Reconnaissance d'expressions corporelles dans des mouvements de personnes en vue de la synthèse de style

Arthur Crenn

N° d'ordre NNT : 2019LYSE1341

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale ED 512

InfoMaths

Spécialité de doctorat :

Discipline :

Soutenue publiquement le 17/12/2019, par :

Arthur Crenn

Reconnaissance d'expressions corporelles dans des mouvements de personnes en vue de la synthèse de style

---

Devant le jury composé de :

SEGUIER Renaud, professer à l'Université de Windsor	Rapporteur
BOUFAMA Boubakeur, professeur à Centrale Supelec, Rennes	Rapporteur
BOYER Edmond, directeur de recherche à l'INRIA Grenoble Rhônes-Alpes	Examinateur
BURDIN Valérie, professeure à l'Université de Bretagne Occidentale	Examinatrice
HASSAS Salima, professeure à l'Université Claude Bernard Lyon 1	Examinatrice
BOUAKAZ Saida, professeure à l'Université Claude Bernard Lyon 1	Directrice de thèse
KONIK Hubert, maitre de conférences à l'Université Jean Monnet	Co-directeur de thèse
MEYER Alexandre, maitre de conférences à l'Université Claude Bernard Lyon 1	Co-directeur de thèse

UNIVERSITÉ CLAUDE BERNARD - LYON 1

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-président du Conseil d'Administration	M. Didier REVEL
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALIER
Vice-président de la Commission de Recherche	
Directrice Générale des Services	M. Damien VERHAEGHE

COMPOSANTES SANTÉ

Faculté de Médecine Lyon-Est - Claude Bernard	Doyen : M. Gilles RODE
Faculté de Médecine et Maïeutique Lyon Sud Charles. Mérieux	Doyenne : Mme Carole BURILLON
UFR d'Odontologie	Doyenne : Mme Dominique SEUX
Institut des Sciences Pharmaceutiques et Biologiques	Directrice : Mme Christine VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. Xavier PERROT
Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

UFR Biosciences	Directrice : Mme Kathrin GIESELER
Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur : M. Marc BUFFAT
UFR - Faculté des Sciences	Administrateur provisoire : M. Bruno ANDRIOLETTI
UFR (STAPS)	Directeur : M. Yannick VANPOULLE
Observatoire de Lyon	Directeur : Mme Isabelle DANIEL
Ecole Polytechnique Universitaire Lyon 1	Directeur : M. Emmanuel PERRIN
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : M. Gérard PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Institut de Science Financière et d'Assurances	Directeur : M. Nicolas LEBOISNE
ESPE	Administrateur provisoire : M. Pierre CHAREYRON

Table des matières

Résumé

Le thème de ma thèse concerne la reconnaissance et la synthèse d'expressions faciales et corporelles. Notre problématique est d'étudier, de comprendre et d'extraire les éléments qui traduisent l'état émotionnel d'une personne à partir des expressions de son visage et de son corps, dans le but de la reconnaissance et également de la synthèse de l'émotion ou du style dans un geste. Ce double objectif de reconnaissance et de synthèse permettra de généraliser de nouveaux modes d'interactions dans des applications comme les jeux-vidéos, l'interaction homme-machine, etc. En effet, ces applications pourront s'enrichir de scénarios qui pourraient s'adapter à l'état émotionnel de l'utilisateur. Pour répondre à ce problème, le point crucial est de comprendre « où » se situe l'information de style dans une action ou une animation. En effet, un humain sait en quelques secondes caractériser une expression qu'il voit alors que les algorithmes de reconnaissance en sont encore loin, notamment pour les expressions corporelles où à notre connaissance, peu de travaux ont été réalisés comparé à la reconnaissance des expressions faciales. Concernant la reconnaissance des expressions, notre objectif est, dans un premier temps, de proposer des descripteurs capables de reconnaître l'expression portée par un mouvement. Pour cela, le principal verrou est d'arriver à séparer le mouvement réalisé de l'expression perçue. Concernant la reconnaissance d'expressions faciales, nous nous intéressons au problème sociétal lié la protection parentale. Pour cela, il est nécessaire de comprendre et de savoir reconnaitre des expressions faciales chez des enfants. Afin de résoudre ce problème, nous avons construit et proposé une nouvelle base de données dans le but d'aider la communauté de la vision par ordinateur à comprendre les spécificités des expressions faciales de visages d'enfants. Dans un second temps, nous souhaitons également que les différents descripteurs proposés en reconnaissance d'expressions corporelles puissent être utilisés dans le domaine de la synthèse d'animations. En effet, dans le domaine de l'animation, la création d'une action porteuse d'une émotion ou d'un style nécessite énormément de travail, de savoir-faire et de temps afin de proposer des animations stylisées. Par exemple, dans un jeu-vidéo, pour créer de telles animations, il est souvent nécessaire de disposer d'une énorme base de données de mouvements incluant chaque style pour chaque personnage virtuel. Pour disposer d'une telle base de données, on a souvent recours à l'une des 2 méthodes suivantes. La première consiste à procéder à la capture de tous les mouvements réalisés par différents acteurs jouant différents styles. Dans la seconde, c'est le graphiste qui doit réaliser les différentes

animations à la main en utilisant un logiciel d'animation. Notre objectif dans ce cadre est de mettre à profit les descripteurs quantifiant l'expression détectée en reconnaissance d'expressions faciales et corporelles afin de développer des outils capables de changer / éditer le style ou l'expression d'une animation. Ces outils permettront d'assister et de faciliter le travail des graphistes en leur permettant de synthétiser rapidement une « animation primale » stylisée. Ces animations stylisées pourront être affinées en post-processing en apportant une touche artistique à l'animation générée.

Abstract

The theme of my thesis concerns the recognition and synthesis of facial and body expressions. Our problem is to study, understand and extract the elements that convey a person's emotional state from the expressions of his face and body, in order to recognize and also to synthesize the emotion or style in a gesture. This dual objective of recognition and synthesis will make it possible to generalize new modes of interaction in applications such as video games, human-machine interaction, etc. Indeed, these applications can be enriched with scenarios that could be adapted to the emotional state of the user. To answer this problem, the crucial point is to understand "where" the expression information is located in an action. Indeed, in a few seconds, a human knows how to characterize an expression he sees while recognition algorithms are still far from it, especially for body expressions where, to our knowledge, little work has been done compared to the recognition of facial expressions. Concerning the recognition of expressions, our objective is, first of all, to propose features capable of recognizing the expression carried by a movement. To do this, the main problem is to separate the movement achieved from the perceived expression. Concerning the recognition of facial expressions, we are interested in the societal problem of parental protection. To do this, it is necessary to understand and know how to recognize facial expressions of children. To solve this problem, we have built and proposed a new database to help the computer vision community understand the specificities of facial expressions of children's faces. Secondly, we also hope that the various descriptors proposed in recognition of body expressions can be used in the field of animation synthesis. Indeed, in the field of animation, the creation of an action that conveys an emotion or a style requires a lot of work, know-how and time for an animator to propose stylized animations. For example, in a video game, to create such animations, it is often necessary to have a huge database of movements including each style for each virtual character. To have such a database, one of the following two methods is often used. The first consists in capturing all the movements made by different actors playing different styles. In the second, it is the graphic designer who must create the various animations by hand using animation software. Our objective in this context is to use the descriptors quantifying the expression detected in recognition of facial and body expressions in order to develop tools capable of changing / editing the style or expression of an animation. These tools will assist and facilitate the work of graphic designers by allowing them to quickly synthesize a stylized "primal animation". These styli-

zed animations can be refined in post-processing by adding an artistic touch to the generated animation.

Table des figures

Liste des tableaux

1 Introduction

Contents

1.1 Motivation

Dans la vie de tous les jours, la communication, qu'elle soit verbale ou non verbale, joue un rôle essentiel dans nos échanges avec d'autres personnes. Les expressions corporelles et faciales consistent un aspect important de la communication et permettent de fournir des informations sur l'état émotionnel d'une personne afin de mieux comprendre ses intentions. Longtemps, on a pensé que les expressions faciales fournissaient la majorité des informations sur l'état émotionnel d'une personne [106]. Cependant, des travaux dans le domaine de la psychologie ont montré que les expressions corporelles fournissent autant voire plus d'informations sur l'état émotionnel d'une personne. Ceci montre l'intérêt de considérer les deux types d'expressions afin de pouvoir fournir un système capable de décoder l'état émotionnel d'un individu.

Au niveau de l'informatique, nous assistons à un véritable tournant concernant les systèmes d'interaction homme-machine. En effet, dépassant la simple interaction gestuelle, les nouveaux systèmes s'orientent vers des interactions conversationnelles où les expressions prennent une place importante. Pour répondre à cette attente, ces systèmes doivent être capables d'analyser et de comprendre le comportement, c'est-à-dire l'état émotionnel, d'une personne. Il a été montré que des systèmes informatiques ou robotiques plus sensibles à ce que ressent l'utilisateur permettent d'améliorer l'interaction homme-machine. Donner une sorte d'empathie à une machine est le prochain grand défi des systèmes informatiques. En effet, de nombreux domaines d'applications peuvent bénéficier de tels systèmes capables de reconnaître l'expression d'une personne à travers l'analyse de sa posture, de ses gestes et de son visage. Les domaines concernés sont divers et nombreux : la science comportementale, l'éducation, la médecine, le commerce, la sécurité et les loisirs (cinéma, jeux-vidéos, etc.).

Si depuis longtemps les expressions faciales ont été considérées comme les plus significatives, ce qui a suscité l'intérêt des chercheurs, l'analyse des expressions corporelles n'a reçu que peu d'attention. Cela explique la richesse de la littérature sur la reconnaissance d'expressions faciales autour de cette question. Cette différence s'explique aussi par la complexité du problème posé par la reconnaissance d'expressions corporelles. En effet, une expression corporelle s'exprime à travers un mouvement contrairement à une expression faciale. Dans le domaine de la vision par ordinateur, la reconnaissance d'expressions se fait à travers des descripteurs que l'on va chercher à classer afin de détecter l'une des six expressions primaires ou universelles que sont la joie, la tristesse, la peur, le dégoût, la surprise et la colère. Dans cette thèse, nous nous sommes intéressés à l'étude, la compréhension et l'extraction des éléments qui traduisent l'état émotionnel d'une personne à partir

des expressions de son visage et de son corps, dans le but de reconnaissance et, dans une moindre mesure, de synthétiser une émotion dans un geste. Bien que les travaux réalisés dans cette thèse ont porté essentiellement sur la reconnaissance des expressions, faciales et corporelles, ils ouvrent une large perspective pour la création de mouvements expressifs dans le domaine de la synthèse d'animations.

1.2 Approches existantes

L'objectif de reconnaître des expressions et de comprendre un état émotionnel consiste à formaliser des observations, provenant du domaine de la psychologie, par le calcul de descripteurs. Ces derniers vont être calculés sur le corps et le visage. Afin de réaliser ces calculs, il est nécessaire de détecter le visage et le corps de la personne. A ce propos, la reconnaissance d'expressions faciales peut se faire à partir d'images statiques, contrairement aux expressions corporelles où il très difficile voire impossible de reconnaître l'expression corporelle exacte d'une personne à partir d'une image statique. Ceci est lié au fait que le mouvement joue un rôle primordial lors d'une expression corporelle. La vitesse et l'accélération des différentes articulations du corps va nous donner une information importante sur l'état émotionnel d'un individu. C'est pourquoi nous travaillons sur une séquence vidéo ou des données provenant de la capture de mouvements, c'est-à-dire une séquence de posture de squelette. A noter qu'il existe deux familles de méthodes pour la reconnaissance d'expressions corporelles. La première famille va travailler directement sur la séquence vidéo afin d'extraire différents descripteurs basés images. La seconde approche consiste à tout d'abord chercher à extraire le squelette de la personne à partir de l'image pour ensuite extraire différents descripteurs à partir de ce squelette. Nous nous situons dans ce second cas où nous travaillons directement sur une séquence de postures d'un squelette.

Une fois le visage ou le corps détecté, il est nécessaire de construire un ensemble de descripteurs pertinents capable de discriminer correctement les différentes expressions. La Section 2.3 présente en détail les différentes approches pour l'extraction de descripteurs caractérisant les expressions faciales tandis que la Section 2.4 présente les différentes approches pour l'extraction de descripteurs caractérisant les expressions corporelles. La dernière étape d'un système de reconnaissance automatique consiste à donner les descripteurs extraits à un classifieur.

1.3 Verrous scientifiques

Expression faciales

La reconnaissance d'expressions faciales est un sujet qui a été largement étudié pour des personnes adultes dans un environnement contrôlé. Par environnement contrôlé on entend des expressions non spontanées, jouées par des acteurs, dont la position du visage est fixe, dans un décor neutre où la lumière et la caméra sont statiques. L'une des questions posées dans le domaine de la vision par ordinateur est l'étude des expressions faciales dans des conditions dites réelles. Cela concerne aussi bien le quotidien ou des scènes simulées ou jouées : films, jeux vidéos, etc. Dans nos travaux, nous nous sommes intéressés à un problème sociétal qui est le contrôle parental de contenu multimédia. En effet, un enfant peut prendre peur devant divers contenus multimédias que ce soit des films ou des jeux-vidéos par exemple. C'est pourquoi il est nécessaire de caractériser et de comprendre les expressions faciales de visages d'enfants. Le premier problème qui se pose afin de répondre à cette question est le manque de données concernant les expressions faciales chez de jeunes enfants. Nous avons cherché à résoudre ce problème par la construction d'une base de données d'expressions faciales de visages d'enfants spontanées et dynamiques. Cette base de données permettra d'analyser et de comprendre les différences d'expressions faciales entre des visages d'adultes et d'enfants.

Expression corporelles

Le point crucial pour la reconnaissance d'expressions corporelles est de comprendre « où » se situe l'information expressive dans un geste ou une action. Par exemple, une personne qui marche peut le faire de façon joyeuse, de façon rêveuse ou encore en colère. Un humain sait identifier le mouvement instantanément et l'expression en quelques secondes alors que les machines en sont encore très loin. Dans le domaine de la psychologie, les chercheurs ont essayé de comprendre comment s'exprime l'expression corporelle d'une personne à travers divers mouvements. Dans le domaine de la vision par ordinateur, les méthodes sont spécialisées sur un mouvement spécifique. On peut citer à titre d'exemple des méthodes travaillant sur des mouvements de marche [77], de danse [114] ou encore de frapper à une porte [117]. L'ambition de cette thèse est d'avoir une approche générique pour la reconnaissance d'expressions corporelles. Cette expression peut accompagner divers mouvements (course, saut, marche, coup de poing, coup de pied, personne assise, etc.). Dans le but d'atteindre cette généricité, il a été nécessaire de formaliser un

ensemble de descripteurs capable de séparer le mouvement réalisé de l'expression perçue dans le but d'avoir une approche générique. En plus d'avoir une méthode capable de reconnaître l'expression corporelle, nous souhaitons pouvoir utiliser notre ensemble de descripteurs dans le domaine de la synthèse d'animations afin de pouvoir synthétiser des mouvements porteurs d'une expression.

1.4 Dissémination de nos contributions

Article de journal

— A novel database of children's spontaneous facial expressions (LIRIS-CSE). R.A. Khan, A. Crenn, A. Meyer et S. Bouakaz. In Image and Vision Computing. Volumes 83-84, Mars-Avril 2019, Pages 61-69. https ://doi.org/10.1016/j.imavis.2019.02.004.

Article de journal, en cours d'évaluation

— Generic body expression recognition based on the synthesis of realistic neutral motion. A. Crenn, A. Meyer, H. Konik et S. Bouakaz. Soumis à Multimedia Tools and Applications.

Conférences internationales

— Body expression recognition from animated 3D skeleton. A. Crenn, R.A. Khan, A. Meyer et S. Bouakaz. In International Conference on 3D Imaging (IC3D), Lièges, Belgique, 2016. Doi : https ://dx.doi.org/10.1109/IC3D.2016.7823448.

— Toward an efficient body expression recognition based on the synthesis of a neutral movement. A. Crenn, A. Meyer R.A. Khan, H. Konik et S. Bouakaz. In International Conference on Multimodal Interaction (ICMI), Glasgow, Ecosse, 2017. Pages : 15-22. Doi : https ://dx.doi.org/10.1145/3136755.3136763.

Conférence nationale

— Reconnaissance d'expressions corporelles à l'aide d'un mouvement neutre synthétisé. A. Crenn, H. Konik, A. Meyer et S. Bouakaz. In COmpression et REpresentation des Signaux Audiovisuels (Coresa), Caen, 2017, France.

État de l'art

2

Contents

2.1 Définition émotion, expression et état affectif

Avec le développement des systèmes d'interaction hommes-machines, les demandes des applications nous orientent vers des systèmes plus développés qui cherchent à prendre en compte l'état émotionnel de l'utilisateur afin d'adapter leur contenu. En outre la généralisation du numérique aide à l'acceptation des outils utilisant la reconnaissance faciale en vue d'adapter le contenu proposé. A titre d'exemple, on peut citer Facebook, Microsoft et Google qui commencent la détection du visage et la reconnaissance d'expressions faciales afin de proposer une expérience d'interaction nouvelle à l'utilisateur. Bien que longtemps considérés comme intrusifs, ces systèmes commencent à se démocratiser et à être accepté par les utilisateurs. Avant d'aller plus loin, il nous semble nécessaire d'établir une distinction entre une émotion, un état affectif et une expression.

A notre connaissance, il n'y a pas de définition formelle pour la notion d'émotion. Bien qu'il semble que tout le monde sache ce qu'est une émotion lorsque l'on demande une caractérisation concrète, personne n'est capable d'y répondre. Dans une étude [82], Kleinginna et Kleinginna ont recensé 92 définitions différentes d'une émotion et ont tenté d'en extraire différentes caractéristiques communes en vue de proposer la définition suivante : *« les émotions sont le résultat de l'interaction de facteurs subjectifs et objectifs, réalisés par des systèmes neuronaux ou endocriniens, qui peuvent : a) induire des expériences telles que des sentiments d'éveil, de plaisir ou de déplaisir ; b) générer des processus cognitifs tels que des réorientations pertinentes sur le plan perceptif, des évaluations, des étiquetages ; c) activer des ajustements physiologiques globaux ; d) induire des comportements qui sont, le plus souvent, dirigés vers un but adaptatif. »* [55]. Il est souvent souligné qu'une émotion est une réaction physiologique de notre corps face à un évènement extérieur [69]. Cependant, on peut aussi ressentir une émotion lorsque l'on repense à un évènement passé. Cette diversité de points de vue montre à quel point il est difficile de converger sur une définition précise d'une émotion.

