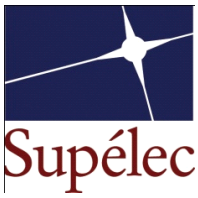
Représentation invariante des expressions faciales. : Application en analyse multimodale des émotions.

Catherine Soladie



N° d'ordre : 2013-32-TH

SUPELEC

École Doctorale MATISSE
« Mathématiques, Télécommunications, Informatique, Signal, Systèmes Électroniques »

THÈSE DE DOCTORAT

DOMAINE : STIC
Spécialité : Télécommunications

Soutenue le
13 décembre 2013

par :
Catherine SOLADIÉ

Représentation Invariante des Expressions Faciales
Application en Analyse Multimodale des Émotions

Directeur de Thèse :	Renaud SÉGUIER	Professeur, Supélec (IETR)

Composition du jury :

Président du jury :	Alice CAPLIER	Professeur, Grenoble-INP (Gipsa-Lab)
Rapporteurs :	Mohamed DAOUDI	Professeur, Télécom Lille1 (LIFL)
	Patrick LAMBERT	Professeur, Polytech Annecy-Chambery (LISTIC)
Examinateur :	Patrice DALLE	Professeur, Université Paul Sabatier (IRIT)

Le visage est l'image de l'âme.

Cicéron - *De oratore, III, 221 - env. 50 av. J.-C.*

Résumé

De plus en plus d'applications ont pour objectif d'automatiser l'analyse des comportements humains afin d'aider ou de remplacer les experts qui réalisent actuellement ces analyses. Cette thèse traite de l'analyse des expressions faciales qui fournissent des informations clefs sur ces comportements.

Les travaux réalisés portent sur une solution innovante permettant de définir efficacement une expression d'un visage, indépendamment de la morphologie du sujet. Pour s'affranchir des différences de morphologies entre les personnes, nous utilisons des modèles d'apparence spécifiques à la personne. Nous proposons une solution qui permet à la fois de tenir compte de l'aspect continu de l'espace des expressions et de la cohérence des différentes parties du visage entre elles.

Pour ce faire, nous proposons une approche originale basée sur l'organisation des expressions. Nous montrons que l'**organisation des expressions, telle que définie, est universelle** et qu'elle peut être efficacement utilisée pour définir de façon unique une expression : une expression est caractérisée par son intensité et sa position relative par rapport aux autres expressions.

La solution est comparée aux méthodes classiques basées sur l'apparence (ICIP 2012) et montre une augmentation significative des résultats de reconnaissance sur 14 expressions non basiques. La méthode a été étendue à des sujets inconnus. L'idée principale est de créer un **espace d'apparence plausible** spécifique à la personne inconnue en **synthétisant ses expressions basiques** à partir de déformations apprises sur d'autres sujets et appliquées sur le neutre du sujet inconnu (CVIU 2013). La solution est aussi mise à l'épreuve dans un environnement multimodal plus complet dont l'objectif est la **reconnaissance d'émotions lors de conversations spontanées**. Les résultats montrent que la solution est efficace sur des données réelles et qu'elle permet l'extraction d'informations essentielles à l'analyse des émotions (ICMI 2012). Notre méthode a été mise en œuvre dans le cadre du **challenge international AVEC 2012 (Audio/Visual Emotion Challenge) où nous avons fini 2nd**, avec des taux de reconnaissance très proches de ceux obtenus par les vainqueurs. La comparaison des deux méthodes (la nôtre et celles des vainqueurs) semble montrer que l'extraction des caractéristiques pertinentes est la clef de tels systèmes (IJACSci 2013).

Mots clefs Analyse des expressions faciales, Représentation invariante, Tessellation de Delaunay, Variété des expressions, Warping linéaire par morceau, Application à la reconnaissance d'émotions, Contexte multimodal, Système d'inférence floue, Contagion d'émotions

Remerciements

Je souhaite remercier en tout premier lieu Patrick Lambert, Professeur à Polytech Annecy-Chambery (LISTIC) et Mohamed Daoudi, Professeur à Télécom Lille (LIFL) d'avoir accepté d'être rapporteurs de cette thèse.

Je remercie également Alice Caplier, Professeur au Gipsa-Lab de Grenoble, et Patrice DALLE, Professeur à l'université Paul Sabatier de Toulouse, d'avoir accepté de juger mon travail en tant que membres du jury.

Ce travail de thèse a vu le jour de façon fortuite, suite à un échange avec Bernard Jouga, délégué à la recherche et aux relations industrielles de Supélec pour le campus de Rennes, qui m'a mis en relation avec Renaud Séguier, Professeur à Supélec et futur directeur de thèse. Je tiens à remercier tout particulièrement ces deux acteurs, le premier pour l'opportunité qu'il m'a offerte et le second pour m'avoir proposé un sujet passionnant, en phase avec mes attentes. Je remercie aussi Renaud pour son dynamisme, ses conseils et son soutien lors de ces trois années. Merci également à Nicolas Stoiber, ancien doctorant de l'équipe, pour ces travaux sur les expressions faciales émotionnelles et pour les échanges fructueux que nous avons pu avoir tout au long de ces trois années. Je souhaite également remercier mes deux autres co-auteurs, Catherine Pelachaud, Directeur de recherches CNRS à TELECOM ParisTech, et Hanan Salam, doctorante à Supélec, avec qui j'ai participé au challenge AVEC 2012 et qui furent deux très belles rencontres.

Je tiens à remercier mes cobayes pour s'être prêté au jeu de reproduire des expressions faciales devant une caméra, expressions qui m'ont parfois donné du fil à retordre : Abel, Christelle, Christophe, Clairette, David, Fred, Gaëlle, Jacques, Nicolas, Ophélie, Pierre, Pierre-Albert, Sylvain, Wassim, Yves et Ziad. J'ai une pensée chaleureuse pour l'ensemble des membres de l'équipe SCEE de Jacques Palicot et pour les membres des autres équipes que j'ai pu côtoyer lors de mon parcours et qui ont fait de ces trois années des moments très agréables. Je les en remercie.

Merci aussi aux personnes qui m'ont envoyé leurs encouragements pour ma soutenance de thèse et à celles qui sont venues assister à ce jour très important pour moi.

Je ne peux terminer ces remerciements sans une pensée pour mes amis et ma famille, tout particulièrement Antoine, mon mari, et ma fille, Louna, qui a vu le jour pendant mes recherches.

Catherine SOLADIE.

Table des matières

Introduction

Sourire jusqu'aux oreilles, avoir les yeux aussi gros que le ventre, faire la bouche en cœur, avoir les dents longues, tendre l'oreille, avoir des yeux de merlan fris, rire de toutes ses dents, faire la bouche en cul de poule, ne pas avoir les yeux en face des trous. Que d'expressions utilisant des traits du visage ! Est-ce bien des expressions (faciales) réalistes ? Non. Et pourtant ce sont des expressions (françaises) très populaires. Certaines *décrivent* l'expression faciale en l'exagérant. D'autres *utilisent* une expression faciale impossible pour indiquer un trait de caractère. En tout état de cause, ces expression indiquent que le visage est vecteur de nombreuses informations, bien au delà de l'expression faciale affichée ; et que le visage, même s'il est hautement déformable, suit certaines lois.

Le visage est vecteur de parole, d'intentions et d'émotions. Tout d'abord, les mouvements bucco-faciaux permettent de s'exprimer correctement. Après un apprentissage lors de la petite enfance, ces mouvements sont devenus des réflexes. Le visage permet aussi de donner volontairement de l'emphase à certains propos ou à certains sentiments et ainsi montrer de plein gré à son interlocuteur ses intentions. Finalement, le visage transmet nos émotions, souvent de façon involontaire.

Ces signes peuvent être analysés par des experts, mais la détection et l'analyse automatique de ces signes restent un domaine émergent, intéressant de nombreuses applications.

C'est dans ce cadre que s'inscrivent les travaux de cette thèse. Nous préciserons tout d'abord, dans cette introduction, le contexte et les motivations en présentant les projets dans lesquels nos recherches s'inscrivent. Ensuite, nous présenterons les enjeux et les contraintes de la mise en œuvre de tels systèmes. Nous ferons alors un focus sur l'acquisition des données, un élément clef du processus. Pour finir, nous présenterons l'organisation de ce document qui s'articule autour des principales contributions.

Sommaire

Contexte et Motivations

Analyse automatique des comportements humains

De nombreuses applications émergentes visent à remplacer l'expert humain dans l'analyse des comportements humains. Nous pouvons identifier trois types d'applications, qui répondent à des besoins différents :

- Les applications dont le but est d'**agir en fonction d'un comportement attendu** d'une personne. Il s'agit par exemple de jeux vidéo, qui adaptent leur contexte de jeu à l'état émotionnel du joueur, ou encore d'application marketing qui visent soit à offrir des produits adaptés aux comportements des clients, soit en amont, à créer des produits plus attractifs. Ce type d'application est basé sur la détection du comportement en cours de l'utilisateur, l'analyse et la classification de ce comportement, puis finalement, l'action en fonction du comportement reconnu. Dans de telles applications, le système cherche à détecter des comportements connus afin d'agir en conséquence.

- Les applications dont le but est d'**agir en fonction d'un changement de comportement** (potentiellement vers un comportement inconnu) d'une personne. Il s'agit par exemple de systèmes d'aide au maintien à domicile des personnes âgées dont le but est de lever une alerte lors d'un changement de comportement ou qu'un comportement non attendu survient ; ou encore des systèmes de surveillance. Ce type d'application est basé sur la détection d'une variation dans le comportement des sujets. Le nouveau comportement est potentiellement un comportement qui n'est pas connu du système ni interprétable par le système.

- Les applications dont le but est d'**évaluer un comportement** (potentiellement un comportement inconnu) d'une personne. Il s'agit essentiellement d'applications à but d'enseignement, basée sur la proposition d'une consigne comportementale que l'utilisateur doit suivre. Ces applications détectent ensuite le comportement effectif de l'utilisateur (comportement qui n'est pas forcément attendu par le système) et évalue ce comportement par rapport à la consigne.

Toutes ces applications sont basées sur l'hypothèse qu'il existe un invariant entre les comportements des différentes personnes, c'est-à-dire que le comportement peut être défini de façon unique quelque soit la personne. Une fois ce comportement défini, l'application l'analyse soit en le classifiant (premier type d'applications), soit en détectant un changement (second type d'applications), soit en évaluant une différence (troisième type d'applications).

Les modalités permettant d'analyser un comportement humain sont vastes, citons par exemple la vidéo, l'audio, l'électromyographie (EMG). Dans cette thèse, nous nous focalisons sur l'analyse des expressions du visage à partir de données vidéo même si pour certaines applications, nous nous sommes penchés sur d'autres aspects du comportement humain (tels que l'empathie) afin d'élargir notre vision et de pouvoir tester la véracité et les limites de notre système dans un cadre global.

Contextes applicatifs

Les travaux présentés dans cette thèse ont été motivés par différents contextes applicatifs, certains réalisés dans le cadre de projets collaboratifs (terminés ou en cours). Ces cas applicatifs sont présentés ci-dessous.

IMMEMO (2010-2013) ANR CONTINT, IMMersion 3D basée sur l'interaction EMOtionnelle, labellisé par le pôle Image & Réseaux.
`http://www.rennes.supelec.fr/immemo/`
Le cas applicatif de ce projet collaboratif fait partie du premier type d'application (agir en fonction d'un comportement attendu d'une personne).
Le contexte est celui du serious game. Il s'agit de créer un environnement virtuel immersif 3D dans lequel est plongé un apprenant. L'environnement virtuel permet de mettre l'apprenant dans des situations particulières afin qu'il s'entraine à manifester les comportements et émotions adaptées à cet environnement. L'objectif plus précis de ce projet est que le formateur puisse manipuler des agents conversationnels émotionnels dans ce monde virtuel immersif 3D. Le formateur ne pouvant pas définir l'ensemble du comportement attendu, celui-ci doit être en parti réalisé de façon automatique. Pour cela, l'idée est de capturer et d'analyser les expressions du visage de la personne immergée afin d'aider le formateur à manipuler le comportement de l'agent conversationnel en lui donnant des commandes simples.
Il s'agit d'un projet industriel ANR, réalisé par le consortium Supélec (porteur du projet), Telecom Paris Tech, ISIR (Institut des Systèmes Intelligents et de Robotique) et Artefacto.
Nos travaux de recherche s'inscrivent directement dans le cadre de ce projet :

- Participation aux réunions projets
- Acquisition de données de visages expressifs (22 expressions, 17 sujets) dans un environnement mono-caméra (RGB) et mise à disposition de cette base de données
- Acquisition de données dans un environnement multi-caméra (caméras infrarouge et caméras RGB) chez Artefacto
- Participation au Challenge AVEC 2012 (voir chapitre 9)

Maintient à domicile des personnes âgées (2013-2015, soumis) Projet PME Bretagne, soumis
Il fait partie du second type d'application (agir en fonction d'un changement de comportement).
Un objectif des programmes gouvernementaux actuels consiste à trouver des moyens permettant aux personnes âgées de rester à leur domicile plus longtemps, plutôt que d'aller en établissement de soins. L'une des préoccupations concernant le maintient à domicile des personnes âgées est la sécurité des personnes. Une solution consiste à lever une alarme lorsque des comportements particuliers se produisent. Parmi les vecteurs permettant de détecter un changement de comportement, les expressions faciales jouent un rôle clef : elles véhiculent les émotions et les humeurs. L'idée est d'avoir un système adapté à la personne permettant d'analyser *ses* expressions. Ce système devra alors être capable de détecter des variations émotionnelles. Il s'agit d'un projet PME Bretagne, soumis par le consortium Neotec-Vision (porteur du projet), Supelec, Dynamixyz, ESC Rennes et INSA Rennes.

REPLICA (2012-2015) ANR Techsan, Rééducation des praxies faciales chez des paralysés cérébraux via un avatar interactif, labellisé par le pôle Image & Réseaux.

Le cas applicatif de ce projet collaboratif fait partie du troisième type d'application (évaluer un comportement).
Le contexte de ce projet est celui de l'aide médicale pour les enfants atteint de paralysie cérébrale. L'objectif est de leur fournir un outil ludique d'entrainement à la parole. Le système visé consiste à pouvoir animer un avatar qui donnera ainsi des consignes aux enfants (principalement, la prononciation de mots), analyser les déformations faciales réalisées par les enfants lors de la reproduction de ces consignes, reproduire ces déformations sur l'avatar afin que les enfants puissent comparer ce qu'ils ont fait aux consignes données.
Il s'agit d'un projet industriel ANR, regroupant HSM (Hôpitaux de Saint-Maurice), l'Université de Rennes 2 - Laboratoire mouvement sport santé (porteur du projet), Dynamixyz et Supélec.
Mes travaux de recherche s'inscrivent là aussi dans le cadre de ce projet :

- Participation aux réunions projets et à la spécification du projet
- Acquisition de données de visages sur les enfants atteint de paralysie cérébrale lors de la prononciation de phonèmes et mots courts, ainsi que d'expressions émotionnelles
- Réunions aux Hôpitaux parisiens Saint Maurice

Les expressions faciales

Parmi les méthodes permettant de caractériser un comportement (qu'il soit effectif ou souhaité ou qu'il s'agisse d'un changement de comportement), les expressions faciales jouent un rôle majeur, dans la mesure où elles sont un vecteur de la parole, de l'emphase des conversations et des émotions [1, 2]. Certaines études ont même montré que les expressions faciales aident la compréhension des conversations [3], [4], voire même sont la principale modalité des communications humaines : en effet, Mehrabian [5] indique que la partie verbale d'un message (c'est-à-dire les mots prononcés) constituent seulement 7% des effets du message, la partie vocale (par exemple les intonations), contribuent à 38% et les expressions faciales et le langage corporel du locuteur ont un impact de 55% sur les effets du message verbal.

L'analyse automatique des expressions faciales reste actuellement un challenge. La principale difficulté est liée au fait que les expressions faciales sont le *résultat* de la déformation du visage par les muscles. L'enjeu majeur pour caractériser une expression faciale consiste alors à identifier de façon unique une expression à partir des informations visibles, qui mixent l'information utile (expression) à d'autres informations (identité, éclairage, ...).

Enoncé du problème

Contraintes

Les systèmes décrits précédemment doivent se plier à un certains nombre de contraintes. Pour commencer, afin d'être acceptés par les utilisateurs, les systèmes doivent être non intrusifs, faciles à utiliser et à bas coût. C'est pourquoi nous avons opté pour l'utilisation de caméras RGB (voir section 1.1).

Par ailleurs, les contraintes suivantes doivent être respectées :

- **Précision** de façon à distinguer des expressions proches
- **Exhaustivité** de façon à distinguer des expressions non connues
- **Robustesse** pour gérer les différentes morphologies
- **Flexibilité** pour s'adapter aux différents individus sans phase préalable d'apprentissage

L'acquisition des données

La disponibilité de données sur lesquelles entraîner et tester les algorithmes est un point transverse important. Nous y reviendrons à plusieurs reprises dans l'état de l'art (partie I). Il s'agit à la fois d'avoir des données, c'est-à-dire dans le cas qui nous intéresse des visages, mais aussi de pouvoir avoir une vérité terrain avec laquelle se comparer.

Nous abordons dans ce paragraphe quelques unes des notions relatives à ce sujet. Nous proposons tout d'abord une réflexion sur l'influence de la façon de *capter* les visages : faut-il des expressions spontanées ou bien des expressions posées sont-elles intéressantes ? Nous présenterons ensuite quelques bases de données existantes. Pour finir, nous expliciterons les problématiques liées à la labellisation de la vérité terrain.

Expression Posée ou Spontanée D'un premier abord, il semble plus intéressant de travailler sur des données spontanées, dans la mesure où elles représentent mieux la réalité, dans sa diversité notamment ; et que les systèmes finaux travailleront sur ce type de données. De plus, certaines modifications unitaires du visage sont difficilement faisables de façon volontaire. C'est par exemple le cas de la déformation liée à la remontée des joues (Action Unit 6 - *Cheek Raiser and Lid Compressor*) qui est difficilement faisable sans une tension au niveau de la paupière (AU7 - *Lid Tightener*). Les muscles s'activent simultanément (voir la description du système FACS [6] dans la section 2.1.1). Néanmoins, les données spontanées possèdent deux inconvénients majeurs. Tout d'abord, les vérités terrains ne sont pas toujours faciles à obtenir avec des données issues de réactions humaines spontanées. Ensuite, l'acquisition de telles données n'est pas aisée d'un point de vue *mise en scène*. En effet, les expressions spontanées nécessitent de limiter au maximum l'intrusion d'outils permettant de *capter* des informations, l'objectif étant que le sujet soit dans un environnement le moins *expérimental* possible. Il est donc intéressant de prendre en considération ces 2 points lorsque l'on souhaite établir une base de données ou utiliser une base existante pour entrainer et/ou tester un système.

Les Bases de Données Les bases de données sont classées en fonction du mode de représentation souhaité. Nous expliciterons les différents modes de représentations (AUs, expressions prototypiques ou *de base*, dimensions représentant l'émotion) dans la partie I.

La base Cohn-Kanade [7] est certainement la base de données la plus répandue pour la reconnaissance des Actions Units (AUs). La base MMI [8] propose des expressions dont les vérités

terrains sont sous la forme d'une unique Action Unit (AU) ou d'une combinaison minimale d'AUs ou encore d'une expression prototypique (6 expressions de base : joie, surprise, peur, colère, tristesse et dégoût). Les bases de données Bosphorus [9] et GEMEP-FERA [10] proposent elles aussi des visages étiquetés soit avec un des 6 labels émotionnels de base, soit avec une combinaison d'AUs. Ces bases de données sont utilisées dans le cadre de l'analyse des expressions.

Pour ce qui est des bases de données possédant une vérité terrain sous la forme de dimensions émotionnelles, nous pouvons citer Sensitive Artificial Listener Database (SAL-DB) [11, 12] qui propose une labellisation sous la forme valence - arousal. La base SEMAINE [13], dont un sous ensemble a été utilisé pour les challenges AVEC 2011 et 2012, a quant à elle été annotée sous la forme de nombreuses dimensions représentant l'émotion, incluant arousal, valence, power et expectation.

La Vérité Terrain La vérité terrain est une question problématique. En effet, elle nécessite que les labels soient très précisément définis. C'est l'une des raisons du succès des Actions Units et du système FACS [6]. Une formation et une évaluation doivent être réalisées pour devenir un codeur FACS certifié.

Dès lors qu'il s'agit d'émotion, la vérité terrain est plus compliquée à obtenir. Certains systèmes utilisent l'émotion ressentie par la personne. Dans ce cas, c'est le sujet qui qualifie son émotion ou qui réalise l'émotion selon une consigne. Dans d'autres cas, la vérité terrain est issue de l'évaluation humaine. C'est alors un ensemble d'annotateurs qui qualifient l'émotion vraisemblablement ressentie par le sujet. Dans ce cas de figure, l'empathie des annotateurs permet de qualifier l'émotion du sujet. Hupont et al. [14] ont montré que le label de l'émotion ressentie par la personne et celui perçu par d'autres personnes ne correspondent pas toujours. Selon Bassili [15], un observateur entraîné classifie correctement une émotion faciale avec un taux de 87%. Par ailleurs, les études interculturelles ont montré que le jugement porté sur une expression faciale dépendait de la culture du sujet étudié [16, 17].

L'ensemble de ces analyses montrent la difficulté à obtenir une vérité terrain fiable sur laquelle s'appuyer pour tester les systèmes.

Organisation de la Thèse

La thèse est organisée en 9 chapitres regroupés en 3 parties de 3 chapitres chacune.

Partie I

La première partie (Partie I) présente un état de l'art des techniques d'analyse des expressions. Nous y exposons notre vision de l'analyse des visages et les différentes représentations précédemment étudiées. De notre point de vue, trois niveaux d'analyse sont à prendre en compte :

- La **description d'un visage**, c'est-à-dire la représentation d'une image représentant un visage et récupérée par le système de capture (caméra RGB dans notre cas).
- La **description d'une expression**, c'est-à-dire la caractérisation unique de chaque expression, quelque soit le sujet.
- L'**interprétation d'une expression**, c'est-à-dire la signification de l'expression, comme par exemple l'émotion du sujet.

Partie II

Concernant la description d'un visage, nous utilisons des données issues de modèles actifs d'apparences [18]. Ces travaux ne font pas l'objet de cette thèse. Les principales contributions concernent le second niveau de description : la représentation d'une expression. Les travaux sont présentés dans la seconde partie de cette thèse (Partie II) et s'articulent autour de trois originalités répondant aux contraintes énoncées précédemment :

- Exhaustivité : une expression est définie par sa **position relative par rapport aux autres expressions** (chapitre 5.1)
- Robustesse (identité vs. expression) : l'organisation des expressions les unes par rapport aux autres est **indépendante de la personne** (chapitre 4)
- Précision et Flexibilité : les modèles utilisés sont spécifiques à la personne et créés à partir des **déformations plausibles du visage d'une personne inconnue** (chapitre 5.3)

La pertinence du modèle proposé est testée et comparée à des méthodes existantes en fin de partie (chapitre 6).

Partie III

Pour finir, nous avons testé et amélioré notre modèle de représentation dans un environnement plus complet, c'est-à-dire sur des expressions spontanées, réalisées dans un contexte de conversation entre un sujet et un agent émotionnel. Nous avons ainsi élargi nos travaux à l'interprétation en terme d'émotion des expressions faciales et introduit d'autres modes de caractérisation de l'émotion (informations vocales, contexte de la conversation). Ces travaux font l'objet de la dernière partie de cette thèse (Partie III).

Première partie

Analyse des Expressions Faciales : État de l'Art

Faut-il peindre ce qu'il y a sur un visage ? Ce qu'il y a dans un visage ? Ou ce qui se cache derrière un visage ?

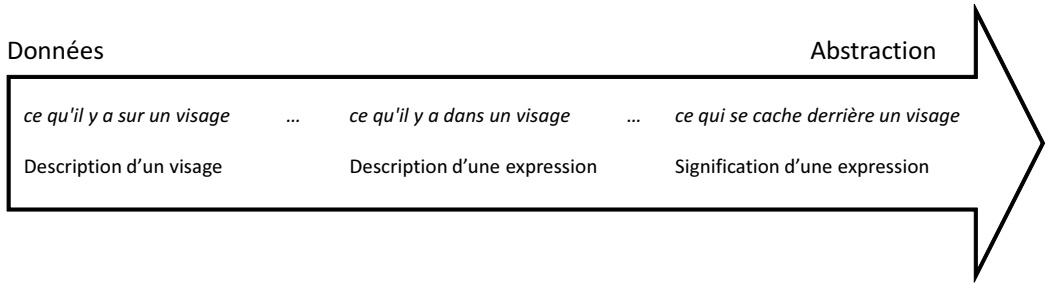
Pablo Picasso

Comme l'indique Picasso, lorsque l'on voit un visage, plusieurs niveaux de description sont possibles, allant du plus près des données au plus abstrait. Nous allons voir comment cette distinction en trois niveaux est aussi applicable à la vision par ordinateur.

Nous pouvons tout d'abord décrire *ce qu'il y a sur un visage*, de façon à représenter le plus fidèlement possible ses traits, ses couleurs, ses aspérités. Tous les aspects du visage sont pris en compte, aussi bien les caractéristiques liées à l'identité de la personne que celles représentant sa mimique. C'est le premier niveau de description, que nous appellerons **description d'un visage**.

Nous pouvons ensuite décrire *ce qu'il y a dans ce visage*. Il s'agit alors d'extraire certaines caractéristiques, telles que la morphologie ou encore l'expression. C'est ce qui est fait dans une caricature où le visage n'est pas représenté fidèlement mais les traits représentant l'identité de la personne sont accentués. Pour notre part, nous nous intéresserons aux caractéristiques de l'expression de la personne (sourire, haussement de sourcils, ...). C'est le second niveau de description, que nous appellerons **description d'une expression**.

Nous pouvons pour finir décrire *ce qui se cache derrière ce visage*, de façon à interpréter et à donner une signification au visage expressif. Cela peut permettre d'identifier la personne par son nom ou encore indiquer l'émotion affichée par la personne (par exemple : *elle est joyeuse*). C'est le troisième niveau de description, que nous appellerons **signification d'une expression**.



La principale contribution de cette thèse concerne le niveau intermédiaire : la description d'une expression faciale. Le système s'appuie sur la description d'un visage (premier niveau). Nous effectuerons donc dans un premier temps un bref tour d'horizon des méthodes de description d'un visage afin de positionner celles utilisées dans ce document et de préciser le vocabulaire utilisé par la suite (chapitre 1). Nous effectuerons ensuite un état de l'art sur les méthodes de description d'une expression faciale afin de positionner notre méthode parmi les techniques existantes (chapitre 2). Pour finir, pour tester la pertinence du modèle proposé, nous avons aussi réalisé des expérimentations relatives à la signification d'une expression. Nous donnerons donc une vue d'ensemble de ce thème dans un chapitre dédié (chapitre 3).

La figure 1 donne une vision synthétique du découpage en 3 étapes tel que nous le proposons. Pour des facilités de navigation dans le document, les numéros de section sont indiqués.

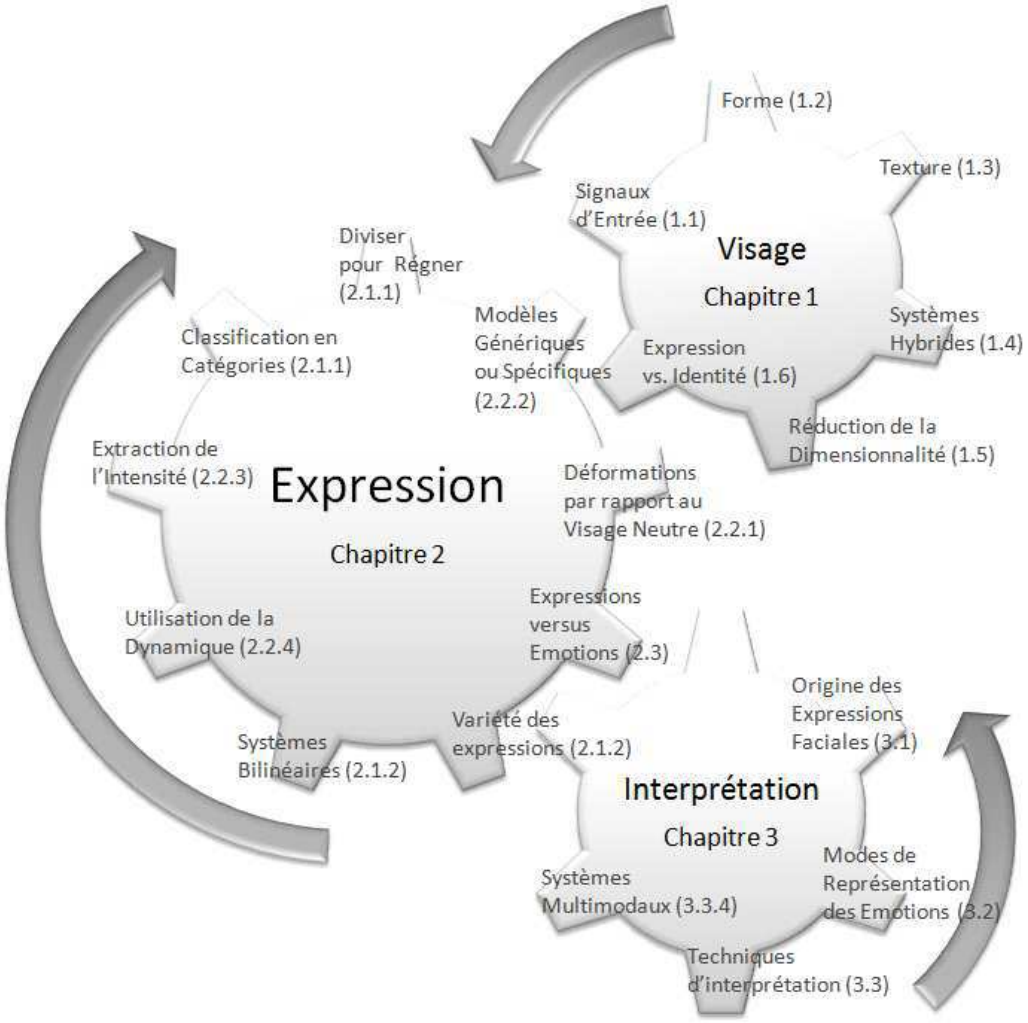


FIGURE 1 – Vision synthétique des 3 étapes d'analyse d'une expression faciale.

Chapitre 1

Description d'un Visage

La *description d'un visage* s'attache à le représenter le plus fidèlement possible. Plusieurs caractéristiques sont utilisées pour le décrire : sa forme et celle de ses composantes (forme des yeux, de la bouche, de la mâchoire), la texture du visage (couleur des yeux, de la peau, présence de fossettes, de rides). Nous les expliciterons dans cette section.

Cette thèse ne propose pas de contribution dans ce niveau de description. Néanmoins, nous utiliserons certaines de ces méthodes en pré requis pour décrire et analyser les expressions faciales. Elles sont présentées dans ce chapitre. Par ailleurs, nous introduisons aussi dans la dernière section la notion d'expression (qui fera l'objet du chapitre suivant) afin de mettre en évidence les difficultés de la description objective d'une expression.

Sommaire

1.1 Les signaux d'entrée

Les données permettant de décrire un visage sont obtenues à partir de capteurs vidéo (caméras) permettant de fournir différents types d'information.

Les capteurs non intrusifs Il est courant de distinguer les capteurs selon qu'ils sont intrusifs ou non. Les capteurs intrusifs sont des capteurs dont au moins une partie est placée sur ou dans le corps humain (ici le visage). Il s'agit par exemple de placer des marqueurs sur le visage des sujets et d'utiliser une caméra infrarouge pour détecter la position de ces marqueurs (et donc du visage du sujet). Dans nos cas applicatifs, nous souhaitons limiter l'impact sur les sujets afin d'augmenter l'acceptabilité du système. C'est pourquoi nous utilisons des capteurs non intrusifs. Les caméras RGB sont une solution lorsque ce type de contraintes est présent. C'est ce que nous utilisons dans cette thèse.

Information 2D ou 3D Une caméra RGB permet de fournir une image en 2D de la scène. Afin d'obtenir une information en 3D de la scène, plusieurs techniques existent. La stéréoscopie (mise en place de 2 ou plus caméras RGB) permet de reconstruire une image 3D à partir d'images 2D ayant des points de vue différents. Plus récemment, les caméras produisant une carte de profondeur sont disponibles à bon marché et possèdent une précision croissante. Elles permettent de fournir une information sur la profondeur de la scène en plus de l'image 2D. Elles peuvent être couplées avec une caméra RGB qui donne les informations de texture (c'est le cas de la Kinect par exemple). Dans nos expérimentations, nous utilisons des informations 2D issues d'une caméra RGB.

1.2 Extraction de la Forme du Visage

La *forme* du visage décrit l'emplacement des principales composantes du visage que sont le front, les sourcils, les yeux, le nez, la mâchoire et la bouche. Cette forme peut être donnée par la localisation de points caractéristiques (en 2 ou 3 dimensions), par des maillages ou encore par des formes géométriques simples.

Points caractéristiques Un visage peut être décrit par un vecteur formé par les coordonnées 2D ou 3D d'un certain nombre de points caractéristiques (par exemple le coin gauche des lèvres).

Dans la figure 1.1, le visage est annoté par 73 points caractéristiques en 2D. Le vecteur résultant possède donc 73*2 composantes.

Des procédures mathématiques permettent d'extraire les principales déformations et de réduire la dimensionnalité de ces vecteurs. Pour ce faire, plusieurs visages sont annotés et les procédures sont appliqués sur l'ensemble des vecteurs associés. Parmi ces méthodes, nous pouvons citer l'**analyse en composantes principales (ACP)** [19] qui permet de transformer des variables corrélées en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées *composantes principales*. Les axes correspondants sont orthogonaux et correspondent aux principales déformations allant de la plus grande à la plus petite. Cette technique a souvent été utilisée pour décrire des visages. Néanmoins, elle est très dépendante des données d'apprentissage et les déformations obtenues ne possèdent pas d'interprétation physique évidente. Une autre technique est l'**analyse en composantes indépendantes (ACI)** [20] qui décompose les données en variables statistiquement indépendantes.

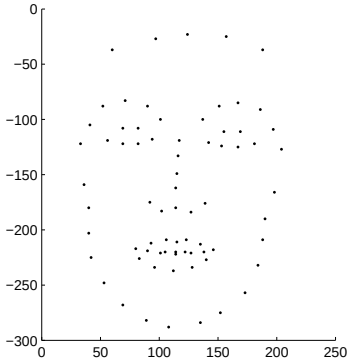


FIGURE 1.1 – Description de la forme d'un visage par les coordonnées 2D de 73 points caractéristiques.

Les **modèles de forme actifs (ASM)** [21], basés sur l'ACP, permettent de retrouver automatiquement les points caractéristiques du visage. Le modèle est appris sur un ensemble de visages annotés manuellement. Le principe de l'ASM est de faire correspondre le modèle de forme à une nouvelle image, en trouvant la transformation (recalage global puis local) permettant d'optimiser la correspondance entre le modèle et la nouvelle image [22].

Formes géométriques simples Il s'agit ici de définir la forme du visage par des équations simples, correspondant à des formes simples. Ces techniques sont appliquées pour la description et la détection de certaines parties du visage, notamment l'iris, modélisé par une ellipse ou un cercle. Pour détecter l'iris sur un nouveau visage, le modèle de l'iris créé est appliqué sur la nouvelle image et les paramètres du modèle sont modifiés de façon à correspondre au mieux au modèle [23, 24, 25]. De telles méthodes sont efficaces lorsque les yeux sont ouverts mais la fermeture et le clignement des yeux posent souvent problème. De plus, ces techniques nécessitent une résolution de haut niveau pour fournir de bons résultats. Et à part quelques exceptions [26], elles sont rarement appliquées à des formes complexes, ne pouvant pas prendre en compte la complexité des déformations du visage.

Maillages et courbures Avec l'arrivée récente sur le marché des caméras de profondeur à bas coût et la définition 3D des visages, les techniques de description de surfaces 3D sont en plein essor.

Une première méthode consiste à définir la forme du visage par un **maillage**, c'est-à-dire que la surface du visage est approximée par des facettes, la plupart du temps des triangles [27]. La figure 1.2 propose une représentation sous la forme de maillage à facettes triangulaires. Une représentation classique d'un maillage consiste à avoir deux vecteurs, le premier contenant les coordonnées 3D de très nombreux points du visage et le second contenant la liste des triangles (c'est-à-dire des points permettant de définir une facette). Contrairement aux *points caractéristiques*, les maillages utilisent un grand nombre de points n'ayant pas de localisation physique particulière.

Une autre méthode permettant d'approximer la surface consiste à définir la **courbure** du visage en de nombreux points [28]. La courbure en un point du visage peut être définie par le Laplacien des coordonnées des points qui sont dans son voisinage. Les coordonnées ainsi obte-

nues forment un vecteur 3D en chacun des points. La direction du vecteur donne l'inclinaison de la surface du visage et la norme du vecteur donne la force de la courbure.

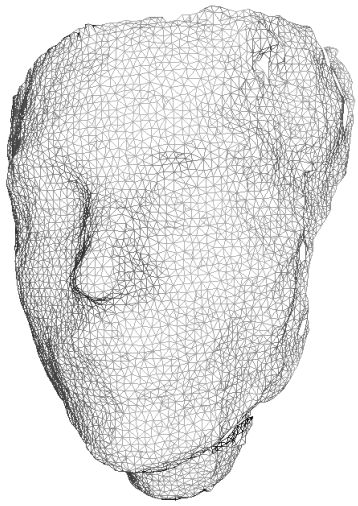


FIGURE 1.2 – Représentation d'un visage sous la forme d'un maillage à facettes triangulaires.

1.3 Extraction de la Texture du Visage

La *texture* peut tout d'abord être décrite par le niveau de gris de l'ensemble des pixels de l'image contenant le visage. Afin de limiter l'information relative à l'arrière plan, le plus souvent l'image est préalablement recentrée sur le visage et rognée de façon à ce que les bords du visage touchent (ou presque) les bords de l'image. Parfois, la forme du visage (contour) est utilisée et seules les informations de texture à l'intérieur de cette forme sont prises en compte. Ces informations sont représentées dans un vecteur dit vecteur de texture. Comme dans le cas des points caractéristiques, des transformations (ACP, ICA) peuvent être appliquées pour extraire les principales déformations et réduire la dimensionnalité de ces vecteurs.

La texture peut aussi être codée par l'application de filtres de Gabor sur les valeurs des pixels. C'est le cas du système proposé par Lyons et al. [29] qui code les visages par l'utilisation d'un ensemble de filtres de Gabor multi-orientés et multi-résolutions. Cette représentation a pour avantage d'avoir des similitudes avec les annotations sémantiques d'observateurs humains.

Une autre façon de décrire un visage consiste à définir un certain nombre de motifs et d'en extraire leur répartition dans l'image. C'est le cas de la méthode LBP [30]. Cette méthode consiste à associer un motif à chaque pixel, selon la nature des pixels environnants. Un histogramme des motifs est alors créé pour une zone de l'image, indiquant la répartition de ces motifs dans cette zone. Ces informations permettent de caractériser les différentes zones du visage et de décrire ainsi les données visuelles de l'image. Des contributions plus récentes continuent d'améliorer cette technique largement employée [31]. Ces méthodes ne permettent pas de reconstruire l'image du visage mais fournissent une signature intéressante permettant d'analyser des visages.

1.4 Systèmes Hybrides

Sur un visage, la forme et la texture sont corrélées dans la mesure où la couleur du visage dépend des différentes parties de ce visage. Par exemple, les lèvres ont une forme particulière et dans cette forme, la couleur est spécifique. De même, les ombres changent (modification de la texture) lorsqu'une personne réalise un sourire (modification de la forme). Certains systèmes proposent de coupler les informations de forme et de texture afin de pouvoir à la fois décrire le visage de façon succincte et plus précise mais aussi de pouvoir retrouver automatiquement la forme d'un visage inconnu.

C'est le principe utilisé par les modèles flexibles d'apparence [32] qui ont abouti aux **modèles actifs d'apparence (AAM)** [18]. Ceux-ci proposent de réaliser sur des visages d'une base d'entrainement, une première ACP sur la forme, une seconde sur la texture et finalement une ACP sur le vecteur concaténant les vecteurs de forme et de texture obtenus. Le modèle ainsi créé permet de retrouver automatiquement la forme d'un nouveau visage. Pour un nouveau visage, la forme est obtenue en faisant varier le vecteur d'apparence (forme+texture) de façon à ce qu'il corresponde au mieux au modèle.

1.5 Réduction de la Dimensionnalité

Deux principaux problèmes se posent avec les données *brutes* de forme et de texture. Le premier concerne le nombre de données qui peut très vite être très important lorsque l'on augmente la résolution. Le second concerne la pertinence des informations. Plusieurs techniques ont été utilisées pour réduire la dimensionnalité des vecteurs de description d'un visage tout en gardant les informations pertinentes.

Les techniques telles que l'ACP [19] ou l'ACI [20] répondent à ce besoin. Elles permettent de réduire la dimensionnalité en indiquant les principales déformations par rapport à un visage moyen. On parle alors de vecteurs de forme ou de texture du visage vu qu'il ne s'agit plus directement de la forme (resp. de texture) du visage qui est décrite mais du poids des principales déformations de cette forme (resp. de cette texture) par rapport à un visage moyen. Des techniques de réductions non linéaires peuvent aussi être utilisées pour diminuer la dimensionnalité de l'espace (par exemple Isomap [33] ou LLE [34]). Ces techniques donnent de bonnes approximations de l'espace sur des structures spécifiques construites sur mesure (données artificielles) mais sont moins pertinentes que l'ACP sur des données réelles [35].

Une autre technique permettant de diminuer la dimensionnalité de l'espace consiste à sélectionner les composantes par corrélation [14]. La corrélation est effectuée à la fois entre chaque composante et l'information de plus haut niveau souhaitée, mais aussi entre composantes (intercorrélation) afin de détecter les caractéristiques redondantes. Seules les composantes ayant une forte corrélation avec l'information de plus haut niveau tout en ayant une faible intercorrélation sont gardées [36].

1.6 Expression versus Identité

L'une des principales difficultés en analyse des expressions du visage est de s'affranchir de l'identité du sujet, c'est-à-dire de sa morphologie en ce qui concerne la forme et la texture de son visage. Il est communément admis que les expressions sont similaires entre les personnes [37] et que l'identité est spécifique à la personne.

Les descriptions de visages présentées précédemment souffrent de ce mélange : les vecteurs obtenus possèdent des informations regroupant à la fois les informations de l'expression effectuée par la personne mais aussi de son identité. Par exemple, la forme d'un visage indique si la personne est ou pas souriante, hausse, fronce les sourcils (expression), mais indique aussi si le visage est rond ou plutôt ovale, si les yeux sont grands, bridés (identité). De même, les informations de texture indiquent les fossettes, le froncement des sourcils (expression), mais indiquent aussi la couleur de la peau, des yeux, si la personne possède une barbe ou pas (identité).

Ainsi les données de visage sont difficilement comparables entre des personnes différentes. A titre d'exemple, la figure ci-dessous affiche les données de visages expressifs de 2 sujets différents. Plus précisément, le principe des modèles actifs d'apparence est utilisé pour décrire les visages (voir annexe A). Chaque point correspond à un visage expressif. Les trois premières composantes, donc les trois principales déformations, sont retenues pour la description des visages. Les visages expressifs d'une première personne sont en rouge, et ceux d'une seconde personne sont en noir.

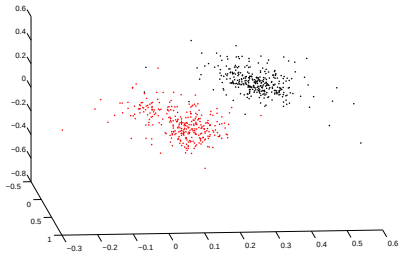


FIGURE 1.3 – Nuages d'expressions faciales de 2 sujets. Affichage pour chaque expression de chaque sujet des trois premières composantes des vecteurs d'apparence obtenus par un AAM.

Nous constatons deux nuages de points, chacun correspondant à un des deux sujets. Cela signifie que des expressions identiques ont des vecteurs d'apparence différents. Les données de visage ne sont pas directement exploitables pour définir une expression et par la suite pouvoir l'interpréter. C'est pourquoi nous proposons d'utiliser une étape intermédiaire permettant de donner une signature unique à une expression quelque soit le sujet. C'est l'objet du chapitre suivant.

