apprentissage de variété est le nombre important de données labellisées nécessaires pour l'apprentissage. Les méthodes d'adaptation de domaine viennent contrecarrer cet inconvénient en adaptant la distribution de l'espace des caractéristiques du sujet à modéliser avec la distribution de l'espace des caractéristiques de plusieurs

sujets connus, ce qui permet de modéliser le sujet à partir d'un nombre moins important de données. Selon les méthodes, les données du sujet à modéliser doivent être labellisées (la méthode est alors supervisée) ou non (la méthode est alors non supervisée).

Dans notre cadre applicatif, nous sommes amené à analyser des expressions non prototypiques. L'apprentissage de variété est adapté car il permet une modélisation continue des expressions faciales et nous pouvons ainsi analyser une grande variété d'expressions faciales. Au contraire, l'adaptation de domaine permet de paramétrer des classifieurs et est donc utilisée pour une analyse discrète des expressions faciales.

1.2.6 Contributions de la thèse

A notre connaissance, il n'existe pas de méthode pour apprendre une variété spécifique à la personne de manière non supervisée, *i.e.* sans connaissance *a priori* de la morphologie du sujet. La principale contribution de cette thèse porte sur un apprentissage non supervisé d'un modèle spécifique à la personne prenant la forme d'une variété apprise sur des caractéristiques géométriques. Nous reprenons le modèle spécifique à la personne développé dans [14], où le modèle est construit de manière faiblement supervisée puisque la seule information nécessaire *a priori* est le visage neutre du sujet. A partir du visage neutre sont synthétisées des expressions de base utiles à la construction du modèle. Nous appelons « modèle plausible » le modèle construit à partir d'expressions de base synthétisées. Comparativement à [14], nous utilisons 5 expressions de base au lieu de 8 et nous proposons une représentation du modèle invariante à la personne, appelée « espace des signatures ».

Notre première contribution est de construire le modèle plausible de [14] de manière non supervisée à l'aide d'une détection automatique du visage neutre.

A ce stade, le modèle plausible est construit sur des expressions de base synthétisées, il ne rend donc pas compte des déformations faciales réelles du sujet. Notre seconde contribution est d'adapter le modèle aux expressions de base réelles du sujet de manière non supervisée. Nous pouvons faire un parallèle avec l'adaptation de domaine dans le sens où nous avons un modèle pré-entraîné avec le visage neutre automatiquement détecté et les expressions de base synthétisées et nous l'adaptons à partir de données non labellisées. La différence avec l'adaptation de domaine est que l'on fait une analyse continue et non discrète des expressions faciales.

Pour réaliser cette adaptation, nous avons besoin d'un système de reconnaissance d'expressions faciales. Dans un environnement non contraint (variation de la pose de la tête, occlusions, sujet en train de parler, ...), les performances d'un tel système peuvent chuter s'il n'est pas assez robuste. Par conséquent, les performances de l'adaptation

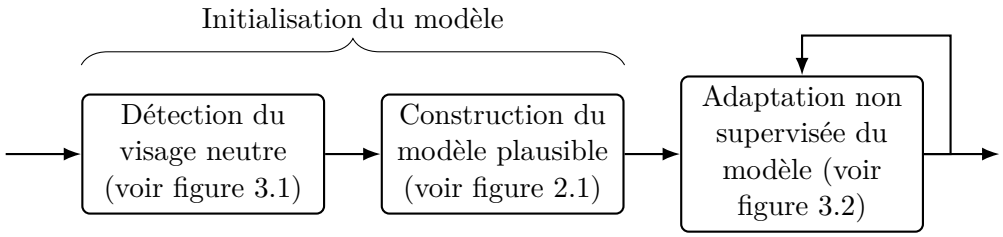


FIGURE 1.9 – Vue d'ensemble du système de construction non supervisée du modèle expressif spécifique à la personne. Dans un premier temps, le modèle est initialisé. Pour cela, le visage neutre du sujet est d'abord détecté automatiquement (voir figure 3.1). Ensuite, le modèle est construit à partir de ce visage neutre et des expressions de base synthétisées (voir figure 2.1), ce qui clôt l'initialisation du modèle. Nous passons ensuite à l'adaptation non supervisée du modèle (voir figure 3.2).

peuvent aussi chuter. Notre troisième contribution est d'assurer la robustesse de notre méthode d'adaptation du modèle dans un environnement non contraint à l'aide d'un système de reconnaissance d'expressions faciales robuste à la pose.

La figure 1.9 donne une vue d'ensemble de notre système pour construire de manière non supervisée le modèle expressif spécifique à la personne. Dans un premier temps, le modèle est initialisé. Pour cela, le visage neutre du sujet est d'abord détecté automatiquement (voir figure 3.1). Ensuite, le modèle est construit à partir de ce visage neutre et des expressions de base synthétisées (voir figure 2.1), ce qui clôt l'initialisation du modèle. Nous passons ensuite à l'adaptation non supervisée du modèle (voir figure 3.2).

Chapitre 2

Modèle expressif et reconnaissance d'expressions robuste à la pose

Sommaire

Comme nous l'avons mentionné dans l'introduction et dans la sous-section 1.2.6, la principale contribution de cette thèse, qui sera présentée dans le chapitre 3, est de construire un modèle expressif spécifique à la personne de manière non supervisée, *i.e.* sans information *a priori* de la morphologie du sujet. Pour ce faire, nous nous basons sur le modèle proposé dans [14] et nous l'adaptons automatiquement de manière non supervisée à la morphologie et aux expressions réelles du sujet.

Dans ce chapitre, nous proposons deux outils qui seront nécessaires pour notre méthode d'adaptation. Dans la section 2.1, nous présentons le modèle expressif spécifique à la personne basé sur [14]. Dans le cadre de la thèse, nous nous sommes réapproprié ce

modèle pour répondre à notre problématique de modélisation spécifique à la personne de manière non supervisée et nous avons apporté quelques modifications. Ensuite, dans la section 2.2, nous présentons notre méthode de reconnaissance d'expressions faciales robuste à la pose qui sera utilisée à deux reprises dans l'adaptation non supervisée du modèle spécifique à la personne : pour la détection du visage neutre lors de l'initialisation du modèle et pour détecter les expressions de base du modèle lors de l'adaptation en tant que telle.

2.1 Modèle expressif spécifique à la personne

Nous reprenons les travaux de [14] sur la représentation invariante des expressions faciales pour définir le modèle expressif spécifique à la personne. Cette représentation invariante repose sur le fait que les déformations faciales des expressions sont organisées de la même manière les unes par rapport aux autres, indépendemment de la personne. Ainsi, il est possible de créer un modèle spécifique à la personne tout en ayant une représentation indépendante de la personne. La figure 2.1 illustre les étapes de la construction du modèle spécifique à la personne basé sur [14] de manière faiblement supervisée. Le modèle est construit à partir du visage neutre et de plusieurs expressions de base dont on connaît la localisation des points caractéristiques. Dans [14], 8 expressions de base sont utilisées, alors que nous proposons dans cette thèse d'en utiliser 5. Les expressions de base sont synthétisées à partir du visage neutre, ce qui rend la construction du modèle faiblement supervisée. Une variété est ensuite apprise sur les points caractéristiques de ces expressions avec une analyse en composantes principales (« Principal Component Analysis », PCA). Nous obtenons alors un espace PCA spécifique à la personne. Grâce à une tessellation de Delaunay calculée sur les projections des expressions de base dans l'espace PCA, nous transformons l'espace PCA en une représentation indépendante de la personne, que nous appelons « espace des signatures ». Cette représentation indépendante résulte du fait que la tessellation de Delaunay donne la même structure de simplexes quel que soit le sujet. Une nouvelle expression est ensuite caractérisée par ce que nous appelons sa « signature », qui correspond aux coordonnées de sa projection dans l'espace des signatures. Nous proposons un calcul de la signature différent de celui proposé dans [14].

Le processus complet de construction faiblement supervisée du modèle expressif est détaillé dans la sous-section 2.1.1. Puis le calcul de la signature d'une expression est explicité dans la sous-section 2.1.2.

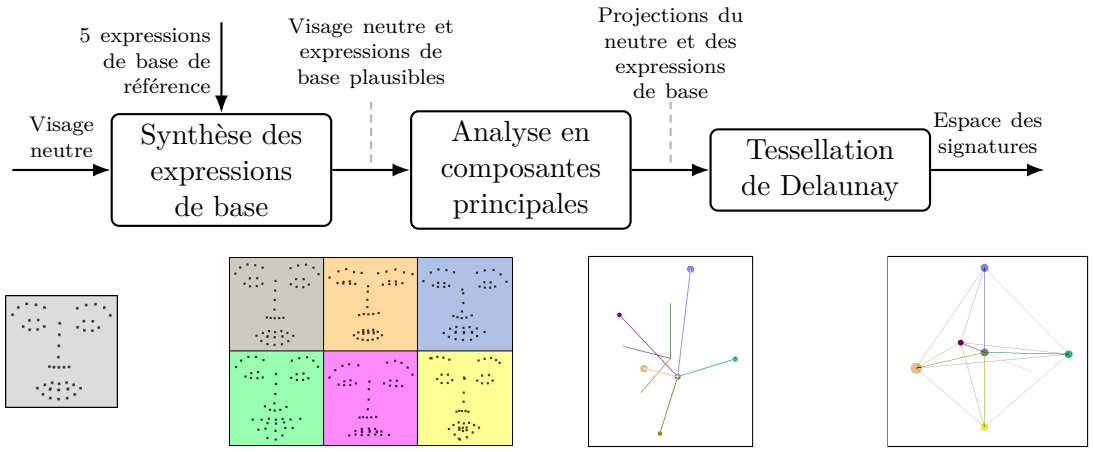


FIGURE 2.1 – Construction faiblement supervisée du modèle expressif spécifique à la personne basé sur [14]. La seule information nécessaire *a priori* est l'ensemble des points caractéristiques du visage neutre du sujet. La première étape est de synthétiser les expressions de base plausibles en déformant le visage neutre du sujet avec les déformations faciales d'un sujet de référence. Une analyse en composantes principales (« Principal Component Analysis », PCA) est ensuite calculée sur l'ensemble des points caractéristiques du visage neutre et des expressions de base plausibles, puis les 3 premiers axes sont conservés. Enfin, une tessellation de Delaunay est appliquée sur l'ensemble des projections des expressions de base plausibles dans l'espace PCA. Cela permet d'obtenir la structure du modèle qui décrit l'organisation des expressions de base les unes par rapport aux autres dans l'espace PCA. Cette structure est indépendante de la personne. Nous pouvons donc transformer l'espace PCA, qui est spécifique à la personne, en une représentation indépendante de la personne appelée espace des signatures.

2.1.1 Construction faiblement supervisée du modèle

Synthèse des expressions de base

Dans le but d'avoir un modèle faiblement supervisé, il est proposé dans [14] de synthétiser les expressions de base du modèle à partir du visage neutre. Ainsi, le visage neutre est la seule information nécessaire *a priori*. La synthèse des expressions se fait à l'aide d'une fonction de déformation affine par morceau (« Piece-Wise Affine Warping ») [165]. Le principe est de transférer les déformations faciales correspondant aux 5 expressions de base d'un sujet de référence sur le visage neutre du sujet dont on veut synthétiser les expressions de base. Les expressions ainsi synthétisées sont appelées « expressions de base plausibles ». Avant de passer au calcul de la PCA, un alignement de Procruste est appliqué sur l'ensemble des expressions de base (neutre inclus).

Apprentissage de la variété (analyse en composantes principales)

L'étape suivante est d'apprendre une variété à l'aide d'une PCA sur les points caractéristiques du visage neutre et des 5 expressions de base plausibles.

Soient $n_e \in \mathbb{N}$ le nombre d'expressions de base du modèle (neutre inclus) et $n_p \in \mathbb{N}$ le nombre de points caractéristiques utilisés pour décrire le visage. Pour tout $i \in [\![1, n_e]\!]$, on note $X^i \in \mathbb{R}^{2n_p}$ le vecteur contenant les points caractéristiques de la i-ième expression de base après alignement. On note $\overline{X^e} \in \mathbb{R}^{2n_p}$ la moyenne de l'ensemble des points caractéristiques des expressions de base du modèle (neutre inclus) après alignement : $\overline{X^e} = moyenne_{i \in [\![1,n_e]\!]}(X^i)$. Dans un premier temps nous calculons la matrice de données centrées $D \in M_{2n_p,n_e}(\mathbb{R})$:

$$D = (X^1 - \overline{X^e} \ \dots \ X^{n_e} - \overline{X^e}). \tag{2.1}$$

La PCA est calculée à l'aide d'une décomposition en valeurs singulières (« Singular Value Decomposition », SVD) de la matrice de corrélation $\Delta \in M_{n_e,n_e}(\mathbb{R})$:

$$\Delta = \frac{1}{n_e} D^T D. \tag{2.2}$$

La SVD permet d'écrire la matrice de corrélation sous la forme $\Delta = U\Sigma V^T$ où $\Sigma \in M_{n_e,n_e}(\mathbb{R}_+)$ est une matrice diagonale et $U, V \in M_{n_e,n_e}(\mathbb{R})$ sont des matrices unitaires orthogonales. Nous pouvons extraire les vecteurs propres $(W_1, \dots, W_{n_e}) \in (\mathbb{R}^{2n_p})^{n_e}$ et les valeurs propres associées $(\lambda_1, \dots, \lambda_{n_e}) \in \mathbb{R}^{n_e}$ de la matrice Δ grâce à la SVD :

$$(W_1 \ \dots \ W_{n_e}) = DU, \tag{2.3}$$

$$\forall i \in [\![1, n_e]\!], \lambda_i = \Sigma_{i,i}, \tag{2.4}$$

où $\forall i \in [\![1, n_e]\!], \Sigma_{i,i}$ est le i-ième élément de la diagonale de Σ.

Il faut ensuite choisir le nombre d'axes à conserver pour la PCA, c'est-à-dire le nombre de vecteurs propres à conserver. Cela peut être fixé par un pourcentage sur l'énergie totale des valeurs propres ($\sum_{i=1}^{n_e} \lambda_i$) ou arbitrairement. Dans [14], il est montré que les 3 premiers axes sont suffisants pour obtenir de bons taux de reconnaissance d'expressions faciales par la suite, nous conservons donc l'énergie des 3 premiers axes de la PCA. En outre, cela permet une représentation visuelle des expressions faciales. La matrice de projection des expressions faciales dans l'espace PCA est notée $\Phi \in M_{2n_p,3}(\mathbb{R})$:

$$\Phi = (W_1 W_2 W_3). \tag{2.5}$$

Auto-organisation de l'espace PCA et tessellation de Delaunay

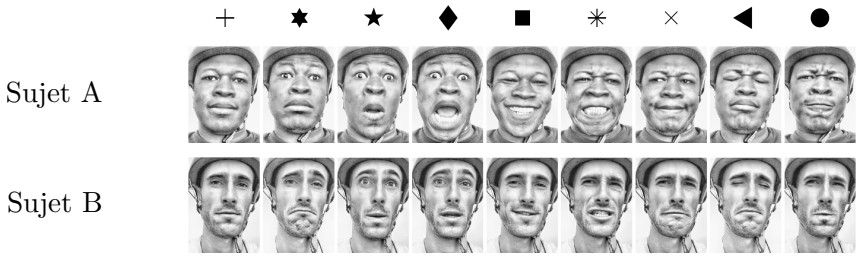
La PCA nous permet d'obtenir un espace continu couvrant les déformations faciales contenues dans les expressions de base plausibles. Le choix des expressions de base doit alors se faire en sachant que l'objectif est de modéliser l'ensemble des déformations faciales possibles autour du visage neutre. Prenons l'exemple des sourcils, deux déformations sont possibles autour de l'état neutre : froncés et levés. Si parmi les expressions de base aucune ne contient des sourcils froncés, le visage neutre ne sera pas au centre des déformations. Il faut donc choisir les expressions de base de telle sorte que l'état neutre soit au centre de chaque déformation possible dans l'espace PCA. Si cette condition est respectée, il est montré dans [14] que les projections des expressions de base dans l'espace PCA sont toujours organisées de la même manière autour du neutre, indépendamment du sujet. Il y a donc une auto-organisation de l'espace PCA.

Pour tirer profit de cette auto-organisation, nous effectuons une tessellation de Delaunay sur l'ensemble des projections des expressions de base plausibles et du visage neutre dans l'espace PCA. Il en résulte un ensemble de simplexes dont la structure est invariante et dans lequel le neutre est connecté à chaque expression de base. Cette structure de simplexes est appelée « structure du modèle ». Les expressions de base sont choisies pour la stabilité qu'elles confèrent à cette auto-organisation.

La figure 2.2 illustre un exemple d'auto-organisation de l'espace PCA pour deux sujets avec 9 expressions (neutre inclus). Deux sujets ont effectué 9 expressions identiques, dont le visage neutre. Pour chacun des sujets, la variété est apprise avec une PCA sur ces 9 expressions. Les expressions sont ensuite projetées dans l'espace PCA (partie gauche de la figure), puis la tessellation de Delaunay est effectuée sur les expressions projetées (partie droite de la figure). Pour les deux sujets, les expressions sont organisées dans le même ordre autour du visage neutre dans l'espace PCA. La structure du modèle est donc identique pour chacun des deux sujets.

Choix des expressions de base

La table 2.1 illustre les 5 expressions de base que nous avons sélectionnées pour construire le modèle : colère, dégoût, joie, tristesse et surprise. Elles sont inspirées des expressions prototypiques d'Ekman [3], mais pour pouvoir garantir la stabilité de l'auto-organisation du modèle les expressions de base du dégoût et de la tristesse sont différentes de celles identifiées par Ekman. Pour le dégoût, au lieu d'avoir une bouche ouverte avec le soulèvement de la lèvre supérieure (voir figure 1.1), la bouche est fermée avec les lèvres contractées l'une contre l'autre et étirées horizontalement. Pour la tristesse, nous avons



Deux premières composantes des projections des expressions dans l'espace PCA avant la tessellation de Delaunay (gauche) et après (droite)

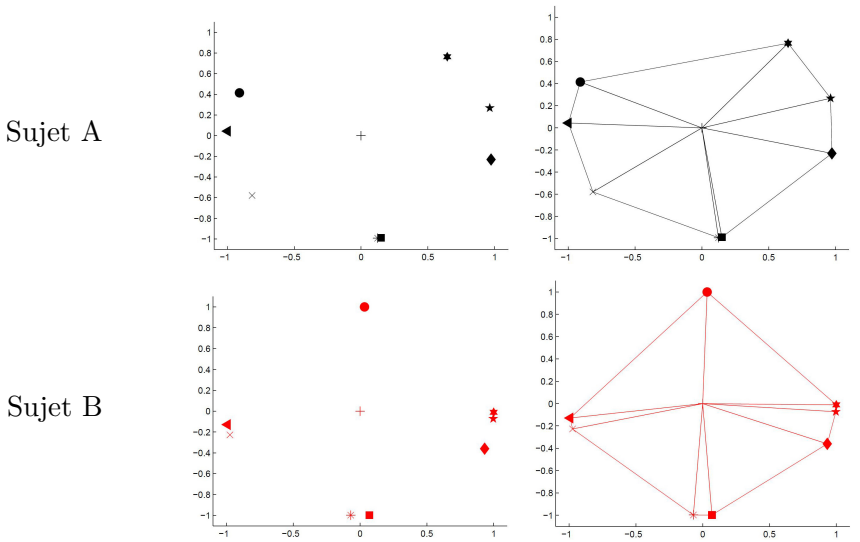


FIGURE 2.2 – Exemple d'auto-organisation de l'espace PCA pour deux sujets. L'espace PCA est appris avec le visage neutre et 8 expressions. Celles-ci sont ensuite projetées dans l'espace PCA et les deux premiers axes sont affichés (partie gauche). Puis la tessellation de Delaunay est effectuée sur la projection des expressions dans l'espace PCA (partie droite). Pour les deux sujets, les 8 expressions sont agencées de la même manière autour du neutre. La structure du modèle est donc identique pour chacun des deux sujets. La figure est extraite de [166].

choisi de la définir avec les yeux fermés pour éviter d'avoir un déséquilibre au niveau de cette déformation, alors que ce n'est pas le cas pour l'expression prototypique d'Ekman (voir figure 1.1). De plus, nous avons choisi de ne pas intégrer la peur, également une expression prototypique d'Ekman, car elle partage des unités d'action en commun avec la surprise dans chaque zone du visage, les déformations sont donc très proches.

TABLE 2.1 – Définition des 5 expressions de base du modèle. Les déformations locales des sourcils, des yeux et de la bouche sont traduites en termes d'unités d'action du système FACS [6].

Expression de base	Sourcils	Yeux	Bouche	Illustration
Colère	4	5+7	17+23 ou 17+24	
Dégoût	4	44	20+24	
Joie	0	6	12+26	
Tristesse	1+2	43	15	
Surprise	1+2	5	26 ou 27	

Parmi les 5 expressions de base, l'état neutre des sourcils est présent dans le visage neutre et dans la joie, l'état froncé des sourcils est présent la colère et le dégoût, et l'état levé des sourcils est présent dans la tristesse et la surprise. Concernant les yeux, il y a plusieurs niveaux allant de la fermeture complète (tristesse) à l'ouverture maximale (surprise). Au niveau de la bouche, la colère s'oppose au dégoût et à la tristesse sur l'étirement horizontal, alors que l'ouverture de la bouche est présente dans la joie et la surprise.

Espace des signatures

Étant donné que la structure du modèle est indépendante de la personne, nous pouvons transformer l'espace PCA spécifique à la personne en un espace normalisé indépendant de la personne, appelé « espace des signatures ». Ainsi, il n'est pas nécessaire d'aligner les modèles de deux sujets différents afin de les comparer. La projection d'une expression dans l'espace des signatures est appelée « signature ». L'espace des signatures est défini de telle sorte que l'expression neutre soit à l'origine et que les expressions de base se trouvent sur une surface sphérique tout en respectant la structure du modèle obtenue avec la tessellation de Delaunay. La figure 2.3 illustre la structure invariante du modèle dans l'espace des signatures. L'expression neutre se situe à l'origine et est connectée aux 5 expressions de base. Les expressions de base de la colère, de la joie et

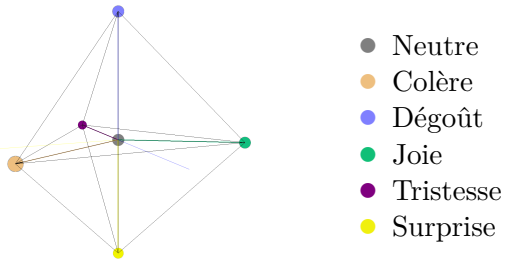


FIGURE 2.3 – Structure du modèle dans l'espace des signatures. Le neutre est à l'origine. Chaque expression de base est reliée aux autres (y compris le neutre) à l'exception du dégoût et de la surprise qui sont opposés.

de la tristesse sont liées à toutes les autres expressions de base. Les expressions de base du dégoût et de la surprise sont liées aux trois autres mais pas entre elles : ce sont les sommets de deux pyramides à base triangulaire commune.

En utilisant la codification Neutre = 1, Colère = 2, Dégoût = 3, Joie = 4, Tristesse = 5 et Surprise = 6, nous représentons l'ensemble des simplexes de la structure du modèle avec la matrice suivante :

$$S = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 5 \\ 1 & 2 & 4 & 6 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 4 & 5 \\ 1 & 4 & 5 & 6 \end{pmatrix}, \tag{2.6}$$

où chaque ligne contient les expressions de base du modèle formant un simplexe.

Les coordonnées des expressions de base du modèle dans l'espace des signatures sont regroupées dans la matrice suivante :

$$C = \begin{pmatrix} 0 & 0 & 0 \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \tag{2.7}$$

où $\forall i \in [\![1, n_e]\!]$, la i-ième ligne de C est notée C_i et contient les coordonnées de la signature de la i-ième expression de base du modèle (en reprenant la codification des expressions utilisée pour la matrice S).

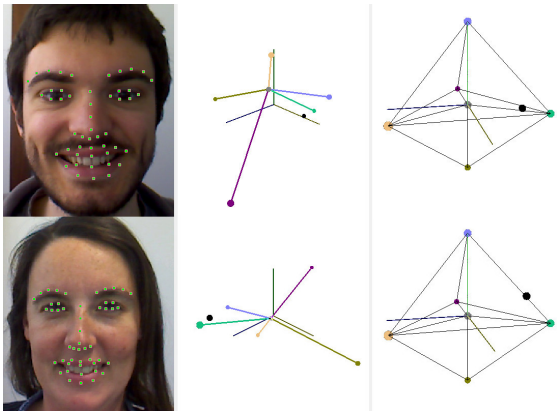


FIGURE 2.4 – Illustration de l'invariance de la structure du modèle. L'expression de joie de deux sujets est projetée dans leur espace PCA respectif (affichée par un point noir dans la partie centrale). Lorsque l'expression est transférée dans l'espace des signatures, l'expression projetée se trouve dans la même zone de l'espace pour les deux sujets (affichée par un point noir dans la partie de droite).

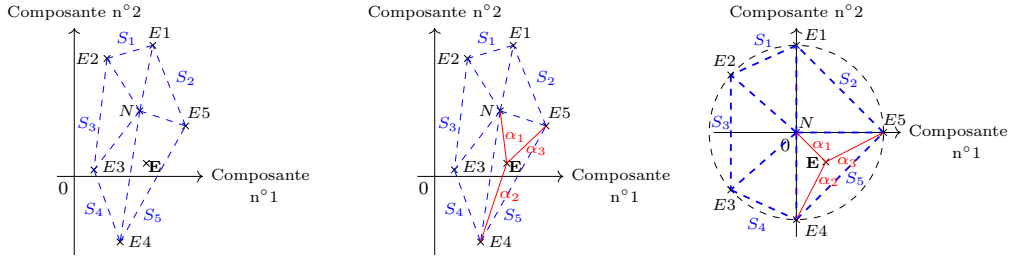
La figure 2.4 illustre la propriété d'invariance de la structure du modèle. Deux sujets font l'expression de la joie. Elle est projetée dans leur espace PCA respectif, qui est spécifique à la personne. La projection de l'expression dans l'espace PCA est affichée par un point noir dans la partie centrale de la figure. On peut voir qu'elle se situe près de l'expression de base de la joie (point cyan) pour les deux sujets. Ensuite, cet espace est transformé en l'espace des signatures. Grâce à la structure invariante du modèle, la signature de la joie (affichée par un point noir dans la partie droite de la figure) est maintenant située dans la même zone de l'espace pour les deux sujets.

Pour récapituler, le modèle est défini par :

- Les points caractéristiques du visage neutre et des 5 expressions de base,
- L'espace PCA et sa matrice de projection Φ,
- La structure du modèle,
- L'espace des signatures.

2.1.2 Calcul de la signature d'une expression

Le calcul de la signature d'une expression se fait en 4 étapes : projection dans l'espace PCA, détermination du simplexe (résultant de la tessellation de Delaunay) dans lequel la projection PCA de l'expression est contenue, calcul des coordonnées barycentriques de la projection PCA de l'expression dans le simplexe et transposition des coordonnées barycentriques dans l'espace des signatures. La figure 2.5 illustre les étapes du calcul.



(a) Projection de l'expression dans l'espace PCA (b) Calcul des coordonnées barycentriques (c) Transposition dans l'espace des signatures

FIGURE 2.5 – Calcul de la signature d'une expression. Bien que cette grandeur soit calculée dans un espace à 3 dimensions, les figures sont en deux dimensions, et ce à des fins illustratives. Le neutre est noté N, les 5 expressions de base du modèle $E1, E2, E3, E4, E5$ et l'expression dont on veut calculer la signature E. Les simplexes résultant de la tessellation de Delaunay sont représentés par les lignes bleues en pointillé et notés S_1, S_2, S_3, S_4, S_5. Les points caractéristiques de l'expression sont d'abord projetés dans l'espace PCA du modèle expressif spécifique à la personne (voir sous-figure 2.5a). Ensuite, nous déterminons quel simplexe de la structure du modèle contient la projection PCA de l'expression, dans cet exemple c'est le simplexe S_5 (voir sous-figure 2.5a). Nous calculons alors les coordonnées barycentriques de la projection PCA de l'expression dans ce simplexe : $X_{PCA} = \alpha_1 X^N_{PCA} + \alpha_2 X^{E4}_{PCA} + \alpha_3 X^{E5}_{PCA}$ (voir sous-figure 2.5b). Enfin, nous transposons les coordonnées barycentriques ainsi obtenues dans l'espace des signatures : $X_{sig} = \alpha_1 C_N + \alpha_2 C_{E4} + \alpha_3 C_{E5}$ (voir sous-figure 2.5c).

Soit $X \in \mathbb{R}^{2n_p}$ le vecteur contenant les points caractéristiques (après alignement avec les expressions de base du modèle) de l'expression dont on veut calculer la signature, où n_p est le nombre de points caractéristiques. L'expression est d'abord projetée dans l'espace PCA :

$$X_{PCA} = (X - \overline{X^e})\Phi, \tag{2.8}$$

où $X_{PCA} \in \mathbb{R}^3$ est appelé vecteur PCA de l'expression, $\overline{X^e} \in \mathbb{R}^{2n_p}$ est la moyenne de l'ensemble des points caractéristiques du visage neutre et des expressions de base du modèle et $\Phi \in M_{2n_p,3}(\mathbb{R})$ est la matrice de projection dans l'espace PCA (voir équation 2.5).

Ensuite, nous déterminons à quel simplexe de la structure du modèle appartient le vecteur PCA de l'expression et nous calculons les coordonnées barycentriques de ce dernier dans le simplexe. Soit $n_{si} \in \mathbb{N}$ le nombre de simplexes dans la structure du modèle, correspondant au nombre de ligne de la matrice S (voir équation 2.6), ici $n_{si} = 6$. Soit $j \in [\![1, n_{si}]\!]$ l'indice du simplexe contentant le vecteur PCA de l'expression. S_j, la j-ième ligne de la matrice S, est le simplexe en question. Pour tout $i \in [\![1, n_e]\!]$, on

note X^i_{PCA} le vecteur PCA de la i-ième expression de base du modèle (on reprend la codification des expressions de base utilisée pour la matrice S). Les coordonnées barycentriques de l'expression dont on veut calculer la signature sont les coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{R}^4$ tels que :

$$X_{PCA} = \sum_{k=1}^{4} \alpha_k X_{PCA}^{S_{j,k}}, \tag{2.9}$$

où $\forall k \in [\![1, 4]\!], S_{j,k}$ est la k-ième colonne de la j-ième ligne de S, c'est-à-dire l'indice de l'expression de base du modèle correspondant au k-ième sommet du simplexe S_j.

Enfin, la signature de l'expression est calculée en transposant les coordonnées barycentriques de l'expression dans l'espace des signatures :

$$X_{sig} = \sum_{k=1}^{4} \alpha_k C_{S_{j,k}}, \tag{2.10}$$

où $\forall k \in [\![1, 4]\!], C_{S_{j,k}}$ est la ligne de la matrice C (voir équation 2.7) contenant les coordonnées de la signature de l'expression de base du modèle correspondant au k-ième sommet du simplexe S_j.

2.2 Reconnaissance d'expressions faciales robuste à la pose

Dans cette section, nous présentons notre méthode de reconnaissance d'expressions faciales. Elle est utilisée à deux reprises : dans la section 3.1 pour détecter le visage neutre et dans les sections 3.2 et 3.3 pour détecter les expressions de base. Nous adoptons l'architecture classique du système de reconnaissance d'expressions faciales présentée dans la sous-section 1.2.3 : extraction des caractéristiques puis classification (voir figure 1.5). Dans notre cadre applicatif, il est nécessaire que la reconnaissance d'expressions faciales soit robuste à la pose car le sujet est dans un environnement non contraint où il est fort probable qu'il bouge la tête. Pour cela, nous proposons de définir les caractéristiques comme la concaténation d'angles et de distances entre certains points caractéristiques du visage et nous les présentons dans la sous-section 2.2.1. Ensuite nous présentons l'outil de classification choisi dans la sous-section 2.2.2. Après cela, nous réalisons dans la sous-section 2.2.3 une expérimentation pour montrer la robustesse à la pose des caractéristiques angle-distance. Enfin, nous présentons les résultats obtenus pour la reconnaissance d'expressions faciales dans la sous-section 2.2.4 et nous terminons par une conclusion dans la sous-section 2.2.5. Le travail présenté dans cette section a été

effectué en collaboration avec Vincent Barrielle, alors doctorant CIFRE chez Dynamixyz en co-tutelle avec l'équipe FAST.

2.2.1 Extraction des caractéristiques angle-distance

Pour l'extraction des caractéristiques faciales, nous choisissons de concaténer des angles et des distances calculés entre certains points caractéristiques du visage. Ce type de caractéristiques a déjà été utilisé pour la reconnaissance d'expressions faciales, notamment dans [167, 125, 45]. A notre connaissance, les caractéristiques angle-distance ont été utilisées pour la reconnaissance d'expressions faciales robuste à la pose uniquement dans [153] où elles sont combinées à des caractéristiques d'apparence. Nous proposons ici une méthode de reconnaissance d'expressions faciales robuste à la pose qui se base uniquement sur les caractéristiques angle-distance. Elle repose sur l'invariance de ces caractéristiques à de faibles variations de pose sous l'hypothèse de faible perspective (voir annexe B). L'angle de « roll » n'est pas mentioné car c'est une rotation dans le plan de la caméra. De plus, nous faisons l'hypothèse que ces caractéristiques ont une capacité de généralisation suffisante par rapport à la variabilité inter-personnelle.

Une autre raison qui nous pousse à choisir les caractéristiques angle-distance est la possibilité de faire une analyse aussi bien globale (sur le visage en entier) que locale (sur une zone spécifique du visage comme les sourcils, les yeux et la bouche), ce qui sera utile dans le chapitre 3. Étant donné que ces caractéristiques ont déjà été utilisées pour la reconnaissance d'unités d'action [49], elles sont adaptées à une analyse locale.

Dans cette thèse, nous utilisons Intraface [168] pour le suivi des points caractéristiques du visage, car il offre une bonne précision sur des images dans un environnement non contraint. Il renvoie 49 points caractéristiques dans les régions des sourcils, des yeux, du nez et de la bouche. Il est montré dans [168] qu'Intraface est capable d'effectuer de manière fiable le suivi de points caractéristiques sur des visages avec une forte pose (« yaw » +/-45°, « roll » +/-90° et « pitch » +/-30°).

Soient n_p le nombre de points caractéristiques et $X \in M_{n_p,2}(\mathbb{R})$ la matrice contenant les points caractéristiques du visage. Pour tout $i \in [\![1, n_p]\!]$, on note $X_i \in \mathbb{R}^2$ le i-ième point caractéristique de X.

Soit $(j, k, l) \in [\![1, n_p]\!]^3$ tel que $j \neq k \neq l$. L'angle $\widehat{X_j X_k X_l}$ est calculé par la fonction arc tangente « atan2 » du sinus et du cosinus de l'angle entre les vecteurs $X_j - X_k$ et $X_l - X_k$.