Dans une étude [126], Scherer et al. ont tenté de caractériser un état affectif selon différents facteurs :

- Intensité de l'état affectif.
- Durée de l'état affectif.
- Synchronisation des sous-systèmes de l'organisme.
- Focalisation sur l'évènement.
- Rapidité du changement d'état.
- Impact comportemental.

Dans ces travaux, ils considèrent les émotions comme un groupe particulier parmi les différents types d'états affectifs. Celles-ci sont plus intenses mais de plus courtes durées. En particulier, les émotions sont caractérisées par un haut degré de synchronisation entre les différents sous-systèmes. De plus, elles sont susceptibles d'être très axées sur les événements déclencheurs et produites par l'évaluation cognitive. Du point de vue de leur effet, elles ont un fort impact sur le comportement et sont susceptibles de changer rapidement. Dans ces travaux, les chercheurs définissent cinq sous-systèmes : l'évaluation cognitive, les changements psychophysiologiques, l'expression motrice, les tendances à l'action et le sentiment subjectif. Un état affectif va donc être caractérisé par ces cinq composants ainsi que les différents facteurs présentés ci-dessus.

Dans ce manuscrit, nous considérons une expression comme étant ce qu'une personne peut percevoir sur un visage, un mouvement. C'est une qualité qui exprime de manière spontanée une émotion. Dans la littérature, en travaillant sur le visage, Ekman [41] a proposé six expressions. Ces expressions sont la joie, la tristesse, la peur, le dégoût, la surprise et la colère. Ces expressions sont qualifiées d'universelles car elles se sont avérées universelles pour toutes les ethnies et cultures humaines. Bien qu'elles soient introduites pour les expressions faciales, cette catégorisation a également servi pour les expressions corporelles. Elles sont à la base de différentes méthodes permettant la reconnaissance d'expressions corporelles. Nous suivrons aussi cette classification dans nos travaux. Une expression est donc un moyen de communication non verbale, c'est une manifestation visuelle d'une émotion.

2.2 Représentation des expressions

2.2.1 Le modèle discret d'Ekman

Le système de codage des expressions faciales et corporelles le plus couramment utilisé dans l'approche de reconnaissance d'expressions est celui proposé par Ekman [41]. Ce modèle propose six expressions basiques : la joie, la colère, la tristesse, la peur, le dégoût et la surprise. Pour le visage, ces expressions sont considérées comme universelles car il est avéré qu'elles sont les mêmes à travers diverses ethnies et cultures humaines. La Figure 2.1 présente ces six expressions basiques sur des visages. Comme on peut le voir sur cette image, une simple photographie permet de reconnaitre l'expression faciale, ce qui n'est pas le cas des expressions corporelles où le mouvement est indispensable.

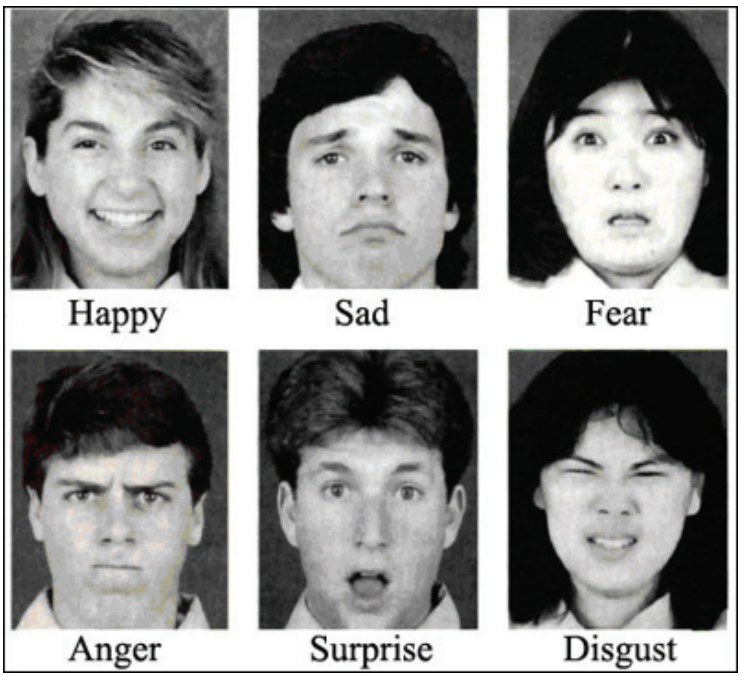


Figure 2.1 : Présentation des six expressions basiques sur des visages. De haut en bas : la joie, la tristesse, la peur, la colère, la surprise et le dégout

La théorie d'Ekman a été inspirée des travaux réalisés par Darwin [29]. Selon Darwin, l'expression des émotions a évolué chez l'homme à partir d'animaux. Darwin soutenait que les expressions n'étaient pas apprises mais innées dans la nature humaine et qu'elles étaient donc importantes pour la survie en raison de l'évolution. Ainsi, Darwin a réussi à démontrer l'universalité des émotions. Cela veut dire que tous les individus du monde, quelque soit leur peuple d'appartenance, ressentent la peur ou la joie de la même manière. Ekman avait émis un doute quant aux thèses de Darwin : il pensait qu'il y avait un biais dans sa recherche et a voulu démontrer qu'il avait tort. Mais après une contre étude celui-ci a dû admettre qu'effectivement l'émotion est universelle et a ainsi proposé six émotions de base qui sont largement utilisées pour tester les différentes méthodes proposées dans la reconnaissance automatique d'expressions faciales et corporelles.

Ekman et Wallace Friesen [44] ont proposé un système de codage d'actions faciales, facial action coding system (FACS) dans le but de décrire les mouvements du visage. Dans ce système, les contractions ou décontractions du visage sont décomposées en unités d'action (AU). Le système FACS repose sur la description de 46 AU identifiées par un numéro dans la nomenclature FACS. Ainsi, l'AU 1 va correspondre à l'action de lever les sourcils. Chaque AU correspond à la contraction ou à la détente d'un ou plusieurs muscles, qui se traduit par un mouvement d'une partie donnée du visage. Une expression faciale correspond donc à la mise en jeu de plusieurs unités d'actions. Par exemple, dans le cas d'un visage apeuré, les AU mises en jeu

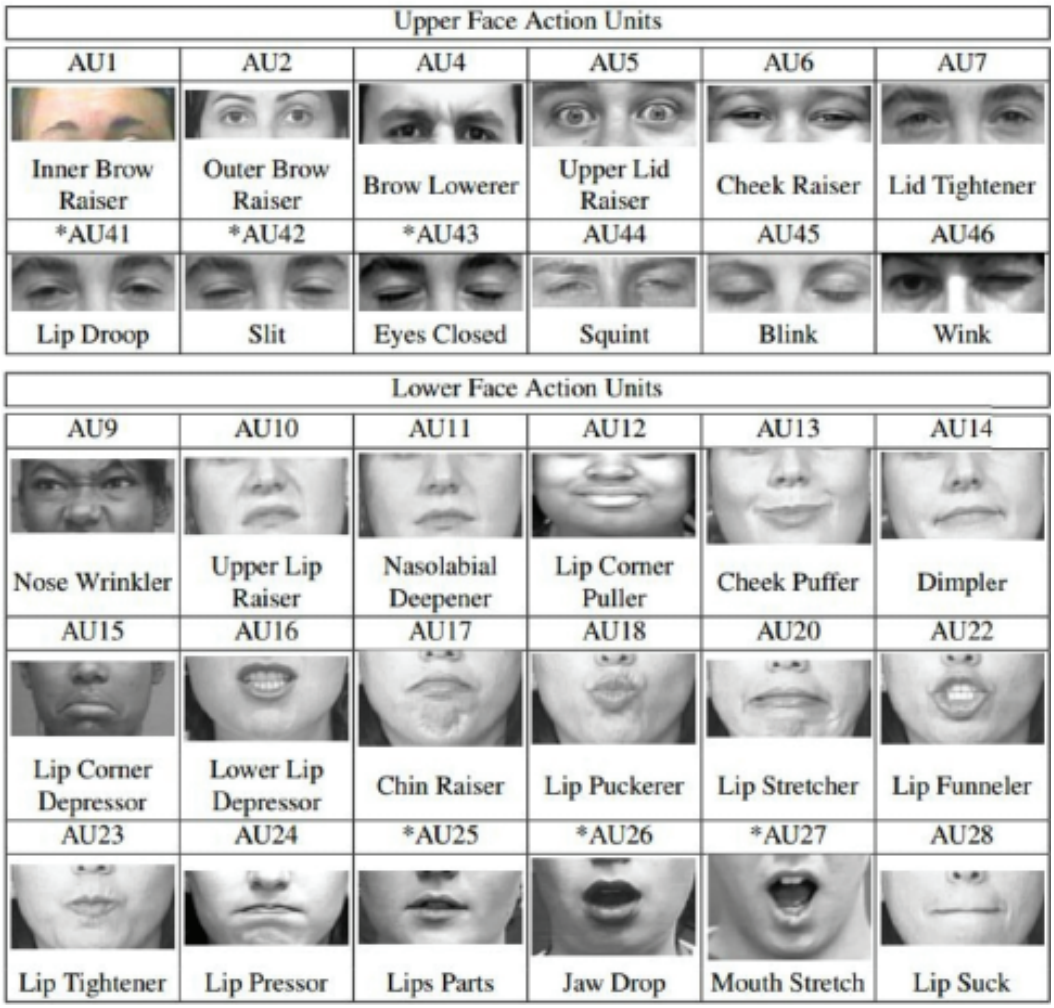


Figure 2.2 : Présentation de différentes unités d'actions pour la partie haute et basse du visage. [30]

sont : lever les sourcils intérieurs, lever les sourcils extérieurs, tension des lèvres, abaissement de la mâchoire inférieure. La Figure 2.2 représente différentes unités d'actions pour la partie haute et basse du visage.

2.2.2 Le modèle arousal - valence

Un autre modèle, afin de représenter les expressions, a émergé et diffère du modèle proposé par Ekman en proposant une représentation non discrète et en liant les expressions l'une à l'autre [18]. Le modèle le plus abouti a été proposé par James A. Russel [123]. Il a proposé un système à deux axes bipolaires, à savoir l'arousal (ou excitation) et la valence. La valence va de la tristesse au bonheur, tandis que

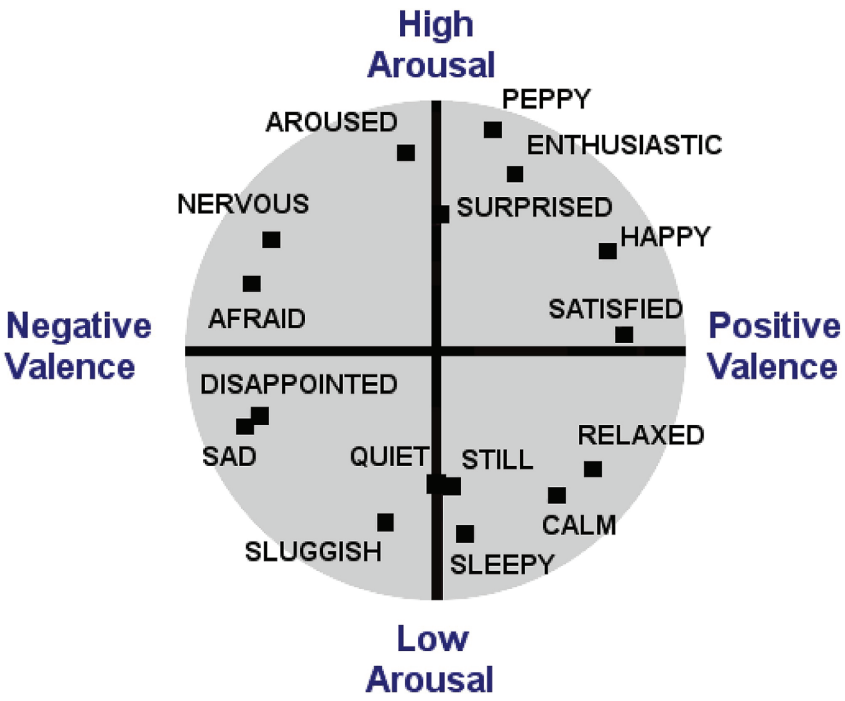


Figure 2.3 : Espace de modélisation arousal - valence.

l'excitation va de l'ennui ou de la somnolence à l'excitation frénétique. Selon Russell, les émotions se situent sur un cercle dans cet espace bidimensionnel, et sont caractérisées par des catégories floues regroupées sur les axes arousal et valence. La Figure 2.3 présente cet espace de représentation des expressions.

Bien qu'un espace continu puisse représenter toutes les expressions possibles, cela n'est pas garanti par la théorie de Russell et n'a d'ailleurs pas été démontré. Il n'est pas clair comment une expression doit être projetée sur l'espace ou, vice versa, comment définir les régions dans l'espace de valence/arousal qui correspondent à une certaine expression. Ce modèle est assez subjectif, ce qui est problématique lorsque l'on cherche à reconnaître de manière automatique des expressions. Nous somme donc arrivés à la conclusion que ce système n'est pas adapté pour la vision par ordinateur.

2.2.3 La roue des émotions de Plutchik

Une dernière représentation significative des émotions a été proposée par Plutchik [116]. Dans son modèle, les émotions humaines sont modélisées selon huit émotions primaires. Les émotions sont organisées en paires d'opposés : la joie contre la tristesse, la confiance contre le dégoût, la peur contre la colère et l'anticipation contre la surprise. La Figure 2.4 représente la roue des émotions de Plutchik qui

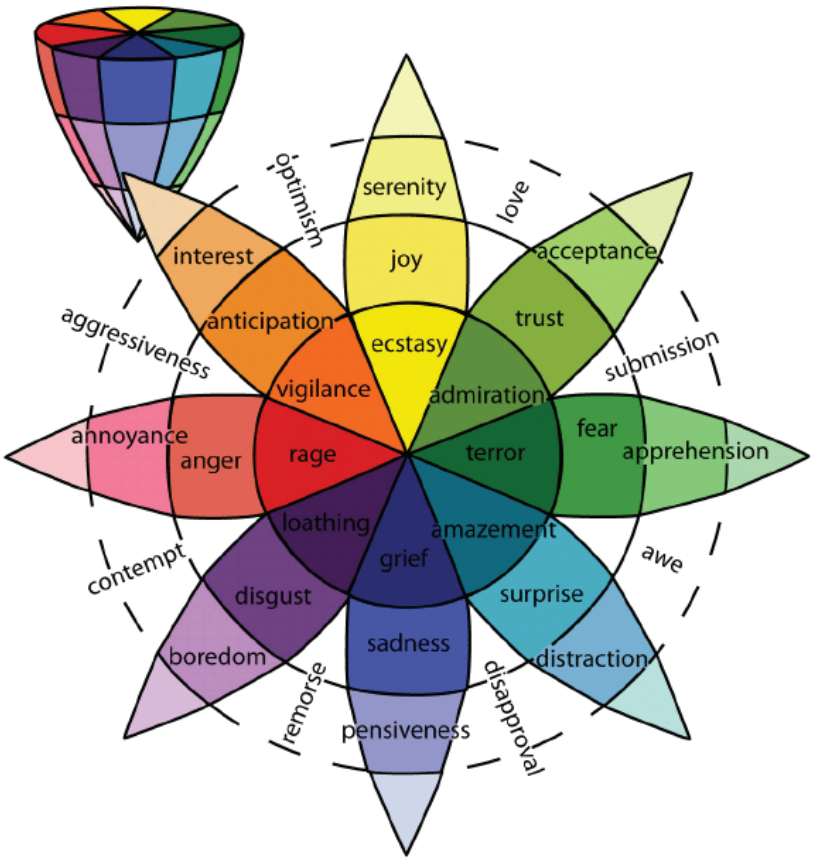


Figure 2.4 : La roue des émotions de Plutchik.

peut être considérée comme un modèle hybride entre le modèle d'Ekman et le modèle arousal/valence. Ainsi, les émotions les plus complexes sont définies comme une combinaison d'émotions basiques. Par exemple, l'amour est considéré comme la combinaison de la joie et de la confiance. Dans ce travail de recherche, les six émotions basiques d'Ekman sont largement utilisées pour tester nos diverses méthodes pour la reconnaissance automatique d'expressions faciales et corporelles. Ceci est dû au fait que les six émotions de base d'Ekman sont bien adaptées à l'application de la vision par ordinateur (classification en classes discrètes) comparées aux modèles de l'arousal / valence et à la roue des émotions de Plutchik.

Dans ce manuscrit, nous nous intéressons à la reconnaissance d'expressions faciales et corporelles. La première étape de la reconnaissance d'expressions consiste à la détection et au suivi soit du visage soit du corps, afin d'extraire le squelette de la personne. Après avoir localisé la région d'intérêt, l'étape suivante consiste à extraire des informations significatives ou discriminatoires à travers la formalisation de différents descripteurs. Ces derniers sont ensuite fournis à un classifieur, que l'on pré-entraîne sur différentes bases de données, afin de détecter l'expression fa-

Figure 2.5 : Approche générique pour la détection d'expressions faciales.

ciale ou corporelle fournie en entrée du système. L'image 2.5 présente le pipeline générique pour la détection d'expressions faciales et corporelles. En entrée du système, nous utilisons soit le visage soit le corps, ici le visage dans l'image, puis nous calculons des descripteurs qui sont ensuite fournis à un classifieur.

2.3 Reconnaissance d'expressions faciales

L'étude des expressions faciales a suscité beaucoup d'intérêt, il serait difficile de faire une étude exhaustive sur ce thème. Comme énoncé précédemment, l'approche générale de la reconnaissance automatique des expressions faciales consiste en trois étapes : détection et suivi du visage, extraction de différents descripteurs et classification/reconnaissance des expressions. Dans cette étude, nous n'abordons pas la détection et le suivi du visage. Une description qualitative de cette étape a été donnée dans la thèse de Rizwan Amhed Khan [78]. Notons qu'avec le développement des approches basées sur l'apprentissage profond, les étapes d'extraction des descripteurs et de classification sont fusionnées en une seule étape.