Chapitre 2

Description d'une Expression Faciale

Les expressions du visage sont principalement générées par la contraction des muscles qui induisent des modifications temporaires des caractéristiques de forme du visage telles que le clignement des paupières, le haussement des sourcils, la forme de la bouche ou encore des modifications de la texture de la peau telle que l'apparition de rides ou de fossettes. Les changements sont souvent brefs, de l'ordre de quelques secondes (entre 250 ms et 5 secondes).

Les informations importantes permettant de caractériser une expression résident dans la localisation des déformations, dans l'intensité de ces déformations ainsi que dans la dynamique de ces déformations. La principale problématique, lorsqu'il s'agit d'analyser automatiquement une expression, est de s'affranchir des différences de morphologie entre les personnes. Nous parlerons ici de morphologie au sens large, c'est-à-dire les différences de forme et de texture entre les sujets qui sont caractéristiques de leur identité.

Ce chapitre décrit tout d'abord les méthodes permettant de représenter une expression faciale, en proposant une découpe selon le mode de représentation choisi (discret ou continu). Il propose ensuite d'aborder différents thèmes autour de la spécificité des individus et de leurs caractéristiques communes. L'analyse des expressions faciales ne doit pas être confondue avec l'analyse des émotions, comme cela est très souvent le cas. Nous reviendrons sur ce sujet en fin de chapitre.

Sommaire

Il est à noter que nous effectuons une simplification en indiquant que les caractéristiques du visage ne contiennent que des informations de l'identité du sujet et des informations de l'expression du sujet. D'autres facteurs viennent *polluer* ces données. Parmi ces facteurs, citons l'illumination. Les variations d'illumination entre différentes scènes peuvent grandement impacter les données de texture. De même, les rotations du visage modifient l'illumination de celui-ci et perturbe la forme perçue du visage et de ses composantes. L'emplacement de la caméra (trop près) peut aussi déformer le visage et impacter la forme extraite. Nous ne traiterons pas de ces aspects dans cette thèse.

2.1 Représentations Discrètes versus Représentations Continues

De très nombreuses méthodes ont été proposées pour décrire une expression faciale. Cette section propose de les présenter selon leur mode de représentation.

2.1.1 Représentations Discrètes des Expressions

Méthodes Diviser pour Régner

Les intellectuels démontent le visage, pour l'expliquer en morceaux, mais ils ne voient plus le sourire.

Antoine de Saint-Exupéry - *Pilote de guerre - 1942*

Une description permettant de définir de façon unique et intelligible une expression s'est largement répandue, principalement sous l'influence des travaux d'Ekman et son utilisation importante en psychologie. Il s'agit du système FACS [6]. L'idée de base est de s'inspirer des connaissances de l'anatomie du visage et des principaux muscles entrant en action lorsque les expressions sont réalisées. Les déformations visuelles correspondantes sont alors identifiées et définissent des Unités d'Actions (AUs). Le système FACS compte 44 unités d'action (Actions Units - AUs) pour la description des actions faciales. A certaines AUs est associée une intensité ayant de 3 à 5 niveaux de magnitude. Dans le système FACS, une expression est caractérisée par une combinaison des AUs. Par exemple, une expression de surprise est définie par AU1 (haussement du sourcil intérieur) + AU2 (haussement du sourcil extérieur) + AU5 (haussement de la lèvre supérieure) + AU25 (écartement des lèvres).

La présence simultanée de plusieurs AUs est un point problématique. Même si les muscles ne peuvent pas tous se contracter de manière indépendante (par exemple le Levator Labii Superioris, qui tire la lèvre supérieure en dehors de la lèvre inférieure et l'Orbicularis Oris, qui comprime les lèvres, ne peuvent se contracter ensemble), il existe d'éventuels mouvements conflictuels en terme d'AUs. Ils interviennent, pour la plupart d'entre eux, dans la région de la bouche. Le système FACS les décrit. Pour traiter cette question dans leur système d'animation faciale, Wojdel et al. [38] ont proposé une approche utilisant la logique floue pour définir les dépendances entre les AUs.

Les personnes réalisant de la synthèse d'image sont bien conscientes qu'une expression est bien plus complexe qu'une simple somme d'unités d'action, notamment lorsque plusieurs canaux d'expression sont présents simultanément (par exemple, émotion et parole). Bui et al. [2] proposent par exemple un système en deux couches. La première couche permet de réaliser des transitions lisses pour chaque canal d'expression. Six canaux d'expressions sont définis (signaux physiologiques, visèmes, interactions sociales, émotions, direction du regard et position de la tête). La seconde couche permet ensuite la combinaison de mouvements de ces différents canaux

par la résolution de conflits entre les muscles. En effet, ces différents canaux peuvent porter sur des muscles identiques (par exemple, certains muscles des lèvres sont sollicités à la fois pour le sourire et pour la parole).

Néanmoins, le présence simultanée de plusieurs AUs reste un point problématique en analyse. La méthode FACS est basée sur de nombreuses unités d'actions séparées, ce qui fait que la corrélation qui existe entre les différentes déformations unitaires lors des expressions du visage est ignorée. Liu and Wu [39] ont montré que, avec leur méthode, pour la détection du sourire de tromperie, prendre en compte les AU6 et AU12 simultanément lors de l'apprentissage du système donnait de meilleures performances que prendre en compte les AU6 et AU12 séparément. Cela semble indiquer que même si la méthode est pertinente et très utilisée dans le domaine de la psychologie, elle ne semble pas adaptée à l'analyse automatique.

L'utilisation des AUs afin d'extraire une information de plus haut niveau (voir chapitre 3) est aussi un point de questionnement. La plupart du temps, les systèmes qui détectent les AUs possèdent deux étapes : la détection des mouvements faciaux et la classification en AUs. La question des expressions non prototypiques est rarement abordée, considérant que la classification du système FACS suffit. Néanmoins, dans la vie de tous les jours, de très nombreuses expressions sont réalisées. La combinatoire des 44 mouvements faciaux donne un nombre important d'expressions possibles dont beaucoup ne sont pas décrites dans le manuel FACS. Pour adresser cette question, Pantic et Rothkrantz [40] ont proposé un système en 3 étapes basé sur les AUs. La première étape est une détection hybride des caractéristiques faciales, la seconde concerne la détection des AUs et la troisième est un système basé sur des règles pour la détection d'expressions émotionnelles basiques. Ils ont postulé que chaque expression faciale émotionnelle non prototypique pouvait être classifiée dans l'une des 6 émotions de base.

En synthèse, la détection des AUs reste un défi comme le montre les résultats sur challenge FERA 2011 [41]. Même si les résultats sont encourageants, les taux de reconnaissance restent bas : la détection des AUs atteint seulement 62%. On peut alors se demander si cette représentation, certe pratique et répandue en psychologie, est adaptée à la vision par ordinateur.

Une autre description, également répandue, est le standard de compression MPEG-4 [42]. Le standard définit 66 paramètres d'animation faciale (FAPs) de bas niveau, issus de l'étude des actions faciales minimales, assez proche des actions des muscles. Les FAPs sont définis en fonction de deux caractéristiques : un ensemble de points clef permettant de définir la forme du visage (FPs) ainsi que des unités (Face Animation Parameter Units - FAPUs) permettant de normaliser les déformations entre les sujets. L'information d'expression est alors contenue dans un ensemble de distances et angles normalisés entre les points caractéristiques. Par exemple le FAP numéro 3 (mâchoire ouverte) correspond au déplacement de FP2.1 (bas du menton) vers le bas exprimé en FAPU MNS (séparation entre la bouche et le nez). Ainsi un FAP numéro 3 à 0.5 indique que le menton a bougé vers le bas par rapport au visage neutre d'une distance équivalente à la moitié de la distance séparant la bouche du nez. Le logiciel faceAPI [43] permet d'extraire ce type d'information.

Ce système est défini essentiellement pour compresser des données. C'est pourquoi, certains systèmes utilisent cette notation et ces informations de distances et angles afin de réduire la dimensionnalité de l'espace [14]. Même si le visage neutre est pris en compte, on peut se demander si la représentation est suffisamment précise pour l'analyse comparative des expressions mélangées entre les sujets.

Classification en Catégories Une autre façon de décrire une expression faciale consiste à lui donner un label. Cette catégorisation a eu un franc succès lorsque la communauté d'analyse d'image s'est emparé des travaux d'Ekman indiquant l'universalité des expressions faciales correspondant aux 6 émotions de base (joie, sourire, peur, dégoût, tristesse et colère). Le label utilisé est d'ailleurs souvent l'émotion correspondante (joie) et non la déformation réelle observée (sourire, plissement des yeux). Dans ce type de représentation, la caractérisation d'une expression est sa sémantique et non la déformation réelle observée.

L'analyse automatique des expressions faciales, basée sur cette classification, a commencé dans les années 90 [44]. De très nombreux systèmes ont proposé des classifieurs toujours plus évolués permettant de fournir un label comme le montre les études [45, 1, 46].

Cette représentation des expressions a atteint de très bons taux de reconnaissance sur des personnes connues (sujet présents dans la base d'apprentissage) comme le montre le challenge récent FERA 2011 [41] avec les meilleurs scores atteignant 100% pour la classification en 5 catégories. Néanmoins, les taux de reconnaissance chutent fortement dès lors que le sujet n'est pas présent dans la base d'apprentissage (meilleur score à 75.2% - moyenne sur les 5 catégories). Ces résultats montrent que la généralisation à différentes morphologies n'est pas bien réalisée par les classifieurs.

Une autre difficulté principale est liée à la représentation par elle-même et à la labellisation de la vérité terrain. De nombreuses déformations différentes (notamment d'intensité variable) se retrouvent dans la même catégorie. C'est pourquoi cette représentation est peu à peu laissée de côté pour faire place à des représentations continues (voir section 2.1.2), que ce soit en terme d'expression ou en terme d'émotion.

A noter que les systèmes proposés nécessitent une grande base d'apprentissage. Cela est vraisemblablement dû aux deux points mentionnés ci-dessus : les systèmes doivent apprendre les différentes morphologies des sujets et les différentes façons de réaliser une expression.

2.1.2 Représentations Continues des Expressions

Systèmes Bilinéaires Une autre façon d'extraire l'information utile concernant l'expression est de découpler l'expression de l'identité via des modèles bilinéaires. L'avantage des méthodes bilinéaires est qu'elles nécessitent peu de données d'entrainement comparé aux méthodes définissant des variétés (voir paragraphe suivant). Wang et Ahuja [47] ont utilisé HOSVD (High Order Singular Value Description) pour décomposer les caractéristiques d'apparences similaires aux vecteurs AAM en un sous espace des identités et un sous espace des expressions. Ils ont utilisé le modèle résultant pour synthétiser des expressions faciales d'un nouveau sujet et pour reconnaitre simultanément l'identité et l'expression. Abboud et Davoine [48] ont utilisé les modèles bilinéaires symétriques et asymétriques pour réaliser à la fois de la reconnaissance et de la synthèse des expressions. Mpiperis et al. [49] ont découplé le visage et l'expression faciale par des modèles bilinéaires appliqués sur des modèles élastiquement déformables. Ils ont utilisé le modèle créé pour effectuer la reconnaissance simultanée de l'identité et de l'expression. Même si la représentation est continue, dans chacune de ces méthodes, les tests de reconnaissance ont été réalisés sur les 6 expressions de base et le visage neutre. La façon dont le système se comporte avec des expressions plus complexes n'a pas été menée.

Variétés des Expressions Il est de plus en plus courant de définir l'espace des expressions faciales comme une variété de dimension plus faible que celle des composantes définissant les visages. Dans les années 2000, nous avons vu naître des méthodes utilisant l'apprentissage de

variétés pour la représentation des expressions faciales. Les caractéristiques du visage sont projetées sur un espace de dimension plus petite. De telles variétés peuvent prendre en compte le caractère continu et mélangé des expressions ainsi que la notion d'intensité. Stoiber et al. [50] ont mappé les vecteurs issus des modèles actifs d'apparence (AAM) sur un disque. Ce mapping est réalisé de façon non supervisée en trouvant les directions dominantes de l'espace d'apparence d'origine et en les organisant sur un disque. Cette organisation est réalisée en minimisant l'angle entre les directions dominantes de façon à garder la proximité des expressions proches dans l'espace d'apparence. L'espace résultant donne des résultats prometteurs pour l'animation mais est labellisé manuellement et dédié à un sujet. Aucune expérimentation n'a été réalisée sur la reconnaissance d'expressions ni sur la similarité des disques entre des personnes différentes.

Chang et al. [51] ont testé la méthode de réduction non linéaire nommée *local linear embedding (LLE) - réduction localement linéaire* ainsi que la réduction de Lipschitz (*Lipschitz embedding*) pour apprendre la variété des expressions. Ils ont extrait une variété pour chaque sujet et les ont ensuite alignées sur l'ensemble des sujets. Un classifieur de plus proches voisins (*k-Nearest Neighbor*) a été utilisé ensuite pour reconnaître les expressions. Dans [52], ils abordent le sujet des expressions mélangées qui ne sont pas inclues dans les bases d'apprentissage. Ils créent une variété spécifique à la personne pour chaque sujet par réduction de Lipschitz appliquée sur des séquences vidéo réalisant des transitions entre le visage neutre et l'une des 6 expressions de base. Ces transitions représentent 6 chemins sur la variété, les expressions mélangées d'intensité variable se situant entre ces chemins. Ils ont appris un modèle probabiliste pour reconnaitre chaque expression, modèle qui prend en compte l'information temporelle des séquences vidéo. Les expressions mélangées sont alors classifiées quantitativement dans les 6 catégories. Leurs expérimentations ont été réalisées sur 5 sujets connus et ne traitent pas des sujets non présents dans la base d'apprentissage. Aucun test n'a été réalisé sur l'adéquation entre les représentations des expressions mélangées similaires de différents sujets. Shan et al. [53] ont proposé une méthode appelée *Supervised LPP (Locality Preserving Projections)* pour extraire une unique variété pour tous les individus. Comme [51], un algorithme de plus proches voisins est ensuite utilisé pour classifier et reconnaître les expressions. Comme pour la réduction de Lipschitz [51, 52], cette réduction nécessite une grande quantité de données d'apprentissage pour calculer une variété qui approxime correctement l'espace des expressions. Ils ne mentionnent pas non plus le comportement de leur système sur des sujets inconnus.

2.2 Spécificités des Visages versus Généricité des Expressions

Il est fréquent de dire que l'identité est spécifique à la personne et que l'expression est commune à tous. Cette section introduit un certain nombre de thématiques et d'axes de réflexions liées à ce postulat.

2.2.1 Déformations par rapport au Visage Neutre

Une méthode très répandue pour minimiser l'impact de l'identité sur les données du visage consiste à aligner les caractéristiques faciales des différents sujets en soustrayant les caractéristiques du visage neutre. Cheon & Kim [54] ont proposé d'aligner les expressions du visage définies par des vecteurs AAM en utilisant la méthode Diff-AAM. Dans cette méthode, les caractéristiques des expressions sont calculées en effectuant la différence entre les caractéristiques

des modèles actifs d'apparence (AAM) du visage expressif et celles du visage de référence (visage neutre).

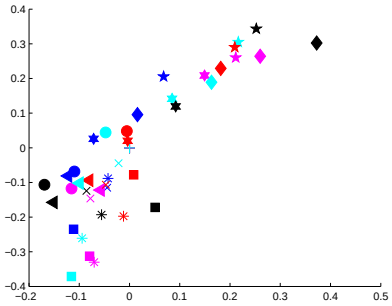


FIGURE 2.1 – 8 expressions similaires de 5 sujets dont les vecteurs d'apparence ont été alignés en soustrayant le vecteur du visage neutre. Affichage des deux premières dimensions de l'espace d'apparence, c'est-à-dire des 2 principales déformations faciales.

Ces méthodes assument que les personnes ont des motifs d'expression similaires lorsqu'ils passent du visage neutre à un visage expressif, et linéaires (c'est-à-dire que le passage du visage neutre à une expression est une droite dans l'espace créé) ce qui n'est qu'une approximation. A titre d'exemple, certaines personnes réalisent des sourires plats alors que d'autres réalisent des sourire en croissant. La figure 2.1 montre, dans les 2 premières dimensions de l'espace d'apparence, 8 expressions similaires de 5 sujets dont les vecteurs d'apparence ont été alignés en

soustrayant le vecteur du visage neutre. Nous constatons que les 2 premières déformations ne sont pas suffisamment discriminantes.

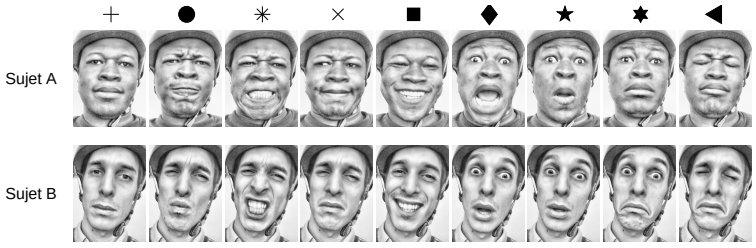
Il est aussi à noter que cette méthode est souvent couplée avec d'autres méthodes. Cheon & Kim [54] réalisent dans leurs travaux un apprentissage de variété sur les paramètres Diff-AAM avant d'effectuer des tâches de reconnaissance (voir section 2.1.2).

Une autre technique prenant en considération le visage neutre consiste à normaliser les informations par rapport aux données d'identité caractéristiques de la personne (écart entre les yeux, distance de la bouche au menton,...). C'est le cas des unités FAPUs (Face Animation Parameter Units) du système MPEG4 (voir description du MPEG4 dans la section 2.1.1).

Ces méthodes s'appuient sur les données du visage neutre du sujet pour définir l'identité de la personne. Dans un système entièrement automatisé, la détection du visage neutre et de ses composantes (par exemple un ensemble de points caractéristiques) sur un sujet inconnu est encore un objet d'études [54].

2.2.2 Modèles Génériques ou Spécifiques

Une seconde méthode, pour s'affranchir des différences d'identité, consiste à définir des modèles spécifiques à chaque personne. Dans le cadre de la description du visage par extraction des déformations de forme et/ou de texture (ACP, IAC), un modèle créé sur un sujet contient dans ses paramètres uniquement des informations de déformations liées aux expressions. Nous parlons alors de **modèles spécifiques à la personne**, par opposition aux modèles appris sur des visages expressifs de différentes personnes que nous nommons **modèles génériques**.



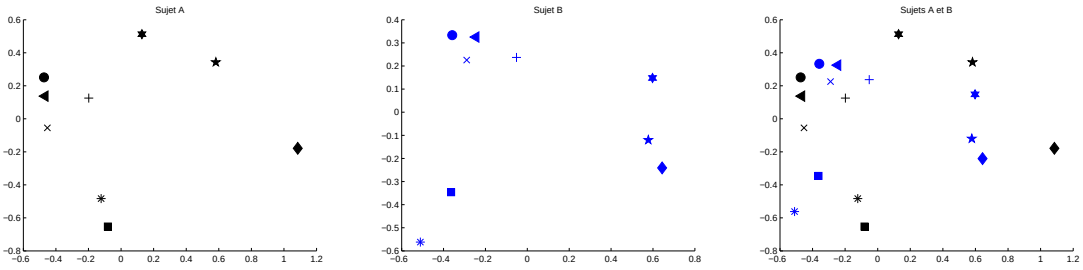


FIGURE 2.2 – Vecteurs d'apparence de 8 expressions similaires et visage neutre de 2 sujets avec des modèles spécifiques. Affichage des deux premières dimensions de l'espace d'apparence, c'est-à-dire des 2 principales déformations faciales.

L'inconvénient des modèles spécifiques est que les paramètres n'ont à priori pas de signification et ne sont de fait pas comparables entre différents sujets. La figure 2.2 montre les deux premières composantes de 8 expressions similaires de deux sujets. Les caractéristiques sont issues de modèles actifs d'apparence créés sur ces 8 expressions. Nous constatons que les premières déformations (principales déformations) ne sont pas toujours les mêmes entre les sujets.

Cette technique nécessite alors d'*aligner* à posteriori les différents sujets. Une première méthode consiste à labelliser manuellement à posteriori l'espace créé [50] et à réaliser des traitements qui sont spécifiques à la personne. Cette méthode a pour avantage de donner des informations très précises sur les expressions de la personne mais signifie que l'apprentissage est à réaliser à chaque nouveau sujet.

Une autre technique pour aligner les sujets consiste à utiliser des algorithmes d'alignement basés sur la similarité sémantique des expressions de forte intensité [51, 53]. [51] souligne la difficulté à aligner les espaces spécifiques. En effet, dans ces travaux, il indique qu'un alignement linéaire sur ses variétés créées par réduction lipchitzienne ne donne pas de résultats satisfaisants car ce type d'alignement ne préserve pas la similarité sémantique des points. Une labellisation manuelle sémantique des variétés semble donc nécessaire pour pouvoir aligner les espaces. A noter que dans leurs travaux [51, 53], aucun test de similarité sur des expressions non prototypiques n'a été réalisé. Nous n'avons donc pas d'information quantitative sur la pertinence de la variété finale générique obtenue.

A noter aussi que les modèles créés sont dépendants des données d'apprentissage de chaque personne. Le nombre d'images ainsi que les exemples utilisés impactent directement les modèles créés et rendent l'alignement d'autant plus difficile.

2.2.3 Extraction de l'Intensité

Il n'est pas facile de définir la notion d'intensité d'une expression. Nous considérons ici que plus la déformation faciale du visage par rapport à sa position au repos (neutre) est importante, plus l'intensité de l'expression est élevée. Tout d'abord, nous pouvons considérer que l'intensité est spécifique à chaque sujet dans la mesure où chaque personne possède une déformation maximale du visage, liée à l'élasticité de chacun de ses muscles. Elle peut être mesurée en utilisant les déformations géométriques du visage ou encore la densité de rides apparaissant sur le visage. Nous définissons donc l'intensité en prenant en compte le visage neutre de la personne ainsi que la déformation maximale possible pour cette personne.

Esau et al. [55] ont traité de cette notion en utilisant un modèle d'émotion floue qui s'adapte aux caractéristiques des visages mais nécessite une phase préalable d'apprentissage du visage.

De nombreux systèmes normalisent leurs données par rapport à la déformation maximale des expressions, permettant ainsi de *gérer* cette notion d'intensité. Chang et al. [51] alignent les variétés spécifiques aux sujets en prenant en compte les expressions à leur intensité maximale et en les mappant de sorte que les expressions d'intensité maximale aient des composantes non nulles valant 1.

2.2.4 Utilisation de la Dynamique

L'importance du chronométrage dans la définition d'une expression est désormais couramment accepté [56]. L'analyse des expressions au travers de leur dynamique se base sur les représentations des sections précédentes en ajoutant la prise en compte du facteur temps. C'est-à-dire qu'il ne s'agit plus de décrire une expression à partir d'une image mais à partir d'une

séquence d'images. Une première méthode consiste à utiliser les informations faciales de nombreuses images consécutives d'une séquence vidéo. C'est le cas des techniques telles que les HMM [57] ou les réseaux de neurones récurrents [58]. Une autre technique consiste à identifier 3 phases, bien que les mouvements du visage soient continus : l'attaque (onset), le maintient (apex), la terminaison (offset) [59], dont les définitions sont approximatives. Une troisième technique consiste à coder l'information temporelle dans la description des visages. C'est le cas des Volume Local Binary Patterns, pour lesquels les données de texture du visage (LBP) incluent alors directement les informations des images précédentes et suivantes de la séquence vidéo [60].

Schmidt & Cohn [61] ont montré qu'il existe des informations dynamiques universelles concernant la réalisation des expressions, ainsi que des spécificités liées aux individus. Ils ont étudié l'ordre d'apparence des AUs et la durée de l'onset lors de sourires et ont constaté une invariance entre les personnes différentes. En revanche, les AUs exprimées dans un contexte précis semblent stables pour un individu au cours du temps mais diffèrent potentiellement entre les sujets.

2.3 Expressions versus Émotions

L'analyse des expressions faciales cherche à décrire le mouvement du visage et les déformations issues de l'activation des muscles à partir d'informations purement visuelles. L'analyse des émotions, quant à elle, est une tentative d'interprétation qui demande souvent la compréhension de la situation ainsi que des informations de contexte. En effet, les émotions sont le résultat de différents facteurs qui peuvent ou non déclencher des modifications des canaux de transmissions que sont la position du visage, la voix, les gestes, la direction du regard et les expressions faciales.

2.3.1 Structure globale d'un système d'analyse des émotions

De nombreux systèmes d'analyse des expressions se décomposent en deux étapes : la description du visage (extraction de caractéristiques faciales) et la classification des expressions (en Action Units ou en Catégories émotionnelles). De nombreux états de l'art sur l'analyse des expressions faciales sont construits de la sorte [45, 1, 46].

Cette façon d'aborder le sujet est certainement due au poids très important des travaux d'Ekman, tant sur la description d'une expression (système FACS [6]) que sur l'universalité des émotions [37]. Pourtant, il s'agit bien de deux niveaux différents. Le premier (AU) concerne la description d'une expression alors que le second (catégories émotionnelles) se propose d'aborder le domaine de la signification d'une expression.

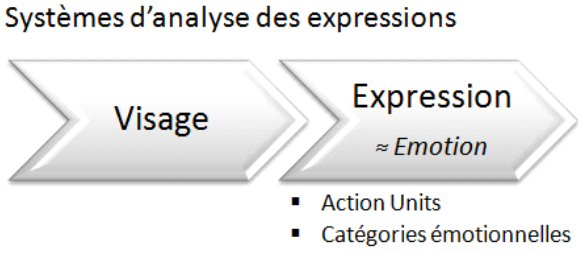


FIGURE 2.3 – Étapes des systèmes d'analyse des expressions.

En indiquant une relation un-un entre expression et émotion, les systèmes se sont focalisés sur ces deux étapes, réduisant la seconde étape à une classification en un nombre restreint de catégories. Voyant les limites de cette représentation pour l'analyse des émotions, les systèmes se sont tournés vers une représentation continue des émotions (voir la section 3.2 sur les modes de représentation des émotions). Ces systèmes ont alors souvent gardé deux étapes [14] : la première concerne la description du visage, la seconde est une classification en terme d'émotions (représentation dimensionnelle). La notion d'expression faciale a été supprimée (et parfois confondue) avec la description du visage. C'est le cas des systèmes parlant d'*informations de bas niveau*. En revanche, lorsque les systèmes utilisent des informations *de haut niveau*, ce qui est le cas dans cette thèse, ils se découpent en trois étapes : la description du visage, la description de l'expression et l'interprétation en terme d'émotion. Nous y reviendrons dans la section 3.3.4.

FIGURE 2.4 – Étapes des systèmes d'analyse des émotions. Deux méthodes : dans la première, l'analyse des émotions se fait à partir des informations des visages (informations de bas niveau) ; dans la seconde, l'analyse des émotions se fait à partir des informations des expressions (informations de haut niveau).

2.3.2 Notre Représentation des Expressions Faciales

La représentation par Action Units (AUs), largement utilisée par les psychologues, ne semble pas adaptée à la vision par ordinateur, dans la mesure où la reconnaissance des AUs donne encore de faibles résultats et que la combinaison des AUs est problématique. Les représentations des expressions prenant en compte la globalité des déformations du visage semblent plus performantes. Dans cette thèse, nous proposons une représentation qui utilise les déformations du visage dans sa globalité. Séparer l'expression de l'identité est alors le problème clef. Les méthodes bilinéaires donnent de bons résultats lorsque l'une des deux composantes (identité ou expression) est connue de la base d'apprentissage mais montrent leurs limites lorsqu'à la fois le visage et l'expression sont inconnus du système, ce qui est le cas dans de nombreuses applications de la vie de tous les jours. C'est pourquoi nous avons opté pour une méthode basée sur la définition d'une variété des expressions. Ces méthodes montrent des résultats prometteurs pour la représentation d'expressions mélangées.

Variété des expressions Les méthodes actuelles basées sur les variétés ont besoin d'une grande quantité de données d'apprentissage afin de construire une variété qui approxime correctement l'espace des expressions. Ici, nous explorons une méthode permettant de *créer l'espace des expressions avec très peu de données* (visage neutre et 8 visages expressifs). Nous nous focalisons sur une représentation permettant de prendre en compte les expressions qui ne sont pas présentes dans les bases de données existantes. Contrairement aux autres techniques qui traitent des expressions *mélangées*, nous mesurons la pertinence de la représentation en calculant un taux de reconnaissance sur des expressions qui sont inconnues de la base d'apprentissage.

La principale contribution est la spécification d'une nouvelle représentation universelle des expressions faciales émotionnelles. L'originalité de l'approche est que nous ne nous focalisons pas sur les caractéristiques d'une expression mais sur *l'organisation des expressions les unes par rapport aux autres*. L'organisation des expressions est extraite des données de description des visages et non pas issues de modèles de représentation des émotions. Nous avons donc bien un espace des expressions et non un espace des émotions.

Identité vs. Expression Nous montrons aussi dans cette thèse que l'organisation que nous proposons est indépendante de l'identité des personnes, ce qui nous permet de qualifier de façon unique chaque expression. Notre système permet de traiter ainsi des différences d'identité entre les nouveaux sujets et les sujets qui sont présents dans la base d'apprentissage.

Nous utilisons cette représentation universelle pour transformer l'espace des déformations du visage, espace qui est spécifique à chaque personne, en un espace des expressions, commun à tous. Le système proposé s'inscrit dans un processus en 3 étapes : description du visage, description de l'expression, interprétation de l'expression.

Modèles spécifiques L'alignement des sujets est aussi une question comme nous l'avons vu dans la section 2.2.2. La plupart des systèmes actuels utilisant la notion de variété créent d'abord la variété qui est spécifique à la personne et ont ensuite besoin d'une seconde phase permettant d'aligner ces variétés. Dans cette thèse, nous analysons d'abord la structure de l'espace des expressions et montrons que la représentation que nous proposons est universelle. Dans un second temps, nous créons une variété conforme à cette structure, si bien que nous n'avons pas besoin de réaliser un alignement supervisé des différents sujets.

Nous adressons aussi l'extension aux sujets inconnus de la base d'apprentissage. Nous proposons de créer une variété des expressions présumée pour ces nouveaux sujets. Pour cela, nous construisons un modèle spécifique de visage pour chaque personne. Ce modèle s'adapte à la morphologie du sujet sans phase préalable d'apprentissage du sujet. La particularité de l'approche est que nous synthétisons les expressions des sujets inconnus pour créer leur espace d'apparence complet. Les expressions synthétisées sont conformes avec l'organisation des expressions, ce qui signifie que l'espace d'apparence contient toutes les expressions définies dans l'espace des expressions ; même si le nouveau sujet ne les a pas réalisées.

Chapitre 3

Signification d'une Expression

Les expressions faciales ont des origines différentes (signaux physiologiques, visèmes, interactions sociales, états mentaux) [1, 2]. De nombreux systèmes utilisent alors l'analyse des expressions faciales pour extraire ce type d'information. Nous parlons dans ce cas de *signification* d'une expression faciale. Il ne s'agit plus de *description* des données mais d'*interprétation* des observations. Nous donnons alors un *label* (continu ou discret) à l'expression. Comme indiqué en fin de chapitre précédent, il n'existe pas de bijection entre l'expression faciale et son interprétation : une même expression peut avoir différentes origines. De plus, plusieurs facteurs agissent simultanément pour produire une expression : par exemple parole et émotion simultanées. Par ailleurs, l'expression du visage n'est pas le seul moyen de communication. C'est pourquoi, nous observons récemment un engouement pour les systèmes multimodaux, prenant en compte aussi bien les informations des expressions faciales que les autres modes d'expression (voix, langage du corps) ainsi que des données de contexte (environnement).

Cette section s'attache dans un premier temps à présenter les différentes origines des expressions faciales. Dans le cadre des travaux présentés dans cette thèse, nous appliquerons l'analyse des expressions faciales à la détection d'émotion. Nous nous focaliserons donc ici plus particulièrement sur les états émotionnels en précisant quels sont les modes de représentation existants. Nous présenterons ensuite quelques systèmes d'interprétation.

Sommaire

3.1 Les Différentes Origines des Expressions Faciales

Nous observons fréquemment un découpage en 4 catégories [1, 2] :

- Signaux physiologiques
- Visèmes
- Interactions sociales
- États mentaux

Bien que nous distinguions plusieurs origines aux expressions faciales, les mouvements correspondant se réalisent sur le visage le plus souvent de façon simultanée. Il s'agit donc ici d'une catégorisation par fonction et non par ensemble disjoint. Certains de ces mouvements d'ailleurs sont conflictuels comme par exemple l'expression de la joie par le sourire et simultanément la parole.

Nous donnons ci-dessous une brève description de ces 4 catégories.

Signaux physiologiques Les signaux physiologiques sont des signaux automatiques produits par la personne pour satisfaire des besoins physiques. Il s'agit par exemple du clignement des yeux, permettant l'hydratation de ceux-ci, ou encore la respiration.

Nous pouvons aussi inclure dans cette catégorie la direction du regard et les mouvements de la tête lorsque le but est d'obtenir un contact visuel. A noter que la direction du regard, comme les mouvements de la tête, ne sont pas toujours d'origine physiologique mais peuvent fournir aussi des informations sur l'état émotionnel. Cet exemple montre la limite d'un tel découpage et la complexité de l'interprétation d'une expression.

Visèmes Les visèmes sont la représentation visuelle des phonèmes. Ce canal correspond donc au mouvement des lèvres dus à la parole lors de conversations.

Interactions sociales Certains mouvements faciaux sont provoqués pour réaliser ou amplifier la transmission d'information. Il s'agit par exemple de haussement des sourcils permettant d'accentuer le discours, de froncement des sourcils pour montrer son mécontentement ou son désaccord, ou encore du hochement de tête permettant d'acquiescer.

Nous pouvons aussi ajouter dans cette catégorie les émotions non ressenties, telles que les faux sourires, c'est-à-dire les sourires volontaires non initiés par une joie sincère.

États mentaux

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

Paul Brulat - *Pensées - 1919*

Dans les interactions de la vie courante, les personnes expriment des états mentaux ou affectifs qui sont souvent visibles sur le visage.

Darwin proposait déjà dès 1874 [62] l'universalité des expressions faciales chez l'homme et les animaux. Par la suite, Ekman a montré que les 6 émotions primaires que sont la joie, la tristesse, la colère, la surprise, le dégoût et la peur s'expriment sur les visages de façon universelle et sont représentés par une unique expression faciale [37]. Elles sont appelées couramment les émotions basiques.

Néanmoins, dans la vie de tous les jours, les états émotionnels sont souvent plus complexes que ces 6 émotions, citons par exemple la réflexion, l'embarras ou la dépression [63]. C'est pourquoi, après s'être intéressé à des classifications en un petit nombre de classes discrètes, les chercheurs se sont penchés sur des représentations continues de l'espace des émotions. Nous proposons de passer en revue, dans la section suivante, les modes de représentations des émotions.

3.2 Mode de Représentations des Émotions

La représentation des émotions a d'abord été étudiée par les psychologues [64]. Nous présentons ici 3 modes classiques de représentation des émotions, selon qu'il soit discret ou continu. Le premier mode (*Catégories)* représente les émotions sous forme discrète. Le troisième mode (*Représentations dimensionnelles*) les représente sous forme continue. Le second mode (*Organisation d'un nombre fini d'émotions*) est un mode intermédiaire dans lequel la représentation discrète est *structurée*, sans pour autant pouvoir parler de réel aspect continu de la représentation.

D'autres approches des psychologues ne tentent pas de représenter les émotions en tant que telles mais partent du principe que les émotions sont le résultat de notre évaluation (*appraisal*) des évènements, des situations [65] ; c'est-à-dire de nos interprétations et explications des circonstances. Ces évaluations causent une réponse émotionnelle. Cette théorie, nommée *appraisal théorie* ou *théorie de l'évaluation*, met aussi en avant que cette évaluation influence non seulement nos émotions mais aussi les futures évaluations des évènements et situations qui se produiront. Elle nécessite une grande quantité de données pour être mise en œuvre : il est nécessaire d'avoir à la fois des données de nature différente pour l'analyse globale des situations (pas seulement les expressions faciales), mais aussi des données permettant d'analyser des moments différents pour la prise en compte des évaluations précédentes des situations.

3.2.1 Catégories

La façon la plus courante de définir une émotion est d'utiliser un label (joie, peur, ...). Le langage humain est très prolifique dans la production de labels permettant de décrire un état mental. Le dictionnaire de Sweeney et Whissell liste plus de 4000 mots permettant de définir un état mental [66]. Dans l'analyse automatique des expressions faciales, la plupart des systèmes se focalisent sur la catégorisation des 6 émotions de base proposées par Ekman [37]. En effet, les études d'Ekman sont séduisantes pour la mise en œuvre de systèmes automatisés dans la mesure où elles indiquent une correspondance 1-1 entre expression et émotion (voir section 2.1.1). Dès lors que l'on sort de ces émotions associées à des expressions universelles ou que l'on doit gérer des canaux simultanés (émotion et parole), l'extraction de label émotionnel est plus compliquée. Nous y reviendrons dans la section suivante (paragraphe 3.3.1).

3.2.2 Organisation d'un nombre fini d'émotions

La représentation sous forme de catégories est certes séduisante mais elle est limitée. Elle ne permet pas de modéliser l'ensemble des états émotionnels, et l'on arrive très vite à un nombre de labels difficile à gérer [66]. Les psychologues se sont alors penchés sur les relations de proximité qui pouvaient exister entre ces labels.

Plutchik [67] a créé une roue des émotions en explorant les relations entre les concepts émotionnels. Cette roue est composée de huit émotions basiques arrangées en quatre paires d'opposés (joie-tristesse, confiance-dégoût, peur-colère, surprise-anticipation), et de huit autres émotions

avancées, chacune étant un mélange de deux émotions basiques (amour étant la somme de joie et confiance, soumission de confiance et peur, crainte de peur et surprise, désapprobation de surprise et tristesse, remords de tristesse et dégoût, mépris de dégoût et colère, agressivité de colère et anticipation, optimisme d'anticipation et joie). Il a aussi étendu son modèle pour prendre en compte 3 niveaux d'intensité pour ses 8 émotions basiques (extase-joie-sérénité, adoration-confiance-résignation, terreur-peur-appréhension, stupeur-surprise-distraction, chagrin-tristesse-songerie, aversion-dégoût-ennui, rage-colère-contrariété, vigilance-anticipation-intérêt).

Meftah et al. [68] ont proposé une modélisation algébrique basée sur l'approche de Plutchik. Une émotion est un vecteur de 8 composantes, chaque composante indiquant le taux correspondant à l'émotion de base de la roue de Plutchik (joie, tristesse, confiance, dégoût, peur, colère, surprise, anticipation). Cette représentation permet à la fois de représenter les émotions mélangées, de gérer la notion d'intensité et d'avoir une représentation simple et intuitive des états émotionnels complexes.

Dans le même ordre d'idée, Ruttkay et al. [69] ont mis en œuvre un outil pour contrôler les émotions d'agents virtuels. Cet outil prend la forme d'un disque qui est basé sur l'interpolation entre les six visages émotionnels de base. Ce disque permet de contrôler l'intensité et la transition entre les émotions. Néanmoins, il ne permet pas de gérer l'ensemble des expressions : la navigation ne s'effectue que dans une partie limitée de l'espace des expressions. Pour ne pas restreindre l'espace des expressions, ils ont proposé un second outil basé sur les représentations dimensionnelles des émotions.

3.2.3 Représentations dimensionnelles

Dans le même temps, d'autres psychologues se sont penchés sur l'aspect continu des émotions et sur leur organisation en vue de définir les dimensions de l'espace des émotions.

Russel s'est posé la question de la dépendance des émotions entre elles et a proposé un modèle spatial permettant de représenter les concepts affectifs sous forme de cercle [70]. Ce modèle a été repris par Cowie et al. [71] qui ont proposé un outil appelé FEELTRACE pour permettre à des observateurs d'annoter le contenu émotionnel des épisodes de la parole telle qu'ils la perçoivent au fil du temps. Cet outil est constitué d'un espace circulaire basé sur les deux dimensions *activation-évaluation*. Whissell [72] propose un dictionnaire permettant de quantifier les valeurs d'évaluation et d'activation pour presque 9 000 concepts émotionnels donnant ainsi un lien entre les concepts et les dimensions émotionnelles.

Le second outil proposé par Ruttkay et al. [69] est constitué de deux carrés, permettant de représenter 4 dimensions (deux dimensions par carré). Les dimensions émotionnelles sont basées sur les quatre premières composantes d'une analyse en composantes principale (ACP) de 15 paramètres d'action du visage. Cet outil permet de créer une plus grande variété d'expressions nouvelles que leur disque (cf. paragraphe 3.2.2), mais peut conduire à des expressions faciales non réalistes.

Plus récemment, Fontaine et al. [73] se sont penchés sur la question de la dimensionnalité de l'espace des émotions en proposant un article dont le titre suscite la curiosité (nous pourrions traduire le titre par *Le monde des émotions n'est pas bidimensionnel*). Dans leurs travaux, ils ont réalisé une analyse en composantes principales (ACP) sur 24 émotions caractérisées par 144 composantes. Ils ont apporté un label aux quatre premières dimensions trouvées (c'est-à-dire les 4 principales déformations) : la valence, le pouvoir (power), l'activité (arousal) et l'attente (expectation).

3.3 Systèmes d'Interprétation

Les systèmes permettant de détecter l'émotion en se basant sur l'analyse des expressions faciales dépendent grandement du mode de représentation des émotions choisi. En effet, le résultat escompté peut être discret ou continu, sous forme de concepts émotionnels ou de dimensions. Nous n'aborderons pas la mise en œuvre de la théorie d'évaluation (*appraisal theory*). Nous pouvons distinguer trois types de techniques correspondant aux trois premiers paragraphes de la section précédente :

- les classifieurs, qui fournissent en général une information discrète, le plus souvent sous la forme de concepts émotionnels,
- les méthodes de régression, qui transforment un espace continu des expressions en un espace continu des émotions, le plus souvent sous forme dimensionnelle,
- les méthodes hybrides qui utilisent des classifieurs pour fournir une information continue, le plus souvent sous la forme d'organisation de concepts.

Ces trois types de techniques sont présentés ci-dessous.

3.3.1 Classifieurs pour la représentation sous forme de catégories

Comme nous l'avons déjà mentionné au chapitre précédent, de nombreux systèmes se sont appuyés sur les travaux d'Ekman, associant l'expression faciale à l'émotion. De bons états de l'art sont disponibles dans [45, 1, 46]. Ces méthodes sont basées sur l'analyse des expressions faciales et effectuent souvent leur apprentissage sur une base de données contenant une grande quantité de visages expressifs. Cela permet aux classifieurs d'apprendre les déformations du visage liées à l'expression, et cela pour tous les types d'identité et pour toutes les intensités. Ce type d'approche peut entraîner un surapprentissage, c'est-à-dire que les systèmes finissent par se spécialiser sur les visages de la base d'apprentissage et perdent leur capacité à généraliser à d'autres visages. Ces méthodes se focalisent sur la reconnaissance d'un petit nombre d'émotions, le plus souvent les six émotions universellement associées à des expressions faciales distinctes [37], mais ne sont pas adaptés à traiter les émotions inconnues, les émotions mélangées ou encore à gérer les différences d'intensité dans les émotions.

Quelques tentatives ont été faites pour prendre en compte d'autres états mentaux tels que la douleur [74, 75], la fatigue [76], l'intérêt [77], la frustration [78], mais il s'agit là encore d'indiquer si l'émotion est ou pas présente. Ces systèmes ne savent pas traiter les autres types d'émotions.

Certaines méthodes fusionnent les informations de différents classifieurs afin d'améliorer leur taux de reconnaissance. Hupont et al. [14] proposent une stratégie de vote majoritaire pondéré utilisant les matrices de confusion de chaque classifieur (5 classifieurs dans leurs travaux) pour pondérer la prise en compte de l'information issue de chaque classifieur et en ajoutant des règles d'incompatibilité qui suivent les lois de Plutchik. L'émotion est alors représentée comme un vecteur de 7 composantes (6 émotions de base plus l'émotion neutre), donnant les valeurs de confiance comprises entre 0 et 1 pour chacune de ces 7 émotions.

D'autres systèmes se sont tournés vers une représentation des émotions sous la forme de dimensions. C'est le cas du challenge AVEC 2011 [10] qui propose une comparaison des méthodes d'analyse automatique d'émotion. Les émotions sont labellisées par l'outil FEELTRACE [71] en quatre dimensions : valence, arousal, power et expectancy [73]. La vérité terrain est proposée sous forme binaire pour chacune de ces quatre dimensions (valence positive ou négative, personne

active ou passive, ...). Les résultats du challenge sont peu encourageants (pas meilleurs qu'un tirage aléatoire), spécialement pour la dimension power. Ils tendent à montrer que l'utilisation de classifieurs afin de fournir une information discrète sur une émotion ayant une représentation dimensionnelle n'est pas adaptée. Nous pouvons aussi nous poser la question de la pertinence du mode de représentation des émotions (sous forme dimensionnelle).