Avant de calculer la distance, nous calculons un facteur d'échelle s pour rendre la distance invariante aux changements d'échelle du visage dans l'image. Il est défini comme étant la distance euclidienne moyenne entre les points caractéristiques et le point carac-

téristique moyen du visage $X_{moy} = moyenne_{i \in [\![1,n_p]\!]}(X_i)$:

$$s = moyenne_{i \in [\![1,n_p]\!]} \|X_i - X_{moy}\|_2. \tag{2.11}$$

Soit $(m, o) \in [\![1, n_p]\!]^2$ tel que $m \neq o$, la distance entre les points caractéristiques X_m et X_o est calculée comme la distance euclidienne entre les points caractéristiques, normalisée par le facteur d'échelle s :

$$d_{m,o} = \frac{\|X_m - X_o\|_2}{s}. \tag{2.12}$$

La conception de nos caractéristiques angle-distance consiste à définir les triplets (X_j, X_k, X_l) pour le calcul des angles et les paires (X_m, X_o) pour le calcul des distances. Nous définissons empiriquement ces triplets et ces paires en fonction de leur capacité à discriminer les expressions faciales. Nos caractéristiques angle-distance sont composées de 3 angles et 6 distances dans la zone des sourcils, 4 angles dans la zone des yeux, 8 angles et 3 distances dans la zone de la bouche, comme indiqué dans la table 2.2.

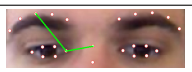
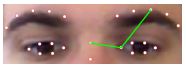
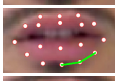
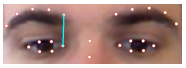
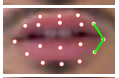
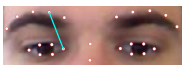
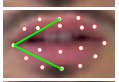
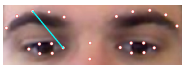
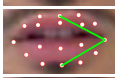
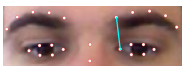
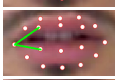
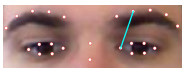
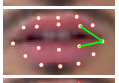
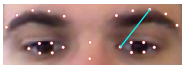
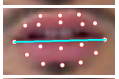
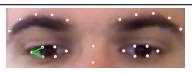
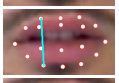
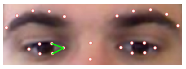
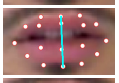
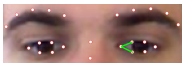
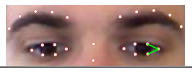
2.2.2 Classification globale et locale

Pour la classification, nous choisissons un SVM avec pour noyau la fonction de base radiale (« Radial Basis Function », RBF). Nous utilisons l'implémentation de scikit-learn [169]. Selon le sous-ensemble des caractéristiques angle-distance que nous utilisons, nous entraînons un classifieur global et trois classifieurs locaux :

- Visage entier avec 6 classes (neutre, colère, dégoût, joie, tristesse, surprise), voir table 2.1,
- Sourcils avec 3 classes (neutres, froncés, levés),
- Yeux avec 2 classes (ouverts, moitié-fermés/fermés),
- Bouche avec 6 classes (neutre, colère, dégoût, joie, tristesse, surprise).

Nous allons tester deux configurations pour l'apprentissage du classifieur. Soit la base d'apprentissage n'est composée que de visages de face, auquel cas on se ramène à un système avec des caractéristiques faciales robustes à la pose (voir partie « Extraction de caractéristiques invariantes à la pose » de la sous-section 1.2.4) ; soit la base d'apprentissage est composée de visages de face et avec 8 poses différentes (+/- 14° et +/- 27° yaw, +/- 10° et +/- 20° pitch), auquel cas on se ramène à un système à classifieur unique avec plusieurs poses (voir partie « Apprentissage d'un unique classifieur avec plusieurs poses » de la sous-section 1.2.4).

TABLE 2.2 – Définition de nos caractéristiques angle-distance. Les points caractéristiques sont affichés par des points blancs, les angles par des lignes vertes et les distances par des lignes bleues.

Indice	Illustration	Indice	Illustration
1		14	
2		15	
3		16	
4		17	
5		18	
6		19	
7		20	
8		21	
9		22	
10		23	
11		24	
12		25	
13			

2.2.3 Résultats sur la robustesse

Dans cette sous-section, nous réalisons une expérimentation pour montrer la robustesse à la pose des caractéristiques angle-distance. Dans la première partie, nous décrivons le protocole expérimental. Dans la seconde partie, nous présentons les résultats.

Protocole expérimental

Pour les expérimentations, nous avons sélectionné 2 sujets de notre base maison FAST (voir partie « Bases de données » de la sous-section 2.2.4 pour la description de la base). Nous enregistrons une vidéo supplémentaire pour ces 2 sujets. Dans cette vidéo, le sujet effectue 3 expressions faciales à la suite (neutre, colère bouche fermée et joie sourcils

levés) et pour chaque expression la pose de la tête varie selon les angles de « yaw » et de « pitch ». Les expressions ont été choisies de telles sortes que les déformations faciales soient opposées par rapport au visage neutre. Pour chaque expression, le sujet démarre de face avec un visage neutre avant de faire l'expression. Une fois l'« apex » atteint, la pose de la tête varie selon le « yaw » dans un sens puis l'autre. Ensuite il se remet de face. Puis la pose de la tête varie selon le « pitch » dans un sens puis l'autre. Enfin le visage se remet de face et retourne à l'expression neutre. A partir du moment où l'« apex » est atteint, le sujet le maintient pendant toute la variation de pose.

Pour chaque sujet, les caractéristiques suivantes sont calculées sur l'intégralité de la vidéo :

- points caractéristiques sans alignement,
- points caractéristiques avec alignement de Procruste,
- caractéristiques angle-distance.

Les points caractéristiques sont calculés à l'aide d'Intraface [168]. Une analyse en composantes principales (PCA) est calculée sur chacun de ces ensembles de caractéristiques. Nous pouvons ensuite visualiser le nuage de points obtenu en projetant l'ensemble des caractéristiques dans leur espace PCA respectif.

Résultats

Dans la table 2.3, nous reportons la répartition de l'énergie de la PCA sur les trois premiers axes et ce pour les 3 ensembles de caractéristiques définis ci-dessus et pour les 2 sujets. Nous remarquons que la majorité de l'énergie de la PCA est contenue dans les deux premiers axes, nous ne visualisons que les deux premiers axes du nuage de points des caractéristiques projetées.

TABLE 2.3 – Énergie des trois premiers axes de la PCA.

Sujet	Ensemble de caractéristiques	1er axe PCA	2e axe PCA	3e axe PCA
Sujet 1	Points caractéristiques sans alignement	56.36%	32.60%	9.73%
	Points caractéristiques avec alignement	65.87%	11.97%	8.54%
	Caractéristiques angle-distance	79.54%	10.31%	5.75%
Sujet 2	Points caractéristiques sans alignement	63.62%	33.26%	2.10%
	Points caractéristiques avec alignement	40.59%	21.81%	14.29%
	Caractéristiques angle-distance	48.35%	31.14%	10.76%

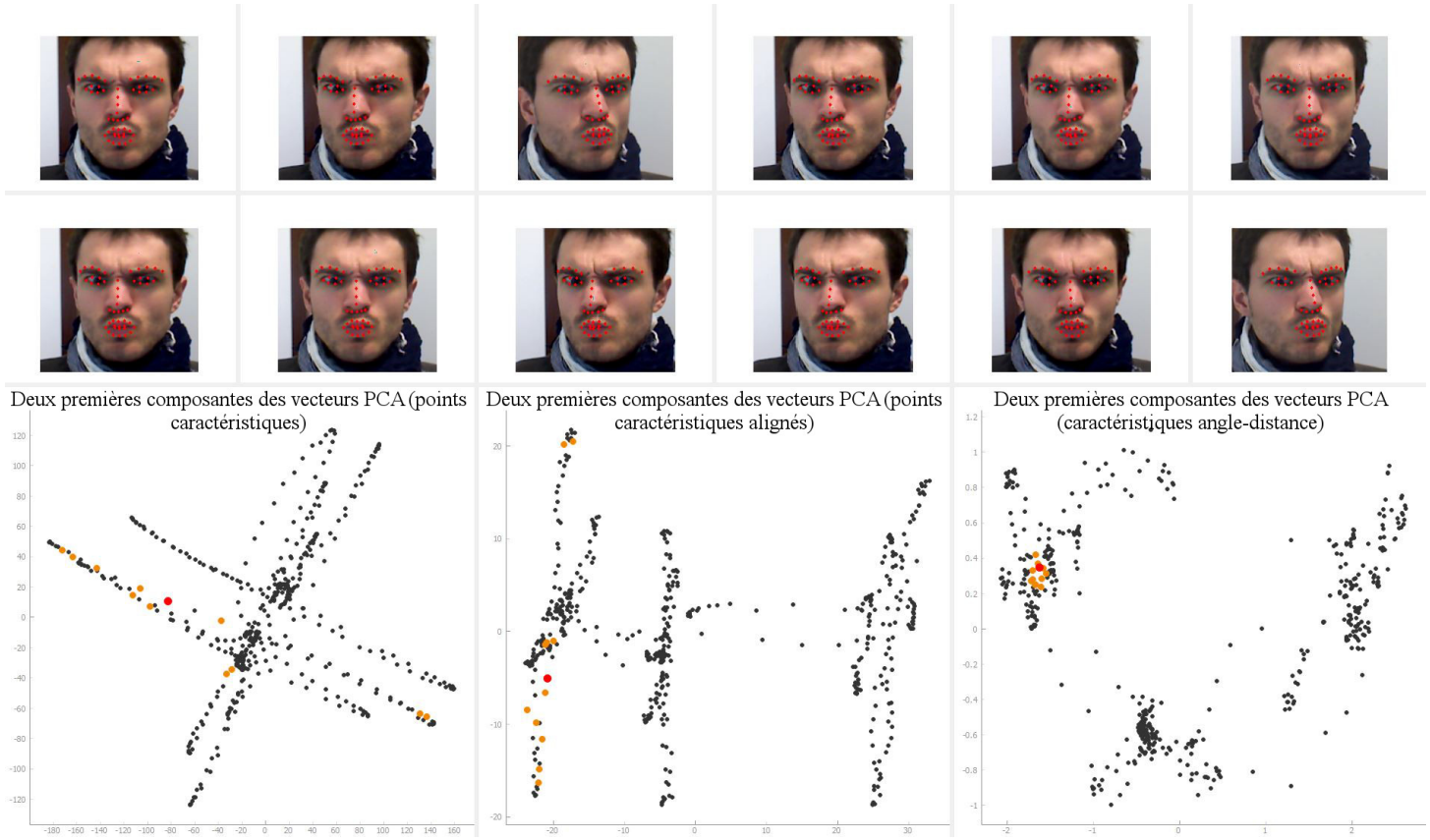


FIGURE 2.6 – Robustesse des caractéristiques angle-distance à de faibles variations de pose. De gauche à droite sont affichés les nuages de points des projections des points caractéristiques sans alignement, des points caractéristiques avec alignement de Procruste et des caractéristiques angle-distance respectivement. Dans chacune de ces sous-figures, nous affichons en rouge le point correspondant à l'expression de colère affichée en haut à gauche. Nous affichons également en orange ses 11 plus proches voisins dans l'espace PCA des caractéristiques angle-distance, correspondant aux autres expressions affichées. Ces mêmes 11 voisins sont aussi affichés en orange dans les deux autres nuages de points.

Robustesse à de faibles variations de pose Nous présentons ici les résultats pour le premier sujet. La variation de pose reste faible : le yaw varie entre -16° et +13° et le pitch varie entre -16° et +21°. La pose de la tête est calculée à l'aide d'Intraface [168]. La figure 2.6 regroupe les nuages de points des 3 ensembles de caractéristiques projetés dans leur espace PCA respectif. Dans chacune des sous-figures, nous affichons en rouge le point correspondant à l'expression de colère affichée en haut à gauche de la figure. Nous affichons également en orange ses 11 plus proches voisins dans l'espace PCA des caractéristiques angle-distance, correspondant aux autres expressions affichées. Ces mêmes 11 voisins sont aussi affichés en orange dans les deux autres nuages de points.

Pour les points caractéristiques sans alignement, les premiers axes de l'espace PCA contiennent essentiellement l'information de pose. Nous pouvons voir que le nuage de points forme une croix, chaque axe de la croix correspondant à une variation de pose selon un angle spécifique (yaw ou pitch).

Pour les points caractéristiques avec alignement, le premier axe contient l'information expressive et le deuxième axe l'information contient l'information de pose. Nous pouvons voir que le nuage de points est constitué grossièrement de 3 segments, chacun contenant une expression spécifique.

Pour les caractéristiques angle-distance, les deux premiers axes contiennent essentiellement l'information expressive. Nous pouvons voir que le nuage de points est constitué de 3 groupes de points, chacun correspondant à une expression spécifique.

Les 12 expressions voisines dans le nuage de points des caractéristiques angle-distance sont toutes de la colère, avec différentes poses selon l'angle du yaw. Nous pouvons voir que ces mêmes expressions se situent toutes sur le premier segment dans le nuage de points des points caractéristiques après alignement, leur positionnement respectif sur ce segment dépendant de la pose. Dans le nuage de points des points caractéristiques sans alignement, ces mêmes expressions se situent toutes sur le même axe de la croix, leur positionnement respectif sur cet axe dépendant de la pose.

Cette première expérimentation montre donc que les caractéristiques angle-distance sont robustes à de faibles variations de pose, contrairement aux points caractéristiques utilisés directement comme caractéristiques faciales avec ou sans alignement.

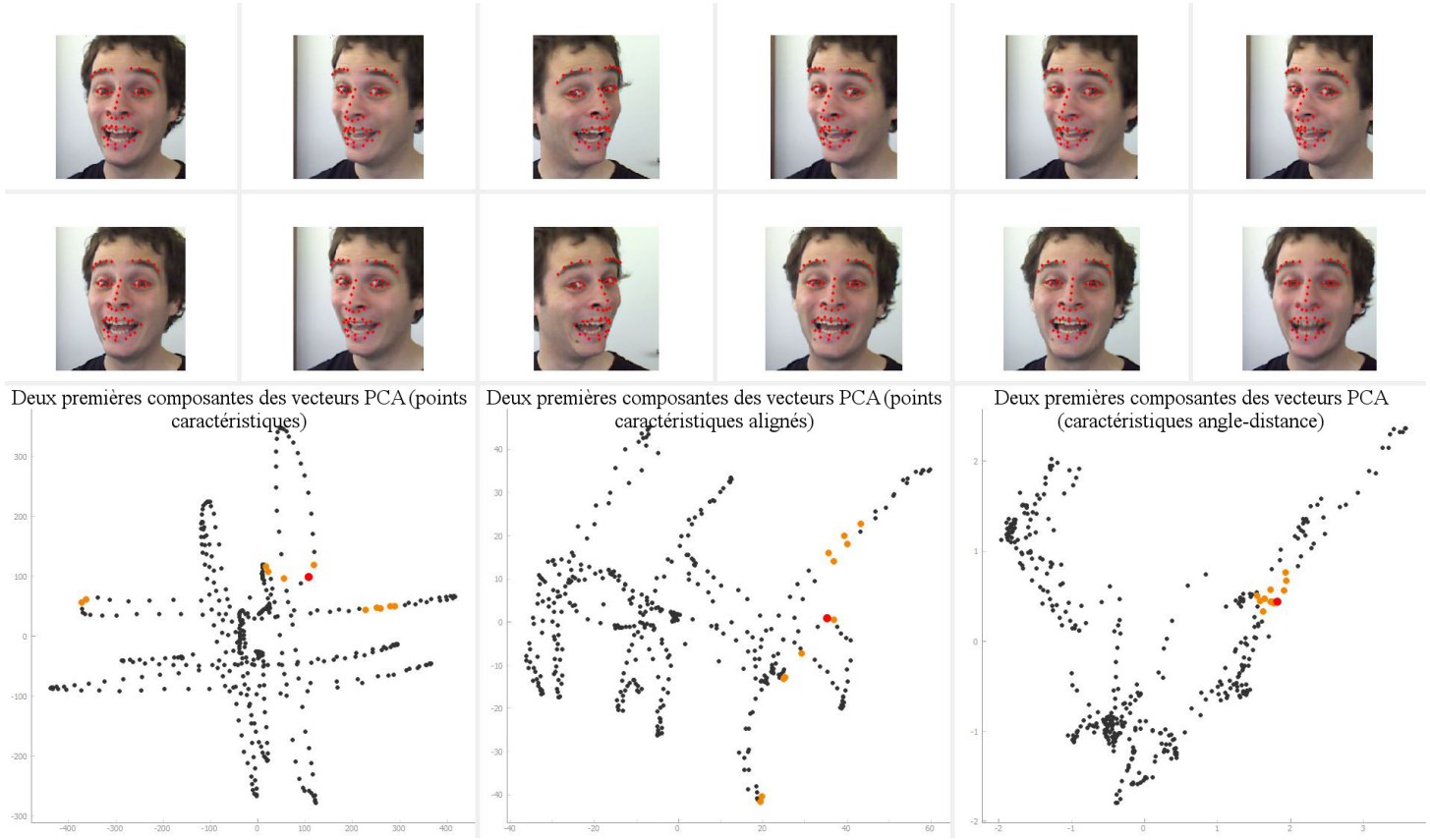


FIGURE 2.7 – Robustesse des caractéristiques angle-distance à des variations de pose importantes. De gauche à droite sont affichés les nuages de points des projections des points caractéristiques sans alignement, des points caractéristiques avec alignement de Procruste et des caractéristiques angle-distance respectivement. Dans chacune de ces sous-figures, nous affichons en rouge le point correspondant à l'expression de joie sourcils levés affichée en haut à gauche. Nous affichons également en orange ses 11 plus proches voisins dans l'espace PCA des caractéristiques angle-distance, correspondant aux autres expressions affichées. Ces mêmes 11 voisins sont aussi affichés en orange dans les deux autres nuages de points.

Robustesse à des variations de pose importantes Nous présentons ici les résultats pour le second sujet. La variation de pose est plus importante que pour le premier sujet : le yaw varie entre -35° et +40° et le pitch varie entre -25° et +30°. La pose de la tête est calculée à l'aide d'Intraface [168]. La figure 2.7 regroupe les nuages de points des 3 ensembles de caractéristiques projetés dans leur espace PCA respectif. Dans chacune des sous-figures, nous affichons en rouge le point correspondant à l'expression de joie sourcils levés affichée en haut à gauche de la figure. Nous affichons également en orange ses 11 plus proches voisins dans l'espace PCA des caractéristiques angle-distance, correspondant aux autres expressions affichées. Ces mêmes 11 voisins sont aussi affichés en orange dans les deux autres nuages de points.

Par rapport à la figure 2.6 (avec de faibles variations de pose), nous pouvons voir que la forme du nuage de points reste grossièrement la même pour les 3 ensembles de caractéristiques. A noter toutefois que pour les caractéristiques angle-distance, les 3 groupes de points (chacun correspondant à une expression spécifique) sont plus disparates et la distinction entre eux est moins marquée. Cela montre que lorsque les variations de pose sont importantes, la robustesse des caractéristiques angle-distance est mise à mal. L'approximation faite dans l'équation B.11 n'est plus valide.

2.2.4 Résultats sur la reconnaissance d'expressions faciales

Puisque la contribution de cette thèse ne porte pas sur la reconnaissance d'expressions faciales, nous ne cherchons pas ici à dépasser les performances de l'état de l'art mais simplement à atteindre une performance suffisante pour pouvoir utiliser cette brique dans l'adaptation non supervisée du modèle spécifique à la personne (voir chapitre 3). Notre objectif est de détecter des expressions faciales qui ne sont pas forcément les expressions prototypiques d'Ekman, dans le but de pouvoir remplacer les expressions de base du modèle spécifique à la personne présenté dans la section 2.1. De plus, nous avons besoin de détecter les expressions de base localement, ce qui n'est pas proposé par les méthodes de l'état de l'art. Dans un premier temps, nous décrivons brièvement les bases de données utilisées pour évaluer la performance de notre méthode. Ensuite, nous comparons notre méthode avec des méthodes de l'état de l'art sur 3 bases de données ne contentant que des expressions frontales. Puis, nous testons la robustesse à la pose de notre méthode sur une base de données maison.

Bases de données

Deux bases de données publiques et une base de données maison sont utilisées pour les expérimentations.

- CK+ [58] : Cette base de données contient 593 vidéos de 123 sujets âgés entre 18 et 50 ans. Les expressions sont posées et frontales, acquises dans l'environnement du laboratoire. Nous gardons les vidéos contenant les 6 expressions protoypiques d'Ekman [3], ce qui donne 309 vidéos de 118 sujets (82 femmes et 35 hommes).
- MUG [22] : Cette base de données contient 934 vidéos d'expressions posées de 49 sujets (21 femmes, 28 hommes) et 45 vidéos d'expressions spontanées d'un sous-ensemble de 45 sujets (19 femmes, 26 hommes). Les sujets sont âgés entre 20 et 35 ans. Les expressions posées sont le visage neutre et les 6 expressions prototypiques d'Ekman [3]. La méthode d'induction émotionnelle pour les expressions spontanées est le visionnage de vidéos censées induire les 6 émotions de base d'Ekman [3]. Les vidéos sont enregistrées dans l'environnement du laboratoire. Nous appelons « MUG-posé » le sous-ensemble d'expressions posées et « MUG-spontané » le sous-ensemble d'expressions spontanées.
- FAST : Etant donné que dans notre modèle les expressions de dégoût et de tristesse sont différentes des expressions prototypiques d'Ekman [3] (voir table 2.1) et ne sont pas présentes dans les bases de données publiques, nous avons construit notre propre base de données. Elle est composée de 38 sujets (12 femmes, 26 hommes) âgés de 20 à 50 ans. Pour chaque sujet, un ensemble de 7 vidéos (neutre, colère, dégoût, peur, joie, tristesse et surprise) est enregistré dans le laboratoire avec un visage de face, où le dégoût et la tristesse sont définis comme dans la table 2.1. Un deuxième ensemble de 7 vidéos défini de la même manière est enregistré pour 32 sujets, ce qui donne 70 ensembles de vidéos en tout. L'expression de la peur est manquante pour 9 de ces ensembles de vidéos. Pour 8 sujets (2 femmes, 6 hommes), un ensemble supplémentaire de vidéos est enregistré où la pose de la tête varie continûment pour les angles de « yaw » et de « pitch ». Nous appelons « FAST-frontal » le sous-ensemble des expressions frontales et « FAST-pose » le sous-ensemble des expressions avec variation de la pose de la tête.

Résultats sur des expressions frontales

Pour la reconnaissance d'expressions faciales frontales, nous comparons notre méthode avec des méthodes récentes utilisant aussi des caractéristiques géométriques. Dans [45], un modèle de distribution de points (« Point Distribution Model », PDM) est utilisé

pour le suivi de points caractéristiques du visage. Ensuite, toutes les distances possibles entre les points caractéristiques sont calculées et deux techniques de sélection de caractéristiques sont testées : sélection empirique et sélection de caractéristiques par corrélation (« Correlation Features Selection », CFS). La classification est assurée par une SVM avec un noyau RBF. Dans [44], la méthode de correspondance élastique par graphes groupés (« Elastic Bunch Graph Matching », EBGM) est utilisée pour l'initialisation des points caractéristiques, puis la méthode de Kanade-Lucas-Tomaci (KLT) est utilisée pour le suivi des points caractéristiques. Trois caractéristiques géométriques sont testées : basées sur des points, basées sur des lignes (équivalent aux angles et distances) et basées sur des triangles. A titre de comparaison, nous ne reportons que les résultats obtenus avec les caractéristiques basées sur des lignes. Ensuite, la sélection de caractéristiques est effectuée à l'aide d'un AdaBoost multi-classe avec de l'apprentissage machine extrême (« Extreme Machine Learning », EML). La classification est assurée par une SVM avec un noyau RBF. Dans [170], les caractéristiques sont définies comme étant toutes les distances possibles entre 18 points caractéristiques. Ensuite, la sélection de caractéristiques est effectuée avec une CFS. La classification est assurée par un réseau de neurones artificiel (« Artificial Neural Network », ANN).

Pour les bases de données FAST-frontal, CK+ et MUG-posé, nous extrayons une image à l'« apex » (intensité maximale de l'expression) dans chaque vidéo. Dans la base CK+, il n'y a pas de vidéos contenant uniquement l'expression neutre mais chaque vidéo commence par le visage neutre avant de montrer l'expression, nous extrayons donc la première image d'une des vidéos de chaque sujet pour avoir une image du visage neutre. Dans ces expérimentations, les expressions sont frontales pour l'apprentissage et le test. Deux problèmes sont considérés : reconnaissance de 6 classes (colère, dégoût, peur, joie, tristesse et surprise) et reconnaissance de 7 classes (ajout du neutre). Nous utilisons la validation croisée par k-dossiers (« k-Fold Cross Validation »), avec $k = 10$, pour calculer la performance de notre méthode de reconnaissance d'expressions faciales et la comparer aux autres méthodes.

Les résultats sont reportés dans la table 2.4. Les méthodes utilisant la sélection de caractéristiques (CFS [45] et ligne [44]) ont une meilleure performance que les méthodes avec des caractéristiques empiriques (empirique [45] et la nôtre). Sur la base CK+, notre méthode présente une meilleure performance que [45] avec des caractéristiques empiriques. Avec notre méthode, nous n'obtenons pas les mêmes performances que l'état de l'art mais cela n'est pas un problème pour l'adaptation non supervisée du modèle spécifique à la personne car l'étape de vérification des contraintes est là pour rejeter les mauvaises détections (voir la sous-section 3.2.4).

TABLE 2.4 – Comparaison du taux de reconnaissance de notre méthode et de méthodes existantes de reconnaissance d'expressions faciales sur des expressions frontales. La performance est calculée avec la validation croisée par k-dossiers, avec $k = 10$. Deux résultats sont reportés : reconnaissance de 6 classes (colère, dégoût, peur, joie, tristesse et surprise) et reconnaissance de 7 classes (ajout du neutre).

Nombre de classes	Méthode	FAST-frontal	CK+	MUG-posé
6	empirique [45]	-	89.75%	-
	CFS [45]	-	94.60%	-
	ligne [44]	-	**96.58%**	**94.13%**
	Notre méthode	85.28%	91.26%	89.07%
7	empirique [45]	-	85.03%	-
	CFS [45]	-	**96.11%**	-
	[170]	-	-	**91.22%**
	Notre méthode	84.22%	89.00%	85.21%

Résultats sur des expressions avec variation de pose

Nous conduisons une seconde expérimentation sur notre base de données FAST-pose afin d'évaluer la robustesse aux variations de pose de notre méthode de reconnaissance d'expressions faciales. La base FAST-pose est composée de 8 sujets pour lesquels 6 vidéos sont enregistrées, correspondant aux expressions de base du modèle expressif spécifique à la personne (voir table 2.1) : neutre, colère, dégoût, joie, tristesse et surprise. Dans chaque vidéo le sujet effectue l'expression avec le visage de face puis conserve l'expression tout en bougeant la tête successivement selon le « yaw » et le « pitch ».

Pour chaque vidéo, nous extrayons 9 images : visage frontal, $+/-14°$ yaw, $+/-27°$ yaw, $+/-10°$ pitch et $+/-20°$ pitch (voir figure 2.8). Les images ainsi extraites sont utilisées pour l'apprentissage. Quatre configurations sont considérées :

- Frontal : pour chaque expression, seule l'image frontale est utilisée,
- 4 poses 14°/10° : pour chaque expression, l'image frontale et 4 images avec de la pose (+/-14° yaw et +/-10° pitch) sont utilisées,
- 4 poses 27°/20° : pour chaque expression, l'image frontale et 4 images avec de la pose (+/-27° yaw et +/-20° pitch) sont utilisées,
- 8 poses : pour chaque expression, l'image frontale et 8 images avec de la pose (+/-14° yaw, +/-27° yaw, +/-10° pitch et +/-20° pitch) sont utilisées.

Les vidéos complètes sont utilisées pour le test. Au préalable, les images où le sujet n'affiche pas l'expression demandée et où le suivi de points caractéristiques est défaillant sont supprimées. Ensuite, nous séparons les vidéos en deux sous-parties : « Test frontal »

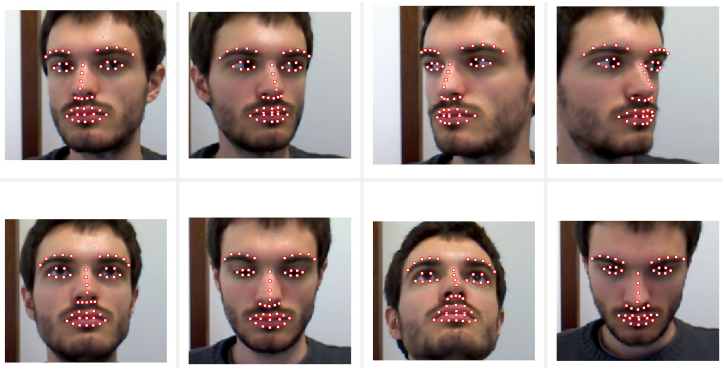


FIGURE 2.8 – Visage neutre d'un sujet de la base FAST-pose avec de les 8 poses extraites pour l'apprentissage du système de reconnaissance d'expressions faciales (de gauche à droite et de haut en bas : -14° yaw, +14° yaw, -27° yaw, +27° yaw, -10° pitch, +10° pitch, -20° pitch, +20° pitch). Les points caractéristiques du visage sont affichés par des points blancs.

contenant les images avec une pose de la tête frontale ($-8° < \text{yaw} < 8°$ et $-5° < \text{pitch} < 5°$) et « Test pose » contenant les images avec de la pose ($-35° < \text{yaw} < -8°$, $8° < \text{yaw} < 35°$, $-35° < \text{pitch} < -5°$ ou $5° < \text{pitch} < 25°$). La réunion des sous-ensembles « Test frontal » et « Test pose » est appelée « Test total ». La performance de notre méthode est évaluée à l'aide d'une validation croisée « Leave-One-Subject-Out ». Le système est entraîné séparément sur le visage entier, les sourcils, les yeux et la bouche (voir sous-section 2.2.2).

Les résultats sont reportés dans la table 2.5. Les différentes configurations d'apprentissage donnent des résultats similaires sur le visage entier et les yeux. Sur les sourcils, les configurations « 4 poses 27°/20° » et « 8 poses » offrent une meilleure performance que les deux autres configurations, en particulier sur le test frontal. Sur la bouche, les configurations « Multi-vues 14°/10° » et « 8 poses » offrent une meilleure performance que les deux autres configurations, et ce quel que soit le sous-ensemble testé. La configuration « 8 poses » sera donc celle retenue par la suite. Les taux de reconnaissance obtenus sur « Test total » sont de 84.88% sur le visage entier, 81.56% sur les sourcils, 98.25% sur les yeux et 71.39% sur la bouche. De plus, nous obtenons des résultats similaires sur « Test frontal » et « Test pose », à part sur la bouche. La différence des résultats entre « Test frontal » et « Test pose » est de 3.13% sur le visage entier, de -1.81% sur les sourcils, 0.39% sur les yeux et 5.06% sur la bouche, ce qui montre la robustesse à la pose de notre système de reconnaissance d'expressions faciales. Les résultats sont moins bons sur la bouche car les classes du dégoût et de la tristesse sont proches dans cette zone, elles sont donc plus difficiles à discriminer l'une de l'autre.

TABLE 2.5 – Taux de reconnaissance de notre système de reconnaissance d'expressions faciales globale et locale sur la base de données FAST-pose (8 sujets). La performance est calculée avec la validation croisée « leave-one-subject-out ». Quatre systèmes sont considérés : sur le visage entier (6 classes), sur les sourcils (3 classes), les yeux (2 classes) et la bouche (6 classes) (voir sous-section 2.2.2 pour la définition des classes dans chaque zone). Pour chaque zone, 4 configurations sont considérées pour l'ensemble d'apprentissage : frontal (1 expression = 1 image de face), 4 poses 14°/10° (1 expression = 1 image de face + 4 images avec de la pose : +/-14° yaw et +/-10° pitch), 4 poses 27°/20° (1 expression = 1 image de face + 4 images avec de la pose : +/-27° yaw et +/-20° pitch), 8 poses (1 expression = 1 image de face + 8 images avec de la pose : +/-14° yaw, +/-27° yaw, +/-10° pitch et +/-20° pitch). Le test est effectué sur une vidéo complète où le sujet effectue l'expression de face puis conserve l'expression en bougeant la tête successivement selon le « yaw » et le « pitch ». De la vidéo, nous extrayons le sous-ensemble avec les images de face et le sous-ensemble avec les images avec de la pose, la performance calculée sur chacun de ces sous-ensemble est reportée dans la colonne « Test frontral » et « Test pose » respectivement. La performance calculée sur la vidéo en entier est reportée dans la colonne « Test total ».

Zone	Apprentissage	Test frontal	Test pose	Test total
Visage entier	Frontal	85.91%	82.04%	83.06%
	4 poses 14°/10°	86.25%	83.85%	84.64%
	4 poses 27°/20°	85.59%	83.04%	83.51%
	8 poses	**87.25%**	**84.12%**	**84.88%**
Sourcils	Frontal	70.57%	73.96%	75.48%
	4 poses 14°/10°	73.83%	77.56%	78.22%
	4 poses 27°/20°	**82.33%**	77.94%	80.16%
	8 poses	79.01%	**80.82%**	**81.56%**
Yeux	Frontal	98.96%	**98.02%**	**98.26%**
	4 poses 14°/10°	99.14%	97.92%	98.17%
	4 poses 27°/20°	**99.38%**	79.70%	97.98%
	8 poses	**99.38%**	97.99%	98.25%
Bouche	Frontal	68.28%	63.42%	64.71%
	4 poses 14°/10°	**75.97%**	**70.83%**	**71.94%**
	4 poses 27°/20°	69.99%	65.14%	66.73%
	8 poses	74.92%	69.86%	71.39%

Dans le chapitre 4, contenant les résultats de notre méthode d'adaptation du modèle expressif spécifique à la personne (voir chapitre 3), le détecteur global et les détecteurs locaux seront entraînés sur les 8 sujets de la base FAST-pose avec les classes définies dans la sous-section 2.2.2 en utilisant la configuration « 8 poses ».

2.2.5 Conclusion

Dans cette section, nous avons présenté notre système de reconnaissance d'expressions faciales robuste à la pose. Cet outil nous sera utile dans le chapitre suivant pour notre méthode d'adaptation non supervisée du modèle spécifique à la personne présenté dans la section 2.1. L'adaptation se déroule comme suit. Dans un premier temps, nous initialisons de manière supervisée le modèle avec les expressions de base plausibles, *i.e.* synthétisées (voir figure 2.1). Pour ce faire, nous détectons automatiquement le visage neutre du sujet avant de synthétiser les expressions de base plausibles. Ensuite, nous effectuons l'adaptation en tant que telle en remplaçant les expressions de base plausibles par les expressions de base réelles du sujet puis en les affinant. Le système de reconnaissance d'expressions faciales nous est utile à deux reprises : pour détecter automatiquement le visage neutre lors de l'initialisation du modèle et pour détecter les expressions de base réelles du sujet lors de l'adaptation en tant que telle.