Concernant l'étape de l'extraction des descripteurs porteurs de différentes informations discriminatives, les méthodes peuvent être classées en cinq types : les méthodes basées sur des descripteurs de texture, des méthodes basées sur la détection de contours, des méthodes basées sur l'extraction globale et locale de descripteurs, les méthodes basées sur des descripteurs géométriques et les méthodes basées sur l'application de patch.

2.3.1 Approches basées descripteurs de textures

Le descripteur le plus classique est le filtre de Gabor qui est un descripteur de texture. Il inclut des informations de magnitude et de phase. La fonction de magnitude du filtre de Gabor contient les informations sur l'organisation de l'image du visage.

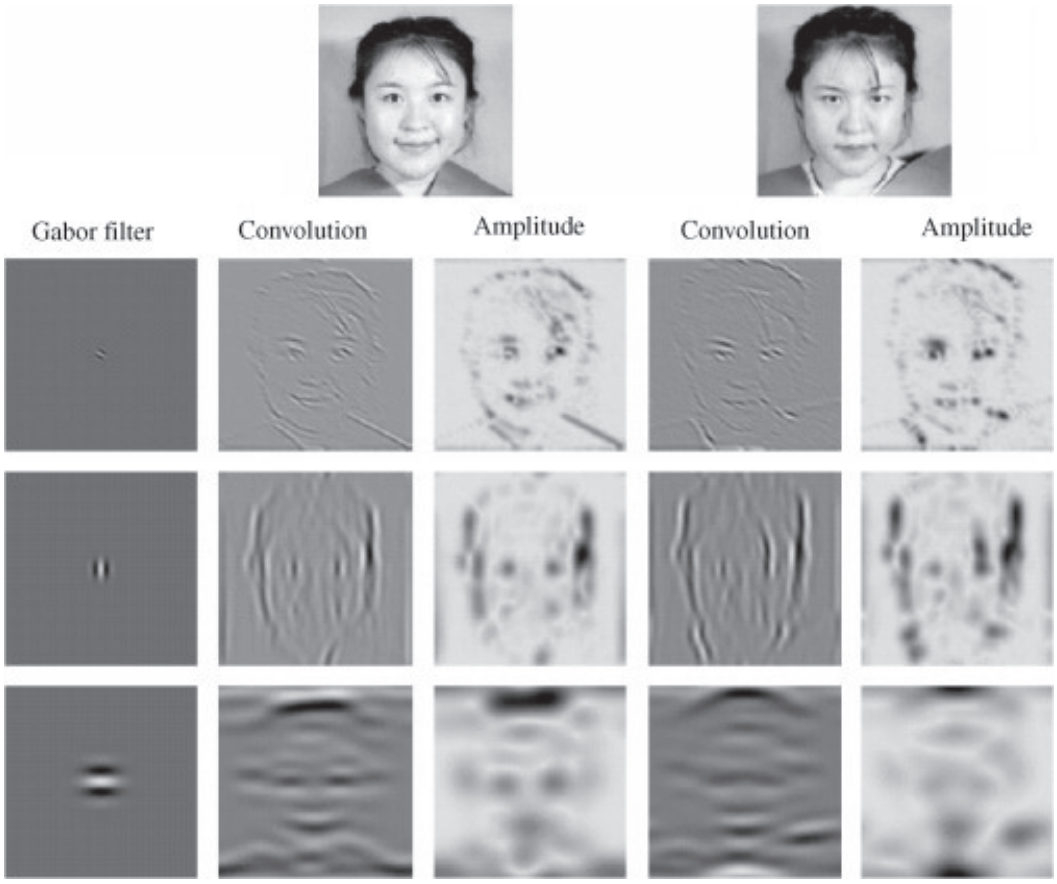


Figure 2.6 : Application d'un filtre de Gabor pour la détection d'expression faciales. Le filtre est tourné selon plusieurs directions afin d'avoir différentes réponses et donc produire différentes textures orientées. Image issue de [11]

La caractéristique de phase précise l'information sur la description complète des descripteurs de magnitude [11, 109, 151, 62, 61]. La Figure 2.6 présente l'application d'un filtre de Gabor sur un visage. On peut observer les différents descripteurs qui vont être obtenus par convolution du filtre sur l'image originale.

Les motifs binaires locaux (Local Binary Pattern - LBP) sont aussi utilisés pour l'extraction de descripteurs de texture. Le principe général est de comparer le niveau de luminance d'un pixel avec les niveaux de ses voisins. Cela permet d'avoir une information relative à des motifs réguliers dans l'image, autrement dit une texture. Selon l'échelle du voisinage utilisé, certaines zones d'intérêt telles que des coins ou des bords peuvent être détectées par ce descripteur [59, 23]. De même, les LBP combinés avec les descripteurs des trois plans orthogonaux, voir l'image 2.7 qui représente ces trois plans sur le visage, permettent de proposer des approches multi-résolution [152]. D'autres descripteurs de texture du visage peuvent aussi être extraits en utilisant des fonctions de Gauss-Laguerre (GL). Ces fonctions donnent une structure pyramidale de direction qui extrait les caractéristiques de texture et l'in-

formation sur l'occurrence du visage. Comparé au filtre de Gabor, GL utilise un filtre simple au lieu de différents filtres multiples [118]. Un autre descripteur utilisé et qui extrait également les caractéristiques de texture des images de visage est le Vertical Time Backward (VTB). De même, le descripteur Moments extrait les caractéristiques liées à la forme des composants faciaux significatifs. Ces deux descripteurs, VTB et Moments, sont efficaces sur les plans spatio-temporels [71]. Dans la continuité des descripteurs de texture, nous trouvons la technique Weber Local Descriptor (WLD) qui extrait les caractéristiques de texture discriminantes des images de visages segmentées [23]. L'extraction de ces descripteurs se fait en trois étapes à l'aide de la méthode supervisée de descente (Supervised Descent Method). Dans un premier temps, les positions principales du visage sont extraites. Ensuite, les positions suivantes sont sélectionnées. Finalement, cette méthode va estimer la distance entre les différentes composantes du visage [124]. Une autre méthode pour l'extraction de caractéristiques consiste en la combinaison de la LBP avec des projections pondérées (Weighted Projection based LBP). Cette méthode est basée sur la détection de régions instructives sur lesquelles les caractéristiques de la LBP vont être extraites. Ensuite, en fonction de l'importance des différentes régions instructives, ces différentes caractéristiques vont être pondérées [90]. Une dernière méthode pour l'extraction de caractéristiques de textures est la Discrete Contourlet Transform. Cette transformée est exécutée par décomposition en deux étapes. Celles-ci sont une pyramide laplacienne (Laplacian Pyramid) et une banque de filtres directionnels (Directional Filter Bank). La première étape, celle de la pyramide laplacienne, va partitionner l'image en passe-bas, passe-bande permettant de confiner la position des discontinuités. Par la suite, l'application des différents filtres directionnels va traiter le passe-bande et former la composition linéaire en associant la position des différentes discontinuités [15].

2.3.2 Approches basées sur la détection de contours

Dans ce paragraphe nous décrivons les méthodes basées sur la détection de contours pour l'extraction de descripteurs. Le descripteur LEM (Line Edge Map) est un descripteur d'expressions faciales qui améliore les caractéristiques structurelles géométriques en utilisant l'algorithme dynamique à deux bandes [147]. Noh et al. [107] ont proposé l'extraction de caractéristiques faciales basés sur une analyse de mouvement. En utilisant les Graphics-processing unit based Active Shape Model (GASM), Song et al.[128] ont proposé une méthode d'extraction de descripteurs. Les GASM permettent l'extraction de différents descripteurs basés sur la détection de contours, l'amélioration, la mise en correspondance tonale et la correspondance locale d'un modèle d'apparence. Un dernier descripteur utilisé est l'histogramme des gradients orientés (HOG) qui calcule des histogrammes locaux de l'orientation du gradient

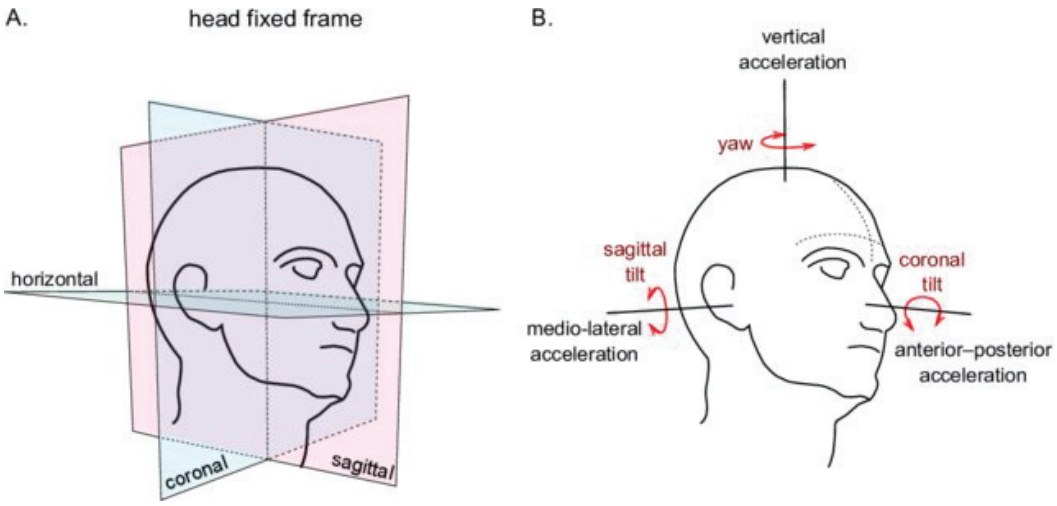


Figure 2.7 : Tête humaine et son corps fixe. A : Trois plans orthogonaux sont définis : sagittal, coronal et horizontal. B : La tête peut être tournée et déplacée dans trois directions orthogonales. Image issue de [47]

sur une grille dense, c'est-à-dire sur des zones régulièrement réparties sur l'image. Dahmane et Meunier [27] utilisent HoG pour extraire des caractéristiques visuelles, par exemple une expression de joie va être transcrite par une courbure au niveau des yeux.

2.3.3 Approches combinant plusieurs descripteurs

Le paragraphe suivant présente les méthodes qui cherchent à extraire des caractéristiques à partir de différentes méthodes globales et locales. Pour cela, la méthode d'analyse en composantes principales (ACP) est utilisée pour l'extraction de différents descripteurs. En effet, elle permet d'extraire des caractéristiques globales et à basses dimensions. L'analyse en composantes indépendantes (ICA) est aussi utilisée dans le but d'extraire des caractéristiques locales en utilisant des observations multicanaux [130]. Siddiqi et al. [127] ont proposé d'extraire différents descripteurs en utilisant une analyse discriminante linéaire pas à pas (Stepwise Linear Discriminant Analysis). Cette méthode permet d'extraire des descripteurs locaux en utilisant deux modèles de régression, un descendant et un ascendant. L'objectif de ces méthodes est de construire des eigenfaces à partir de l'image fournie en entrée. Une eigenface est un ensemble de vecteurs propres et a notamment été utilisée pour la détection et la reconnaissance de visage humain. Pour la reconnaissance d'expressions corporelles, certaines méthodes vont utiliser les eigenfaces de manière globale ou vont appliquer les analyses en composantes sur différentes régions du visage (les yeux, le nez et la bouche par exemple).

Le paragraphe qui suit présente les méthodes basées sur l'extraction de descripteurs géométriques. Pour cela, les méthodes utilisent la Local Curvelet Transform (LCT) qui est une généralisation dimensionnelle supérieure de la transformée en ondelettes. La LCT est conçue pour représenter des images à différentes échelles et sous différents angles. Demir [32] utilise la moyenne, l'entropie et la déviation standard de la LCT pour construire un ensemble de descripteurs. En plus de ces descripteurs, Mahersia et Hamrouni [100] vont calculer l'énergie et le kurtosis de la LCT à l'aide d'une représentation pyramidale paramétrable en trois niveaux.

Pour finir, les derniers types de méthode cherchant à proposer des descripteurs sont basées sur des patchs. En effet, les caractéristiques des mouvements du visage sont extraites sous forme de patchs en fonction de la distance. Celles-ci sont réalisées à l'aide de deux processus tels que l'extraction des correctifs et l'appariement des correctifs. L'appariement des patchs est effectué en traduisant les patchs extraits en caractéristiques de distance Zhang et al. [150]. Des recherches ont été menées avec succès ces derniers temps afin de combiner différents types de descripteurs [89, 37]. Le problème clé avec ces méthodes est de savoir précisément localiser le point de repère et de le suivre. Dans les applications réelles, en raison de la pose et des variations d'éclairement, des images d'entrées à faible résolution et du bruit, il est encore très difficile de localiser avec précision les points de repère.

La classification est l'étape finale du système de la reconnaissance d'expression faciale dans laquelle le classificateur catégorise les six expressions de base. La Figure 2.8 présente une étape de classification classique pour la reconnaissance d'expressions faciales. L'apprentissage automatique comporte généralement deux phases. La première étape consiste à estimer un modèle par apprentissage d'un classifieur. Dans notre cas, les bases de données étant étiquetées en classes, nous nous situons donc dans un apprentissage supervisé. Via cet apprentissage, le classifieur sélectionne les descripteurs les plus pertinents afin de correctement séparer les données et classes. Lors de la seconde étape, nous cherchons à prédire la classe ou l'étiquette d'une nouvelle observation. Pour cela, nous fournissons les descripteurs extraits au classifieur qui va nous donner en retour la classe d'appartenance de cette observation. La première étape d'apprentissage peut être plus ou moins longue en fonction de la taille du jeu de données d'apprentissage et aussi du nombre de descripteurs présents. La seconde étape tourne en temps-réel une fois le modèle correctement entraîné.

2.3.4 Différentes bases de données d'expressions faciales

Nous avons présenté de manière non exhaustive le pipeline global pour la reconnaissance d'expressions faciales en insistant sur les différentes méthodes d'extrac-

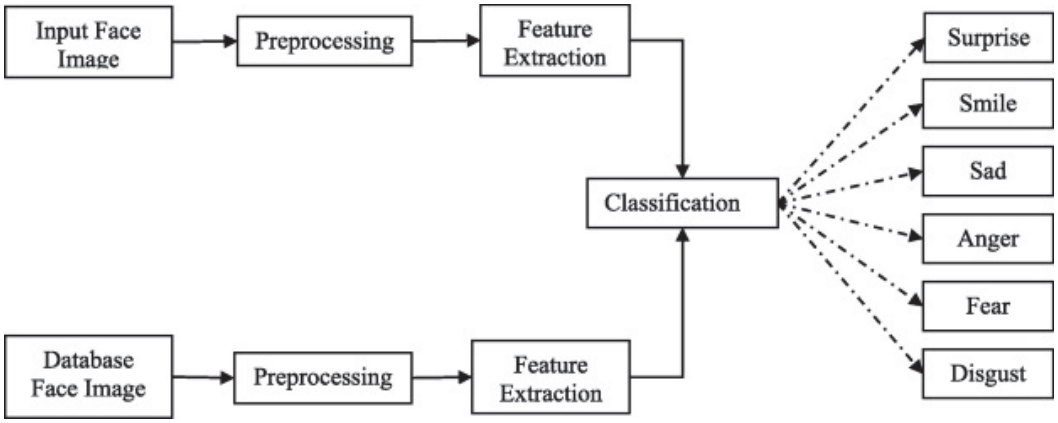


Figure 2.8 : Une fois les descripteurs extraits, ils sont donnés à un classifieur en vue de la détection de l'expression réalisée en entrée. [119]

tion de descripteurs car c'est le point sur lequel nous nous sommes intéressés pour la reconnaissance d'expressions corporelles. Cela va être intéressant de pouvoir mettre en parallèle les deux domaines, c'est pourquoi nous avons commencé cet état de l'art sur ce point. Notre contribution sur la reconnaissance d'expressions faciales concerne par contre la présentation et le partage d'une nouvelle base de données de visages d'enfants. Il est donc nécessaire de faire le point sur les différentes bases de données existantes pour la reconnaissance d'expressions faciales. La grande majorité des travaux sur les FER utilisent les bases de données suivantes : Cohn-Kanade (CK) [74], Extended Cohn - Kanade (CK+) [96], Japanese Female Facial Expressions (JAFFE) [73], MMI [111], Multimedia Understanding Group (MUG) [1], Yale [53], AR face database [101], etc. Les paramètres qui changent en fonction des différentes bases de données sont les suivants :

— Individualité des sujets. La forme du visage, la texture de la peau (poils par exemple), lunettes de vue ou de soleil, le sexe, l'origine ethnique et l'âge des sujets sont des paramètres sur lesquels on peut jouer pour créer une base de données variées ou non.

— Expression actée contre expression spontanée. La plupart des bases de données ont été réalisées en demandant à des personnes de jouer une série d'expressions. Ces expressions dirigées peuvent différer dans leurs caractéristiques, leur dynamique temporelle et leur spontanéité [46].

— Résolution de la séquence d'images. Il est nécessaire de faire varier le format des images. Pour cela, certaines bases de données contiennent des images à haute résolution et à basse résolution dans le but de se rapprocher du monde réel.

— Orientation tête/visage. L'orientation du visage par rapport à la caméra influence énormément la performance de différents algorithmes de reconnaissances d'expressions faciales.

— Complexité de l'arrière-plan. La séquence d'images enregistrées dans un arrière-plan complexe rend la tâche de la reconnaissance automatique des expressions faciales encore plus difficile car l'arrière-plan complexe influence la précision de la détection automatique des visages, le suivi des caractéristiques et la reconnaissance des expressions. La plupart des bases de données disponibles ont un fond neutre ou très persistant.

— Variation d'éclairage. Il est souhaitable que les algorithmes de reconnaissance automatique des expressions soient invariables en fonction des conditions d'éclairage. Très peu de bases de données accessibles au public enregistrent les stimuli d'éclairage variable.