3.3.2 Classifieurs pour la représentation sous forme continue

Certaines méthodes proposent de partir de classifieurs permettant de distinguer quelques émotions de base et d'étendre à un système continu. Les travaux de Hupont et al. [14] présentés dans le paragraphe précédent ont aussi abordé ce point. A partir des valeurs de confiance de chacune des 7 émotions, ils déterminent une valeur de valence et d'arousal en calculant le centre de masse des coordonnées des 7 émotions mappées sur l'espace de Whissel [72] et pondérées par les valeurs de confiance. Néanmoins, ce type de système se base sur les erreurs des classifieurs (confusion) afin de déterminer des émotions mélangées. Nous pouvons penser que plus les classifieurs seront bons, moins il y aura détection d'expressions mélangées. De plus, l'espace produit n'est pas continu mais résulte de la combinaison possible des confusions des classifieurs. Aux limites (classifieurs ayant un taux de reconnaissance de 100%), l'espace produit se résumera à 7 possibilités correspondant aux 6 émotions de base et à l'émotion neutre.

Ozkan et al. [79], dans le cadre du challenge AVEC 2012 [80], proposent de modéliser les variations continues des 4 dimensions émotionnelles par un ensemble de valeurs discrètes (6 valeurs pour chaque label émotionnel) et ainsi d'appliquer leur classifieur HMM sur ces données. Ce système leur a permis d'obtenir la 4ième place du challenge. Néanmoins, les résultats sont difficilement comparables entre les équipes. Nous y reviendrons dans la section 9.2.1.

3.3.3 Régression pour la représentation sous forme continue

Plus récemment, des méthodes de régression ont été utilisées pour permettre de donner une prédiction de l'émotion sous la forme de dimension continue. Ishii et al. [81] ont proposé une méthode pour générer une carte émotionnelle spécifique à la personne et compatible avec le modèle de Russell [70]. Le mapping est alors appris lors d'une phase préalable d'apprentissage pour chaque personne et par un algorithme d'apprentissage supervisé.

Le principal frein à ces études est la disponibilité de données labellisées pour les systèmes nécessitant des apprentissages et pour l'évaluation de ces systèmes. Les bases de données HUMAINE [12] puis SEMAINE [13] ont permis de combler ce manque. En 2012, le challenge AVEC [80] est organisé. Il reprend le principe de AVEC 2011 (voir paragraphe précédent) mais cette fois, la vérité terrain n'est plus binaire mais continue. L'objectif du challenge est alors de reconnaître les variations d'états émotionnels au travers des quatre dimensions : valence, arousal, power et expectancy [73]. Les méthodes mises en œuvre par les concurrents sont variées [82, 79, 83]. Elles sont présentées dans le paragraphe suivant dans la mesure où elles prennent en compte des données multimodales.

3.3.4 Les Systèmes Multimodaux

Pour exprimer son âme, on n'a que son visage.
Jean Cocteau - *Renaud et Armide - 1943*

Les expressions faciales ne sont qu'un vecteur parmi d'autres permettant d'avoir une information globale sur le comportement humain. C'est pourquoi l'analyse des expressions faciales est souvent une brique d'un système plus global. Même si cela n'est pas le cœur de nos recherches, il nous a paru intéressant de nous immerger dans ce contexte plus global, afin notamment de vérifier l'utilité et la pertinence du système proposé. Nous faisons donc dans ce paragraphe, une brève présentation des systèmes multimodaux existants.

La reconnaissance des émotions via des informations multimodales a connu un grand essor ces dernières années [84]. Comme pour l'analyse d'un seul mode, la majorité des systèmes se concentre sur la classification des émotions discrètes [85, 86, 87]. Certains systèmes ont évolué vers une représentation tridimensionnelle des émotions (activation, valence, évaluation), mais la valeur de sortie restait discrète [88]. Plus récemment, avec la mise à disposition de bases de données appropriées, nous constatons un engouement pour les systèmes permettant de fournir une sortie continue [82, 79, 83].

Caractéristiques pertinentes La plupart des systèmes de fusion utilisent uniquement deux modalités : acoustique (prosodie) et visuelle (expressions faciales). Le contexte est rarement pris en compte. Nous pouvons citer Ozkan et al. [79] qui utilisent l'observation suivante : plus les sujets sont engagés dans une conversation, plus leurs émotions sont intenses.

L'un des points clefs dans les systèmes de reconnaissance d'émotion est le niveau des caractéristiques utilisées pour la prédiction. Certaines méthodes utilisent des informations de bas niveau telles que les paramètres de forme du visage proches des paramètres des modèles ASM [82], ou encore des descripteurs de texture LBP (Local Binary Patterns) [83]. D'autres systèmes utilisent des caractéristiques de haut niveau telles que l'intensité du sourire ou la direction du regard [79], permettant d'avoir moins de données à traiter et ayant souvent l'avantage d'avoir une correspondance intelligible entre la caractéristique et l'émotion.

Systèmes de fusion La fusion des modalités peut être effectuée à différents stades. Nous parlons de *early data fusion* lorsque la fusion est effectuée sur les caractéristiques multimodales, avant le processus de reconnaissance. Nous parlons de *late data fusion* lorsque la fusion est réalisée sur les prédictions, c'est-à-dire après une première étape de reconnaissance réalisée pour chaque modalité.

Dans le premier cas, une première méthode consiste à concaténer les caractéristiques provenant des différents modes [85]. Une autre façon de faire consiste à utiliser la corrélation entre les caractéristiques. Cela se fait par exemple par des méthodes HMM [89], les méthodes de réseaux de neurones [90, 88] ou les méthodes de réseau Bayésien [87].

Dans le cas d'une fusion tardive (*late data fusion*), les résultats de la reconnaissance des différentes modalités sont fusionnés. Les systèmes pondèrent généralement les différentes prédictions, par exemple avec des poids empiriques [86] ou des règles [91].

Dans tous les cas, les principales contraintes pour les systèmes de fusion utilisant des données réelles sont qu'ils doivent être capables de pouvoir traiter des informations manquantes (modalité non disponible sur une période) et être robustes aux données d'entrée erronées (fiabilité de l'extraction des informations du mode).

Impact de la multimodalité sur les performances Il est difficile de se faire une idée précise et générale de l'impact de la multimodalité sur les résultats globaux des systèmes. En effet, les résultats sont très variables selon les méthodes (allant de l'extraction des caractéristiques à la fusion) et selon les dimensions prédites. Voici quelques exemples.

Le challenge AVEC 2011 [10] proposait trois sous challenges pour l'analyse des émotions : un sous challenge audio, un sous challenge vidéo et un sous challenge audio-vidéo. Seuls deux participants ont concouru pour le sous-challenge audio-vidéo [92, 93]. Les résultats de Ramirez et al. [92] montrent une amélioration de la précision pour la dimension power (passage de 47.1% pour le mode visuel seul et 19.8% pour le mode audio seul à 62.9% pour la fusion des deux modes). En revanche, pour les trois autres dimensions (valence, expectancy et arousal), les résultats sont meilleurs ou équivalents avec uniquement l'analyse de la vidéo (pour arousal : 65.5% en vidéo versus 65.6% en multimodal ; pour expectancy : 61.7% en vidéo versus 53.4% en multimodal ; pour valence : 69.8% en vidéo versus 59.5% en multimodal). Les résultats de Glodek et al. [93] montrent quant à eux une amélioration pour la dimension expectancy (passage de 47.5% pour le mode visuel seul et 41.1% pour le mode audio seul à 58.5% pour la fusion des deux modes). En revanche, pour les trois autres dimensions (valence, power et arousal), les résultats sont meilleurs ou équivalents avec uniquement l'analyse de l'un deux modes (pour arousal : 63.5% en audio versus 54.2% en multimodal ; pour power : 47.3% en vidéo versus 42.7% en multimodal ; pour valence : 65.4% en audio versus 44.8% en multimodal). Les résultats de ces deux équipes montrent la difficulté à analyser ces informations. Il n'est pas possible actuellement de définir l'impact de chaque modalité pour chacune des dimensions.

Le challenge AVEC 2012, quant à lui, n'offrait pas de discutions sur les méthodes de fusion et l'apport de la multimodalité. Certains participants ont néanmoins réalisés ces tests sur la base de développement. Les vainqueurs du challenge [82] ont constaté que leur méthode de fusion améliorait les résultats pour les dimensions arousal et power mais que pour les dimensions valence et expectancy, les résultats étaient meilleurs ou équivalents avec uniquement les informations visuelles, les informations provenant de l'audio donnant de mauvaises prédictions. Il est donc là aussi très difficile de pouvoir ressortir des conclusions.

Deuxième partie

Représentation Universelle des Expressions Faciales

En traitement du signal, il est connu que le choix de la représentation des données influence les analyses réalisées. Ce chapitre propose une nouvelle approche permettant de décrire de façon unique une expression, qu'elle soit connue ou inconnue du système, sur un visage lui-même connu ou inconnu.

Comme indiqué dans l'état de l'art (section 1.6), l'analyse des expressions faciales se focalise sur les caractéristiques communes des expressions, indépendamment de l'identité du sujet. Elle s'appuie aussi sur les données du visage. Dans cette partie, nous passerons donc de l'un à l'autre de ces deux espaces (espace d'apparence pour les données du visage et espace des expressions pour les caractéristiques des expressions - voir figure 3.1).

Nous allons tout d'abord présenter la description des visages sur laquelle nous avons porté notre choix et en montrer les limites pour l'analyse des expressions (section 4.1). A partir de l'étude d'expressions similaires de différentes personnes, nous allons décrire une caractéristique invariante de l'espace des expressions faciales (sections 4.2 à 4.4) et nous utiliserons cette caractéristique invariante pour qualifier de façon unique une expression, connue ou inconnue du système (sections 5.1 à 5.2). Nous étendrons ensuite la méthode à des personnes inconnues (section 5.3). Pour finir, nous analyserons la pertinence de la représentation proposée en calculant les taux de reconnaissance d'expressions non présentes dans les bases d'apprentissage (chapitre 6).

Dans cette partie, nous parlerons de représentation *universelle* des expressions faciales. Il s'agit là de faire écho aux travaux d'Ekman qui parlent d'*universalité* des expressions faciales réalisées lors d'émotions particulières. Ici, nous ne traitons en aucun cas d'émotions. Il s'agit de définir une représentation de l'espace des expressions qui soit *indépendante des personnes*, c'est-à-dire *indépendante de leur identité*. C'est en cela que nous employons le terme *universel*.

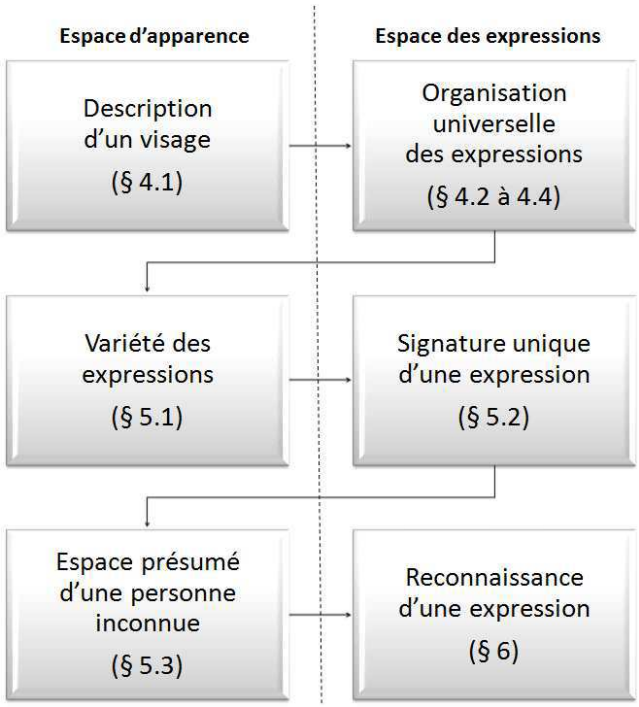


FIGURE 3.1 – Processus décrivant l'organisation de la partie II.

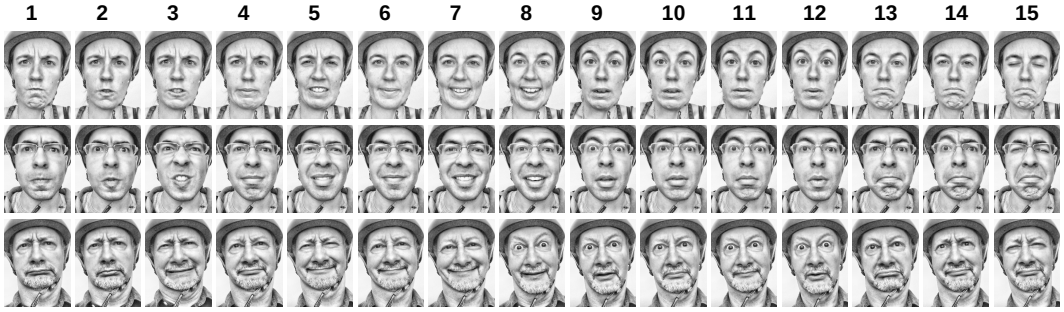


FIGURE 3.2 – Sujets réalisant des expressions similaires.

Visage neutre Nous assumons dans cette partie que le visage neutre des personnes est connu. Nous proposerons au chapitre 5.3.5 une méthode simple permettant d'extraire le visage neutre lorsque des vidéos de la personne sont disponibles.

Description des visages Comme nous avons besoin d'identifier des différences subtiles entre les expressions d'une même personne, notre méthode est basée sur des vecteurs d'apparence issus de modèles actifs d'apparence (AAMs) [18]. Les AAMs sont connus pour fournir des informations spatiales précises sur des points clefs du visage. Néanmoins, le suivi précis par les AAMs, notamment sur des sujets inconnus, reste un sujet d'étude. Ici, nous ne traiterons pas de ce point. Les travaux de cette partie ont été réalisés à partir de visages annotés manuellement, afin d'éviter des éventuels problèmes liés à un mauvais suivi automatique.

Contributions La principale contribution est la spécification d'une nouvelle représentation invariante de l'espace des expressions faciales.

Les systèmes actuels décrivent une expression en cherchant ses caractéristiques, communes à tous les individus, indépendamment des autres expressions. En d'autres termes, ils se focalisent sur une colonne de la figure 3.2.

Dans ce manuscrit, nous explorons un autre aspect, il s'agit non plus de comparer ce qu'il y a de communs entre une expression réalisée par une personne et cette même expression réalisée par une autre personne ; mais entre un ensemble d'expressions réalisées par une personne et ce même ensemble d'expressions réalisées par une autre personne. En d'autres termes, notre système se focalise sur les lignes de la figure 3.2. Ces expressions sont toutes *ordonnancées* de la même façon pour chaque individu. Cette partie va mettre en évidence le fait que l'organisation des expressions entre elles est universelle, c'est-à-dire indépendante du sujet. Nous attesterons que cette organisation est indépendante des personnes en calculant un indice de similarité.

Nous utilisons cette organisation invariante pour transformer l'espace d'apparence spécifique à la personne en un espace des expressions indépendant des personnes. Dans cet espace, nous donnons une signature *direction-intensité* à chaque expression.

Une autre contribution importante de notre travail consiste à créer un espace présumé pour les personnes qui sont inconnues du système. Cet espace est adapté à la morphologie des sujets, sans phase préalable d'apprentissage du sujet. La particularité de l'approche est que nous synthétisons les expressions des sujets inconnus pour créer leur propre espace d'apparence complet. Les expressions synthétisées sont conformes à celles utilisées dans la création de l'espace

des expressions, ce qui signifie que l'espace présumé de la nouvelle personne est complet, même si cette personne n'a pas réalisé l'ensemble des expressions possibles.

Chapitre 4

L'Organisation des Expressions Faciales est Universelle

Dans ce chapitre, nous allons tout d'abord présenter les limites des modèles de description des visages pour la représentation des expressions. Cette étude nous permettra d'introduire notre proposition de représentation des expressions qui est basée sur l'*organisation des expressions* les unes par rapport aux autres. Nous allons ensuite présenter comment nous avons mesuré la similarité des organisations de deux personnes différentes. Pour finir, nous montrerons que cette organisation est *universelle*.

Sommaire

4.1 Les Limites de la Description d'un Visage

Comme indiqué dans l'état de l'art (partie I), la description d'un visage possède à la fois des informations sur les expressions des visages mais aussi sur l'identité des personnes. Nous revenons ici plus en détail sur ces aspects et nous introduisons les données sur lesquelles nos travaux ont été réalisés ainsi que les méthodes traditionnelles utilisant ces données.

La section 4.1.1 présente les caractéristiques de description du visage que nous utilisons dans notre système : les vecteurs d'apparence issus des modèles actifs d'apparence AAMs [18]. Deux méthodes traditionnelles d'analyse des expressions faciales utilisent ce type de vecteurs d'apparence. Il s'agit des méthodes AAM et Differential-AAM que nous présenterons et comparerons dans la section 4.1.2. Ces deux méthodes seront par la suite comparées à la nouvelle méthode proposée (résultats et discussions dans le chapitre 6). Pour finir, nous indiquerons en quoi les caractéristiques AAM ne sont pas suffisantes pour l'analyse des expressions et justifierons l'utilisation des modèles spécifiques aux personnes (section 4.1.3).

4.1.1 Les Caractéristiques de Description des Visages Utilisées

Nous utilisons les paramètres d'apparence (forme et/ou texture) conformément aux modèles actifs d'apparence [18]. Les vecteurs d'apparence sont obtenus en utilisant une base de données de K images de P sujets réalisant différentes expressions faciales (voir figure 3.2).

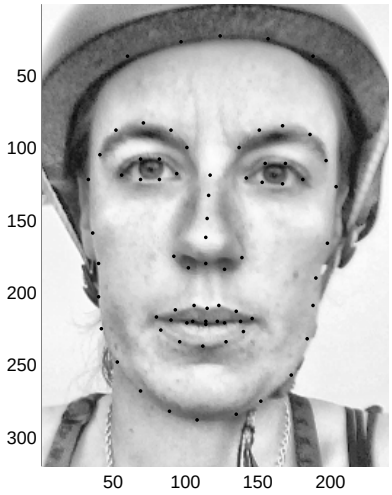


FIGURE 4.1 – 73 points caractéristiques des visages.

Les $K * P$ visages sont annotés par 73 points caractéristiques (par exemple, un point caractéristique est le coin gauche de la bouche), présentés dans la figure 4.1. Pour chaque image $\mathbf{i}$, les coordonnées de ces points caractéristiques sont concaténés pour former un vecteur $\mathbf{s_i}$, qui représente la forme du visage. Les intensités des pixels contenus dans la zone intérieure de la forme du visage moyen (évalué sur toute la base d'apprentissage) forment le vecteur $\mathbf{g_i}$, qui représente la texture. Pour détecter les distorsions de forme et de texture, une analyse en com-

posantes principales (ACP) est réalisée sur chacun des deux vecteurs :

$$\mathbf{s_i} = \overline{\mathbf{s}} + \mathbf{\Phi_s}.\mathbf{b_i^s} \tag{4.1}$$

$$\mathbf{g_i} = \overline{\mathbf{g}} + \mathbf{\Phi_t}.\mathbf{b_i^t} \tag{4.2}$$

où $\overline{\mathbf{s}}$ et $\overline{\mathbf{g}}$ sont les formes et texture moyennes, $\mathbf{\Phi_s}$ et $\mathbf{\Phi_t}$ les matrices formées par les vecteurs propres issus de l'ACP et $\mathbf{b_i^s}$ et $\mathbf{b_i^t}$ sont la décomposition de $\mathbf{s_i}$ et $\mathbf{g_i}$ sur les modes propres. $\mathbf{s_i}$ et $\mathbf{g_i}$ sont appelées vecteurs de forme et vecteurs de texture.

Pour prendre en compte la corrélation qui existe entre la forme et la texture, une troisième ACP est réalisée sur un vecteur qui concatène le vecteur de forme et le vecteur de texture $\mathbf{b_i} = [w_s.\mathbf{b_i^s}|\mathbf{b_i^t}]$ (w_s est un facteur de mise à l'échelle qui assure que les vecteurs de forme et de texture ont des variances comparables).

$$\mathbf{b_i} = \mathbf{\Phi}.\mathbf{c_i} \tag{4.3}$$

où $\mathbf{\Phi}$ est la matrice formée par les vecteurs propres de l'ACP, et $\mathbf{c_i}$ est le vecteur d'apparence.

Pour plus d'informations sur les modèles actifs d'apparence, se reporter à l'annexe A.

4.1.2 Méthode Basée sur les Vecteurs AAM (ABM) et Méthode Basée sur les Vecteurs AAM Différentiels (DABM)

Les méthodes classiques basées sur les vecteurs AAM utilisent des modèles actifs d'apparences génériques, c'est-à-dire que le modèle AAM est appris sur une base de données contenant différentes expressions de différents sujets. Dans les équations de la section précédente, $\mathbf{i}$ varie de 1 à $K * P$ ($\mathbf{i} = 1..KP$). Comme le modèle est générique, il est utilisé pour aligner et comparer les expressions de sujets inconnus. Par construction, les vecteurs d'apparence contiennent dans leurs premières composantes les déformations principales de forme et de texture. Ces déformations sont dues à la fois aux différences entre les expressions réalisées, mais aussi aux différences entre les identités des sujets de la base d'apprentissage. Dans la méthode basée sur les vecteurs AAM (*Appearance Based Method* notée **ABM**), les vecteurs d'apparence sont directement utilisés pour décrire les expressions, alors que dans la méthode basée sur les vecteurs AAM différentiels (*Differential Appearance Based Method* notée **DABM**), les vecteurs d'apparence sont alignés entre les sujets en soustrayant le vecteur d'apparence du visage neutre de la personne (voir figure 4.2).

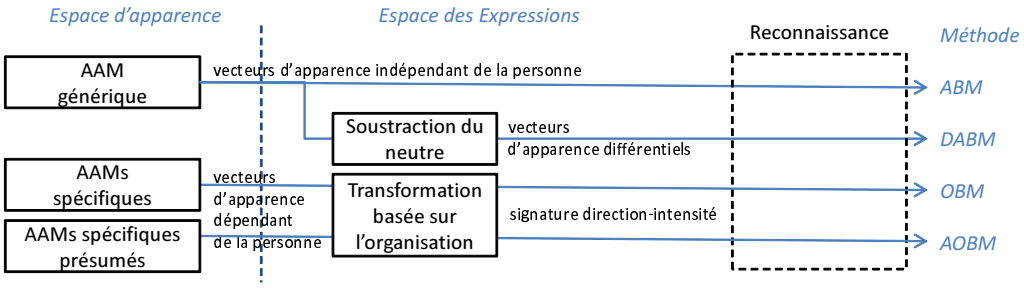


FIGURE 4.2 – Vue d'ensemble des méthodes comparées.

Comme nous le verrons dans le chapitre 6, la méthode ABM donne de faibles résultats pour la description d'une expression. Cela peut être justifié par le fait que les vecteurs d'apparence portent à la fois des informations d'identité et d'expression, si bien que la comparaison de ces vecteurs

sur des personnes différentes réalisant la même expression n'est pas pertinente. De meilleurs résultats sont obtenus avec la méthode DABM car une partie des différences d'identité entre les sujets a été soustrait. Néanmoins, dans les deux cas, le nombre de paramètres d'apparence optimal dépend beaucoup de la base de données d'apprentissage. Sur notre base de données (voir section 6), une vingtaine de paramètres d'apparence sont nécessaires pour conserver l'essentiel de l'information contenue dans les visages. Cheon et Kim [54] obtiennent les meilleurs résultats avec seulement 6 paramètres (leur base de données globale contient 16 personnes réalisant 4 expressions, la répartition pour la base d'apprentissage n'est pas fournie). Abboud et Davoine [48], quant à eux, obtiennent les meilleurs résultats en prenant les 20 premières composantes (leur base de données d'apprentissage contient 10 personnes réalisant 7 expressions). De plus, nous le verrons dans le chapitre 6, les résultats varient beaucoup selon le type de données utilisées (forme ou texture ou les deux en même temps). Ces méthodes ne permettent donc pas de caractériser une expression inconnue de façon unique.

Pour dépasser ces limites, nous proposons une méthode qui utilise des modèles actifs d'apparence spécifiques à la personne, c'est-à-dire qu'un modèle AAM est appris pour chaque sujet de la base de données à partir des différentes expressions réalisées par ce sujet (voir figure 4.2). Nous avons donc P modèles qui sont créés, où P est le nombre de sujets. Dans les équations de la section précédente, $\mathbf{i}$ varie alors de 1 à K ($\mathbf{i} = 1..K$) pour chacun des P modèles.

4.1.3 Pourquoi Utiliser des Modèles Spécifiques à la Personne ?

Il existe plus de variabilités de forme et de texture entre les visages neutres d'individus différents qu'entre les expressions d'un même individu. C'est ce que nous avons mis en évidence dans le graphique (a) de la figure 4.3. Nous observons que les vecteurs d'apparence de 8 expressions réalisées par deux sujets différents forment deux nuages de points distincts. Toutes les expressions d'un même individu sont dans la même zone de l'espace et chaque individu se trouve dans une zone différente. Un modèle générique indique donc des différences d'expressions mais aussi de morphologie (que ce soit sur la forme ou sur la texture). Bien que le graphique 4.3(a) ne montre que les deux premières composantes, correspondant aux deux principales déformations, cette constatation s'étend lorsque davantage de composantes sont prises en compte. Les résultats du paragraphe 6.3 confirment ce point.

Lorsque l'on soustrait les informations d'identité des sujets en soustrayant le vecteur du visage neutre, les deux zones se chevauchent. C'est ce qu'illustre le second graphique (figure 4.3(b)). Néanmoins, nous constatons que certaines expressions, pourtant similaires lorsque l'on observe les images, ne sont pas *proches* dans l'espace ainsi créé. C'est le cas par exemple de l'expression de sourire (représentée par un carré). Nous observons que le sourire (carré) du sujet B est plus proche de l'expression de tristesse (triangle) du sujet A et de l'expression de colère (étoile à 8 branches) du sujet A que de l'expression de sourire (carré) du sujet A. Ces composantes ne caractérisent donc pas uniquement l'expression. Prendre davantage de composantes (plus de deux) permet mieux distinguer les expressions. Néanmoins, comme nous le verrons dans le chapitre 6.3, la description obtenue conserve ne nombreux cas de confusion. L'alignement entre les différents sujets par soustraction des informations du visage neutre ne permet pas totalement de s'affranchir des différences de morphologies.

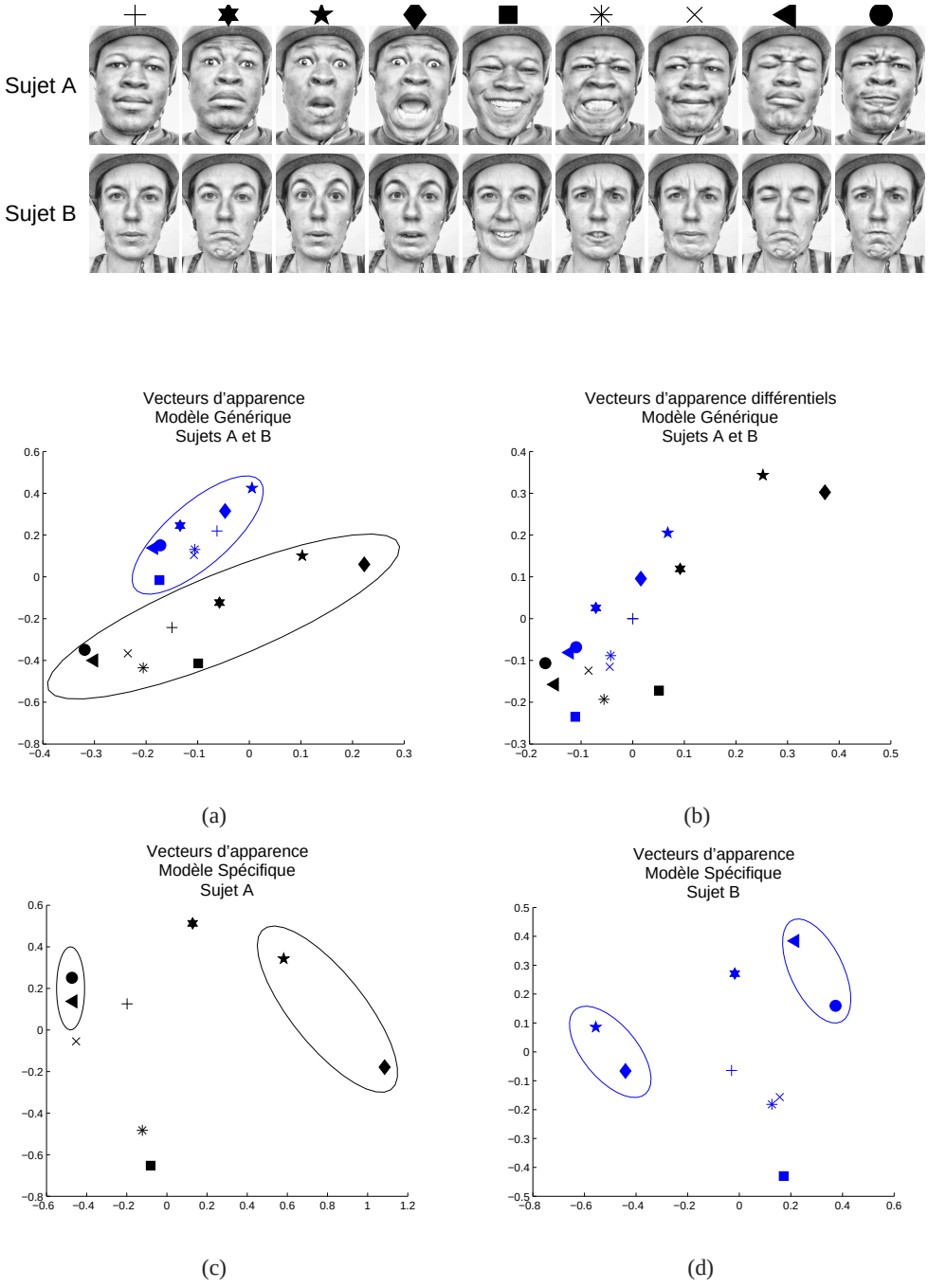


FIGURE 4.3 – Comparaison des vecteurs d'apparence de deux sujets A et B avec des modèles génériques et des modèles spécifiques. Affichage des deux premières dimensions de l'espace d'apparence, c'est-à-dire des 2 principales déformations faciales. 9 points correspondants aux 8 expressions plus le visage neutre.

Un modèle spécifique se focalise sur les déformations du visage d'un individu. Les premières composantes reflètent donc directement les variations de forme et de texture dues aux différentes expressions. C'est ce que montre les deux derniers graphiques de la figure 4.3. Nous pouvons

d'ailleurs constater que le visage neutre (représenté par un +) se situe plus ou moins au centre de ces déformations. Cela se comprend dans la mesure où les principales déformations du visage par rapport au neutre peuvent se réaliser dans les deux sens (par exemple, les yeux peuvent s'écarquiller ou se plisser, les lèvres se pincer ou s'ouvrir, ...). Nous utiliserons cette propriété dans la suite de notre étude.

L'utilisation de modèles spécifiques à la personne permet donc de se focaliser sur les déformations dues aux expressions. Par contre, il n'est pas possible de comparer directement les paramètres d'apparence issus des modèles spécifiques. Dans notre exemple, les composantes des graphiques (c) et (d) ne sont pas comparables. Même si les principales déformations sont souvent liées au sourire et au haussement des sourcils, ces composantes sont créées à partir de données différentes. Elles ne sont donc pas identiques et ne correspondent pas aux mêmes déformations du visage d'un individu à l'autre. A titre d'exemple, un premier individu peut avoir une déformation principale qui va être due au sourire et une secondaire au haussement de sourcils (son premier axe correspondra au sourire et son second au haussement de sourcils) alors qu'un autre individu pourra avoir une déformation principale correspondant au haussement de sourcils et une secondaire au sourire (son premier axe correspondra au haussement de sourcils et son second au sourire).

Pour pallier ce problème, nous n'allons pas comparer les valeurs des composantes d'apparence mais l'organisation des expressions entre elles. En effet, nous constatons que des expressions *proches*, c'est-à-dire ayant certaines caractéristiques de forme et de texture similaires, vont se situer dans la même zone de l'espace d'apparence. Sur la figure 4.3, les expressions représentées par l'étoile à 5 branches et le losange se situent dans la même zone, qui est différente de la zone contenant les expressions représentées par le triangle et le rond. Ainsi, une expression pourra se situer *entre* deux autres expressions (ou plus).

4.2 Définition de l'Organisation des Expressions d'un Sujet

Pour donner une idée de ce que nous avons appelé l'*organisation des expressions*, imaginez que les graphiques (c) et (d) de la figure 4.3 représentent des horloges. Posez alors l'aiguille sur le neutre (représenté par un +) et faites la tourner dans le sens des aiguilles d'une montre. Pour le sujet A, vous obtenez l'organisation des expressions suivante : étoile à 6 branches, étoile à 5 branches, losange, carré, étoile à 8 branches, croix, triangle, rond. Pour le sujet B, vous obtenez (à une inversion près) la même organisation en sens inverse : étoile à 6 branches, triangle, rond, croix, étoile à 8 branches, carré, losange, étoile à 5 branches.

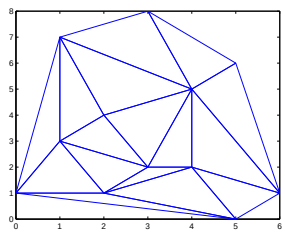


FIGURE 4.4 – Exemple de triangulation de Delaunay de 12 points en 2D.

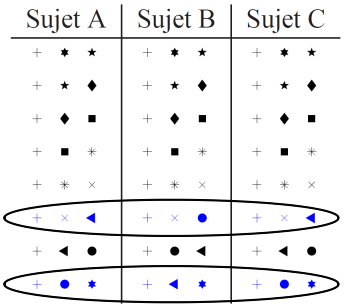
Dans cette section, nous allons présenter de façon plus mathématique cette *organisation des expressions*. Pour cartographier l'espace des expressions, nous utilisons la triangulation de Delau-

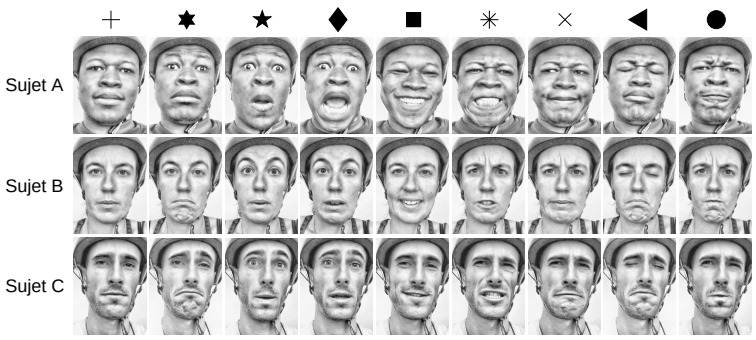
nay. La triangulation de Delaunay maximise l'angle minimum de tous les angles des triangles de la triangulation. Elle tend à éviter les angles fins et les angles ouverts (exemple sur la figure 4.4).

Avant de réaliser une triangulation de Delaunay sur les vecteurs des expressions, l'espace est recentré sur le visage neutre et les n premières composantes des paramètres d'apparence sont prises en compte et normalisées. Ainsi le neutre est au centre de l'espace et les K expressions se situent sur une hyper-sphère (exemple sur la figure 4.5). Une tessellation de Delaunay est effectuée sur ces K expressions plus le neutre. La figure 4.5 montre le résultat.

Nous avons constaté, au chapitre précédent (figures 4.3 (c) et (d)), que les expressions similaires de sujets différents ont des valeurs des composantes d'apparence très différentes. En revanche, lorsque nous regardons les simplexes (ici, en 2D, des triangles) issus des tessellations de Delaunay des sujets A et C, ceux-ci sont identiques.

Dans l'exemple de la figure, la liste des simplexes est la suivante :





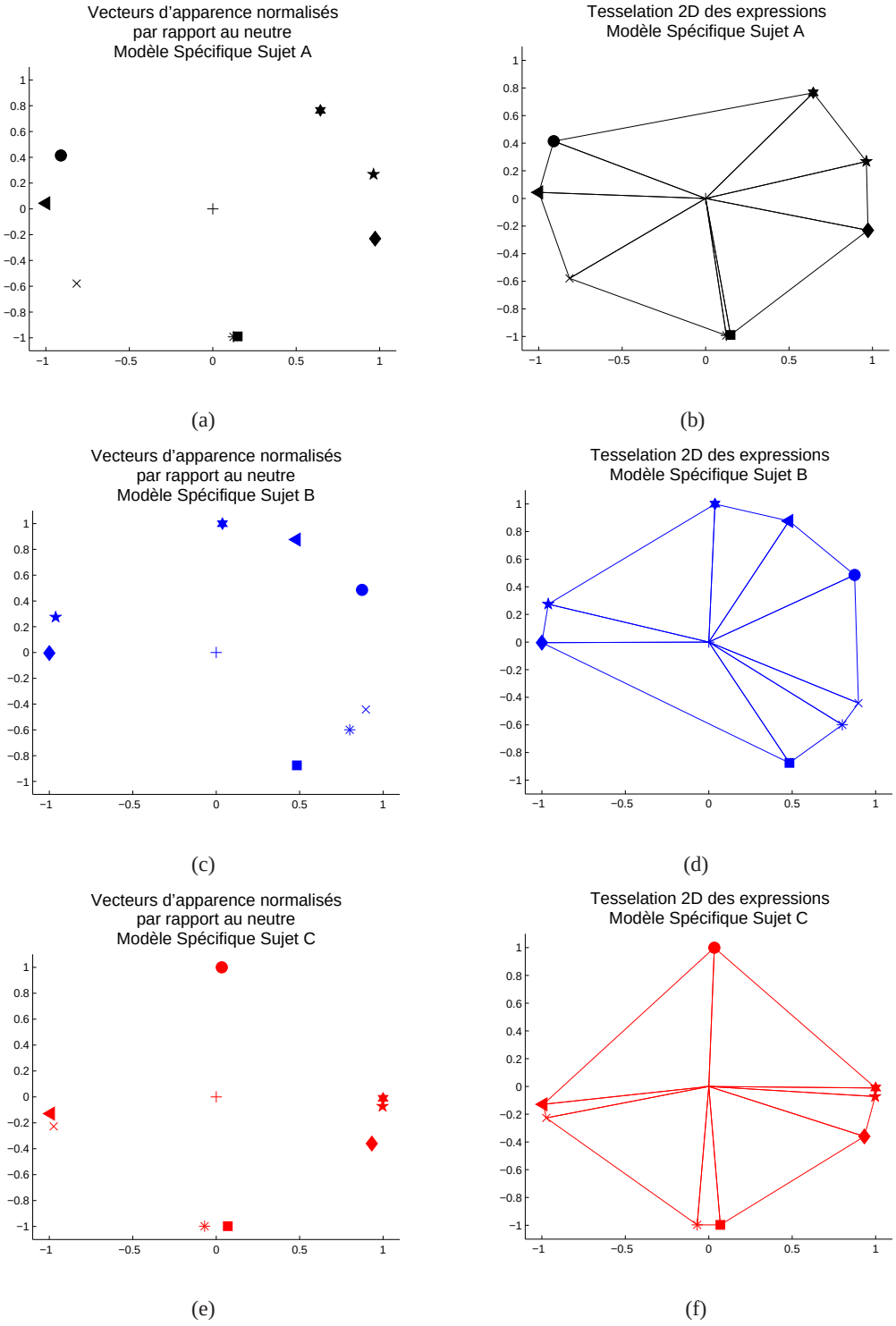


FIGURE 4.5 – Exemples de triangulations de Delaunay sur les sujets A, B et C avec $n = 2$ et $K = 8$. Affichage des deux premières dimensions de l'espace d'apparence normalisées, c'est-à-dire des 2 principales déformations faciales normalisées.

La figure 4.6 montre les résultats sur les 3 mêmes sujets en dimension 3, c'est-à-dire en prenant en compte les trois principales déformations issues des expressions ($n = 3$).

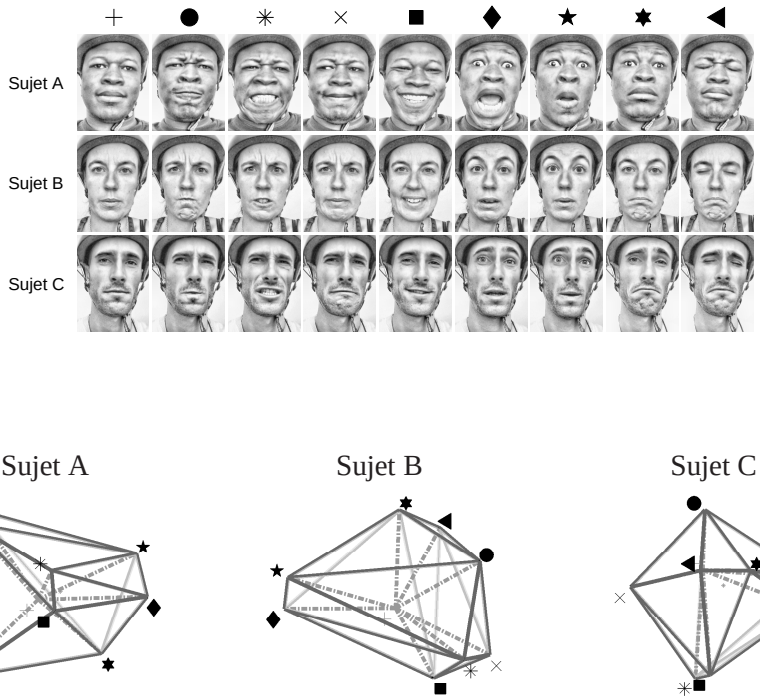
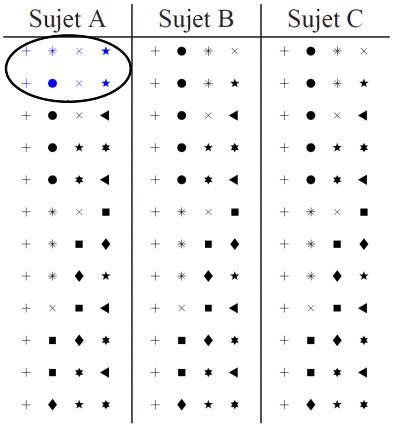


FIGURE 4.6 – Organisation de K = 8 expressions de 3 sujets. Le neutre est au centre et les 8 expressions sont sur une sphère.

Nous constatons là aussi que les valeurs des composantes d'apparence ne sont pas identiques entre les personnes. En revanche, lorsque nous regardons les simplexes (ici, en 3D, des tétraèdres) issus des tessellations de Delaunay, ceux-ci sont identiques pour les sujets B et C.

Dans l'exemple de la figure, la liste des simplexes est la suivante :



Cette propriété peut paraître assez surprenante. Elle peut s'expliquer par trois phénomènes. Le premier concerne le nombre de composantes principales permettant de caractériser les expressions. Dans les modèles spécifiques créés à partir de nos 8 expressions, quasiment 90% de l'énergie des déformations est atteinte à partir de 3 composantes. Ces 3 composantes suffisent donc pour qualifier efficacement un visage expressif pour chaque personne. La figure 4.7 illustre ce point.

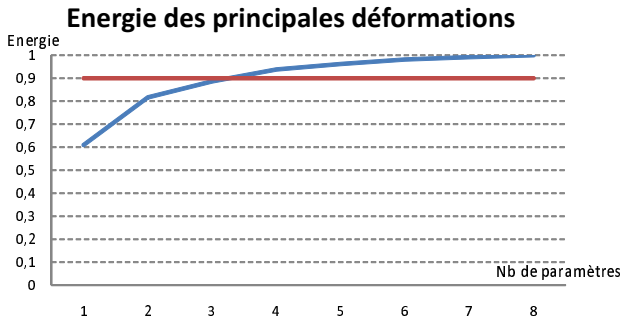


FIGURE 4.7 – Énergie des déformations pour un modèle spécifique créé à partir de 8 expressions plus neutre. Quasiment 90% de l'énergie est obtenue à partir des 3 principales déformations.

Le second phénomène est du au fait que les expressions sont issues de l'activation et du relâchement des muscles faciaux, ce qui implique une certaine cohérence entre les différents individus. Pour finir, les 8 expressions prises en compte ont certaines déformations communes ou proches, même si elles ne sont pas exprimées avec la même intensité. Par exemple, l'expression de dégoût et l'expression de colère vont toutes deux faire apparaitre un froncement des sourcils et donc une direction semblable liée à cette déformation. Ainsi, même si la façon de froncer les sourcils est différente entre deux individus, chaque individu fronce les sourcils de la même façon à chaque expression réalisée. L'ordonnancement des expressions entre les individus est ainsi identique.