Nous avons besoin à la fois d'une détection globale et locale pour pouvoir mettre à jour les expressions de base du modèle globalement (visage entier) ou localement (une seule zone du visage). Nous nous plaçons donc à un niveau intermédiaire entre les unités d'action du système FACS [6] et les expressions prototypiques d'Ekman [3] puisque nous devons détecter des déformations locales dans la zone des sourcils, des yeux ou de la bouche correspondant localement aux expressions de base du modèle. A notre connaissance, les méthodes de l'état de l'art pour la reconnaissance d'expressions faciales ne proposent pas de détection locale, nous avons donc développé notre propre système.

Nous suivons le schéma classique d'un système de reconnaissance d'expressions faciales (voir figure 1.5) : extraction des caractéristiques puis classification. Étant donné que dans notre cadre applicatif le sujet est dans un environnement non contraint, la pose de la tête est susceptible de varier, ce qui peut impacter négativement les performances du système de reconnaissance d'expressions faciales et donc les performances de l'adaptation. Nous avons donc cherché à développer un système de reconnaissance d'expressions faciales robuste à la pose. Pour ce faire, nous utilisons des caractéristiques composées d'angles et de distances entre certains points caractéristiques du visage. En effet, ces caractéristiques sont invariantes aux faibles variations de pose (voir annexe B). L'autre avantage de ces caractéristiques est qu'elles s'appliquent naturellement à la détection locale puisqu'il suffit de sélectionner les caractéristiques de la zone voulue pour faire une analyse locale. Nous entraînons donc 4 classifieurs : 1 global (visage entier) et 3 locaux (zone des sourcils, des yeux et de la bouche). Nous avons choisi comme classifieur une SVM avec pour noyau la RBF. Pour augmenter la robustesse à la pose, nous proposons

d'entraîner les classifieurs avec 1 image de face et 8 images avec de la pose pour chaque expression.

Nous avons d'abord comparé les performances de notre méthode avec des méthodes existantes pour une détection globale et frontale, *i.e.* les images d'apprentissage et de test sont toutes frontales. Les résultats montrent que notre méthode n'est pas aussi performante : l'écart du taux de reconnaissance avec la meilleure méthode est compris entre 4% et 7% selon la base de données et le nombre de classes considéré. Cependant, aucune des méthodes existantes ne propose une détection locale.

Nous avons ensuite testé la robustesse à la pose de notre méthode sur une base de données maison. Les résultats montrent que le fait d'avoir plusieurs poses différentes dans les images d'apprentissage permet d'augmenter la robustesse, en particulier sur les détecteurs locaux des sourcils et de la bouche. Si l'on compare les taux de reconnaissance obtenus sur les images de test frontales et les images de test avec de la pose, on constate que l'écart est compris entre 5% (pour la bouche) et 1% (pour les sourcils et les yeux), ce qui montre la robustesse à la pose de notre méthode.

Chapitre 3

Adaptation non supervisée du modèle expressif spécifique à la personne

Sommaire

Comme nous l'avons indiqué dans le chapitre introductif, le cadre applicatif visé par cette thèse est le milieu médical. L'application envisagée est l'analyse automatique de variation de comportement pour le maintien à domicile de personnes âgées. Il est alors nécessaire que l'apprentissage du modèle spécifique à la personne se fasse de manière non

supervisée. En effet, l'étape de calibration, nécessaire lors d'un apprentissage supervisé d'un modèle spécifique à la personne, peut s'étaler sur une durée non négligeable et être désagréable pour le sujet.

L'apprentissage du modèle basé sur [14] est faiblement supervisé dans le sens où uniquement le visage neutre du sujet est nécessaire pour l'apprentissage (voir section 2.1). Les expressions de base du modèle sont synthétisées à partir du visage neutre. Pour rendre l'apprentissage du modèle non supervisé, nous proposons dans un premier temps de détecter automatiquement le visage neutre du sujet (section 3.1). A ce stade, le modèle est construit sur des expressions de base synthétisées, elles ne rendent donc pas compte des déformations faciales réelles du sujet. Pour y remédier, nous proposons une méthode d'adaptation non supervisée du modèle spécifique à la personne pour pouvoir remplacer les expressions de base du modèle par les expressions de base réelles du sujet et ainsi mieux rendre compte de ses déformations faciales. Dans un premier temps cette adaptation se fait de manière séquentielle (section 3.2), c'est-à-dire que les expressions de base sont remplacées au fur et à mesure qu'elles apparaissent dans la vidéo. Dans un second temps, nous proposons une extension de l'adaptation non supervisée du modèle sur une fenêtre temporelle (section 3.3).

Le système final se présente de la manière suivante :

- Le sujet se présente devant la « webcam »,
- Le visage neutre du sujet est détecté automatiquement (voir section 3.1),
- Le modèle est initialisé avec le visage neutre détecté automatiquement et les 5 expressions de base synthétisées (aussi appelées expressions de base plausible, voir section 2.1 pour la construction du modèle),
- Le modèle est adapté avec les expressions du sujet contenues dans le flux vidéo de la « webcam » (voir section 3.2 pour l'adaptation séquentielle et la section 3.3 pour l'adaptation sur une fenêtre temporelle).

La preuve de concept du système final est réalisée sur des vidéos de plusieurs bases de données. Les résultats seront présentés dans le chapitre 4.

3.1 Initialisation du modèle spécifique à la personne

Pour l'initialisation du modèle, nous utilisons la méthode présentée dans la section 2.1. Si l'on se réfère à la figure 2.1, nous ajoutons la détection automatique du visage neutre en amont pour rendre la construction du modèle non supervisée. Lors de l'initialisation, le modèle est construit sur les expressions de base plausibles qui ont été synthétisées.

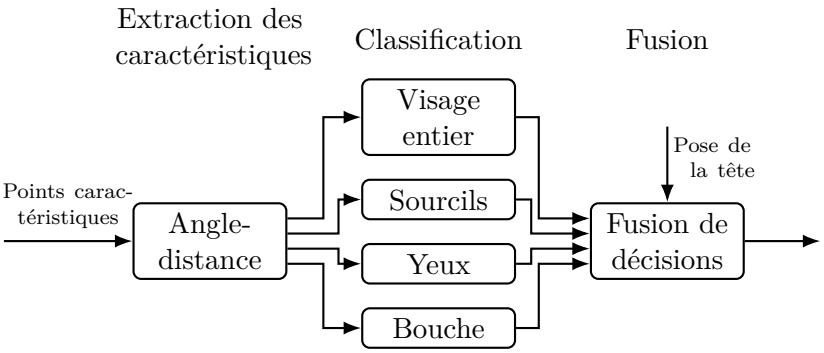


FIGURE 3.1 – Présentation générale de la détection automatique du visage neutre. Tout d'abord, les caractéristiques faciales sont extraites, elles se composent de 15 angles et 10 distances calculés entre des points caractéristiques spécifiques (voir table 2.2). Ensuite, le classifieur global (visage entier) et les 3 classifieurs locaux (sourcils, yeux et bouche) sont fusionnés pour détecter le visage neutre. La pose de la tête est prise en compte pour s'assurer que celle-ci soit minimale.

Pour détecter automatiquement le visage neutre à partir d'une vidéo, nous fusionnons les décisions d'un détecteur global et de 3 détecteurs locaux (sourcils, yeux et bouche) d'expressions faciales (voir figure 3.1). Nous utilisons notre système de reconnaissance d'expressions faciales robuste à la pose (voir section 2.2) pour chacun de ces détecteurs. La fusion de décisions consiste à ne retenir que les images de la vidéo où l'ensemble des 4 détecteurs globaux et locaux renvoient la classe neutre. L'idée est de s'assurer qu'un visage classé comme neutre globalement l'est aussi localement. Étant donné que les détecteurs locaux ne sont pas exempts d'erreur et peuvent faire l'objet d'un défaut de généralisation sur des sujets inconnus, il est possible qu'une expression soit localement incorrectement classifiée comme non neutre alors qu'elle l'est globalement et alors aucune image n'est retenue. Dans ce cas, nous sélectionnons les images avec une détection globale neutre et une seule détection locale non neutre. Une fois que toutes les images contenant un visage neutre sont sélectionnées, l'image retenue est choisie de telle sorte que la pose de la tête soit minime pour les angles de yaw et pitch. La pose de la tête est calculée par Intraface [168] lors du suivi des points caractéristiques.

3.2 Adaptation non supervisée du modèle spécifique à la personne de manière séquentielle

L'idée de l'adaptation non supervisée du modèle spécifique à la personne est de détecter, à la fois globalement et localement, les expressions de base réelles du sujet

afin de remplacer les expressions de base plausibles (synthétisées) du modèle puis de les affiner, tout en maintenant un ensemble de contraintes. Une présentation générale de la méthode est faite dans la sous-section 3.2.1. Les sous-sections suivantes détaillent chacune des étapes de la méthode.

3.2.1 Présentation générale

Si l'on se réfère à la figure 2.1, l'adaptation du modèle consiste à modifier les points caractéristiques des expressions de base avec les expressions de base réelles du sujet avant de calculer la PCA. Ainsi, ce que nous appelons la mise à jour du modèle est le fait de remplacer certaines zones des expressions de base du modèle avant le calcul d'une nouvelle PCA. La figure 3.2 donne une vue globale de notre méthode d'adaptation non supervisée du modèle spécifique à la personne de manière séquentielle. Le modèle est initialisé de manière non supervisée comme indiqué dans la section 3.1. L'adaptation est incrémentale et séquentielle : le modèle est mis à jour image par image, dans l'ordre d'apparence des expressions de base affichées par le sujet, globalement ou localement, dans le flux vidéo.

On appelle « expression courante » l'expression du sujet à l'instant courant dans le flux vidéo. Dans un premier temps, le système de reconnaissance des expressions de base associe à l'expression courante un label global et des labels locaux qui indiquent à quelle expression de base elle correspond (voir sous-section 3.2.2). Afin de s'assurer que l'expression courante est appropriée pour modifier le modèle, nous vérifions un ensemble de contraintes sur l'expression courante (voir sous-section 3.2.4). Ces contraintes sont en partie vérifiées dans l'espace PCA, c'est pourquoi nous projetons d'abord les points caractéristiques de l'expression courante dans l'espace PCA (voir sous-section 3.2.3). Si les contraintes sont vérifiées, alors l'expression courante devient une expression candidate et les points caractéristiques des expressions de base du modèle sont mis à jour avec l'expression candidate (voir sous-section 3.2.5). Ensuite, un nouveau modèle est calculé à partir des expressions de base mises à jour (voir sous-section 3.2.6). Enfin, nous vérifions une dernière contrainte sur la structure du nouveau modèle (voir sous-section 3.2.7, voir équation 2.6 pour la définition de la structure du modèle). Si la structure reste inchangée, alors le nouveau modèle devient le modèle courant et boucle sur les étapes précédentes.

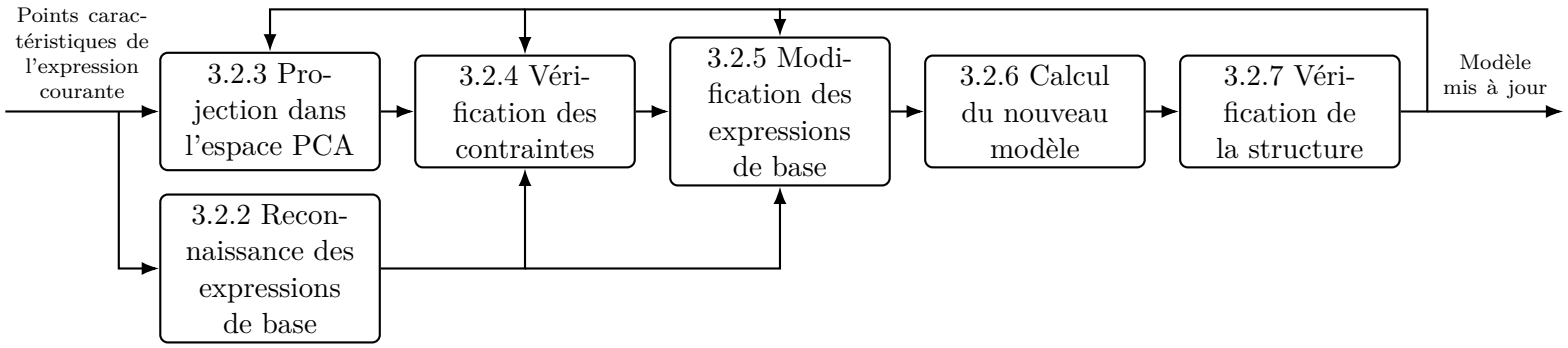


FIGURE 3.2 – Présentation générale de l'adaptation non supervisée du modèle spécifique à la personne. A chaque brique de l'algorithme correspond le numéro de la sous-section dans laquelle elle est décrite. Le but de l'adaptation est de modifier les expressions de base du modèle avec les expressions de base globales ou locales présentes dans le flux vidéo. Premièrement, l'expression courante du flux vidéo est classée globalement et localement parmi les classes correspondant aux expressions de base du modèle (voir sous-section 3.2.2). En parallèle, l'expression courante est projetée dans l'espace PCA (voir sous-section 3.2.3). Afin de s'assurer que le modèle ne sera pas mis à jour avec une expression courante entachée d'erreur, nous vérifions un ensemble de contraintes globales et locales (voir sous-section 3.2.4). Si les contraintes sont vérifiées, alors l'expression courante devient une expression candidate. Les points caractéristiques des expressions de base du modèle sont mis à jour avec l'expression candidate (voir sous-section 3.2.5). Ensuite, un nouveau modèle est calculé avec les expressions de base mises à jour (voir sous-section 3.2.6). Enfin, nous vérifions si la structure du modèle reste inchangée (voir sous-section 3.2.7, voir équation 2.6 pour la définition de la structure du modèle). Si tel est le cas, alors le nouveau modèle devient le modèle courant.

3.2.2 Reconnaissance des expressions de base

La reconnaissance des expressions de base est assurée par notre méthode présentée dans la section 2.2. Nous apprenons un classifieur global (visage entier) et 3 classifieurs locaux (sourcils, yeux et bouche) avec les caractéristiques angle-distance. Pour l'apprentissage, neuf images sont utilisées pour chaque expression : frontal, +/-14° yaw, +/-27° yaw, +/-10° pitch et +/-20° pitch. Nous associons à l'expression courante 4 labels : l_{global} (visage entier), $l_{sourcils}$ (sourcils), l_{yeux} (yeux) et l_{bouche} (bouche). Ces labels seront utilisés pour la vérification de contraintes et la modification des expressions de base.

3.2.3 Projection dans l'espace PCA

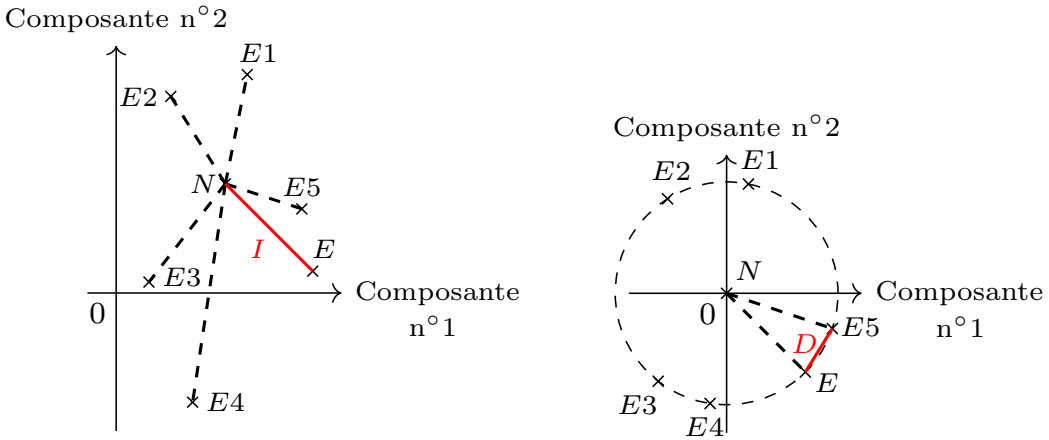
Afin d'analyser l'expression courante et de décider si elle peut devenir une expression candidate pour mettre à jour le modèle, nous projetons l'expression courante dans l'espace PCA du modèle courant et nous définissons dans cet espace deux variables qui seront utilisées pour la vérification de contraintes (voir sous-section 3.2.4). Ces deux variables donnent respectivement de l'information sur l'intensité et les déformations faciales de l'expression courante.

Projection de l'expression courante

Si la pose de la tête est trop forte, l'expression courante est frontalisée (voir annexe C) avant d'être projetée dans l'espace PCA. La pose de la tête est considérée comme trop forte si le « yaw » est inférieur à $-12°$ ou supérieur à $+12°$, ou si le « pitch » est inférieur à $-15°$ ou supérieur à $+15°$. Soient n_p le nombre de points caractéristiques et n_e le nombre d'expressions de base dans le modèle (neutre inclus), ici $n_p = 49$ et $n_e = 6$. Soit $X \in \mathbb{R}^{2n_p}$ le vecteur contenant les points caractéristiques de l'expression courante, frontalisée si nécessaire et alignée avec les expressions de base du modèle. On note $\overline{X^e} \in \mathbb{R}^{2n_p}$ la moyenne de l'ensemble des points caractéristiques des expressions de base du modèle (neutre inclus) après alignement et $\Phi \in M_{2n_p,3}(\mathbb{R})$ la matrice de projection dans l'espace PCA (voir section 2.1). On note $X_{PCA} \in \mathbb{R}^3$ la projection de X dans l'espace PCA et nous l'appelons le vecteur PCA de X :

$$X_{PCA} = (X - \overline{X^e})\Phi. \tag{3.1}$$

Pour tout $i \in [\![1, n_e]\!]$, on note $X^i_{PCA} \in \mathbb{R}^3$ le vecteur PCA de la i-ième expression de base du modèle courant. A partir du vecteur PCA, nous calculons l'intensité de l'expression



(a) Espace PCA

(b) Espace PCA centré normalisé

FIGURE 3.3 – Illustration de l'intensité et de la distance de déformation faciale. Bien que ces deux grandeurs soient calculées dans un espace à 3 dimensions, les deux figures sont des projections de l'espace PCA sur les deux premières dimensions, et ce à des fins illustratives. Le neutre est noté N, les 5 expressions de base du modèle $E1, E2, E3, E4, E5$ et l'expression courante E. Dans la figure 3.3a, les intensités des expressions de base sont représentées par des lignes noires pointillées et l'intensité de l'expression courante (I) est représentée par une ligne rouge. Dans la figure 3.3b, les vecteurs PCA sont centrés sur le neutre et normalisés. La distance de déformation faciale (D) entre l'expression courante et l'expression de base avec le label $l_{global} = E_5$ est représentée par une ligne rouge.

et la distance de déformation faciale de l'expression dans l'espace PCA relativement à l'expression de base du modèle avec le label l_{global} (voir figure 3.3).

Calcul de l'intensité

Puisque le neutre est au centre des déformations faciales dans l'espace PCA, l'intensité de l'expression est calculée comme la distance euclidienne entre son vecteur PCA et le vecteur PCA du neutre. Ainsi le visage neutre a une intensité nulle. On note $I \in \mathbb{R}$ l'intensité de l'expression courante :

$$I = \|X_{PCA} - X_{PCA}^{neutral}\|_2. \tag{3.2}$$

Pour tout $i \in [\![1, n_e]\!]$, on note $I^i \in \mathbb{R}$ l'intensité de la i-ième expression de base du modèle courant.

Calcul de la distance de déformation faciale

La seconde grandeur que nous calculons donne de l'information sur la proximité de la déformation faciale de l'expression courante avec celle de l'expression de base ayant le label l_{global}, et ce indépendamment de leur intensité respective. Cette grandeur est calculée comme la distance euclidienne dans l'espace PCA centré normalisé entre les vecteurs PCA de l'expression courante et de l'expression de base du modèle courant avec le label l_{global}.

On note $X_{PCAcn} \in \mathbb{R}^3$ le vecteur PCA centré normalisé de l'expression courante :

$$X_{PCAcn} = \frac{X_{PCA} - X_{PCA}^{neutral}}{I}. \tag{3.3}$$

Pour tout $i \in [\![1, n_e]\!]$, on note $X_{PCAcn}^{i} \in \mathbb{R}^3$ le vecteur PCA centré normalisé de la i-ième expression de base du modèle courant.

On note $D_{l_{global}} \in \mathbb{R}$ la distance entre les vecteurs PCA centrés normalisés de l'expression courante et de l'expression de base du modèle courant avec le label l_{global} :

$$D_{l_{global}} = \|X_{PCAcn} - X_{PCAcn}^{l_{global}}\|_2, \tag{3.4}$$

où X_{PCAcn} et $X_{PCAcn}^{l_{global}}$ sont définis par l'équation 3.3. On appelle cette grandeur la « distance de déformation faciale » entre l'expression courante et l'expression de base du modèle avec le label l_{global}.

3.2.4 Vérification des contraintes

Pour s'assurer que l'expression courante n'est pas entachée d'erreurs dues au suivi de points caractéristiques ou à une mauvaise détection et qu'elle est cohérente avec la structure du modèle, nous vérifions trois ensembles de contraintes : contraintes sur l'erreur de suivi des points caractéristiques et sur la pose, contraintes globales et contraintes locales. L'étape de vérification des contraintes se déroule comme suit. Si les contraintes sur l'erreur de suivi des points caractéristiques et sur la pose ne sont pas vérifiées, alors les contraintes globales et locales sont ignorées et l'expression courante est rejetée. Sinon, nous vérifions les contraintes globales. Si elles sont vérifiées, alors l'expression courante devient une expression candidate pour modifier globalement l'expression de base du modèle courant avec le label l_{global} et les contraintes locales sont ignorées. Sinon, nous vérifions les contraintes locales pour les zones des sourcils et de la bouche. Si elles sont vérifiées, alors l'expression courante devient une expression candidate pour mettre à jour localement les expressions de base du modèle courante avec le label local l_z, où

Algorithme 1 Étape de vérification des contraintes

Entrées: • Expression courante
• Modèle courant

if contraintes sur l'erreur de suivi des points caractéristiques et sur pose vérifiées **then**
  if contraintes globales vérifiées **then**
    - Expression courante devient expression candidate pour mettre à jour le modèle courant globalement
  else
    for $z \in \{sourcils, bouche\}$ **do**
      if contraintes locales vérifiées dans la zone z **then**
        - Expression courante devient expression candidate pour mettre à jour le modèle localement dans la zone z
      else
        - Expression courante rejetée
      end if
    end for
  end if
else
  - Expression courante rejetée
end if

$z \in \{sourcils, bouche\}$. L'algorithme 1 récapitule l'étape de vérification des contraintes. Les paragraphes suivants détaillent le calcul des différentes contraintes.

Contraintes sur l'erreur de suivi des points caractéristiques et sur la pose

D'abord, nous vérifions qu'il n'y a pas d'erreur manifeste dans le suivi des points caractéristiques de l'expression courante en imposant à l'erreur de reconstruction dans l'espace PCA d'être inférieure à un certain seuil :

$$\widetilde{X_{PCA}}(W_1 \ \dots \ W_{n_e})^T + \overline{X^e} \leq \epsilon, \tag{3.5}$$

où $\widetilde{X_{PCA}} \in \mathbb{R}^{n_e}$ est la projection de l'expression courante dans l'espace PCA sans la réduction de dimension, $\forall i \in [\![1, n_e]\!], W_i \in \mathbb{R}^{2n_p}$ est le i-ième vecteur propre de la PCA (voir équation 2.3) et $\epsilon \in \mathbb{R}$ est le seuil.

De plus, nous rejetons l'expression courante si la pose de la tête est si importante que l'information expressive commence à disparaître du visage. Nous considérons que cela est le cas si le « yaw » est inférieur à $-35°$ ou supérieur à $+35°$, ou si le « pitch » est inférieur à $-35°$ ou supérieur à $+25°$.

Contraintes globales

Premièrement, l'expression courante doit être détectée comme une expression non neutre, *i.e.* $l_{global} \neq neutre$.

Deuxièmement, nous vérifions la cohérence globale de l'expression courante : les labels locaux $l_{sourcils}$, l_{yeux} et l_{bouche} associés ensemble doivent correspondre au label global l_{global} (voir table 2.1).

Troisièmement, nous empêchons la mise à jour du modèle avec une expression courante de plus faible intensité que l'expression de base du modèle courant avec le label l_{global}. Ainsi, le modèle mis à jour ne peut que couvrir une plus grande plage d'intensité (voir équation 3.2 pour la définition de l'intensité). Pour ce faire, nous imposons un seuil minimum sur l'intensité de l'expression courante. Le seuil est défini comme l'intensité de l'expression de base du modèle courant avec le label l_{global} (voir équation 3.6). Parce que le sujet ne va pas nécessairement afficher des expressions plus intenses que les expressions de base plausibles (utilisées pour l'initialisation du modèle, voir section 2.1), ce seuil est utilisé seulement après avoir mis à jour l'expression de base avec le label l_{global} une première fois. Autrement, si l'expression de base avec le label l_{global} n'a jamais été mise à jour, *i.e.* est toujours une expression de base plausible, le seuil est défini comme le tiers de l'intensité minimum parmi les expressions de base du modèle courant (voir équation 3.6). Ainsi, il n'est pas possible de mettre à jour le modèle avec une expression de très faible intensité.

$$I > \begin{cases} min_{\substack{i \in [\![1, n_e]\!] \\ i \neq neutre}} (I^i)/3, & \text{si l'expression de base avec le label } l_{global} \text{ pas encore mise à jour} \\ I^{l_{global}}, & \text{sinon} \end{cases} \tag{3.6}$$

La dernière contrainte globale porte sur les déformations faciales de l'expression courante. Nous faisons l'hypothèse que l'initialisation du modèle avec les expressions de base plausibles donne une bonne approximation des expressions faciales réelles du sujet [14]. Nous devons donc mettre à jour le modèle avec l'expression courante si elle présente des déformations faciales assez proches de celles de l'expression de base du modèle courant avec le label l_{global}. Dans ce but, nous imposons un seuil maximal sur la distance de déformation faciale entre l'expression courante et l'expression de base du modèle avec le label l_{global} (voir équation 3.7 pour la contrainte et équation 3.4 pour la définition de la distance de déformation faciale). Dans le cas où $l_{global} = joie$, nous définissons un seuil spécifique qui est plus petit que le seuil pour les autres expressions

de base. Cela vise à rendre l'adaptation plus robuste à la parole puisque des expressions de bouche proches de la joie peuvent apparaître lors de la parole et être détectées à tort comme telles.

$$D_{l_{global}} < \begin{cases} d_{joie}, & \text{si } l_{global} = joie \\ d_{global}, & \text{sinon} \end{cases} \tag{3.7}$$
$$\text{où } (d_{joie}, d_{global}) \in \mathbb{R}^2 \text{ et } d_{joie} < d_{global}$$

Contraintes locales

Soit $z \in \{sourcils, bouche\}$ la zone dans laquelle vérifier les contraintes. Premièrement, nous vérifions que le label local l_z de l'expression courante n'est pas neutre.

Deuxièmement, nous nous prémunissons d'une mauvaise détection sur l'expression locale de la bouche en imposant un seuil maximal sur la distance de déformation faciale (voir équation 3.8). Partant du constat que la bouche rend compte d'une grande partie de l'information de déformation faciale dans l'espace PCA, cette contrainte assure que la bouche de l'expression courante est assez proche de l'expression de base du modèle courante avec le label local l_{bouche}. De même que pour la contrainte globale sur la distance de déformation faciale (voir équation 3.7), nous utilisons un seuil spécifique dans le cas où $l_{bouche} = joie$. De plus, nous vérifions aussi dans ce cas si $l_{global} = joie$ afin de gagner en robustesse face aux mauvaises détection de l'expression de joie.

$$D_{l_{bouche}} < \begin{cases} d_{joie}, & \text{si } l_{bouche} = joie \\ d_{local}, & \text{sinon} \end{cases} \tag{3.8}$$
$$\text{où } (d_{joie}, d_{global}) \in \mathbb{R}^2 \text{ et } d_{joie} < d_{local}$$

Troisièmement, si la zone z des différentes expressions de base avec le même label local l_z n'a jamais été mise à jour, aucune contrainte supplémentaire n'est vérifiée. L'expression courante locale devient une expression candidate pour mettre à jour localement la zone z des expressions de base du modèle courant avec label local l_z. Sinon, nous imposons à l'intensité locale de la zone z de l'expression courante d'être supérieure à celle des expressions de base du modèle courant avec le label local l_z.

Pour vérifier cette contrainte sur l'intensité locale, nous comparons les valeurs d'un sous-ensemble des caractéristiques angle-distance (voir section 2.2) de l'expression courante et des expressions de base concernées. Selon la valeur du label l_z, ces valeurs doivent

être minimum ou maximum pour l'expression la plus intense localement. Ci-dessous, nous détaillons comment cette comparaison est faite.

Soit $F \in \mathbb{R}^{n_c}$ le vecteur contenant les caractéristiques angle-distance de l'expression courante, où n_c est le nombre de caractéristiques angle-distance (voir table 2.2), ici $n_c = 25$. Pour tout $j \in [\![1, n_c]\!]$, on note F_j la j-ième caractéristique de F. On note $e_1, ..., e_m$ les indices correspondant aux expressions de bases du modèle courant avec le même label local l_z. Pour tout $k \in [\![1, m]\!], F^{e_k} \in \mathbb{R}^{n_c}$ est le vecteur contenant les caractéristiques angle-distance de l'expression de base e_k. Soit $Min(l_z)$ (respectivement $Max(l_z)$) l'ensemble des indices des caractéristiques angle-distance à minimiser (respectivement maximiser) pour avoir l'expression avec le label local l_z la plus intense localement. L'expression courante est localement plus intense que les expressions de base du modèle courant avec le label local l_z si $\forall j \in Min(l_z), min(F_j, F_j^{e_1}, ..., F_j^{e_m}) = F_j$ et $\forall j \in Max(l_z), max(F_j, F_j^{e_1}, ..., F_j^{e_m}) = F_j$.

La table 3.1 donne la définition des ensembles $Min(l_z)$ et $Max(l_z)$ avec $z \in \{$ *sourcils, bouche* $\}$ et la figure 3.4 illustre un exemple de vérification de la contrainte locale sur l'intensité. Dans cet exemple, les deux premières images sont les expressions de base du modèle de tristesse et de surprise. La troisième image (à droite) est l'expression courante (avec les labels locaux $l_{sourcils} = levés$ et $l_{bouche} = surprise$) sur laquelle on vérifie la contrainte locale d'intensité. Pour vérifier la contrainte d'intensité sur les sourcils, nous comparons les ensembles $Min(l_{sourcils} = levés)$ et $Max(l_{sourcils} = levés)$ des expressions de base du modèle ayant le label local $l_{sourcils} = levés$ (donc la tristesse et la surprise) et de l'expression courante. Dans l'exemple, l'expression courante ne minimise qu'une seule valeur de $Min(l_{sourcils} = levés)$ et ne maximise aucune valeur de l'ensemble $Max(l_{sourcils} = levés)$, elle ne vérifie donc pas la contrainte locale sur l'intensité des sourcils. Pour vérifier la contrainte d'intensité sur la bouche, nous comparons les ensembles $Min(l_{bouche} = surprise)$ et $Max(l_{bouche} = surprise)$ de l'expression de base du modèle ayant le label local $l_{bouche} = surprise$ (donc la surprise) et de l'expression courante. Dans l'exemple, l'expression courante minimise toutes les valeurs de $Min(l_{sourcils} = levés)$ et maximise toutes les valeurs de l'ensemble $Max(l_{sourcils} = levés)$, elle vérifie donc la contrainte locale sur l'intensité de la bouche.

L'ensemble des contraintes locales n'est pas vérifié sur la région des yeux car le détecteur local associé n'est pas entraîné pour discriminer plusieurs niveaux de fermeture et d'ouverture des yeux. Il ne donne que les labels « ouverts » et « moitié-fermés/fermés », ce n'est donc pas assez précis pour proposer une expression candidate qui puisse mettre à jour localement les expressions de base du modèle courant.

TABLE 3.1 – Définition des ensembles $Min(l_z)$ et $Max(l_z)$ pour $z \in \{eyeborws, bouche\}$. $Min(l_z)$, respectivement $Max(l_z)$, est l'ensemble des indices des caractéristiques angle-distance à minimiser, respectivement maximiser, pour avoir l'expression avec le label local l_z la plus intense localement. La définition des caractéristiques angle-distance est donnée dans la table 2.2.

Zone	Label	Ensemble des caractéristiques angle-distance à minimiser	Ensemble des caractéristiques angle-distance à maximiser
Sourcils	Froncés	$\{4, 5, 6, 7, 8, 9\}$	$\{1, 2, 3\}$
	Levés	$\{1, 2, 3\}$	$\{4, 5, 6, 7, 8, 9\}$
Bouche	Colère	$\{14, 17, 22, 23, 24, 25\}$	$\{\}$
	Dégoût	$\{14, 17, 18, 19, 23, 24, 25\}$	$\{15, 16, 22\}$
	Joie	$\{15, 16\}$	$\{20, 21, 22, 23, 24, 25\}$
	Tristesse	$\{14, 17, 18, 19, 23, 24, 25\}$	$\{15, 16, 22\}$
	Surprise	$\{15, 16\}$	$\{14, 17, 18, 19, 20, 21, 23, 24, 25\}$

3.2.5 Modification des expressions de base

Une fois qu'une expression candidate a vérifié les contraintes, nous calculons de nouvelles expressions de base afin de mettre à jour le modèle.

Dans le cas où l'expression candidate met à jour globalement l'expression de base du modèle avec le label l_{global}, les points caractéristiques de ce dernier sont entièrement remplacés par ceux de l'expression candidate. Si l'on se réfère à la définition des expressions de base dans la table 2.1, les sourcils sont les mêmes pour la colère et le dégoût ($l_{sourcils} = froncés$, unité d'action 4) et pour la tristesse et la surprise ($l_{sourcils} = levés$, unités d'action 1+2). Dans le but de maintenir la cohérence de la structure du modèle, si l'une de ces expressions de base est remplacée globalement par l'expression candidate, alors l'autre expression de base avec le même label local $l_{sourcils}$ est mis à jour localement dans la zone des sourcils. La figure 3.5 donne un exemple de la mise à jour globale de la surprise, on peut voir que les sourcils de la tristesse sont aussi mis à jour.

Soit $z \in \{sourcils, bouche\}$. Dans le cas où l'expression candidate met à jour localement les expressions de base avec le même label local l_z, uniquement les points caractéristiques dans la zone z sont remplacés par les points caractéristiques de l'expression candidate dans la zone z. De même que pour la mise à jour globale, si $z = sourcils$, les deux expressions de base avec le même label local $l_{sourcils}$ (froncés ou levés) sont mis à jour dans la zone des sourcils.