2.3.5 Approches à base de réseaux de neurones

Avec l'arrivée des méthodes basées sur des réseaux de neurones profonds, des nouvelles bases de données ont commencé à émerger, les bases de données dîtes "sauvages" (in the wild) [38, 36, 35]. Ces bases de données contiennent des images / vidéos capturées en faisant varier les différents paramètres explicités ci-dessous en capturant des expressions faciales variées, des mouvements naturels de pose de la tête, des occlusions, des sujets de différentes races, du sexe, d'âges différents et avec des sujets multiples dans une scène. De même, la généralisation des réseaux de neurones profonds a permis le développement de l'apprentissage par transfert. L'apprentissage par transfert est une méthode d'apprentissage machine où un modèle développé pour une tâche est réutilisé comme point de départ pour un modèle sur une seconde tâche. Il s'agit d'une approche populaire dans l'apprentissage profond où les modèles pré-entraînés sont utilisés comme point de départ pour les tâches de vision par ordinateur, étant donné les vastes ressources en calcul et en temps nécessaires pour développer des modèles de réseaux neuronaux sur ces problèmes. L'apprentissage par transfert permet aussi de travailler sur des tâches dont on a très peu de données. Ceci marche uniquement si les caractéristiques du modèle apprises lors de la première tâche sont générales ou similaires à la seconde tâche. Dans nos travaux, nous utilisons l'apprentissage par transfert afin de proposer une méthode de reconnaissance automatique d'expressions faciales sur la base de données que nous avons proposé.

2.3.6 Bilan de l'état de l'art sur la reconnaissance d'expressions faciales

Pour conclure, le domaine de la reconnaissance d'expressions faciales a beaucoup évolué que ce soit en matière de méthodes ou de bases de données. Les méthodes étaient conçues pour une base de données très contrôlées au début et maintenant à travers différents concours, les méthodes se sont généralisées afin d'être invariant au plus grand nombre de paramètres possible. Bien que nous n'avons pas présenté les méthodes basées sur l'apprentissage profond, ce type de méthode est le plus utilisé actuellement car leurs résultats surpassent celles des autres méthodes. Une autre observation que l'on peut noter, à la suite de cet état de l'art, est que la plupart des méthodes travaillent sur des visages d'adultes. Il existe très peu d'études qui ont cherchées à analyser et reconnaitre les expressions faciales sur des visages de jeunes enfants. Ce constat est lié au fait que la sensibilité des données ainsi que la complexité de la création d'une telle base de données ont entrainé un manque de données de visages d'enfants. En nous intéressant à la question du contrôle parentale, nous avons cherché à construire une telle base de données. Cette dernière permettra de comprendre et d'observer des expressions faciales d'enfants et ainsi d'analyser les différences d'expressions faciales entre des visages d'adultes et d'enfants.

2.4 Reconnaissance d'expressions corporelles

2.4.1 Études psychologiques

Avant de présenter les approches informatiques liées aux expressions corporelles, nous allons présenter un aperçu des différents travaux réalisés dans le domaine de la psychologie afin de comprendre comment se perçoit une expression corporelle lors d'un mouvement. Les différentes études réalisées dans ce domaine ont cherché à comprendre une expression corporelle selon le mouvement réalisé et la posture du squelette. Pour cela, les chercheurs du domaine de la psychologie ont observés deux niveaux de détails corporels, une description haut niveau et bas niveau des expressions corporelles.

Selon une étude réalisée en neurosciences par Giese et Poggio [54] et Vaina et al [133], il existe deux voies séparées dans le cerveau pour reconnaître des informations de postures et de mouvements. La première voie se concentre sur l'information de forme en cherchant à analyser la posture du corps, tandis que la seconde

voie va analyser les informations de mouvement afin de reconnaitre et d'analyser l'action réalisée par la personne.

D'autres études dans le domaine de la psychologie [63, 115, 103] ont confirmé cette hypothèse en observant le fait que les informations de forme sont cruciales pour la reconnaissance de mouvements corporels. De même [7, 108] ont aussi montré qu'uniquement l'information de mouvement permet de reconnaitre l'expression corporelle d'un mouvement particulier. En effet, en analysant un stimuli lumineux posé sur une articulation, par exemple sur la main lors d'un mouvement de frapper à une porte, il est possible de discriminer différentes expressions. Ces études montrent que le mouvement joue un rôle primordial pour reconnaître une action. Couplé à des informations sur la posture, on peut donc construire un système permettant la reconnaissance d'expression corporelles. En effet, l'analyse de la posture permet de discriminer deux expressions différentes qui vont posséder des dynamiques et un mouvement similaire [120]. C'est pourquoi les chercheurs dans le domaine de la psychologie ont cherché à combiner ces deux types d'observations afin de comprendre comment un humain reconnait une expression corporelle. Pour cela, une approche consiste à comprendre la relation entre un état affectif et un haut-niveau de descripteur caractérisant soit le mouvement, soit la posture. Afin de discriminer différents états affectifs en fonction de la posture, différentes études ont cherché à caractériser des descripteurs corporels. James [70] a découvert que l'ouverture du corps est un critère important lorsqu'un humain cherche à reconnaitre une expression corporelle, l'inclinaison du corps et la position de la tête (penché en avant, en arrière ou tourné). De même, d'autres études [60, 104] ont confirmé l'importance de l'inclinaison du corps afin de reconnaître des expressions corporelles. En utilisant des mouvements de danse, Aronoff et al.[6] ont montré que des postures dites arrondies vont être plus chaleureuses tandis que des postures avec des configurations angulaires et diagonales vont signifier un comportement menaçant. Pour l'importance du mouvement lors de la discrimination de différents états affectifs, Gross et al [57] ont cherché à analyser le mouvement d'une action spécifique, frapper à une porte. Leur étude a montré que l'analyse de la vitesse et de l'accélération de la main permet de reconnaître l'expression corporelle d'une personne pour les expressions à forte activation, c'est-à-dire la joie, la colère et la fierté. Un autre objectif de cette étude était d'évaluer quantitativement la valeur de différentes caractéristiques sur différentes expressions corporelles. Les résultats ont été positifs, ce qui signifie qu'une comparaison quantitative des expressions était possible. Par exemple, ils ont constaté que le bras était levé au moins 17 degrés plus haut pour les mouvements de colère que pour les autres expressions. Cela monte la difficulté de quantifier les expressions corporelles car elles différent en fonction de la présence ou de l'absence d'une caractéristique particulière ainsi que de la valeur quantitative de chaque caractéristique. Glowinski et al [19] ont

émis l'hypothèse que l'utilisation d'un ensemble réduit de caractéristiques, dans le haut du corps seulement, serait suffisante pour classer un grand nombre de comportements affectifs. Les expressions émotionnelles de la partie supérieure du corps qui ont été jouées à partir du corpus GEMEP [9] ont été statistiquement regroupées selon les quatre quadrants du plan valence-arousal. Les auteurs ont conclu que des "groupes significatifs d'émotions" pourraient être regroupés dans chaque quadrant et que les résultats sont similaires à ceux de la recherche sur le comportement non verbal existante [138, 84].

Plus récemment, et grâce à la possibilité d'automatiser l'analyse des expressions corporelles, les chercheurs du domaine de la psychologie ont essayé de transformer les expressions corporelles affectives en descriptions de configurations de bas niveau afin d'examiner quelles postures permettent aux humains de distinguer les émotions spécifiques. Dans l'étude de Wallbott [138], il a construit un système de catégories qui comprenait les mouvements du corps, les postures et la qualité du mouvement. Dans son étude, il a constaté qu'il existe des différences dans la façon dont les gens évaluent la posture afin de faire la distinction entre les émotions. Cependant, Wallbott lui-même a déclaré qu'il s'agissait d'une étude initiale et a affirmé que d'autres études devaient être menées pour faire la distinction entre les émotions. En particulier, il a souligné la nécessité d'études examinant les expressions non jouées et les questions inter-culturelles. De Meijer [105] a mené une étude pour déterminer si des mouvements corporels spécifiques étaient porteurs d'émotions spécifiques et quelles caractéristiques du mouvement étaient à l'origine de ces attributions. Pour cela, des danseurs ont été filmés pendant qu'ils exécutaient des caractéristiques spécifiques des mouvements au lieu d'exprimer des émotions de façon explicite. Un groupe distinct d'observateurs a évalué chaque mouvement en fonction de sa compatibilité avec chaque émotion. Les résultats ont montré que des mouvements spécifiques ont été attribués à chaque catégorie d'émotions à l'exception du dégoût et que des caractéristiques spécifiques pouvaient être attribuées à des mouvements spécifiques. Le mouvement du tronc (le mouvement effectué du tronc aux jambes, le mouvement créé par le tronc allant du fléchissement de la tête aux genoux) était le plus prédictif pour toutes les émotions, sauf pour la colère, et se révélait capable de distinguer les émotions positives des négatives. Par exemple, la surprise est caractérisée par le torse droit et des jambes droites, un pas en arrière et des mouvements rapides, alors que la peur est caractérisée par un tronc et une tête abattus, des genoux légèrement fléchis, des mouvements du corps vers le bas, en arrière, rapides et des muscles tendus. En utilisant une approche fondée sur l'information, De Silva et Bianchi-Berthouze [31] ont étudié la pertinence de différents descripteurs caractérisant la posture corporelle dans la transmission et la discrimination entre quatre émotions fondamentales. Vingt-quatre descripteurs ont été utilisés pour décrire la position des articulations du haut du corps et l'orien-

tation des épaules, de la tête et des pieds pour analyser les postures efficaces de la base de données affective de l'UCLIC [86]. L'analyse statistique a montré que peu de dimensions étaient nécessaires pour expliquer la variabilité de la configuration des descripteurs et des résultats similaires ont été obtenus en regroupant les postures selon les étiquettes moyennes des observateurs. Les descripteurs verticaux étaient les plus pertinents pour séparer le bonheur de la tristesse. Plus précisément, les mains ont été levées pour le bonheur et sont restées basses le long du corps pour la tristesse. Les traits indiquant l'ouverture latérale du corps étaient les seconds plus informatifs, avec les mains significativement plus étendues pour la joie et la peur comparée à la tristesse. En utilisant le même ensemble de descripteurs, Kleinsmith et Bianchi-Berthouze [85] ont étendu cette analyse en examinant comment les descripteurs ont contribué à la discrimination entre différents niveaux des quatre dimensions affectives. Dans une étude récente, Kleinsmith et al. [86] ont cherché à aborder la question de l'obtention de postures affectives non actées. Ils ont collecté des données de mouvements de personnes jouant à des jeux de sport avec la Nintendo Wii (qui fait partie de la base de données affective UCLIC [86]) et ont utilisé des postures statiques à partir des données lorsqu'un point a été gagné ou perdu dans le jeu. Ensuite, chaque posture a été associée à un vecteur contenant une description de la posture au niveau le plus bas. Une analyse statistique des descripteurs a montré que les descripteurs les plus importants étaient principalement les bras et le haut du corps. Bien qu'il y ait eu une discrimination significative entre les quatre émotions distinctes, une plus grande discrimination a été obtenue entre les états affectifs plus "actifs" (frustrés et triomphants) et les états moins "actifs" (concentrés et défaite). Par exemple, les épaules étaient affaissées vers l'avant avec le bras tendu vers le bas et en diagonale à travers le corps pour la concentration et la défaite. Les postures frustrantes et triomphantes étaient indiquées avec les épaules droites vers le haut ou vers l'arrière et les bras levés et tendus latéralement.

Le modèle LABAN

Le modèle LABAN a été proposé par Rudolf Laban [136], un chorégraphe hongrois, dans le but de caractériser l'expressivité de mouvements de danse. En effet, il a développé une analyse de mouvement du nom de Laban Movement Analysis (LMA). Cette analyse se base sur 4 grands axes : le flux (flow), l'espace (space), le poids (weight) et le temps (time). La figure 2.9 représente ces quatres axes. Dans son étude, Laban définit différents plans du mouvement : le plan de la table (horizontal), de la porte (vertical) et de la roue (sagittal). Il construit une sphère du mouvement, la kinéspère, qui désigne l'espace accessible directement aux membres d'une personne (voir Figure 2.10). Elle s'étend tout autour d'elle, jusqu'à l'extrémité de

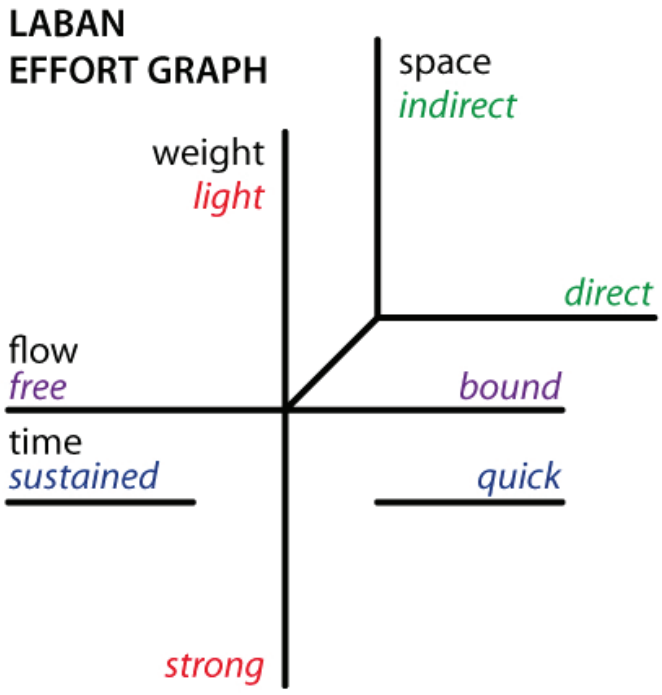


Figure 2.9 : Diagramme de l'effort selon Laban : poids (léger ou fort), espace (indirect ou direct), temps (soutenu ou urgent) et flux (libre ou contenu).

ses doigts et pieds tendus dans toutes les directions. Le modèle LABAN a été étendu et se base maintenant sur 6 catégories distinctes.

— Le corps. Qu'est-ce qui bouge, et comment ? Quel est le mouvement produit ?

— L'espace. Où va le mouvement ? Dans quel espace s'inscrit-il ?

— L'effort. Comment le mouvement est-il exécuté ? Avec quelles qualités d'énergie ?

— La forme. Quels sont les différents chemins empruntés par le mouvement ?

— Le phrasé ou le rythme. Dans quel laps de temps et suivant quel rythme s'effectue le mouvement ?

— L'interrelation. Comment l'individu en mouvement est-il en relation avec son entourage ?

Ce modèle est devenu très populaire en vision par ordinateur tant pour la reconnaissance d'actions que pour la reconnaissance d'expressions corporelles.

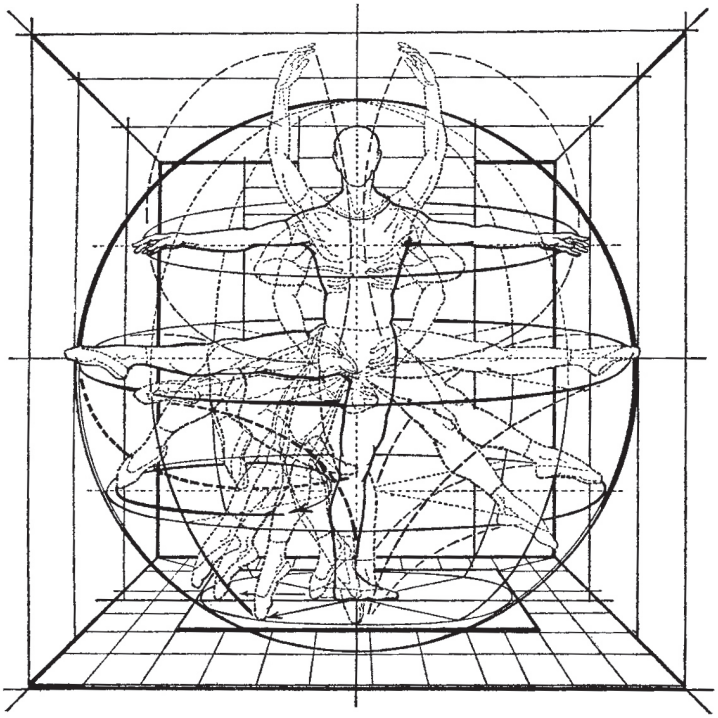


Figure 2.10 : Présentation de la kinésphère. Elle représente la globalité des endroits de l'espace que l'on peut atteindre lorsque l'on se tient sur un pied.

Bilan des études en psychologie

En général, les différentes études du domaine de la psychologie montrent que l'analyse de la posture peut permettre une discrimination entre différents états affectifs. Cependant, cela est loin de signifier qu'il existe une relation unique entre une émotion discrète et une expression corporelle. De même, la plupart des études sur les expressions corporelles semblent également présenter des schémas assez discriminatoires lorsqu'on considère des combinaisons de descripteurs plutôt que des descripteurs individuels. Par exemple, Wallbott [137] montre que dans certains cas, comme pour le dégoût et la fierté, les bras vont être croisés, mais la caractéristique discriminante semble être la position de la tête (penchée vers l'avant pour le dégoût et vers l'arrière pour la fierté). Typiquement, les éléments centraux de l'expression corporelle de la détresse sont les bras tendus vers le bas, près du côté du corps.

Alors que certaines catégories d'expressions (ou familles d'expressions) partagent un ensemble de descripteurs, elles présentent également un certain nombre de variations pour d'autres parties du corps. Par exemple, le bonheur et la joie partagent la caractéristique de la tête penchée en arrière dans plusieurs études [24, 86, 138]. Cependant, bien que les bras soient levés dans plusieurs cas, ils restent droits dans l'étude de Roether et coll. [121] ce qui pourrait être du au contexte

de la démarche. Tout au long de notre étude de la bibliographie, nous avons vu divers exemples où plus d'un modèle d'expression corporelle peut être associé à la même catégorie d'émotions. Dans certains cas, les modèles partagent certains descripteurs (voir ci-dessus pour le bonheur). Dans d'autres cas, les schémas d'une même classe d'émotions semblent très différents les uns des autres. Cela renforce l'idée qu'il existe des facteurs contextuels qui peuvent affecter la façon dont un état émotionnel est exprimé. Tout cela semble en accord avec Russell [122], qui affirme que les expressions prototypiques sont en fait assez rares. Selon Russell, les descripteurs qui forment l'expression d'une émotion ne sont pas fixes et chaque réaction émotionnelle n'est pas unique. Ce constat renforce notre idée qu'il est nécessaire de proposer une méthode capable de séparer le mouvement réalisé de l'expression perçue plutôt que de chercher à proposer une méthode caractérisant l'expression corporelle à partir d'un mouvement.