Néanmoins, il arrive que pour certains sujets, l'organisation des expressions soit légèrement différente. Les figures 4.5 et 4.6 donnent des exemples. Le sujet B pour la dimension 2 et le sujet A pour la dimension 3 ont des organisations légèrement différentes. Afin d'estimer cette différence, nous avons défini un indice de similarité entre les organisations de deux individus.

4.3 Indice de Similarité entre deux Organisations

Comparer des tessellations n'est pas évident [94]. Les méthodes basées sur les caractéristiques permettant d'aligner des maillages utilisent les distances entre les points. Or, nous avons vu précédemment que la distance entre les vecteurs des expressions n'était pas pertinente (dans la mesure où nous utilisons des modèles spécifiques). D'autres mesures telles que le calcul de la distance *edit distance* [95] ou encore le calcul du sous graphe le plus grand [96] ne prennent pas en compte l'organisation spécifique (points sur une sphère) de notre structure.

Pour comparer les organisations des expressions entre 2 individus, nous avons comparé la liste des expressions connectées dans les tessellations de Delaunay de dimension n. Pour la tessellation

de Delaunay en dimension $n = 3$ du sujet i, la liste des simplexes est la suivante (où $+$ est le visage neutre) :

$$DT_i = \{\{ + \ * \ \times \ \star \}, \{ + \ \bullet \ \times \ \star \}, \ldots\}$$

La liste des expressions connectées est :

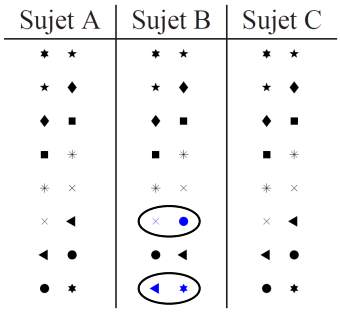
$$DT_i = \{\{ \bullet \ * \}, \{ \bullet \ \times \}, \{ * \ \times \}, \{ \bullet \ \star \}, \{ * \ \star \}, \ldots\}$$

Nous avons utilisé l'indice de similarité de Sorensen [97] afin de comparer les ensembles d'expressions connectés de deux individus. Chaque connexion a une force identique de façon à ne pas prendre en compte la distance entre les vecteurs d'apparence. L'indice de similarité de Sorensen est défini par :

$$Q(S_i, S_j) = \frac{2.|S_i \cap S_j|}{|S_i| + |S_j|} \tag{4.4}$$

où $|S_i|$ est le nombre de connections de la n-tessellation de Delaunay du sujet i et $|S_i \cap S_j|$ est le nombre de connections qu'ont en commun les n-tessellation de Delaunay des sujets i et j. Le facteur 2 permet d'avoir un index entre 0 et 1.

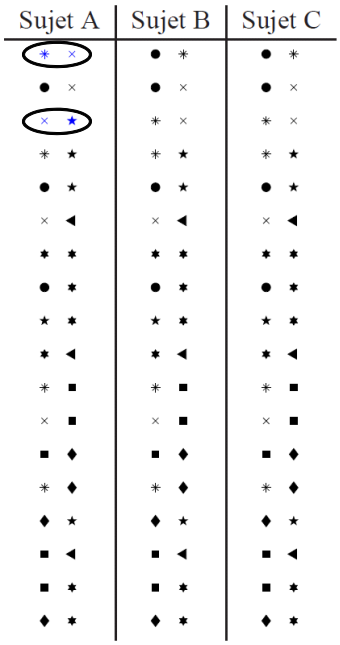
Dans l'exemple de la figure 4.5, les ensembles des expressions connectées sont :



L'indice de similarité entre A et C vaut 1, celui entre A et B vaut 0.75 et celui entre B et C vaut aussi 0.75 :

$$Q(S_A, S_B) = \frac{2 * 6}{8 + 8} = 0.75 \tag{4.5}$$

Dans l'exemple de la figure 4.6, les ensembles des expressions connectées sont :



L'indice de similarité entre B et C vaut 1, celui entre A et B vaut 0.94 et celui entre A et C vaut aussi 0.94 :

$$Q(S_A, S_B) = \frac{2 * 17}{18 + 18} \simeq 0.94 \tag{4.6}$$

4.4 Le Caractère Universel de l'Organisation des Expressions

Nous définissons l'indice de similarité d'une organisation avec les $P - 1$ autres organisations par la valeur moyenne des indices de similarité :

$$Q(S_i) = \frac{1}{P-1} \sum_{k=1, k \neq i}^{P} Q(S_i, S_k) \tag{4.7}$$

L'organisation universelle des expressions est alors définie par l'organisation S_s dont l'indice de similarité est le plus élevé :

$$Q(S_s) = \max_{i=1..P} Q(S_i) \tag{4.8}$$

Nous avons réalisé une expérience sur 8 expressions affichées par 17 sujets ($K = 8$, $P = 17$). La figure 4.8 montre 14 de ces 17 sujets (les 3 autres n'ayant pas souhaité que leur photo soit publiée). La figure 4.11 illustre les 8 expressions sur ces sujets. Les origines différentes (Afrique, Caucase), les attributs (barbe, lunettes, ...) ainsi que les différences d'âges (sujets âgés de 20 à 55 ans) des sujets permettent d'avoir une grande variabilité en terme de forme et de texture des visages, accentuant ainsi la difficulté de trouver un critère commun pour analyser les expressions faciales. Une description plus complète de la base de données ainsi que des expressions est réalisée dans la section 6.1.

FIGURE 4.8 – Visages neutres de 14 sujets ayant servis à l'extraction de l'organisation des expressions.

Nous avons utilisé une tessellation de Delaunay en 3D ($n = 3$) pour notre étude. Le nombre de connexions entre les expressions augmente avec n. $n = 2$ donne peu d'expressions connectées. $n > 3$ donne un grand nombre de connexions avec à la limite (pour $n = 8$) toutes les expressions entièrement connectées entre elles. Choisir $n = 3$ permet d'avoir suffisamment de connexions entre les expressions sans que celles-ci soient toutes connectées (environ 18 connexions parmi 28 possibles). Sur ces données, nous avons extrait l'organisation universelle des expressions et calculé l'indice de similarité entre cette organisation universelle et l'organisation de chacun des 17 sujets.

La figure 4.9 montre en gris la distribution de ces 17 indices de similarité ($Q(S_i, S_s)$, $i = 1..17$). L'indice de similarité se situe entre 0.82 et 1. Les différences entre l'organisation S_i des expressions d'un individu et l'organisation S_s universelle sont principalement dues à la substitution d'une connexion. Dans ce cas (cf. figure 4.10(a)), les configurations gardent le même voisinage.

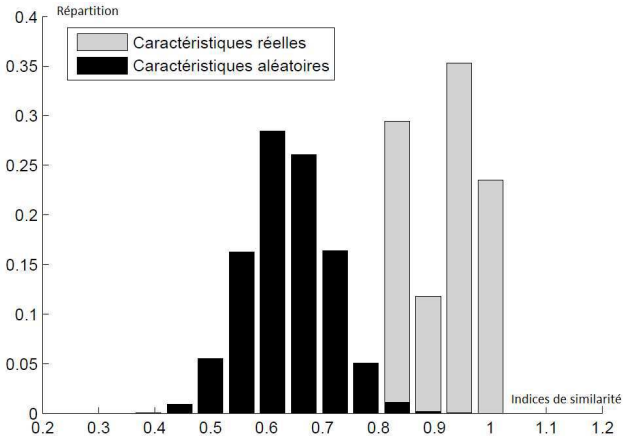
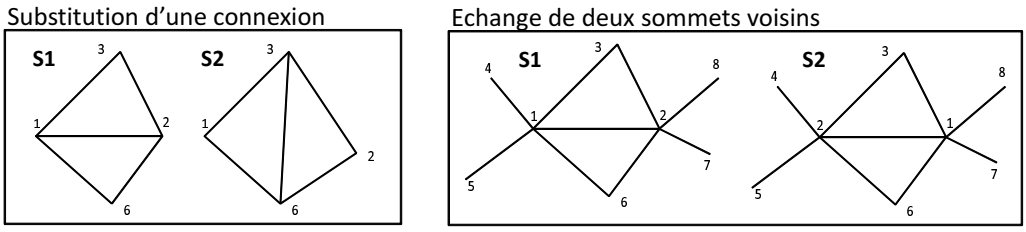


FIGURE 4.9 – Distribution des indices de similarité entre les organisations spécifiques aux personnes et l'organisation universelle extraite des données. En gris, les organisations de 17 sujets réels. En noir, les organisations de paramètres aléatoires (pour comparaison).

A titre de comparaison, une transposition de deux sommets voisins ayant chacun 5 voisins (voir figure 4.10(b)) aurait donné un indice de 0.78. Les organisations ayant un indice compris entre 0.8 et 1 peuvent donc être considérées comme similaires à S_s.



(a) (b)

FIGURE 4.10 – Pertinence de l'indice utilisé : exemple d'indice de similarité entre deux configurations S_1 et S_2. (a) : une substitution d'un bord ($Q(S_1, S_2) = 0.94$). (b) : une transposition de deux sommets voisins ($Q(S_1, S_2) = 0.78$).

Nous avons aussi calculé la distribution de l'indice de similarité de 10 000 organisations aléatoires issues de 8 expressions. Cette distribution est représentée en noir sur la figure 4.9. Nous constatons que tous les indices de similarité des organisations réelles (en gris) sont dans les 1.5% des indices de similarité les plus élevés des organisations aléatoires. Les organisations aléatoires sont donc très différentes des organisations réelles. De plus, les organisations réelles sont similaires à l'organisation S_s (indice de similarité supérieur à 0.8). Nous pouvons donc considérer que l'organisation S_s est une organisation *universelle*, c'est-à-dire indépendante de la personne, ou du moins indépendante pour nos 17 sujets de test.

FIGURE 4.11 – Visages neutres et 8 expressions de 14 sujets ayant servis à l'extraction de l'organisation des expressions.

Chapitre 5

Espace des Expressions : une Expression est Définie par sa Position Relative par rapport aux Autres Expressions

Cette nuit, en regardant le ciel, je suis arrivé à la conclusion qu'il y a beaucoup plus d'étoiles qu'on en a besoin.

Quino - *Extrait de la bande dessinée Mafalda*

Imaginez-vous en train de regarder un ciel étoilé. Une étoile particulière vous intrigue et vous souhaitez en parler avec votre ami. Mais comment lui faire comprendre de quelle étoile il s'agit ? Peut-être pointerez-vous le doigt dans la direction de cette étoile en indiquant : *Tu vois, dans cette direction, il y a trois étoiles qui brillent plus que les autres. Elles forment un triangle avec la pointe en bas. Et bien, l'étoile dont je veux te parler, c'est celle qui est dans ce triangle, assez proche de l'étoile qui forme la pointe en bas et un peu sur la droite.*

En d'autres termes, vous localisez des objets de forte intensité dans une direction donnée et ensuite, vous donnez la position relative de votre objet par rapport à ces objets de référence. Nous allons faire de même avec les expressions. Nous allons définir la direction d'une expression par rapport à sa position relative par rapport à d'autres expressions, de forte intensité et connues du système. En plus de sa direction, nous indiquerons aussi l'intensité de cette expression.

Dans ce chapitre, nous expliciterons tout d'abord la variété (au sens mathématique du terme) des expressions, créée à partir de l'organisation des expressions identifiée dans le chapitre précédent. Nous définirons ensuite le formalisme utilisé pour caractériser de façon unique une nouvelle expression. Pour finir, nous proposerons une visualisation sur une carte 2D de la signature d'expressions inconnues.

Sommaire

5.1 La Variété des Expressions

5.1.1 Définition de la variété

Dans l'espace d'apparence spécifique au sujet, les 8 expressions connues plus le visage neutre utilisées pour définir l'organisation universelle S_s des expressions forment une variété linéaire par morceaux DT_s. En dimension 2 ($n = 2$), la variété est une collection de triangles connectés, chaque triangle étant formé par les paramètres du visage neutre et de 2 expressions connues (voir figure 5.1). En dimension 3 ($n = 3$), la variété est une collection de tétraèdres connectés, chaque tétraèdre étant formé par les paramètres du visage neutre et de 3 expressions connues. De façon plus générale, en dimension n, la variété est une collection de n-simplexes, chaque n-simplexe étant formé par le visage neutre et de n autres expressions connues. Ces simplexes sont connectés les uns aux autres par un de leur bord, c'est-à-dire par un simplexe de dimension $n - 1$ (visage neutre et $n - 1$ expressions). Cette variété est conforme à l'organisation universelle des expressions, puisqu'elle est composée des 8 expressions connues plus du neutre ayant permis de définir l'organisation universelle.

Notez la différence entre la dimensionnalité de la variété et la dimensionnalité de l'espace d'apparence. Les vecteurs d'apparence ont K composantes (où K est le nombre d'expressions utilisées pour la création du modèle d'apparence). La dimensionnalité de la variété est notée n. Nous avons $2 \leq n \leq K$.

5.1.2 Projection d'une expression sur la variété

Une expression nouvelle (appelée *expression inconnue* ou *expression non prototypique*), elle, est caractérisée par son vecteur d'apparence dans l'espace d'apparence spécifique au sujet (voir annexe A.3 pour le calcul du vecteur d'apparence). Elle possède donc K composantes. Elle peut être approximée par un point sur la variété linéaire par morceau décrite précédemment. Cette approximation est réalisée par projection du vecteur d'apparence sur la variété adaptée à la personne définie précédemment. Comme la variété est linéaire par morceau, chaque simplexe se comporte localement comme un espace linéaire. Nous utilisons donc la projection orthogonale sur cet espace.

La figure 5.1 montre un exemple de cette projection pour une variété de dimension 2. Le vecteur d'apparence est représenté dans l'espace d'apparence par le point *expression inconnue*. Sa projection sur la variété formée par l'organisation des expressions est appelée *approximation* et est représentée par une croix sur la figure.

L'algorithme est le suivant :

Algorithme 1 Calcul de la projection sur la variété

for chaque simplexe **do**
  Calcul de la projection orthogonale de l'expression sur le simplexe
end for
Sélection de la projection la plus petite

A noter que les simplexes sont limités dans l'espace. Ainsi, la projection orthogonale sur le simplexe peut être située en dehors des limites du simplexe. Sur la figure 5.2, la projection orthogonale du point en rouge sur le triangle formé par le plus, la croix et l'étoile à huit branches est situé en dehors du triangle (croix rouge). Nous calculons alors la projection orthogonale sur

chacun des $n+1$ simplexes de dimension $n-1$ qui composent chaque simplexe de dimension n et réitérons l'opération jusqu'à ce que la projection se fasse à l'intérieur de la variété. Dans l'exemple de la figure 5.2, la projection sur la variété est la projection sur le segment formé par la croix et l'étoile à huit branches (croix bleue).

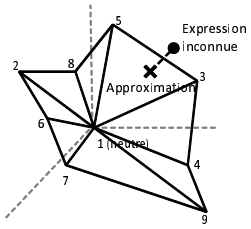


FIGURE 5.1 – Exemple de la variété des expressions avec une tessellation de Delaunay en dimension 2 ($K=8$, $n=2$). Chaque expression possède $K=8$ composantes. Représentation dans un espace de dimension 3 (trois premières composantes).

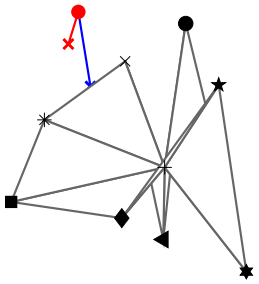


FIGURE 5.2 – Projection orthogonale sur la variété.

5.2 La Signature d'une Expression

Les expressions similaires sont organisées de la même façon d'une personne à une autre, si bien qu'une expression peut être définie par sa position relative par rapport aux autres expressions. Nous utilisons cette propriété pour associer une signature indépendante de la personne à une nouvelle expression.

5.2.1 Définition de la signature

La signature d'une expression est définie selon deux caractéristiques : la direction et l'intensité. Ces deux caractéristiques sont calculées par rapport aux valeurs des 8 expressions connues formant l'organisation des expressions.

Le visage neutre étant au centre de la variété, la direction est donnée par l'inclinaison entre le neutre et l'expression ; et l'intensité correspond à la distance entre le neutre et l'expression (voir figures 5.3 et 5.4).

Plus précisément, ces coordonnées sont calculées de la façon suivante :

- la direction est donnée par les coordonnées barycentriques du point d'intersection entre le vecteur *neutre / expression* et la surface extérieure de la variété,
- l'intensité est donnée par la norme du vecteur entre le neutre et l'expression, normalisée de façon à ce que la surface de la variété ait une intensité de 1.

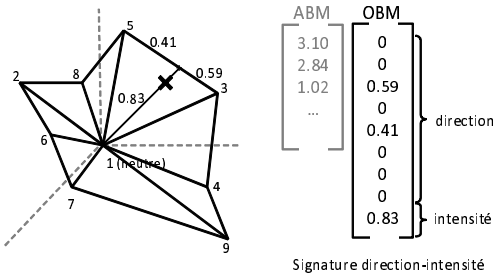


FIGURE 5.3 – Transformation des vecteurs d'apparence (ABM) en une signature direction-intensité (OBM) ($K = 8$, $n = 2$).

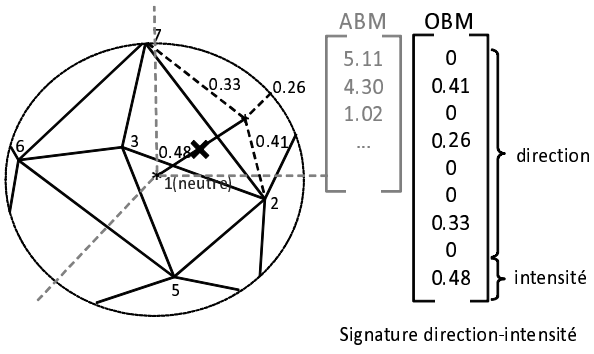


FIGURE 5.4 – Transformation des vecteurs d'apparence (ABM) en une signature direction-intensité (OBM) ($K = 8$, $n = 3$).

5.2.2 Exemples en dimensions 2 et 3

Dans l'exemple de la figure 5.3, l'expression est projetée sur la surface extérieure de la variété. Elle tombe dans le segment formé par les expressions 3-5. Les coordonnées barycentriques de la projection dans ce segment sont 0.59-0.41, c'est-à-dire que

$$0.59 * (Expr_{Surface} - Expr_3) + 0.41 * (Expr_{Surface} - Exp_5) = 0 \tag{5.1}$$

où $Expr_{Surface}$ correspond aux coordonnées de la nouvelle expression projetée sur la surface extérieur de la variété et $Expr_i$ aux coordonnées de l'expression i formant la variété.

La direction est donc le vecteur de 8 composantes [0 0 0.59 0 0.41 0 0 0].

L'intensité est de 0.83 car l'expression $Expr_{Approximation}$ est située à 0.83 de la surface de la variété :

$$Expr_{Neutre} - Expr_{Approximation} = 0.83(Expr_{Neutre} - Expr_{Surface}) \tag{5.2}$$

Dans l'exemple de la figure 5.4, l'expression est projetée sur la surface extérieure de la variété. Elle tombe dans le triangle formé par les expressions 2-4-7 (l'expression 4 n'est pas visible sur le schéma, elle se trouve de l'autre côté de la sphère). Les coordonnées barycentriques de la projection dans ce triangle sont 0.41-0.26-0.33, c'est-à-dire que

$$0.41*(Expr_{Surface}-Expr_2)+0.26*(Expr_{Surface}-Exp_4)+0.33*(Expr_{Surface}-Exp_7) = 0 \tag{5.3}$$

où $Expr_{Surface}$ correspond aux coordonnées de la nouvelle expression projetée sur la surface extérieur de la variété et $Expr_i$ aux coordonnées de l'expression i formant la variété.

La direction est donc le vecteur de 8 composantes [0 0.41 0 0.26 0 0 0.33 0].

L'intensité est de 0.48 car l'expression $Expr_{Approximation}$ est située à 0.48 de la surface de la variété :

$$Expr_{Neutre} - Expr_{Approximation} = 0.48(Expr_{Neutre} - Expr_{Surface}) \tag{5.4}$$

De façon plus générale, pour chaque expression, la direction est donnée par K composantes formées par les n coordonnées barycentriques et $K - n$ coordonnées nulles (où K est le nombre d'expressions connues formant la variété et n la dimension de la variété). L'intensité est donnée par une composante. La signature direction-intensité est alors un vecteur de $K + 1$ éléments.

5.2.3 Les signatures sur une carte 2D

Afin de pouvoir comparer visuellement les signatures d'expressions inconnues similaires de sujets différents, nous avons réalisé une carte 2D de la variété 3D. La surface extérieure de la variété a été dépliée. La figure 5.5 montre cette carte. Le bord droit et le bord gauche sont identiques, il s'agit de la même arrête de la variété lorsque celle-ci est repliée. De même, les bords haut droit et haut gauche sont identiques ainsi que les bords bas droit et bas gauche. Les expressions connues sont représentées par des émoticônes, caractérisant l'expression que les sujets devaient reproduire (voir la description de la base de données dans la section 6.1).

La figure montre aussi les signatures de 8 expressions inconnues similaires de 4 sujets. Chaque type d'expression est donné par un motif différent ($+, \square, \lozenge, ...$). L'intensité est donnée par la taille du motif. Lorsque l'intensité est très faible, nous n'avons pas affiché l'expression car sa direction

n'a alors plus de signification. Nous constatons que les signatures des expressions similaires se retrouvent dans les mêmes zones de l'espace. Elles peuvent donc être utilisées pour définir de façon unique une expression.

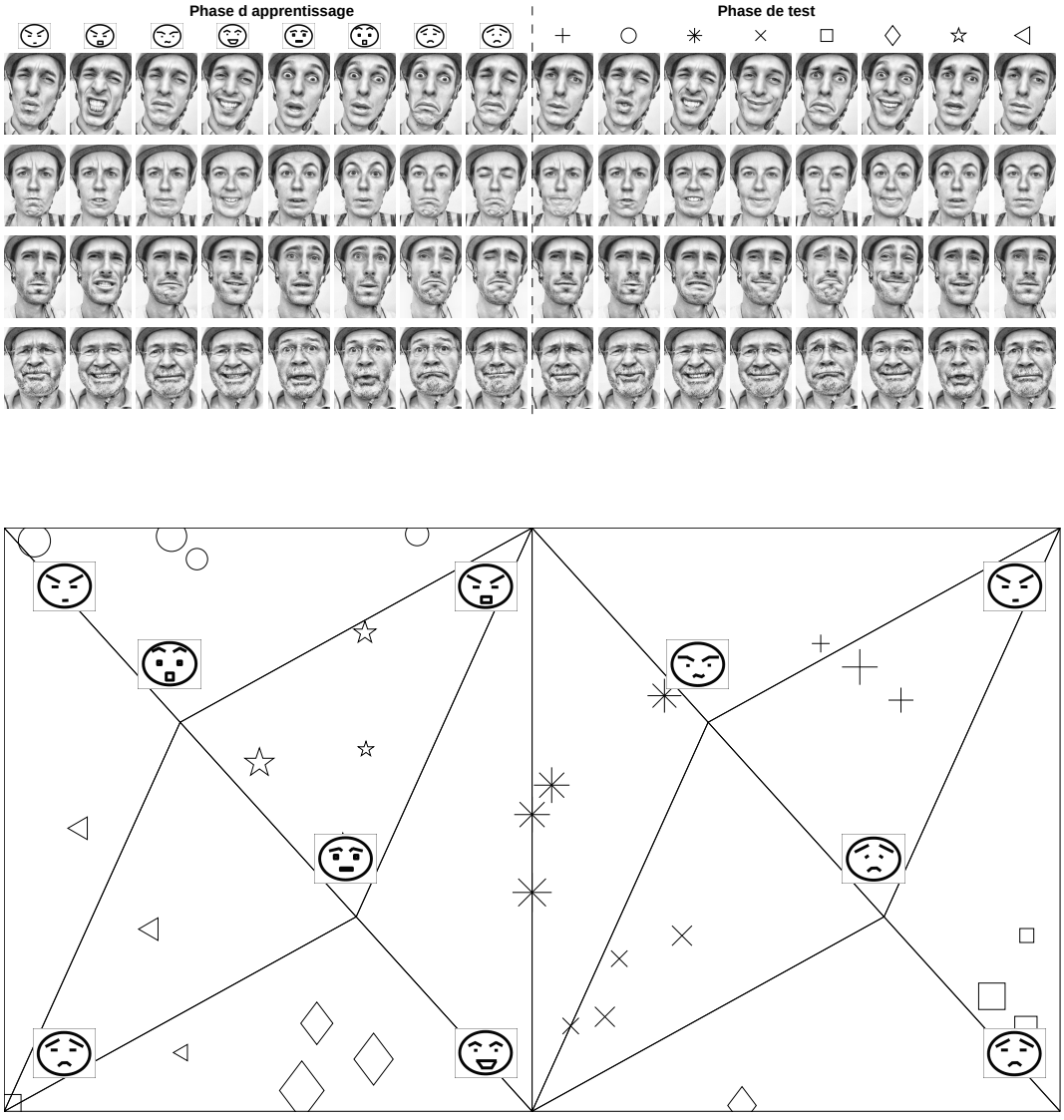


FIGURE 5.5 – Signatures de 8 expressions inconnues similaires de 4 sujets. Variété de dimension 3 ($n = 3$) représentée sur une carte 2D.

5.2.4 Reconnaissance par un algorithme de vote

Nous avons vu sur la figure 5.5 que des expressions similaires de personnes différentes avaient des signatures proches. Nous proposons dans ce paragraphe un algorithme permettant de vérifier cette constatation en effectuant la reconnaissance des expressions non incluses dans les bases d'apprentissage.

La méthode est basée sur le calcul du plus proche voisin couplé avec un algorithme de vote.

Algorithme 2 Pertinence de la signature d'une expression non incluse dans les bases d'apprentissage

Require: Signatures de M expressions non incluses dans les bases d'apprentissage pour P sujets

for chaque expression e non incluse dans les bases d'apprentissage ($e = 1..M$) **do**
  for chaque sujet $i, i = 1..Pe$ **do**
    for chaque sujet $j, j = 1..P, j \neq i$ **do**
      Calcul de l'expression $p_{i,j}$ du sujet j qui est la plus proche de l'expression e du sujet i (algorithme du plus proche voisin)
    end for
    Calcul de l'expression p_i la plus fréquente des $p_{i,j}$ lorsque j varie (algorithme de vote)
  end for
  Calcul du taux de reconnaissance de l'expression e : $\frac{\mathbf{1}}{\mathbf{Pe}} * \mathbf{sum_{i=1}^{P} e|p_i = e|}$.
end for

Dans l'algorithme de plus proche voisin, la norme 1 a été utilisée. Des tests ont aussi été réalisés en norme 2 ; les résultats ne sont que peu impactés. Nous avons aussi testé la pondération de l'intensité par rapport à la direction en ajoutant un poids à cette composante ; les meilleurs résultats sont obtenus pour un poids de 1. A noter que seules les expressions considérées comme correctement réalisées par les sujets sont prises en compte pour les tests. C'est pourquoi nous avons introduit Pe ($Pe <= P$) qui détermine la liste des sujets pour laquelle l'expression e est réalisée correctement.

La figure 5.6 illustre l'algorithme pour l'expression 2.

Nous verrons dans le chapitre 6 une étude chiffrée sur l'utilisation de cette signature direction-intensité. Ce chapitre 6 propose des résultats expérimentaux de reconnaissance d'expressions inconnues. Il montre que l'utilisation de la signature d'une expression telle que définie précédemment améliore les performances de reconnaissance par rapport aux méthodes traditionnelles basées sur l'apparence, tout en diminuant le nombre de caractéristiques nécessaires (dimensionnalité de l'espace). Par ailleurs, ce chapitre met aussi en évidence la robustesse de l'espace des expressions, selon les différents types de paramètres d'apparences (forme et/ou texture).

Avant de présenter ces résultats, nous allons étendre la méthode précédente aux personnes inconnues du système.

5.3 Et sur une Personne Inconnue...

Ce chapitre étend l'analyse effectuée précédemment aux sujets inconnus. Il propose une méthode permettant de créer l'espace d'apparence spécifique à la personne et conforme à l'organisation des expressions. Cet espace est créé sans phase d'apprentissage préalable du sujet. Ce chapitre présente aussi comment adapter le calcul de la signature d'une expression à cet espace.

La méthode d'analyse d'une expression sur un sujet connu (voir chapitre précédent) utilisait K expressions connues du sujet pour créer l'espace d'apparence spécifique du sujet. Nous devons maintenant nous affranchir de cette donnée. Nous allons donc créer artificiellement ces K expressions. Nous appellerons ces expressions créées artificiellement, des expressions *plausibles* du sujet.

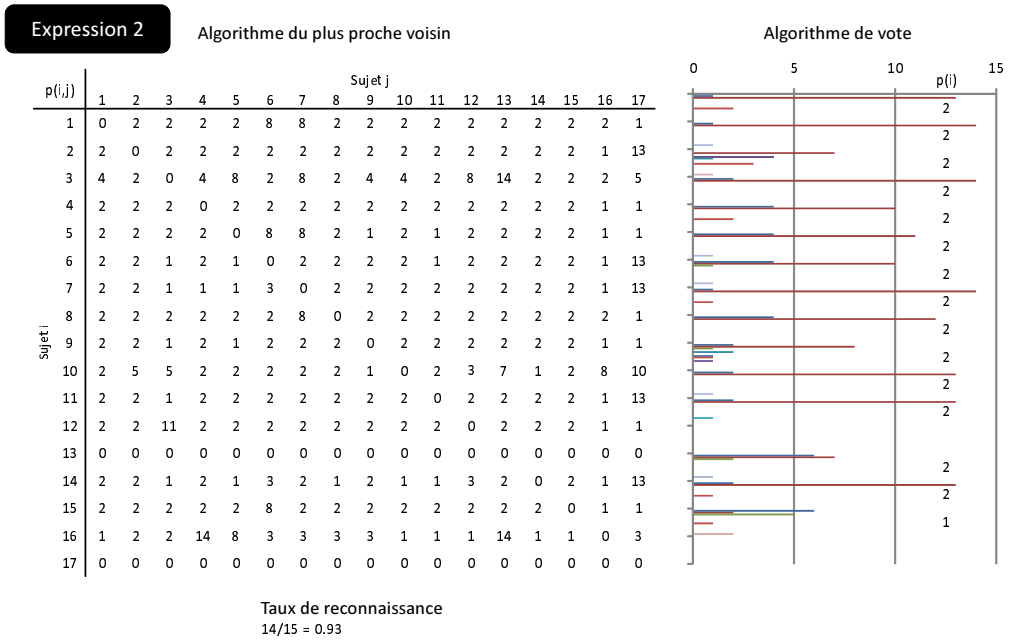


FIGURE 5.6 – Illustration de l'algorithme de calcul de la pertinence de la signature d'une expression non incluse dans les bases d'apprentissage. Cas de l'expression $e = 2$ avec $M = 14$ expressions inconnues et $P = 17$ sujets dont $Pe = 15$ sujets ont réalisé l'expression e correctement.

Dans cette section, nous supposons que le visage neutre du sujet est connu. Une méthode permettant de trouver automatiquement le visage neutre du sujet est proposée dans la section 5.3.5.

L'originalité de l'approche consiste donc à créer artificiellement les K expressions *connues* du sujet (c'est-à-dire les K expressions utiles pour la création de la variété). Cette création est réalisée en appliquant des déformations plausibles correspondant à l'expression *souhaitée* sur le visage neutre du sujet inconnu. Les déformations plausibles sont apprises sur des sujets connus. Nous introduisons dans un premier temps la méthode permettant d'apprendre les déformations plausibles d'une expression souhaitée et leur application sur un sujet inconnu, avant de présenter la méthode de création de l'espace présumé et le calcul de la signature dans cet espace. Nous proposerons aussi, dans la section 5.3.4, une méthode simplifiée pour réaliser ce traitement.

5.3.1 Déformations plausibles appliquées sur le visage neutre

Nous avons choisi d'utiliser une fonction de warping affine par morceau pour créer les expressions plausibles d'un sujet inconnu ([98]). Ce type de warping n'est pas adapté pour réaliser la synthèse d'expressions réalistes mais est suffisamment efficace pour calculer la direction d'une déformation faciale. Il a aussi l'avantage d'être plus rapide que des warping plus réalistes ([99]). Park et Kim [100] l'ont utilisé pour amplifier les déformations des expressions faciales afin de pouvoir reconnaitre des expressions de faible intensité.

Cette méthode a pour avantage de garder les spécificités morphologiques des sujets. Le principe consiste à apprendre les déformations sur des sujets connus et à les appliquer sur des sujets

inconnus. L'une des approches du warping linéaire par morceau consiste à partitionner l'enveloppe convexe des points correspondants à la forme du visage en utilisant une triangulation (voir figure 5.8). La déformation de la forme du visage causée par une expression est alors décrite relativement à cette triangulation.

La figure 5.7 montre 3 formes de visage plausibles correspondant à la même expression : le visage neutre (a) et une expression (b) sont affichés pour 3 sujets connus. Les déformations entre le visage neutre et l'expression sont calculées pour chacun de ces trois sujets, et sont appliquées sur le visage neutre du sujet inconnu (c).

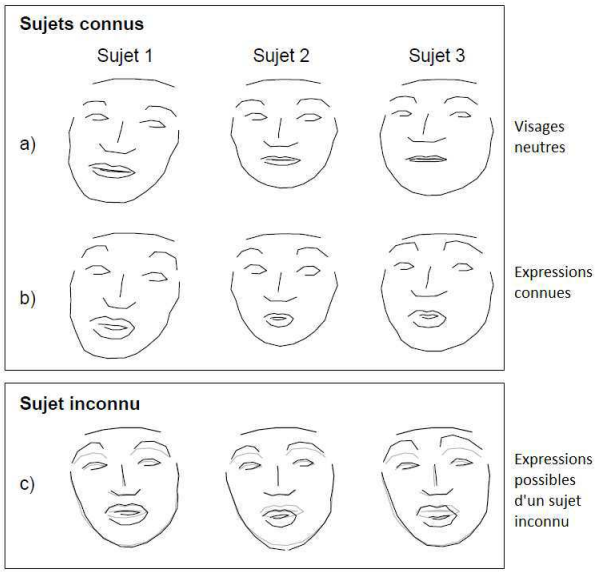


FIGURE 5.7 – Les déformations possibles sont apprises sur des sujets connus. (a) Visage neutre de 3 sujets connus. (b) Expression similaire de 3 sujets connus. (c) Expressions possibles (noir) et visage neutre (gris pointillé) d'un sujet inconnu.

Pour des raisons de simplicité, nous explicitons tout d'abord comment est calculée une déformation plausible apprise sur un seul sujet, puis nous étendons la méthode à une déformation moyenne apprise sur P sujets. Pour finir, nous indiquerons comment ces déformations sont utilisées pour créer des expressions plausibles sur des sujets inconnus. Nous utilisons ces deux types de calculs (déformations apprises sur un sujet et déformations apprises sur P sujets) dans notre méthode. Nous en expliciterons les raisons dans les autres sections de ce chapitre.

Calcul d'une déformation plausible apprise sur un sujet connu La forme du visage est décrite par m points de contrôle (cf. figure 5.8). Pour apprendre la déformation due à une expression sur un sujet connu, nous prenons chacun des points de contrôle $\mathbf{x_{expr},}_i$ de l'expression $\mathbf{I_{expr}}$ du sujet connu et détectons sa position sur le visage neutre $\mathbf{I_{neutre}}$ de ce sujet (voir figure 5.8) ; c'est-à-dire que nous détectons quel triangle du visage neutre contient le point de contrôle i de l'expression. Supposons que $\mathbf{x_{neutre,1}}$, $\mathbf{x_{neutre,2}}$ et $\mathbf{x_{neutre,3}}$ sont les trois sommets d'un tel triangle, le point de contrôle $\mathbf{x_{expr},}_i$ de l'expression peut être écrit par :

$$\mathbf{x_{expr},}_i = \alpha \mathbf{x_{neutre,1}} + \beta \mathbf{x_{neutre,2}} + \gamma \mathbf{x_{neutre,3}} \tag{5.5}$$

L'ensemble des coefficients α, β, γ et des triangles associés donne les caractéristiques de la déformation plausible.

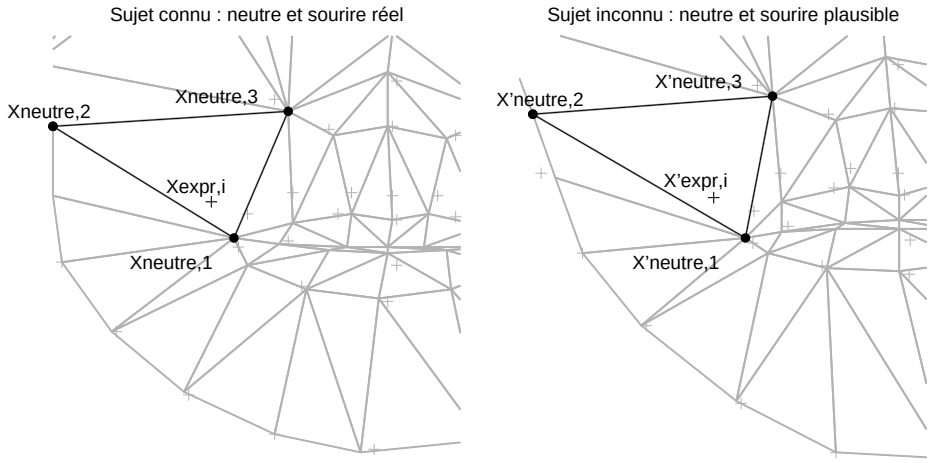


FIGURE 5.8 – Sur la gauche, la triangulation du visage neutre $\mathbf{I_{neutre}}$ et les points de contrôle de l'expression $\mathbf{I_{expr}}$ du sujet connu. Sur la droite, la triangulation du visage neutre $\mathbf{I'_{neutre}}$ et les points de contrôles warpés de l'expression $\mathbf{I'_{expr}}$ du sujet inconnu.

A noter que le warping linéaire par morceau classique est contraint à l'enveloppe convexe des points de contrôle. Or, il est possible que l'expression sorte de cette enveloppe convexe. C'est par exemple le cas des points de contrôle de la mâchoire lorsque la mâchoire est ouverte. Ils sont *plus bas* que ceux du visage neutre. Pour pallier ce comportement, les coefficients négatifs sont pris en compte dans l'algorithme. La relation $\alpha + \beta + \gamma = 1$ est conservée et le triangle choisi est celui qui minimise la somme des distances entre α, β, γ et $[0, 1]$. La figure 5.9 montre un exemple de calcul. Le triangle le plus proche est à une distance de $(1.07 - 1) + 0.19 + 0 = 0.26$.

Calcul d'une déformation plausible moyenne apprise sur plusieurs sujets connus Dans le cas où nous souhaitons apprendre une déformation à partir de plusieurs sujets connus, nous étendons l'algorithme précédent en prenant les moyennes sur les P sujets des coefficients α, β et γ. La relation $\bar{\alpha} + \bar{\beta} + \bar{\gamma} = 1$ est de fait gardée et nous déterminons le triangle qui minimise la distance entre $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ et $[0, 1]$.

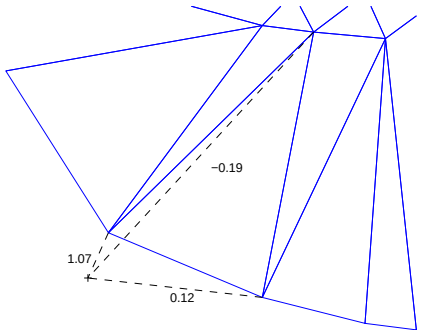


FIGURE 5.9 – Cas des déformations en dehors de l'enveloppe convexe des points.

Algorithme 3 Calcul d'une déformation moyenne

```
for chaque expression do
  for chaque point caractéristique do
    for chaque triangle do
      for chaque sujet connu do
        Calcul de α, β, γ
      end for
      Calcul de la moyenne ᾱ, β̄, γ̄
    end for
    Calcul du triangle qui minimise la distance entre ᾱ, β̄, γ̄ et [0, 1]
  end for
end for
```

Application d'une déformation plausible sur un sujet inconnu Qu'il s'agisse d'une déformation apprise sur un seul sujet ou d'une déformation moyenne apprise sur plusieurs sujets, la méthode d'application de la déformation sur le sujet inconnu reste la même.

Dans le cas d'une déformation apprise sur un seul sujet, les m points de contrôle $\{\mathbf{x}_{\mathbf{neutre},i}\}$ du visage neutre $\mathbf{I}_{\mathbf{neutre}}$ du sujet connu correspondent aux m points de contrôle $\{\mathbf{x'}_{\mathbf{neutre},i}\}$ du visage neutre $\mathbf{I'}_{\mathbf{neutre}}$ du sujet inconnu (voir figure 5.8). Nous utilisons les coefficients α, β et γ trouvés précédemment pour trouver le point équivalent $\mathbf{x'}_{\mathbf{expr}}$ de l'expression $\mathbf{I'}_{\mathbf{expr}}$ du sujet inconnu. Dans le cas d'une déformation apprise sur plusieurs sujets, il est plus difficile de représenter le mapping sous forme de figure mais le traitement est identique : nous utilisons les coefficients $\bar{\alpha}$, $\bar{\beta}$ et $\bar{\gamma}$ trouvés précédemment pour trouver le point équivalent $\mathbf{x'}_{\mathbf{expr}}$ de l'expression $\mathbf{I'}_{\mathbf{expr}}$ du sujet inconnu.

$$\mathbf{x'}_{\mathbf{expr},i} = \alpha \mathbf{x'}_{\mathbf{neutre,1}} + \beta \mathbf{x'}_{\mathbf{neutre,2}} + \gamma \mathbf{x'}_{\mathbf{neutre,3}} \tag{5.6}$$

s'il s'agit d'une déformation apprise sur un seul sujet, ou

$$\mathbf{x'}_{\mathbf{expr},i} = \bar{\alpha} \mathbf{x'}_{\mathbf{neutre,1}} + \bar{\beta} \mathbf{x'}_{\mathbf{neutre,2}} + \bar{\gamma} \mathbf{x'}_{\mathbf{neutre,3}} \tag{5.7}$$

s'il s'agit d'une déformation apprise sur plusieurs sujets.

5.3.2 Espace d'apparence présumé d'un sujet inconnu

Pour créer l'espace spécifique à la personne, nous utilisons la déformation plausible moyenne des K expressions formant la variété (voir figure 5.10(a)) et non les P déformations plausibles issues de chaque sujet pour chacune des K expressions. Nous avons fait ce choix afin de garder un espace restreint en termes de dimensions. L'espace présumé possède alors une dimensionnalité correspondant au nombre d'expressions (donc K). Le nombre de personnes P servant à l'apprentissage n'entre pas en compte. Si nous avions pris l'ensemble des déformations plausible issues de la base d'apprentissage, nous aurions obtenu un espace de dimension $P * K$, plus important et dépendant de la base d'apprentissage. A noter que nous aurions pu réduire la dimensionnalité de cet espace en ne conservant que les K premières composantes. Néanmoins, ces composantes auraient reflété en partie les différences de pattern de déformations inter-individu.