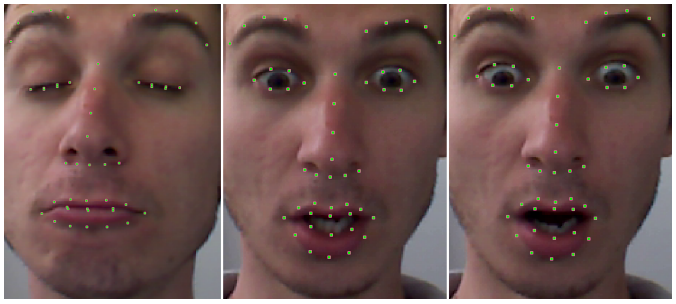
	Expression de base du modèle (tristesse)	Expression de base du modèle (surprise)	Expression courante ($l_{sourcils} = levés$ et $l_{bouche} = surprise$)
			
$Min(l_{sourcils} = levés)$	$\mathbf{F_1 = 1.480}$ $\mathbf{F_2 = 1.634}$ $F_3 = 1.877$	$F_1 = 1.851$ $F_2 = 1.901$ $F_3 = 1.831$	$F_1 = 1.722$ $F_2 = 1.719$ $\mathbf{F_3 = 1.745}$
$Max(l_{sourcils} = levés)$	$\mathbf{F_4 = 1.652}$ $\mathbf{F_5 = 1.887}$ $\mathbf{F_6 = 2.024}$ $\mathbf{F_7 = 1.718}$ $\mathbf{F_8 = 1.891}$ $\mathbf{F_9 = 2.056}$	$F_4 = 1.357$ $F_5 = 1.609$ $F_6 = 1.856$ $F_7 = 1.230$ $F_8 = 1.519$ $F_9 = 1.772$	$F_4 = 1.461$ $F_5 = 1.702$ $F_6 = 1.919$ $F_7 = 1.348$ $F_8 = 1.631$ $F_9 = 1.864$
$Min(l_{bouche} = surprise)$		$F_{15} = 2.586$ $F_{16} = 2.558$	$\mathbf{F_{15} = 2.577}$ $\mathbf{F_{16} = 2.546}$
$Max(l_{bouche} = surprise)$		$F_{14} = 1.618$ $F_{17} = 1.724$ $F_{18} = 0.919$ $F_{19} = 0.957$ $F_{20} = 0.495$ $F_{21} = 0.671$ $F_{23} = 1.178$ $F_{24} = 1.191$ $F_{25} = 1.176$	$\mathbf{F_{14} = 1.833}$ $\mathbf{F_{17} = 2.108}$ $\mathbf{F_{18} = 1.057}$ $\mathbf{F_{19} = 1.126}$ $\mathbf{F_{20} = 0.758}$ $\mathbf{F_{21} = 0.964}$ $\mathbf{F_{23} = 1.295}$ $\mathbf{F_{24} = 1.347}$ $\mathbf{F_{25} = 1.317}$

FIGURE 3.4 – Exemple de vérification de la contrainte locale sur l'intensité. Sous chaque expression se trouvent les valeurs des caractéristiques angle-distance (voir 2.2 pour leur définition) regroupées par ensemble ($Min(l_{sourcils} = levés)$, $Max(l_{sourcils} = levés)$, $Min(l_{bouche} = surprise)$ et $Max(l_{bouche} = surprise)$). Dans l'exemple, l'expression courante ne vérifie pas la contrainte locale sur l'intensité des sourcils, mais vérifie la contrainte locale sur l'intensité de la bouche.

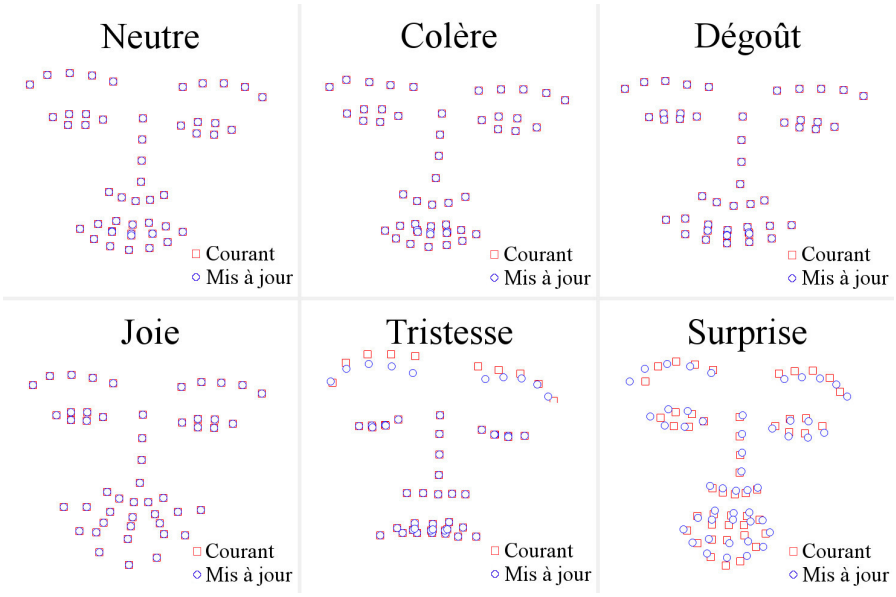


FIGURE 3.5 – Exemple de modification des expressions de base. L'expression candidate met à jour globalement l'expression de base avec le label $l_{global} = surprise$ pour laquelle $l_{sourcils} = levés$. Les sourcils de l'expression de base avec le label $l_{global} = tristesse$ (pour laquelle $l_{sourcils} = levés$ également) sont aussi remplacés par les sourcils de l'expression courante.

3.2.6 Mise à jour du modèle

L'étape suivante est le calcul d'un nouveau modèle. Si l'on se réfère à la figure 2.1, nous avons juste à calculer à la PCA et la tessellation de Delaunay à partir des expressions de bases mises à jour avec l'expression candidate.

3.2.7 Contrainte sur la structure du modèle

Après avoir calculé le nouveau modèle, nous vérifions une dernière contrainte sur sa structure (voir équation 2.6). Si la structure du nouveau modèle reste inchangée, alors le nouveau modèle remplace le modèle courant et boucle sur les étapes précédentes de l'adaptation (voir figure 3.2). Sinon, deux cas de figures se présentent. Soit l'expression candidate a vérifié les contraintes globales et dans ce cas l'expression candidate redevient l'expression courante et nous reprenons le processus à la vérification des contraintes locales. Soit l'expression candidate a vérifié les contraintes locales et dans ce cas le nouveau modèle est rejeté. Ainsi, nous pouvons détecter une expression candidate locale qui vérifie les contraintes globales mais change la structure du modèle après une mise à jour globale.

3.3 Adaptation non supervisée du modèle spécifique à la personne sur une fenêtre temporelle

Dans la section précédente, nous avons présenté notre méthode d'adaptation non supervisée du modèle expressif spécifique à la personne de manière séquentielle, c'est-à-dire image par image. Le défaut de l'adaptation séquentielle est le nombre potentiellement important de mises à jour. En effet, une expression de base peut être mise à jour autant de fois que nécessaire entre l'« onset » de l'expression (formation de l'expression sur le visage) et son « apex » (expression à son intensité maximale) dans la vidéo. Nous proposons donc dans cette section une extension de l'adaptation sur une fenêtre temporelle. L'idée est de sélectionner, avant d'effectuer l'adaptation, quelles sont les expressions candidates les plus intenses pour chaque expression de base du modèle.

Dans la sous-section 3.3.1 nous décrivons le principe général de l'adaptation sur une fenêtre temporelle. Ensuite, dans la sous-section 3.3.2 nous présentons plus en détail comment s'effectue la sélection des expressions candidates sur une fenêtre temporelle.

3.3.1 Présentation générale

Dans cette sous-section nous présentons le principe général de l'adaptation du modèle sur une fenêtre temporelle. Dans un premier temps, nous décrivons la méthode adoptée. Nous l'appliquons ensuite sur un exemple.

Méthode

Très schématiquement, l'adaptation du modèle sur une fenêtre temporelle se déroule en deux temps : sélection des expressions candidates globales sur la fenêtre temporelle et mise à jour globale du modèle avec ces expressions candidates, puis sélection des expressions candidates locales sur la fenêtre temporelle et mise à jour locale du modèle avec ces expressions candidates. L'algorithme 2 décrit l'adaptation du modèle sur une fenêtre temporelle.

Soit $T \in \mathbb{N}$ la longueur de la fenêtre temporelle en nombre d'images. La fenêtre démarrant à l'image d'indice t se termine à l'image d'indice $t+T-1$. La fenêtre suivante commence à l'image d'indice $t+T$ et se termine à l'image d'indice $t+2T-1$.

Soit Y la liste contenant les expressions courantes sur une fenêtre temporelle de taille T. La première étape est de déterminer les labels $l_{global}, l_{sourcils}, l_{yeux}$ et l_{bouche} de chacune des expressions de Y. Comme dans l'adaptation du modèle de manière séquentielle (voir sous-section 3.2.2), le système utilisé pour la reconnaissance d'expressions faciales

Algorithme 2 Adaptation du modèle sur une fenêtre temporelle

Entrées:
- Liste des expressions courantes sur la fenêtre temporelle (Y)
- Liste ordonnée arbitrairement des labels globaux des expressions de base à mettre à jour (L_{global})
- Liste ordonnée arbitrairement des labels locaux des expressions de base à mettre à jour dans la zone des sourcils ($L_{sourcils}$)
- Liste ordonnée arbitrairement des labels locaux des expressions de base à mettre à jour dans la zone de la bouche (L_{bouche})

for expression courante $\in Y$ **do**
- Détermination des labels $l_{global}, l_{sourcils}, l_{yeux}, l_{bouche}$ de l'expression courante

end for
for $l_g \in L_{global}$ **do**
- Sélection de l'expression candidate globale $\in Y$ avec le label $l_{global} = l_g$ et mise à jour globale du modèle

end for
- Suppression des expressions candidates globales dans Y

for $z \in \{sourcils, bouche\}$ (ou $\{bouche, sourcils\}$ selon l'ordre choisi) **do**
  for $l_l \in L_z$ **do**
    - Sélection de l'expression candidate locale $\in Y$ avec le label $l_z = l_l$ et mise à jour locale du modèle
  end for
end for

est notre méthode présentée dans la section 2.2 avec un classifieur global (visage entier) et 3 classifieurs locaux (sourcils, yeux et bouche). Pour l'apprentissage, neuf images différentes sont utilisées pour chaque expression : frontal, +/-14° yaw, +/-27° yaw, +/-10° pitch et +/-20° pitch.

Ensuite, nous effectuons la mise à jour globale du modèle sur la fenêtre temporelle. Soit L_{global} la liste des labels globaux des expressions de base à mettre à jour. Les labels globaux sont {*colère, dégoût, joie, surprise, tristesse*}. L'ordre des labels dans L_{global} est fixé arbitrairement. Pour chaque label $l_g \in L_{global}$, nous sélectionnons dans la liste Y l'expression candidate globale avec le label $l_{global} = l_g$ et nous mettons à jour le modèle globalement avec cette expression candidate.

Après cela, nous répétons les mêmes opérations de sélection d'expressions candidates pour la mise à jour locale du modèle sur la fenêtre temporelle. Seuls les labels de la bouche et des sourcils sont investigués pour la mise à jour locale. Pour éviter de sélectionner une expression candidate locale qui a déjà été sélectionnée comme expression candidate globale (et a donc déjà mis à jour le modèle), nous supprimons les expressions candidates

globales de Y avant de sélectionner les expressions candidates locales. Soit $L_{sourcils}$ (respectivement L_{bouche}) la liste des labels locaux des expressions de base à mettre à jour, les labels locaux sont {*froncés, levés*} (respectivement {*colère, dégoût, joie, surprise, tristesse*}). L'ordre des labels dans $L_{sourcils}$ et L_{bouche} est fixé arbitrairement. Nous fixons aussi arbitrairement laquelle des deux zones sera mise à jour en première sur la fenêtre temporelle.

Exemple

La figure 3.6 illustre un exemple d'adaptation non supervisée du modèle sur une fenêtre temporelle de taille $T = 9$ démarrant à l'image d'indice t. Pour tout $i \in [\![1, 9]\!]$, on note E_i l'expression à l'image d'indice $t + i - 1$ et on lui associe un label global $l_{global} \in \{l_g^0, l_1^g, l_g^2, l_g^3, l_g^4, l_g^5\} = \{$*neutre, colère, dégoût, joie, tristesse, surprise*$\}$, un label local au niveau de la bouche $l_{bouche} \in \{l_b^0, l_b^1, l_b^2, l_b^3, l_b^4, l_b^5\} = \{$*neutre, colère, dégoût, joie, tristesse, surprise*$\}$ et un label local au niveau des sourcils $l_{sourcils} \in \{l_s^0, l_s^1, l_s^2\} = \{$*neutre, froncés, levés*$\}$.

Dans un premier temps, nous sélectionnons les expressions candidates globales sur la fenêtre temporelle et nous mettons successivement le modèle à jour globalement avec ces expressions candidates. Dans cet exemple, l'ordre de sélection est $L_{global} = \{l_1^g, l_g^2, l_g^3, l_g^4, l_g^5\}$ et seules les expressions de base du modèle avec les labels globaux $\{l_g^1, l_g^2, l_g^5\}$ sont mises à jour.

Après avoir mis à jour le modèle avec les expressions candidates globales, ces dernières sont supprimées de la fenêtre temporelle. Ensuite, nous sélectionnons les expressions candidates locales sur la fenêtre temporelle et nous mettons successivement le modèle à jour localement avec ces expressions candidates. Dans cet exemple, l'ordre de sélection est $L_{bouche} = \{l_b^1, l_b^2, l_b^3, l_b^4, l_b^5\}$ suivi de $L_{sourcils} = \{l_s^1, l_s^2\}$. Au niveau de la bouche, les expressions de base du modèle avec les labels locaux $\{l_2^b, l_b^3, l_b^4\}$ sont mises à jour. Au niveau des sourcils, les expressions de base du modèle avec les labels locaux $\{l_s^1, l_s^2\}$ sont mises à jour. Cet exemple montre qu'il est possible de mettre à jour localement un expression de base qui a déjà été mise à jour avec une expression candidate globale lors de la sélection des expressions candidates globales. Cela s'illustre par la sélection de l'expression candidate locale E_7 avec le label local l_b^2 alors que l'expression candidate E_2 avec le label global l_g^2 (partageant le même label local au niveau de la bouche) a déjà été sélectionnée. De même au niveau des sourcils, l'expression candidate locale E_3 (respectivement E_6) met à jour localement les expressions de base avec le label l_s^1 (respectivement l_s^2) alors qu'elles ont déjà été mises à jour avec les expressions candidates globales E_4 et E_2 (respectivement l'expression candidate gobale E_9).

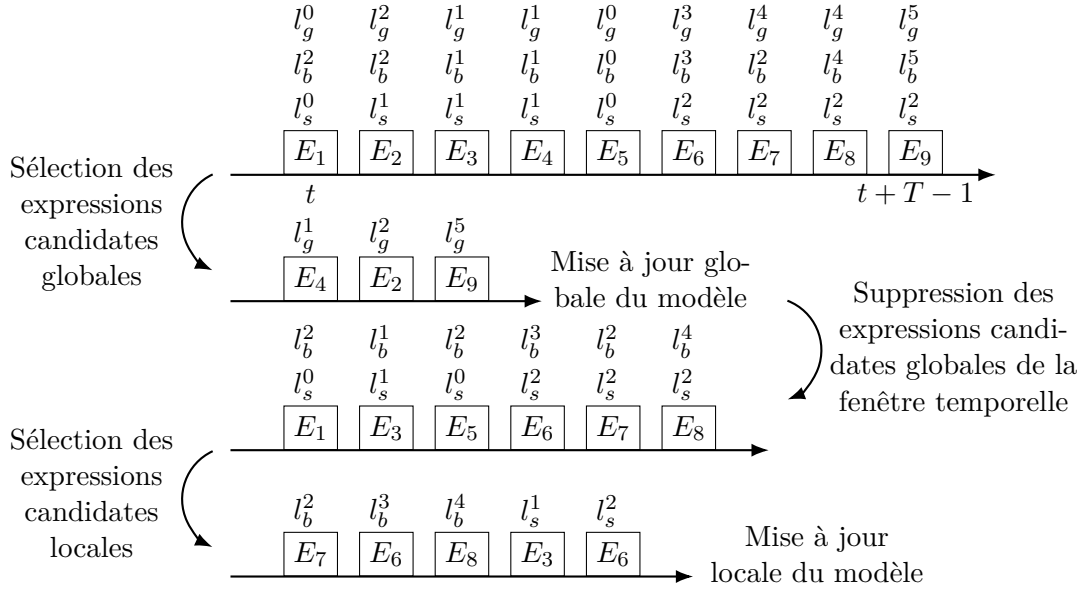


FIGURE 3.6 – Exemple d'adaptation sur une fenêtre temporelle de longueur T démarrant à l'instant t (dans cet exemple $T = 9$). Pour tout $i \in [\![1, 9]\!]$, on fait référence à l'expression courante à l'instant $t + i - 1$ par E_i et on lui associe un label global $l_{global} \in \{l_g^0, l_1^g, l_g^2, l_g^3, l_g^4, l_g^5\}$, un label local au niveau de la bouche $l_{bouche} \in \{l_b^0, l_b^1, l_b^2, l_b^3, l_b^4, l_b^5\}$ et un label local au niveau des sourcils $l_{sourcils} \in \{l_s^0, l_s^1, l_s^2\}$. Dans chaque cas, l^0 est le label de l'état neutre. Dans un premier temps, nous sélectionnons les expressions candidates globales sur la fenêtre temporelle et nous mettons successivement le modèle à jour globalement avec ces expressions candidates. Dans cet exemple, l'ordre de sélection est $L_{global} = \{l_g^1, l_g^2, l_g^3, l_g^4, l_g^5\}$ et seules les expressions de base du modèle avec les labels globaux $\{l_g^1, l_g^2, l_g^5\}$ sont mises à jour. Ensuite, nous supprimons les expressions candidates globales de la fenêtre temporelle. Enfin, nous sélectionnons les expressions candidates locales sur la fenêtre temporelle et nous mettons successivement le modèle à jour localement avec ces expressions candidates. Dans cet exemple, l'ordre de sélection est $L_{bouche} = \{l_b^1, l_b^2, l_b^3, l_b^4, l_b^5\}$ suivi de $L_{sourcils} = \{l_s^1, l_s^2\}$ et seules les expressions de base du modèle avec les labels locaux $\{l_2^b, l_b^3, l_b^4, l_s^1, l_s^2\}$ sont mises à jour.

3.3.2 Sélection de l'expression candidate sur une fenêtre temporelle

Dans cette sous-section nous décrivons comment est sélectionnée l'expression candidate sur une fenêtre temporelle. Dans un premier temps, nous décrivons la méthode adoptée. Nous l'appliquons ensuite sur un exemple.

Méthode

Le principe de base de la sélection de l'expression candidate sur une fenêtre temporelle est de chercher, pour l'expression de base à mettre à jour, quelle est l'expression candidate (*i.e.* l'expression courante vérifiant les contraintes dans le modèle courant, voir sous-section 3.2.4) d'intensité la plus grande dans la fenêtre temporelle et de mettre à jour le modèle avec cette expression candidate. L'algorithme 3 décrit la sélection d'une expression candidate sur une fenêtre temporelle.

Soit Y la liste contenant les expressions courantes sur la fenêtre temporelle. Pour chaque expression de Y, les labels l_{global}, $l_{sourcils}$, l_{yeux} et l_{bouche} ont déjà été déterminés. Soit l le label (global ou local) de l'expression de base à mettre à jour. Selon la nature du label l, nous sélectionnons une expression candidate globale ou locale.

La première étape est d'extraire de Y la sous-liste Y_l ne contenant que les expressions courantes avec le label l.

Ensuite, chaque expression courante de Y_l est projetée dans l'espace PCA du modèle courant (voir sous-section 3.2.3) et son intensité est calculée. Si l'expression candidate à sélectionner est globale (*i.e.* le label l est global), nous reprenons la définition de l'intensité présentée dans la sous-section 3.2.3 : l'intensité des expressions est calculée dans le modèle courant avec l'équation 3.2. Si l'expression candidate à sélectionner est locale (*i.e.* le label l est local), nous ne pouvons pas utiliser la même définition de l'intensité car celle-ci rend uniquement compte de l'intensité de l'expression dans sa globalité et ne contient aucune information locale. Nous proposons donc une définition de l'intensité locale basée sur les valeurs des caractéristiques angle-distance dans la zone concernée. Nous reprenons les ensembles $Min(l_{sourcils})$, $Max(l_{sourcils})$, $Min(l_{bouche})$ et $Max(l_{bouche})$ définis dans la table 3.1. On note $F \in \mathbb{R}^{n_c}$ le vecteur contenant les caractéristiques angle-distance de l'expression, où $n_c = 25$ est le nombre de caractéristiques (voir table 2.2 pour la définition des caractéristiques angle-distance). Pour tout $j \in [\![1, n_c]\!]$, on note F_j la j-ième caractéristique de F. L'intensité locale est calculée par l'équation suivante :

$$I_{local} = \sum_{j \in Max(l)} F_j - \sum_{j \in Min(l)} F_j. \tag{3.9}$$

Après avoir projeté toutes les expressions courantes de la liste Y_l et avoir calculé leur intensité globale ou locale, nous trions les éléments la liste Y_l par intensité décroissante.

Nous cherchons ensuite à mettre à jour le modèle (globalement ou localement selon la nature du label l) avec la première expression candidate de Y_l vérifiant la contrainte sur la structure du modèle. Pour cela, nous prenons la première expression courante de Y_l (étant donné que la liste Y_l a été triée, le premier élément est l'expression avec

la plus grande intensité), nous vérifions les contraintes sur l'erreur de suivi des points caractéristiques et sur la pose, puis les contraintes globales ou locales selon la nature du label l (voir sous-section 3.2.4). Si les contraintes sont vérifiées, alors l'expression courante devient une expression candidate. Nous calculons alors les nouvelles expressions de base avec cette expression de base (voir sous-section 3.2.5) et nous calculons le nouveau modèle (voir sous-section 3.2.6). Si la contrainte sur la structure du modèle (voir sous-section 3.2.7) est vérifiée pour le nouveau modèle, alors le modèle courant est remplacé par le nouveau modèle. Ainsi le modèle est mis à jour et le processus de sélection de l'expression candidate s'arrête. Sinon, nous reprenons le processus à l'étape de vérification des contraintes en prenant l'expression courante suivante dans la liste Y_l.

Exemple

La figure 3.7 illustre un exemple de sélection de l'expression candidate sur une fenêtre temporelle de taille $T = 9$ démarrant à l'image d'indice t. Pour tout $i \in [\![1, 9]\!]$, on note E_i l'expression à l'image d'indice $t + i - 1$, on lui associe un label $l \in \{l^0, l^1, l^2\}$ et son intensité I_i. Dans cet exemple, nous sélectionnons l'expression candidate pour le label l^1. Il peut s'appliquer à la fois pour la sélection d'une expression candidate globale ou d'une expression candidate locale.

La première étape est d'extraire les expressions avec le label l^1, nous obtenons alors la liste d'expressions $\{E_3, E_4, E_5, E_6, E_9\}$. La liste est ensuite triée par intensité décroissante, elle devient alors $\{E_4, E_5, E_3, E_6, E_9\}$.

Ensuite, nous vérifions les contraintes sur le premier élément de la liste, l'expression E_4. Dans cet exemple, les contraintes ne sont pas vérifiées, nous passons donc à l'expression suivante, E_5. Dans cet exemple, les contraintes sont vérifiées pour l'expression E_5. Elle devient donc une expression candidate et nous mettons à jour le modèle avec elle. Dans cet exemple, la contrainte sur la structure du modèle n'est pas vérifiée pour le modèle mis à jour avec E_5, elle est donc rejetée et nous passons à l'expression E_3. Les contraintes sont vérifiées pour E_3, elle devient donc une expression candidate. Le modèle est mis à jour avec E_3 et la contrainte sur la structure est vérifiée. L'expression E_3 est donc sélectionnée comme l'expression candidate pour le label l^1 sur la fenêtre temporelle. Le modèle courant est remplacé par le modèle mis à jour avec l'expression candidate E_3.

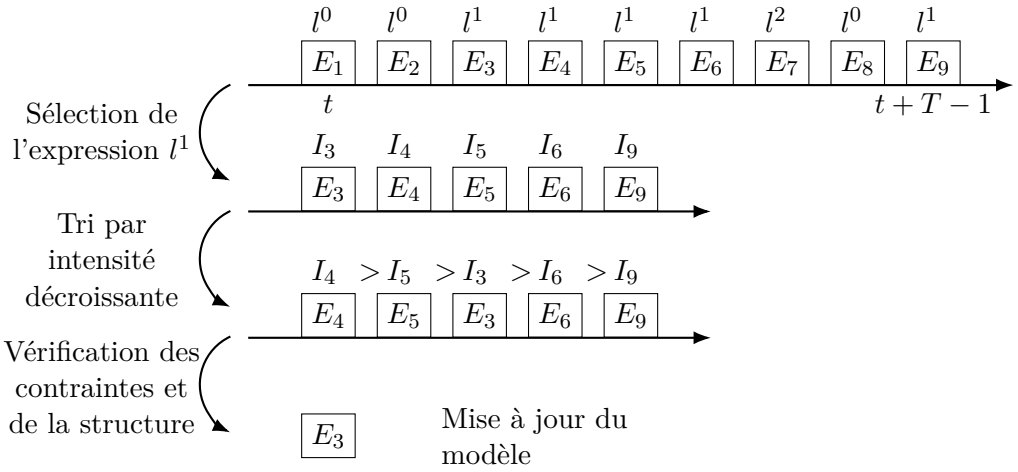


FIGURE 3.7 – Exemple de sélection de l'expression candidate pour le label l^1 sur une fenêtre temporelle de longueur T démarrant à l'image d'indice t (dans cet exemple $T = 9$). Pour tout $i \in [\![1, 9]\!]$, on fait référence à l'expression courante à l'image d'indice $t+i-1$ par E_i et on lui associe un label $l \in \{l^0, l^1, l^2\}$ et son intensité I_i. Dans un premier temps, nous extrayons de la fenêtre temporelle uniquement les expressions avec le label l^1. Puis la sous-liste d'expressions ainsi extraite est triée par intensité décroissante. Ensuite, nous retenons la première expression de la sous-liste triée vérifiant les contraintes et laissant la structure du modèle inchangée après la mise à jour de celui-ci. A l'issue de la sélection de l'expression candidate, le modèle est mis à jour.

Algorithme 3 Sélection de l'expression candidate avec le label l sur une fenêtre temporelle et mise à jour du modèle

Entrées: • Liste des expressions courantes (dont les labels ont déjà été déterminés) sur la fenêtre temporelle (Y)
• Label l (global ou local) de l'expression de base à mettre à jour

- Extraction de la sous-liste $Y_l \subseteq Y$ des expressions courantes avec le label l
for expression courante $\in Y_l$ **do**
  - Projection de l'expression courante dans l'espace PCA du modèle courant (voir sous-section 3.2.3)
  - Calcul de l'intensité (globale ou locale selon la nature du label l) de l'expression courante (voir équation 3.2 pour la définition de l'intensité globale et équation 3.9 pour la définition de l'intensité locale)
end for
- Tri des éléments de Y_l par intensité décroissante des expressions
while modèle non mis à jour **do**
  - Expression courante $\leftarrow$ premier élément de Y_l
  - Vérification des contraintes sur l'erreur de suivi des points caractéristiques et sur la pose (voir sous-section 3.2.4)
  - Vérification des contraintes globales ou locales selon la nature du label l (voir sous-section 3.2.4)
  if contraintes vérifiées **then**
    - Expression candidate avec le label $l \leftarrow$ expression courante
    - Modification des expressions de base avec l'expression candidate (voir sous-section 3.2.5)
    - Calcul du nouveau modèle (voir sous-section 3.2.6)
    - Vérification de la contrainte sur la structure du nouveau modèle (voir sous-section 3.2.7)
    if contrainte sur la structure du nouveau modèle vérifiée **then**
      - Modèle courant $\leftarrow$ nouveau modèle
      - Modèle mis à jour
    else
      - Suppression du premier élément de Y_l
    end if
  end if
end while

Chapitre 4

Résultats de l'adaptation non supervisée

Sommaire

Dans le chapitre précédent, nous avons présenté notre méthode d'adaptation non supervisée du modèle expressif spécifique à la personne. Nous nous basons sur la représentation invariante des expressions faciales de [14] pour construire le modèle expressif spécifique à la personne (voir section 2.1). La première contribution de cette thèse est de construire le modèle de manière non supervisée en détectant automatiquement le visage neutre du sujet (voir section 3.1). La seconde contribution de cette thèse est d'adapter le modèle de manière non supervisée aux expressions de base réelles du sujet. Nous avons d'abord proposé de réaliser cette adaptation de manière séquentielle (voir section 3.2) et nous avons ensuite proposé une extension de la méthode d'adaptation sur une fenêtre temporelle (voir section 3.3).

Dans ce chapitre, nous présentons les résultats obtenus pour ces deux contributions de la thèse. Dans un premier temps, nous présentons le protocole expérimental (section 4.1). Ensuite, nous reportons les résultats de la détection automatique du visage neutre, utile pour l'initialisation du modèle spécifique à la personne (section 4.2). Les résultats de l'adaptation sont présentés dans les deux sections suivantes : dans un premier temps avec l'adaptation séquentielle (section 4.3) et dans un second temps avec l'adaptation sur une fenêtre temporelle (section 4.4).

4.1 Protocole

Dans cette section nous présentons les bases de données utilisées (sous-section 4.1.1), les paramètres choisis pour l'étape de vérification des contraintes (sous-section 4.1.2), les métriques que nous proposons pour pouvoir évaluer la performance de l'adaptation (sous-section 4.1.3) et le protocole suivi pour évaluer la performance de l'adaptation (sous-section 4.1.4).

4.1.1 Bases de données

Pour les expérimentations, nous utilisons de nouveau notre base de données maison FAST et la base de données publique MUG [22] (voir partie « Bases de données » de la sous-section 2.2.4 pour la description) sur lesquelles nous avons testé notre système de reconnaissance d'expressions faciales (voir sous-section 2.2.4). Dans ce chapitre nous utilisons le sous-ensemble FAST-frontal de la base FAST et les sous-ensembles MUG-posé et MUG-spontané de la base MUG. Dans ces bases, les expressions sont acquises dans l'environnement du laboratoire avec peu, ou pas, de variation de pose de la tête et les sujets ne sont pas en train de parler pendant l'acquisition. Les expressions sont posées

dans les bases FAST-frontal et MUG-posé, alors que les expressions sont spontanées dans la base MUG-spontané.

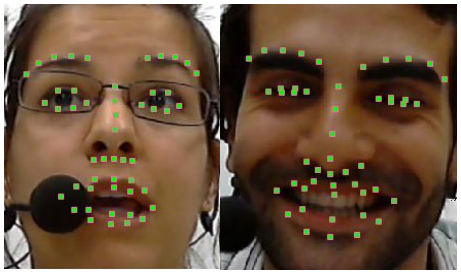
Nous utilisons également la base de données RECOLA [98]. Cette base de données contient 27 vidéos de 5 minutes contenant des expressions spontanées de 27 sujets (15 femmes, 12 hommes) telles qu'elles ont été fournies pour le « challenge » AVEC 2016 [53]. Deux sujets doivent interagir par ordinateur interposé pour résoudre une tâche dite de survie. Cette tâche consiste à arriver à un compromis entre les deux sujets quant aux objets à posséder en priorité dans un contexte de survie. Pour plus de détails, le lecteur peut se référer à [98]. Chaque sujet est enregistré pendant l'interaction, il est donc amené à parler. De plus, la variation de pose de la tête est naturelle et fréquente et des occlusions dues aux cheveux ou à une main sur le visage peuvent se produire. Cette base de données permet donc de se confronter à certains des défis de la vie réelle que nous avons identifiés dans la sous-section 1.2.1. Nous pouvons ainsi tester la robustesse de notre méthode d'adaptation non supervisée du modèle spécifique à la personne dans un environnement non contraint.

4.1.2 Paramètres de l'adaptation

Quatre paramètres sont à fixer pour l'étape de vérification des contraintes de l'adaptation non supervisée du modèle (voir sous-section 3.2.4) : ϵ (voir équation 3.5), d_{global} (voir équation 3.7), d_{joie} (voir équations 3.7 et 3.8) et d_{local} (voir équation 3.8). Nous fixons la valeur de ces paramètres empiriquement sur la base de données RECOLA. Dans nos expérimentations, nous utilisons les valeurs suivantes :

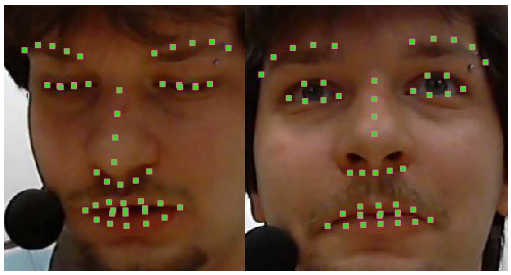
- $\epsilon = \|J_{n_p,2}\|_2$, où $J_{n_p,2}$ est une matrice unitaire de dimension $(n_p, 2)$ avec $n_p = 49$ le nombre de points caractéristiques du visage,
- $d_{global} = 0.5$,
- $d_{joy} = 0.2$,
- $d_{local} = 0.7$.

Dans la figure 4.1, nous donnons deux exemples d'expressions qui nous ont servies pour fixer le seuil d_{joy} sur le visage en entier. Les deux expressions sont détectées globalement comme de la joie ($l_{global} = joie$). Cependant seule l'expression de droite correspond effectivement à de la joie, celle de gauche étant une mauvaise détection de notre système de reconnaissance des expressions de base (voir sous-section 3.2.2). La distance de déformation faciale D_{joie} est de 0.847 (respectivement 0.190) pour l'expression de gauche (respectivement de droite). Le seuil d_{joy} doit être fixé entre ces deux valeurs, de telle sorte que l'expression de gauche ne vérifie pas la contrainte sur la distance de déformation faciale (voir équation 3.7) et que l'expression de droite la vérifie.



$D_{joie} = 0.847$ $D_{joie} = 0.190$

FIGURE 4.1 – Exemples d'expressions de la base RECOLA détectées globalement comme de la joie ($l_{global} = joie$) avec leur distance de déformation faciale D_{joie} respective.



$D_{tristesse} = 0.869$ $D_{tristesse} = 0.526$

FIGURE 4.2 – Exemples d'expressions de la base RECOLA détectées localement comme de la tristesse ($l_{bouche} = tristesse$) avec leur distance de déformation faciale $D_{tristesse}$ respective.

Dans la figure 4.2, nous donnons deux exemples d'expressions qui nous ont servies pour fixer le seuil d_{local} sur la bouche. Les deux expressions sont détectées localement comme de la tristesse ($l_{bouche} = tristesse$). Cependant seule la bouche de l'expression de droite correspond effectivement à de la tristesse, celle de gauche étant une mauvaise détection de notre système de reconnaissance des expressions de base (voir sous-section 3.2.2). La distance de déformation faciale $D_{tristesse}$ est de 0.869 (respectivement 0.526) pour l'expression de gauche (respectivement de droite). Le seuil d_{local} doit être fixé entre ces deux valeurs, de telle sorte que l'expression de gauche ne vérifie pas la contrainte sur la distance de déformation faciale (voir équation 3.7) et que l'expression de droite la vérifie.