De plus, le principal problème des différentes études effectuées est qu'elles sont basées sur des expressions jouées. Pour cela, il est nécessaire d'augmenter le nombre d'études réalisées sur des expressions non actées, on pourra trouver des modèles moins distincts mais toujours discriminatoires lorsque le contexte n'est pas pris en compte comme le montre Kleinsmith et al. [83]. Une autre question importante qui ressort immédiatement de cet état de l'art est le manque de vocabulaire commun utilisé par le domaine de la psychologie pour décrire les descripteurs. Les descriptions semblent souvent fondées sur des évaluations subjectives et qualitatives et sont donc difficiles à comparer d'une étude à l'autre. De plus, les descripteurs de haut niveau sont très dépendants du contexte de l'action réalisée et difficiles à comparer sans décomposer et interpréter les termes. Dans l'ensemble, pour les descriptions de haut et de bas niveau, il est nécessaire d'utiliser systématiquement des descripteurs communs, et éventuellement numériques, afin de comparer plus objectivement les expressions du corps, comme le montrent Gross et al [57].

2.4.2 Reconnaissance automatique d'expressions corporelles

Comme présenté dans la Section précédente de ce chapitre, les chercheurs en psychologie ont été les premiers à s'intéresser aux expressions corporelles selon la posture et le mouvement du corps. Ainsi, les différentes tentatives de formalisation ont dégagé deux niveaux de descripteurs : les descripteurs de haut niveau et ceux de bas niveau. Les descripteurs de haut niveau dépendent du contexte de l'action et sont liés au mouvement. De façon formelle, les descripteurs de bas niveau se réfèrent aux articulations : angle de rotation, position 3D, vitesse, accélération, distance entre articulations, etc. Le domaine de la vision par ordinateur a donc commencé par chercher à formaliser ces différentes observations afin de proposer

un ensemble de descripteurs capables de reconnaître automatiquement l'expression corporelle d'un mouvement spécifique. Dans cette section, nous allons présenter les différentes méthodes existantes pour la reconnaissance automatique d'expressions corporelles.

Karg et al. [76] ont examiné la reconnaissance automatique d'expressions corporelles en fonction de la valence et de l'arousal (voir Section 2.2.2) dans des mouvements de marche. Les taux de reconnaissance étaient les meilleurs pour l'arousal, et les pires pour la valence. Kleinsmith et Bianchi-Berthouze ont d'abord examiné la reconnaissance automatique de l'expressions des postures du corps entier à l'aide d'une description bas niveau de la posture dans une situation jouée[14, 85], puis dans une situation spontanée [83]. Dans leur deuxième méthode, des modèles automatiques ont été construits pour reconnaître les niveaux de quatre dimensions affectives [85]. Bien que ces modèles aient atteint des taux de classification élevés, ils ont obtenu des scores de classification plus bas que d'autres modèles en utilisant des catégories discrètes. Dans leur dernière méthode, Kleinsmith et Bianchi-Berthouze [83] ont utilisé des postures non jouées et des états affectifs subtils obtenus à partir de personnes jouant à des jeux vidéo. Leurs modèles ont atteint des taux de reconnaissance inférieurs à ceux de leurs études jouées, mais similaires au taux cible fixé en calculant le niveau d'accord entre les observateurs.

Nous allons maintenant présenter des méthodes se focalisant sur la reconnaissance d'expressions dans des mouvements de danse [20, 113, 21]. Camurri et al [20] ont examiné différents descripteurs impliqués dans l'expression des émotions en danse pour quatre états affectifs. Après avoir supprimé les informations faciales, un ensemble de repères de mouvement a été extrait et utilisé pour construire des systèmes de reconnaissance automatique. La reconnaissance de la peur a été la plus mauvaise, atteignant des taux de classification inférieurs au hasard. La peur était le plus souvent confondue avec la colère. Il s'agit d'un résultat intriguant parce que le mouvement du corps a été utilisé à l'opposé de postures statiques et, comme le suppose Coulson [105], l'information dynamique peut aider à augmenter le taux de reconnaissance de la peur. La classification erronée de la tristesse en tant que joie est également intéressante compte tenu de l'examen par les auteurs de la qualité du mouvement, qui a montré que les mouvements de joie étaient très fluides et que les mouvements de tristesse étaient tout à fait le contraire. Kapur et al [75] ont utilisé des mouvements de danse réalisés à la fois par des danseurs professionnels et aussi par des amateurs de danse. Les observateurs ont correctement classé la majorité des mouvements et les modèles de reconnaissance automatique ont obtenu des taux de reconnaissance comparables. L'utilisation de mouvements de danse pour construire des systèmes de reconnaissance d'expressions corporelles est intéressante, cependant, ces mouvements sont exagérés et visent délibérément à transmettre les expressions corporelles.

Truong et al. [131] ont proposé un nouvel ensemble de descripteurs basé sur la formalisation du modèle Laban, décrit dans la Section 2.4.1. Leur objectif était de faire de la reconnaissance d'actions. Cependant ils ont aussi construit une base de données de mouvements expressifs et ont appliqué leur méthode dessus. Les résultats obtenus sont très prometteurs mais les mouvements expressifs évalués sont assez limités en terme de diversité d'actions. Une autre méthode qui a proposé une formalisation du modèle Laban a été proposée par Dewan et al. [34]. Ils ont étendu le modèle Laban en le combinant avec une fenêtre temporelle afin de segmenter le mouvement en entrée. Ils ont évalué leur méthode sur la base de données UCLIC [86]. Les résultats sont prometteurs mais la méthode est très spécialisée pour cette base de données impliquant un type d'action.

Bilan de la reconnaissance d'expressions corporelles

Comparé à la reconnaissance d'expressions faciales, le domaine de la reconnaissance d'expressions corporelles est encore très jeune. Il n'existe pas encore une ou plusieurs bases de données communes pour comparer les différentes méthodes. En effet, beaucoup de chercheurs utilisent des bases de données internes afin d'évaluer leur méthode. Du coup, la plupart des méthodes sont, pour l'instant, évaluées sur une unique base de données. Les bases de données publiques contiennent encore très peu de mouvements et / ou d'expressions différentes. C'est pourquoi les méthodes basées sur l'extraction de différents descripteurs obtiennent de très bons résultats. En effet, un ensemble de descripteurs est souvent construit pour une base de données spécifiques, i.e. pour un ensemble de mouvements connus. Notre objectif dans cette thèse est de proposer une méthode générique, en terme d'actions, de reconnaissance d'expressions corporelles. C'est pourquoi toutes nos méthodes proposées ont été évaluées sur différentes bases de données disponibles. La description des bases de données utilisées dans ce manuscrit est présentée dans la Section 2.6. L'objectif de reconnaitre l'expression corporelle pour différents mouvements nous a conduit à nous intéresser au domaine de la synthèse d'animations afin de comprendre comment les chercheurs dans ce domaine arrivent à extraire l'expression d'une animation pour la transférer sur une autre animation.

2.5 Synthèse d'expressions dans des animations

Nos travaux portant sur l'inverse du problème du domaine de la synthèse d'animations, nous avons voulu analyser les différentes méthodes de ce domaine afin de

nous en inspirer dans le but d'extraire l'expression d'un mouvement. De même, l'objectif à long terme de cette thèse est d'utiliser les différents outils présentés pour la reconnaissance d'expressions corporelles afin de pouvoir modéliser des variations de style dans des mouvements de personnages 3D. Pour l'instant, les applications actuelles que nous visons (jeux vidéos ou films) possèdent une énorme base de données de capture de mouvements réalisée à l'aide de nombreux acteurs. On demande à chaque acteur de réaliser un mouvement avec un certain style et on répète ce processus pour chaque combinaison de mouvement / style nécessaire à l'application. Cependant, ce processus est très coûteux en temps et en moyens. L'autre méthode pour réaliser cette base de données est de demander à différents artistes de créer ces animations à la main ce qui est très long et répétitif. C'est pourquoi la recherche s'est intéressée aux méthodes de translation de style afin de pouvoir rapidement transformer un mouvement existant en différents styles, tout en préservant son contenu. Le problème de la translation de style peut être résumé de la manière suivante. Supposons que nous avons une animation, A, et une autre animation, A', réalisant la même action que A mais dans un style différent. Nous voulons transformer une autre animation B, réalisant une action différente des deux premières animations, A et A', mais dans le style de A'. En d'autres termes, nous voulons trouver l'animation B' qui relie B de la même manière que A' relie A. Dans cette section, nous allons présenter les différents outils et espaces de représentations utilisés dans le domaine de la synthèse d'animation afin de résoudre ce problème.

Approches basées statistiques

Des méthodes vont chercher à construire un modèle statistique de mouvement afin d'interpréter des variations de mouvements d'un facteur [16] ou de plusieurs facteurs [139, 98]. La méthode de Brand et Hertzman [16] va chercher à apprendre un espace linéaire de style spécifique à partir d'un modèle de Markov caché pour différents individus. Dans la continuité de la méthode de Hsu et al. [68], des méthodes basées sur les processus gaussiens ont été introduites. Les processus gaussiens sont des techniques non paramétriques qui permettent d'aboutir à une distribution de probabilité suivant une loi gaussienne dans l'espace de sortie en fonction d'un vecteur en entrée. Ces modèles ont été introduits par Lawrence [91]. Ainsi, Grochow et al. [56] ont présenté « the Scaled Gaussian process latent variable model » qui permet d'apprendre une fonction de mapping entre un espace latent à basses dimensions et un espace à hautes dimensions contenant les poses d'un caractère. En combinant ce modèle avec un système de cinématique inverse, ils peuvent choisir de nouvelles poses expressives. Wang et al. [139] ont séparé le contenu du style en utilisant des processus gaussiens à plusieurs facteurs qui représentent l'identité de la personne, la démarche et l'état courant de l'animation. L'approche proposée

par Ma et al. [98] diffère de Wang et al. [139] car ils appliquent des techniques d'analyses multilinéaires afin de construire un espace à basses dimensions qui soit compact pour une séquence de mouvement complète. Ainsi, leur méthode permet d'assurer aux animations produites d'avoir une structure de mouvement humaine correcte de même pour les informations de contact avec l'environnement ce qui permet de réduire les artefacts tels que le footsliding, c'est-à-dire que les pieds vont glisser lors d'un mouvement de marche par exemple. Leur travail est en rapport avec l'approche proposé par Liu et al. [94] qui utilisent des techniques d'optimisation inverse afin d'apprendre les paramètres physiques du mouvement en entrée afin de modéliser différents styles d'animations. Parmi les méthodes permettant de modifier le style d'une animation de manière robuste sur des données très hétérogènes, c'est-à-dire contenant des actions différentes et éloignées, nous pouvons citer la méthode de Xia et al. [146]. Leur méthode est proche du travail de Hsu et al.[68] et cherche à modéliser la relation entre différents styles de mouvement en utilisant une base de données d'apprentissage et, ensuite, utilise le modèle appris afin de transformer un mouvement en entrée en différents styles. Cependant leur modèle utilisé est différent des autres car il modélise les différences de style en utilisant une série de mixtures de modèles régressifs locaux leur permettant de gérer des données non labellisées et hétérogènes.

Approches basées traitement de signal

Une autre famille de méthode permettant de modifier le style d'une animation est celle travaillant sur le signal de l'animation. Ainsi, ces méthodes vont traiter les différentes valeurs d'angles pour chaque articulation comme un signal temporel sur lequel on va appliquer différents filtres et méthodes. Bruderlin et Williams [17] présentent l'utilisation de techniques de traitement de signal et d'image dans le domaine de l'animation. Ainsi, ils montrent que ces méthodes peuvent être appliquées afin de désigner, modifier et d'adapter des animations. Pour cela, ils présentent différentes méthodes de traitement de signal tel que le filtrage multirésolution, l'interpolation entre différentes cibles, etc. Unuma et al. [132] utilisent le domaine de Fourier afin de calculer le spectre de différentes données d'animations d'un comportement humain. Cette base va leur servir à interpoler ou extrapoler différentes animations d'humains. Ainsi, ils peuvent créer une transition entre la marche et la course de manière fluide et réaliste. De plus, ils arrivent à extraire « l'humeur » via leur méthode ce qui leur permet de générer des animations riches et variées. Leur méthode offre une interface permettant d'éditer des mouvements de motion capture afin de transformer leur style. De même, Witkin et Popovic [144] ont introduit une méthode permettant de facilement éditer des données de motion capture

ou des animations utilisant des images clés. Pour cela, ils ont présenté une méthode appelée motion warping qui permet de déformer les paramètres des courbes d'animations. Amaya et al. [3] vont chercher à calculer des transformations d'émotions qui vont être ensuite appliquées à des animations existantes afin de modifier l'expression du mouvement. Pour cela, les transformations cherchées capturent la différence entre un mouvement neutre et un mouvement expressif en respectant deux composants. Ces derniers sont la vitesse et l'amplitude spatiale du mouvement. Une méthode plus récente permettant de transférer le style d'une animation sur différentes actions hétérogènes a été proposée par Yumer et Mitra [148]. L'idée principale de cette méthode est basée sur la transformée de Fourier. En effet, en calculant la différence dans le domaine spectral entre une animation neutre réalisant un mouvement et une animation stylisée réalisant le même mouvement, ils obtiennent le résidu contenant uniquement la différence entre ces animations, c'est-à-dire le style. En appliquant ce résidu, toujours dans le domaine spectral, sur une autre animation neutre, on peut donc modifier le style de celle-ci afin de lui appliquer le style voulu. La Figure 2.11 présente la méthode proposée par Yumer et Mitra. Ce principe marche très bien lorsque les actions réalisées par les animations sont proches. Afin de pouvoir transférer le style sur des animations réalisant des mouvements différents, ils ont proposé une fonction de coût synthétisant le mouvement stylisé. Cette fonction de coût va reconstruire la phase et la magnitude du mouvement en respectant des contraintes sur le timing afin de pouvoir correctement transférer le résidu sur les articulations effectuant le mouvement. En effet, lorsque l'on cherche à transférer le style d'un mouvement de coup de poing à un mouvement de coup de pied, le résidu que l'on va extraire contenant le style souhaité va être principalement localisé soit dans le haut du corps, pour le mouvement de coup de poing, soit dans le bas du corps, pour le mouvement de coup de pied. Leur fonction de coût va donc permettre de résoudre ce problème.

Approches basées sur l'apprentissage

Plus récemment, les méthodes d'apprentissage profond ont montré une bonne expressivité sur de grandes bases de données tout en nécessitant peu de supervision. [67] emploie un auto-encodeur convolutionnel et montre que l'interpolation sphérique dans la couche de code permet, contrairement à l'interpolation des coordonnées ou des quaternions, de produire des mouvements réalistes. Le potentiel de cette architecture est augmenté dans [65], en adjoignant deux blocs successifs à la partie décodeur pré-entraînée (figure 2.12). Le premier d'entre eux est une simple couche chargée de produire une représentation des paramètres de contrôle haut niveau tenant compte des contacts des pieds avec le sol. Le second est composé de plusieurs couches de convolution chargées de traduire la sortie du premier bloc

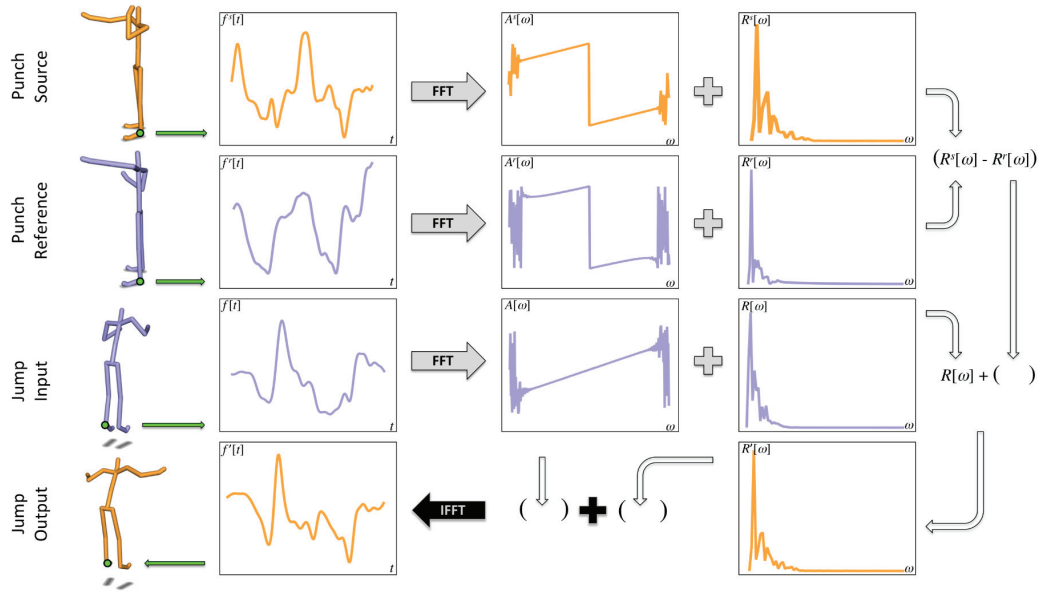


Figure 2.11 : L'extraction du résidu est obtenue par soustraction dans le domaine spectral entre le mouvement neutre (Source) et le mouvement contenant l'expression souhaitée (Référence). Le résidu est ensuite appliqué sur le mouvement neutre souhaité (Input) et permet d'obtenir le mouvement souhaité avec l'expression choisie (Output).

dans la variété de l'auto-encodeur. Cette méthode a l'inconvénient de produire des glissements des pieds sur le sol, des variations dans la longueur des os, ou de ne pas respecter précisément les paramètres de contrôle. Il est cependant possible d'appliquer des contraintes à une animation ainsi générée : on n'utilise plus que le bloc vert dans la figure 2.12, et on rétropropage le gradient de l'erreur entre l'animation obtenue et les contraintes que l'on souhaite appliquer pour modifier les valeurs de la couche de code, c'est à dire l'entrée du bloc vert. On détermine ensuite l'animation correspondant à ce nouveau code par un passage à travers le décodeur. Ce processus est répété autant que nécessaire. On obtient ainsi l'animation la plus proche possible, respectant au mieux les contraintes posées. Le fait d'optimiser dans la variété de l'auto-encodeur présente l'avantage de produire des mouvements toujours plausibles. Dans cette méthode, Holden applique avec succès la méthode basée sur les matrices de Gram utilisée dans [52] afin de réussir à transférer le style d'une animation à une autre.