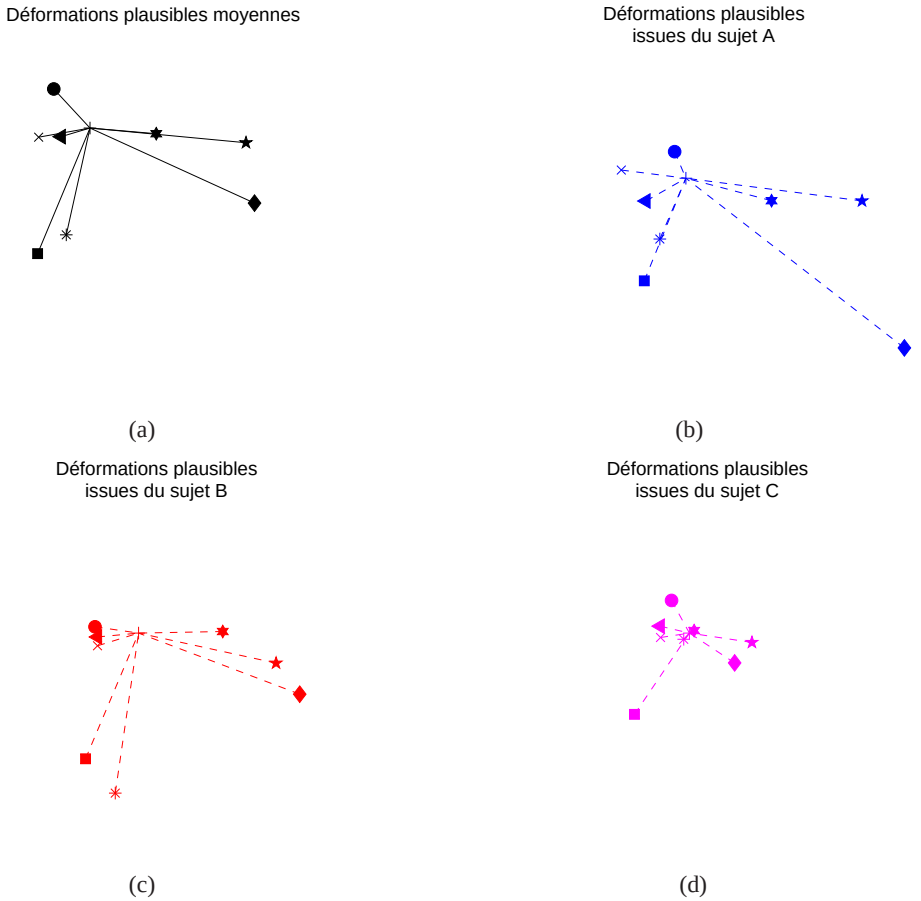


FIGURE 5.10 – (a) Espace présumé créé à partir des K déformations moyennes apprises sur P sujets ($K = 8$, $P = 16$). (b) à (d) Projection des K expressions plausibles issues de 3 sujets sur l'espace présumé. Affichage des deux premières dimensions.

Le modèle d'apparence spécifique à la personne est créé en 3 étapes :

- pour chaque expression e ($e = 1..K$) de la variété des expressions, la déformation moyenne plausible de l'expression est apprise à partir de P sujets connus

- pour chaque expression e, la déformation moyenne est appliquée sur le visage neutre du sujet inconnu et donne l'expression plausible moyenne du sujet
- l'espace d'apparence est calculé à partir de ces K expressions plausibles moyennes, les K vecteurs propres sont conservés

5.3.3 Calcul de la signature d'une expression dans le modèle présumé

Ce paragraphe propose une méthode permettant d'adapter le calcul de la signature à l'espace d'apparence présumé trouvé précédemment.

Nous utilisons ici, pour chacune des K expressions, les P déformations plausibles issues chacune d'un des P sujets connus (et non la déformation moyenne). Cela permet de définir une zone de déformations plausibles pour chacune des K expressions ne sachant pas quelle est la déformation réelle du sujet. Ainsi, pour chacun des sujets connus, K expressions plausibles sont calculées, chacune correspondant à l'une des K expressions de l'organisation des expressions (voir figures 5.10(b) à (d)). Ces K expressions plausibles forment une variété conforme à l'organisation des expressions. Nous avons donc P variétés plausibles, chacune créée à partir des déformations d'un sujet connu. Nous calculons P signatures direction-intensité conformément à l'algorithme du paragraphe 5.2. Nous définissons la signature finale d'une expression comme la moyenne de ces P signatures.

La signature direction-intensité est donc calculée en 3 étapes :

- P variétés sont calculées à partir des K expressions plausibles, une variété pour chacun des P sujets connus
- la signature direction-intensité est calculée comme dans la section 5.2 pour chacune de ces P variétés
- la valeur moyenne des P signatures direction-intensité est calculée.

Les figures 5.10(b) à (d) donnent trois exemples d'organisation plausibles des expressions. Nous constatons que les déformations plausibles apprises sur différents sujets ont des directions relativement proches. Par exemple, la déformation correspondant à l'expression représentée par un carré a une direction (pour ses deux premières composantes) en bas à droite, qui est similaire pour les trois sujets A, B et C. A noter que les variations d'intensité peuvent être importantes. C'est le cas de la déformation correspondant à l'expression représentée par un losange. C'est pour cette raison que nous parlons d'espace *présumé* et de déformations plausibles. Nous n'avons pas de connaissance sur la déformation (direction et intensité maximale) des expressions réelles de la personne inconnue.

5.3.4 Modèle présumé simplifié

Nous pouvons simplifier le modèle pour un sujet inconnu en ne prenant en compte qu'un seul sujet connu ou alors le sujet moyen. Ainsi, le modèle présumé est directement calculé à partir des expressions plausibles issues de ce sujet. Pour l'étape de calcul de la signature, une seule variété plausible est disponible, comme dans le cas d'un sujet connu. Cette étape se réduit donc directement au calcul de la signature direction-intensité conformément à la procédure d'un sujet connu.

5.3.5 Détection automatique du neutre

Lorsque nous avons à disposition des séquences vidéo suffisamment longues et variées des personnes inconnues, il est possible d'extraire efficacement le visage neutre de la personne inconnue en calculant la valeur moyenne des paramètres d'apparence sur un modèle d'apparence générique. La figure 5.11 présente la répartition des vecteurs d'apparence (2 première composantes) des séquences audio-vidéo variées d'un seul sujet. La croix rouge indique la valeur moyenne.

Nous constatons une forte concentration d'expressions au centre du nuage de points. Ces expressions correspondent au visage neutre de la personne. En effet, pour une séquence vidéo suffisamment longue, les visages neutres vont être extrêmement fréquents.

Nous constatons quelques expressions à la périphérie du nuage de points, relativement bien réparties au niveau des directions. Nous avons vu au paragraphe 4.1.3 que le vecteur d'apparence du visage neutre est situé au milieu des vecteurs d'apparence des autres expressions, lorsque des expressions très différentes sont réalisées. Dans notre exemple (figure 5.11), les expressions à la périphérie sont les expressions de forte intensité. En prenant en compte de nombreuses séquences audio-vidéo variées, quasiment l'ensemble des expressions sont réalisées. Les expressions de forte intensité ont donc peu d'impact sur le calcul de la moyenne des vecteurs d'apparence pour deux raisons : elles sont peu fréquentes et elles se compensent.

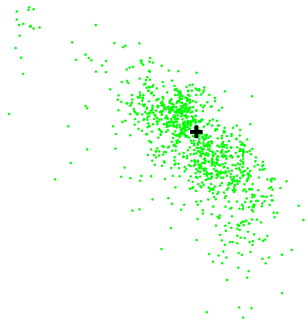


FIGURE 5.11 – Répartitions des visages expressifs sur une séquence audio-vidéo lorsque le sujet ne parle pas. La croix marque la position du neutre obtenue par le calcul de la moyenne des vecteurs d'apparence.

En synthèse, le calcul de la valeur moyenne des vecteurs d'apparence est influencé par deux facteurs :

- la position du visage neutre par rapport aux autres expressions (nécessité d'expressions variées qui se compensent)
- la fréquence d'apparition du visage neutre dans la séquence vidéo (nécessité de longues séquences audio-vidéo permettant d'avoir beaucoup de visages neutres)

Lorsque ces conditions sont réunies, nous considérons que le visage neutre est l'image qui a les paramètres d'apparence les plus proches de cette valeur moyenne.

Nous avons réalisé une expérimentation sur les données issues de la base de données publique AVEC2012 [80] qui réunie ces conditions. La figure 5.12 montre quelques exemples de visages neutres trouvés avec cette méthode.

A noter que nous avons supprimé de la séquence vidéo les passages pendant lesquels le sujet était en train de parler. En effet, lorsque le sujet parle, les déformations faciales dues à la parole sont importantes, fréquentes et nous savons de fait qu'il ne s'agit pas d'un visage neutre. Cela impacte le nuage de points et le rend plus diffus. La figure 5.13 donne un exemple. Il s'agit du même sujet et des mêmes séquences audio-vidéo que ceux de la figure 5.11. Pour information, dans le cadre du challenge AVEC 2012, nous avons utilisé les transcriptions de paroles qui étaient fournies par les organisateurs du challenge et qui étaient horodatées pour déterminer les moments où le sujet était en train de parler des moments où il était muet.

FIGURE 5.12 – Visages neutres extraits de séquences vidéo.

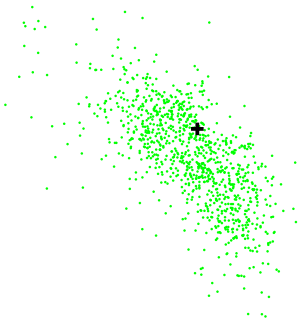


FIGURE 5.13 – Répartitions des visages expressifs sur une séquence audio-vidéo. Les images pendant le temps de parole du sujet sont conservées. La croix marque la position du neutre obtenue par le calcul de la moyenne des vecteurs d'apparence.

Chapitre 6

Reconnaissance d'une expression par signature intensité-direction

Ce qui est admirable, ce n'est pas que le champ des étoiles soit si vaste, c'est que l'homme l'ait mesuré.

Anatole France

Ce chapitre a pour objectif de tester la validité de l'espace des expressions créé et de la signature obtenue. Pour cela, nous allons principalement utiliser l'algorithme défini précédemment en 5.2.4 et nous comparerons notre méthode aux méthodes traditionnelles basées sur les paramètres d'apparence (voir section 4.1).

Nous présenterons dans ce chapitre les données utilisées. Puis nous analyserons la robustesse de la représentation des expressions proposée (représentation par organisation des expressions) selon 2 axes : la dimensionnalité de la variété et le type de données du visage (forme et/ou texture). Nous étudierons aussi les résultats selon qu'il s'agit d'un sujet connu ou d'un sujet inconnu. Pour finir, nous analyserons les cas de confusion.

Sommaire

Le processus global est présenté sur la figure 6.1.

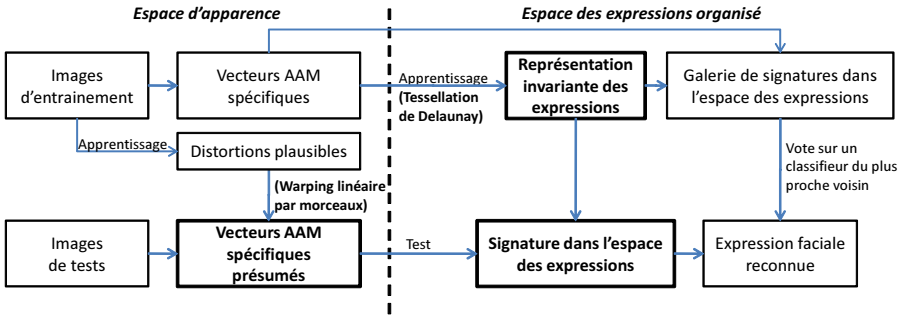


FIGURE 6.1 – Vision globale du processus de description d'une expression faciale.

6.1 Les Données : 14 Expressions Mélangées et Inconnues de 17 Sujets Connus

Les bases de données publiques existantes n'ont pas suffisamment d'expressions similaires différentes pour tester notre méthode. Pour cette raison, les expérimentations ont été réalisées sur une base de données spécifique accessible à l'adresse suivante

`http://www.rennes.supelec.fr/immemo/`.

Cette base de données consiste en 22 expressions similaires jouées par 17 sujets âgés de 20 à 55 ans, dont 30% de femmes. La plupart des sujets sont Caucasiens, certains portent des lunettes ou ont une barbe (voir figure 6.2). Ces 22 expressions sont issues de travaux réalisés par Nicolas Stoiber [50] qui a identifié automatiquement 25 expressions incluant les principales déformations du visage d'un sujet réalisant des expressions émotionnelles variées. Seules 22 expressions ont été conservées pour nos travaux car les 3 autres déformations étaient rarement réalisées correctement par les sujets. Il s'agit notamment d'expressions asymétriques telles que des expressions nécessitant la levée d'un seul sourcil, expressions que peu de sujet arrivent à produire. Le caractère automatique de l'analyse de Nicolas Stoiber fait que les expressions identifiées ne sont pas des expressions *basiques* au sens d'Ekman [37], mais des expressions mélangées avec des intensités variables.

FIGURE 6.2 – Visages neutres de 14 des 17 sujets.

La base de données a été séparée en deux parties : pour chaque sujet 8 expressions ont été utilisées pour l'apprentissage et les 14 autres expressions ont été utilisées pour les tests. La figure 6.3 montre cette répartition sur 3 sujets.

Le processus d'acquisition est le suivant : 22 photos de visages expressifs ont été montrées aux sujets et les sujets avaient pour consigne de reproduire le plus fidèlement possible ces expressions. Par ailleurs, une indication précisant la nature de l'expression était fournie pour certaines expressions (par exemple : *il s'agit d'une expression de joie et de surprise simultanée*). L'ordre des consignes était toujours le même et une photo était prise au moment de l'expression (il ne s'agit pas de vidéo). Ces 22 expressions sont des expressions émotionnelles mélangées avec des intensités variables (voir figure 6.4). Les différences entre les expressions sont dues à l'intensité de l'expression (par exemple une intensité croissante de joie pour les expressions 5 - D - 6) ou bien elles sont dues au mélange d'émotions (par exemple l'expression 7 correspond à un mélange de joie et de surprise).

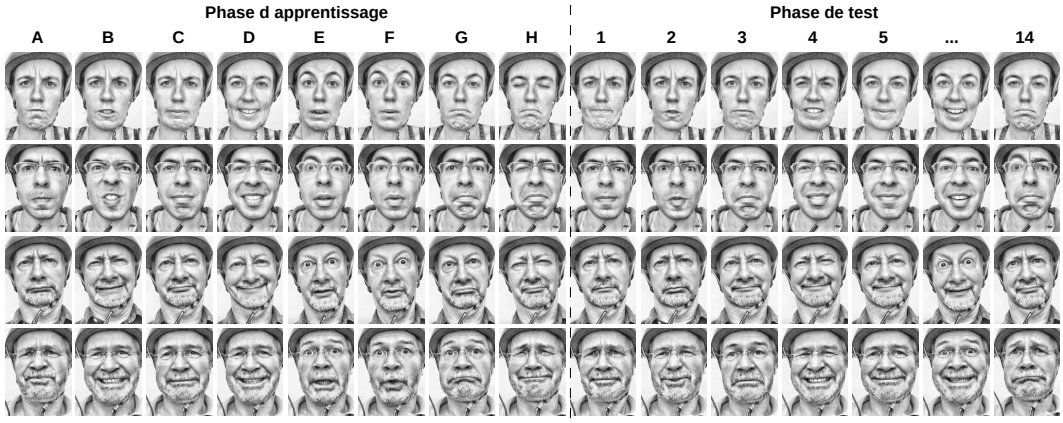


FIGURE 6.3 – Expressions similaires réalisées par différents sujets. 8 expressions ont été utilisées pour l'apprentissage et 14 pour les tests.

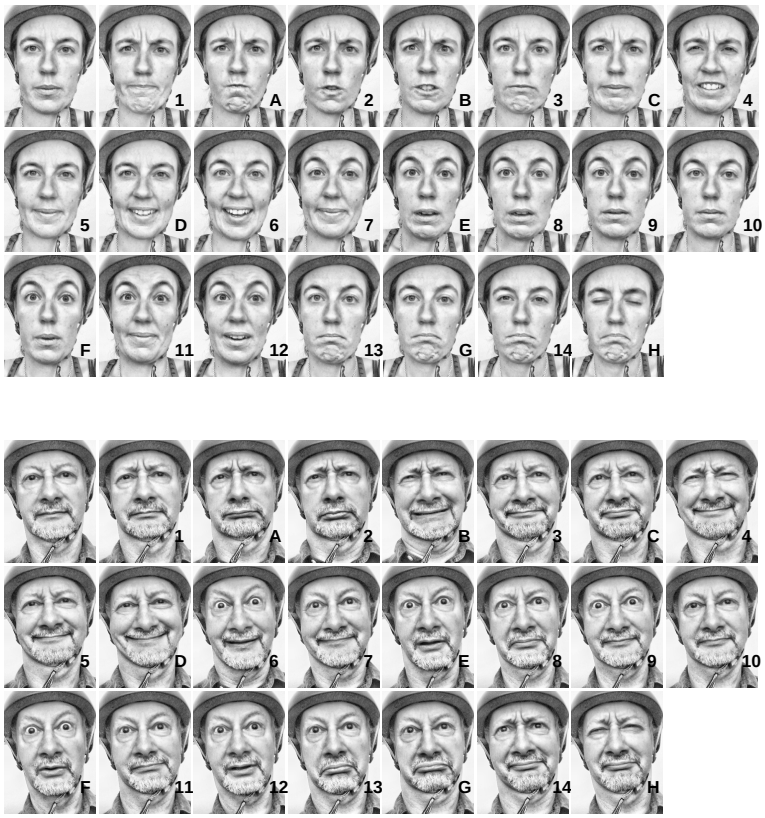


FIGURE 6.4 – Exemple des 22 expressions plus visage neutre réalisées par deux sujets.

Même si les consignes étaient les mêmes pour l'ensemble des sujets, nous avons constaté que certaines expressions sensées être similaires étaient réalisées différemment. Par exemple, certains sujets ont la bouche ouverte alors que d'autres non pour la même expression (voir expression D

sur la figure 6.3). De plus, certains sujets ont eu des difficultés à réaliser certaines expressions, si bien que certaines expressions sont très proches les unes des autres (voir les expressions 7 et 11 ou E et 8 du premier sujet sur la figure 6.4). Lorsque les expressions de certains sujets étaient très éloignées de la consigne, ces expressions ont été retirées des tests de reconnaissance. Ce traitement a été réalisé lors de la capture des expressions. La personne chargée de prendre les photos notait aussi si l'expression avait ou pas été correctement réalisée par le sujet. Cette information a donc un caractère subjectif.

6.2 Robustesse de la Représentation

6.2.1 Robustesse selon la dimensionnalité de la variété

Cette section présente les résultats de reconnaissance de la méthode basée organisation (OBM) sur 14 expressions inconnues de sujets connus et analyse l'impact de la dimensionnalité de la variété sur ces résultats. Pour chacun des sujets, les 8 expressions connues plus le neutre permettent de calculer les modèles d'apparence spécifiques aux sujets. La figure 6.5 montre les taux moyens de reconnaissance sur les 14 expressions inconnues, selon la dimensionnalité de la variété et la figure 6.6 donne le taux de reconnaissance moyen. Nous remarquons que la dimensionnalité de la variété influence peu les résultats de reconnaissance. En revanche, les résultats varient beaucoup selon l'expression, avec des taux de reconnaissance allant de moins de 20% pour l'expression 9 à plus de 90% pour les expressions 2 et 6. Cela s'explique par le fait que les expressions 8 et 9 sont très peu différentiables, il s'agit dans les deux cas d'une expression de surprise ; alors que les expressions 2 et 6 n'ont pas d'expression correspondant à la même émotion dans la base de test : l'expression 2 est liée à la colère et l'expression 6 représente un mélange de joie et de surprise.

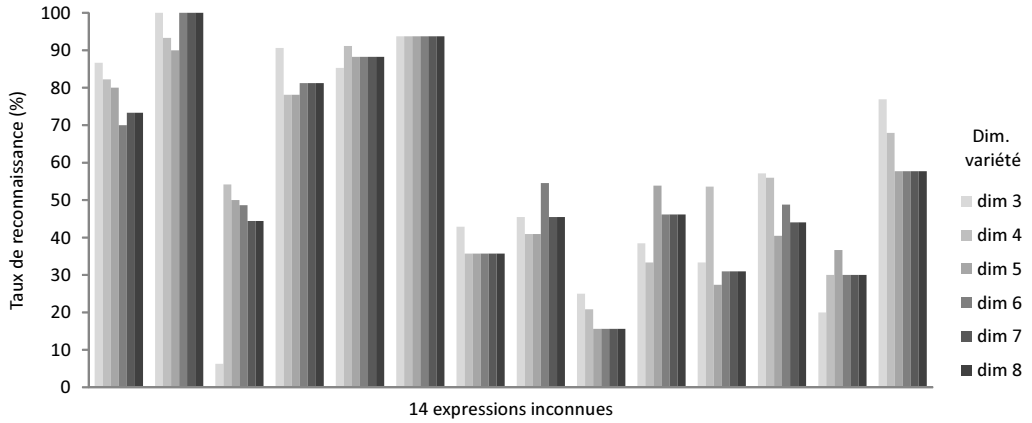


FIGURE 6.5 – Taux de reconnaissance de 14 expressions inconnues sur des sujets connus avec la méthode OBM, utilisant les caractéristiques de forme et de texture du visage. Données calculées selon la dimension de la variété.

Ces résultats montrent que 4 composantes non nulles (dimension de la variété égale à 3 et intensité - voir figure 5.4) sont suffisantes pour caractériser efficacement une expression avec cette méthode.

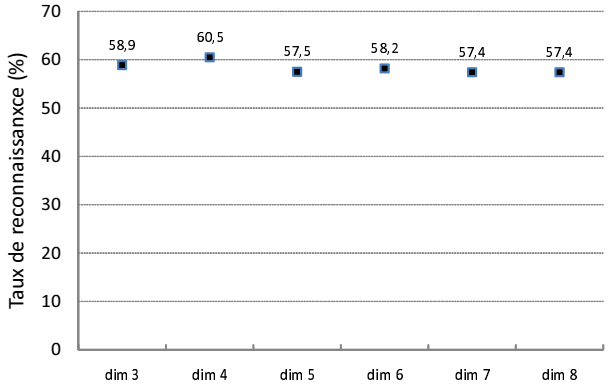


FIGURE 6.6 – Taux de reconnaissance moyen de 14 expressions inconnues sur des sujets connus avec la méthode OBM, utilisant les caractéristiques de forme et de texture du visage.

6.2.2 Robustesse par rapport aux types de données du visage (forme et/ou texture)

Cette section va montrer que, contrairement aux paramètres d'apparence, l'organisation des expressions et la signature direction-intensité sont robustes aux différences de types de données du visage (forme et/ou texture).

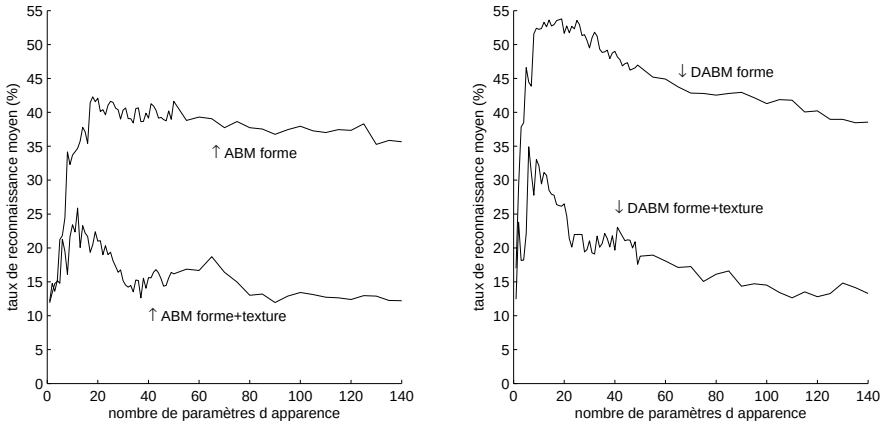


FIGURE 6.7 – Taux de reconnaissance des méthodes ABM et DABM avec ou sans les informations de texture (14 expressions inconnues de sujets connus).

L'organisation des expressions est robuste par rapport au type de données Nous allons tout d'abord étudier l'impact du type de données sur l'organisation des expressions. En d'autres termes, nous allons répondre à la question : est-ce que l'organisation des expressions varie si l'on prend d'un côté uniquement des vecteurs de forme du visage, d'un autre uniquement des vecteurs de texture du visage et pour finir des vecteurs de forme et de texture du visage ?

Pour évaluer l'impact de la forme du visage versus la texture du visage, nous avons calculé l'organisation des expressions et la distribution de l'indice de similarité. Exactement la même organisation a été trouvée avec les paramètres de forme seulement et les paramètres de forme et de texture simultanément ($Q(S_s^{shape+texture}, S_s^{shape}) = 1$). L'organisation trouvée avec seulement les paramètres de texture diffère par seulement deux substitutions d'un bord, cf. substitution sur la figure 4.10 ($Q(S_s^{shape+texture}, S_s^{texture}) = 0.88$. Ces structures peuvent donc être considérées comme similaires, conformément à l'analyse faite dans la section 4.4 puisque l'indice de similarité est supérieur à 0.8.

Le type de données n'influence pas les taux de reconnaissance Dans ce paragraphe, nous étudions l'impact du type de données sur le taux de reconnaissance.

Contrairement à notre méthode, les méthodes classiques basées apparence (ABM et DABM) sont très sensibles au type de données utilisées. ABM et DABM appliquées sur notre base de données montrent que l'ajout de l'information de texture diminue les taux de reconnaissance d'environ 16% pour ABM et 19% pour DABM (voir figure 6.7).

Contrairement à nos résultats, Cheon et Kim [54] faisaient le constat inverse. Sur leur base de données, l'ajout des informations de texture améliorait les résultats de reconnaissance à la fois sur des vecteurs d'apparence (AAM) et des vecteurs d'apparence différentiels (Diff-AAM). Cela peut être expliqué par le fait que notre base de données contient une variété plus large de sujets comparée à la leur (qui ne contenait que des Coréens). Ainsi, dans notre base de données, les caractéristiques de texture portent trop d'informations d'identité (barbe, moustache, lunette, couleur de la peau, ...) pour être pertinentes et comparables entre les sujets. Les résultats des modèles d'apparence sur notre base de données peuvent aussi être comparés aux travaux de Martin et al. [101]. Ils utilisent un classifieur MLP et un classifieur SVM sur des données d'apparence (AAM) pour effectuer la reconnaissance des expressions. Ils ont eux aussi comparé les taux de reconnaissance en utilisant des données de forme uniquement d'une part et de forme et texture d'autre part. Ils ont constaté que l'ajout de la texture améliorait la reconnaissance. L'une des raisons de ces résultats peut être que les classifieurs (MLP ou SVM) apprennent, en plus des expressions, les différents types d'identité pendant la phase d'apprentissage. Les tests n'ont pas été effectués en généralisation sur des sujets inconnus. Les sujets de tests étaient des sujets de la base d'apprentissage. L'ajout de la texture permet alors de mieux séparer les différentes identités des sujets et améliorent ainsi les résultats des classifieurs. Ces comparaisons soulignent le fait que ces méthodes basées sur les vecteurs d'apparence sont fortement impactées par le type de données utilisées (forme et/ou texture) et dépendent grandement de la base d'apprentissage.

Qu'en est-il de notre méthode basée organisation ? La figure 6.8 montre les taux de reconnaissance de notre méthode OBM basée d'abord sur les caractéristiques de forme uniquement et ensuite sur les données de forme et de texture. Ces informations sont fournies selon la dimensionnalité de la variété. Nous pouvons constater que le taux de reconnaissance sur les 14 expressions inconnues se situe entre 55.9% et 61.4% indépendamment du type de caractéristiques pris en compte (forme ou forme et texture). Dans ces expérimentations, le taux de reconnaissance moyen avec OBM varie de moins de 3% quelque soit le type de données utilisées. La méthode OBM est donc stable vis-à-vis du type de données.

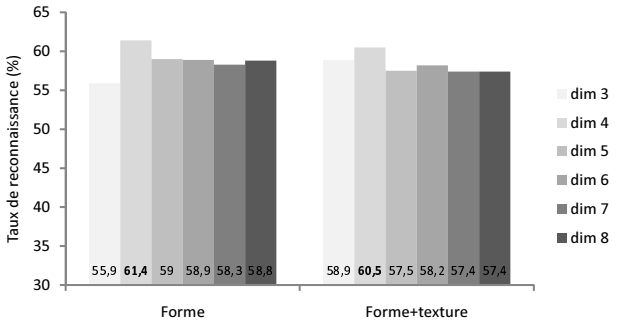


FIGURE 6.8 – Taux de reconnaissance moyen de 14 expressions inconnues sur des sujets connus avec la méthode OBM. Résultats donnés selon la dimensionnalité de la variété et réalisés sur deux types de données : forme et forme+texture. Le nombre de paramètres d'apparence varie entre 1 et 136 = 17 (nombre de sujets) * 8 (nombre d'expressions).

Le type de données n'influence pas la signature direction-intensité Dans ce paragraphe, nous étudions l'impact du type de données sur la signature direction-intensité.

Comme l'organisation est robuste au type de données, nous nous attendons à ce que la signature d'une expression issue des informations de forme, celle issue des informations de texture et celle issue des informations de forme et texture aient des valeurs proches.

Pour cela, nous avons effectué des tests pour lesquels les signatures sont apprises sur un premier type de données et sont testées sur un autre type de données. Dans l'algorithme de la section 5.2.4, cela implique de prendre des signatures issues d'un premier type de donnée pour les sujets i et des signatures d'un autre type de données pour les sujets j. Par exemple, la signature de l'expression 5 du sujet 17 calculée à partir des données de forme+texture est comparée aux signatures des expressions 1 à 14 des sujets 1 à 16 calculées à partir des données de texture uniquement. La figure 6.9 montre cet exemple. Pour les sujets 1, 2 et 16, la signature est correctement reconnue alors que pour le sujet 3, l'expression la plus proche est l'expression 13.

Les résultats de ces tests sur les 14 expressions non basiques sont montrés dans le tableau 6.1. Ces résultats montrent que le type de données (entre la phase d'apprentissage et la phase de test) influence peu les résultats de reconnaissance avec une variation moyenne des taux de reconnaissance entre les techniques de 2.3%.

Test\Apprentissage	Texture	Forme	Forme+Texture
Texture	56.8	55.9	59.1
Forme	59.2	**61.4**	58.9
Forme+Texture	**61.9**	60.3	60.5

TABLE 6.1 – Taux de reconnaissance moyen de 14 expressions inconnues sur des sujets connus avec la méthode OBM. Résultats fournis selon le type de données des phases d'apprentissage et de test.

Cette robustesse peut être analysée par le fait que les caractéristiques de forme et de texture sont corrélées et que, contrairement aux méthodes basées sur les vecteurs d'apparence (ABM et DABM), les paramètres de signature de la méthode basée sur l'organisation des expressions

(OBM) ont une signification qui est connue par construction : une expression est située entre n autres expressions (où n est la dimension de la variété).

De façon générale, nous proposons d'utiliser la méthode OBM sur les données de forme car le calcul de la forme diminue la complexité des traitements et donne les meilleures performances (meilleurs taux moyen de 61.4% en dimension 4).

Forme + Texture

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Sujet 17	0,68	0,4	0,37	0,08	0,41	0	0	0,13	0	0,42	0	0	0,61	0
	0	0	0	0,15	0	0	0	0	0	0	0	0	0	0
	0	0	0,35	0	0	0	0	0	0	0	0	0	0,13	0
	0	0	0,13	0,62	0,5	0,6	0,36	0,09	0,14	0,4	0,24	0,3	0	0,22
	0,07	0,6	0	0,14	0,01	0	0	0,78	0,36	0,09	0,13	0,62	0	0,17
	0,25	0	0	0	0,08	0,4	0,64	0	0,5	0,08	0,62	0,09	0	0,03
	0	0	0,15	0	0	0	0	0	0	0	0,01	0	0,16	0,58
	0	0	0	0	0	0	0	0	0	0	0	0	0,11	0
	0,46	1,21	0,89	0,85	0,52	1,06	1,02	0,75	0,71	0,51	0,73	1,12	0,81	0,69

Texture

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Sujet 1	0,29	0,37	0,33	0	0	0	0	0	0	0	0	0	0	0,27
	0,2	0,34	0	0,69	0	0	0	0	0	0	0	0	0	0
	0,45	0,11	0,26	0	0,41	0	0,13	0	0	0,29	0	0	0,41	0
	0,06	0	0	0	0,48	0,5	0,25	0,01	0,12	0,01	0,08	0,22	0	0,06
	0	0	0	0,17	0	0,43	0	0,18	0	0	0	0,29	0	0
	0	0,18	0	0	0	0	0	0,21	0,04	0	0,06	0,21	0	0
	0	0	0	0	0,1	0,07	0,62	0,59	0,84	0,42	0,86	0,29	0,25	0,33
	0	0	0,41	0,14	0,01	0	0	0	0	0,28	0	0	0,34	0,34
	0,68	0,69	0,85	0,88	0,54	0,91	0,66	0,77	0,9	0,51	1	0,6	0,5	0,77
Sujet 2	0,24	0,65	0,31	0,05	0,23	0,06	0	0	0	0	0	0	0,3	0,28
	0	0,3	0	0,49	0	0	0	0,27	0	0	0	0	0	0
	0,3	0	0	0,06	0	0	0	0	0	0,93	0	0	0	0
	0	0	0	0,4	0,77	0	0,52	0	0,01	0	0,15	0,38	0	0
	0	0	0	0	0	0,52	0,25	0,44	0,2	0	0,2	0,3	0	0
	0,46	0,04	0	0	0	0,42	0	0,18	0,27	0,07	0,3	0	0	0
	0	0	0,21	0	0	0	0,24	0	0,52	0	0,35	0,32	0,32	0,56
	0	0	0,47	0	0	0	0	0,1	0	0	0	0	0,38	0,16
	0,27	0,88	0,65	0,96	0,69	0,95	1,15	0,44	0,85	0,24	0,63	1	0,45	0,5
Sujet 3	0,77	0,33	0	0	0,09	0	0	0,03	0	0	0	0	0,36	0
	0	0,14	0,11	0,48	0	0	0	0,54	0	0	0	0	0	0,14
	0,08	0,26	0,28	0,25	0	0	0	0,3	0	0	0	0	0	0
	0	0	0	0	0,37	0,36	0,22	0	0	0	0,08	0,06	0	0
	0	0	0	0	0	0,4	0,23	0	0,53	0,54	0,08	0,75	0	0,2
	0,02	0,26	0	0,01	0	0	0,05	0,14	0,16	0,24	0,14	0,16	0,31	0
	0	0	0,43	0	0,18	0,24	0,5	0	0,31	0,23	0,7	0,03	0	0,51
	0,13	0	0,17	0,26	0,36	0	0	0	0	0	0	0	0,34	0,15
	0,42	0,7	0,67	0,93	0,86	1,1	0,86	0,84	0,84	0,67	0,96	0,96	0,45	0,77
⋮														
Sujet 16	0,379	0,216	0	0	0,071	0	0	0	0	0	0	0	0,093	0,048
	0	0,768	0,096	0,392	0	0	0	0,059	0	0	0	0	0	0
	0,451	0	0,477	0,04	0,294	0	0	0	0	0	0	0	0	0
	0,148	0	0	0,568	0,616	0,786	0,394	0	0	0	0,305	0,261	0	0
	0	0	0	0	0	0,213	0,015	0,684	0,388	0,248	0	0,189	0,16	0
	0	0,016	0	0	0	0,002	0,563	0,125	0,419	0,342	0,695	0,55	0	0
	0	0	0,427	0	0	0	0,028	0,133	0,193	0,411	0	0	0,747	0,777
	0,022	0	0	0	0,018	0	0	0	0	0	0	0	0	0,175
	0,96	0,72	0,71	1,05	0,63	1,07	0,88	0,92	0,87	0,6	0,97	1,05	0,81	0,77

FIGURE 6.9 – Exemples de signatures ($n = 4$) de différents sujets issues des données de forme et de texture et des données de texture uniquement. La signature de l'expression 5 du sujet 17 calculée à partir des données de forme+texture est comparée aux signatures des expressions 1 à 14 des sujets 1 à 16 calculées à partir des données de texture uniquement.

6.3 Résultats Comparatifs

6.3.1 Résultats sur des sujets connus

Cette section compare les taux de reconnaissance de la méthode basée sur l'organisation des expressions (OBM) avec les méthodes basées les vecteurs d'apparence (ABM et DABM). La figure 6.10 montre les résultats des méthodes ABM et DABM selon leur nombre de paramètres ainsi que les résultats précédents issus de la méthode OBM (meilleurs taux de reconnaissance de 61.4% avec des paramètres de forme uniquement).

Nous constatons tout d'abord que la méthode DABM donne de bien meilleurs résultats que la méthode ABM, avec une augmentation du taux de reconnaissance d'environ 10%. Cela confirme les résultats de Cheon et Kim [54] sur le fait que la soustraction du neutre diminue l'impact de la morphologie des sujets.

Nous constatons aussi que la méthode basée sur l'organisation des expressions (OBM) donne de meilleurs résultats que la méthode DABM. La figure 6.10 montre le meilleur taux de reconnaissance pour OBM, mais cette constatation est valable quelle que soit la dimensionnalité de la variété et quel que soit le type de donnée (forme ou forme+texture) de la méthode OBM (taux de reconnaissance allant de 55.9% à 61.4% - voir figure 6.8). Le meilleur taux de reconnaissance obtenu avec la méthode OBM est 61.4% soit une amélioration de 7.6% sur le meilleur résultat de la méthode DABM.

Ces résultats montrent que la position d'une expression par rapport aux autres est plus pertinente pour qualifier une expression que la position absolue de cette expression. Ces résultats confirment aussi le fait que les expressions sont organisées de la même façon selon les personnes.

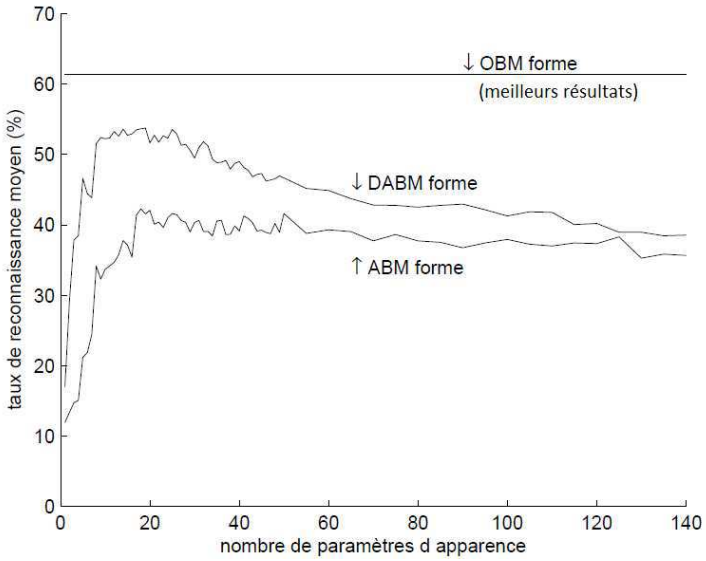


FIGURE 6.10 – Comparaison des taux de reconnaissance d'expressions faciales entre les méthodes ABM, DABM et OBM. Résultats fournis avec ou sans les informations de texture et pour 14 expressions inconnues de sujets connus. Pour la méthode OBM, les meilleurs résultats sont affichés. Les résultats en fonction du nombre de paramètres d'apparence (variant de 1 à 8) sont présentés sur la figure 6.7.

6.3.2 Résultats sur des sujets inconnus

Les analyses précédentes ont été effectuées sur 14 expressions non basiques de sujets connus. Nous allons maintenant étudier les résultats de reconnaissance sur des sujets inconnus.

Description de l'algorithme L'algorithme utilisé est le même que celui de la section 5.2.4. Une méthode de leave-one-out sur les sujets a été mise en œuvre pour créer les espaces d'apparence (modèles AAM). Cela permet de gérer le fait que les sujets soient inconnus, tout en ayant suffisamment de données : le modèle AAM est créé sur 16 sujets et le $17^{\text{ième}}$ sujet est testé ; ce mécanisme est réalisé pour chacun des 17 sujets. Les résultats sont présentés avec et sans l'algorithme de vote.

Les données : 22 expressions non basiques et inconnues de 17 sujets inconnus Nous utilisons la même base de données que précédemment. Dans la mesure où les sujets sont inconnus, nous pouvons utiliser l'ensemble des 22 expressions pour effectuer les tests de reconnaissance.

La méthode basée sur les vecteurs d'apparence n'est pas impactée par la dimensionnalité de la variété Le tableau 6.2 montre les résultats de reconnaissance de la méthode basée sur l'organisation des expressions sur des sujets inconnus (AOBM - *Assumed Organization Based Method*) selon la dimensionnalité de la variété. Comme sur des sujets connus, les résultats sont stables pour une dimension supérieure ou égale à 3.

Dim. variété	3	4	5	6	7	8
Sans vote	**32.5**	32.3	30.7	30.4	30.3	30.1
Avec vote	45.2	**45.9**	43.8	44.6	42.7	42.1

TABLE 6.2 – Taux moyen de reconnaissance de 22 expressions inconnues de sujets inconnus avec la méthode AOBM. Résultats fournis selon la dimensionnalité de la variété.

La méthode basée sur les vecteurs d'apparence améliore les résultats de reconnaissance Le tableau 6.3 compare les résultats des méthodes basées sur les vecteurs d'apparence (ABM), basées sur les vecteurs d'apparence différentiels (DABM) et basées sur l'organisation des expressions (AOBM). Les meilleurs résultats sont obtenus avec la méthode basée sur l'organisation des expressions. Elle améliore les meilleurs résultats de reconnaissance des méthodes basée sur les vecteurs d'apparence de 1.3% avec l'algorithme de vote et 4.9% sans l'algorithme de vote. Le meilleur taux de reconnaissance sur 22 expressions de sujets inconnus est obtenu par la méthode basée sur l'organisation des expressions (AOBM) avec un taux de 45.9%. A titre de comparaison, un tirage aléatoire donne un taux de reconnaissance de 4.5% (1 chance sur 22).

Méthode	ABM	DABM	AOBM
Sans vote	17.4	27.6	**32.5**
Avec vote	29.9	44.6	**45.9**

TABLE 6.3 – Comparaison des taux de reconnaissance de 22 expressions inconnues de sujets inconnus avec les méthodes ABM, DABM et AOBM.

A noter que sur des sujets inconnus, avec l'algorithme de vote, la méthode AOBM n'améliore que faiblement les résultats par rapport à la méthode DABM. Cela est dû au fait que les $K = 8$ expressions ayant servies à définir la structure sont des déformations plausibles et non des déformations réelles. Avoir les déformations réelles nous ramènerait aux résultats de la section 6.3.1 avec des taux de l'ordre de 60%.

Impact de l'algorithme de vote Comme attendu, l'algorithme de vote mis en œuvre dans l'algorithme de la section 5.2.4 améliore les taux de reconnaissance, quelle que soit la méthode utilisée (basée sur les vecteurs d'apparence ou sur l'organisation des expressions). Nous constatons quand même que sans algorithme de vote, la méthode AOBM est nettement plus performante que la méthode DABM, ce qui signifie que la signature proposée est plus pertinente.

6.3.3 Étude des cas de confusions : émotions proches

Les 22 expressions (voir description de la base de données dans la section 6.1) ont été labellisées manuellement avec l'une des 6 catégories émotionnelles de base [37] (voir figure 6.11). Les expressions 1 à 4 ont été labellisées comme 4 types de colère, les expressions 5 à 7 comme 3 types de dégoût, les expressions 8 à 11 comme 4 types de joie, 12 et 13 comme 2 types de peur, 14 à 18 comme 5 types de surprise et 19 à 22 comme 4 types de tristesse.

La figure 6.12 montre la matrice de confusion correspondant au meilleur résultat obtenu par la méthode basée sur les vecteurs d'apparence sur des sujets inconnus (AOBM).

Nous constatons que les confusions apparaissent principalement sur des expressions liées à la même émotion. Pour certaines expressions non basiques, nous constatons la confusion classique entre des expressions de colère et de dégoût et entre des expressions de peur et de surprise [102].

Cela confirme qu'il est possible de se baser sur l'analyse des expressions faciales pour en déduire une information sur les émotions. La partie III étudie cet aspect.

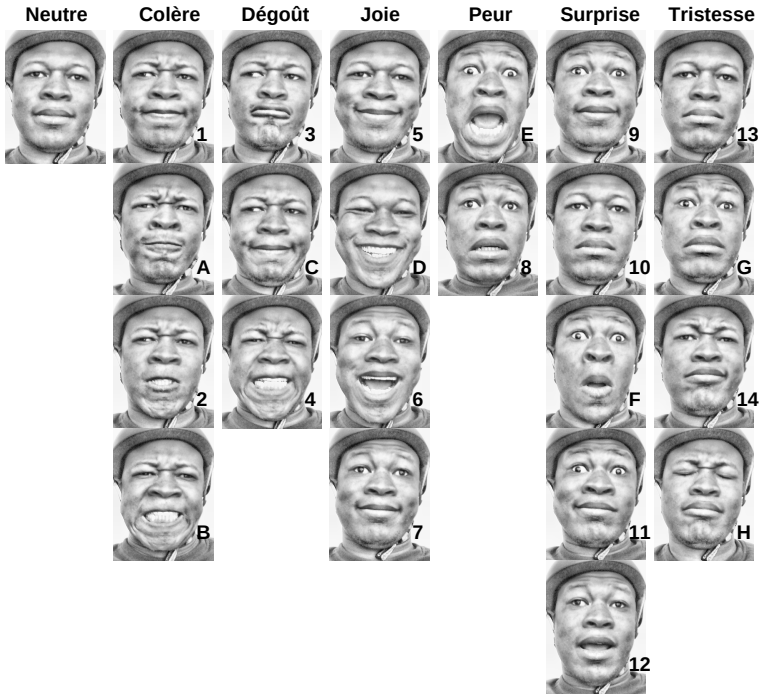


FIGURE 6.11 – Exemple des 22 expressions plus visage neutre réalisées par un sujet. Ces 22 expressions sont labellisées avec l'une des 6 catégories émotionnelles de base.