4.1.3 Métriques proposées

Dans cette thèse, nous nous basons sur la représentation invariante des expressions faciales de [14] pour définir le modèle expressif spécifique à la personne. La principale

contribution de cette thèse est l'adaptation non supervisée du modèle aux expressions de base réelles du sujet, dans le but de construire un modèle expressif spécifique à la personne de manière non supervisée. Puisqu'il n'existe pas de travaux similaires s'inscrivant dans le cadre bien particulier de cette représentation invariante des expressions faciales, il nous faut définir nos propres métriques pour évaluer la performance de l'adaptation. Une manière de procéder est de comparer les expressions de base du modèle après adaptation avec les expressions de base réelles du sujet que l'on souhaite détecter et que l'on considère comme la vérité terrain.

Étant donné que les bases de données FAST et MUG contiennent des expressions posées, nous pouvons utiliser ces expressions à l'« apex » comme vérité terrain, *i.e.* comme les expressions de base réelles du sujet vers lesquelles on doit s'approcher avec l'adaptation du modèle. Ainsi, nous pouvons définir des métriques qui quantifient l'erreur entre les expressions de base d'un modèle donné et les expressions de base de la vérité terrain. Cela nous permet de faire une analyse quantitative de la performance de l'adaptation. Nous présentons les métriques quantitatives, *i.e.* définies avec la vérité terrain, dans la première partie de cette sous-section.

En ce qui concerne la base de données RECOLA, seules des vidéos d'expressions spontanées sont disponibles. Nous ne pouvons donc pas définir de vérité terrain et il nous est alors impossible d'utiliser les métriques quantitatives. Nous proposons donc de faire une analyse qualitative de la performance de l'adaptation avec des métriques qui sont calculées à partir d'annotations manuelles. Nous présentons les métriques qualitatives dans la seconde partie de cette sous-section.

Métriques quantitatives

Nous définissons une métrique pour rendre compte de la proximité entre les expressions de base du modèle et les expressions de base de la vérité terrain. Elle s'applique soit sur le modèle adapté, soit sur le modèle plausible (qui correspond à l'initialisation). Nous définissons ensuite deux variables à partir de cette métrique : le taux d'amélioration et le facteur d'amélioration. Ces deux variables permettent de rendre compte de l'efficacité de l'adaptation par rapport aux modèles plausibles sur l'ensemble de la base de donnée testée.

La métrique est calculée comme la distance euclidienne entre les points caractéristiques d'une expression de base de la vérité terrain et de l'expression de base du modèle correspondante. Nous appelons cette métrique la « métrique des points caractéristiques ». La figure 4.3 l'illustre. Soit pl (respectivement ad) l'indice faisant référence au modèle plausible (respectivement au modèle après adaptation). Soit n_s le nombre de sujets dans

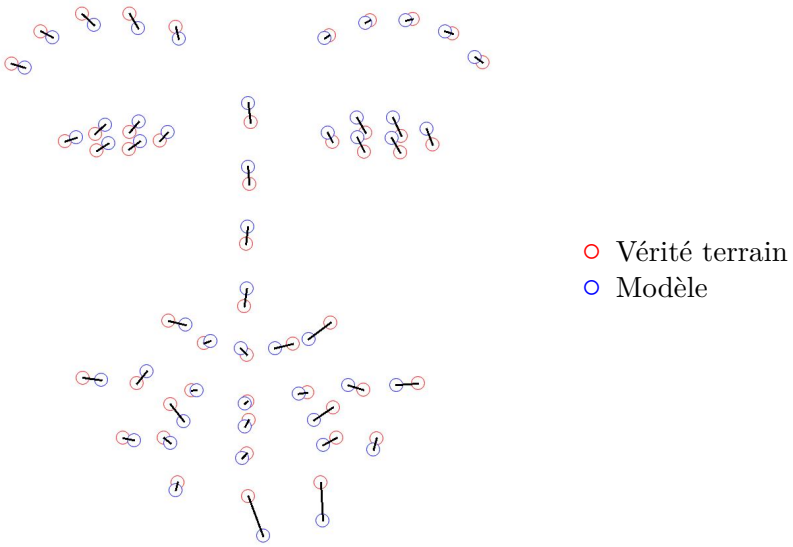


FIGURE 4.3 – Illustration de la métrique des points caractéristiques sur l'expression de la joie. L'expression de vérité terrain alignée avec les expressions de base du modèle est affichée avec les points rouges. L'expression de base du modèle alignée est affichée avec les points bleus. La métrique est calculée comme la distance euclidienne entre les points caractéristiques de l'expression de vérité terrain et de l'expression de base du modèle, c'est-à-dire la somme des distances affichées en noir.

la base de données. Soit $n_e \in \mathbb{R}$ le nombre d'expressions de base du modèle (neutre inclus), on note $N_e^* \subset [\![1, n_e]\!]$ le sous-ensemble des indices des expressions de base du modèle sur lesquelles la métrique est calculée (neutre systématiquement exclus). Pour tout $i \in N_e^*$, pour tout $j \in [\![1, n_s]\!]$ et pour tout $k \in \{pl, ad\}$, on note $X_k^{i,j,k}$ (respectivement $X_k^{i,j,vt}$) les points caractéristiques de la i-ième expression de base du modèle k (respectivement de la vérité terrain) après alignement avec les expressions de base du modèle k. La métrique des points caractéristiques se calcule comme :

$$\forall i \in N_e^*, \forall j \in [\![1, n_s]\!], \forall k \in \{pl, ad\}, m_{i,j,k} = \|X_k^{i,j,k} - X_k^{i,j,vt}\|_2. \tag{4.1}$$

A partir de cette métrique, nous définissons deux variables : le taux d'amélioration et le facteur d'amélioration. Le taux d'amélioration donne le pourcentage de sujets sur la base de données pour lesquels l'adaptation résulte en un modèle plus proche de la vérité terrain que ne l'est le modèle plausible, conformément à la métrique utilisée. Le facteur d'amélioration mesure à quel point les modèles adaptés sur la base de données sont proches de la vérité terrain, relativement aux modèles plausibles et conformément à la métrique utilisée.

Soit $j \in [\![1, n_s]\!]$. Le modèle adapté du sujet j est plus proche de la vérité terrain que ne l'est le modèle plausible, conformément à la métrique des points caractéristiques m, si la somme sur les expressions de base des valeurs des métriques du modèle adapté est plus petite que la somme sur les expressions de base des valeurs des métriques du modèle plausible, *i.e.* $\sum_{i \in N_e^*} m_{i,j,ad} < \sum_{i \in N_e^*} m_{i,j,pl}$. Le taux d'amélioration est calculé comme le pourcentage de sujets sur la base de données pour lesquels cette inégalité est vérifiée :

$$\alpha = \frac{card\left(\left\{ j \in [\![1, n_s]\!] \text{ tel que } \dfrac{\sum\limits_{i \in N_e^*} m_{i,j,ad}}{\sum\limits_{i \in N_e^*} m_{i,j,pl}} < 1 \right\}\right)}{n_s}. \tag{4.2}$$

La fraction $\sum_{i \in N_e^*} m_{i,j,ad} / \sum_{i \in N_e^*} m_{i,j,pl}$ donne le pourcentage d'amélioration des métriques du modèle adapté par rapport aux métriques du modèle plausible du sujet j. Nous définissons le facteur d'amélioration du sujet j par $\beta^j = 1 - \sum_{i \in N_e^*} m_{i,j,ad} / \sum_{i \in N_e^*} m_{i,j,pl}$. β^j est égal à 0 si la somme sur les expressions de base des valeurs des métriques du modèle adapté est égale à la somme sur les expressions de base des valeurs des métriques du modèle plausible, ce qui signifie qu'aucune amélioration n'est apportée par l'adaptation du modèle. Plus le modèle adapté est proche de la vérité terrain comparativement au modèle plausible, plus β^j est proche de 1. Au contraire, β^j devient négatif si le modèle plausible est plus proche de la vérité terrain que ne l'est le modèle adapté. Le facteur d'amélioration β est équivalent à la moyenne de l'ensemble des β^j pour $j \in [\![1, n_s]\!]$ et est calculé comme :

$$\beta = 1 - moyenne_{j \in [\![1, n_s]\!]} \left(\frac{\sum\limits_{i \in N_e^*} m_{i,j,ad}}{\sum\limits_{i \in N_e^*} m_{i,j,pl}} \right). \tag{4.3}$$

Métriques qualitatives

En l'absence de vérité terrain, il nous est impossible de calculer les métriques quantitatives présentées ci-dessus. Nous proposons alors de faire une analyse qualitative.

Tout au long de l'adaptation d'un sujet donné, nous comptons manuellement le nombre de mises à jour. Nous prenons en compte à la fois les mises à jour globales et locales.

A la fin de l'adaptation d'un sujet donné, nous comptons manuellement le nombre d'expressions de base du modèle qui ont été mises à jour (« nombre de mises à jour finales ») et le nombre d'expressions de base qui ont été mises à jour correctement

(« nombre de mises à jour finales correctes »). Nous prenons en compte les expressions de base qui ont été mises à jour globalement ou qui ont été mises à jour localement à la fois dans la zone des yeux et de la bouche (pour l'expression de base de la joie, étant donné que les sourcils sont neutres, une mise à jour locale de la bouche suffit pour être comptabilisée comme une mise à jour globale). Nous entendons par mise à jour correcte une mise à jour du modèle avec une expression qui correspond bien, en termes de déformations faciales, à la définition de l'expression de base du modèle attendue. Ces deux nombres sont nécessairement inférieurs ou égaux à $n_e - 1$ (le nombre d'expressions de base dans le modèle sans le neutre, dans notre cas $n_e - 1{=}5$). Nous calculons ensuite le « pourcentage de mises à jour finales correctes » comme la fraction $\frac{\text{nombre de mises à jour finales correctes}}{\text{nombre de mises à jour finales}}$.

Enfin, nous faisons la moyenne sur l'ensemble des sujets du nombre de mises à jour, du nombre de mises à jour finales et du pourcentage de mises à jour finales correctes. Nous avons donc 3 métriques qualitatives :

- Nombre moyen de mises à jour,
- Nombre moyen de mises à jour finales,
- Pourcentage moyen de mises à jour finales correctes.

Le « pourcentage moyen de mises à jour finales correctes » permet d'évaluer la capacité de l'adaptation à mettre à jour le modèle avec des expressions candidates correspondant aux déformations faciales des expressions de base du modèle (voir table 2.1). Le « nombre moyen de mises à jour finales » permet de savoir, au terme de l'adaptation, combien d'expressions de base du modèle sont mises à jour en moyenne. Le « nombre moyen de mises à jour » permet quant à lui de comparer le coût calculatoire de plusieurs variantes de l'adaptation.

4.1.4 Déroulement de l'adaptation

Notre méthode d'adaptation non supervisée du modèle spécifique à la personne est testée sur 4 bases de données (voir sous-section 4.1.1) : FAST-frontal, MUG-posé, MUG-spontané et RECOLA. La table 4.1 récapitule les données sur lesquelles nous testons l'adaptation et les métriques que nous utilisons pour évaluer la performance.

Pour les bases FAST-frontal et MUG-posé, l'adaptation est effectuée sur la concaténation de 6 vidéos d'expressions posées (appelée ensemble de vidéos) : neutre, colère, dégoût, joie, tristesse et surprise. Pour les bases MUG-spontané et RECOLA, l'adaptation est effectuée sur une seule vidéo d'expressions spontanées.

Pour la base FAST-frontal, l'adaptation est effectuée sur les sujets pour lesquels deux ensembles de vidéos sont disponibles (voir la partie « Base de données » de la sous-section 2.2.4) et qui ne sont pas inclus dans l'ensemble d'apprentissage du système

TABLE 4.1 – Récapitulatif des données sur lesquelles l'adaptation est testée et des métriques utilisées. La détection automatique du visage neutre pour l'initialisation du modèle et l'adaptation proprement dite sont effectuées sur le même ensemble de vidéos pour chaque sujet. Entre parenthèses est indiquée la nature des expressions dans les vidéos. Pour les métriques quantitatives, le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés soit sur les 5 expressions de base du modèle, soit sur 3 expressions de base du modèle (colère, joie, surprise).

Base de données	Nombre de sujets	Ensemble de vidéos pour l'adaptation	Nombre d'ensembles de vidéos par sujet	Métriques
FAST-frontal	26	6 vidéos (posées)	2	Quantitatives (5 expressions)
MUG-posé	49	6 vidéos (posées)	2	Quantitatives (3 expressions)
MUG-spontané	45	1 vidéo (spontanées)	1	Quantitatives (3 expressions, vérité terrain de MUG-posé)
RECOLA	27	1 vidéo (spontanées)	1	Qualitatives

de reconnaissance d'expressions faciales, *i.e.* les 8 sujets de la base FAST-pose. Cela laisse 26 sujets sur lesquels nous testons l'adaptation. Pour la base MUG-posé, nous extrayons pour chaque sujet deux ensembles de 6 vidéos correspondant aux 6 expressions prototypiques d'Ekman [3]. Pour 45 sujets, seule une vidéo du neutre est disponible, dans ce cas la même vidéo du neutre est utilisée pour les deux ensembles de 6 vidéos. La vidéo de la tristesse est absente pour 2 sujets et la vidéo du dégoût est absente pour 1 sujet.

La première étape est l'initialisation du modèle spécifique à la personne (voir section 3.1). Pour ce faire, nous détectons automatiquement le visage neutre des sujets avec notre méthode de reconnaissance d'expressions faciales (voir figure 3.1). Pour les bases FAST-frontal et MUG-posé, la détection est faite sur la concaténation des 6 vidéos. Nous détectons un visage neutre pour chacun des deux ensembles de 6 vidéos, il y a donc deux visages neutre détectés pour chaque sujet. Pour les bases MUG-spontané et RECOLA, la détection est effectuée sur la vidéo d'expressions spontanées.

La seconde étape est l'adaptation à proprement parler. Elle est effectuée sur les mêmes ensembles de vidéos que pour la détection automatique du visage neutre. Pour les bases FAST-frontal, MUG-posé et MUG-spontané, nous utilisons les métriques quantitatives pour évaluer la performance de l'adaptation, alors que nous utilisons les métriques qualitatives pour la base RECOLA (voir sous-section 4.1.3).

Pour définir la vérité terrain de FAST-frontal et MUG-posé, nous extrayons l'image à l'« apex » de l'expression de base pour chacune des 6 vidéos. Puisqu'il y a deux ensembles de 6 vidéos pour chaque sujet dans les bases FAST-frontal et MUG-posé, lorsque l'adaptation est effectuée sur le premier ensemble, nous utilisons la vérité terrain du second ensemble et inversement. Pour la base MUG-spontané, nous utilisons la vérité terrain de la base MUG-posé pour calculer les métriques quantitatives. Pour la base FAST-frontal, le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur les 5 expressions de base du modèle (colère, dégoût, joie, tristesse et surprise, voir table 2.1). Puisque dans la base MUG les expressions du dégoût et de la tristesse sont différentes de celles définies pour le modèle, le taux d'amélioration α et le facteur d'amélioration β sont calculés sur 3 expressions de base du modèle (colère, joie et surprise), bien que les 5 expressions soient présentes dans les ensembles de vidéos.

4.2 Inititalisation du modèle spécifique à la personne

L'initialisation du modèle se déroule en deux temps. D'abord, le visage neutre de chaque sujet est détecté automatiquement. Il est ensuite utilisé pour synthétiser les expressions de base plausibles avec lesquelles le modèle plausible est construit.

Nous utilisons notre système de reconnaissance d'expressions faciales pour la détection automatique du visage neutre (voir figure 3.1). Pour chaque sujet, nous annotons manuellement si le visage détecté correspond bien à un visage neutre ou non. Puis le taux de reconnaissance est calculé comme le pourcentage de sujets de la base de données pour lesquels le visage détecté est effectivement neutre. Ensuite, nous construisons le modèle plausible de chaque sujet avec le visage neutre détecté et les expressions de base plausibles, ce même si le visage détecté n'est pas neutre. Enfin, nous calculons le pourcentage de modèles plausibles correctement construits, *i.e.* dont la structure est celle attendue (voir équation 2.6).

Les taux de reconnaissance du visage neutre obtenus sur les différentes bases de données sont reportés dans la table 4.2, ainsi que le pourcentage de modèles plausibles dont la structure correspond à l'équation 2.6. Les mauvaises détections sont surtout dues à un manque de généralisation de notre système de reconnaissance d'expressions faciales. A l'exception de quelques rares sujets, ces mauvaises détections sont subtiles (par exemple, les sourcils sont légèrement froncés ou la bouche est légèrement étirée), de telle sorte que cela ne pose aucun problème majeur pour la construction du modèle ensuite. Preuve s'il en est, tous les modèles plausibles, sans exception, ont la structure attendue. Cela signifie que les visages incorrectement détectés comme neutre restent assez proches

TABLE 4.2 – Taux de reconnaissance de la détection du visage neutre avec notre méthode de reconnaissance d'expressions faciales (voir figure 3.1). La détection est effectuée sur la concaténation de 6 vidéos d'expressions posées (neutre, colère, dégoût, joie, tristesse et surprise) pour les bases FAST-frontal et MUG-posé et sur la vidéo d'expressions spontanées pour les bases MUG-spontané et RECOLA. Le pourcentage de modèles correctement construits correspond au pourcentage de modèles plausibles (construits à partir du visage neutre détecté) dont la structure correspond à l'équation 2.6.

Base de données	Nombre d'ensemble de vidéos	Taux de reconnaissance	Pourcentage de modèles correctement construits
FAST-frontal	52	86.54%	100%
MUG-posé	98	87.76%	100%
MUG-spontané	46	78.26%	100%
RECOLA	27	100.00%	100%

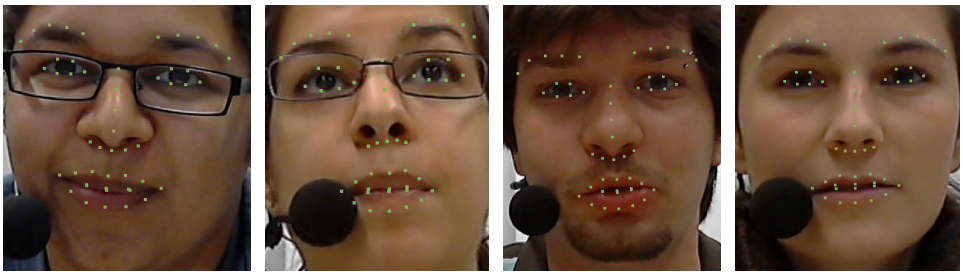


FIGURE 4.4 – Exemples de visages correctement détectés comme neutres dans la base RECOLA [98]

du visage neutre pour qu'il soient au centre des déformations dans le modèle construit sur les expressions plausibles.

La figure 4.4 donne 4 exemples de visages correctement détectés comme neutres dans la base RECOLA. La figure 4.5 donne 4 exemples de visages incorrectement détectés comme neutres dans la base MUG-posé. Le premier visage à gauche est un exemple de mauvaise détection subtile, ici au niveau de la bouche. Les trois autres mauvaises détections montrent la limite de notre système de reconnaissance d'expressions faciales basés sur les points caractéristiques pour la détection du neutre.

4.3 Résultats de l'adaptation séquentielle

Dans cette section, nous présentons les résultats obtenus pour l'adaptation séquentielle sur des expressions posées dans un environnement contraint (sous-section 4.3.1)

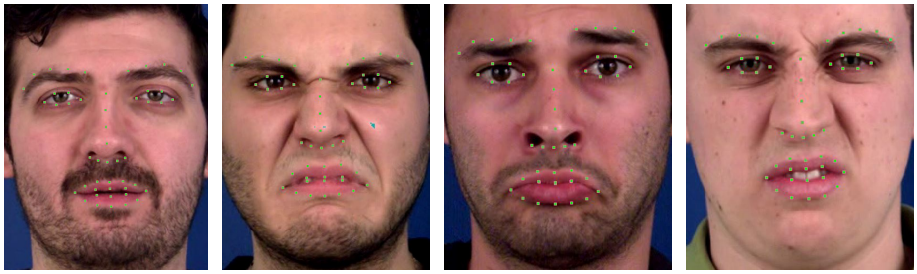


FIGURE 4.5 – Exemples de visages incorrectement détectés comme neutres dans la base MUG-posé [22]

puis sur des expressions spontanées dans un environnement contraint (sous-section 4.3.2) et enfin sur des expressions spontanées dans un environnement non contraint (sous-section 4.3.3). Nous terminons par une conclusion (sous-section 4.3.4).

Pour évaluer l'efficacité de l'étape de vérification des contraintes (voir sous-section 3.2.4), nous effectuons également une adaptation avec uniquement deux contraintes globales et aucune contrainte locale. Pour une expression courante avec le label global l_{global} et l'intensité I (voir équation 3.2), ces contraintes sont $l_{global} \neq neutre$ et $I > I^{l_{global}}$ si l'expression de base du modèle avec le label l_{global} a déjà été mise à jour. De plus, nous conservons la contrainte sur la structure après la mise à jour des expressions de base. Nous faisons référence à l'adaptation avec uniquement deux contraintes globales par le schéma d'adaptation « Sous-contraint » et à l'adaptation avec la vérification de contraintes au complet par le schéma d'adaptation « Contraint ».

4.3.1 Adaptation sur des expressions posées

Dans cette sous-section, nous présentons les résultats de l'adaptation séquentielle obtenus sur les bases FAST-frontal et MUG-posé, contenant des expressions posées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives.

Les résultats sont reportés dans la table 4.3 et attestent de l'efficacité de l'adaptation avec le schéma contraint sur les deux bases de données d'expressions posées : le taux d'amélioration α est supérieur à 94% et le facteur d'amélioration β est positif, ce qui veut dire que la grande majorité des modèles adaptés sont plus proches de la vérité terrain que ne le sont les modèles plausibles.

Sur les deux bases de données, l'adaptation avec le schéma contraint donne un meilleur taux d'amélioration α que l'adaptation avec le schéma sous-contraint. Il y a 19.23% (respectivement 7.14%) de sujets en plus dans la base FAST-frontal (respectivement MUG-posé) pour lesquels le modèle adapté est plus proche de la vérité terrain

TABLE 4.3 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur les bases FAST-frontal et MUG-posé. L'adaptation est effectuée sur la concaténation de 6 vidéos d'expressions posées (neutre, colère, dégoût, joie, tristesse et surprise). Le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur les 5 expressions de base du modèle pour la base FAST-frontal et sur 3 expressions de base du modèle (colère, joie et surprise) pour la base MUG-posé.

Base de données	Schéma d'adaptation	Taux d'amélioration α	Facteur d'amélioration β
FAST-frontal	Sous-contraint	75.00%	13.28%
	Contraint	**94.23%**	**17.42%**
MUG-posé	Sous-contraint	89.80%	**32.02%**
	Contraint	**96.94%**	23.87%

que ne l'est le modèle plausible lorsque nous vérifions la totalité des contraintes. Cela montre donc que l'étape de vérification des contraintes permet une bonne détection des expressions candidates de telle sorte que la majorité des modèles adaptés est plus proche de la vérité terrain que ne le sont les modèles plausibles.

Le facteur d'amélioration β augmente également lorsque nous passons du schéma sous-contraint au schéma contraint sur la base FAST-frontal, ce qui montre que les modèles adaptés sont globalement plus proche de la vérité terrain avec le schéma contraint qu'avec le schéma sous-contraint. En revanche, cela n'est pas le cas sur la base MUG-posé. Cela peut s'expliquer par le fait que les expressions candidates détectées pour la mise à jour du modèle dans le schéma sous-contraint sont rejetées par le schéma contraint, ces expressions candidates étant plus proches de la vérité terrain que les expressions candidates détectées dans le schéma contraint.

4.3.2 Adaptation sur des expressions spontanées dans un environnement contraint

Dans cette sous-section, nous présentons les résultats de l'adaptation séquentielle obtenus sur la base de données MUG-spontané, contenant des expressions spontanées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives.

Les résultats sont reportés dans la table 4.4. Le taux d'amélioration α et le facteur d'amélioration β diminuent radicalement comparés aux résultats obtenus sur la base MUG-posé (voir table 4.3). Ceci peut s'expliquer par deux facteurs. Premièrement, les 3 expressions de base utilisées pour calculer les métriques ne sont pas toujours présentes dans les vidéos spontanées (en particulier la colère et la surprise), ce qui impacte négativement le taux d'amélioration α car ces expressions de base ne sont pas toujours mises

TABLE 4.4 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base MUG-spontané. L'adaptation est effectuée sur les vidéos d'expressions spontanées. Le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur 3 expressions de base du modèle (colère, joie et surprise).

Schéma d'adaptation	Taux d'amélioration α	Facteur d'amélioration β
Sous-contraint	**54.35%**	-0.90%
Contraint	**54.35%**	**1.56%**

à jour. Deuxièmement, le sujet peut potentiellement afficher des expressions de base légèrement différentes de celles de la vérité terrain puisque les premières sont spontanées et les secondes sont posées, ce qui impacte négativement le facteur d'amélioration β. Cela montre les limites que nous pouvons rencontrer en définissant les métriques avec des expressions de base de vérité terrain. En effet, les expressions de la vérité terrain ne sont pas nécessairement affichées par le sujet dans une base d'expressions spontanées.

Il est à noter que lorsqu'on passe du schéma sous-contraint au schéma contraint, le taux d'amélioration α reste le même, ce qui veut dire qu'il y autant de modèles adaptés plus proches de la vérité terrain que ne le sont les modèles plausibles lorsqu'on passe d'un schéma à l'autre. Le facteur d'amélioration β, quant à lui, devient positif lorsqu'on passe du schéma sous-contraint au schéma contraint. Cela montre que le schéma contraint permet d'avoir en moyenne des modèles adaptés plus proches de la vérité terrain que ne le sont les modèles plausibles, alors que ce n'est pas le cas avec le schéma sous-contraint ($\beta < 0$).

Dans la table 4.5 nous avons reporté les pourcentages de sujets pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation. Les résultats montrent que l'étape de vérification des contraintes est surtout utile pour rejeter des mauvaises détections de la joie et de la tristesse. En effet, le pourcentage de mise à jour chute de 21.74% (respectivement 15.22%) pour la joie (respectivement la tristesse) lorsqu'on passe du schéma sous-contraint au schéma contraint. Ces expressions qui sont rejetées avec le schéma contraint et pas le schéma sous-contraint sont pour la plupart de mauvaises détections, ce qui est cohérent avec le fait que le facteur d'amélioration β devient positif en passant du schéma sous-contraint au schéma contraint (voir table 4.4). Les résultats nous montrent également que l'expression de la joie est mise à jour pour la plupart des sujets. Pour les autres expressions de base, le pourcentage chute en-dessous de 50%. Cela est cohérent avec le contenu expressif des vidéos de la base MUG-spontané.

TABLE 4.5 – Pourcentages de sujets dans la base MUG-spontané pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation.

Schéma d'adaption	Colère	Dégoût	Joie	Tristesse	Surprise
Sous-contraint	19.57%	30.43%	95.65%	50.00%	43.48%
Contraint	13.04%	34.78%	73.91%	34.78%	45.65%

TABLE 4.6 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base RECOLA. Les expressions sont spontanées. Nous comptons manuellement le nombre de mises à jour et nous calculons le pourcentage de mises à jour correctes durant l'adaptation. A la fin de l'adaptation, nous comptons manuellement le nombre d'expressions de base qui ont été mises à jour (« mises à jour finales ») et nous calculons le pourcentage d'expressions de base qui ont été mises à jour correctement (« mises à jour finales correctes »). Nous calculons la moyenne de ces comptages et pourcentages sur les sujets de la base de données.

Schéma d'adaptation	Nombre moyen de mises à jour	Nombre moyen de mises à jour finales	Pourcentage moyen de mises à jour finales correctes
Sous-contraint	13.59	3.33	23.89%
Contraint	11.81	2.67	**59.23%**

4.3.3 Adaptation sur des expressions spontanées dans un environnement non contraint

Dans cette sous-section, nous présentons les résultats de l'adaptation séquentielle obtenus sur la base RECOLA, contenant des expressions spontanées dans un environnement non contraint. La vérité terrain n'étant pas disponible, nous utilisons les métriques qualitatives.

Les résultats sont reportés dans la table 4.6. Comme nous pouvions nous y attendre, le nombre moyen de mises à jour décroît lorsqu'on passe du schéma sous-contraint au schéma contraint. Dans une moindre mesure, cela est aussi le cas pour le nombre moyen de mises à jour finales. Nous notons au contraire que le pourcentage moyen de mises à jour finales correctes augmente de 35.34% lorsqu'on passe du schéma sous-contraint au schéma contraint. Alors que le nombre moyen de mises à jour diminue, le pourcentage moyen moyen de mises à jour finales correctes augmente, le schéma contraint permet donc de rejeter les mauvaises expressions candidates détectées avec le schéma sous-contraint. Cela montre l'efficacité de l'étape de vérification des contraintes lorsque nous testons l'adaptation dans un environnement non contraint.

TABLE 4.7 – Pourcentages de sujets dans la base RECOLA pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation.

Schéma d'adaptation	Colère	Dégoût	Joie	Tristesse	Surprise
Sous-contraint	44.44%	40.74%	96.30%	88.89%	59.26%
Contraint	29.63%	25.93%	74.07%	59.26%	77.78%

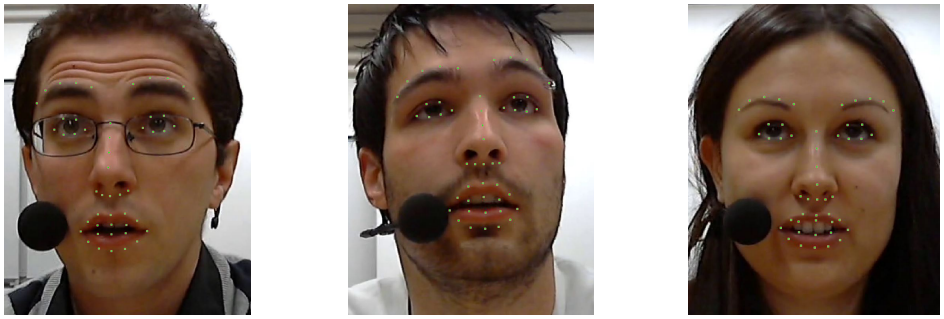
Dans la table 4.7, nous avons reporté pour chaque expression de base du modèle le pourcentage de sujets pour lesquels l'expression de base a été mise à jour à l'issue de l'adaptation. Les résultats montrent que la plupart des mises à jour concernent les expressions de la surprise et de la joie, suivies par la tristesse. Cela est cohérent avec le contenu expressif des vidéos spontanées de la base RECOLA. Bien que peu de mises à jour finales concernent la colère et le dégoût, il est à noter que la plupart des mises à jour incorrectes sont dues à une mauvaise détection des expressions candidates de la colère, du dégoût et dans une moindre mesure de la tristesse, ce qui impacte négativement le pourcentage moyen de mises à jour finales correctes.

Cette mauvaise détection des expressions candidates s'explique surtout par la présence d'expressions faciales résultant de la parole et s'approchant des expressions de base de la colère, du dégoût, de la tristesse ou de la surprise. Ces expressions candidates correspondent rarement à l'émotion ressentie et cela montre que notre méthode d'adaptation, en particulier l'étape de vérification des contraintes, n'est pas encore tout à fait robuste à la parole. Le problème est que notre détecteur d'expressions faciales classifie toutes les expressions dans l'une des 6 classes (neutre, colère, dégoût, joie, tristesse, surprise) mais ne peut pas déterminer si une expression n'appartient à aucune de ces classes. C'est donc à l'étape de vérification des contraintes de déterminer si l'expression courante correspond bien à la classe détectée.

Ceci étant dit, certaines des expressions candidates de la tristesse et la majorité des expressions candidates de la surprise présentent des déformations faciales correspondant bel et bien à leur expression de base respective, même si ce n'est pas l'émotion ressentie. La figure 4.6 (respectivement la figure 4.7) illustre quelques exemples d'expressions de surprise (respectivement de tristesse) qui ont été détectées comme expressions candidates pour la mise à jour du modèle.

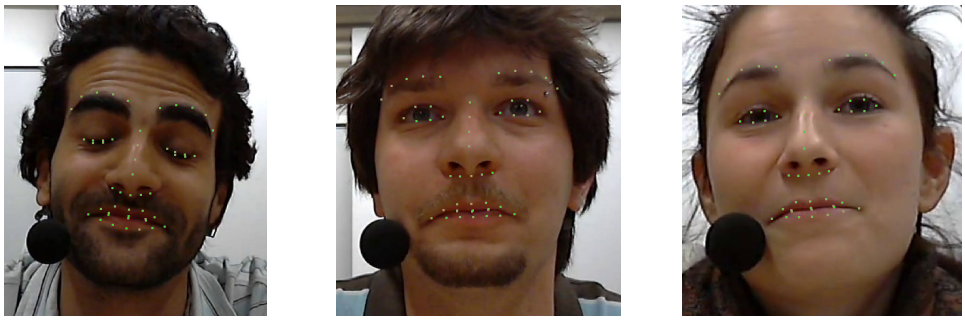
4.3.4 Conclusion sur l'adaptation séquentielle

Dans cette section, nous avons présenté les résultats de notre méthode d'adaptation non supervisée du modèle spécifique à la personne de manière séquentielle. Pour



(a) Mise à jour globale (b) Mise à jour globale (c) Mise à jour locale

FIGURE 4.6 – Exemple d'expressions de surprise détectées comme expressions candidates pour la mise à jour du modèle dans la base RECOLA [98]. La mise à jour locale concerne la zone de la bouche.



(a) Mise à jour globale (b) Mise à jour locale (c) Mise à jour locale

FIGURE 4.7 – Exemple d'expressions de tristesse détectées comme expressions candidates pour la mise à jour du modèle dans la base RECOLA [98]. La mise à jour locale concerne la zone de la bouche.

montrer l'efficacité et l'importance de l'étape de vérification des contraintes dans l'adaptation (voir sous-section 3.2.4), nous avons considéré deux schémas d'adaptation : sous-contraint (avec uniquement deux contraintes globales et aucune contrainte locale) et contraint (avec la totalité des contraintes).

Nous avons d'abord conduit les expérimentations sur les bases FAST-frontal et MUG-posé, contenant des expressions posées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. Les résultats montrent que l'adaptation résulte en un modèle plus proche de la vérité terrain que ne l'est le modèle plausible pour la majorité des sujets. En effet, le taux d'amélioration α obtenu avec le schéma contraint est de l'ordre de 95% sur les deux bases de données. De plus, le taux d'amélioration α est supérieur à celui obtenu avec le schéma sous-contraint, ce qui montre l'efficacité

de l'étape de vérification des contraintes.. Les facteurs d'amélioration β obtenus sur les deux bases de données sont positifs, ce qui montre que sur l'ensemble des deux bases de données l'adaptation permet d'être plus proche des expressions de base de la vérité terrain, comparativement aux modèles plausibles.

Nous avons ensuite conduit les expérimentations sur la base MUG-spontané, contenant des expressions spontanées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. Le taux d'amélioration α et le facteur d'amélioration β diminuent grandement par rapport à ceux obtenus sur les bases d'expressions posées. Le taux d'amélioration α est de l'ordre de 55% avec le schéma d'adaptation contraint. Cela peut s'expliquer par le fait que le sujet ne va pas nécessairement effectuer les expressions de base spontanément de la même manière que la vérité terrain (qui sont des expressions posées en l'occurrence). Cela montre les limites de la définition de nos métriques à l'aide de la vérité terrain. Le facteur d'amélioration β devient positif lorsqu'on passe du schéma sous-contraint au schéma contraint, ce qui montre l'efficacité de l'étape de vérification des contraintes.