Bilan de la synthèse d'animations

En analysant les différentes méthodes de la synthèse d'animations expressives avec l'objectif de proposer une méthode de reconnaissance, nous avons pu tirer énormément d'idée. En effet, le domaine de la synthèse d'animations a proposé de

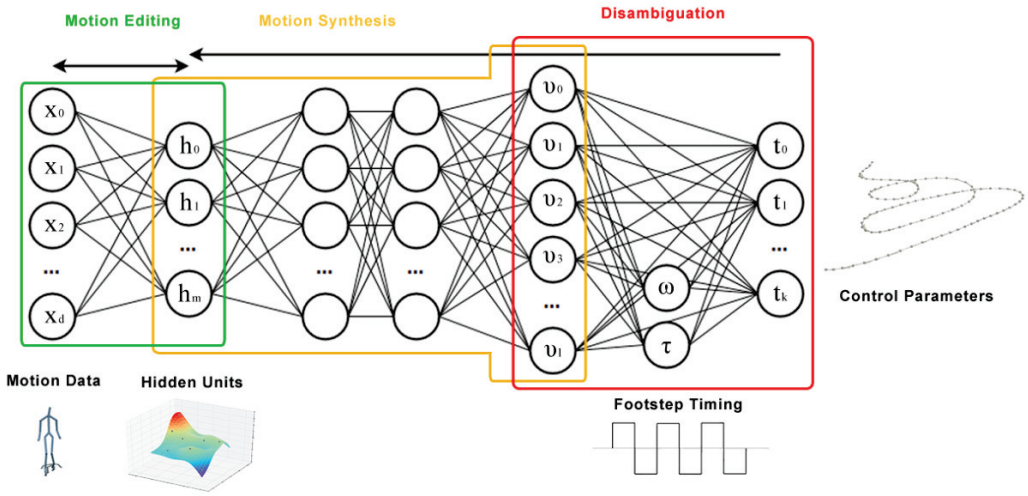


Figure 2.12 : Schéma global de la méthode d'Holden et al. [65]

nombreuses méthodes permettant de représenter une animation dans un autre espace. En appliquant différents traitements sur cet espace de représentation, les chercheurs du domaine de la synthèse d'animations arrivent à appliquer une expression sur une animation neutre dans le but de changer son expression. Contrairement aux approches du domaine de la vision par ordinateur qui vont chercher à formaliser différents descripteurs pour extraire l'expression d'un mouvement, le domaine de la synthèse d'animations cherche à extraire l'expression d'un mouvement en se basant sur un mouvement neutre et un mouvement expressif. En nous inspirant de ce domaine et en combinant des connaissances de la vision par ordinateur ainsi que du domaine de la psychologie, nous avons cherché à proposer une méthode générique afin de séparer le mouvement réalisé de l'expression perçue.

2.6 Bases de données disponibles et utilisées

Nous avons validé nos différentes méthodes de reconnaissances d'expressions corporelles sur différentes bases de données afin de tester la généricité et la robustesse des approches proposées dans ce manuscrit. Les bases de données sont Emilya [49], UCLIC Affective Body Posture and Motion [86], MPI Emotional Body Expressions Database for Narrative Scenarios [135], Biological Motion [99] et la base de données provenant de la méthode de [146] que l'on nommera SIGGRAPH dans la suite de ce manuscrit.

Dans la suite de cette Section, nous détaillerons les différentes caractéristiques des bases de données que nous avons utilisées dans nos travaux sur la reconnaissance

d'expressions corporelles. Une synthèse des bases de données est disponible dans le Tableau 2.1.

1. Base de données Emilya [49]. Cette base de données contient 10001 mouvements réalisant 7 actions différentes : marcher, s'asseoir, frapper à une porte, porter et jeter un objet avec une main et déplacer des objets à deux mains. Ces différents mouvements sont réalisés par 11 acteurs (6 femmes et 5 hommes) avec une moyenne d'âge de 26 ans. Les mouvements sont réalisés au travers de 8 expressions : joie, colère, panique / peur, anxiété, tristesse, honte, fierté et neutre. Il s'agit de la plus fournie des bases de données disponible publiquement.
2. UCLIC Affective Body Posture and Motion [86]. Cette base de données comporte 183 mouvements réalisées par 13 personnes de différentes cultures. Ces mouvements sont particuliers car il a été demandé aux acteurs de réaliser différentes postures en fonction d'une émotion. C'est pourquoi les mouvements réalisés sont extrêmement durs à décrire comme une action de la vie de tous les jours. Les acteurs ont réalisé différents mouvements selon quatre expression : la peur, la tristesse, la joie et la colère. Les mouvements ont été capturés par un système de motion capture aboutissant à un squelette contentant 32 articulations.
3. MPI Emotional Body Expressions Database for Narrative Scenarios [135]. Cette base de données rassemble 1447 mouvements réalisés par 8 acteurs, 4 femmes et 4 hommes, pour une moyenne d'âge de 25 ans. Il a été demandé aux acteurs d'imaginer qu'ils racontent différentes histoires à un enfant. Contrairement aux autres bases de données, les acteurs sont donc assis et discutent. L'émotion est donc présente dans le haut du corps principalement. Cette base de données contient 11 expressions : l'amusement, la colère, le dégoût, la peur, la joie, le neutre, la fierté, la tristesse, la honte, la surprise et le soulagement. Les postures ont été capturées par un système de motion capture résultant à un squelette contenant 22 articulations. Il est à noter que cette base de données est non équilibrée en terme d'expressions. En effet, la joie est l'expression la plus représentée avec 227 mouvements, contrairement à la honte qui est présente lors de 58 mouvements. Cette disparité entraîne une complexité lors de l'apprentissage pour la reconnaissance d'expressions corporelles.
4. Biological Motion [99]. Cette base de données contient 4080 mouvements : marcher, frapper à une porte, porter et lancer un objet. Ces mouvements sont réalisés par 30 acteurs, 15 hommes et 15 femmes, avec une moyenne d'âge de 22 ans. Ces acteurs ont réalisé les différentes actions avec 4 expressions, le neutre, la colère, la joie et la tristesse. Malheureusement, nous avons utilisé une sous partie de cette base de données car c'est la seule partie disponible

sur internet. Les méthodes de l'état de l'art ont aussi utilisé cette sous partie de la base de données afin de proposer des approches pour la reconnaissance d'expressions corporelles. Cette sous partie contient 1356 mouvements, principalement frapper à une porte. Le système de motion capture utilisé pour lors de la création de cette base de données fournit un squelette contenant 35 articulations.

5. SIGGRAPH database [146]. Cette base de données provient du domaine de la synthèse d'animations. Cette spécificité fait que les expressions possèdent un style très cartoon facilitant la reconnaissance d'expressions. Elle contient 572 mouvements : course, la marche, le saut, le coup de pied, le coup de poing et les transitions entre ces mouvements. Cette variété de mouvements nous a motivé à utiliser cette base de données. De plus, les mouvements sont réalisés avec 8 expressions ou styles différents : la colère, la dépression, la fierté, le neutre, le style enfantin, le style vieux, le style agressif et le style sexy.

Table 2.1 : Description des base de données utilisées lors de la thèse.

Base de données	Nombre de mouvements	Nombre d'expressions
Emilya [49]	10001	8
UCLIC [86]	183	4
MPI [135]	1443	11
Biological [99]	1356	4
SIGGRAPH [146]	572	4 and 4 styles

3 Reconnaissance d'expressions corporelles

Contents

Dans ce chapitre nous allons présenter les différentes approches pour la reconnaissance d'expressions corporelles que nous avons proposé lors de cette thèse. Tout d'abord, nous allons parler de notre nouvel ensemble de descripteurs afin d'extraire l'expression corporelle perçue lors d'un mouvement. Nous allons ensuite présenter les deux méthodes d'extraction d'expressions corporelles basées sur la synthèse d'un mouvement neutre. Pour finir, nous présenterons les bases de données utilisées pour la validation de ces différentes méthodes ainsi que les résultats obtenus. Le but de nos approches est de confirmer l'hypothèse présentée lors de l'état de l'art, à savoir qu'il est nécessaire de réussir à séparer le mouvement réalisé de l'expression perçue afin d'avoir une reconnaissance générique, c'est-à-dire qui soit invariante au mouvement présenté en entrée du système.

3.1 Approche basées descripteurs

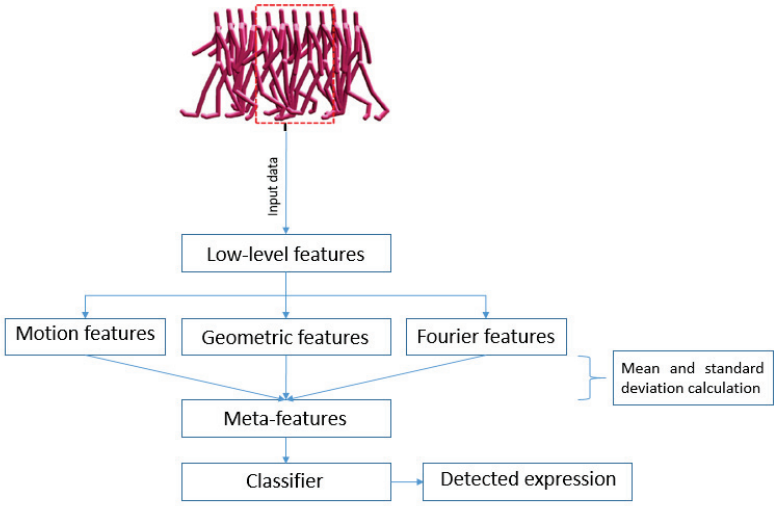


Figure 3.1 : Schéma global de notre approche

Dans cette partie, nous allons présenter notre première contribution pour la reconnaissance d'expressions corporelles. Cette méthode est basée sur l'extraction de différents descripteurs inspirés du domaine de la psychologie. L'objectif est de proposer un nouvel ensemble de descripteurs capable d'extraire l'expression corporelle sur des gestes très hétérogènes. Nous allons comparer les résultats obtenus avec les différentes méthodes de l'état de l'art. L'objectif de cette section est de montrer les forces et faiblesses des approches basées descripteurs. La principale force de ces méthodes réside dans leur spécialisation. En effet, lorsque l'on cherche à extraire l'expression corporelle d'un mouvement particulier, nous pouvons construire un ensemble de descripteurs spécialisés pour ce mouvement. Grâce à cette spécialisation,

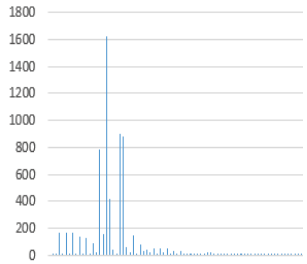
nous pouvons obtenir de très bon taux de reconnaissance d'expressions corporelles lorsque le mouvement est connu. Cependant, dans le scénario inverse, c'est à dire lorsque que le mouvement est inconnu ou lorsque nous travaillons sur différents mouvements très hétérogènes, ces méthodes se montrent limitées. En effet, les descripteurs que l'on cherche à extraire sont liés au mouvement réalisé et il est très dur de séparer le mouvement et l'expression. Notre approche est de résoudre ce problème en proposant un nouvel ensemble de descripteurs afin d'avoir une reconnaissance d'expressions corporelles la plus générique possible.

Comme énoncé précédemment, notre motivation est de proposer une méthode pour la reconnaissance d'expressions corporelles sur des mouvements hétérogènes qui rivalisent avec les approches spécialisées de l'état de l'art. En se basant sur des travaux existants [87, 131, 48], nous avons proposé un nouvel ensemble de descripteurs basés sur des descripteurs géométriques, des descripteurs du mouvement et des descripteurs fréquentiels. Le schéma global de cette méthode est proposé sur la Figure 3.1

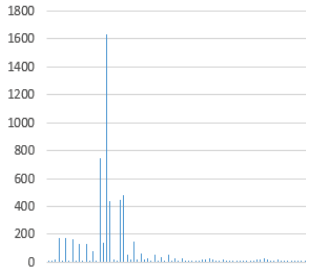
1. La première étape de notre modèle consiste à calculer des descripteurs que nous qualifierons de bas-niveaux pour chaque pose du mouvement fournie en entrée. Nous avons trié ces descripteurs en trois catégories : descripteurs géométriques, descripteurs de mouvement et des descripteurs de Fourier. Le détail de tous ces descripteurs est décrit ci-dessous.
2. A partir de cette série de descripteurs bas-niveaux obtenus sur une séquence temporelle, nous calculons des meta-descripteurs compatactant l'aspect temporel en un nombre réduit de valeurs. Nous proposons d'utiliser uniquement la moyenne et l'écart-type pour chaque descripteurs. Vu que les meta-descripteurs sont indépendant du temps, ce calcul nous permet d'éviter des étapes coûteuses en temps de calcul comme le time-warping pour synchroniser nos différents mouvements.
3. Ces meta-descripteurs sont ensuite fournis à un classifieur qui va, une fois entrainé, nous fournira l'expression corporelle correspondante au mouvement fourni en entrée de notre système.

En se référant aux études de [77] et [13], nous avons décidé d'utiliser uniquement les articulations suivantes pour l'extraction de descripteurs : la tête, le bassin, les épaules, les coudes et les mains. En effet, les résultats expérimentaux obtenus par [77] ont montré que la tête, le bassin, les coudes et les épaules sont les articulations qui sont les plus représentatives lors de la transmission d'une expression à travers un mouvement. Les résultats obtenus par [13] ont aussi prouvés que les mains sont très importantes afin d'interpréter une expression corporelle. C'est pourquoi, nous pensons qu'il est très important de se focaliser sur ces 5 articulations pour la reconnaissance d'expressions corporelles et ainsi gagner en compacité et

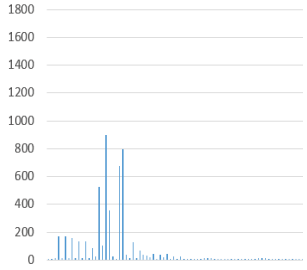
donc en performance. Les descripteurs bas-niveaux proposés sont basés sur des distances entre différentes articulations du corps, des aires de triangle construites entre des articulations spécifiques, vitesse et accélération des articulations et sur la fréquence du mouvement de différentes articulations. Le tableau 3.1 décrit l'ensemble des différents descripteurs. Nous utilisons différents types de descripteurs afin de correctement respecter différentes observations effectuées par des études de psychologies que vous pouvez retrouver dans l'état de l'art. L'ensemble des descripteurs proposé comporte 68 descripteurs. Ensuite, pour chaque descripteur nous calculons un meta-descripteur. Ce qui nous amène à un total de 136 descripteurs pour reconnaître l'expression corporelle d'un mouvement. La figure 3.2 présente l'ensemble des meta-descripteurs pour trois expressions différentes lors d'un mouvement consistant à frapper à une porte. Pour des raisons de présentations, nous affichons seulement les 76 premiers descripteurs de notre vecteur de 136 descripteurs. Dans la figure 3.2, nous pouvons observer que les descripteurs utilisés sont globalement discriminants pour les expressions présentées. Cette capacité discriminante est utilisée dans la méthode proposée afin de détecter de manière robuste les expressions corporelles lors de mouvement hétérogène.



(a) Histogramme de nos descripteurs pour l'action frapper à une porte en étant content



(b) Histogramme de nos descripteurs pour l'action frapper à une porte de manière triste.



(c) Histogramme de nos descripteurs pour l'action frapper à une porte en colère

Figure 3.2 : Présentation des descripteurs utilisés lors de la même action avec des expressions différentes. Ces histogramme montrent que nos descripteurs sont discriminants.

Id.	Type de descripteur	Description
V	Espace occupé par le squelette	La taille de la bouding box du squelette.
θ	Angle	Les 3 angles induits par le triangle formé entre les deux épaules et le cou. Angle entre la direction verticale de notre monde y et l'axe formé par le centre du bassin à la tête.
$\mathcal{D}$	Distance	Main droite et le bassin. Main gauche et le bassin. Main droite et épaule droite. Main gauche et épaule gauche. Coude droit et le bassin. Coude gauche et le bassin.
A	Aire	Triangle formé par les deux mains et le cou. Triangle formé par les deux épaules et le cou. Triangle formé par les deux mains et le bassin. Triangle formé par les deux coudes et le bassin.
$\vec{v}$	Vitesse	Mains. Épaules. Bassin. Tête. Coudes.
$\vec{a}$	Accélération	Mains. Épaules. Bassin. Tête. Coudes.
$\mathcal{F}$	Fréquentiel	Transformée de Fourier appliquée au signal de rotation des articulations suivantes : Mains. Épaules. Bassin. Tête. Coudes.

Table 3.1 : Ensemble des descripteurs que nous extrayons lors d'un mouvement pour reconnaître l'expression corporelle.

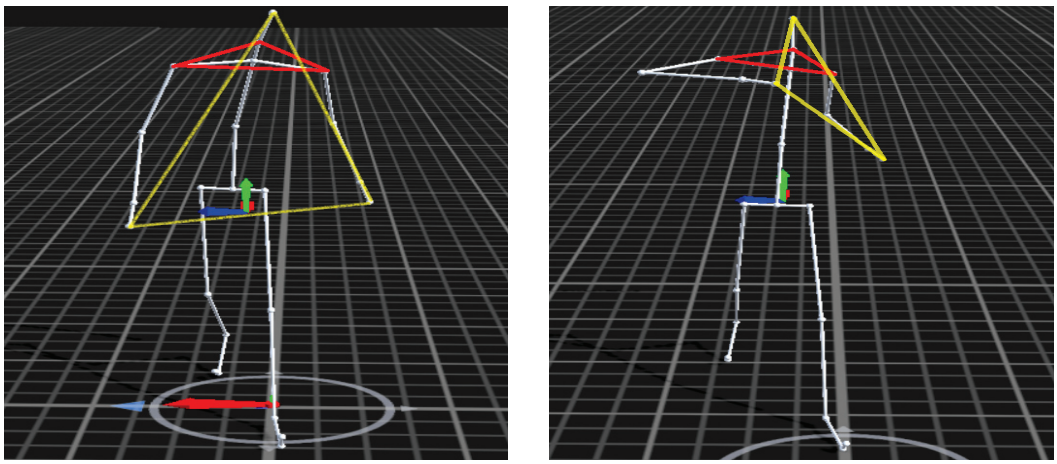
Les deux premiers types de descripteur que nous allons présenter dans ce paragraphe cherchent à caractériser la posture du corps de manière globale. Le premier descripteur, *V*, va calculer l'espace global occupé par le squelette pour chaque posture de notre mouvement. Nous utilisons la taille de la bounding box du squelette dans les 3 directions de notre monde. Le second descripteur, θ, est l'angle défini entre l'axe vertical de notre monde, y et l'axe formé à partir du centre du bassin au

cou du squelette. Ce descripteur nous permet d'avoir l'information sur la courbure du corps. En effet, en se référant à plusieurs études dans le domaine de la psychologie, une personne a tendance à étendre son corps lors d'expression positive. Au contraire, une personne qui cherche à transmettre une expression négative va adopter une posture plus compacte notamment en se penchant un peu vers l'avant.