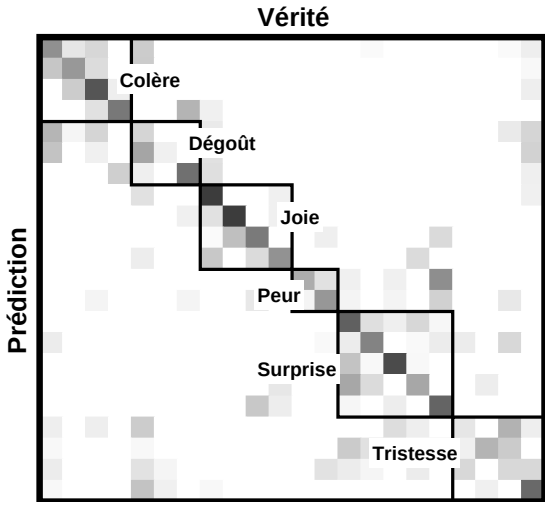


FIGURE 6.12 – Matrice de confusion de la reconnaissance de 22 expressions mélangées de 17 sujets inconnus. Plus la case est foncée, plus le taux de reconnaissance est élevé.

6.4 Conclusion Intermédiaire

Cette partie a présenté une nouvelle représentation invariante de l'espace des expressions faciales et, en conséquence, une nouvelle approche pour l'analyse des expressions faciales. La représentation est basée sur l'organisation des expressions les unes par rapport aux autres. Une expression est définie par sa position relative par rapport à d'autres expressions connues du système. Ainsi, chaque nouvelle expression non prototypique possède une unique signature. Le système permet ainsi de traiter des expressions non incluses dans les bases d'apprentissage. Pour répondre aux exigences de flexibilité, la méthode a d'abord été appliquée sur des sujets connus avant d'être étendue aux sujets inconnus. La méthode proposée montre de meilleures performances que les méthodes traditionnelles basées sur des vecteurs d'apparence AAM.

Les techniques permettant de créer un espace des déformations basées sur l'ACP sont très sensibles à la distribution sous-jacente des données et aux principales variations de cette distribution. Dans la méthode proposée, les espaces spécifiques sont tous créés avec le même nombre d'expressions et le même type d'expressions (expressions similaires). Cela permet de contrecarrer les problématiques liées à la distribution des données.

Contrairement aux méthodes basées sur les vecteurs d'apparence AAMs, cette méthode est robuste au type de données utilisées (forme et/ou texture). Par ailleurs, les paramètres ont un sens et peu de paramètres sont nécessaires pour caractériser de façon unique une expression.

Le taux de reconnaissance de 45.9% peut paraître faible et sans intérêt pratique. Néanmoins, dans un système complet, l'expression est calculée pour chaque image d'une séquence vidéo et est intégrée dans le temps avec une période d'au moins 1 seconde (c'est-à-dire environ 30 images). L'information est alors suffisamment précise pour détecter des informations de plus haut niveau (telles que le rire) des sujets. Nous allons, dans la partie suivante, proposer de mettre en œuvre cette méthode dans un système plus complet (voire figure 6.13), basé sur des séquences vidéo affichant des émotions spontanées.

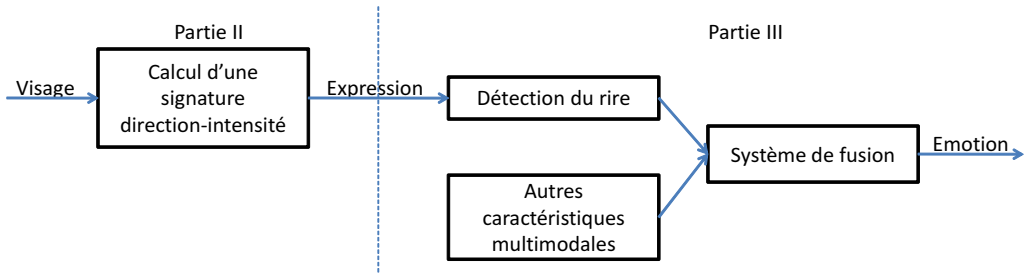


FIGURE 6.13 – Intégration de la méthode de description d'une expression dans un système plus complet.

Troisième partie

Analyse d’Émotions

Cette partie a pour objectif de tester la représentation des expressions faciale présentée dans la partie précédente (partie II) sur des expressions spontanées, en utilisant les données de séquences vidéo et dans un système plus complet. Comme nous l'avons vu dans l'état de l'art (section 3.1), une expression faciale a plusieurs origines : signaux physiologiques, visèmes, interactions sociales, états mentaux. Nous nous proposons, dans cette partie, de nous intéresser aux *états mentaux* et de détecter les variations d'*émotions* de sujets à partir de séquences audio-vidéo.

Ces travaux ont été réalisés dans le cadre du challenge AVEC 2012 [80]. Il s'agit ici de données de personnes réelles en situation d'interaction (dialogue) avec un agent émotionnel. Le mode de représentation choisi est une représentation des émotions sous la forme de 4 dimensions : valence, arousal, power et expectancy [73].

Comme indiqué dans la section 3.3.4, les expressions faciales ne sont qu'un *canal* parmi d'autres permettant d'obtenir des informations sur l'émotion des sujets. Nous avons donc extrait des informations émotionnelles provenant d'autres canaux de transmission et les avons fusionnées afin d'obtenir un système multimodal d'analyse d'émotions. La pertinence du système multimodal a été testé via le calcul de la corrélation entre les prédictions obtenues par notre système et les vérités terrain, nous parlerons donc de *variations* émotionnelles.

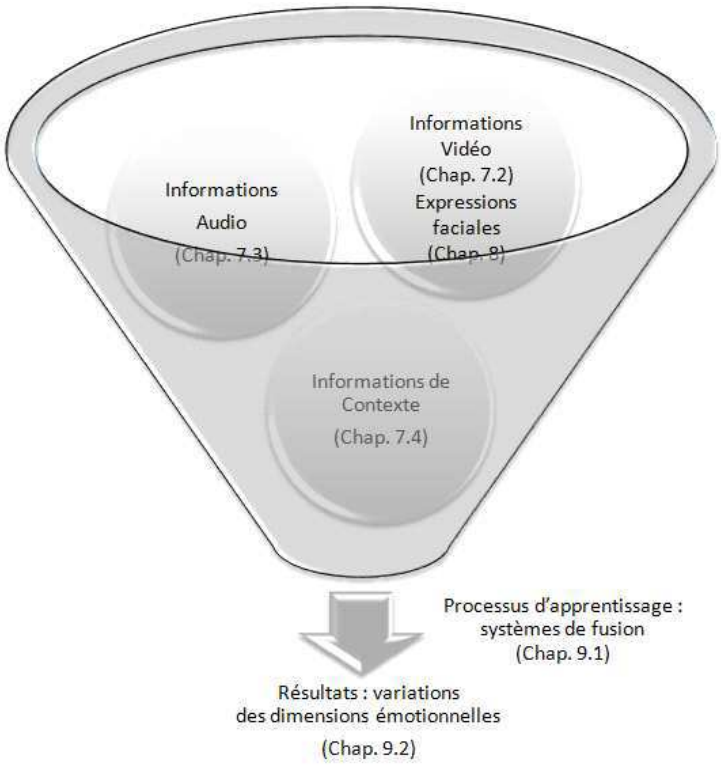


FIGURE 6.14 – Présentation du processus d'interprétation des expressions en émotions (système multimodal).

Dans cette partie, nous présentons tout d'abord en introduction une vue d'ensemble du système multimodal proposé. Nous exposerons ensuite l'extraction des caractéristiques pertinentes pour l'analyse des émotions qui ne relèvent pas des expressions du visage (chapitre 7) puis nous nous focaliserons sur les expressions faciales (chapitre 8). Pour finir, nous présenterons et comparerons deux systèmes de fusion des données (chapitre 9) et indiquerons les résultats obtenus (section 9.2).

Le challenge AVEC 2012 Les travaux se sont inscrits dans le cadre du challenge AVEC 2012 [80] où nous sommes arrivés second sur 10 équipes en compétition, derrière [82]. Le challenge AVEC (Audio/Visual Emotion Challenge and Workshop) a pour objectif principal de comparer des méthodes de traitement de données multimédia et de systèmes d'apprentissage dans le contexte d'analyse automatique d'émotion à partir de la vidéo, de l'audio ou des deux médias. La mise en œuvre du challenge permet de mettre les candidats dans des conditions identiques et de comparer leurs méthodes sur les mêmes données.

Un second objectif du challenge est l'amélioration des systèmes de reconnaissance d'émotion afin qu'ils soient capables de traiter des comportements naturels, de gros volumes de données non segmentées, des données réelles non présélectionnées et non prototypées. L'objectif est de se rapprocher de contextes réels de communication entre humains et machines.

Ce challenge est une **triple opportunité**. Il permet de tester et d'améliorer notre représentation des expressions faciales :

- sur des données réelles issues de séquences vidéo,
- dans un contexte de reconnaissance d'émotions,
- dans un environnement multimodal.

Les données du challenge Les données du challenge sont **95 séquences audio-vidéo** d'une durée de 2 à 10 minutes affichant une conversation entre un sujet et un agent émotionnel (réalisé par une personne humaine). Il existe 4 agents émotionnels ayant chacun un comportement émotionnel défini :

- Poppy est joyeux,
- Spike est agressif,
- Odadiah est sombre, triste,
- Prudence est pragmatique.

Lors de ces conversations, le sujet est filmé de face et la conversation est enregistrée à l'aide d'un micro porté par le sujet.

Les données sont structurées en 3 **bases de données**. La base d'entrainement (training) a pour objectif d'être utilisée pour l'apprentissage du système (31 séquences audio-vidéo). La base de développement (development) a pour objectif d'être utilisée pour tester le système (32 séquences audio-vidéo). La base de test (test) sert à l'évaluation des participants (32 séquences audio-vidéo).

Les **labels de vérité terrain** sont fournis pour les bases d'entrainement et de développement. Les données ont été labellisées par 3 à 8 annotateurs grâce à l'outil FEELTRACE [71] et les vérités terrains sont les valeurs moyennes de ces 3 à 8 labels. L'émotion doit être reconnue sous la forme de 4 dimensions continues [73] :

- Arousal : niveau d'excitation du sujet. Plus le niveau est élevé, plus la personne est active. Inversement, plus le niveau est bas, plus le sujet est passif.

- Expectation : niveau d'attente du sujet. Il s'agit ici d'évaluer si le sujet est ou non surpris lors de la conversation.
- Valence : niveau de positivité du sujet. Plus le niveau est élevé, plus le sujet est positif. Inversement, plus le niveau est bas, plus le sujet est négatif.
- Power : niveau de puissance du sujet. Il s'agit ici d'évaluer si le sujet domine ou pas la conversation.

La **performance** du système est calculée en termes de corrélation entre les prédictions de ces 4 dimensions trouvées par les participants et les labels de vérité terrain. Cela signifie que ce qui importe n'est pas la valeur trouvée mais la variation de l'émotion représentée sous forme dimensionnelle au cours du temps.

$$perf = \left| \frac{1}{4} \sum_{dim.} \frac{1}{nbSeq.} \sum_{seq.} \frac{\sigma_{pt}}{\sigma_p \sigma_t} \right| \tag{6.1}$$

où $nbSeq.$ correspond au nombre de séquences audio-vidéo $seq.$, $dim.$ correspond aux 4 dimensions, σ_{pt} est la covariance entre la prédiction p et la vérité terrain t de la séquence audio-vidéo pour la dimension choisie, et σ_p et σ_t sont les écarts type des prédictions p et vérités terrains t de la séquence audio-vidéo pour la dimension choisie.

Les acteurs et leurs contributions Nous avons réalisé ce challenge dans le cadre du projet collaboratif IMMEMO (1). Les partenaires sont les suivants :

- Catherine Pelachaud de TELECOM ParisTech qui nous a précisé les dimensions émotionnelles et les caractéristiques audio-visuelles mises en jeu lors de ces variations émotionnelles (section 7.1.1),
- Nicolas Stoïber de Dynamixyz qui nous a explicité un ensemble de techniques de fusion des données en précisant les avantages/inconvénients de ces techniques,
- Hanan Salam de Supélec qui s'est chargé de l'extraction des données du visage (section 8.2),
- moi-même (Catherine Soladié - Supélec) qui a réalisé l'extraction des caractéristiques pertinentes (chapitres 7 et 8), la mise en œuvre du système de fusion (chapitre 9) et la coordination des partenaires du challenge et
- Renaud Séguier de Supélec (responsable de thèse et coordinateur du projet IMMEMO) qui a supervisé le déroulement et les différentes étapes du challenge.

La démarche Nous nous sommes rapidement interrogés sur la pertinence de la vérité terrain fournie dans le cadre du Challenge, et notamment sur l'impact de l'outil de sa labellisation et sur les désaccords entre les annotateurs. Nous détaillerons en chiffres cet aspect dans le chapitre 9.2. Nous voulions avant tout mettre en œuvre un système réutilisable permettant de réaliser de la reconnaissance d'émotion à partir de séquences audio-vidéo et non un système dédié au challenge. Pour ne pas se lier aux données du challenge, nous nous sommes détournés des systèmes *boîtes noires* dans lesquels les résultats sont obtenus par apprentissage massif de données. Nous nous sommes orientés vers la mise en œuvre d'un système *boîte blanche*, ce qui signifie que le système propose des résultats intelligibles à chacune de ces étapes. Pour ce qui est des *spécificités du challenge*, nous les avons inclues dans notre étude et dans notre système afin d'avoir des résultats

(1). `http://www.rennes.supelec.fr/immemo/`

suffisamment élevés pour concourir. Pour autant, nous avons veillé à garder la possibilité de facilement les supprimer et obtenir ainsi un système non dépendant des données du challenge.

Le système global Nous avons opté pour l'utilisation d'un système de règles permettant de transformer des caractéristiques multimodales en variations émotionnelles. Pour cela, nous avons mis en œuvre un système d'inférence floue (*Fuzzy Inference System* FIS). Nous avons aussi comparé cette méthode à celle utilisée par les vainqueurs du challenge [82], qui ont utilisé un modèle de fonctions de base radiales (*Radial Basis Function* RBF).

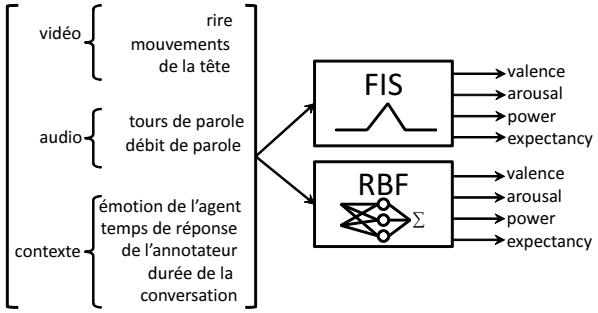


FIGURE 6.15 – Vue d'ensemble de la méthode proposée : un système d'inférence floue ou un système de fonctions de base radiales transforme les caractéristiques pertinentes des fichiers vidéo, des fichiers audio et des données de contexte en 4 dimensions : valence, arousal, power et expectancy.

Chapitre 7

L'Extraction des Caractéristiques Pertinentes

Qu'est-ce que cette étoile ? Et on lit son nom dans un livre, et on croit la connaître.
Jules Renard - *Extrait de son Journal*

Ce chapitre traite de l'extraction automatique des caractéristiques pertinentes permettant de détecter une variation d'émotion. Plusieurs sources d'informations sont disponibles : les séquences vidéo, les fichiers audio, les transcriptions de paroles et les labels de vérité terrain.

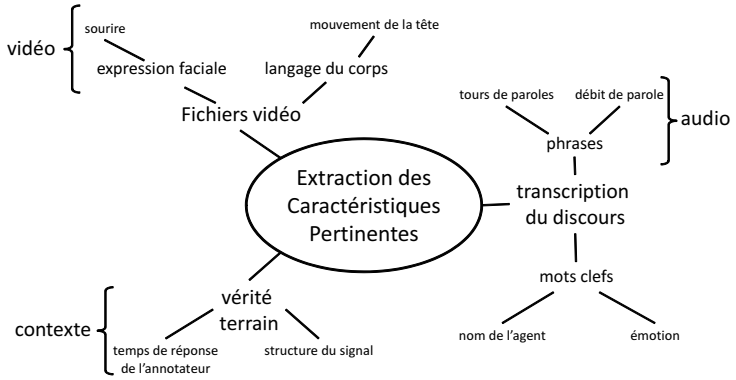


FIGURE 7.1 – Sources des caractéristiques pertinentes : fichiers vidéo, transcription du discours et labels émotionnels.

Avant de définir les méthodes d'extraction des caractéristiques pertinentes, nous avons réalisé une phase préliminaire d'analyse (section 7.1) qui nous a permis d'identifier ces caractéristiques.

Les fichiers vidéo Après une analyse automatique que nous détaillerons dans le chapitre suivant, ils fournissent des informations sur l'expression faciale de la personne ainsi que sur certaines caractéristiques du mouvement du corps (notamment les mouvements de la tête).

Les fichiers audio L'utilisation des fichiers audio pour l'extraction de caractéristiques pertinentes (telles que le ton de la voix) nécessite un savoir faire à part entière qui n'est pas l'objet de notre étude. Nous n'avons donc pas utilisé ce type d'information.

La transcription du discours des sujets Dans le cadre du challenge, cette information est disponible. Dans un système complètement automatisé, une reconnaissance vocale automatique aurait dû être préalablement effectuée. Les informations fournies concernent le discours des sujets uniquement (pas les discours des agents émotionnels). L'ensemble des mots prononcés ainsi que les dates de début et de fin de mot et la structuration en phrase est disponible. Ces données permettent d'extraire facilement des informations concernant les tours de parole (phrases courtes ou longues) et le débit de parole. De plus, ces données fournissent aussi des informations de contexte sur l'agent émotionnel avec lequel la conversation est effectuée. Nous y reviendrons dans la section 7.3.

Les labels de vérité terrain Ils sont disponibles pour les 4 dimensions définissant l'émotion pour les bases d'entrainement et de développement. L'analyse de ces vérités terrains permet d'extraire des informations générales, indépendantes des sujets, caractérisant la structure du signal et le temps de réponse de l'annotateur.

Sommaire

Après avoir présenté la phase préliminaire d'analyse des données (section 7.1), ce chapitre expose les différentes techniques d'extraction des caractéristiques pertinentes mises en œuvre (cf. figure 7.1). Chaque caractéristique est décrite suivant le même modèle. Tout d'abord, le résultat de la phase préliminaire d'analyse est indiqué. Ensuite, la méthode d'extraction automatique de la caractéristique est précisée et son impact sur l'émotion du sujet est mesuré par corrélation : quelque soit la source d'information (fichiers vidéo, transcription de parole, ...), une corrélation est réalisée entre la valeur continue de l'information extraite et la vérité terrain. La valeur de la corrélation est calculée pour chaque séquence de la base de développement et nous affichons la valeur moyenne sur l'ensemble des séquences. Une valeur élevée de la moyenne signifie une forte corrélation et donc un signal pertinent pour définir les variations émotionnelles. Il est à noter que cette corrélation est effectuée sur l'ensemble des séquences. Ainsi, si la corrélation est faible, cela peut être dû à différents facteurs :

- La corrélation n'est pas systématique : dans certains cas, la caractéristique est présente mais n'engendre pas la variation émotionnelle attendue (voire parfois même engendre la variation opposée).
- La caractéristique n'apparait que rarement.

Dans ces différents cas, si la corrélation est faible, la composante ne sera pas considérée comme pertinente pour définir l'allure globale des variations émotionnelles. Elle ne sera pas prise en compte dans le système de fusion décrit au chapitre 9.

7.1 Phase d'Analyse Préliminaire

Afin de définir des caractéristiques pertinentes permettant de détecter une variation d'émotion ainsi que des règles permettant de transformer ces caractéristiques en variations d'émotion, nous avons réalisé une phase préliminaire d'analyse des données du challenge.

7.1.1 Visualisation des séquences audio-vidéo

Une première phase d'analyse des données a été réalisée avec Catherine Pelachaud. L'objectif était de la questionner pour récupérer l'expertise qu'elle avait dans le domaine et la traduire dans un système d'inférence floue. La méthode a été la suivante : Catherine Pelachaud nous a tout d'abord précisé le sens des dimensions caractérisant l'émotion et quelques caractéristiques permettant de les détecter. Puis nous avons visualisé avec elle les vidéos des bases d'entrainement et de développement ainsi que les labels de vérité terrain associés, fournis dans le cadre du challenge. Lors de changement émotionnel, les caractéristiques multimodales entrant en jeu ont été notées. Par exemple :

Sujet Devel-07, 2 min 36 : le sujet ne prend pas la parole, il écoute. Cela implique une dimension power faible et une hausse de l'unexpectancy.

Ces premières analyses ont abouti à un document préliminaire d'étude (voir annexe C). Pour chacune des dimensions, nous avons donné :

- une définition,
- des exemples d'émotions,
- des indices pour détecter l'offset de l'émotion dans une séquence,
- des indices pour détecter la variation de l'émotion autour de cet offset,
- des commentaires.

Lors de cette étude, nous nous sommes aussi interrogés sur la façon dont les annotateurs ont labellisé les séquences. Nous pensons que les annotateurs ont d'abord regardé la séquence dans son ensemble pour se faire une idée du niveau moyen de chaque dimension caractérisant l'émotion et qu'ensuite, lors de l'annotation, ils ont effectué des variations autour de ce niveau moyen.

Quelques **points de vigilance** sont néanmoins à prendre en considération vis-à-vis de cette étude préliminaire :

- Certaines modifications émotionnelles du sujet, constatées sur les labels de vérité terrain, n'ont pas pu être caractérisées.
- La démarche réalisée permet d'identifier les caractéristiques susceptibles de déclencher un changement émotionnel. En revanche, il se peut, dans d'autres séquences ou à d'autres moments de la séquence, que la caractéristique soit présente mais n'engendre pas le changement attendu.
- Seules les principales variations d'émotion ont été analysées et non les petites variations subtiles afin de définir l'allure globale des changements émotionnels.

7.1.2 Calcul de statistiques

Une seconde phase d'analyse a été réalisée. Elle a consisté à calculer des statistiques sur les séquences vidéo. Cette analyse nous a permis :

- de définir le niveau d'offset moyen des personnes,

- de montrer que le caractère de l'avatar déteint sur l'interlocuteur par empathie. Nous y reviendrons dans la section 7.4.1.

Les caractéristiques multimodales identifiées lors de notre phase d'analyse préalable et extraites automatiquement sont listées ci-après. A noter que certaines caractéristiques identifiées lors de l'analyse préliminaire n'ont pas été extraites et utilisées dans notre système, l'extraction nécessitant un trop grand investissement de temps et n'étant pas directement reliée à notre thématique. C'est par exemple le cas de la réactivité du sujet (laps de temps entre la fin de la phrase de l'agent émotionnel et le début de la réponse du sujet) ; cette information aurait nécessité une analyse automatique des séquences audio afin de déterminer les fins de phrases de l'agent émotionnel, information que nous n'avions pas de façon directe dans les données du challenge.

7.2 Analyse des Fichiers Vidéo

Les caractéristiques extraites des fichiers vidéo sont les expressions faciales et le langage du corps.

7.2.1 Le rire (vidéo et audio)

Phase préliminaire d'analyse : Le rire semble jouer un rôle important sur les dimensions valence et arousal. Un rire provoque une augmentation quasi systématique de la perception de valence et d'arousal. Cela s'explique par le fait que le rire est souvent lié à une émotion positive (sauf dans certains cas tels que le rire nerveux) et provoque une augmentation du niveau d'excitation de la personne. La figure 7.2 montre quelques exemples.

Extraction de la caractéristique : La détection du rire a été retenue suite à la phase d'analyse. Elle est présentée dans un chapitre dédié (chapitre 8).

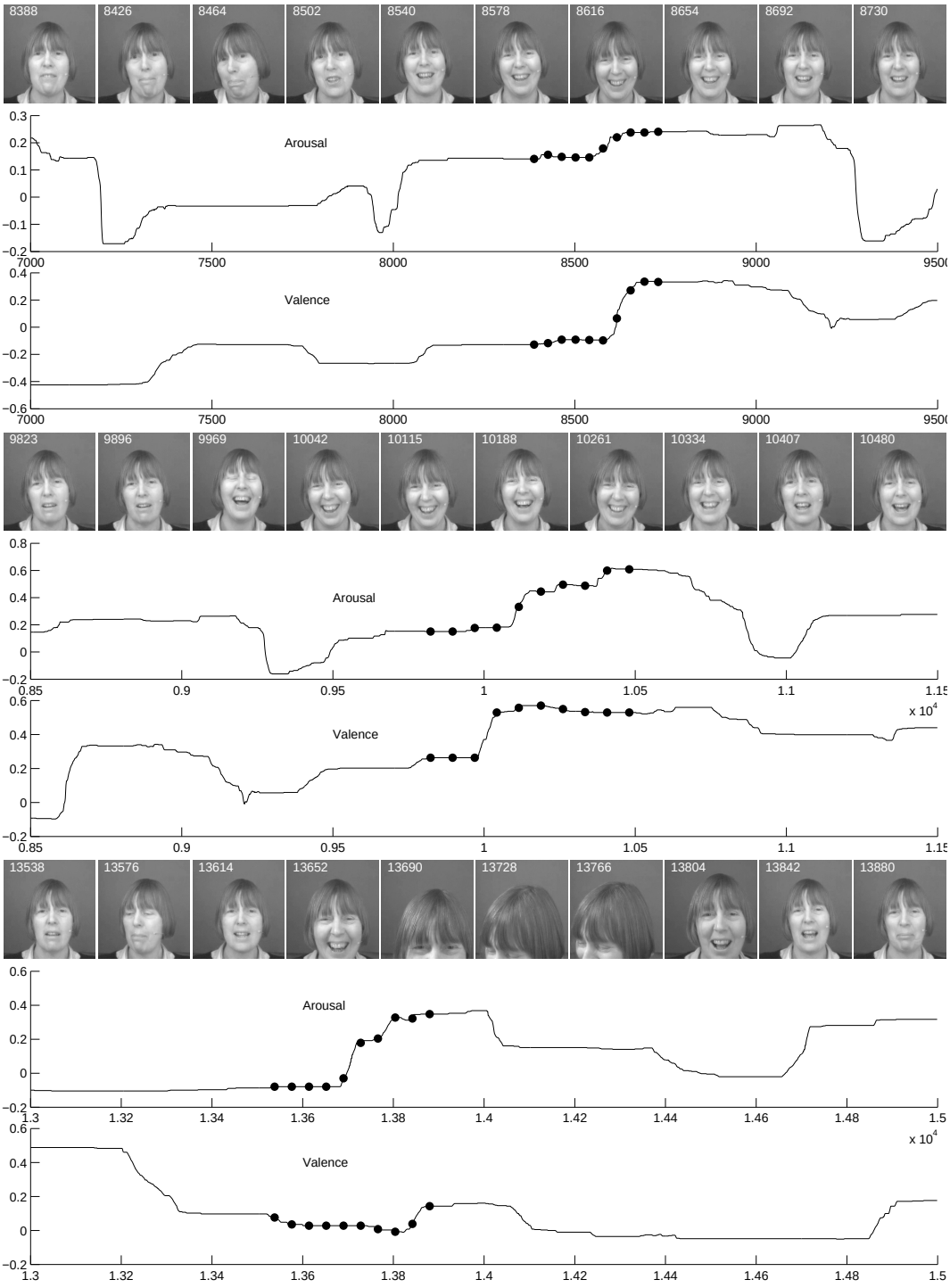


FIGURE 7.2 – Impact du rire sur l'arousal et sur la valence. Les courbes représentent les dimensions arousal et valence sur une période donnée. Les points correspondent aux images affichées au dessus de la courbe.

7.2.2 Le mouvement de la tête (vidéo)

Phase préliminaire d'analyse : Le mouvement du haut du corps et de la tête semble impacter la dimension arousal. Lorsque le sujet se met à bouger beaucoup, la valeur de arousal augmente, et inversement (voir deux premiers exemples de la figure 7.3). Néanmoins, ce constat n'est pas systématique. Par exemple, sur le troisième exemple de la figure 7.3, le sujet se met à bouger et pourtant la valeur de arousal reste quasi constante.

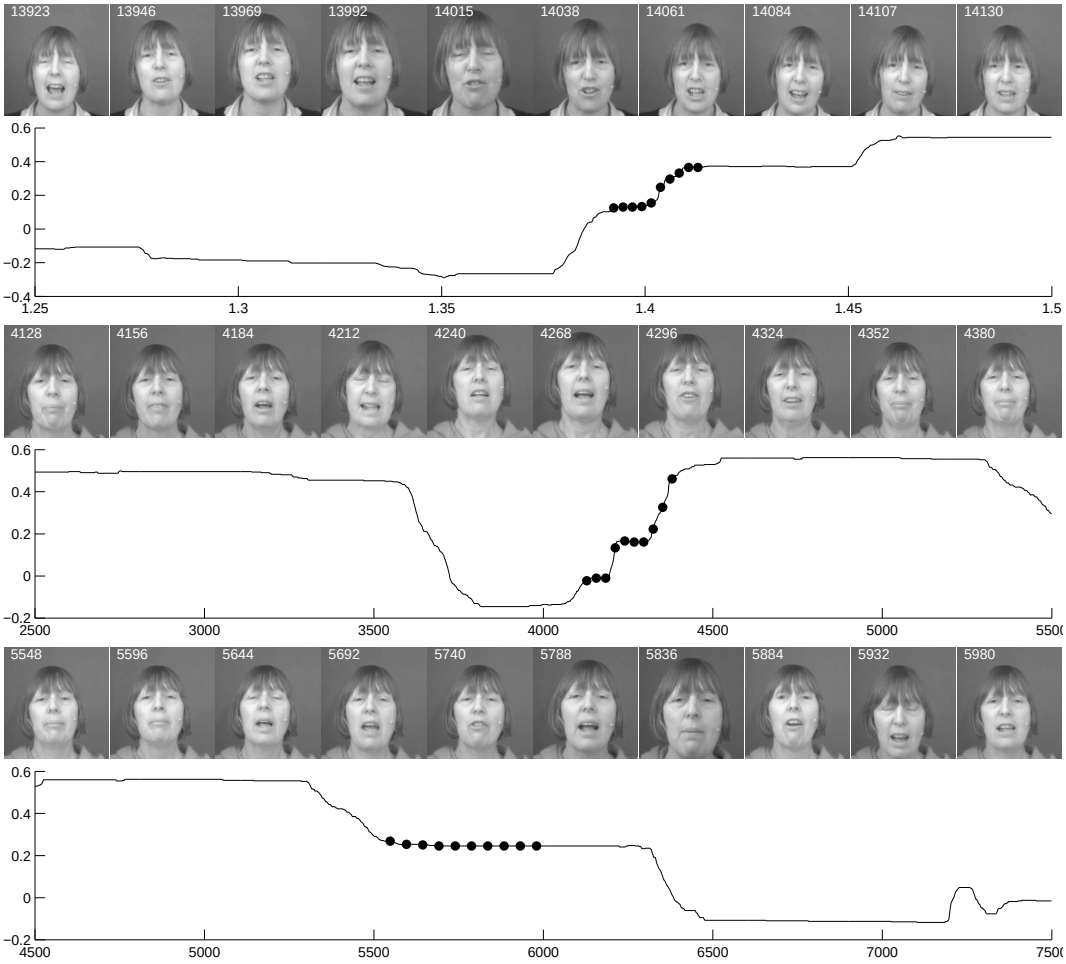


FIGURE 7.3 – Impact du mouvement du haut du corps et de la tête sur l'arousal. La courbe représente la dimension arousal sur une période donnée. Les points correspondent aux images affichées au dessus de la courbe. Dans les trois séquences vidéo, le sujet bouge.

Extraction de la caractéristique : Concernant le langage du corps, nous avons calculé le mouvement global de la position de la tête dans la scène. La détection du visage a été réalisée par Hanan Salam. Elle permet de fournir des informations sur la pose du visage de la façon suivante. Les vidéos sont analysées en utilisant un modèle AAM [18] indépendant de la personne et appris sur les bases d'entrainement et de développement. Lors de la phase de test, les paramètres de pose

du visage sont calculés à partir de ce modèle AAM. Pour la description détaillée de ces travaux, voir les articles [103, 104].

Pour définir le mouvement du corps, nous avons calculé les variations (écart type) des paramètres de pose de la tête dans la séquence vidéo sur une durée glissante de 40 secondes. Plus le sujet va bouger et faire de larges mouvements, plus cette valeur sera élevée. Le tableau 7.1 donne la valeur moyenne des corrélations entre le signal trouvé et les vérités terrains. Nous constatons que cette information est élevée pour l'arousal, ce qui se justifie puisque l'arousal donne un degré d'agitation.

Dimensions	Mouvement de la tête
Arousal	**0.15**
Valence	0.08
Power	-0.02
Expectancy	0.03

TABLE 7.1 – Corrélation moyenne entre le mouvement de la tête et les dimensions décrivant l'émotion.

Néanmoins, nous constatons aussi fréquemment, pour de nombreuses séquences, une corrélation négative d'arousal (29% des séquences de la base de développement ont une corrélation négative). Cela signifie que dans certaines séquences, bien qu'une augmentation des mouvements du corps soit observée, une diminution de l'arousal est constatée. L'information de mouvement du corps, telle qu'elle a été calculée, ne nous semble donc pas suffisamment pertinente. Nous avons choisi de ne pas la prendre en compte lors de la définition de nos règles floues (cf. chapitre 9).

7.3 Analyse des Transcriptions de Parole

Les caractéristiques issues du mode audio proviennent de l'analyse de la transcription des discours réalisés par les sujets. Les phrases et mots clefs de ces discours sont analysés.

7.3.1 Les tours de paroles (audio)

Phase préliminaire d'analyse : Nous avons constaté que la durée des tours de paroles jouait un rôle dans l'émotion des sujets sur la dimension expectancy (pour rappel, il s'agit de conversations entre un agent émotionnel et le sujet). Lorsque les sujets parlent beaucoup en réalisant de longues phrases, ils ne sont pas surpris, vu que l'agent produit alors uniquement des retours courts et ne change pas le sujet de conversation. En revanche, lorsque les sujets répondent, notamment par des phrases courtes, cela peut signifier que les sujets de conversation ne sont pas attendus.

Extraction de la caractéristique (longueur des phrases) : L'analyse des discours permet de définir la longueur des phrases prononcées par le sujet. Dans notre système, nous utilisons une information binaire. Lors d'un tour de parole, la phrase est dite longue si le nombre de mots prononcés est élevé (supérieur à 35), sinon, la phrase est dite courte. Cette valeur de 35 mots a été trouvée empiriquement. C'est celle qui maximise la corrélation entre le signal obtenu et la vérité terrain. La figure 7.4 fournit des exemples de séquences audio-vidéo. Elle présente la valeur du signal (signal binaire de longueur des phrases) ainsi que la vérité terrain (trait plein) de la dimension expectancy. Elle fournit aussi pour la séquence, la corrélation entre ces deux signaux.

A noter que la vérité terrain correspond à l'*unexpectancy* et non à l'*expectancy*. Ainsi, une valeur élevée du label de vérité terrain signifie la surprise.

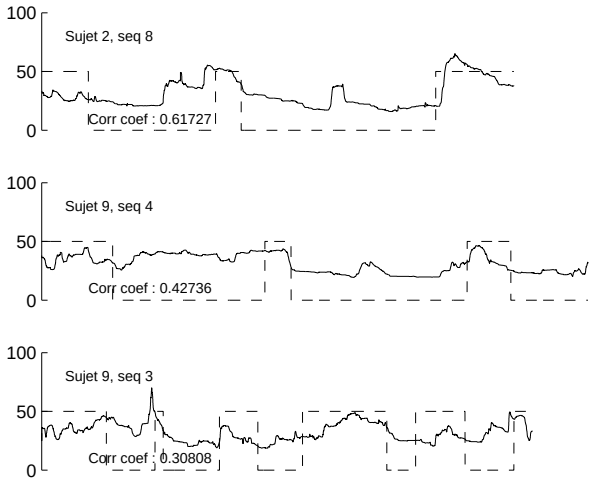


FIGURE 7.4 – Corrélation entre la vérité terrain de la dimension expectancy (trait plein) et la longueur des phrases (trait pointillé) sur trois séquences audio-vidéo différentes.

Le tableau 7.2 donne la valeur moyenne des corrélations entre le signal trouvé et les vérités terrains. Nous constatons, comme observé dans la phase préliminaire d'analyse, que cette information est pertinente pour l'expectancy.

Dimensions	Tours de parole
Arousal	-0.02
Valence	0.09
Power	-0.13
Expectancy	**0.25**

TABLE 7.2 – Corrélation moyenne entre les tours de paroles et les dimensions décrivant l'émotion.

7.3.2 Le débit de paroles (audio)

Phase préliminaire d'analyse : Le débit de parole semble jouer un rôle. Néanmoins, l'analyse préliminaire ne permet pas de définir de règle précise et générale concernant ce ressenti.

Extraction de la caractéristique : Nous avons utilisé l'analyse des phrases pour calculer le débit de parole. Ce débit a été calculé comme le nombre de mots par unité de temps. Le tableau 7.3 donne la valeur moyenne des corrélations entre le signal trouvé (débit de parole) et les vérités terrains. Nous constatons que cette information est élevée pour power, ce qui ne nous semble pas facilement interprétable. Nous constatons aussi fréquemment, pour de nombreuses séquences, une corrélation négative (28% des séquences de la base de développement ont une corrélation négative). Cela signifie que dans certaines séquences, bien qu'une augmentation du débit de parole

soit observée, une diminution du power est constatée. L'information du débit de parole, telle qu'elle a été calculée, ne nous semble donc suffisamment pertinente. Nous avons choisi de ne pas la prendre en compte lors de la définition de nos règles floues (cf. chapitre 9).

Dimensions	Débit de parole
Arousal	0.08
Valence	0.03
Power	0.11
Expectancy	-0.03

TABLE 7.3 – Corrélation moyenne entre le débit de paroles et les dimensions décrivant les émotions.

7.4 Analyse des Labels de Vérité Terrain

7.4.1 L'agent émotionnel (contexte)

Phase préliminaire d'analyse : Les conversations sont réalisées entre un sujet et l'agent émotionnel dont le caractère émotionnel fait partie d'un des 4 quadrants de l'espace valence-arousal (voir figure 7.5) :

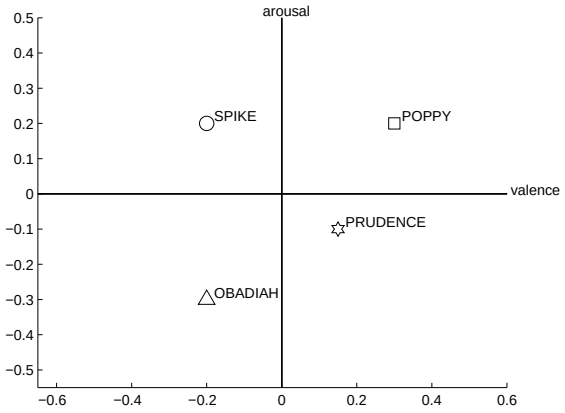


FIGURE 7.5 – Caractère émotionnel des agents : Spike est agressif, Poppy est enjoué, Obadiah est mélancolique et Prudence est pragmatique.

Nous avons constaté que l'émotion moyenne du sujet sur une séquence était généralement liée au caractère émotionnel de l'agent. En effet, si le sujet parle à Poppy, il aura tendance à avoir un comportement avec une valence élevée et un arousal élevé (comme Poppy). S'il parle à Spike, il aura un arousal élevé et une valence faible. En face d'Obadiah, l'arousal et la valence seront faibles. Pour finir, avec Prudence, l'arousal sera moyen et la valence élevée. C'est ce que nous observons sur la figure 7.6.

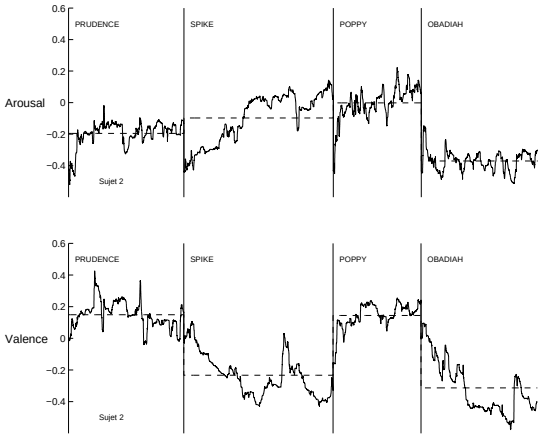


FIGURE 7.6 – Vérité terrain (ligne pleine) comparée à la valeur moyenne des vérités terrains lorsque les sujets parlent à ce même agent émotionnel (ligne pointillée). Exemple de 4 conversations du sujet 2. Arousal sur le premier graphique, valence sur le second graphique.

La figure 7.7 synthétise cette information en analysant l'ensemble des sujets des bases d'entrainement et de développement. Elle montre que l'état émotionnel affiché par le sujet correspond à celui de l'agent émotionnel auquel il parle.

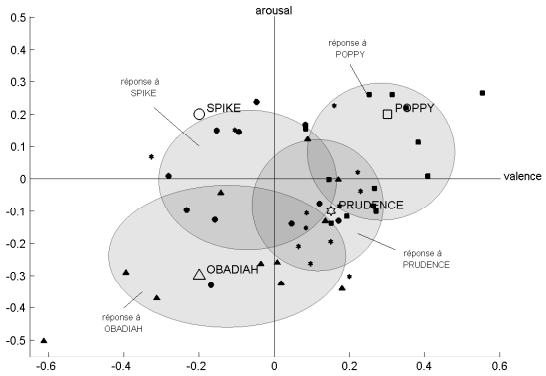


FIGURE 7.7 – Réponse émotionnelle des sujets selon l'agent avec lequel ils interagissent. L'émotion de l'agent est représentée par un marqueur vide (Spike est agressif, Poppy est enjoué, Obadiah est mélancolique et Prudence est pragmatique). La valeur moyenne des labels de vérité terrain de chaque sujet (13 sujets au total) est représentée par un marqueur plein et leur distribution est représentée par les ellipses (valeur moyenne et écart type).

L'impact de l'agent émotionnel peut être interprété comme de l'empathie et de la contagion d'émotion, ce qui peut être une information clef lorsque l'information audio-visuelle n'est pas disponible ou non fiable ; ou alors une information complémentaire lorsque les données audio-visuelles sont exploitables.

Extraction de la caractéristique : Comme nous venons de le voir dans la section 7.4.1, la connaissance de l'agent émotionnel permet d'avoir une information statistique sur l'état émotionnel supposé du sujet. Pour trouver automatiquement l'agent émotionnel, nous avons extrait des mots clefs des discours. Les mots clefs sont soit des termes émotionnels tels que *angry* ou *annoy* qui sont fréquemment utilisés dans les conversations avec Spike soit le nom de l'agent lors de phrases telles que *Bonjour Poppy*. Pour ce qui est du nom de l'agent, au début de chaque conversation, le sujet sélectionne l'agent émotionnel avec lequel il veut discuter. De même, en fin de conversation, il quitte cet agent et choisi le nom de l'agent suivant. Pour ce qui est des mots clefs tels que *angry* ou *annoy*, ils ont été identifiés suite à une analyse statistique des mots prononcés dans les conversations des bases d'entrainement et de développement. Cette analyse a abouti à la liste de mots suivants :

- Pour Poppy : *Poppy*, *fun*
- Pour Spike : *Spike*, *angry*, *annoy*
- Pour Obadiah : *Obadiah*, *sad*
- Pour Prudence : *Prudence*

Il est donc possible, à partir de la transcription d'une conversation, de trouver automatiquement l'agent émotionnel et ainsi d'avoir une information statistique moyenne sur la valence et l'arousal du sujet sur la séquence.

7.4.2 Le temps de réponse de l'annotateur (contexte)

Phase préliminaire d'analyse : L'analyse des vérités terrain souligne un délai dans le temps de réponse des annotateurs. Nous avons calculé la valeur moyenne et l'écart type des vérités terrain sur les bases d'entrainement et de développement, pour chaque dimension caractérisant l'émotion. Nous avons aussi extrait la valeur de départ des vérités terrain au démarrage des séquences.