Enfin, nous avons conduit les expérimentations sur la base RECOLA, contenant des expressions spontanées dans un environnement non contraint. La vérité terrain n'étant pas disponible, nous avons utilisé les métriques qualitatives. Le pourcentage moyen de mises à jour finales correctes passe de 24% à 59% en passant du schéma d'adaptation sous-contraint au schéma d'adaptation contraint. Cela montre donc l'importance de l'étape de vérification des contraintes dans un environnement non contraint. Les mises à jour incorrectes résultent de détections erronées d'expressions candidates (en particulier pour les expressions de base de la colère, du dégoût et de la tristesse) qui sont dues au fait que notre méthode d'adaptation n'est pas tout à fait robuste à la parole.

4.4 Résultats de l'adaptation sur une fenêtre temporelle

Dans cette section, nous présentons les résultats obtenus pour l'adaptation sur une fenêtre temporelle sur des expressions posées dans un environnement contraint (sous-section 4.4.1), puis sur des expressions spontanées dans un environnement contraint (sous-section 4.4.2) et enfin sur des expressions spontanées dans un environnement non contrain (sous-section 4.4.3). Nous terminons par une conclusion (sous-section 4.4.4).

Pour rappel, avec l'adaptation séquentielle (voir section 3.2) le modèle est mis à jour image par image, dans l'ordre d'apparence des expressions de base affichées par le sujet. Le défaut de cette méthode est le nombre potentiellement important de mises à jour. En effet, une expression de base peut être mise à jour autant de fois que nécessaire

TABLE 4.8 – Définition des 3 configurations de sélection des expressions candidates sur la fenêtre temporelle. L_{global} est la liste ordonnée des labels globaux des expressions de base à mettre à jour, $L_{sourcils}$ est la liste ordonnée des labels locaux des expressions de base à mettre à jour dans la zone des sourcils, L_{bouche} est la liste ordonnée des labels locaux des expressions de base à mettre à jour dans la zone de la bouche et l'ordre des zones indique quelle liste entre $L_{sourcils}$ et L_{bouche} sera parcourue en première pour la sélection des expressions candidates locales (voir sous-section 3.3.1).

Configuration	L_{global}	$L_{sourcils}$	L_{bouche}	Ordre des zones
1	{*colère, dégoût, joie, tristesse, surprise*}	{*froncés, levés*}	{*colère, dégoût, joie, tristesse, surprise*}	{*bouche, sourcils*}
2	{*colère, dégoût, joie, tristesse, surprise*}	{*froncés, levés*}	{*colère, dégoût, joie, tristesse, surprise*}	{*sourcils, bouche*}
3	{*surprise, tristesse, joie, dégoût, colère*}	{*levés, froncés*}	{*surprise, tristesse, joie, dégoût, colère*}	{*bouche, sourcils*}

entre l'« onset » de l'expression (formation de l'expression sur le visage) et son « apex » (expression à son intensité maximale) dans la vidéo. Avec l'adaptation sur une fenêtre temporelle, nous faisons une sélection de l'expression candidate sur une fenêtre temporelle avant de faire la mise à jour, ce qui permet de mettre à jour le modèle directement avec l'expression candidate à son « apex ».

Les expérimentations menées dans la section précédente ont montré l'importance de l'étape de vérification des contraintes pour mener à bien l'adaptation, en particulier dans un environnement non contraint. Dans cette section, nous ne présentons donc que les résultats obtenus avec le schéma d'adaptation contraint et nous cherchons à savoir si la méthode d'adaptation sur une fenêtre temporelle permet d'améliorer les résultats par rapport à la méthode d'adaptation séquentielle. De plus, nous voulons montrer que l'ordre de sélection des expressions candidates sur la fenêtre temporelle (L_{global}, $L_{sourcils}$ et L_{bouche}, voir sous-section 3.3.1) a peu d'influence sur les résultats obtenus. Nous définissons donc 3 configurations pour la sélection des expressions candidates sur la fenêtre temporelle, qui sont décrites dans la table 4.8. Sur chaque base de données, chacune de ces 3 configurations est testée avec 2 ou 3 valeurs de longueur de fenêtre temporelle T (voir sous-section 3.3.1).

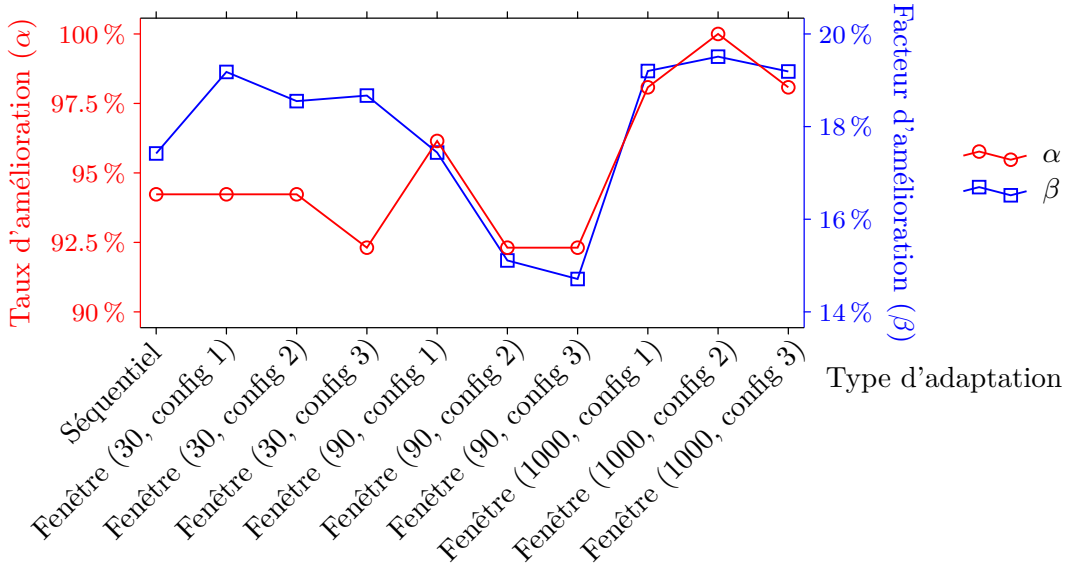


FIGURE 4.8 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base FAST-frontal. L'adaptation est effectuée sur la concaténation de 6 vidéos d'expressions posées (neutre, colère, dégoût, joie, tristesse et surprise). Le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur les 5 expressions de base du modèle. Le type d'adaptation « Séquentiel » correspond à l'adaptation séquentielle avec le schéma d'adaptation contraint. Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

4.4.1 Adaptation sur des expressions posées

Dans cette sous-section, nous présentons les résultats de l'adaptation sur une fenêtre temporelle obtenus sur les bases FAST-frontal et MUG-posé, contenant des expressions posées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. Les résultats sont reportés dans les figures 4.8 et 4.9. Le type d'adaptation fait référence à la méthode d'adaptation utilisée : séquentielle ou sur une fenêtre temporelle. Pour l'adaptation sur une fenêtre temporelle, les 3 configurations de la table 4.8 sont considérées, chacune avec 3 longueurs de fenêtre T différentes. Les courbes rouges avec des points circulaires (respectivement bleues avec des points carrés) correspondent aux taux d'amélioration α (respectivement facteurs d'amélioration β) obtenus pour les différents types d'adaptation.

La figure 4.8 contient les résultats sur la base FAST-frontal. Les 3 longueurs de fenêtre T testées sont 30, 90 et 1000. Cela correspond respectivement à environ 1 seconde, 3.5

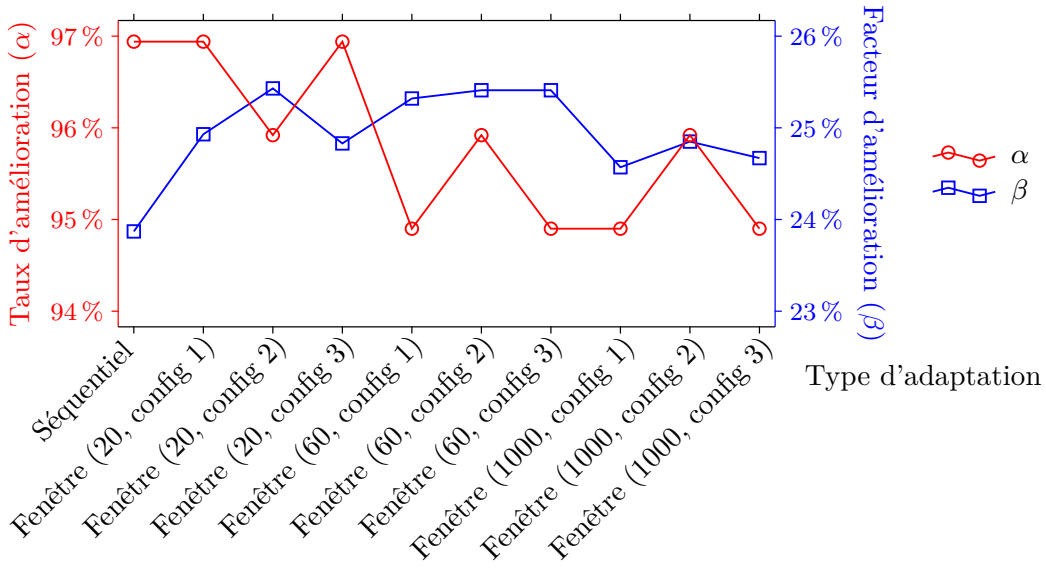


FIGURE 4.9 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base MUG-posé. L'adaptation est effectuée sur la concaténation de 6 vidéos d'expressions posées (neutre, colère, dégoût, joie, tristesse et surprise). Le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur 3 expressions de base du modèle (colère, joie et surprise). Le type d'adaptation « Séquentiel » correspond à l'adaptation séquentielle avec le schéma d'adaptation contraint. Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

secondes et 40 secondes. La longueur $T = 90$ correspond environ à la longueur d'une vidéo d'une expression posée. La longueur $T = 1000$ est choisie de telle sorte que la fenêtre englobe l'intégralité de l'ensemble de 6 vidéos. Les résultats sur α et β ne montrent pas d'amélioration significative en passant de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle ($T = 30, 90, 1000$) : pour α, au mieux +5.77% (on atteint alors α=100%) et au pire -1.92% ; pour β, au mieux +2.09% et au pire -2.71%.

La figure 4.9 contient les résultats sur la base MUG-posé. Puisque le taux d'image est différent par rapport à la base FAST-frontal, nous testons 3 longueurs de fenêtre T différentes : 20, 60 et 1000. Cela correspond respectivement à environ 1 seconde, 3 secondes et 40 secondes. La longueur $T = 60$ correspond environ à la longueur d'une vidéo d'une expression posée. La longueur $T = 1000$ est choisie de telle sorte que la fenêtre englobe l'intégralité de l'ensemble de 6 vidéos. Les résultats sur α et β ne montrent

pas d'amélioration significative en passant de l'adaptation séquentielle à l'adaptation temporelle sur une fenêtre temporelle ($T = 20, 60, 1000$) : pour α, au mieux +0% et au pire -2.04% ; pour β, au mieux +1.56% et au pire +0.70%.

L'adaptation sur une fenêtre temporelle ne permet donc pas d'augmenter significativement le taux d'amélioration α sur des expressions posées, qui était déjà assez élevé avec l'adaptation séquentielle. L'intérêt de l'adaptation sur une fenêtre temporelle se fait plutôt sentir sur le facteur d'amélioration β. Mis à part sur la base FAST-frontal pour l'adaptation avec une fenêtre temporelle de taille $T = 90$, lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle, le facteur d'amélioration β ne fait qu'augmenter et ce quelles que soient la longueur de la fenêtre T et la configuration considérée. Enfin, les résultats montrent que la performance de l'adaptation sur une fenêtre temporelle varie peu selon le choix de la taille de la fenêtre temporelle T et de la configuration de sélection des expressions candidates (voir table 4.8).

4.4.2 Adaptation sur des expressions spontanées dans un environnement contraint

Dans cette sous-section, nous présentons les résultats de l'adaptation sur une fenêtre temporelle obtenus sur la base MUG-spontané, contenant des expressions spontanées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. Les résultats sont reportés dans la figure 4.10. Le type d'adaptation fait référence à la méthode d'adaptation utilisée : séquentielle ou sur une fenêtre temporelle. Pour l'adaptation sur une fenêtre temporelle, les 3 configurations de la table 4.8 sont considérées, chacune avec 3 longueurs de fenêtre T différentes. Les courbes rouges avec des points circulaires (respectivement bleues avec des points carrés) correspondent aux taux d'amélioration α (respectivement facteurs d'amélioration β) obtenus pour les différents types d'adaptation.

Dans la sous-section précédente, l'ensemble des vidéos d'expressions posées sur lesquelles nous faisions l'adaptation durait environ 40 secondes. Ici, une vidéo d'expression spontanées dure environ 75 secondes, nous testons donc de nouvelles longueurs de fenêtre T. Les 3 longueurs de fenêtre T testées sont 100, 200 et 1500. Cela correspond respectivement à environ 5 secondes, 10 secondes et 75 secondes. La longueur $T = 1500$ est choisie de telle sorte que la fenêtre englobe l'intégralité de la vidéo.

Contrairement aux résultats obtenus sur des expressions posées (voir sous-section 4.4.1), la longueur de la fenêtre temporelle T a une influence sur la performance de l'adaptation. Pour les longueurs $T = 100$ et $T = 200$, les résultats restent à peu près constants. En revanche, pour la longueur $T = 1500$, le taux d'amélioration α augmente

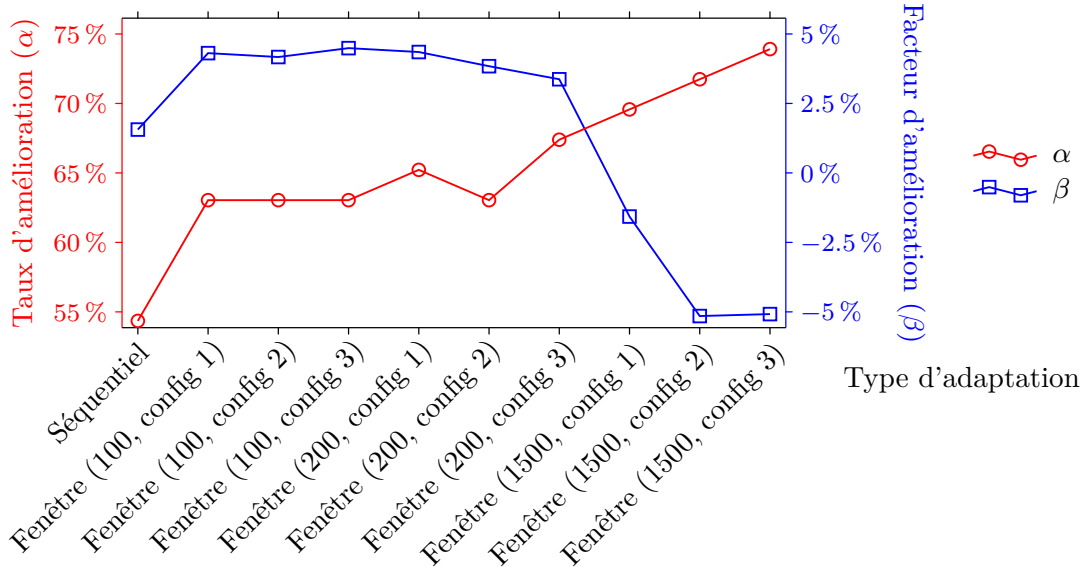


FIGURE 4.10 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base MUG-spontané. L'adaptation est effectuée sur la vidéo d'expressions spontanées. Le taux d'amélioration α (voir équation 4.2) et le facteur d'amélioration β (voir équation 4.3) sont calculés sur 3 expressions de base du modèle (colère, joie et surprise). Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

alors que le facteur d'amélioration β diminue jusqu'à être inférieur à celui obtenu avec l'adaptation séquentielle. Ces mauvais résultats pour la longueur $T = 1500$ montrent que notre méthode n'est pas capable de sélectionner les meilleures expressions candidates avec une longueur de fenêtre temporelle T trop importante.

En ce qui concerne les longueurs $T = 100$ et $T = 200$, l'adaptation sur une fenêtre temporelle permet d'augmenter à la fois le taux d'amélioration α et le facteur d'amélioration β par rapport à l'adaptation séquentielle : pour α, jusqu'à +13.04% ; pour β, jusqu'à +2.93%.

Dans la figure 4.11 nous avons reporté les pourcentages de sujets pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation, et ce pour chaque type d'adaptation. Les résultats nous montrent que la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8) n'ont pas d'influence notable sur les expressions de base qui sont mises à jour à l'issue de l'adaptation. Les pourcentages restent dans le même ordre de grandeur lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle.

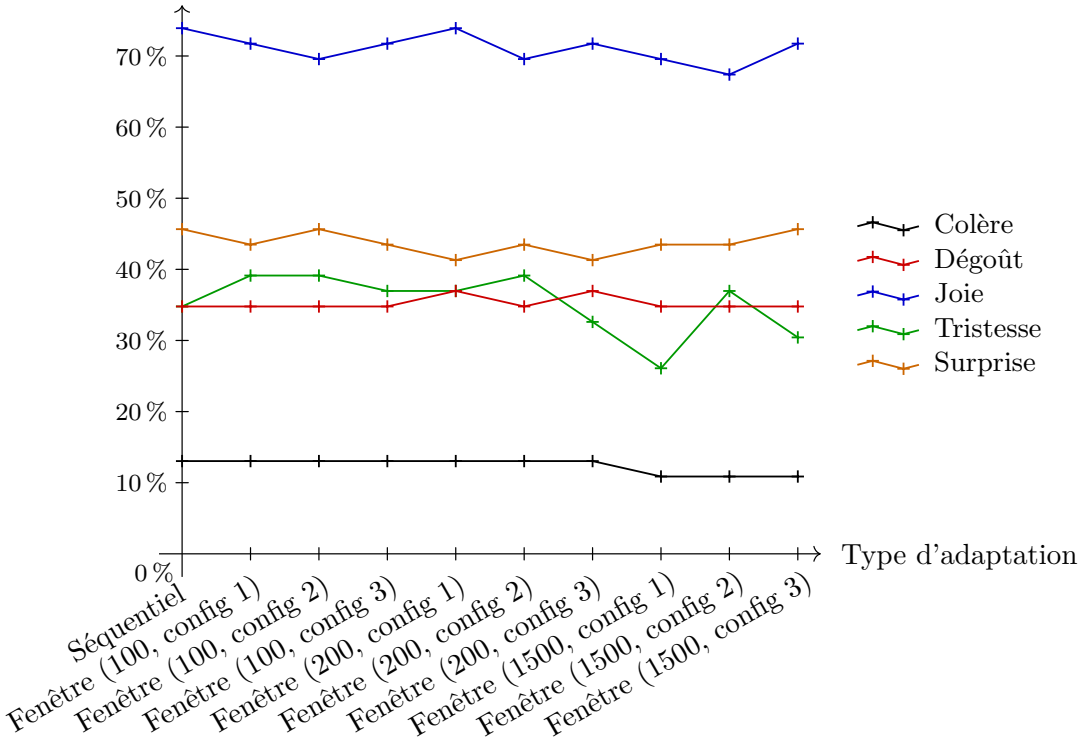


FIGURE 4.11 – Pourcentages de sujets dans la base MUG-spontané pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation. Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

4.4.3 Adaptation sur des expressions spontanées dans un environnement non contraint

Dans cette sous-section, nous présentons les résultats de l'adaptation sur une fenêtre temporelle obtenus sur la base RECOLA, contenant des expressions spontanées dans un environnement non contraint. La vérité terrain n'étant pas disponible, nous avons utilisé les métriques qualitatives. Les résultats sont reportés dans la figure 4.12. Le type d'adaptation fait référence à la méthode d'adaptation utilisée : séquentielle ou sur une fenêtre temporelle. Pour l'adaptation sur une fenêtre temporelle, les 3 configurations de la table 4.8 sont considérées. La longueur d'une vidéo est maintenant de 5 minutes, nous testons donc 2 nouvelles valeurs de longueur de fenêtre T : 750 et 1500. Cela correspond respectivement à 30 secondes et 60 secondes. La courbe verte pointillée avec des points circulaires correspond au nombre moyen de mises à jour, la courbe verte pleine avec des

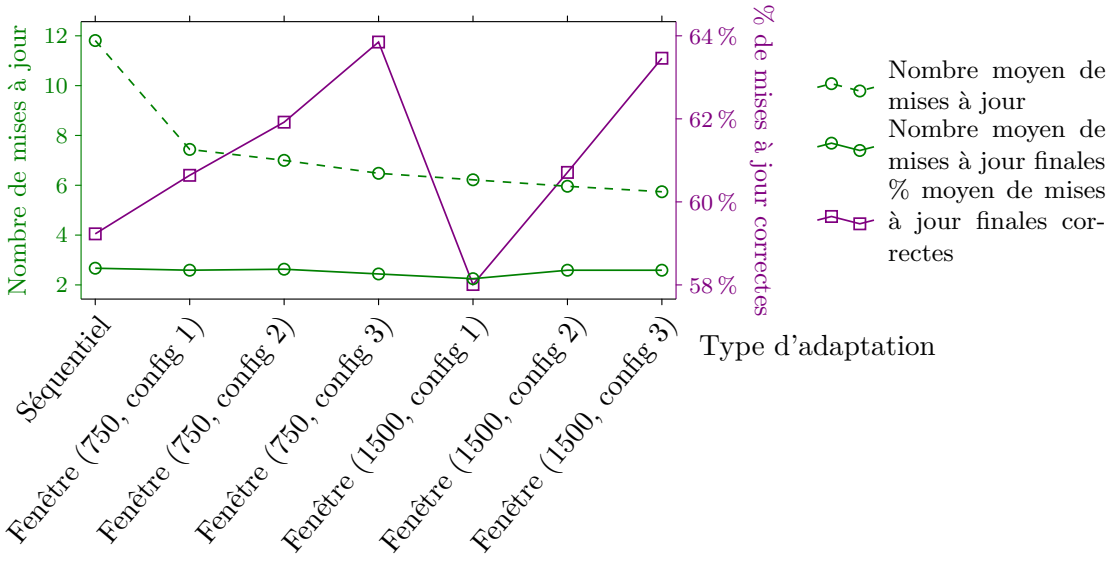


FIGURE 4.12 – Résultats de l'adaptation non supervisée du modèle spécifique à la personne sur la base RECOLA. L'adaptation est effectuée sur la vidéo d'expressions spontanées. Nous comptons manuellement le nombre de mises à jour (« nombre moyen de mises à jour »). A la fin de l'adaptation, nous comptons manuellement le nombre d'expressions de base qui ont été mises à jour (« nombre moyen de mises à jour finales ») et nous calculons le pourcentage d'expressions de base qui ont été mises à jour correctement (« % moyen de mises à jour finales correctes »). Nous calculons la moyenne de ces comptages et pourcentages sur les sujets de la base de données. Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

points circulaires au nombre moyen de mises à jour finales et la courbe violette pleine au pourcentage moyen de mises à jour finales correctes.

Comme nous pouvions nous y attendre, le nombre moyen de mises à jour (courbe verte pointillée) diminue de manière significative (environ de moitié) lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle. Le nombre moyen de mises à jour finales (courbe verte pleine), quant à lui, reste quasi constant lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle. Cela montre que quel que soit le type d'adaptation, nous avons toujours le même nombre d'expressions de base qui sont mises à jour à l'issue de l'adaptation.

Le pourcentage moyen de mises à jour finales correctes (courbe violette pleine) est égal à 59.23% pour l'adaptation séquentielle et est compris entre 58.01% et 63.46% pour l'adaptation sur une fenêtre temporelle. A part pour la fenêtre de longueur $T = 1500$ avec

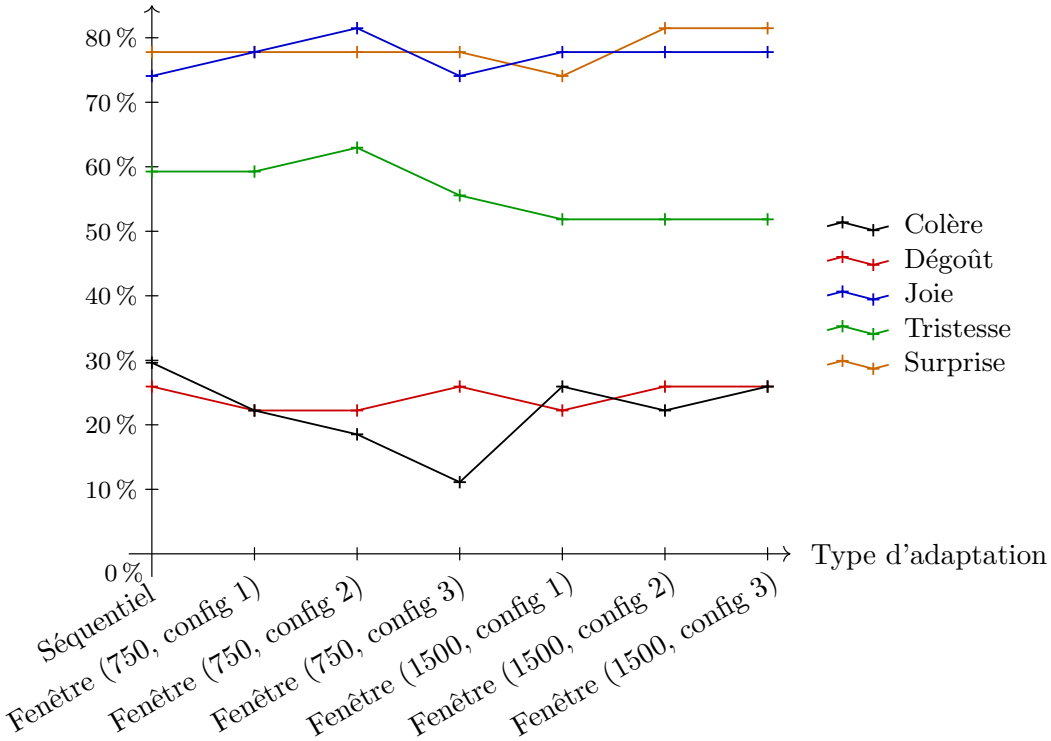


FIGURE 4.13 – Pourcentages de sujets dans la base RECOLA pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation. Le type d'adaptation « Fenêtre » correspond à l'adaptation sur une fenêtre temporelle avec entre parenthèses la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8).

la configuration 1, l'adaptation sur une fenêtre temporelle donne toujours un pourcentage moyen de mises à jour finales correctes qui est supérieur à celui obtenu avec l'adaptation séquentielle. Cela montre l'efficacité de l'adaptation sur une fenêtre temporelle dans un environnement non contraint.

Dans la figure 4.13 nous avons reporté les pourcentages de sujets pour lesquels les différentes expressions de base du modèle ont été mises à jour après l'adaptation, et ce pour chaque type d'adaptation. De même que sur la base MUG-spontané (voir figure 4.11), les résultats nous montrent que la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates (voir table 4.8) n'ont pas d'influence notable sur les expressions de base qui sont mises à jour à l'issue de l'adaptation. Les pourcentages restent dans le même ordre de grandeur lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle.

4.4.4 Conclusion sur l'adaptation sur une fenêtre temporelle

Dans cette section, nous avons présenté les résultats de notre méthode d'adaptation non supervisée du modèle spécifique à la personne sur une fenêtre temporelle. Plusieurs longueurs de fenêtre T ainsi que 3 configurations de sélection des expressions candidates (voir table 4.8) ont été considérées. Nous avons comparé la performance de l'adaptation sur une fenêtre temporelle et de l'adaptation séquentielle avec un schéma contraint.

Nous avons d'abord conduit les expérimentations sur les bases FAST-frontal et MUG-posé, contenant des expressions posées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. Les résultats montrent que l'adaptation sur une fenêtre temporelle ne permet pas d'augmenter significativement le taux d'amélioration α sur des expressions posées, qui était déjà important avec l'adaptation séquentielle. En revanche, nous observons une augmentation du facteur d'amélioration β lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle. Cela montre que même s'il n'y a pas d'augmentation du nombre de sujets pour lesquels le modèle adapté est plus proche de la vérité terrain que ne l'est le modèle plausible (*i.e.* α n'augmente pas), les modèles adaptés sont en moyenne plus proches de la vérité terrain qu'avec l'adaptation séquentielle (*i.e.* β augmente). En outre, les résultats montrent que la performance de l'adaptation sur une fenêtre temporelle dépend peu du choix de la longueur de fenêtre T et de la configuration de sélection des expressions candidates et reste toujours dans le même ordre de grandeur.

Nous avons ensuite conduit les expérimentations sur la base MUG-spontané, contenant des expressions spontanées dans un environnement contraint. La performance est évaluée avec les métriques quantitatives. L'adaptation sur une fenêtre temporelle présente de meilleurs résultats que l'adaptation séquentielle lorsque la longueur de la fenêtre T est équivalente à 5 ou 10 secondes. De plus, les résultats restent dans le même ordre de grandeur quelle que soit la configuration de sélection des expressions candidates. En revanche, les résultats se dégradent par rapport à l'adaptation séquentielle lorsque la longueur de la fenêtre T permet d'englober la vidéo en entier (75 secondes), ce qui montre que sur cette base de données, notre méthode d'adaptation sur une fenêtre temporelle n'est pas capable de sélectionner les meilleures expressions candidates avec une longueur de fenêtre temporelle T trop importante.

Enfin, nous avons conduit les expérimentations sur la base RECOLA, contenant des expressions spontanées dans un environnement non contraint. La vérité terrain n'étant pas disponible, nous avons utilisé les métriques qualitatives. Les résultats montrent que le nombre moyen de mises à jour obtenu avec l'adaptation sur une fenêtre temporelle peut valoir jusqu'à la moitié du nombre moyen de mises à jour obtenu avec l'adaptation

séquentielle. Concernant le nombre moyen de mises à jour finales, il n'évolue pas lorsque nous passons de l'adaptation séquentielle à l'adaptation sur une fenêtre temporelle et reste quasi constant lorsque nous faisons varier la longueur de la fenêtre temporelle T et la configuration de sélection des expressions candidates. Donc quel que soit le type d'adaptation utilisée, nous avons au final toujours le même nombre d'expressions de base mises à jour. A propos du pourcentage moyen de mises à jour finales correctes, l'adaptation sur une fenêtre temporelle amène toujours une amélioration par rapport à l'adaptation séquentielle, à une configuration près. Cela montre que notre méthode d'adaptation sur une fenêtre temporelle est efficace dans un environnement non contraint et que celle-ci est plus performante que l'adaptation séquentielle.

Si l'adaptation sur une fenêtre temporelle permet de réduire le nombre de mises à jour de l'adaptation par rapport à l'adaptation séquentielle, il y a une opération supplémentaire de tri par intensité décroissante sur la fenêtre temporelle, ce qui va impacter le temps de calcul de l'adaptation. Dans la table 4.9, nous reportons le temps de calcul moyen sur chaque base de données pour l'adaptation séquentielle et l'adaptation sur une fenêtre temporelle. Pour les bases MUG-posé et MUG-spontané, le temps de calcul moyen est plus court pour l'adaptation sur une fenêtre temporelle, mais le gain par rapport à l'adaptation séquentielle n'est pas significatif. Nous notons que le temps de calcul est plus court pour l'adaptation séquentielle sur les bases FAST-frontal et RECOLA. L'écart entre l'adaptation séquentielle et l'adaptation sur une fenêtre temporelle est particulièrement important sur la base RECOLA. Cela montre que même si le nombre de mises à jour diminue avec l'adaptation sur une fenêtre temporelle, l'opération de tri est trop coûteuse. Celle-ci est effectuée avec la fonction « argsort » de la bibliothèque « numpy » pour le langage de programmation Python.

TABLE 4.9 – Temps de calcul de l'adaptation séquentielle et de l'adaptation sur une fenêtre temporelle. Le type d'adaptation T correspond à l'adaptation sur une fenêtre temporelle, suivi de la longueur de la fenêtre en nombre d'images. La configuration n° 1 de sélection des expressions candidates est utilisée (voir table 4.8). Le temps de calcul est moyenné sur la base de données. La durée de vidéo correspond à la longueur de la vidéo (ou de l'ensemble de vidéos) sur laquelle est effectuée l'adaptation.

Base de données	Durée de vidéo (secondes)	Type d'adaptation	Temps de calcul moyen (secondes)
FAST-frontal	≈ 24	Séquentiel	**7.054**
		T-30	7.923
		T-90	8.046
		T-1000	7.787
MUG-posé	≈ 24	Séquentiel	3.073
		T-20	2.917
		T-60	**2.796**
		T-1000	2.849
MUG-spontané	≈ 60	Séquentiel	9.205
		T-100	9.122
		T-200	9.201
		T-1500	**8.934**
RECOLA	≈ 300	Séquentiel	**174.245**
		T-750	300.287
		T-1500	297.421

Conclusion

Sommaire

L'objectif de cette thèse était de proposer une méthode pour construire un modèle expressif continu et spécifique à la personne de manière non supervisée, *i.e.* sans connaissance *a priori* de la morphologie du sujet. Dans ce chapitre, nous rappelons dans un premier temps les contributions, nous revenons ensuite sur les résultats obtenus et nous terminons par les perspectives.

Contributions

Le cadre applicatif visé par les travaux de cette thèse est le milieu médical et plus particulièrement le maintien à domicile de personnes âgées. L'idée à terme serait de développer un système capable de lever une alarme lorsqu'un comportement anormal est détecté. Dans ce cadre applicatif, notre système doit répondre aux besoins suivants :

- Analyse d'un seul sujet,
- Aucune étape de calibration nécessaire,
- Analyse d'expressions non prototypiques,
- Analyse dans un environnement non contraint en termes de pose de la tête et de parole.

Le système que nous proposons est basé sur la représentation invariante des expressions faciales de [14] qui permet de construire un modèle expressif, continu et spécifique à la personne à partir de caractéristiques faciales géométriques. Le choix d'un modèle spécifique à la personne est justifié par le fait qu'un seul sujet est analysé par notre système et cela permet alors une analyse plus fine. Le choix d'un modèle continu est justifié par l'analyse d'expressions non prototypiques. Les limites du modèle de [14] sont

la nécessité d'acquérir le visage neutre du sujet et le fait que les expressions de base soient synthétisées, ce qui rend le modèle non adapté aux expressions réelles du sujet et pas assez spécifique.

Notre principale contribution réside dans l'apprentissage d'une variété spécifique à la personne de manière non supervisée, *i.e.* sans connaissance *a priori* de la morphologie du sujet. L'apprentissage non supervisé permet de répondre au besoin d'absence d'étape de calibration. Pour ce faire, nous avons proposé deux contributions.

Pour notre première contribution, nous avons proposé de construire le modèle spécifique à la personne de [14] de manière non supervisée à l'aide d'une détection automatique du visage neutre (voir section 3.1). Le visage neutre détecté automatiquement permet ensuite de synthétiser les expressions de base modèle, puis de construire le modèle. Le modèle ainsi initialisé est appelé « modèle plausible ».