Les descripteurs suivants vont caractériser une posture de manière plus précise. En effet, les descripteurs présentés ci-dessus caractérisent une posture de manière grossière. Afin de discriminer correctement différentes expressions, il est nécessaire d'affiner ce niveau de détail sur la posture. Pour cela, les distances, $\mathcal{D}$, entre différentes articulations sont très importantes. En effet, ces distances vont raffiner la posture en nous fournissant des informations comme si le corps est renfermé sur lui-même ou non. Nous utilisons la distance entre les mains et les épaules. De même, nous calculons la distance entre les mains et le bassin, les mains et les épaules et les coudes au bassins. En plus de ces distances, nous introduisons des descripteurs dont l'idée repose sur l'utilisation d'aires de triangles, A, pour raffiner les postures. Ces aires de triangle vont renforcer la qualité discriminative de notre approche. En effet, ces triangles vont nous fournir une information sur la forme de notre posture. L'idée de la construction de ces triangles est assez simple. Vu que le corps est symétrique et que pour une grande majorité de mouvement la symétrie est aussi temporelle notamment pour la marche ou la course. Ainsi, pour la création de nos triangles, nous prenons une articulation de chaque côté du corps, gauche et droite, et une articulation sur l'axe centrale du corps, soit le cou ou le bassin.

Comme mentionné lors de la partie État de l'art 2.4.1, les psychologues ont montré que l'étude du mouvement de différentes articulations précises est très important pour discriminer différents état émotionnels. En effet, en analysant uniquement la vitesse de la main lors d'un mouvement de frapper à une porte, les psychologues ont réussi à reconnaître différentes expressions corporelles. C'est pourquoi nous avons décidé d'utiliser les descripteurs suivants : la vitesse absolue, $\vec{v}$, et l'accélération absolue, $\vec{a}$, des deux mains, deux épaules, bassin, de la tête et des deux coudes. La vitesse est la dérivée première de la position de l'articulation cible. L'accélération est la dérivée seconde de cette position.

Pour finir, nous utilisons la transformée de Fourier, $\mathcal{F}$, pour obtenir le composant fréquentiel de l'articulation. Nous calculons la transformée de Fourier discrète via l'algorithme de la FFT sur le signal de rotation de l'articulation.



(a) Première pose d'un mouvement de marche déprimée.

(b) Première pose d'un mouvement de marche fière.

Figure 3.3 : Différentes poses d'un même mouvement mais avec des expressions corporelles différentes. Cette figure montre les variations de la zone triangulaire formée par les deux épaules et le cou.

3.1.1 Résultats et analyse

Nous avons évalué notre méthode sur trois base de données, résumées dans le tableau 3.2, et présentées dans la Section 2.6. Parmi ces bases de données, deux d'entre elles sont des bases de données jouées par des acteurs tandis que la dernière base de données contient des données synthétiques utilisées par la méthode de Xia et al. [146].

Base de données	Nombre de mouvements	Nombre d'expressions
UCLIC [**kleinsmith_cross**]	183	4
Biological [99]	1356	4
SIGGRAPH [146]	572	8

Table 3.2 : Description des bases de données utilisées pour l'évaluation de notre méthode.

Un résumé des résultats obtenus par notre méthode en utilisant différents ensembles de descripteurs est présenté dans le tableau 3.3. Notre approche proposée tourne en temps réel à ~50 fps (sur PC avec i7-4710MQ avec 8GB de RAM). L'extraction des descripteurs proposés, à l'exception des descripteurs de Fourier, prend 10ms pour chaque frame tandis que l'extraction des descripteurs de Fourier prend 500ms pour une séquence de 27 secondes avec 1657 frames. Les résultats présentés ont été obtenus en utilisant une machines à vecteurs de support (SVM en anglais : Support Vector Machine) avec un noyau de fonction de base radiale (noyau SVM

Base de données	Ensemble de descripteurs utilisés	Résultats
UCLIC	Ensemble des descripteurs	**78%**
UCLIC	Descripteurs géométriques	66%
UCLIC	Descripteurs de mouvements	52%
UCLIC	Descripteurs fréquentiels	61%
SIGGRAPH	Ensemble des descripteurs	**93%**
SIGGRAPH	Descripteurs géométriques	92%
SIGGRAPH	Descripteurs de mouvements	75%
SIGGRAPH	Descripteurs fréquentiels	90%
Biological	Ensemble des descripteurs	**57%**
Biological	Descripteurs géométriques	56%
Biological	Descripteurs de mouvements	48%
Biological	Descripteurs fréquentiels	46%

Table 3.3 : Nos résultats sur les différentes bases de données et avec différents ensembles de descripteurs. La méthode de classification est un SVM.

- RBF) pour la classification [22]. Les descripteurs proposés peuvent être classés dans les sous-catégories suivantes.

1. Descripteurs géométriques : distances, surface et angles de triangles.
2. Descripteurs de mouvements : vitesse et accélération.
3. Descripteurs fréquentiels : magnitude du spectre pour les différentes articulations.
4. Ensemble des descripteurs : combinaison des différents descripteurs présents ci-dessus.

Le meilleur résultat est obtenu en combinant les différents descripteurs proposés (voir tableau 3.3). Il est intéressant d'observer que la précision de la classification ne se dégrade pas de manière significative lorsque nous n'utilisons que les descripteurs géométriques. Sur les deux bases de données, SIGGRAPH et Biological Motion, les descripteurs géométriques ont juste perdu 1% de taux de reconnaissance comparé à l'ensemble des descripteurs. Sur la base de données UCLIC, le taux de reconnaissance des descripteurs géométriques est inférieur de 12% à celui de la combinaison de tous les descripteurs. Ceci est probablement dû au fait que cette base de données est significativement plus petite en taille d'échantillon (183 actions) et présente le pire scénario pour un algorithme d'apprentissage supervisé. Notre méthode a obtenu les meilleurs résultats sur la base de données synthétique (SIGGRAPH) qui est due au fait que les animations synthétiques présentent des expressions exagérées avec de fortes variations inter-classes.

Le tableau 3.4 montre que notre méthode est compétitive par rapport à l'état de l'art sur la base de données UCLIC avec un taux de reconnaissance similaire. Pour

Base de données	Résultats de l'état de l'art	Nos résultats
UCLIC	79% [87]	78%
Biological	50% non biaisé [13]	57%
SIGGRAPH	–	93%

Table 3.4 : Comparaison de notre méthode, en utilisant tous les descripteurs mentionnés dans cette section, à l'état de l'art.

la base de données Biological Motion, les mouvements sont principalement l'action de frapper à une porte ($\approx$ 1200 mouvements sur 1356 mouvements). L'état de l'art [13] utilise cette particularité pour calculer le mouvement moyen de frapper à une porte et ensuite soustraire ce mouvement avant d'exécuter la reconnaissance afin de mettre en évidence l'expression. Leur taux de reconnaissance pour cette méthode biaisée est de 81%. Néanmoins, cette astuce est possible lorsque l'on connaît le mouvement sur lequel on va extraire l'expression. Étant donné que le but de nos travaux est de proposer une méthode robuste à des mouvements hétérogènes, nous ne pouvons pas appliquer la même hypothèse qu'eux. Nous pensons que pour comparer [13] et notre approche, leur taux de reconnaissance impartial de 50% doit être comparé à notre approche qui a obtenu un taux de classification correcte de 57%. Avec ces résultats, on peut observer la différence du taux de reconnaissance lorsque l'on connaît le mouvement en entrée ou non. En effet, la différence du taux de reconnaissance est de 31% lorsque l'on injecte des connaissances sur le mouvement fournie en entrée. Enfin, à notre connaissance, aucune méthode de la littérature n'a testé la base de données SIGGRAPH pour la reconnaissance d'expressions corporelles.

3.2 Détection de l'expression corporelle en générant un mouvement neutre

3.2.1 Introduction

Comme on a pu le voir dans la Section 3.1, les approches basées descripteurs fournissent de bons résultats. Cependant, ce type d'approche est limité lorsque l'on travaille sur une grande variété de mouvements. En effet, les approches descripteurs fonctionnement extrêmement bien lorsque le mouvement est connu et que l'on peut spécialiser ces descripteurs pour un mouvement spécifique. Notre but étant de reconnaître l'expression corporelle quelque soit le mouvement en entrée, nous avons cherché à proposer une approche qui soit invariante au mouvement original. En nous inspirant du domaine de la synthèse d'animations où les méthodes se

basent sur un mouvement neutre afin de changer le style du mouvement exécuté, nous avons cherché à utiliser ce mouvement neutre pour la reconnaissance d'expressions corporelles. Pour cela, nous cherchons à synthétiser automatiquement un mouvement neutre à partir du mouvement expressif fourni en entrée. Ensuite, nous analysons les différences entre le mouvement expressif et le mouvement neutre synthétisé afin d'extraire l'expression contenue dans le mouvement réalisé. La synthèse d'un mouvement neutre permet d'avoir une méthode de reconnaissance d'expressions invariante au mouvement corporel exécuté.

3.2.2 Méthode proposée

Le principal objectif de ce travail est la reconnaissance d'expressions corporelles indépendamment du geste exécuté. Dans nos travaux, nous considérons que le résidu obtenu entre un mouvement expressif et un mouvement neutre contient l'information relative au style. Le résidu est calculé comme une différence, dans le domaine fréquentiel, sur chaque degré de liberté entre un mouvement expressif et un mouvement neutre. Cette idée a déjà été exploitée en synthèse d'animations par Yumer et Mitra [148]. A la différence de leur méthode, nous ne disposons pas du mouvement neutre.

3.2.3 Synthèse de mouvement neutre

Le premier verrou consiste donc à obtenir une animation neutre à partir d'un mouvement expressif. La notion de mouvement neutre versus expressif est toujours liée à un contexte. Nous définissons un mouvement neutre comme un mouvement comportant juste une action, c'est-à-dire non porteur d'une marque émotionnelle. Ainsi, un oeil humain ne pourrait qualifier le mouvement que par l'action produite et rien d'autre. Notre approche se base sur un filtrage des trajectoires de chaque articulation afin de réduire les oscillations dans le mouvement expressif. La seconde étape fait appel à la cinématique inverse pour produire un mouvement sans expression respectant les contraintes humaines. Finalement, on introduit une fonction de coût qui va réaliser un compromis entre le filtrage des articulations et la cinématique inverse. Cette fonction de coût va ensuite être optimisée afin de produire le mouvement neutre cherché. Nous passons par une optimisation d'une fonction de coût car nous pouvons obtenir plus de contrôle sur le mouvement produit comparé à des méthodes procédurales. Pour cela, nous proposons une fonction de coût qui caractérise un mouvement neutre. Cette fonction est basée sur des caractéristiques cinématiques (distance, vitesse, accélération) calculées pour chaque articulation durant un mouvement. La figure 3.4 donne le schéma général de notre méthode.

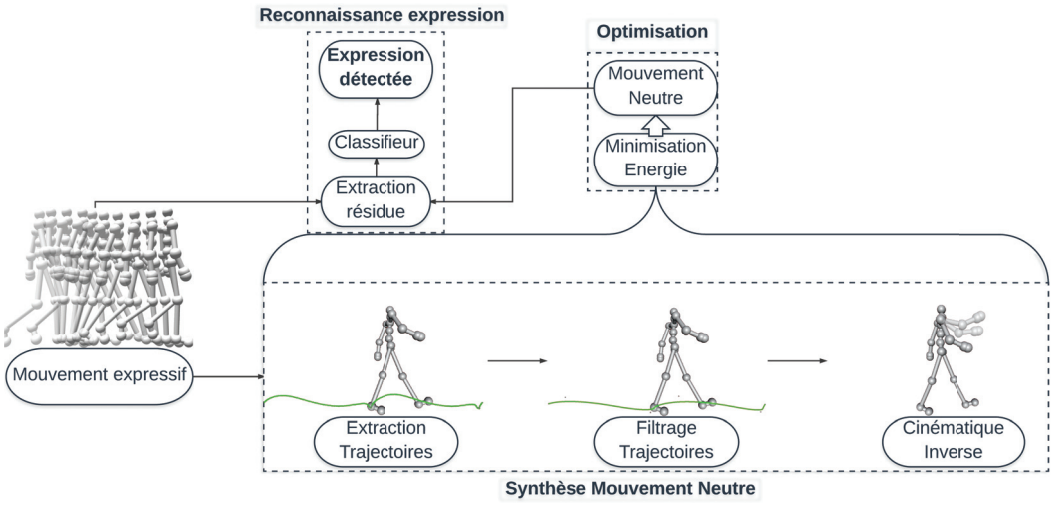


Figure 3.4 : Schéma général de notre méthode. L'objectif est de synthétiser un mouvement neutre correspondant au mouvement expressif fourni en entrée du système afin de séparer le geste réalisé de l'expression perçue.

3.2.4 Cinématique Inverse

Dans cette Section, nous allons présenter ce qu'est la cinématique inverse et les différents algorithmes existants. Le problème de la cinématique inverse est de trouver une pose afin d'atteindre une cible spécifiée. Ainsi, il faut trouver la pose nécessaire à notre chaîne articulée qui permet d'atteindre notre objectif. Pour un modèle de corps humain, l'algorithme doit donc déterminer la torsion des poignets, des coudes, des doigts, etc. (en un mot, de toutes les articulations du corps) afin d'atteindre la posture désirée. Plutôt que de spécifier à la main un ensemble de coordonnées articulaires, les algorithmes de cinématique inverse vont chercher à résoudre ce problème de manière automatique. En plus du domaine de l'animation, ces algorithmes sont utilisés dans le domaine de la robotique. En effet, on peut commander un robot en termes d'objectifs (par exemple la position du centre de masse) et déterminer par cinématique inverse les angles articulaires qui seront envoyés comme commande aux moteurs du robot. La principale difficulté de la cinématique inverse est de gérer l'interdépendance des différents éléments du squelette ou de la chaîne articulée sur laquelle on travaille et de gérer la non unicité de la solution. La Figure 3.5 présente le problème sur une chaîne articulée simple composée de deux articulations.

L'approche numérique la plus populaire consiste à utiliser la matrice jacobienne pour trouver une approximation linéaire au problème de cinématique inverse. Les solutions utilisant la jacobienne modélisent linéairement les mouvements des effecteurs terminaux par rapport aux changements instantanés du système dans la translation des liens et de l'angle de jonction. Plusieurs méthodes différentes ont

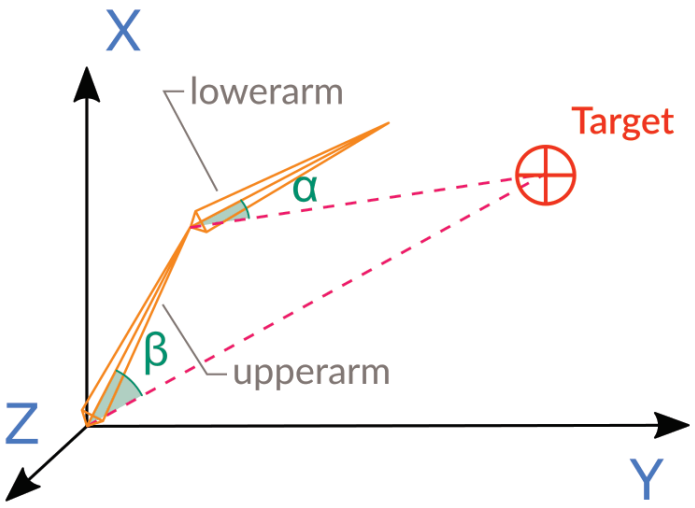


Figure 3.5 : Présentation du problème de la cinématique inverse. L'objectif est d'amener la fin de notre chaîne articulée à la cible (Target dans le schéma) désirée.

été présentées pour calculer ou approximer l'inverse jacobien [8, 145]. Les solutions inverses jacobiennes produisent des postures lisses, cependant, la plupart de ces approches souffrent de coûts de calcul élevés, de calculs matriciels complexes imprécis dans le cas des singularités. Une autre méthode de cinématique inverse très populaire est l'algorithme de descente de coordonnées cycliques (CCD), qui a été introduit pour la première fois par Wang.et Chen [140], puis soumis à des contraintes biomécaniques par Welman [142]. Le CCD est une méthode heuristique itérative à faible coût de calcul pour chaque articulation par itération, ce qui permet de résoudre le problème de cinématique inverse sans manipulations matricielles. Cependant, le CCD présente certains inconvénients, il peut souffrir d'une animation irréaliste, et produit souvent du mouvement avec des discontinuités erratiques. Dans nos différentes méthodes, nous utilisons l'algorithme de FABRIK [5] qui est un solveur de cinématique inverse simple assez récent, rapide et fiable. C'est le premier algorithme à utiliser une méthode itérative avec des points et des lignes pour résoudre le problème de cinématique inverse. De même, cet algorithme a reçu une extension [4] où un modèle humanoïde est mis en œuvre. Ainsi, FABRIK peut être appliqué de manière hiérarchique et séquentiellement pour suivre de multiples cibles tout en s'exécutant en temps-réel. C'est pourquoi nous avons retenu FABRIK qui s'adapte très bien à notre problème tout en étant simple à implémenter et très efficace notamment pour gérer des contraintes humaines.

Filtrage de la trajectoire d'une articulation

Un mouvement est représenté par des échantillons temporels de chaque articulation correspondant aux angles de chaque articulation de notre squelette. La figure3.6 montre la représentation d'une trajectoire d'une articulation. Dans la première figure, nous représentons la trajectoire 3D de chaque articulation par une B-spline. Afin d'atténuer les oscillations dans la trajectoire initiale, nous procédons à un "filtrage" de cette trajectoire. Pour cela, nous réduisons chaque B-spline en retirant un point de contrôle, à chaque itération de notre algorithme, pour chaque courbe de Bézier composant cette B-spline.