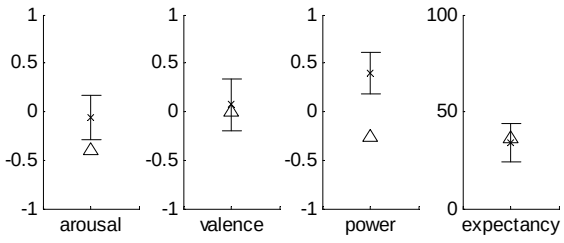


FIGURE 7.8 – Impact du temps de réponse des annotateurs sur les vérités terrains : le triangle montre la valeur du label au début de la séquence audio-vidéo (identique pour toutes les séquences audio-vidéo), la croix montre la moyenne des labels et la plage (I) indique l'écart type des labels. Ces informations sont données pour chaque dimension décrivant l'émotion.

Nous avons tout d'abord constaté que la vérité terrain en début de séquence est la même pour toutes les séquences, et que cette valeur est très différente des valeurs pendant les conversations (cf. figure 7.8), notamment pour arousal et power. Cela est peut être dû à l'initialisation de l'outil d'annotation et au temps de réponse des annotateurs, si bien que les premières secondes des labels de vérité terrain ne sont peut-être pas représentatifs de l'émotion des sujets.

Extraction de la caractéristique : Nous avons modélisé ce comportement (spécifique aux données du challenge) par une fonction linéaire croissante sur les 20 premières secondes de la conversation. La figure 7.9 fournit un exemple de séquence audio-vidéo. Elle présente la valeur du signal (fonction linéaire croissante sur les 20 premières secondes) ainsi que la vérité terrain. Elle fournit aussi pour la séquence, la corrélation entre ces deux signaux.

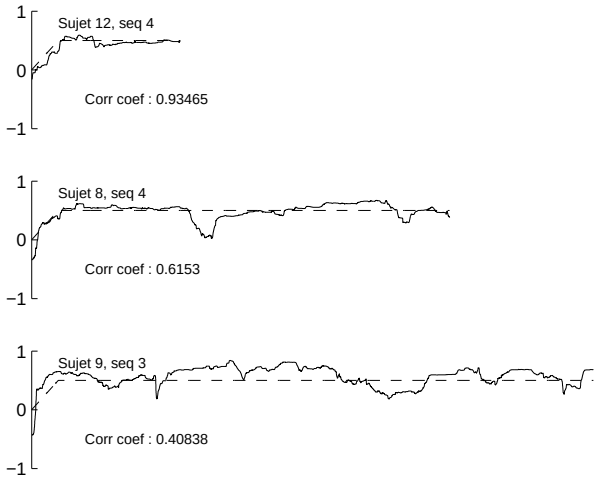


FIGURE 7.9 – Corrélation entre la vérité terrain de la dimension power (trait plein) et le temps de réponse de l'annotateur (trait pointillé) sur trois séquences audio-vidéo différentes.

Le tableau 7.4 donne la valeur moyenne des corrélations entre le signal trouvé et les vérités terrains. Nous constatons que cette information est pertinente pour les dimensions arousal et power, cela confirme l'étude réalisée précédemment (figure 7.8).

Dimensions	Temps de réponse de l'annotateur
Arousal	**0.43**
Valence	0.12
Power	**0.56**
Expectancy	-0.10

TABLE 7.4 – Corrélation moyenne entre le temps de réponse de l'annotateur et les dimensions décrivant l'émotion

La valeur de 20 secondes pour le temps de réponse moyen d'un annotateur a été définie empiriquement par le calcul de corrélation entre ce signal et la vérité terrain.

7.4.3 Le temps depuis le début de la conversation (contexte)

Phase préliminaire d'analyse : L'analyse des labels de vérité terrain montre que l'expectancy varie de façon assez similaire lors des conversations. En début de conversation, l'expectancy est basse. Puis celle-ci est plus élevée. Cela peut-être dû au fait que le sujet découvre l'agent émotionnel dans les premières secondes de la conversation. Il a alors une expectancy basse.

Extraction de la caractéristique : Nous avons modélisé cet aspect par un signal créneau (valeur élevée la première minute, puis valeur basse pour le reste de la séquence). La valeur de 1 minute a été trouvée empiriquement, c'est celle qui maximise la corrélation entre ce signal créneau et la vérité terrain. La figure ci-après 7.10 fournit un exemple de séquence audio-vidéo. Elle présente la valeur du signal (signal créneau) ainsi que la vérité terrain. Elle fournit aussi pour la séquence, la corrélation entre ces deux signaux.

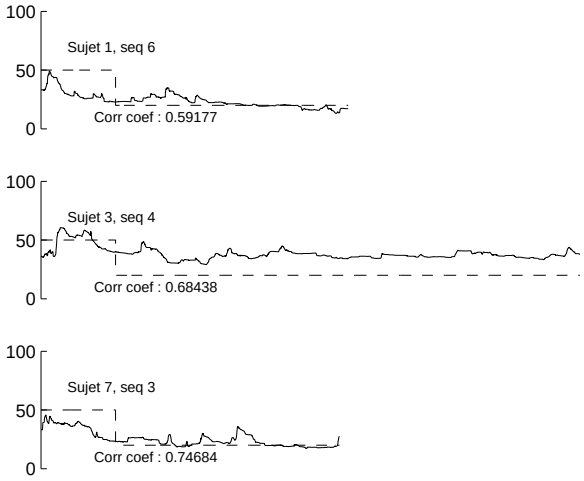


FIGURE 7.10 – Corrélation entre la vérité terrain de la dimension expectancy (trait plein) et le temps depuis le début de la conversation (trait pointillé) sur trois séquences audio-vidéo différentes.

Le tableau 7.5 donne la valeur moyenne des corrélations entre le signal trouvé et les vérités terrains. Nous constatons que cette information est pertinente pour l'expectancy. Les valeurs sont aussi élevées pour arousal et power mais cela est dû au phénomène de temps de réponse de l'annotateur (cf. 7.4).

Dimensions	Début de conversation
Arousal	0.19
Valence	-0.04
Power	0.26
Expectancy	**-0.21**

TABLE 7.5 – Corrélation moyenne entre le début de la conversation et les dimensions décrivant l'émotion

7.5 Synthèse

Pour connaitre les sources pertinentes des variations des dimensions définissant l'émotion, nous avons calculés la corrélation entre les signaux définis précédemment et les labels de vérité terrain pour chaque séquence de la base de développement et avons calculé la moyenne sur ces séquences. Une valeur élevée de la moyenne signifie une forte corrélation et donc un signal pertinent pour définir les variations émotionnelles. Les résultats des sections précédentes (et de la section 8.3 pour le rire) sont synthétisés dans le tableau 7.6.

Dimensions	Rire	Mouvement de la tête	Tours de parole	Débit de paroles	Temps de réponse	Longueur du discours
Arousal	**0.30**	**0.15**	-0.02	0.08	**0.43**	0.19
Valence	**0.41**	0.08	0.09	0.03	0.12	-0.04
Power	0.10	-0.02	-0.13	0.11	**0.56**	0.26
Unexpectancy	0.11	0.03	**0.25**	-0.03	-0.10	**-0.21**

TABLE 7.6 – Corrélation moyenne entre les caractéristiques pertinentes et les dimensions caractérisant l'émotion.

A noter que pour éviter l'impact du temps de réponse des annotateurs sur le calcul de la corrélation, les premières secondes des séquences ont été supprimées du calcul de corrélation des caractéristiques autres que celle de temps de réponse de l'annotateur. Il faut aussi noter que l'impact du type émotionnel de l'agent ne peut pas être mesuré en termes de corrélation puisqu'il donne une information statistique constante sur l'état émotionnel moyen du sujet sur la séquence. Néanmoins, cette valeur va être utilisée dans les règles (chapitre 9) pour définir l'offset de la séquence ou fournir une information lorsqu'aucune autre caractéristique n'est détectée.

Chapitre 8

Les Expressions Faciales

Nous nous focalisons dans ce chapitre sur l'analyse des expressions du visage, qui est l'objet principal de cette thèse. Pour rappel, la participation au challenge avait pour objectif de tester et d'améliorer la méthode d'analyse automatique présentée dans la partie précédente (partie II) sur les données du challenge. Les principales différences avec la partie précédente sont les suivantes :

- Les annotations des visages sont réalisées automatiquement par un modèle actif d'apparence (AAM) générique, et non manuellement (travaux de Hanan Salam). Cela implique des erreurs possibles au niveau de la détection de la forme du visage.
- Les données sont issues de séquences vidéo et ne sont plus uniquement des images indépendantes. Nous utilisons cette nouvelle information afin de déduire une information de plus haut niveau sur l'expression faciale (détection du rire et non plus uniquement du sourire).
- Les données sont issues de conversations réelles entre un sujet et un agent émotionnel. Cela signifie qu'à certains moments, le sujet est en train de parler. L'expression faciale est donc le résultat de plusieurs facteurs (signaux physiologique, visèmes, interactions sociales et états mentaux).
- Les expressions sont spontanées, contrairement aux tests réalisés dans la partie précédente.

Dans ce chapitre, nous présenterons dans un premier temps le système global (section 8.1). Puis nous donnerons quelques informations sur l'analyse automatique des visages (section 8.2) avant de préciser la méthode de détection des rires (8.3).

Sommaire

8.1 Vue Globale de l'Extraction des Expressions du Visage

Le schéma d'analyse des expressions du visage présentée dans la figure 8.1 étend celui de la partie précédente (cf. figure 6.1).

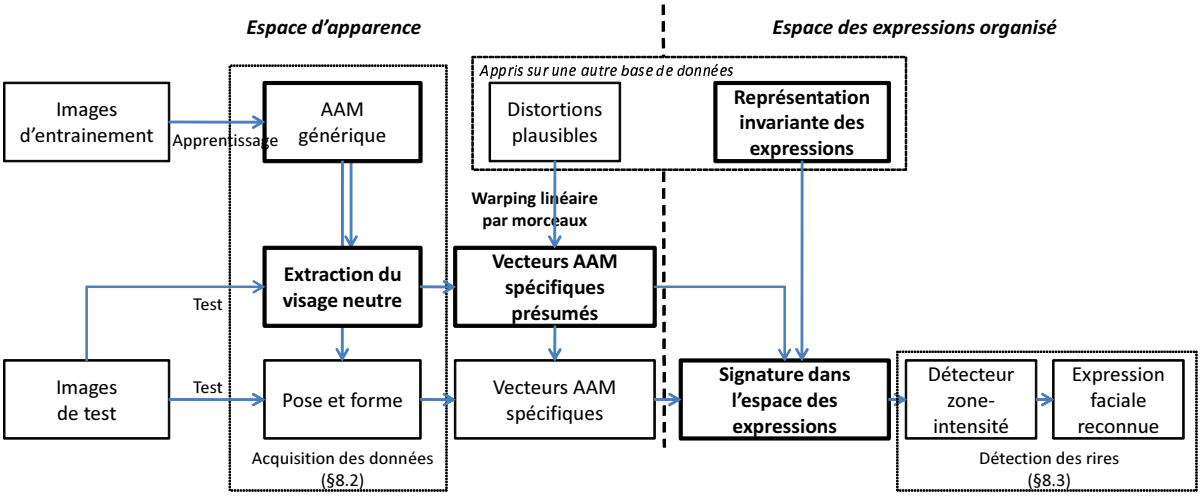


FIGURE 8.1 – Vue globale de l'extraction des expressions faciales. Le visage neutre de chaque sujet ainsi que la forme du visage (points caractéristiques) pour chaque image sont extraits en utilisant un AAM générique. L'espace d'apparence spécifique présumé de la personne est créé en appliquant les distorsions plausibles sur le visage neutre du sujet. L'espace d'apparence spécifique à la personne est transformé dans l'espace des expressions en utilisant l'organisation invariante des expressions faciales.

Nous retrouvons les *distorsions plausibles* qui permettent de créer les *vecteurs AAM spécifiques présumés* par *warping linéaire par morceaux*. Nous retrouvons aussi la *représentation invariante des expressions* faciales utilisée pour définir la *signature* de l'expression.

8.2 L'Acquisition des Données du Visage

Le premier apport concerne l'acquisition des données du visage. Un modèle AAM générique est appris sur les données d'entrainement. Le modèle AAM générique est utilisé pour trouver la forme (points caractéristiques) des visages de tests, inconnus du système. La technique mise en œuvre ainsi que les contributions ne sont pas mentionnées ici. Il s'agit des travaux de Hanan Salam [103]. Le visage neutre est extrait de ces données en appliquant le processus décrit dans la section 5.3.5. La forme des visages (points caractéristiques) est utilisée pour définir les vecteurs d'apparence des expressions inconnues par projection sur le modèle AAM présumé de la personne (voir annexe A.3).

8.3 La Détection des Rires

Le second apport concerne la représentation de l'expression. Plutôt que d'utiliser directement l'information de signature d'une image, nous utilisons les signatures de plusieurs images consécutives. Cela permet à la fois de lisser les résultats obtenus concernant l'expression et d'avoir une information de plus haut niveau sur l'expression. C'est ce que nous avons fait concernant le rire. Le rire est une expression de haut niveau issue des expressions de sourire. La figure 8.2 montre la trajectoire de la signature lors d'un rire.

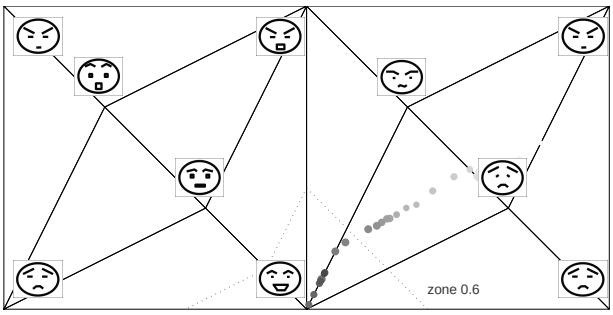


FIGURE 8.2 – Trajectoire de la signature des expressions d'un sujet dans l'espace des expressions lors d'un rire. Chaque image de la séquence vidéo est représentée par un point. Plus le point est clair, plus il est ancien.

Dans notre espace de représentation des expressions du visage, la trajectoire correspondant à un rire est caractérisée de façon suivante :

- Intensité forte des expressions faciales.
- Expressions proches de l'expression correspondant au sourire qui est l'une des 8 expressions utilisées pour la création de l'organisation des expressions.
- Maintien de ces deux composantes (intensité et zone) pendant une durée importante.

Un rire peut donc être caractérisé par une expression de sourire prolongé et de forte intensité. Nous avons défini un rire, de façon continue, en calculant le nombre d'expressions de sourire de forte intensité sur une fenêtre glissante de 40 secondes. Nous avons défini de façon empirique :

- la notion de *forte intensité* en prenant les expressions dont l'intensité est supérieure à 0.3.
- la notion d'*expression de sourire* en prenant les expressions dont la direction vers l'expression de sourire de l'organisation des expressions est supérieure à 0.6 (en d'autres termes, les expressions qui entourent l'expression de sourire).

Nous parlons alors de calcul de zone-intensité. La corrélation entre ce paramètre et les dimensions émotionnelles est présenté dans le tableau 8.1. Nous constatons une forte corrélation pour les composantes arousal et valence, comme attendu. Les figures 8.3 et 8.4 fournissent des exemples de séquences audio-vidéo. Elles présentent la valeur du signal ainsi que la vérité terrain. Elles fournissent aussi pour la séquence, la corrélation entre ces deux signaux.

Dimensions	Rire
Arousal	**0.30**
Valence	**0.41**
Power	0.10
Unexpectancy	0.11

TABLE 8.1 – Corrélation moyenne entre le rire et les dimensions caractérisant l'émotion.

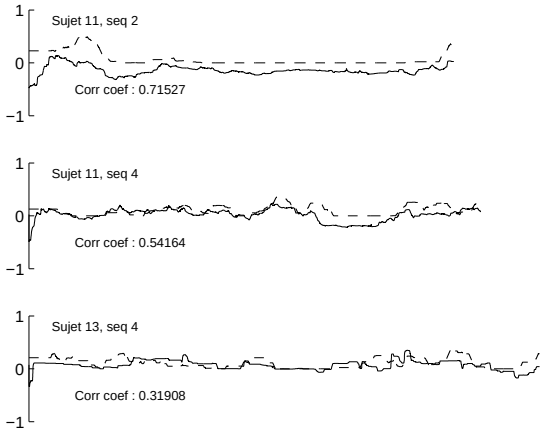


FIGURE 8.3 – Corrélation entre la vérité terrain de la dimension arousal (trait plein) et le rire (trait pointillé) sur trois séquences audio-vidéo différentes.

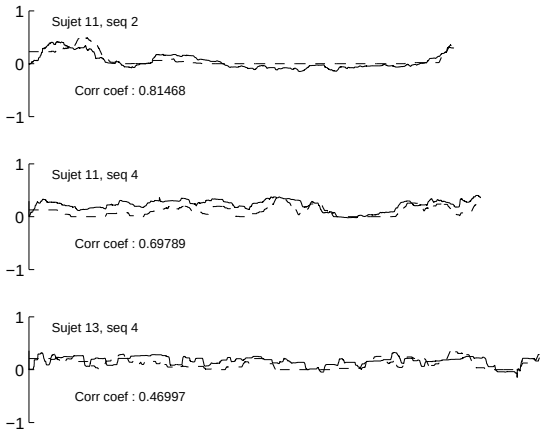


FIGURE 8.4 – Corrélation entre la vérité terrain de la dimension valence (trait plein) et le rire (trait pointillé) sur trois séquences audio-vidéo différentes.

Chapitre 9

Processus d'Apprentissage

Un système multimodal présente deux étapes clefs : l'extraction des caractéristiques pertinentes et la fusion de ces caractéristiques pour fournir l'information souhaitée (voir figure 6.14). En ce qui nous concerne, il s'agit d'une information sur des composantes émotionnelles des sujets. Les chapitres précédents se sont focalisés sur la première étape : l'extraction des caractéristiques pertinentes. Nous allons dans ce chapitre aborder la seconde étape : la fusion de ces caractéristiques.

Nous souhaitons proposer un système mettant en œuvre les observations humaines réalisées lors de la phase initiale d'analyse. Nous proposons ici un système de fusion, basé sur des règles, qui utilise un système d'inférence floue (*Fuzzy inference system* FIS). Par ailleurs, nous souhaitons analyser l'impact de la méthode de fusion sur le système global. Pour cela, nous proposons de mettre en œuvre un second système de fusion inspiré des travaux des gagnants du challenge AVEC 2012 [82]. La méthode est basée sur des fonctions de bases radiales (*Radial Basis Functions* RBF). Après avoir présenté ces deux méthodes de fusion, nous les comparerons et nous discuterons des résultats du challenge.

Sommaire

9.1 Les Systèmes de Fusion

9.1.1 Système d'inférence floue

Pour fusionner ces différentes caractéristiques, nous avons utilisé un système d'inférence floue de type Mamdani [105]. Les caractéristiques utilisées sont les suivantes :

- l'opérateur AND est le produit,
- l'implication floue est le produit,
- l'agrégation est réalisée par la somme,
- la défuzzification est réalisée par la méthode des centroïdes.

Les règles retenues sont décrites dans le tableau 9.1. Elles correspondent aux règles identifiées par les observations humaines ainsi qu'aux résultats des corrélations présentées dans le tableau 7.6.

Ce tableau se lit en ligne de la façon suivante. Première ligne : *Pendant le temps de réponse de l'annotateur, arousal est très bas, la valence est entre moyen bas et moyen et le power est très bas.* Troisième ligne : *En dehors du temps de réponse de l'annotateur, si l'agent est Poppy, l'arousal est élevé et la valence est élevée.*

Le tableau peut aussi se lire en colonne. Troisième colonne : *Le power est très bas pendant le temps de réponse de l'annotateur et moyen sinon.* Quatrième colonne : *L'expectancy est très basse lorsque les phrases sont longues ou que la conversation est établie. Elle est très élevée lorsque les phrases sont courtes ou qu'il s'agit d'un début de conversation.*

	Règles	Arousal	Valence	Power	Expectancy
1	Pendant RT	TB	MMB	TB	
2	Hors RT			M	
3	Hors RT et Agent est Poppy	H	H		
4	Hors RT et Agent est Spike	H	MB		
5	Hors RT et Agent est Obadiah	B	B		
6	Hors RT et Agent est Prudence	M	MH		
7	Hors RT et Agent est inconnu	M	M		
8	Hors RT et Rire est élevé	TH	TH		
9	Phrases longues				TB
10	Phrases courtes				TH
11	Début de conversation				TH
12	Conversation établie				TB

TABLE 9.1 – Règles floues du système pour chaque dimension : Valence, Arousal, Power et Expectancy. RT : Temps de réponse de l'annotateur. TB : Très Bas, B : Bas, MB : Moyen Bas, MMB : entre MB et M, M : Moyen, MH : Moyen Haut, H :Haut, TH : Très Haut.

Le système d'inférence floue, de part sa méthode d'agrégation, gère les cas non explicités, c'est-à-dire par exemple lorsque nous avons des *phrases courtes en conversation établie*. Pour plus de détail sur les fuzzifications, défuzzifications et règles utilisées, voir l'annexe B.

9.1.2 Radial Basis Function

La seconde méthode de fusion mise en œuvre est inspirée de [82]. Un ensemble d'exemples représentatifs de caractéristiques pertinentes est calculé par un algorithme des k-moyennes (*k*-

means). Ces exemples représentatifs sont utilisés comme centres de fonctions de base radiales, pour la prédiction de l'émotion. Ce calcul est réalisé pour chacune des 4 dimensions utilisées pour représenter les émotions. Plus précisément, les données d'entrée sont les caractéristiques pertinentes trouvées dans les chapitres précédents et concaténées dans un vecteur. La première étape est l'extraction d'exemples représentatifs. Pour réaliser cette tâche, nous utilisons la méthode des k-moyennes. Les centres des clusters sont choisis comme exemples représentatifs pour la dimension caractérisant l'émotion. Le label émotionnel associé à chaque exemple est la valeur moyenne des labels du cluster. Le tableau 9.2 montre les 5 exemples représentatifs (5 clusters) calculés pour la dimension arousal. La valeur 5 a été trouvée empiriquement. C'est celle qui permet d'obtenir les meilleures prédictions.

Temps de réponse	0.99	0.99	0.99	**0.32**	0.99
Rire	0.04	0.02	0.03	0.09	**0.25**
Arousal de l'agent émotionnel	**-0.24**	**-0.08**	**0.04**	-0.06	-0.00
Arousal	-0.19	-0.11	0.03	-0.28	0.05

TABLE 9.2 – Exemples obtenus par clusterisation (méthode des k-moyennes) pour la dimension arousal.

La seconde étape est la prédiction. La prédiction est réalisée par des fonctions de bases radiales (RBF) centrées sur les exemples calculés précédemment. Soit $\{\mathbf{x_j} \in \mathbb{R}^n, j \in [1..m]\}$ les vecteurs des m exemples représentatifs obtenus après l'étape de clustering, et $\{\mathbf{y_j}, j \in [1..m]\}$ les labels associés. La prédiction pour un exemple s décrit par le vecteur $\mathbf{x_s} \in \mathbb{R}^n$ est donné par :

$$\hat{y}(s) = \frac{\sum_m^{j=1} e^{\frac{-\left\|x_s - x_j\right\|^2}{\sigma^2}} y_j}{\sum_m^{j=1} e^{\frac{-\left\|x_s - x_j\right\|^2}{\sigma^2}}} \tag{9.1}$$

où la distance utilisée est la distance euclidienne et σ est la largeur de la fonction gaussienne de base radiale.

9.1.3 Comparaison de l'apprentissage des deux méthodes

Nous pouvons remarquer que les mêmes règles sont implémentées dans les deux systèmes. Par exemple, pour arousal, nous pouvons analyser les clusters (tableau 9.2) par les règles (tableau 9.3) :

Cluster	Si	arousal est
1	l'agent est Obadiah	bas
2	l'agent est Prudence ou inconnu	moyen
3	l'agent est Poppy ou Spike	haut
4	durant le temps de réponse de l'annotateur	très bas
5	le rire est élevé	très haut

TABLE 9.3 – Interprétation des exemples représentatifs sous la forme de règles pour la dimension arousal.

En effet, dans le tableau 9.2, le cluster 4 a une valeur faible pour le temps de réponse de l'annotateur (ce qui signifie *durant le temps de réponse de l'annotateur*), une valeur moyenne pour

le rire et l'arousal de l'agent et la plus petite valeur des clusters pour arousal. Nous interprétons ces valeurs comme : *si nous sommes pendant le temps de réponse de l'annotateur, peu importe quelles sont les autres valeurs d'entrée, l'arousal est très bas*, ce qui est la première règle du tableau 9.1. Le cluster 5 (tableau 9.2) a une valeur élevée pour le rire, une valeur moyenne pour l'arousal de l'agent et une valeur proche de 1 pour le temps de réponse de l'annotateur (ce qui signifie *pas pendant le temps de réponse de l'annotateur*) et la plus grande valeur des clusters pour arousal. Nous interprétons ces valeurs comme *si nous ne sommes pas pendant le temps de réponse de l'annotateur et que le rire est élevé, quelque soit les autres valeurs d'entrée, l'arousal est très élevé*, ce qui est la règle 8 du tableau 9.1.

Dans cet exemple, les règles obtenues correspondent exactement à celles du système d'inférence floue (tableau 9.1). Cela est dû au fait que chaque composante d'entrée (temps de réponse de l'annotateur, arousal de l'agent et rire) ont des valeurs qui dictent une valeur de sortie de l'arousal indépendamment des autres composantes d'entrée. Dans le cas général, la k-moyenne nécessite plus de clusters pour prendre en compte la combinatoires des composantes d'entrée. Par exemple, pour la dimension valence, les clusters sont présentés dans le tableau 9.4.

Temps de réponse	**0.32**	0.99	0.99	0.99	0.99	0.99	0.99
Rire	0.09	**0.02**	**0.20**	**0.01**	**0.20**	**0.04**	**0.27**
Valence de l'agent émotionnel	0.05	**-0.10**	**-0.10**	0.10	0.10	**0.28**	**0.28**
Valence	0.06	-0.10	0.20	0.08	0.21	0.25	0.33

TABLE 9.4 – Exemples obtenus par clusterisation par la méthode des k-moyennes pour la dimension valence.

Nous pouvons remarquer que le premier cluster correspond à la première règle du système d'inférence floue (tableau 9.1), mais les autres clusters combinent le rire et la valence de l'agent. Par exemple, les clusters 2 et 3 correspondent tous deux à l'agent avec un faible valence (Obadiah), mais avec un niveau différent de rire (pas de rire pour le cluster 2 et haut niveau de rire pour le cluster 3). Nous constatons la même combinaison pour les clusters 4 et 5 (l'agent avec une valence moyenne) et pour les clusters 6 et 7 (un agent avec une forte valence). Dans le système d'inférence floue, cette combinatoire se fait directement par l'agrégation, de sorte que les règles n'ont pas besoin de prendre en compte plusieurs entrées. Nous pouvons constater le même comportement avec l'expectancy (cf. tableau 9.5) :

Phrases (courtes/longues)	**0**	**0**	**1**	**1**
Conversation (établie/début)	**0**	**1**	0	1
Expectancy	34.5	38.5	30.6	29.4

TABLE 9.5 – Exemples obtenus par clusterisation par la méthode des k-moyennes pour la dimension expectancy.

Les clusters 1 et 2 correspondent à l'impact du discours lorsque les phrases sont courtes : *Quand les phrases sont courtes, si le discours commence (cluster 2), l'expectancy est très élevé, alors que si le discours est établie (cluster 1), l'expectancy est moyenne (combinaison d'une très grande et très faible valeur)*. Les clusters 3 et 4 correspondent à la 9ième règle : *Si phrases sont longues, l'expectancy est très faible*.

Dans le cas général, les règles sont difficiles à extraire à partir de l'analyse des clusters. Au contraire, le système d'inférence floue utilise des règles intelligibles, facilite l'ajout et la suppres-

sion des règles et des données d'entrée (système *boîte blanche*). Ce système nous permet facilement de différencier les règles génériques (celles issues des variations effectives d'émotion des sujets) des règles spécifiques au challenge (voir section 9.2.3 : discussion sur les caractéristiques du contexte). Mais les systèmes d'inférence floue sont aussi connus pour n'être efficaces que lorsqu'il y a peu de données d'entrée.

9.2 Résultats du Challenge

Cette section présente les résultats du challenge et discute de plusieurs aspects : tout d'abord de l'impact des annotateurs et la véracité de la vérité terrain ; ensuite des points clefs des différentes briques d'un tel système.

(Dés)accord des annotateurs : Nous avions effleuré ce point en introduction de cette partie III, dans le paragraphe précisant notre démarche. Nous avions alors indiqué que *nous nous sommes rapidement interrogés sur la pertinence de la vérité terrain fournie dans le cadre du Challenge, et notamment sur l'impact de l'outil de sa labellisation et sur les désaccords entre les annotateurs*. Voici ici quelques données étayant cet aspect. Nous avons calculé la corrélation des labels de vérité terrain inter-annotateurs. Le tableau 9.2 donne ces résultats.

Dimensions	**Annotateurs**
Arousal	0.44
Valence	0.53
Power	0.51
Expectancy	0.33
Moyenne	0.45

TABLE 9.6 – Corrélation moyenne entre un annotateur et les autres annotateurs.

Les valeurs de corrélation entre les évaluateurs humains utilisés pour l'annotation de la réalité terrain sont faibles (moyenne d'environ 0.45), ce qui signifie que les évaluateurs humains sont souvent en désaccord sur les variations de l'émotion. La figure 9.1 illustre cette observation avec deux exemples. Sur la gauche, les deux évaluateurs sont d'accord (corrélation de 0.80), tandis que sur la droite, ils sont en désaccord (corrélation négative de -0.22). Ces exemples montrent la difficulté de labellisation de la réalité terrain en termes de dimension continue. Ce désaccord sur les étiquettes de réalité de terrain peut influencer fortement l'apprentissage automatique des systèmes. C'est pourquoi nous avons choisi de mettre en place un système *boîte blanche* (voir la démarche explicitée en introduction de la partie III). De façon plus générale, cette constatation pose la question de la pertinence d'un tel challenge. Pour notre part, comme indiqué en introduction de cette partie, il s'agissait avant tout de tester et améliorer notre représentation dans un environnement différent (séquences audio-visuelles, données spontanées, interprétation en terme d'émotion, environnement multimodal).

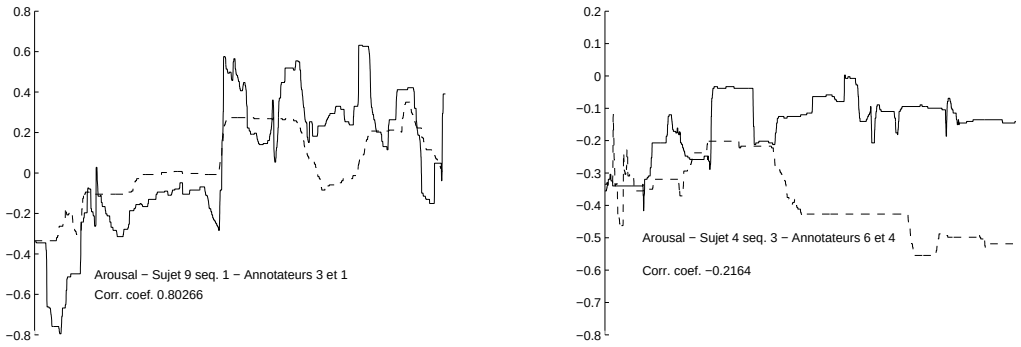


FIGURE 9.1 – Comparaison de la labellisation de la dimension arousal de mêmes séquences audio-vidéo réalisée par deux annotateurs différents (première séquence audio-vidéo à gauche : annotateurs d'accord ; seconde séquence audio-vidéo à droite : annotateurs en désaccord).

9.2.1 Les résultats en chiffres

La figure 9.2 donne les résultats officiels du challenge AVEC 2012. Dans le cadre du challenge, c'est la méthode de fusion FIS qui a été mise en œuvre. L'équipe est arrivée seconde, talonnant les premiers [82] et avec de biens meilleurs résultats que les troisièmes [83]. Nous allons dans cette section analyser ces résultats globaux.

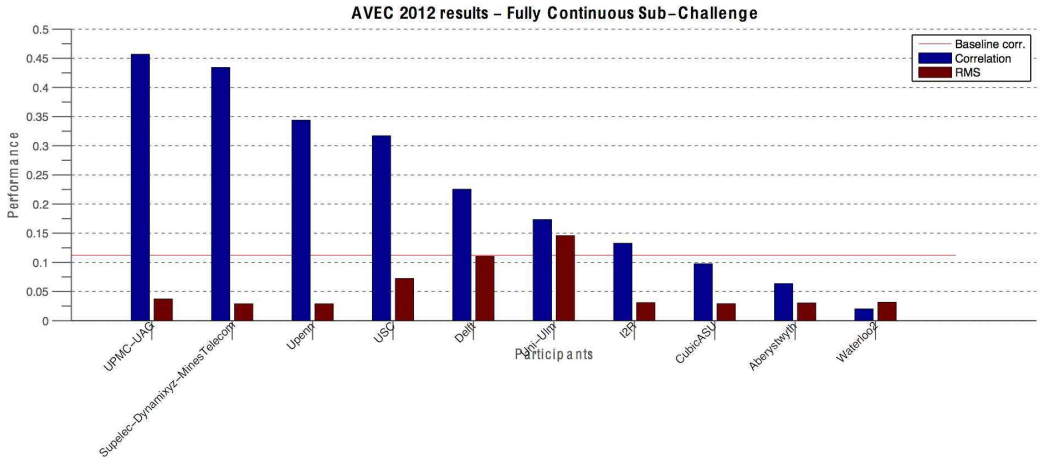


FIGURE 9.2 – Résultats officiels du Challenge AVEC 2012. Résultats disponibles sur le site `http://sspnet.eu/avec2012/`.

Le tableau 9.7 reprend les résultats des trois premières équipes et montre les résultats de notre système (en séparant les deux techniques de fusion) sur les bases d'entrainement, de développement et de test. L'apprentissage a été réalisé sur les bases d'entrainement et de développement. Les résultats sur la base de tests (1er, 2nd et 3^{ième}) sont les résultats officiels du challenge, calculés par

les organisateurs du challenge. Nous affichons aussi à nouveau le coefficient de corrélation moyen entre un évaluateur et les autres (dernière colonne du tableau) pour comparaison.

Dimensions	**Entrain.**		**Dévelop.**		**Test**				**Annot.**
	FIS	RBF	FIS	RBF	FIS	RBF	[82]	[83]	
AVEC 2012					2nd		1er	3ième	
Arousal	0.40	0.36	0.52	0.47	0.42	0.42	0.61	0.36	0.44
Valence	0.39	0.40	0.47	0.43	0.42	0.42	0.34	0.22	0.53
Power	0.61	0.59	0.59	0.58	0.57	0.57	0.56	0.48	0.51
Expectancy	0.37	0.37	0.30	0.32	0.33	0.32	0.31	0.33	0.33
Moyenne	0.44	0.43	0.47	0.45	0.43	0.43	0.46	0.34	0.45

TABLE 9.7 – Résultats globaux des deux systèmes de fusion : système d'inférence floue (FIS) et fonctions de base radiales (RBF). Coefficients de corrélation moyens entre la prédiction et la vérité terrain. A titre de comparaison, la dernière colonne donne la corrélation moyenne entre un annotateur et les autres annotateurs et les deux dernières colonnes de la partie Test donnent les résultats des vainqueurs du challenge AVEC 2012 et des compétiteurs arrivés en 3ième position.

Nous pouvons tout d'abord constater la stabilité de nos résultats sur les différentes bases de données, quel que soit le système de fusion utilisé, ce qui signifie que les deux méthodes se généralisent correctement.

Même si les valeurs restent basses (moyenne aux alentours de 0.44), elles sont similaires aux taux d'un annotateur humain ayant réalisé l'annotation de la vérité terrain (moyenne de 0.45) et des gagnants du challenge (0.46 sur la base de test), les autres challengers étant loin derrière (0.34 sur la base de test pour le troisième).

Les résultats des annotateurs humains montrent que nous ne sommes pas aussi bons sur la dimension de valence. Pour définir la valence, nous avons principalement utilisé le rire (durée d'un sourire d'intensité élevé). D'autres informations sur les expressions faciales, telles que l'abaissement des sourcils (AU4) auraient pu nous donner une information sur une diminution de valence et ainsi améliorer les résultats.

La différence sur valence entre les bases d'entrainement et de développement est principalement due à la détection du rire. Nous n'avons pas pu détecter le sourire pour l'un des sujets de la base d'entrainement car il avait la partie basse du visage en dehors de la vidéo.

9.2.2 Comparaison des performances des systèmes de fusion

La comparaison des résultats des deux systèmes FIS et RBF basés sur les mêmes données (voir figure 9.2) montre que les résultats de prédiction sont similaires pour les deux techniques de fusion. Nous pouvons donc penser que ce n'est pas la technique de fusion qui est une question clé dans ces systèmes, mais les caractéristiques initiales. Cette conclusion est confortée par la comparaison des résultats de notre système avec fusion par fonctions radiales de base et du système des gagnants de AVEC 2012, qui ont également utilisé des k-moyennes et des fonctions de base radiales (voir le tableau 9.7). En effet, ils ont obtenu des résultats assez similaires pour power et expectancy (0.56 contre 0.57 pour power et 0.31 contre 0.32 pour expectancy), mais de bien meilleurs résultats pour la dimension arousal (0.61 contre 0.42) et de moins bons résultats pour la dimension valence (0.34 contre 0.42). Notre analyse est qu'il manque à notre système une ou plusieurs caractéristiques essentielles utiles à une bonne prédiction de l'arousal, et qu'il manque à leur système une ou de plusieurs caractéristiques essentielles pour la prédiction de la

valence. En effet, pour la dimension arousal, [82] ont obtenus des résultats comparables sur les bases de développement et de tests. Sur la base de développement, les prédictions par modalités sont les suivantes : 0.54 à partir des caractéristiques de forme uniquement, 0.50 avec des caractéristiques d'apparence globale, 0.47 avec des caractéristiques d'apparence locale et 0.45 avec les caractéristiques audio. Leur système de fusion donne un résultat global de 0.64 sur cette base de développement. Nous pouvons donc penser que les informations issues de l'audio et/ou des déformations faciales locales fournissent une information pertinente pour l'arousal que nous n'avons pas utilisée. A noter que dans notre système (avec fusion RBF ou FIS), nous nous sommes placés dans le cadre d'une fusion sur des informations de haut niveau (de type *rire*) alors que les gagnants du challenge ont utilisé des informations de bas niveau (forme du visage et texture du visage par analyse en composantes principales par exemple).

9.2.3 Les caractéristiques pertinentes dans un contexte général d'analyse d'émotion

Comme nous venons de le voir, l'extraction des caractéristiques pertinentes est une question clé dans la prédiction de l'émotion. Dans ce paragraphe, nous discutons de leur impact et de leur prise en compte dans un système *générique* d'analyse d'émotion (c'est-à-dire non dédié au challenge).

L'**impact du rire** sur l'arousal peut être analysé par le fait que nous calculons le sourire avec une grande intensité sur une longue durée, qui sont les caractéristiques du rire. D'autres types de sourire pourraient être utilisés pour améliorer les résultats. Ces caractéristiques ne semblent pas liées aux données du challenge.

Le fait que l'**expectancy augmente au cours de la conversation** doit être confirmée par l'analyse d'autres bases de données présentant des conversations pour vérifier si cette information peut être utilisée dans un contexte général. Nous pouvons noter qu'Ozkan et al. [79] ont analysé cette observation d'une manière différente. Ils ont expliqué que les participants perçoivent le contexte de la conversation. Ainsi, plus ils sont engagés dans la conversation, plus leurs émotions (quelles qu'elles soient) deviennent intenses. Par conséquent, ils ont utilisé la même fonction, basée sur la durée de conversation, pour chacune des 4 dimensions caractérisant l'émotion. Dans notre système, nous séparons d'un côté la durée de la conversation utilisée pour l'expectancy (comme les participants perçoivent le contexte de la conversation, ils sont moins surpris) et de l'autre le temps de réponse de l'évaluateur utilisé pour les 3 autres dimensions. Nous avons donc utilisé deux signaux d'entrée différents, nommés *Temps depuis le début de la conversation* et *Temps de réponse de l'annotateur*.

La corrélation entre la **longueur des phrases** et l'expectancy doit elle aussi être confirmée par l'analyse d'autres bases de données présentant des conversations pour vérifier si cette information peut être utilisée dans un contexte général.

L'**impact émotionnel de l'agent** peut être interprété comme l'empathie et la contagion des émotions, qui peuvent être des informations d'indice dans le cas général où l'information audio-visuelle n'est pas disponible ou incertaine. Même si cette caractéristique semble être générale, la méthode d'extraction de l'information, elle, est dans notre système en partie spécifique aux données du challenge (nom de l'agent), et en partie générique (thème de la conversation comme *fun*).

Pour finir, pour ce qui est de la réalité terrain du début des séquences, que nous avons interprété comme un **temps de réponse de l'annotateur**, il s'agit de notre point de vue d'une donnée purement spécifique au challenge.

9.2.4 Boîte noire ou boîte blanche ?

Nous avons vu dans la section 9.2.1 que le système de fusion influençait peu les résultats, faisant penser que le point clef du système concerne plus particulièrement la détection et l'extraction des caractéristiques multimodales.

Contrairement aux systèmes statistiques (de type *boîte noire*) dont le RBF fait partie et pour lesquels la phase d'apprentissage doit être réitérée à chaque nouvelle base de donnée ou à chaque nouveau contexte, notre système utilisant la fusion FIS peut être facilement adapté en ajoutant ou supprimant des règles qui sont soit spécifiques à la base de données (ici, le temps de réponse de l'annotateur), soit spécifiques au contexte lors de scénario réels (ici, le fait d'avoir des conversations avec un agent émotionnel).

9.3 Conclusion Intermédiaire

Ce challenge a tout d'abord permis de tester la représentation des expressions

- sur des données réelles (c'est-à-dire spontanées) et
- sur des séquences vidéo (et non des images statiques).

Les résultats de corrélation entre la réalité terrain et les valeurs obtenues (coefficient de corrélation de 0.43 en moyenne sur l'ensemble de test) montrent qu'il y a encore des améliorations à faire afin de déterminer les variations d'émotions, même si nous effectuons en moyenne aussi bien que les évaluateurs humains. Bien que nos résultats globaux soient encourageants (2nde place), nous pouvons regretter de n'avoir pu mettre en œuvre que peu de caractéristiques faciales (détection du rire uniquement). Cela est principalement dû au fait que nous nous sommes focalisé sur les variations importantes d'émotion et non sur des variations plus subtiles. L'ajout d'autres types de mouvements du visage (mouvements des sourcils, autres types de sourires, mouvement des yeux) pourraient améliorer les résultats.

Pour autant, ce challenge nous a permis de mettre en œuvre un système original. Nous sommes les seuls concurrents à avoir opté pour un système *boîte blanche* et les seuls à avoir utilisé des informations provenant de l'agent émotionnel (empathie) comme caractéristiques pertinentes pour détecter l'émotion des sujets. De plus, ce challenge nous a permis d'utiliser notre représentation des expressions afin de détecter des expressions de plus haut niveau (ici le rire). La méthode proposée consiste à définir une zone et une intensité et à intégrer les expressions correspondant à ces critères sur une période d'une séquence audio-vidéo.

Nous nous sommes aussi intéressés à l'impact des méthodes de fusion dans un système global d'analyse d'émotion. Les mêmes modalités (audio, vidéo et contexte) et les mêmes caractéristiques sont fusionnées, soit avec un système d'inférence floue soit avec un système de fonctions de base radiales. Ils fournissent tous deux la prédiction de 4 dimensions : valence, arousal, power et expectancy. Les expériences montrent que le choix de la technique de fusion affecte peu les résultats, ce qui semble dire que l'extraction de caractéristiques est le point clef de la détection d'émotions.

Concernant les caractéristiques pertinentes permettant de détecter les variations d'émotions, nous avons extrait le rire (impact sur arousal et valence), la gestion des tours de paroles et la longueur du discours (impact sur expectancy), la contagion d'émotion et l'empathie (impact sur arousal et valence). Ces caractéristiques et leurs impacts nous semblent généralisables dans un contexte plus global. Cela doit être confirmé sur d'autres bases de données. Pour ce qui est du temps de réponse de l'annotateur, cette caractéristique nous semble spécifique à la vérité terrain du challenge et à l'outil utilisé pour l'annotation de cette vérité terrain. Elle ne peut pas être prise en compte dans un système global dans la mesure où elle ne reflète pas les variations d'émotion des sujets.