A ce stade, le modèle est construit sur des expressions de base synthétisées, il ne rend donc pas compte des déformations faciales réelles du sujet. Pour notre seconde contribution, nous avons proposé notre méthode d'adaptation non supervisée du modèle spécifique à la personne. L'idée de l'adaptation est de détecter, à la fois globalement et localement, les expressions de base réelles du sujet afin de remplacer les expressions de base synthétisées du modèle puis de les affiner, tout en maintenant un ensemble de contraintes. Dans un premier temps, nous avons proposé une adaptation séquentielle (voir section 3.2). Nous avons ensuite proposé une extension de la méthode d'adaptation sur une fenêtre temporelle (voir section 3.3).

Pour réaliser cette adaptation, nous avons besoin d'un système de reconnaissance d'expressions faciales. Dans un environnement non contraint, les performances d'un tel système chutent s'il n'est pas assez robuste, ce qui impacte négativement les performances de l'adaptation. Pour notre troisième contribution, nous avons donc proposé d'assurer la robustesse de la méthode d'adaptation en utilisant un système de reconnaissance d'expressions faciales robuste à la pose que nous avons développé (voir section 2.2).

Résultats

Nous avons réalisé nos expérimentations sur 3 bases de données : FAST-frontal (base maison), MUG [22] et RECOLA [98]. La base FAST-frontal contient des expressions posées. La base MUG est séparée en deux bases : MUG-posé, contenant des expressions posées, et MUG-spontané, contenant des expressions spontanées dans un environnement non contraint. La base RECOLA contient des expressions spontanées dans un environnement non contraint.

En ce qui concerne l'initialisation du modèle (notre première contribution), nous avons effectué la détection automatique du visage neutre sur les sujets de chacune des bases de données, puis nous avons construit le modèle à partir du visage neutre détecté et des expressions synthétisées. Nous avons relevé quelques mauvaises détections, en particulier sur la base MUG-spontané, mais cela n'empêche pas le modèle d'être systématiquement construit correctement.

Pour pouvoir mesurer la performance de notre méthode d'adaptation, nous avons défini nos propres métriques (voir sous-section 4.1.3). Dans le cas des bases FAST-frontal et MUG, nous avons défini la vérité terrain à atteindre avec les expressions posées à leur « apex ». Nous entendons par vérité terrain à atteindre les expressions de base réelles du sujet que notre méthode d'adaptation devrait détecter pour mettre à jour le modèle. Nous avons alors défini 2 métriques quantiatives qui correspondent à l'erreur entre une expression de base du modèle et l'expression de base correspondante de la vérité terrain. A partir de chacune de ces métriques, nous avons défini le taux d'amélioration et le facteur d'amélioration pour pouvoir comparer les métriques des modèles plausibles et des modèles après adaptation. La base RECOLA ne contenant que des expressions spontanées, la vérité terrain n'est pas disponible. Nous avons donc défini des métriques qualitatives qui sont calculées grâce à une annotation manuelle.

En ce qui concerne notre méthode d'adaptation séquentielle, les résultats sur les bases FAST-frontal et MUG-posé ont montré l'efficacité de l'adaptation sur des expressions posées. En revanche, les résultats sur la base MUG-spontané ont chuté. Cela montre les limites de définir nos métriques quantitatives à l'aide d'une vérité terrain. En effet, il n'est absolument pas garanti que les expressions de base que le sujet affiche spontanément soient proches des expressions de base de la vérité terrain. Les résultats sur la base RECOLA ont montré l'importance de l'étape de vérification des contraintes lorsque nous testons notre méthode d'adaptation dans un environnement non contraint. Les résultats sur cette base ont aussi mis en évidence les limitations de notre méthode. A l'issue de l'adaptation, en moyenne un peu moins de 50% des mises à jour du modèle sont incorrectes. Cela est dû au fait que notre système de reconnaissance d'expressions de base classifie toutes les expressions dans l'une des 6 classes (neutre, colère, dégoût, joie, tristesse, surprise) même si aucune ne convient. Il arrive donc que l'étape de vérification des contraintes reçoive une expression qui ne correspond pas l'expression de base attendue mais qui s'en approche et que les contraintes soient vérifiées malgré tout. C'est ce cas de figure qui explique les mauvaises mises à jour. Il est à noter tout de même qu'à l'issue de l'adaptation, parmi les expressions de base qui ont été mises à jour, 60% des expressions de base ont été mises à jour correctement.

Nous avons ensuite mené les expérimentations sur les mêmes bases de données avec notre méthode d'adaptation sur une fenêtre temporelle. Nous avons testé plusieurs configurations possibles de mise à jour sur la fenêtre temporelle et plusieurs longueurs de fenêtre. Les résultats sur les bases FAST-frontal et MUG-posé ont montré que, sur des expressions posées, l'adaptation sur une fenêtre temporelle n'améliore pas significativement les performances par rapport à l'adaptation séquentielle. Il est à noter également que les résultats restent relativement constants quelle que soit la configuration ou la longueur de fenêtre considérée. Les résultats sur les bases MUG-spontané et RECOLA ont montré que la fenêtre temporelle améliore les performances de l'adaptation par rapport à l'adaptation séquentielle sur des expressions spontanées. En outre, sur la base RECOLA, l'utilisation de la fenêtre temporelle permet de réduire en moyenne de moitié le nombre de mises à jour.

Perspectives

A l'issue des travaux de cette thèse, nous avons relevé quelques perspectives à explorer dans des travaux futurs.

Nous avons montré dans la sous-section 4.3.3 que la limite de notre méthode d'adaptation réside dans la reconnaissance des expression de base (voir figure 3.2), notamment à cause de la parole. Il est donc nécessaire d'améliorer notre système de reconnaissance d'expressions faciales robuste à la pose (voir section 2.2) et d'essayer de le rendre robuste à la parole, par exemple en faisant une analyse bimodale incorporant l'audio. Cela permettrait de savoir si l'utilisation d'un système de reconnaissance d'expressions plus performant permet d'améliorer la performance de l'adaptation. De façon similaire, nous pourrions aussi tester notre méthode d'adaptation en utilisant une autre méthode de rectification de pose (voir Annexe C).

Un axe de recherche qui est très peu étudié, voire pas du tout, est l'analyse des expressions faciales (ou des émotions) sur un axe temporel long. En effet, les recherches se cantonnent jusqu'ici à des analyses sur des vidéos de courte durée puisque c'est ce qui est proposé par les bases de données. Rares sont les bases de données proposant plusieurs acquisitions d'un même sujet espacées dans le temps. La création d'une telle base de données contenant des acquisitions espacées dans le temps sur une durée de l'ordre d'une ou plusieurs années permettrait de lancer des recherches approfondies sur les variations de comportement en adoptant une modélisation spécifique à la personne.

Dans les travaux de cette thèse, nous n'avons pas mis l'accent sur le fait que le modèle spécifique à la personne représenté sous la forme d'espace des signatures (voir

figure 2.1) peut servir comme outil de représentation des émotions. En effet, l'espace des signatures à 3 dimensions permet une représentation visuelle et synthétique des déformations faciales et la localisation des expressions de base permet d'en faire une interprétation sémantique. Il pourrait être intéressant de poursuivre selon cet axe de recherche en incorporant d'autres types de caractéristiques faciales ou d'autres modalités pour tenter de donner une représentation visuelle de l'état émotionnel.

En ce qui concerne le maintien à domicile de personnes âgées, nous pouvons envisager de quelle manière notre système peut être utilisé. L'idée serait d'installer des caméras à des endroits clés de l'habitation (par exemple au niveau du miroir de la salle de bain). Dans un premier temps, notre système détecte automatiquement le visage neutre du sujet et initialise le modèle avec ce visage neutre et les expressions de base plausibles. Ensuite, notre modèle s'adapte aux expressions réelles du sujet grâce à notre méthode d'adaptation non supervisée. Puis nous pouvons analyser les trajectoires des expressions faciales du sujet dans son espace des signatures et en déduire un comportement nominal. Lorsque le sujet effectue des expressions qui sortent de la trajectoire du comportement nominal dans l'espace des signatures, alors un comportement anormal est détecté et une alarme est levée.

« Challenge » AVEC 2016

Au cours de cette thèse, nous avons eu l'occasion de confronter le mode de représentation de l'espace des signatures au mode de représentation arousal/valence proposé par les psychologues (voir partie « Représentation dimensionnelle » de la sous-section 1.1.1) en participant au « challenge » AVEC 2016 (« Audio-Visual Emotion Challenge and Workshop »). Le but du « challenge » était d'estimer de manière continue la valeur des dimensions émotionnelles arousal et valence sur la base RECOLA à l'aide de plusieurs modalités (vidéo, audio et signaux physiologiques). Notre participation au « challenge » était motivée par la volonté de confronter le mode de représentation de l'espace des signatures à un gros volume de données réelles.

Les organisateurs du « challenge » ont repris les travaux des 5 éditions précédentes pour proposer un système de base faisant une estimation multimodale des dimensions arousal et valence ; les résultats obtenus forment la « baseline ». Les participants au « challenge » étaient donc invités à proposer un système d'estimation multimodale des dimensions arousal et valence dont les performances dépassent la « baseline ». Les résultats de celle-ci étaient particulièrement élevés cette année et seules 3 équipes ont réussi à la battre. La « baseline » est arrivée 4^{e} puis nous sommes arrivés 5^{e} sur 13 équipes.

Les caractéristiques faciales géométriques proposées pour la « baseline » étaient les points caractéristiques du visage, des caractéristiques bas-niveau donc. Nous avons proposé d'ajouter des caractéristiques faciales géométriques haut-niveau pour l'estimation multimodale : la pose de la tête et la signature de l'expression. L'avantage des caractéristiques haut-niveau est leur faible dimension comparée à celle des caractéristiques bas-niveau (rapport de 26).

Les résultats de l'estimation unimodale ont montré que les caractéristiques haut-niveau de la signature des expressions permettent d'obtenir des résultats du même ordre de grandeur qu'avec les caractéristiques géométriques bas-niveau. Cela nous laisse penser qu'il existe une cohérence entre notre représentation des expressions faciales à l'aide de l'espace des signatures et la représentation des émotions à l'aide des dimensions arousal et valence, plus particulièrement pour la dimension valence.

Publications

Publication dans une revue internationale avec comité de lecture

Raphaël Weber, Vincent Barrielle, Catherine Soladié, Renaud Séguier. Unsupervised adaptation of a person-specific manifold of facial expressions. Soumis le 6 juin 2017 à *Affective Computing, IEEE Transactions on*. Révision Mineure.

Publications dans des conférences internationales avec comité de lecture et « proceedings »

Raphaël Weber, Vincent Barrielle, Catherine Soladié, Renaud Séguier. High-level geometry-based features of video modality for emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 51-58. ACM, octobre 2016.

Raphaël Weber, Catherine Soladié, Renaud Séguier. A survey on databases for facial expression analysis. Accepté à *VISAPP 2018*.

Communication nationale sans actes

Raphaël Weber. Caractéristiques géométriques haut-niveau de la modalité vidéo pour la prédiction d'émotion. *GdR ISIS, Paris, 15 novembre 2016*.

Annexe A

Recensement des bases de données d'expressions faciales

Dans cette annexe, nous regroupons toutes les bases de données que nous avons recensées pour l'état de l'art de la section 1.2.2. La table A.1 contient les bases de données d'expressions posées, la table A.2 contient les bases de données d'expression spontanées et la table A.3 contient les bases de données d'expressions « in-the-wild ». Dans ces tables, chaque colonne correspond à l'une des 6 catégories que nous avons utilisées pour notre état de l'art : population, modalités, matériel d'acquisition, conditions expérimentales, protocole expérimental et annotations. Ces tables sont réalisées à partir de la lecture des articles décrivant les bases de données, si une information n'est pas précisée dans l'article, nous l'indiquons alors par un point d'interrogation.

Pour la population, 3 informations sont données : le nombre de sujets (avec le pourcentage femmes/hommes), les âges extrêmes et les groupes ethniques.

Pour le matériel d'acquisition, 3 informations sont données : le nombre de caméras, la résolution des images en pixels et le taux d'images par seconde (« FPS »). Si la base de données ne contient que des images mais aucune vidéo, le taux d'images par seconde ne s'applique pas et nous l'indiquons par un tiret. Lorsqu'il y a plusieurs caméras, nous indiquons entre parenthèses leur positionnement, les angles étant des angles de yaw (voir figure 1.8).

Pour les conditions expérimentales, 3 informations sont données : le fond de l'image, l'illumination, les occlusions et la variation de pose de la tête. Quatre types de fond sont rencontrés : uni, laboratoire, extérieur et varié. Le terme « laboratoire » est utilisé lorsque les acquisitions sont effectuées dans le laboratoire de recherche, ainsi le fond reste à peu près constant sur toutes les acquisitions. Le terme « extérieur » est utilisé si les

acquisitions sont effectuées en extérieur. Le terme « varié » est utilisé si les acquisitions sont effectuées dans plusieurs lieux différents, que ce soit en intérieur ou en extérieur. Pour les occlusions, nous indiquons entre parenthèses leur nature si elles ont été clairement identifiées dans la description de la base. Pour la pose de la tête, nous indiquons entre parenthèses le nombre de poses disponibles ; dans la quasi-totalité des cas, cela correspond au nombre de caméras.

Pour le protocole expérimental, les informations données changent d'une table à l'autre. Dans la table A.1, contenant les bases de données d'expressions posées, nous donnons le protocole de reproduction des expressions (pouvant être « libre », « consigne » ou « portrait ») et le nombre d'expressions différentes disponibles avec entre parenthèses les expressions en question. Les 6 émotions de base font alors référence aux 6 expressions protoypiques d'Ekman [3] : colère, dégoût, joie, peur, tristesse, surprise. Dans la table A.2, contenant les bases de données d'expressions spontanées, nous donnons le type d'induction émotionnelle utilisé (pouvant être « tâche passive », « tâche active », « interaction humain-humain » et « interaction humain-machine », où la tâche passive correspond au visionnage de vidéos ou d'images censées induire une émotion spécifique) et le nombre d'expressions différentes disponibles avec entre parenthèses les émotions correspondantes. Dans la table A.3, contenant les bases de données d'expressions « in-the-wild », nous donnons la nature des données (pouvant être des extraits télévisés, des extraits de films ou du « crowdsourcing ») et le type d'induction émotionnelle lorsque les expressions sont spontanées. Si les expressions sont posées, nous mettons un tiret pour le type d'induction.

Enfin, nous indiquons la présence ou non d'annotations dans les bases de données.

TABLE A.1 – Bases de données d'expressions posées - 1/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
University Of Maryland [85]	40 (?/ ?) / ? / Varié	Vidéo 2D du visage	1 / 560x420 / 30	Uni / ? / Non / Naturelle	Libre / 6 (6 émotions de base)	- / - / - / -
JAFFE [56]	10 (100/0) / ? / Japonais	Image 2D du visage	1 / ? / -	Uni / Uniforme / Non / Non	Libre / 7 (neutre, + 6 émotions de base)	- / - / Oui / -
KDEF [57]	70 (50/50) / 20-30 / ?	Image 2D du visage	5 (-90° :45° :90°) / 562x762 / -	Uni / Uniforme / Non / Oui (5)	Consigne / 7 (neutre + 6 émotions de base)	- / - / - / -
CK [58]	182 (69/31) / 18-50 / Caucasien, Africain sub-saharien, autres	Vidéo 2D du visage	1 / 640x490 (N&B) ou 640x480 (couleur) / ?	Uni / 1/3 ambiante et 2/3 uniforme / Non / Non	Consigne / 23 (3 unités d'action seules et combinaisons d'unités d'action dont les 6 émotions de base)	- / Oui / Oui / -

Suite de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 2/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
PICS - Pain Expressions [102]	23 (57/43) / ? / ?	Image 2D du visage	1 / 720x576 / -	Uni / Uniforme / Non / Oui (3)	? / 17 (neutre + 6 émotions de base + 10 expressions de douleur)	- / - / - / -
MMI [106]	19 (56/44) / 19-62 / Caucasien, Asiatique ou Sud-américain	Vidéo et image 2D du visage	2 (face et profil) / 720x576 / 24	1/4 laboratoire et 3/4 uni / 1/4 ambiante et 3/4 uniforme / Oui (lunettes) / Oui (2)	Consigne / 79 (unité d'action seule ou combinaisons d'unités d'action dont les 6 émotions de base)	- / Oui / - / -
FABO [100]	23 (52/48) / 18-24 / Caucasien, Moyen-Oriental, Hispanique, Asiatique	Vidéo 2D du visage et du mouvement du corps	2 (visage de face et haut du corps) / 1024x768 / 15	Uni / Uniforme / ? / Naturelle	Libre / 12 (neutre + 6 émotions de base + 5 émotions secondaires)	- / - / Oui / -

Suite de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 3/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
GEMEP [87]	10 (?/ ?) / ? / Francophone	Vidéo 2D du visage et du mouvement du corps, audio	3 (1 visage de face, 1 corps de face et 1 corps de profil) / ? / ?	Uni / ? / ? / Naturelle	Portrait / 18 (6 émotions de base + 12 émotions secondaires)	- / - / - / Oui
BU-3D FE [66]	100 (60/40) / ? / Caucasien, Africain sub-saharien, Asiatique, Hispanique et autres	Image 3D du visage	6 / 1300x900 (brut, 20000 à 35000 vertexes) et 512x512 (rogné, 13000 à 21000 vertexes) / -	Uni / Uniforme / Non / Oui (3D)	Libre / 7 (neutre + 6 émotions de base)	Oui / - / - / -
BU-4D FE [67]	101 (57/43) / 18-45 / Caucasien, Asiatique, Africain sub-saharien, Hispanique	Vidéo 3D du visage	3 (2 stéréo et 1 pour la texture) / 1040x1329 (35000 vertexes) / 25	Uni / Uniforme / Non / Oui (3D)	Consigne / 6 (6 émotions de base)	Oui / - / - / -
IEMOCAP [88]	10 (50/50) / ? / ?	Capture de mouvement et audio	8 / ? / 120	Laboratoire / Ambiante / Non / Naturelle	Portrait / 5 (neutre, colère, joie, tristesse, frustration)	Oui / - / Oui / Oui

Suite de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 4/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
Bosphorus [68]	105 (43/57) / 25-35 / Caucasian, others	Image 3D du visage	1 (3D) / 1600x1200 / -	Uni / Uniforme / Oui (lunettes, cheveux, main) / Oui (13)	Consigne / 34 (unité d'action seule ou combinaisons d'unités d'action dont les 6 émotions de base)	Oui / - / - / -
OULU-CASIA [104]	80 (26/74) / 23-58 / Finlandais, Chinois	Vidéo 2D du visage	1 / 320x420 / 25	Laboratoire / Normale, faible, sombre / ? / Non	Consigne / 6 (6 émotions de base)	- / - / - / -
Multi-PIE [71]	337 (30/70) / ? / Caucasien, Asiatique, Africain-subsaharien, autres	Image 2D du visage	15 (-90° :15° :90° et 2 en hauteur) / ? / -	Uni / 19 configurations / Oui (lunettes) / Oui (15)	Consigne / 6 (neutre, sourire, surprise, dégoût, yeux plissés, cri)	- / - / - / -
Radboud Faces [80]	67 (?/ ?) / enfants/adultes / Hollandais, Marocains	Image 2D du visage	5 (-90° :45° :90°) / ? / -	Uni / Uniforme / Non / Oui (5)	Consigne / 8 (neutre + 6 émotions de base + mépris)	- / - / Oui / Oui

Suite de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 5/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
MUG [22]	86 (41/59) / 20-35 / ?	Vidéo 2D du visage	1 / 896x896 / 19	Uni / Uniforme / Oui (cheveux) / Non	Consigne / 7 (neutre + 6 émotions de base)	Oui / - / - / -
NVIE [78]	215 (27/73) / 17-31 / Asiatique	Vidéo 2D, image 2D et vidéo infrarouge du visage	2 (visible et infrarouge) / 704x480 (visible) et 320x240 (infrarouge) / 30 (visible) et 25 (infrarouge)	Uni / 3 configurations / Oui (lunettes) / Non	Consigne / 6 (6 émotions de base)	Oui / - / Oui / Oui
B(3D)AC2 [90]	14 (57/43) / 21-53 / Anglophone	Vidéo 3D du visage et audio	1 (3D) / ? / 25	Uni / Uniforme / Non / Oui (3D)	Portrait / 80	Oui / - / - / -
ICT-3DRFE [79]	23 (26/74) / 22-35 / Variés	Image 3D du visage	« Light stage » / 1296x1944 (1200000 vertexes) / -	Uni / Configurable / Non / Oui (3D)	Libre / 15 (neutre + 6 émotions de base + expressions non émotionnelles)	- / Oui / - / -

Suite de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 6/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
D3DFACS [82]	10 (60/40) / 23-41 / Caucasien	Vidéo 3D du visage	6 / 1024x1280 (30000 vertexes) / 60	Uni / Uniforme / Non / Oui (3D)	Consigne / entre 19 et 97 selon les sujets (unité d'action seule ou combinaisons d'unités d'action dont les 6 émotions de base)	- / Oui / - / -
ADFES [107]	22 (45/55) / 18-25 / Caucasien, Méditerranéen	Vidéo et image 2D du visage	2 (face et 45°) / ? / ?	Uni / Uniforme / Non / Oui	Consigne / 10 (neutre + 6 émotions de base + mépris + embarras + fierté)	- / - / - / - / -
MPI [65]	19 (53/47) / 20-30 / Allemand	Vidéo 2D et image 3D du visage et audio	3 (face et +/-23°) / 768x576 / 50	Uni / Uniforme / Non / Oui (3)	Portrait / 56 (6 émotions de base + mépris + émotions secondaires)	- / - / - / - / -

Suite et fin de la table page suivante

TABLE A.1 – Bases de données d'expressions posées - 7/7

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Reproduction / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
PICS - Stirling ESRC 3D Face Database [102]	99 (55/45) / ? / ?	Vidéo et image 2D et 3D du visage	4 (3D et 2D) / 1000x1400, 1200x1200 ou 800x1200 (85000 ou 3746 vertexes) / ?	Uni et extérieur / Ambiante, uniforme et variable / Oui (lunettes et cheveux) / Oui	? / 7 (neutre + 6 émotions de base)	- / - / - / -
DISFA+ [113]	9 (44/56) / ? / Caucasien, Hispanique, Asiatique, Africain sub-saharien	Vidéo 2D du visage	1 / 1024x768 / 20	Uni / Uniforme / Oui (lunettes) / Non	Consigne / 42 (combinaisons d'unités d'action et 6 émotions de base)	- / Oui / - / -
BAUM-1 [84]	31 (55/45) / 19-65 / Turcophone	Vidéo 2D du visage et audio	2 (face et 45°) / 576x720 / 30	Laboratoire / Uniforme / Non / Oui (2)	Portrait / 8 (colère, dégoût, joie, peur, surprise, confusion, ennui, intérêt/curiosité)	- / - / Oui / -

TABLE A.2 – Bases de données d'expressions spontanées - 1/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
Smile Database [60]	94 (?/ ?) / ? / ?	Vidéo 2D du visage et signal EMG	1 / ? / ?	Uni / Ambiante / Oui (lunettes et capteurs EMG) / Naturelle pour 7 sujets	Tâche passive / 1 (sourire)	Oui / Oui / - / -
UT-Dallas [61]	284 (73/27) / ? / Caucasien, Afrcain sub-saharien, Asiatique, Hispanique, autres	Vidéo et image 2D du visage	9 (-90° :22.5° :90°) / 720x480 / 29.97	Uni / Ambiante / Non / Oui (9)	Tâche passive / 10 (6 émotions de base + émotions secondaires)	- / - / Oui / -
ENTERFACE [94]	16 (37/63) / ? / ?	Vidéo 2D du visage et signaux physiologiques	1 / ? / ?	Uni / Ambiante / Oui (capteur fNIRS) / ?	Tâche passive / 3 (neutre, joie et dégoût)	- / - / - / -
RU-FACS [69]	100 (?/ ?) / ? / ?	Vidéo 2D du visage et du mouvement du corps	4 / ? / ?	Uni / Ambiante / ? / Naturelle	Interaction humain-humain / Variées	- / Oui / - / -

Suite de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 2/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
EmoTABOO (HUMAINE) [101]	18 (39/61) / ? / ?	Vidéo 2D du visage, mouvement du corps et audio	2 (face et haut du corps) / ? / ?	Uni / Ambiante / Oui (lunettes) / Naturelle	Interaction humain-humain / 21 (émotions de base et secondaires)	- / - / Oui / Oui
SAL (HUMAINE) [114]	4 (?/ ?) / ? / ?	Vidéo 2D du visage et audio	1 / ? / ?	Uni / Ambiante / ? / Naturelle	Interaction humain-machine / Variée	- / - / - / Oui
IEMOCAP [88]	10 (50/50) / ? / ?	Capture de mouvement (visage et mains) et audio	8 (capture de mouvement) / - / 120	Laboratoire / Ambiante / Non / Naturelle	Improvisation sur un scénario / 5 (neutre, joie, colère, tristesse, frustration)	Oui / - / Oui / Oui
CK+ [116]	182 (69/31) / 18-50 / Caucasien, African sub-saharien, autres	Vidéo 2D du visage	1 / 640x490 (N&B) ou 640x480 (couleur) / ?	Uni / 1/3 ambiante et 2/3 uniforme / Non / Non	Sourire spontané pendant acquisition posée / 1 (sourire)	Oui / Oui / Oui / -

Suite de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 3/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
MMI+ [111]	25 (48/52) / 20-32 / Caucasien, Asiatique, Sud-américain	Vidéo 2D du visage et audio	1 / ? / ?	Uni et laboratoire / Ambiante / Oui (lunettes) / Naturelle	Tâche passive / 3 (dégoût, joie, surprise)	- / Oui / - / -
MUG [22]	45 (42/58) / 20-35 / ?	Vidéo 2D du visage	1 / 896x896 / 19	Uni / Uniforme / Oui (cheveux et lunettes) / Non	Tâche passive / 7 (neutre + 6 émotions de base)	Oui / - / - / -
SEMAINE [89]	21 (62/38) / 22-60 / Varié (8 pays différents)	Vidéo 2D du visage et audio	1 / ? / ?	Uni / Ambiante / ? / Naturelle	Interaction humain-machine / Variées	- / - / Oui / Oui
NVIE [78]	215 (27/73) / 17-31 / Asiatique	Vidéo 2D, image 2D et vidéo infrarouge du visage	2 (visible et infrarouge) / 704x480 (visible) et 320x240 (infrarouge) / 30 (visible) et 25 (infrarouge)	Uni / 3 configurations / Oui (lunettes) / Non	Tâche passive / 7 (neutre + 6 émotions de base)	Oui / - / Oui / Oui

Suite de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 4/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
UNBC-McMaster Shoulder Pain Expression Archive [112]	129 (51/49) / ? / ?	Vidéo 2D du visage	1 / ? / ?	Uni / Ambiante / ? / Naturelle	Tâche active (liée à la douleur) / Liées à la douleur	Oui / Oui / - / Oui
CAM3D [91]	16 (50/50) / 24-50 / Caucasien, Asiatique, Moyen-oriental	Vidéo 3D et 2D du visage et audio	2 (2D et Kinect) / 720x576 (640x480 Kinect) / 30	Laboratoire / Uniforme / Oui / Naturelle	Interaction humain-humain et interaction humain-machine / 10 (émotions de base + émotions secondaires)	- / - / Oui / -
MAHNOB-HCI [97]	27 (59/51) / 19-40 / ?	Vidéo 2D du visage et du mouvement du corps, signaux physiologiques et audio	6 /780x580 / 60	Uni / ? / ? / ?	Tâche passive / 6 (émotions de base + émotions secondaires)	- / - / - / Oui

Suite de la table page suivante

Table A.2 – Bases de données d'expressions spontanées - 5/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
BINED - Ensemble 1 [70]	114 (39/61) / ? / Irlandais du Nord	Vidéo 2D du visage et du mouvement du torse	1 / 720x576 / ?	Laboratoire / Ambiante / Non / Naturelle	Tâche active et tâche passive / 5 (dégoût, peur, surprise, frustration, amusement)	- / - / Oui / Oui
BINED - Ensemble 2 [70]	82 (55/45) / ? / Irlandais du Nord	Vidéo 2D du visage et du mouvement du torse	1 / 720x576 / ?	Laboratoire / Ambiante / Non / Naturelle	Tâche active et tâche passive / 6 (colère, dégoût, peur, tristesse, surprise, amusement)	- / - / Oui / Oui
BINED - Ensemble 3 [70]	60 (50/50) / ? / Irlandais du Nord, Péruvien	Vidéo 2D du visage et du mouvement du torse	1 / 1920x1080 / ?	Laboratoire / Ambiante / Non / Naturelle	Tâche active et tâche passive / 3 (dégoût, peur, amusement)	- / - / Oui / Oui

Suite de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 6/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
DEAP [95]	32 (50/50) / 19-37 / Caucasien, autres	Vidéo 2D du visage et signaux physiologiques	1 / ? / ?	Uni / Uniforme / Oui (lunettes) / ?	Tâche passive / Variées (4 quadrants de l'espace arousal-valence)	- / - / - / Oui
AVEC 2013 AViD-Corpus [92]	292 (?/ ?) / 18-63 / ?	Vidéo 2D du visage et audio	1 / 640x480 / 30	Laboratoire / ? / ? / ?	Tâche active / Variées	Oui / - / - / Oui
DISFA [105]	27 (44/56) / 18-50 / Caucasien, Hispanique, Asiatique, Africain sub-saharien	Vidéo 2D du visage	1 / 1024x7688 / 20	Uni / Uniforme / Non / Naturelle	Tâche passive / Unités d'action	Oui / Oui / - / -
PICS - Stirling ESRC 3D Face Database [102]	99 (55/45) / ? / ?	Vidéo et image 2D et 3D du visage	4 (3D et 2D) / 1000x1400, 1200x1200 ou 800x1200 (85000 ou 3746 vertexes) / ?	Uni et extérieur / Ambiante, uniforme et variable / Oui (lunettes et cheveux) / Oui	Tâche active et tâche passive / 7 (6 émotions de base + horreur)	- / - / - / -

Suite de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 7/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
DynEmo [83]	358 (51/49) / 25-65 / Caucasien	Vidéo 2D du visage et du mouvement du corps	2 (face et 45°) / 768x576 / 25	Uni / Ambiante / Oui (lunettes) / Naturelle	Tâche active et tâche passive / 12 (émotions de base + émotions secondaires)	- / - / Oui / Oui
RECOLA [98]	46 dont 23 disponibles (59/41) / ? / Caucasien	Vidéo 2D du visage, audio et signaux physiologiques	1 / 1280x720 / 25	Laboratoire / Uniforme / Oui / Naturelle	Interaction humain-humain / Variées	- / - / - / Oui
BP4D-Spontaneous [103]	41 (56/44) / 18-29 / Caucasien, Asiatique, Africain sub-saharien, Hispanique	Vidéo 3D du visage	3 (1 paire stéréo et 1 pour la texture) / 1040x1392 (30000 à 50000 vertexes) / 25	Uni / Uniforme / ? / Naturelle	Tâche active et passive / 8 (6 émotions de base + embarras + douleur)	Oui / Oui / - / -
CAS(ME)2 [77]	22 (73/27) / ? / Asiatique	Vidéo 2D du visage	1 / 640x480 / 30	Uni / Uniforme / Non / Non	Tâche passive / 3 (colère, dégoût, joie)	- / Oui / Oui / -

Suite et fin de la table page suivante

TABLE A.2 – Bases de données d'expressions spontanées - 8/8

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Type d'induction émotionnelle / Nombre d'expressions	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
BioVid Emo [96]	94 dont 86 disponibles (53/47) / 18-65 / ?	Vidéo 2D du visage, carte de profondeur et signaux physiologiques	4 (1 kinect et 3 en 2D (face et +/-45°)) / 1388x1038 (640x480 Kinect) / 25	Laboratoire / Uniforme / ? / Oui (3)	Tâche passive / 5 (colère, dégoût, peur, tristesse et amusement)	- / - / Oui / -
BAUM-1 [84]	31 (55/45) / 19-65 / Turcophone	Vidéo 2D du visage et audio	2 (face et 45°) / 576x720 / 30	Laboratoire / Uniforme / ? / Oui (2)	Tâche passive / 12 (6 émotions de base + émotions secondaires)	- / - / Oui / -

TABLE A.3 – Bases de données d'expressions « in-the-wild » - 1/2

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Nature des données / Type d'induction	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
Belfast Naturalistic (HUMAINE) [72]	125 (75/25) / ? / ?	Vidéo 2D du visage et audio	1 / ? / ?	Studio télé / Ambiante / ? / Naturelle	Extraits télévisés / Interaction humain-humain	- / - / Oui / Oui
EmoTV (HUMAINE) [108]	48 (?/ ?) / ?	Vidéo 2D du visage et audio	1 / ? / ?	Varié / Ambiante / Oui / Naturelle	Extraits télévisés / Interaction humain-humain	- / - / Oui / Oui
VAM [73]	104 (?/ ?) / 16-69 / ?	Vidéo et image 2D du visage et audio	1 / 352x288 / 25	Varié / Ambiante / ? / Naturelle	Extraits télévisés / Interaction humain-humain	- / - / Oui / Oui
AFEW [74]	330 (?/ ?) / 1-70 / Varié	Vidéo 2D du visage et audio	1 / ? / ?	Varié / Ambiante / Oui / Naturelle	Extraits de films (neutre + 6 émotions de base) / -	Oui / - / Oui / -
SFEW [109]	330 (?/ ?) / 1-70 / Varié	Image 2D du visage	1 / ? / -	Varié / Ambiante / Oui / Naturelle	Images extraites de AFEW [74] / -	Oui / - / Oui / -

Suite et fin de la table page suivante

TABLE A.3 – Bases de données d'expressions « in-the-wild » - 2/2

Bases de données	Population Nb sujets (% F/H) / Âges extrêmes / Groupe(s) ethnique(s)	Modalité(s)	Matériel d'acquisition Nb caméras / Résolution / FPS	Conditions expérimentales Fond / Illumination / Occlusions / Pose	Protocole expérimental Nature des données / Type d'induction	Annotations Caractéristiques faciales / Unités d'action / Labels / Dimensions émotionnelles
AM-FED [75]	242 (42/58) / ? / ?	Vidéo 2D du visage	1 / 320x240 / 14	Varié / Ambiante / Oui / Naturelle	« Crowdsourcing » / Tâche passive (uniquement pour le sourire)	Oui / Oui / Oui / Oui
Aff-Wild [115]	500+ (vidéos) et 2000+ (images) (?/ ?) / ? / ?	Vidéo et image 2D du visage	1 / ? / ?	Varié / Ambiante / Oui / Naturelle	Extraits de vidéos youtube et images recueillies sur google image / -	- / Oui / - / Oui
Vinereactor [76]	222 (?/ ?) / ? / ?	Vidéo 2D du visage	1 / 320x240 / ?	Varié / Ambiante / Oui / Naturelle	« Crowdsourcing » / Tâche passive	Oui / Oui / - / -
CHEAVD [81]	238 (?/ ?) / 11-62 / Asiatique	Vidéo 2D du visage et audio	1 / 640x480 / 25	Varié / Ambiante / ? / Naturelle	Extraits télévisés et de films / Interaction humain-humain	- / - / Oui / -

Annexe B

Robustesse à la pose des caractéristiques angle-distance

Dans cette annexe, nous faisons la démonstration mathématique de la robustesse à la pose des caractéristiques angle-distance. Ce travail a été réalisé par Vincent Barrielle, alors doctorant CIFRE chez Dynamixyz en co-tutelle avec l'équipe FAST. Dans la première section, nous rappelons le contexte et nous indiquons les hypothèses de travail. Dans la seconde section, nous déroulons la démonstration de la robustesse à la pose des caractéristiques angle-distance.

B.1 Contexte et hypothèses

Avec les caractéristiques angle-distance, nous cherchons à définir des caractéristiques capables de décrire l'expression d'un sujet tout en étant invariantes aux variations de pose et inter-personnelles.

La robustesse à la pose de ces caractéristiques repose sur l'hypothèse de faible perspective (« Weak Perspective »). En partant du principe que le sujet sera assis face à la caméra, nous pouvons considérer que la variation des points caractéristiques du visage selon l'axe Z (axe de la caméra) est faible par rapport à la distance à la caméra, ce qui correspond à l'hypothèse de faible perspective. Pour un point caractéristique du visage

de coordonnées $(x, y, z) \in \mathbb{R}^3$, nous pouvons modéliser cette hypothèse comme suit :

$$\begin{pmatrix} u \\ v \\ s \end{pmatrix} = W \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{B.1}$$

où

$$W = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{Z_{moy}}{f} \end{pmatrix} \tag{B.2}$$

avec f la distance focale de la caméra et Z_{moy} la distance moyenne entre le visage et la caméra.

Nous poussons l'approximation plus loin en faisant également l'hypothèse que les points caractéristiques du visage que nous utilisons pour définir les caractéristiques angle-distance appartiennent au même plan. Cette hypothèse reste vérifiée tant que nous n'utilisons pas les points caractéristiques du bout du nez.

B.2 Démonstration

La démonstration repose sur la transformation des points caractéristiques du visage induite par la rotation d'un plan sous l'hypothèse de faible perspective. Nous considérons les points caractéristiques du visage appartenant au plan Z_0 orthogonal à l'axe Z. Nous allons les transformer en faisant une rotation autour des axes X et Y, suivie d'une translation. Nous ne considérons pas la rotation autour de l'axe Z puisqu'elle induit une rotation dans le plan, elle n'a donc pas d'influence sur les caractéristiques angle-distance.

Soit $\theta \in \mathbb{R}$ l'angle de la rotation autour de l'axe X. La matrice de rotation autour de l'axe X est :

$$R_X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_\theta & -s_\theta \\ 0 & s_\theta & c_\theta \end{pmatrix}, \tag{B.3}$$

où $c_\theta = cos(\theta)$ et $s_\theta = sin(\theta)$. Soit $\phi \in \mathbb{R}$ l'angle de la rotation autour de l'axe Y. La matrice de rotation autour de l'axe Y est :

$$R_Y = \begin{pmatrix} c_\phi & 0 & -s_\phi \\ 0 & 1 & 0 \\ s_\phi & 0 & c_\phi \end{pmatrix}, \tag{B.4}$$

où $c_\phi = cos(\phi)$ et $s_\phi = sin(\phi)$. La matrice de rotation considérée est donc :

$$R = R_X R_Y, \tag{B.5}$$

ce qui donne la matrice de transformation de pose suivante :

$$T_p = \begin{pmatrix} c_\phi & 0 & -s_\phi & t_x \\ -s_\phi s_\theta & c_\theta & c_\theta - s_\theta c_\phi & t_y \\ c_\theta s_\phi & s_\theta & s_\theta + c_\theta c_\phi & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{B.6}$$

où $(t_x, t_y, t_z) \in \mathbb{R}^3$ est un vecteur de translation.

En appliquant la faible perspective après la rotation, nous obtenons :

$$WT_p = \begin{pmatrix} c_\phi & 0 & -s_\phi & t_x \\ -s_\phi s_\theta & c_\theta & c_\theta - s_\theta c_\phi & t_y \\ 0 & 0 & 0 & \frac{Z_{moy}}{f} \end{pmatrix}. \tag{B.7}$$

Ainsi, nous pouvons calculer la projection d'un point de coordonnées $(x, y, z_0) \in \mathbb{R}^3$ appartenant au plan Z_0 après avoir appliqué la transformation de pose :

$$WT_p \begin{pmatrix} x \\ Yy \\ z_0 \\ 1 \end{pmatrix} = \begin{pmatrix} c_\phi x - s_\theta z_0 + t_x \\ -s_\phi s_\theta x + c_\theta y + (c_\theta - s_\theta c_\phi) z_0 + t_y \\ \frac{Z_{moy}}{f} \end{pmatrix}. \tag{B.8}$$

Nous notons A_{XY} la matrice suivante :

$$A_{XY} = \begin{pmatrix} c_\phi & 0 & \frac{f(t_x - z_0 s_\theta)}{Z_{moy}} \\ -s_\phi s_\theta & c_\theta & \frac{f(z_0(c_\theta - s_\theta c_\theta) + t_y)}{Z_{moy}} \\ 0 & 0 & 1 \end{pmatrix}. \tag{B.9}$$

Nous avons alors :

$$A_{XY} W \begin{pmatrix} x \\ y \\ z_0 \\ 1 \end{pmatrix} = WT_p \begin{pmatrix} x \\ y \\ z_0 \\ 1 \end{pmatrix}. \tag{B.10}$$

Cela prouve que l'application d'une transformation rigide sur un plan sous l'hypothèse de faible perspective induit une transformation affine dans l'espace de l'image. De

plus, si nous faisons l'hypothèse de faibles angles de rotation, le développement de Taylor de A_{XY} donne :

$$A_{XY} = \begin{pmatrix} 1 & 0 & k_x \\ 0 & 1 & k_y \\ 0 & 0 & 1 \end{pmatrix} + O(\theta\phi) + O(\theta^2) + O(\phi^2), \tag{B.11}$$

où $(k_x, k_y) \in \mathbb{R}^2$ sont des constantes. Cela montre que, pour de faibles rotations du plan, la transformation induite est une similarité qui préserve les ratios de longueur, ainsi que les angles.

Annexe C

Rectification de pose

Dans cette annexe, nous présentons notre méthode de rectification de pose. La rectification de pose est le fait de normaliser les points caractéristiques du visage par rapport à la pose, *i.e.* remettre le visage de face. Cette méthode de rectification de pose a été implémentée par Vincent Barrielle, alors doctorant CIFRE chez Dynamixyz en co-tutelle avec l'équipe FAST.

La méthode se déroule en deux temps. Premièrement, nous calculons la pose de la tête du visage 2D à l'aide d'un modèle 3D d'un visage neutre de référence. Pour ce faire, nous cherchons quels sont les paramètres de pose qui, une fois appliqués sur le modèle 3D de référence, minimise l'erreur entre ce modèle 3D et le visage 2D. Deuxièmement, nous calculons les points 3D du visage après rectification de pose en appliquant les paramètres de pose obtenus précédemment et en faisant épouser sa forme au visage 2D.

Soit $X \in M_{n_p,2}(\mathbb{R})$ les points caractéristiques du visage dont la pose doit être rectifiée, où n_p est le nombre de points caractéristiques. Soit $X^{ref} \in M_{n_p,3}(\mathbb{R})$ les points 3D d'un visage neutre de référence. On note $X^f \in M_{n_p,3}(\mathbb{R})$ les points 3D du visage après rectification de pose.

Dans un premier temps, nous cherchons le facteur d'échelle $s \in \mathbb{R}$, la matrice de rotation $R \in M_{3,3}(\mathbb{R})$ et la translation $t \in \mathbb{R}^3$ de la pose du visage en minimisant sous l'hypothèse de faible perspective (voir sous-section **??** pour la définition de cette hypothèse) :

$$E_{pose}(s, R, t) = ||sPR(X^{ref} + t) - X||^2 \tag{C.1}$$

où $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ est la matrice de projection orthographique dans le plan orthogonal à l'axe Z.

Ensuite, nous calculons les coordonnées des points 3D après rectification de pose X^f de manière à épouser au mieux la forme des points caractéristiques X. Cela se fait en minimisant l'expression suivante tout en gardant les paramètres R et t fixés :

$$\begin{aligned} E(X^f) = w_1 E_1(X^f) + w_2 E_2(X^f) \\ + w_3 E_3(X^f) + w_4 E_4(X^f). \end{aligned} \tag{C.2}$$

E_1 décrit à quel point X^f épouse la forme des points caractéristiques X :

$$E_1(X^f) = ||W(R)sPR(X^f + t) - X||^2, \tag{C.3}$$

où $W(R)$ est une matrice diagonale qui sous-pondère les vertexes qui sont moins visibles à cause de la rotation.

E_2 cherche à préserver la structure différentielle de X^{ref} :

$$E_2(X^f) = ||GX^f - GX^{ref}||^2, \tag{C.4}$$

avec G la matrice qui transforme un maillage 3D en un ensemble spécifique de bords.

E_3 cherche à préserver la structure absolue de X^{ref} :

$$E_3(X^f) = ||DX^f - DX^{ref}||^2, \tag{C.5}$$

avec D une matrice diagonale qui sur-pondère pour préserver les coordonnées de l'axe Z.

Enfin, E_4 permet d'inférer la position des vertexes non visibles à cause de la pose :

$$E_4(X^f) = ||S(R)X^f||^2, \tag{C.6}$$

où les lignes de $S(R)$ ont chacune deux valeurs non nulles, $w(R)$ et $-w(R)$, situées de telle sorte que la coordonnée de l'axe Y d'un vertexe soit égale à la coordonnée de l'axe Y d'un vertexe symétrique.

Toutes les optimisations sont effectuées avec l'algorithme de Gauss-Newton.

Glossaire

Espace des signatures : Représentation indépendante de la personne du modèle spécifique à la personne.

Espace PCA : Espace obtenu après l'analyse en composante principales sur les expressions de base du modèle.

Expression courante : Expression à l'instant courant lors de l'adaptation du modèle.

Expressions de base : Ensemble des expressions utilisées pour construire le modèle expressif spécifique à la personne.

Expressions de base de la vérité terrain : Expressions de base posées capturées à l'« apex ».

Expressions de base plausibles : Expressions de base synthétisées à partir du visage neutre.

Expressions de base réelles : Expressions de base réalisées par le sujet durant la vidéo sur laquelle l'adaptation est effectuée.

Facteur d'amélioration : Pourcentage d'amélioration moyen sur la base de données quantifiant le rapprochement des modèles adaptés à la vérité terrain, relativement aux modèles plausibles.

Modèle adapté : Modèle obtenu à l'issue de l'adaptation.

Modèle plausible : Modèle construit à partir des expressions de base plausibles.

Nombre moyen de mises à jour : Nombre de mises à jour durant l'adaptation (en prenant en compte les mises à jour globales et locales), moyenné sur la base de données.

Nombre moyen de mises à jour finales : Nombre d'expressions de base du modèle mises à jour à l'issue de l'adaptation, moyenné sur la base de données.

Nombre moyen de mises à jour finales correctes : Nombre d'expressions de base du modèle mises à jour à l'issue de l'adaptation avec une expression correspondant aux déformations faciales de l'expression de base attendue, moyenné sur la base de données.

Pourcentage moyen de mises à jour finales correctes : Rapport entre le nombre de mises à jour finales correctes et le nombre de mises à jour, moyenné sur la base de données.

Signature : Projection de l'expression dans l'espace des signatures.

Structure du modèle : Ensemble de simplexes renvoyées par la tessellation de Delaunay effectuée sur les expressions de base (neutre inclus) dans l'espace PCA du modèle.

Taux d'amélioration : Pourcentage de sujet sur la base de données pour lesquels le modèle adapté est plus proche de la vérité terrain que ne l'est le modèle plausible.

Notations

$n_e \in \mathbb{N}$ **:** Nombre d'expressions de base du modèle (neutre inclus).

$n_p \in \mathbb{N}$ **:** Nombre de points caractéristiques utilisés pour décrire le visage.

$n_{si} \in \mathbb{N}$ **:** Nombre de simplexes dans la structure du modèle.

$T \in \mathbb{N}$ **:** Longueur de la fenêtre temporelle en nombre d'images.

$N_e^* \subset [\![1, n_e]\!]$ **:** Sous-ensemble des indices des expressions de base du modèle sur lesquelles le taux d'amélioration et le facteur d'amélioration sont calculés.

$X \in \mathbb{R}^{2n_p}$ **:** Vecteur contenant les points caractéristiques de l'expression courante.

$\forall i \in [\![1, n_e]\!], X^i \in \mathbb{R}^{2n_p}$ **:** Vecteur contenant les points caractéristiques de la i-ième expression de base du modèle.

$\bar{X^e} \in \mathbb{R}^{2n_p}$ **:** Vecteur contenant la moyenne de l'ensemble des points caractéristiques des expressions de base du modèle (neutre inclus) après alignement.

$\Phi \in M_{2n_p,3}\mathbb{R}$ **:** Matrice de projection des points caractéristiques d'une expression dans l'espace PCA du modèle.

$X_{PCA} \in \mathbb{R}^3$ **:** Vecteur PCA de l'expression courante (projection dans l'espace PCA du modèle).

$\forall i \in [\![1, n_e]\!], X_{PCA}^i \in \mathbb{R}^3$ **:** Vecteur PCA de la i-ème expression de base du modèle (projection dans l'espace PCA du modèle).

$X_{PCAn} \in \mathbb{R}^3$ **:** Vecteur PCA centré sur le neutre et normalisé de l'expression courante.

$\forall i \in [\![1, n_e]\!], X^i_{PCAn} \in \mathbb{R}^3$ **:** Vecteur PCA centré sur le neutre et normalisé de la i-ième expression de base du modèle.

$X_{sig} \in \mathbb{R}^3$ **:** Vecteur contenant la signature de l'expression courante.

$I \in \mathbb{R}$ Intensité de l'expression courante.

$\forall i \in [\![1, n_e]\!], I^i \in \mathbb{R}$ **:** Intensité de la i-ième expression de base du modèle.

$D_{global} \in \mathbb{R}$ **:** Distance de déformation faciale entre l'expression courante et l'expression de base du modèle avec le label l_{global}.

l_{global} **:** Label global (visage entier) de l'expression courante renvoyé par le système de reconnaissance d'expressions faciales robuste à la pose. Les valeurs possibles sont les expressions de base du modèle (neutre, colère, dégoût, joie, surprise, tristesse).

$l_{sourcils}$ **:** Label local (sourcils) de l'expression courante renvoyé par le système de reconnaissance d'expressions faciales robuste à la pose. Les valeurs possibles sont neutre, froncés et levés.

l_{yeux} **:** Label global (yeux) de l'expression courante renvoyé par le système de reconnaissance d'expressions faciales robuste à la pose. Les valeurs possibles sont ouverts et mi-fermés/fermés.

l_{bouche} **:** Label global (bouche) de l'expression courante renvoyé par le système de reconnaissance d'expressions faciales robuste à la pose. Les valeurs possibles sont les expressions de base du modèle (neutre, colère, dégoût, joie, surprise, tristesse).

ϵ **:** Seuil sur l'erreur de reconstruction dans l'espace PCA.

$d_{global} \in \mathbb{R}$ **:** Seuil sur la distance de déformation faciale appliqué dans les contraintes globales, sauf si $l_{global} = joie$.

$d_{joie} \in \mathbb{R}$ **:** Seuil sur la distance de déformation faciale appliqué dans les contraintes locales et globales si $l_{global} = joie$ (ou $l_{bouche} = joie$ si contraintes locales).

$d_{local} \in \mathbb{R}$ **:** Seuil sur la distance de déformation faciale appliqué dans les contraintes locales, sauf si $l_{bouche} = joie$.

$\forall i \in N_e^*, \forall j \in [\![1, n_s]\!], \forall k \in \{pl, ad\}, m_{i,j,k} \in \mathbb{R}$ **:** Métrique du sujet j pour la i-ième expression de base du modèle k (pl : plausible, ad : adapté).

α **:** Taux d'amélioration.

β **:** Facteur d'amélioration.

Table des figures

Liste des tableaux

Bibliographie

[1] Klaus R Scherer, Klaus R Scherer, and Paul Ekman. On the nature and function of emotion : A component process approach. *Approaches to emotion*, 2293 :317, 1984.

[2] Albert Mehrabian. Communication without words. *Psychological today*, 2 :53–55, 1968.

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

[4] `https://en.wikipedia.org/wiki/Theatre_of_ancient_Greece`.

[5] Maja Pantic and Leon JM Rothkrantz. Automatic analysis of facial expressions : The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12) :1424–1445, 2000.

[6] Paul Ekman and Wallace V Friesen. Facial action coding system. 1977.

[7] Paul Ekman. Telling lies : Clues to deceit in the marketplace, marriage, and politics, 1985.

[8] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. Movement differences between deliberate and spontaneous facial expressions : Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, 30(1) :37–52, 2006.

[9] Michel F Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

[10] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. Spontaneous vs. posed facial behavior : automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006.

[11] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michal R Wrobel. Emotion recognition and its applications. In *Human-Computer Systems Interaction : Backgrounds and Applications 3*, pages 51–62. Springer, 2014.

[12] Brais Martinez and Michel F Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in Face Detection and Facial Image Analysis*, pages 63–100. Springer, 2016.

[13] Klaus R Scherer, Tanja Wranik, Janique Sangsue, Véronique Tran, and Ursula Scherer. Emotions in everyday life : probability of occurrence, risk factors, appraisal and reaction patterns. *Social Science Information*, 43(4) :499–570, 2004.

[14] Catherine Soladié, Nicolas Stoiber, and Renaud Séguier. Invariant representation of facial expressions for blended expression recognition on unknown subjects. *Computer Vision and Image Understanding*, 117(11) :1598–1609, 2013.

[15] Charles Darwin. The expression of the emotions in man and animais. *Murray, London*, 1872.

[16] Klaus R Scherer. Psychological models of emotion. *The neuropsychology of emotion*, 137(3) :137–162, 2000.

[17] Paul Ekman. Cross-cultural studies of facial expression. *Darwin and facial expression : A century of research in review*, pages 169–222, 1973.

[18] Paul Ekman. Face of man : Universal expression in a new guinea village. *New York : Garland*, 1980.

[19] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition : a meta-analysis. *Psychological bulletin*, 128(2) :203, 2002.

[20] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4) :169–200, 1992.

[21] Paul Ekman and Wallace V Friesen. A new pan-cultural facial expression of emotion. *Motivation and emotion*, 10(2) :159–168, 1986.

[22] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE, 2010.

[23] S Baron-Cohen, O Golan, S Wheelwright, and JJ Hill. Mind reading : The interactive guide to emotions. london : Jessica kingsley, 2004.

[24] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3) :273–294, 1977.

[25] James A Russell and Geraldine Pratt. A description of the affective quality attributed to environments. *Journal of personality and social psychology*, 38(2) :311, 1980.

[26] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion : dissecting the elephant. *Journal of personality and social psychology*, 76(5) :805, 1999.

[27] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1) :145, 2003.

[28] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12) :1050–1057, 2007.

[29] Faiza Abdat. Reconnaissance automatique des émotions par données multimodales : expressions faciales et signaux physiologiques. *Université de Metz, France*, 2010.

[30] Klaus R Scherer. What are emotions ? and how can they be measured ? *Social science information*, 44(4) :695–729, 2005.

[31] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion : Theory, methods, research*. Oxford University Press, 2001.

[32] Richard S Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8) :819, 1991.

[33] Craig A Smith and Leslie D Kirby. Consequences require antecedents. *Feeling and thinking : The role of affect in social cognition*, page 83, 2001.

[34] Klaus R Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion : Theory, methods, research*, 92(120) :57, 2001.

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

[36] Ashok Samal and Prasana A Iyengar. Automatic recognition and analysis of human faces and facial expressions : A survey. *Pattern recognition*, 25(1) :65–77, 1992.

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

[38] Ying-Li Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer, 2005.

[39] Zhihong Zeng, Maja Pantic, Glenn Roisman, Thomas S Huang, et al. A survey of affect recognition methods : Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1) :39–58, 2009.

[40] Amit Konar, Anisha Halder, and Aruna Chakraborty. Introduction to emotion recognition. In *Emotion Recognition : A Pattern Analysis Approach*. John Wiley & Sons, Inc., 2014.

[41] Muhammad Hameed Siddiqi, Maqbool Ali, Mohamed Elsayed Abdelrahman Eldib, Asfandyar Khan, Oresti Banos, Adil Mehmood Khan, Sungyoung Lee, and Hyunseung Choo. Evaluating real-life performance of the state-of-the-art in facial expression recognition using a novel youtube-based datasets. *Multimedia Tools and Applications*, pages 1–21, 2017.

[42] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4) :1334–1345, 2007.

[43] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 375–388. Springer, 2007.

[44] Deepak Ghimire, Joonwhoan Lee, Ze-Nian Li, and Sunghwan Jeong. Recognition of facial expressions based on salient geometric features and support vector machines. *Multimedia Tools and Applications*, pages 1–26, 2016.

[45] Jovan de Andrade Fernandes, Leonardo Nogueira Matos, and Maria Géssica dos Santos Aragão. Geometrical approaches for facial expression recognition using support vector machines. In *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*, pages 347–354. IEEE, 2016.

[46] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1) :329–349, 2003.

[47] Sidney D'Mello and Rafael A Calvo. Beyond the basic emotions : what should affective computing compute ? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2287–2294. ACM, 2013.

[48] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[49] M Valstar, J Girard, T Almaev, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and J Cohn. Fera 2015 - second facial expression recognition and analysis challenge. *Proc. IEEE ICFG*, 2015.

[50] Jun Wang, Shangfei Wang, and Qiang Ji. Facial action unit classification with hidden knowledge under incomplete annotation. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 75–82. ACM, 2015.

[51] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012 : the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[52] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. Av+ ec 2015 : The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

[53] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of AVEC'16, co-located with ACM MM 2016*, Amsterdam, The Netherlands, October 2016. ACM.

[54] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013.

[55] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge : Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv :1703.01210*, 2017.

[56] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[57] D. Lundqvist, A. Flykt, and A Öhman. The karolinska directed emotional faces - kdef, cd rom from department of clinical neuroscience, psychology section, karolinska institutet, 1998.

[58] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[59] Jeffrey F Cohn and Karen L Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02) :121–132, 2004.

[60] Karen L Schmidt and Jeffrey F Cohn. Dynamics of facial expression : Normative characteristics and individual differences. In *ICME*. Citeseer, 2001.

[61] Alice JO Toole, Joshua Harms, Sarah L Snow, Dawn R Hurst, Matthew R Pappas, Janet H Ayyad, and Hervé Abdi. A video database of moving faces and people. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5) :812–816, 2005.

[62] C Anitha, MK Venkatesha, and B Suryanarayana Adiga. A survey on facial expression databases. *International Journal of Engineering Science and Technology*, 2(10) :5158–5174, 2010.

[63] Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. Beyond emotion archetypes : Databases for emotion modelling using neural networks. *Neural networks*, 18(4) :371–388, 2005.

[64] Siyao Fu, Guosheng Yang, Xinkai Kuai, and Rui Zheng. *A Parametric Survey for Facial Expression Database*, pages 373–381. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[65] Kathrin Kaulard, Douglas W Cunningham, Heinrich H Bülthoff, and Christian Wallraven. The mpi facial expression database a validated database of emotional and conversational facial expressions. *PloS one*, 7(3) :e32321, 2012.

[66] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[67] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008.

[68] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[69] `http://mplab.ucsd.edu/grants/project1/research/rufacs1-dataset.html`, 2006.

[70] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *Affective Computing, IEEE Transactions on*, 3(1) :32–41, 2012.

[71] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5) :807–813, 2010.

[72] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. A new emotion database : considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[73] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE, 2008.

[74] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. 2012.

[75] Daniel McDuff, Rana El Kaliouby, Thibaud Senechal, Mohammed Amr, Jeffrey F Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed) : Naturalistic and spontaneous facial expressions collected" in-the-wild". In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 881–888. IEEE, 2013.

[76] Edward Kim and Shruthika Vangala. Vinereactor : Crowdsourced spontaneous facial expression data. In *International Conference on Multimedia Retrieval (ICMR)*. IEEE, 2016.

[77] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, and Xiaolan Fu. Cas(me)2 : A database of spontaneous macro-expressions and micro-expressions. In *International Conference on Human-Computer Interaction*, pages 48–59. Springer, 2016.

[78] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on*, 12(7) :682–691, 2010.

[79] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition : a novel 3d relightable facial database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 611–618. IEEE, 2011.

[80] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and Ad van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8) :1377–1388, 2010.

[81] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. Cheavd : a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2016.

[82] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2296–2303. IEEE, 2011.

[83] Anna Tcherkassof, Damien Dupré, Brigitte Meillon, Nadine Mandran, Michel Dubois, and Jean-Michel Adam. Dynemo : A video database of natural facial expressions of emotions. *The International Journal of Multimedia & Its Applications*, 5(5) :61–80, 2013.

[84] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. Baum-1 : A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 2016.

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

[86] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database : addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

[87] Tanja Bänziger, Hannes Pirker, and K Scherer. Gemep-geneva multimodal emotion portrayals : A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, volume 6, pages 15–019, 2006.

[88] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4) :335–359, 2008.

[89] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010.

[90] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *Multimedia, IEEE Transactions on*, 12(6) :591–598, 2010.

[91] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek. 3d corpus of spontaneous complex mental states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 205–214. Springer, 2011.

[92] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013 : the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[93] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, Jeffrey F Cohn, and Fernando De la Torre. Sayette group formation task (gft) spontaneous facial expression database. 2017.

[94] Arman Savran, Koray Ciftci, Guillaume Chanel, Javier Mota, Luong Hong Viet, Blent Sankur, Lale Akarun, Alice Caplier, and Michele Rombaut. Emotion detection in the loop from brain signals and facial images. 2006.

[95] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap : A database for emotion analysis ; using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1) :18–31, 2012.

[96] Lin Zhang, Steffen Walter, Xueyao Ma, Philipp Werner, Ayoub Al-Hamadi, Harald C Traue, and Sascha Gruss. "biovid emo db" : A multimodal database for emotion analyses validated by subjective ratings. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–6. IEEE, 2016.

[97] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1) :42–55, 2012.

[98] Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of EmoSPACE 2013, held in conjunction with FG 2013*, Shanghai, China, April 2013. IEEE.

[99] Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities : face, body gesture, speech. In *Affect and emotion in human-computer interaction*, pages 92–103. Springer, 2008.

[100] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Pattern Recognition,*

2006. ICPR 2006. 18th International Conference on, volume 1, pages 1148–1153. IEEE, 2006.

[101] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. Collection and annotation of a corpus of human-human multimodal interactions : Emotion and others anthropomorphic characteristics. In *Affective computing and intelligent interaction*, pages 464–475. Springer, 2007.

[102] `http://pics.stir.ac.uk`, 2013.

[103] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous : a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10) :692–706, 2014.

[104] `http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA`, 2009.

[105] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa : A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2) :151–160, 2013.

[106] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[107] Job Van Der Schalk, Skyler T Hawk, Agneta H Fischer, and Bertjan Doosje. Moving faces, looking places : validation of the amsterdam dynamic facial expression set (adfes). *Emotion*, 11(4) :907, 2011.

[108] Sarkis Abrilian, Laurence Devillers, S Buisine, and Jean-Claude Martin. Emotv1 : Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*, 2005.

[109] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011.

[110] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps) : Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58, 1997.

[111] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise : an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC) : Corpora for Research on Emotion and Affect*, page 65, 2010.

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

[113] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended disfa dataset : Investigating posed and spontaneous facial expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[114] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and DKJ Heylen. The sensitive artificial listner : an induction technique for generating emotionally coloured conversation. 2008.

[115] Stefanos Zafeiriou, Athanasios Papaioannou, Irene Kotsia, Mihalis A Nicolaou, Guoying Zhao, E Antonakos, P Snape, G Trigeorgis, and S Zafeiriou. Facial affect "in-the-wild" : A survey and a new database. In *International Conference on Computer Vision*, 2016.

[116] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[117] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace' : An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[118] Edward J Peacock and Paul TP Wong. The stress appraisal measure (sam) : A multidimensional approach to cognitive appraisal. *Stress and Health*, 6(3) :227–236, 1990.

[119] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.

[120] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks : Coping with few data and the training sample order. *Pattern Recognition*, 61 :610–628, 2017.

[121] Huibin Li, Jian Sun, Dong Wang, Zongben Xu, and Liming Chen. Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition. *arXiv preprint arXiv :1511.03015*, 2015.

[122] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild : past, present and future. *Computer Vision and Image Understanding*, 138 :1–24, 2015.

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

[124] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6) :915–928, 2007.

[125] Arnaud Dapogny, Kevin Bailly, and Séverine Dubuisson. Pairwise conditional random forests for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3783–3791, 2015.

[126] Wei Zhang, Youmei Zhang, Lin Ma, Jingwei Guan, and Shijie Gong. Multimodal learning for facial expression recognition. *Pattern Recognition*, 48(10) :3191–3202, 2015.

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

[128] Deepak Ghimire, Sunghwan Jeong, Joonwhoan Lee, and San Hyun Park. Facial expression recognition based on local region specific features and support vector machines. *Multimedia Tools and Applications*, pages 1–19, 2016.

[129] Xiao Zhang, Mohammad H Mahoor, and S Mohammad Mavadati. Facial expression recognition using {l} _ {p}-norm mkl multiclass-svm. *Machine Vision and Applications*, 26(4) :467–483, 2015.

[130] Alessandra Bandrabur, Laura Florea, Corneliu Florea, and Matei Mancas. Emotion identification by facial landmarks dynamics analysis. In *Intelligent Computer Communication and Processing (ICCP), 2015 IEEE International Conference on*, pages 379–382. IEEE, 2015.

[131] Ya Chang, Changbo Hu, and Matthew Turk. Manifold of facial expression. In *AMFG*, pages 28–35, 2003.

[132] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Appearance manifold of facial expression. In *International Workshop on Human-Computer Interaction*, pages 221–230. Springer, 2005.

[133] Yeongjae Cheon and Daijin Kim. Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7) :1340–1350, 2009.

[134] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *BMVC*, pages 297–306, 2006.

[135] Muhammad Hameed Siddiqi, Md Golam Rabiul Alam, Choong Seon Hong, Adil Mehmood Khan, and Hyunseung Choo. A novel maximum entropy markov model for human facial expression recognition. *PloS one*, 11(9) :e0162702, 2016.

[136] Nicu Sebe, Michael S Lew, Ira Cohen, Ashutosh Garg, and Thomas S Huang. Emotion recognition using a cauchy naive bayes classifier. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 17–20. IEEE, 2002.

[137] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences : temporal and static modeling. *Computer Vision and image understanding*, 91(1) :160–187, 2003.

[138] Aruna Chakraborty, Amit Konar, Uday Kumar Chakraborty, and Amita Chatterjee. Emotion recognition from facial expressions and its control using fuzzy logic. *IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans*, 39(4) :726–743, 2009.

[139] Curtis Padgett and Garrison W Cottrell. Representing face images for emotion classification. *Advances in neural information processing systems*, pages 894–900, 1997.

[140] Liying Ma and Khashayar Khorasani. Facial expression recognition using constructive feedforward neural networks. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 34(3) :1588–1595, 2004.

[141] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, pages 1–17, 2016.

[142] Alberto Fernández, Rubén Usamentiaga, Juan Luis Carús, and Rubén Casado. Driver distraction using visual-based sensors and algorithms. *Sensors*, 16(11) :1805, 2016.

[143] Zhiwei Zhu and Qiang Ji. Robust real-time face pose and facial expression recovery. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 681–688. IEEE, 2006.

[144] Shiro Kumano, Kazuhiro Otsuka, Junji Yamato, Eisaku Maeda, and Yoichi Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International journal of computer vision*, 83(2) :178–194, 2009.

[145] Jing Xiao, Tsuyoshi Moriyama, Takeo Kanade, and Jeffrey F Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology*, 13(1) :85–94, 2003.

[146] Ognjen Rudovic, Maja Pantic, and Ioannis Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(6) :1357–1369, 2013.

[147] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. *arXiv preprint arXiv :1702.04174*, 2017.

[148] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. In *European Conference on Computer Vision*, pages 645–661. Springer, 2016.

[149] Yuxiao Hu, Zhihong Zeng, Lijun Yin, Xiaozhou Wei, Jilin Tu, and Thomas S Huang. A study of non-frontal-view facial expressions recognition. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

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

[151] Nikolas Hesse, Tobias Gehrig, Hua Gao, and Hazim Kemal Ekenel. Multi-view facial expression recognition using local appearance features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3533–3536. IEEE, 2012.

[152] Wenming Zheng, Hao Tang, Zhouchen Lin, and Thomas Huang. Emotion recognition from arbitrary view facial images. *Computer Vision–ECCV 2010*, pages 490–503, 2010.

[153] Arnaud Dapogny, Kévin Bailly, and Séverine Dubuisson. Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests. *IEEE Transactions on Affective Computing*, 2017.

[154] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *Image Processing, IEEE Transactions on*, 24(1) :189–204, 2015.

[155] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1) :188–194, 2016.

[156] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis : A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012.

[157] Gary B Huang and Erik Learned-Miller. Labeled faces in the wild : Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, pages 14–003, 2014.

[158] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Multi-view common space learning for emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 464–471. ACM, 2016.

[159] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016 : Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 427–432. ACM, 2016.

[160] Feifei Zhang, Yongbin Yu, Qirong Mao, Jianping Gou, and Yongzhao Zhan. Pose-robust feature learning for facial expression recognition. *Frontiers of Computer Science*, pages 1–13, 2016.

[161] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 42(4) :966–979, 2012.

[162] Jixu Chen, Xiaoming Liu, Peter Tu, and Amy Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15) :1964–1970, 2013.

[163] Stefanos Eleftheriadis, Ognjen Rudovic, Marc P Deisenroth, and Maja Pantic. Gaussian process domain experts for model adaptation in facial behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26, 2016.

[164] Michael Xuelin Huang, Grace Ngai, Kien A Hua, Stephen CF Chan, and Hong Va Leong. Identifying user-specific facial affects from spontaneous expressions with

minimal annotation. *IEEE Transactions on Affective Computing*, 7(4) :360–373, 2016.

[165] Ardeshir Goshtasby. Piecewise linear mapping functions for image registration. *Pattern Recognition*, 19(6) :459–466, 1986.

[166] Catherine Soladié. *Représentation invariante des expressions faciales. : Application en analyse multimodale des émotions.* PhD thesis, Supélec, 2013.

[167] Anwar Saeed, Ayoub Al-Hamadi, Robert Niese, and Moftah Elzobi. Frame-based facial expression recognition using geometrical features. *Advances in Human-Computer Interaction*, 2014 :4, 2014.

[168] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

[169] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[170] Alptekin Durmuşoğlu and Yavuz Kahraman. Facial expression recognition using geometric features. In *Systems, Signals and Image Processing (IWSSIP), 2016 International Conference on*, pages 1–5. IEEE, 2016.