Ce challenge nous a sensibilisé à la difficulté d'obtenir une vérité terrain *objective*, notamment lorsqu'il s'agit d'émotion. Nous avons observé de grandes variations entre les annotateurs qui induit un bruit important sur la vérité terrain. C'est pour cette raison que nous avons opté pour un système *boîte blanche*. Contrairement aux systèmes statistiques, dont l'apprentissage doit être retraité pour chaque nouvelle base de données ou chaque nouveau contexte, le système d'inférence floue peut être facilement adapté en supprimant ou en ajoutant des règles qui sont spécifiques à la base de données ou au contexte des scénarios de la vie réelle.

Bilan et Perspectives

Garde-toi, tant que tu vivras, de juger des gens sur la mine.
Jean de La Fontaine - *Le cochet, le chat, et le souriceau - 1668*

10.1 Résumé des Contributions et Résultats

Dans ce paragraphe, nous résumons les principales contributions et résultats présentés dans cette thèse. Comme indiqué dans l'introduction, le type d'applications visées impose quatre contraintes pour représenter une expression :

- **Précision** de façon à distinguer des expressions proches
- **Exhaustivité** de façon à distinguer des expressions non connues
- **Robustesse** pour gérer les différentes morphologies
- **Flexibilité** pour s'adapter aux différents individus sans phase préalable d'apprentissage

Pour chacune de ces contraintes, nous avons proposé une solution innovante dans la partie II. Concernant la **précision** du système, nous avons proposé de travailler sur des modèles de visage spécifiques à la personne. Pour s'affranchir de la contrainte de **robustesse**, nous avons défini une *organisation des expressions*, et avons montré que cette organisation était similaire entre les différents sujets. Nous nous sommes alors basé sur cette organisation pour définir l'espace des expressions et représenter une expression *inconnue* par son intensité et sa position relative par rapport aux autres expressions. Nous répondons ainsi à la contrainte d'**exhaustivité**. Pour finir, concernant la **flexibilité** du système, nous avons proposé de générer l'espace d'apparence d'une personne inconnue en synthétisant ses expressions basiques.

Cette représentation a tout d'abord été testée pour la caractérisation des expressions non incluses dans les bases d'apprentissage (chapitre 6). Les résultats montrent que la représentation proposée donne de meilleurs résultats que les méthodes traditionnelles basées sur les vecteurs d'apparence issus des modèles actifs d'apparence. Cette représentation a aussi été mise à l'épreuve dans un système plus complet (Partie III). Ces travaux ont été réalisés dans le cadre du challenge AVEC 2012 [80], dont l'objectif était de détecter des variations émotionnelles de sujets lors de conversations avec un agent émotionnel. Nous avons ainsi pu constater que la méthode était applicable avec des expressions spontanées et nous avons proposé la représentation d'une expression de plus haut niveau : le rire. Cette expression est définie comme un sourire intense de durée importante. Nous avons mis en évidence l'impact de cette expression de rire sur deux composantes émotionnelles : la valence et l'arousal. Pour finir, nous avons extrait d'autres informations permettant de caractériser les variations émotionnelles des sujets, et les avons fusionnées via un système multimodal d'inférence floue. Parmi les autres caractéristiques pertinentes permettant de définir l'émotion, nous pouvons noter la contagion d'émotion (l'émotion de l'agent émotionnel et celle

de l'interlocuteur sont fortement corrélées). Ces travaux nous ont permis d'obtenir la seconde place du challenge, avec des taux avoisinant ceux des vainqueurs.

10.2 Perspectives

Impact du mode de représentation des visages Les travaux ont été réalisés sur une description des visages basée sur les vecteurs d'apparence issus de modèles actifs d'apparence. Bien que nous ayons montré que la structure de l'espace des expressions n'était pas impactée par le type de données (forme et/ou texture), il serait intéressant de remplacer ces vecteurs par une autre représentation des visages (par exemple une représentation sous la forme de Local Binary Patterns - LBP).

Caractérisation des types d'applications Deux axes principaux peuvent être envisagés à partir de la représentation des expressions proposée. Le système peut :

- **Apprendre** Il s'agit alors d'adapter le système à l'espace réel des expressions du sujet. Cela permettrait d'augmenter les performances dans le cadre des sujets inconnus afin d'atteindre des performances équivalentes à celles obtenues sur des sujets connus. L'espace présumé proposé est considéré comme approximatif. Cette étape est intéressante pour les applications nécessitant une connaissance précise du sujet. C'est par exemple le cas pour le maintient à domicile des personnes âgées, dont l'objectif est de lever une alerte lorsque le comportement du sujet change ou encore pour les applications de type *serious game* pour lesquelles l'analyse et l'interprétation des expressions doivent être les plus pertinentes possibles.
- **Faire apprendre** Il s'agit dans ce cas d'utiliser le système pour indiquer une consigne. L'espace présumé proposé est considéré comme référence. C'est au sujet de s'adapter pour correspondre à cet espace. C'est le cas dans le projet REPLICA dont l'objectif est de pouvoir évaluer l'adéquation entre l'expression réelle du sujet et l'expression attendue (celle du système). La définition d'une métrique permettant de mesurer l'écart entre l'expression réalisée et la consigne est actuellement à l'étude.

Représentation des mouvements bucco-faciaux Actuellement, la représentation proposée est en cours de mise en œuvre, non plus sur l'ensemble du visage, mais sur la partie inférieure du visage (analyse des mouvements bucco-faciaux). Ces travaux se font dans le cadre du projet REPLICA, visant à fournir un outil d'aide à l'apprentissage des mouvements bucco-faciaux pour les enfants atteints de paralysie cérébrale. Les premiers résultats montrent que l'*universalité* de l'organisation des expressions semble être vérifiée dans ce contexte. Ces premiers résultats doivent être confirmés sur un plus grand nombre de sujets.

Étude d'autres représentations des expressions D'autres travaux en cours concernent l'utilisation de modèles bilinéaires dans le système global. Les systèmes bilinéaires permettent de séparer l'identité de l'expression du sujet. Une étude préliminaire a été menée pour analyser via ces modèles les expressions non contenues dans les bases d'apprentissage. Ces travaux ont fait l'objet de deux publications (VCIP 2013 et GRETSI 2013). Ces modèles pourraient aussi être utilisés pour synthétiser les expressions plausibles des sujets inconnus.

Publications

Publications dans des revues internationales avec comité de lecture

[CVIU2013] **Catherine Soladié**, Nicolas Stoiber, Renaud Séguier *Invariant Representation of Facial Expressions for Blended Expression Recognition on Unknown Subjects* Computer Vision and Image Understanding (CVIU), Elsevier, Vol 117, Issue 11, Nov. 2013, pp. 1598-1609

[IJACSci2013] **Catherine Soladié**, Nicolas Stoiber, Renaud Séguier *Continuous Facial Expression Representation for Multimodal Emotion Detection* International Journal of Advanced Computer Science (IJACSci), Vol3, No5, Mai 2013, pp. 202-216

Publications dans des conférences internationales avec comité de lectures et proceedings

[VCIP2013] **Catherine Soladié**, Nicolas Stoiber, Renaud Séguier *Bilinear Decomposition for Blended Expression Representation* IEEE Visual Communications and Image Processing (VCIP), Malaysia, Nov. 2013

[ICIP2012] **Catherine Soladié**, Nicolas Stoiber, Renaud Séguier *A new invariant representation of facial expressions : definition and application to blended expression recognition* IEEE International Conference on Image Processing (ICIP), Orlando, Florida, U.S.A., Sept.-Oct. 2012, pp. 2617-2620

[ICMI2012] **Catherine Soladié**, Hanan Salam, Catherine Pelachaud, Nicolas Stoiber, Renaud Séguier *A Multimodal Fuzzy Inference System Using a Continuous Facial Expression Representation for Emotion Detection* ACM International Conference on Multimodal Interaction (ICMI), 2nd International Audio/Visual Emotion Challenge and Workshop -AVEC 2012 , Santa Monica, California, U.S.A., Oct. 2012, pp. 493-500

Publications dans des conférences nationales avec comité de lecture et proceedings

[GRETSI2013] **Catherine Soladié**, Nicolas Stoiber, Renaud Séguier *Création de l'espace des expressions faciales à partir de modèles bilinéaires asymétriques* XXIVème Colloque GRETSI, Brest, France, 3-6 Septembre 2013

Annexe

Annexe A

Modèles Actifs d'Apparence

Les modèles actifs d'apparence [18] permettent d'apprendre un modèle de forme à partir d'images annotées et de retrouver cette forme dans une nouvelle image. Par exemple, un modèle de visage peut être appris à partir de visages dont le contour des yeux, des sourcils, du nez, de la bouche et du visage ont été annotés. Le modèle permet ensuite de trouver automatiquement ces données (contour des yeux, des sourcils, du nez, de la bouche et du visage) sur un visage inconnu.

A.1 Apprentissage du Modèle

Le principe de l'apprentissage du modèle consiste à extraire les déformations principales de la forme ainsi que les déformations principales de la texture (niveau de gris) contenue dans cette forme. Ces deux informations (forme et texture) sont corrélées puisque ce sont les variations de texture qui permettent de définir la forme. Par exemple, dans un visage, c'est la différence de couleur entre la peau et les sourcils qui permet de tracer le contour (donc la forme) des sourcils. Le modèle apprend la corrélation entre les déformations de forme et les déformations de texture.

Les N images sont annotées par m points caractéristiques. Pour chaque image $\mathbf{i}$, les coordonnées de ces points caractéristiques sont concaténés pour former un vecteur $\mathbf{s_i}$, qui représente la forme de l'image. Les intensités des pixels contenu dans la zone intérieure de la forme de l'image forment le vecteur $\mathbf{g_i}$, qui représente la texture. Pour détecter les distorsions de forme et de texture, une analyse en composantes principales (ACP) est réalisée sur chacun des deux vecteurs :

$$\mathbf{s_i} = \overline{\mathbf{s}} + \mathbf{\Phi_s}.\mathbf{b_i^s} \tag{A.1}$$

$$\mathbf{g_i} = \overline{\mathbf{g}} + \mathbf{\Phi_t}.\mathbf{b_i^t} \tag{A.2}$$

où $\overline{\mathbf{s}}$ et $\overline{\mathbf{g}}$ sont les formes et texture moyennes, $\mathbf{\Phi_s}$ et $\mathbf{\Phi_t}$ les matrices formées par les vecteurs propres issus de l'ACP et $\mathbf{b_i^s}$ et $\mathbf{b_i^t}$ sont la décomposition de $\mathbf{s_i}$ et $\mathbf{g_i}$ sur les modes propres. $\mathbf{s_i}$ et $\mathbf{g_i}$ sont appelées vecteurs de forme et vecteurs de texture.

Pour prendre en compte la corrélation qui existe entre la forme et la texture, une troisième ACP est réalisée sur un vecteur qui concatène le vecteur de forme et le vecteur de texture $\mathbf{b_i} = [w_s.\mathbf{b_i^s}|\mathbf{b_i^t}]$ (w_s est un facteur de mise à l'échelle qui assure que les vecteurs de forme et de texture ont des variances comparables).

$$\mathbf{b_i} = \mathbf{\Phi}.\mathbf{c_i} \tag{A.3}$$

où $\mathbf{\Phi}$ est la matrice formée par les vecteurs propres de l'ACP, et $\mathbf{c_i}$ est le vecteur d'apparence.

Seules les principales déformations sont conservées à chaque étape.

Avant l'apprentissage des principales déformations (de forme et de texture), les formes sont préalablement alignées et les textures sont normalisées.

A.2 Prédiction de la Forme

La forme d'une nouvelle image est calculée de façon itérative en utilisant le modèle. A chaque itération j, les paramètres c_j du modèle permettent de définir une texture. En effet, la nature linéaire du modèle permet d'exprimer la texture en fonction du vecteur $\mathbf{c_j}$.

$$\mathbf{b_j} = \mathbf{\Phi}.\mathbf{c_j} = \begin{pmatrix} \mathbf{\Phi_{cs}} \\ \mathbf{\Phi_{ct}} \end{pmatrix}.\mathbf{c_j} \tag{A.4}$$

$$\mathbf{g_j} = \overline{\mathbf{g}} + \mathbf{\Phi_t}.\mathbf{\Phi_{ct}}.\mathbf{c_j} \tag{A.5}$$

Cette texture est comparée à la texture réelle de l'image et l'erreur entre les deux textures est calculée. A partir de la matrice permettant de prédire les modifications à appliquer sur les paramètres c en fonction de l'erreur obtenue sur la texture, une nouvelle prédiction c_{j+1} est réalisée. Les conditions d'arrêt sont soit que l'erreur entre la prédiction de texture et la texture réelle ne diminue plus, soit que le nombre maximum d'itération est atteint.

La forme est alors retrouvée par l'équation suivante :

$$\mathbf{s_j} = \overline{\mathbf{s}} + \mathbf{\Phi_s}.\mathbf{W_s^{-1}}\mathbf{\Phi_{cs}}.\mathbf{c_j} \tag{A.6}$$

A.3 Calcul du Vecteur d'Apparence par Projection

Cette section présente le calcul du vecteur d'apparence d'une image dont on connait les points caractéristiques. Il s'agit de projeter la forme (points caractéristiques) sur le modèle actif d'apparence.

La forme est préalablement alignée sur la forme moyenne du modèle. Lorsque tous les modes sont conservés lors de la création du modèle d'apparence, la matrice $\mathbf{\Phi_{cs}}$ est carrée. Elle est aussi inversible de part la nature des données d'apprentissage. Le vecteur d'apparence peut être alors directement déduit de la forme par l'équation suivante :

$$\mathbf{c_i} = \mathbf{\Phi_{cs}^{-1}}.\mathbf{W_s^{-1}}\mathbf{\Phi_s^T}.(\mathbf{s_i} - \overline{\mathbf{s}}) \tag{A.7}$$

Annexe B

Fuzzification, Règles Floues et Défuzzification Utilisées pour la Détection de l'Émotion

Cette annexe présente la configuration du système d'inférence floue mise en œuvre dans le cadre du challenge AVEC 2012 pour prédire les variations d'émotion des sujets. Les fonctions d'appartenance et les règles sont issues de l'analyse réalisée dans les chapitres 7, 8 et 9.

B.1 Fonctions d'Appartenance

Les fonctions d'appartenance permettent de fuzzifier les données du système.

B.1.1 Fonctions d'appartenance d'entrée

Periode La période caractérise le moment de la séquence. Elle définit s'il s'agit ou pas d'un début de séquence. Une seule fonction d'appartenance est mise en œuvre et elle est associée à la donnée d'entrée définissant le temps de réponse de l'annotateur. .

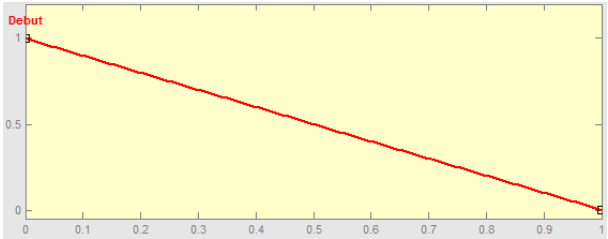


FIGURE B.1 – Fonction d'appartenance PERIODE.

Poppy, Fun, Spike, Angry, Annoy, Obadiah, Sad et Prudence Il s'agit des mêmes formes de fonctions d'appartenance pour chacune des données d'entrée relatives au caractère émotionnel de l'agent. Les données d'entrée prennent deux valeurs : 1 pour indiquer qu'il s'agit de cet agent

émotionnel, 0 pour indiquer qu'il ne s'agit pas de cet agent émotionnel. Deux fonctions d'appartenance sont utilisées : une fonction permettant d'indiquer si l'agent est présent et une autre pour indiquer si l'agent est absent.

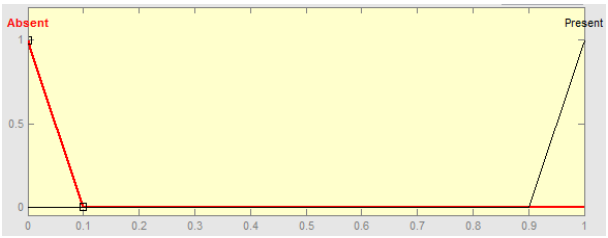


FIGURE B.2 – Fonctions d'appartenance POPPY, FUN, SPIKE, ANGRY, ANNOY, OBADIAH, SAD et PRUDENCE.

Smile La fonction d'appartenance *smile* est définie pour les données d'entrée caractérisant le rire. Une seule fonction d'appartenance est mise en œuvre car seul le fait que le rire soit présent est analysé dans les règles (une absence de rire n'est pas analysée par exemple). La fonction d'appartenance proposée est une fonction linéaire croissante de 0 à 1 permettant de garder le caractère continu de la donnée d'entrée.

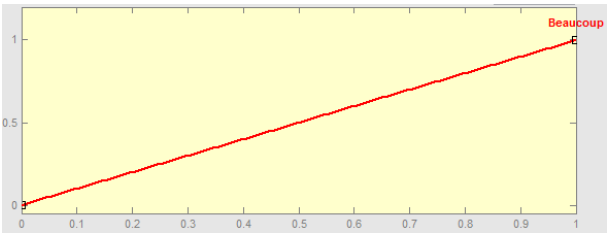


FIGURE B.3 – Fonction d'appartenance SMILE.

Prise de parole Les fonctions d'appartenance *PriseDeParole* permettent de gérer les tours de parole. Deux fonctions d'appartenance sont proposées : une première pour les phrases courtes et une seconde pour les phrases longues. Ces fonctions d'appartenance conservent les données d'entrée qui valent 0 si les phrases sont courtes et 1 si les phrases sont longues.

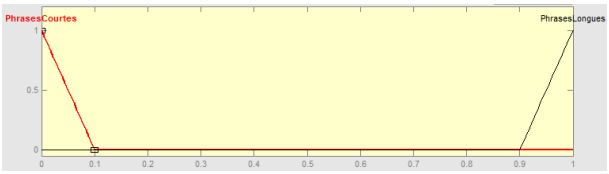


FIGURE B.4 – Fonctions d'appartenance PRISE DE PAROLE.

Discours Les fonctions d'appartenance *DiscoursSignalStructure* permettent de gérer le moment du discours. Deux fonctions d'appartenance sont proposées : une première pour indiquer sur la conversation commence et une seconde pour indiquer que la conversation est établie. Ces fonctions d'appartenance conservent les données d'entrée qui valent 0 si le discours est établi et 1 si le discours commence.

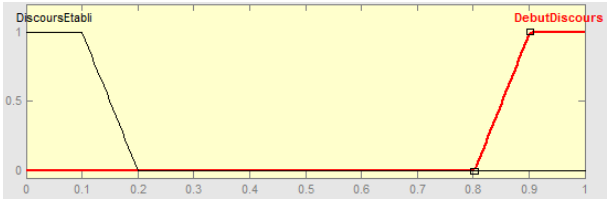


FIGURE B.5 – Fonctions d'appartenance DISCOURS.

B.1.2 Fonctions d'appartenance de sortie

Ces fonctions d'appartenance permettent de donner une valeur aux caractéristiques du tableau 9.1 (TB : Très Bas, B : Bas, ML : Moyen Bas, MMB : entre MB et M, M : Moyen, MH : Moyen Haut, H :Haut, TH : Très Haut). Les valeurs sont issues des analyses statistiques des labels (moyenne, écart type et valeur initiale) réalisée sur toutes les séquences audio-vidéo (voir données sur la figure 7.8). Pour les dimensions émotionnelles arousal et valence, les analyses statistiques (moyenne et écart type) sur les séquences audio-vidéo associée à chaque agent émotionnel ont aussi été prise en compte (voir données sur la figure 7.7).

Arousal Les fonctions d'appartenance *Faible*, *Moyen* et *Fort* correspondent aux valeurs issues des analyses statistiques (moyenne et écart type) sur les séquences audio-vidéo associée à chaque agent émotionnel (voir figure 7.7). La fonction d'appartenance *Très faible* correspond à la valeur de début de séquence (voir figure 7.8). La fonction d'appartenance *Très fort* permet de prendre en compte le rire.

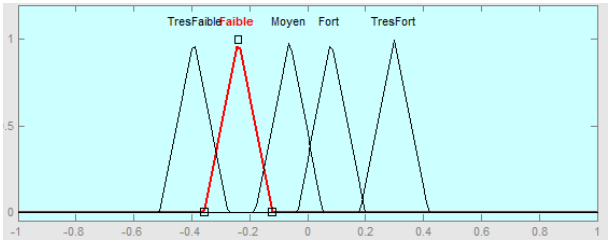


FIGURE B.6 – Fonctions d'appartenance AROUSAL.

Valence Les fonctions d'appartenance *Faible*, *MoyenFaible*, *Moyen*, *Fort* et *Moyen fort* correspondent aux valeurs issues des analyses statistiques (moyenne et écart type) sur les séquences audio-vidéo associée à chaque agent émotionnel (voir figure 7.7). La fonction d'appartenance *Très*

faible correspond à la valeur de début de séquence (voir figure 7.8). Les fonctions d'appartenance *Très fort* et *TTTF* permettent de prendre en compte le rire.

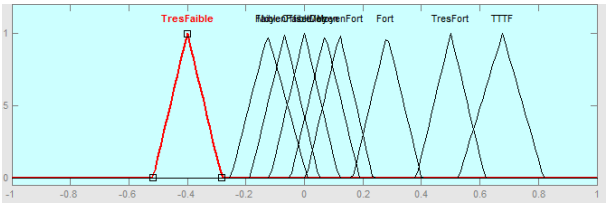


FIGURE B.7 – Fonctions d'appartenance VALENCE.

Power La fonction d'appartenance *Moyen* correspond à la valeur moyenne et la fonction d'appartenance *Très faible* correspond à la valeur de début de séquence (voir figure 7.8).

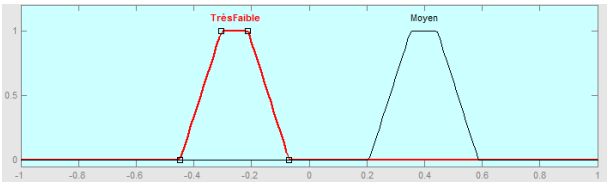


FIGURE B.8 – Fonctions d'appartenance POWER.

Expectancy N'ayant pas d'analyse précise sur l'impact de chacune des deux données d'entrée *PriseDeParole* et *DiscoursSignalStructure*, seules deux fonctions d'appartenance ont été mises en œuvre : *Très faible* et *TresFort*. Elles permettent aussi de gérer la valeur de début de séquence (voir figure 7.8).

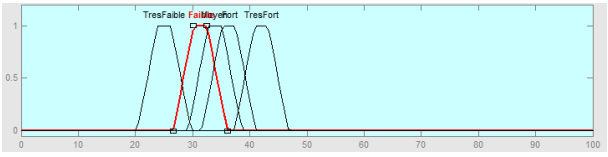


FIGURE B.9 – Fonctions d'appartenance EXPECTANCY.

B.2 Règles Floues

Les règles floues permettent de combiner les entrées. Elles définissent précisément, à partir des éléments précédents, les règles identifiées dans le tableau 9.1.

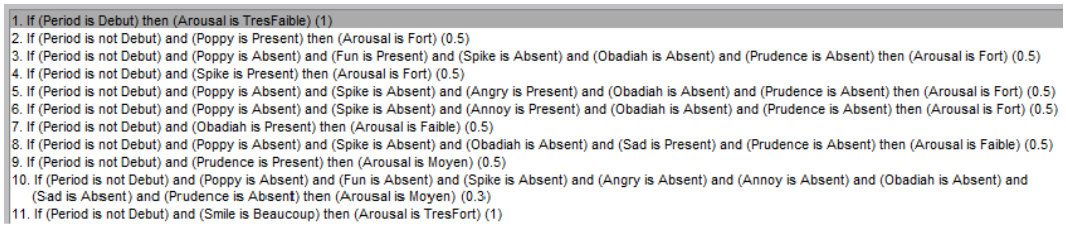


FIGURE B.10 – Règles pour AROUSAL.

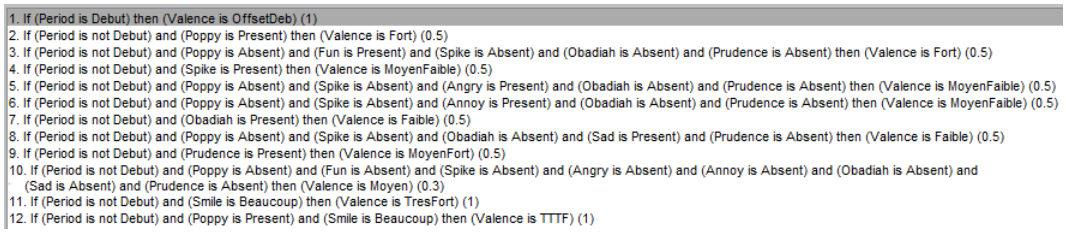


FIGURE B.11 – Règles pour VALENCE.

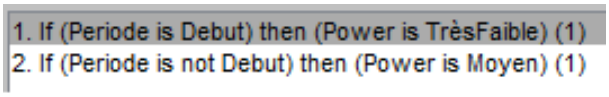


FIGURE B.12 – Règles pour POWER.

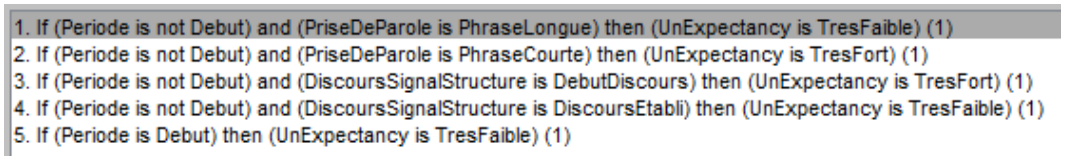


FIGURE B.13 – Règles pour EXPECTANCY.

Annexe C

Analyse préliminaire des séquences audio-visuelle du challenge AVEC 2012

Cette annexe présente l'analyse préliminaire des données du challenge AVEC 2012. Cette analyse a été réalisée en visualisant les séquences audio-vidéo du challenge des bases d'entrainement et de développement, ainsi que les labels émotionnels représentant les vérités terrains associées à ces séquence audio-vidéo. Quatre dimensions émotionnelles sont proposées dans le cadre du challenge : arousal, valence, power et expectancy. L'analyse préliminaire est découpée selon ces quatre dimensions émotionnelles.

Il s'agit ici d'un extrait du document de travail réalisé dans le cadre du challenge.

C.1 Arousal

C.1.1 Définition

Excitation / Activité / Éveil (vs. Mouvements lents)

C.1.2 Exemples d'émotions

+ peur, stress, amour
- contentement (satisfaction), déception, tristesse

C.1.3 Indices pour détecter l'offset

Offset élevé quand :

- La personne parle beaucoup
- La personne bouge beaucoup
- Les gestes sont plus larges
- Les gestes sont plus rapides
- Il y a beaucoup de clignement des yeux
- La personne a un débit élevé de paroles
- La voix est forte (exemples : devel 4 et devel14)

Offset faible quand :

- Il y a des silences entre les tours de parole
- Il y a peu de mouvements de la tête

C.1.4 Indices pour détecter la variation

- Augmentation lors de sourires, de rires (exemple : devel 1, 2 :07)
- Augmentation lorsque le sujet montre de la vivacité (exemple : devel 2, 0 :53)
- Modification du son (exemple : devel 17, au début plutôt atone, lent puis plus réveillé)
- Variation des indices d'offset dans la séquence (exemple : train 24, arousal négatif car il ne bouge pas trop / à 2 :20 l'arousal augmente car il bouge plus, parle plus vite)

C.1.5 Commentaires

Pour le clignement des yeux, ce n'est pas vrai pour la peur par exemple. Ne faudrait-il pas regarder plutôt si le rythme des clignements est ou pas régulier au cours de la séquence ?

Parfois corrélé à Valence. Les deux dimensions ont peut-être été annotées simultanément sur un disque (voir outil FEELTRACE [71]) ?

C.2 Valence

C.2.1 Définition

valence = pleasure
Content (vs. pas content)

C.2.2 Exemples d'émotions

+ joie
- dégout, colère, tristesse

C.2.3 Indices pour détecter l'offset

Offset élevé quand :

- voix plutôt positive c'est pourquoi la valence est élevée alors que le visage n'est pas spécialement souriant (exemple : devel 7)

Offset faible quand :

- voix plus basse (devel 15)
- moins de sourires (devel 15)

C.2.4 Indices pour détecter la variation

Analyse des images pour la détection des rires et utilisation des tags <LAUGH> Exemples :

- devel 1, 2 :07
- train 2, 1 :46
- train 4, 2 :52 et 3 :20

- train 19, 0 :10
- train 21, 6 :30
- train 22, 6 :10

C.2.5 Commentaires

Parfois corrélé à Arousal. Les deux dimensions ont peut-être été annotées simultanément sur un disque ?

C.3 Power

C.3.1 Définition

power = dominance = potency
En contrôle, en maîtrise (vs. pas en contrôle)

C.3.2 Exemples d'émotions

+ l'intérêt, la haine, la colère
- la peur, l'anxiété, la tristesse, la honte

C.3.3 Indices pour détecter l'offset

Offset élevé quand :

- voix forte
- bouge la tête dans tous les sens (exemple : devel 14)
- la personne est positive, elle répond (exemple : train 1)

Offset faible quand :

- surprise (exemple : devel 14, 1 :15)

C.3.4 Indices pour détecter la variation

- Analyse des images pour la détection des rires et utilisation des tags <LAUGH> (exemple : devel 1, 2 :07)
- Lié aux phases de dialogue : lorsque la personne écoute, baisse du power et hausse de l'unexpectancy (exemple : devel 7)
- Lié à la prise de parole : lorsque la personne prend la parole, elle reprend aussi le contrôle (exemple : devel 9 à 3 :00)
- Quand la personne parle, elle a un power élevé
- Lever les yeux en l'air : power faible
- Lever un sourcil : négatif, perte de contrôle

C.3.5 Commentaires

Parfois corrélé à Expectancy. Les deux dimensions ont peut-être été annotées simultanément sur un disque ? On a l'impression que les courbes expectancy et power sont inversées. Exemples :

- devel 27,
- train 8,
- train 19,
- train 12

En général une valeur de 0.5 : on contrôle un peu tout le temps. Proposition : donner une valeur constante 0.5 sauf quand valeur forte d'expectancy.

C.4 Expectancy

C.4.1 Définition

expectancy = predictability
Prévisibilité (vs. Surprise)
Attention, sur les courbes, c'est plutôt un-expectancy

C.4.2 Exemples d'émotions

+ surprise
- les autres

C.4.3 Indices pour détecter l'offset

En général autour de 40.
Offset élevé quand :

- Les personnes parlent beaucoup
- Peu de pauses dans le discours

Offset faible quand :

- Hésitation dans la parole (exemple : devel 8, il ne reprend pas la parole, dit des petites phrases, des *perhaps* ; devel 9 : des *I don't know*)
- Tours de parole très courts (devel 9)
- Réalise des hochements de tête petits et saccadés (exemple : devel 8)
- Reste quasi immobile (exemples : devel 8 et devel 9)
- Regarde en l'air (exemple : devel 9)

C.4.4 Indices pour détecter la variation

- Lié aux phases de dialogue : lorsque la personne écoute, baisse du power et hausse de l'unexpectancy (exemple : devel 7)
- Rupture dans la scène (exemple devel 14, 1 :15)
- Quand il y a des rires provoqués par l'avatar (exemples : devel 22, 0 :27 ; train 4, 3 :21 ; train 17, 1 :46 ; train 19, 2 :16)
- Lié aux pauses - réflexions (exemple devel 10)

C.4.5 Commentaires

Parfois corrélé à Power. Les deux dimensions ont peut-être été annotées simultanément sur un disque ?

Dimension peu utilisée dans la littérature.

Table des figures

Liste des tableaux

Bibliographie

[1] B. Fasel and J. Luettin, “Automatic facial expression analysis : a survey,” *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[2] T. D. Bui, D. Heylen, and A. Nijholt, “Combination of facial movements on a 3D talking head,” in *Computer Graphics International, 2004. Proceedings*, pp. 284–290, 2004.

[3] E. A. Boyle, A. H. Anderson, and A. Newlands, “The effects of visibility on dialogue and performance in a cooperative problem solving task,” *Language and speech*, vol. 37, no. 1, pp. 1–20, 1994.

[4] G. M. Stephenson, K. Ayling, and D. R. Rutter, “The role of visual communication in social exchange,” *British Journal of Social and Clinical Psychology*, vol. 15, no. 2, pp. 113–120, 1976.

[5] A. Mehrabian, “Communication without words,” *Psychological today*, vol. 2, pp. 53–55, 1968.

[6] P. Ekman and W. V. Friesen, “Facial action coding system,” 1977.

[7] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 46–53, 2000.

[8] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 5–pp, 2005.

[9] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, “Bosphorus database for 3d face analysis,” in *Biometrics and Identity Management*, pp. 47–56, Springer, 2008.

[10] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “Avec 2011-the first international audio/visual emotion challenge,” *Affective Computing and Intelligent Interaction*, pp. 415–424, 2011.

[11] R. Cowie, E. Douglas-Cowie, and C. Cox, “Beyond emotion archetypes : Databases for emotion modelling using neural networks,” *Neural Networks*, vol. 18, no. 4, pp. 371–388, 2005.

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

[13] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pp. 1079–1084, 2010.

[14] I. Hupont, S. Baldassarri, and E. Cerezo, "Facial emotional classification : from a discrete perspective to a continuous emotional space," *Pattern Analysis & Applications*, 2013.

[15] J. N. Bassili, "Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face.," *Journal of personality and social psychology*, vol. 37, no. 11, p. 2049, 1979.

[16] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivation and Emotion*, vol. 14, no. 3, pp. 195–214, 1990.

[17] D. Matsumoto, "Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an american sample," *Motivation and emotion*, vol. 17, no. 2, pp. 107–123, 1993.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 681–685, 2001.

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

[20] P. Comon, "Independent component analysis," *Higher-Order Statistics*, pp. 29–38, 1992.

[21] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[22] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image and Vision Computing*, vol. 24, no. 6, pp. 605–614, 2006.

[23] J.-T. Kim and D. Kim, "Gaze tracking with active appearance models," in *Proceeding of The 7th POSTECH-KYUTECH Joint Workshop On Neuroinfomatics*, pp. 90–92, 2007.

[24] W. J. Ryan, D. L. Woodard, A. T. Duchowski, and S. T. Birchfield, "Adapting starburst for elliptical iris segmentation," in *Biometrics : Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pp. 1–7, 2008.

[25] D. W. Hansen and A. E. Pece, "Iris tracking with feature free contours," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pp. 208–214, 2003.

[26] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International journal of computer vision*, vol. 8, no. 2, pp. 99–111, 1992.

[27] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 5, no. 4, pp. 349–359, 1999.

[28] O. Sorkine, "Differential representations for mesh processing," in *Computer Graphics Forum*, vol. 25, pp. 789–807, 2006.

[29] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 200–205, 1998.

[30] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2, pp. II–370, 2005.

[31] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 860–865, 2011.

[32] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 743–756, 1997.

[33] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[34] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[35] L. J. P. Van der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality reduction : A comparative review," *Journal of Machine Learning Research*, vol. 10, pp. 1–41, 2009.

[36] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[37] P. Ekman *et al.*, *Universals and cultural differences in facial expressions of emotion*. University of Nebraska Press Lincoln, 1971.

[38] A. Wojdel, L. J. M. Rothkrantz, and J. C. Wojdel, "Fuzzy-logical implementation of co-occurence rules for combining AUs," *Proc. CGIM 2003*, 2003.

[39] H. Liu and P. Wu, "Comparison of methods for smile deceit detection by training AU6 and AU12 simultaneously," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 1805–1808, 2012.

[40] M. Pantic and L. J. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

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

[42] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation : the standard, implementation and applications*. Wiley, 2003.

[43] S. Machines, "Faceapi," *URL : http ://www. seeingmachines. com/product/faceapi*, 2009.

[44] M. Kenji, "Recognition of facial expression from optical flow," *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.

[45] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions : The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[46] Y. L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," *Handbook of face recognition*, pp. 247–275, 2005.

[47] H. Wang and N. Ahuja, "Facial expression decomposition," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 958–965, 2003.

[48] B. Abboud and F. Davoine, "Bilinear factorisation for facial expression analysis and synthesis," in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 152, pp. 327–333, 2005.

[49] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear elastically deformable models with application to 3d face and facial expression recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pp. 1–8, 2008.

[50] N. Stoiber, R. Seguier, and G. Breton, "Automatic design of a control interface for a synthetic face," in *Proceedings of the 14th international conference on Intelligent user interfaces*, pp. 207–216, 2009.

[51] Y. Chang, C. Hu, M. Turk, *et al.*, "Manifold of facial expression," in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 28–35, 2003.

[52] Y. Chang, C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–520, 2004.

[53] C. Shan, S. Gong, and P. McOwan, "Appearance manifold of facial expression," *Computer Vision in Human-Computer Interaction*, pp. 221–230, 2005.

[54] Y. Cheon and D. Kim, "Natural facial expression recognition using differential-AAM and manifold learning," *Pattern Recognition*, vol. 42, no. 7, pp. 1340–1350, 2009.

[55] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann, "Real-time facial expression recognition using a fuzzy emotion model," in *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pp. 1–6, 2007.

[56] D. Matsumoto, B. Willingham, and A. Olide, "Sequential dynamics of culturally moderated facial expressions of emotion," *Psychological science*, vol. 20, no. 10, pp. 1269–1274, 2009.

[57] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences : temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 160–187, 2003.

[58] H. Kobayashi and F. Hara, "Dynamic recognition of basic facial expressions by discrete-time recurrent neural network," in *Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on*, vol. 1, pp. 155–158, 1993.

[59] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden markov models for modeling facial action temporal dynamics," in *Human-Computer Interaction*, pp. 118–127, Springer, 2007.

[60] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[61] K. Schmidt and J. Cohn, "Dynamics of facial expression : Normative characteristics and individual differences," in *Proceedings of Intel. Conf. On Multimedia and Expo*, 2001.

[62] C. Darwin, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[63] S. Baron-Cohen, "Mind reading : the interactive guide to emotions," *Tead*, 2003.

[64] K. R. Scherer, "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, p. 162, 2000.

[65] K. R. Scherer, "Appraisal theory," *Handbook of cognition and emotion*, pp. 637–663, 1999.

[66] K. Sweeney and C. Whissell, "A dictionary of affect in language : I. establishment and preliminary validation," *Perceptual and motor skills*, vol. 59, no. 3, pp. 695–698, 1984.

[67] R. Plutchik, *Emotion, a psychoevolutionary synthesis*. Harper & Row New York, 1980.

[68] I. T. Meftah, N. Le Thanh, and C. Ben Amar, "Towards an algebraic modeling of emotional states," in *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, pp. 513–518, 2010.

[69] Z. Ruttkay, H. Noot, and P. Ten Hagen, "Emotion disc and emotion squares : Tools to explore the facial expression space," in *Computer Graphics Forum*, vol. 22, pp. 49–53, 2003.

[70] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[71] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "'FEELTRACE' : an instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[72] C. Whissell, "The dictionary of affect in language," *Emotion : Theory, research, and experience*, vol. 4, no. 113-131, p. 94, 1989.

[73] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[74] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B. J. Theobald, "The painful face : pain expression recognition using active appearance models," in *Proceedings of the 9th international conference on Multimodal interfaces*, pp. 9–14, 2007.

[75] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain : automated measurement of spontaneousallfacial expressions of genuine and posed pain," in *Proceedings of the 9th international conference on Multimodal interfaces*, pp. 15–21, 2007.

[76] Q. Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on*, vol. 36, no. 5, pp. 862–875, 2006.

[77] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Multimedia, IEEE Transactions on*, vol. 8, no. 3, pp. 500–508, 2006.

[78] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[79] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-HMM," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 477–484, 2012.

[80] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012 : the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 449–456, 2012.

[81] M. Ishii, K. Sato, H. Madokoro, and M. Nishida, "Generation of emotional feature space based on topological characteristics of facial expression images," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pp. 1–6, 2008.

[82] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 501–508, 2012.

[83] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 485–492, 2012.

[84] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods : Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[85] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 205–211, 2004.

[86] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2, pp. ii–1085, 2005.

[87] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 1136–1139, 2006.

[88] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," *Artifical Intelligence for Human Computing*, pp. 91–112, 2007.

[89] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition-a new approach," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–1020, 2004.

[90] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–406, 2005.

[91] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 5117–5120, 2008.

[92] G. A. Ramirez, T. Baltrusaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction*, pp. 396–406, Springer, 2011.

[93] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kachele, M. Schmidt, H. Neumann, and G. Palm, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*, pp. 359–368, Springer, 2011.

[94] J. W. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," in *Shape Modeling Applications, 2004. Proceedings*, pp. 145–156, 2004.

[95] K. Zhang, J. T.-L. Wang, and D. Shasha, "On the editing distance between undirected acyclic graphs," *International Journal of Foundations of Computer Science*, vol. 7, no. 1, pp. 43–58, 1996.

[96] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph," *Pattern recognition letters*, vol. 19, no. 3, pp. 255–259, 1998.

[97] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons," *Biol. skr.*, vol. 5, pp. 1–34, 1948.

[98] A. Goshtasby, "Piecewise linear mapping functions for image registration," *Pattern Recognition*, vol. 19, no. 6, pp. 459–466, 1986.

[99] F. Bookstein, "Principal warps : Thin-plate splines and the decomposition of deformations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 6, pp. 567–585, 1989.

[100] S. Park and D. Kim, "Subtle facial expression recognition using motion magnification," *Pattern Recognition Letters*, vol. 30, no. 7, pp. 708–716, 2009.

[101] C. Martin, U. Werner, and H. M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pp. 1–6, 2008.

[102] N. Sebe, I. Cohen, and T. Huang, "Multimodal emotion recognition," *Handbook of Pattern Recognition and Computer Vision*, pp. 981–256, 2005.

[103] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 493–500, 2012.

[104] C. Soladié, H. Salam, N. Stoiber, and R. Séguier, "Continuous facial expression representation for multimodal emotion detection," *International Journal of Advanced Computer Science*, vol. 3, no. 5, 2013.

[105] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International journal of man-machine studies*, vol. 7, no. 1, pp. 1–13, 1975.

Résumé

De plus en plus d'applications ont pour objectif d'automatiser l'analyse des comportements humains afin d'aider ou de remplacer les experts qui réalisent actuellement ces analyses. Cette thèse traite de l'analyse des expressions faciales qui fournissent des informations clefs sur ces comportements.

Les travaux réalisés portent sur une solution innovante permettant de définir efficacement une expression d'un visage, indépendamment de la morphologie du sujet. Pour s'affranchir des différences de morphologies entre les personnes, nous utilisons des modèles d'apparence spécifiques à la personne. Nous proposons une solution qui permet à la fois de tenir compte de l'aspect continu de l'espace des expressions et de la cohérence des différentes parties du visage entre elles.

Pour ce faire, nous proposons une approche originale basée sur l'organisation des expressions. Nous montrons que l'**organisation des expressions, telle que définie, est universelle** et qu'elle peut être efficacement utilisée pour définir de façon unique une expression : une expression est caractérisée par son intensité et sa position relative par rapport aux autres expressions.

La solution est comparée aux méthodes classiques basées sur l'apparence (ICIP 2012) et montre une augmentation significative des résultats de reconnaissance sur 14 expressions non basiques. La méthode a été étendue à des sujets inconnus. L'idée principale est de créer un **espace d'apparence plausible** spécifique à la personne inconnue en **synthétisant ses expressions basiques** à partir de déformations apprises sur d'autres sujets et appliquées sur le neutre du sujet inconnu (CVIU 2013). La solution est aussi mise à l'épreuve dans un environnement multimodal plus complet dont l'objectif est la **reconnaissance d'émotions lors de conversations spontanées**. Les résultats montrent que la solution est efficace sur des données réelles et qu'elle permet l'extraction d'informations essentielles à l'analyse des émotions (ICMI 2012). Notre méthode a été mise en œuvre dans le cadre du **challenge international AVEC 2012 (Audio/Visual Emotion Challenge) où nous avons fini 2nd**, avec des taux de reconnaissance très proches de ceux obtenus par les vainqueurs. La comparaison des deux méthodes (la nôtre et celles des vainqueurs) semble montrer que l'extraction des caractéristiques pertinentes est la clef de tels systèmes (IJACSci 2013).

Mots clefs Analyse des expressions faciales, Représentation invariante, Tessellation de Delaunay, Variété des expressions, Warping linéaire par morceau, Application à la reconnaissance d'émotions, Contexte multimodal, Système d'inférence floue, Contagion d'émotions