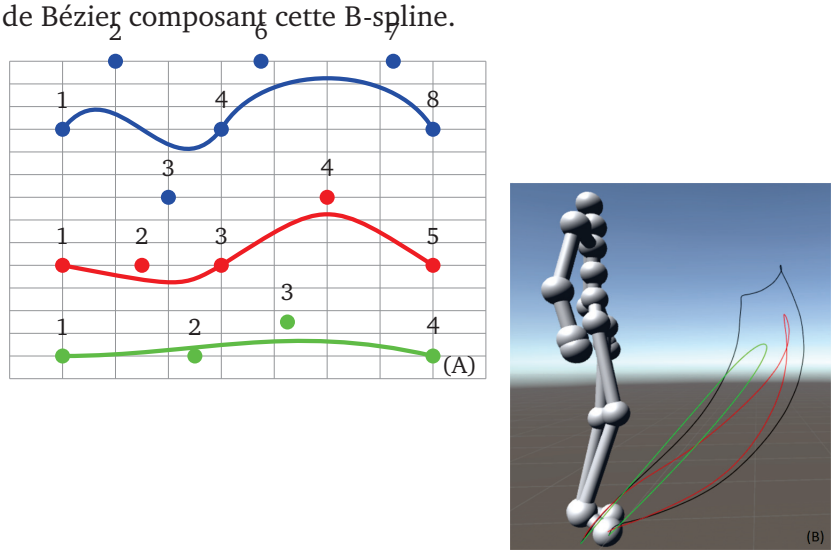


Figure 3.6 : Figure A, réduction de la trajectoire par filtrage de B-spline à partir du mouvement initial (en bleu) : en rouge (la trajectoire simplifiée) après une (resp.deux) itération. Figure B, comparaison de différentes trajectoires lors d'un mouvement de coup de pied. Le mouvement original est en noir (colère). La vérité-terrain est représentée en rouge. La trajectoire "neutre" générée par notre méthode après optimisation est en vert.

L'étape de filtrage permet de réduire les oscillations contenues dans le mouvement expressif de façon incrémentale. En analysant différents mouvements expressifs contenus dans différentes bases de données, nous avons catégorisé deux styles de mouvements expressifs : les styles énergiques (joie, fierté, colère, etc.) et les styles modérés (déprimé, âgé, etc.). Par ailleurs, nous avons également constaté qu'un mouvement neutre est plus "plat" qu'un mouvement énergique, c'est-à-dire qu'il présente une trajectoire avec des oscillations moins présentes. Notre étape de filtrage des trajectoires permet de générer des mouvements neutres dans ce cas-là. Le filtrage de la trajectoire aura peu d'effet sur les mouvements modérés. Pour cela, le recours à une étape de cinématique inverse combinée à la fonction de coût permet de synthétiser des animations neutres. En effet, l'étape de cinématique inverse est tout d'abord nécessaire car les mouvements générés par l'étape de lissage

ne respectent plus différentes contraintes (longueur des articulations, pieds qui glissent). D'autre part, comme il a été signalé ci-dessus, un mouvement basique est une bonne référence pour un mouvement neutre et la cinématique inverse permet de produire ce type de mouvement. C'est pourquoi nous appliquons une étape de cinématique inverse (IK) pour déterminer un ensemble de postures. L'étape de filtrage nous fournit différentes positions pour chacune des articulations de notre squelette. Nous utilisons les positions des deux mains ainsi que des deux pieds comme position cible pour l'étape de cinématique inverse. Nous fournissons également les positions des coudes et genoux comme indice afin de respecter, en partie, la posture obtenu par le filtrage des trajectoires.

L'algorithme 1 présente la génération d'un mouvement à partir de n'importe quel mouvement expressif. Cette méthode peut produire des mouvements peu naturels dans certains cas. En effet, la phase de filtrage des trajectoires peut amener à des mouvements manquant de fluidité, c'est-à-dire que nous retirons de l'information à notre mouvement fournit en entrée de notre système. Vu que l'étape de cinématique inverse est basée sur la phase de filtrage des trajectoires, nous n'arrivons pas à compenser ce manque d'information. Cette observation se retrouve surtout sur des mouvements où les pieds ne touchent plus le sol comme un mouvement de saut. Cependant, comme le montre le taux de reconnaissance obtenu dans la section 3.2.5, elle reste néanmoins efficace. Notre algorithme est générique et fait intervenir différents paramètres qui le rendent adaptables à tous types de mouvements. Nous présentons dans la section suivante la fonction de coût utilisée pour la synthèse de mouvement neutre.

Fonction de coût pour la synthèse d'un mouvement neutre

Nous proposons une fonction de coût qui caractérise un mouvement neutre. Cette dernière est relativement facile à implémenter et permet de produire, après optimisation, des animations neutres pertinentes pour la classification. Cette fonction est basée sur la distance parcourue par chaque articulation et son accélération durant un mouvement. Nous supposons qu'un mouvement neutre correspond à une dépense d'énergie minimale lors d'un mouvement. En effet, une personne cherchant à exécuter un mouvement cherchera à économiser son énergie. En posant $D_s(j)$ (respectivement $D_o(j)$) la distance parcourue par l'articulation j lors du mouvement neutre synthétisé (respectivement le mouvement original). De même, $\mathbf{A_s}(j)$ (respectivement $\mathbf{A_o}(j)$) l'accélération de l'articulation j durant le mouvement neutre synthétisé (respectivement le mouvement original). La fonction de coût est définie comme la somme des différences entre la distance et l'accélération originale pour chaque articulation et celle synthétisée. La minimisation de cette fonction de coût fournit un mouvement neutre grossier utilisé dans la section 12 pour calculer

Algorithm 1 : Algorithme pour la synthèse de mouvement neutre

Input : $M_i i$: Mouvement original
Data : joints : tableau des articulations
trajectories : tableau des trajectoires de chaque articulation
trajectoriesSmooth : tableau des trajectoires filtrées pour chaque articulation
Result : M_n : Mouvement Neutre
Parameters : samplingValue : paramètre temporel afin de déterminer les postures clés pour la phase de cinématique inverse temporelle
weightTargets : tableau de poids permettant d'équilibrer la posture finale obtenu par la cinématique inverse (1 = sur la cible et 0 = squelette au repos).
weightHints : tableau de poids permettant d'équilibrer les indices de la posture finale obtenu par la cinématique inverse (1 = sur la cible et 0 = squelette au repos).

```
7 foreach joint, j_i, in joints do
    trajectories_i ← computeTrajectory(j_i) ;
    trajectoriesSmooth_i ← smoothTrajectory(trajectories_i) ;
end
foreach EndEffector, end_i, in joints do
    indiceHint ← end_i.getParent() ;
    for i =0; i < endTime; i += samplingValue do
        target_{end_i} ← trajectoriesSmooth_i ;
        hint_{end_i} ← trajectoriesSmooth_{indiceHint} ;
        IK_Step( target_{end_i}, hint_{end_i}, weightTargets, weightHints) ;
    end
end
```

le résidu obtenu entre le mouvement original et le mouvement synthétisé. La formulation de la fonction de coût décrivant un mouvement neutre est donné dans l'équation 3.1.

$$\begin{aligned} Cout = & \sum_{j \in \theta} |(1-\lambda) Distance(s,o) + \lambda Acceleration(s,o)|^2 \\ & \text{avec Distance(s,o)} = D_s(j) - D_o(j) \\ & \text{et Acceleration(s,o)} = (\mathbf{A_s}(j) - \mathbf{A_o}(j)) \end{aligned} \tag{3.1}$$

avec j représente une articulation, θ est l'ensemble des articulations du squelette et $\lambda \in [0,1]$ est un paramètre de poids. Les distances $D_s(j)$ et $D_o(j)$ parcourues par une articulation j lors d'un mouvement sont données par la longueur des B-spline. Les accélérations $\mathbf{A_s}(j)$ et $\mathbf{A_o}(j)$ fournissent de l'information sur l'énergie dépensée par une articulation j lors d'un mouvement. Afin de trouver ces accélérations, nous calculons la dérivée seconde de chaque B-spline. La fonction de coût est minimi-

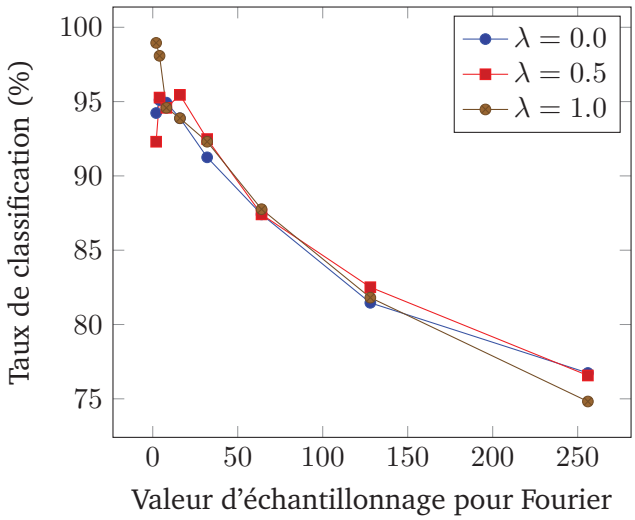


Figure 3.7 : Comparaison de taux de classification avec différentes valeurs de λ et de ré-échantillonnage pour la base SIGGRAPH[146].

sée itérativement en utilisant l'algorithme d'optimisation par essaim de particules (PSO) [39]. PSO possède l'avantage de n'imposer que peu de contraintes voire aucune sur la fonction à optimiser. De plus, PSO n'utilise pas le gradient du problème à optimiser, ce qui signifie que PSO n'exige pas que le problème d'optimisation soit différentiable comme l'exigent les méthodes d'optimisation classiques telles que la descente du gradient et les méthodes quasi Newton. Cependant, PSO ne garantit pas de trouver la solution optimale au problème fournit en entrée. Dans notre cas, cela ne nous pose pas de problème car nous cherchons à trouver un mouvement neutre grossier vu que l'on se trouve dans un cas de reconnaissance d'expressions où l'animation n'a pas à être montrée ni réutilisée.

Résidu entre le mouvement neutre et expressive

À ce stade, nous avons le mouvement original expressif et le mouvement neutre obtenu en minimisant la fonction de coût. Nous calculons la représentation fréquentielle via la transformée de Fourier de ces deux mouvements. Dans la représentation fréquentielle, la magnitude contient principalement l'information de mouvement ainsi que l'expression pour une animation donnée [148]. En calculant la différence de magnitude pour chaque degré de liberté de chaque articulation, nous obtenons le résidu entre l'animation neutre et l'animation expressive. Nous utilisons la transformée de Fourier car les résultats obtenus par Yumer et Mitra [148] dans le domaine de la synthèse d'animations montrent qu'il s'agit d'un espace de représentation prometteur afin d'extraire l'expressivité d'un mouvement. De manière formelle, ce résidu est décrit dans l'équation 3.2 où $M_s(\omega, j, l)$ (respective-

ment $M_o(\omega, j, l)$) est la magnitude fréquentielle de l'articulation j et pour le degré de liberté l à la fréquence ω pour l'action neutre synthétisée (respectivement pour le mouvement original).

$$\begin{aligned} \textbf{Résidu} =& (|M_o(\omega, j, l) - M_s(\omega, j, l)|) \\ & j \in \theta \\ & l \in DOF \\ & \omega \in 1..N \quad \text{avec N la longueur de notre signal} \end{aligned} \tag{3.2}$$

Le résidu forme le vecteur de descripteurs utilisé en entrée pour la classification afin de détecter l'expression corporelle. Dans notre approche, nous avons un paramètre à définir concernant le nombre d'échantillon de notre signal d'entrée. La figure 3.7 présente la variation du taux de classification en fonction de différentes valeurs de ré-échantillonnage. Le taux de classification diminue lorsqu'on augmente le nombre d'échantillons. Cette observation est liée au fait que lorsqu'on augmente la taille de notre signal d'entrée, les valeurs du résidu obtenues sont très proches de zéro dans les basses fréquences créant du bruit pour notre classifieur.

3.2.5 Résultats

Nous avons évalué notre méthode sur quatre bases de données présentées dans la Section 2.6. Trois de ces bases ont été réalisées par capture de mouvements d'acteurs jouant diverses actions comme marcher, frapper, porter, lancer, etc. La dernière base, nommée SIGGRAPH, est utilisée en synthèse d'animation.

La figure 3.8 montre l'influence de la taille de l'ensemble d'apprentissage sur la performance des trois classifieurs utilisés dans notre méthode. Nous avons comparé la performance de notre méthode en utilisant différents algorithmes de classifications : SVM avec un noyau χ^2, Random Forest avec 100 arbres et 2-Nearest neighbor basé sur la distance Euclidienne. La figure 3.8 a été produite sur la base SIGGRAPH. Pour chaque classifieur, nous avons calculé le taux de classification en utilisant différentes valeurs d'échantillons par la méthode de validation croisée *k*-fold. La performance de notre système a été évaluée en utilisant l'algorithme Random Forest avec 10 échantillons et 100 arbres qui fournit le meilleur taux de classification. Le fait que de bons résultats soient obtenus avec cet algorithme montre que notre espace de descripteurs est bien discriminant.

Le tableau 3.5 compare notre méthode avec les méthodes de l'état de l'art sur les mêmes bases de données. Il démontre que notre approche dépasse les autres

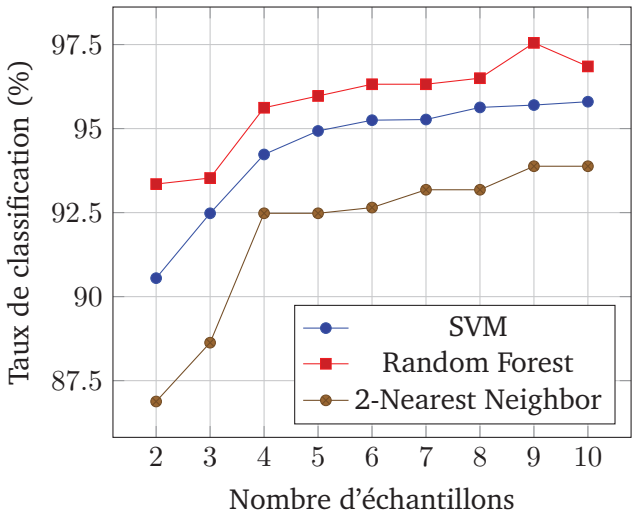


Figure 3.8 : Évolution du taux de classification pour la base SIGGRAPH en faisant varier le nombre d'échantillons pour la validation croisée par *k*-fold

BDD	Résultats état de l'art	Nos résultats
UCLIC	78% [141]	**83%**
Biological	50% à 80% [13]	**57%**
MPI	–	**50%**
SIGGRAPH	93% [25]	**98%**

Table 3.5 : Présentation des bases de données utilisées ainsi que la comparaison de notre méthode par rapport aux méthodes de l'état de l'art.

approches de l'état de l'art en terme de taux de reconnaissance d'expressions corporelles. De plus, notre méthode est générique alors que l'état de l'art comporte des méthodes spécifiques à certains gestes. Notre taux de reconnaissance d'expressions corporelles est meilleur que l'état de l'art pour les bases SIGGRAPH et UCLIC. Sur la base Biological Motion, la méthode [13] suppose que tous les gestes sont du même type pour calculer un mouvement moyen, ici tous les mouvements répondent à l'action de frapper à une porte. Cette hypothèse empêche de généraliser leur approche à des gestes dont le type n'est pas connu à l'avance, contrairement à notre approche. Nous pensons donc que pour comparer notre approche à la leur il est cohérent d'utiliser leur taux de reconnaissance générique de 50% alors que la notre est de 57%. Qui plus est, dans la littérature, aucune méthode de reconnaissance d'expressions corporelles n'a utilisé cette base MPI. Notre méthode obtient un taux de reconnaissance de 50%, ce qui est fort louable avec cette base extrêmement difficile car elle contient de nombreuses expressions avec un nombre d'exemples très variable selon l'expression. En utilisant un filtre de ré-échantillonnage classique afin de gérer des bases de données déséquilibrées, nous obtenons même un taux de classification de 67%.

Dans cette section, nous avons présenté une approche pour la reconnaissance automatique d'expressions corporelles à partir d'un squelette 3D obtenu par capture de mouvements. Nous soutenons l'idée que l'expression corporelle peut être détectée de manière robuste en analysant la différence obtenue entre un mouvement neutre et un mouvement expressif. Nous avons proposé un algorithme qui permet de synthétiser un mouvement neutre à partir d'un mouvement expressif. A partir du mouvement neutre synthétisé, notre méthode classifie l'expression du mouvement original en calculant la différence dans le domaine fréquentiel entre l'animation neutre synthétisée et le mouvement expressif. Les résultats obtenus sur différentes bases de données montrent que cette approche est très prometteuse. En effet, les résultats obtenus dépassent notre première approche qui consistait à l'extraction de différents descripteurs. Ces résultats confirment l'idée que les approches basées descripteurs sont limitées lorsque l'on cherche à détecter l'expression corporelle sur des mouvements hétérogènes voire inconnus. Cependant, le réalisme de l'animation neutre produite est encore problématique. En effet, le mouvement neutre synthétisé est assez robotique et comporte différents artefacts : le mouvement global n'est pas très fluide, les pieds glissent et pour certaines expressions nous avons du mal à neutraliser l'expression. Nous pensons que ces différents problèmes peuvent nuire à la reconnaissance d'expressions corporelles, c'est pourquoi nous avons cherché à proposer une seconde formalisation d'un mouvement neutre.

3.3 Seconde formalisation du mouvement neutre

La méthode précédemment présentée pour la reconnaissance d'expressions corporelles, basée sur l'extraction de l'expression par un mouvement neutre, nous a permis d'obtenir de bons résultats de classification malgré quelques défauts. Notre objectif ici est d'améliorer le mouvement neutre afin d'augmenter le taux de reconnaissance d'expressions corporelles.

3.3.1 Présentation de la méthode

Le but de cette méthode va être de proposer une nouvelle formalisation du mouvement neutre afin d'obtenir un mouvement plus humainement réaliste. Pour cela, nous allons combiner les connaissances de notre approche basée descripteurs et les connaissances de synthèse obtenus lors de la première méthode. Le schéma global de cette méthode est donné dans la Figure 3.9. Contrairement à notre première méthode dans laquelle nous nous étions focalisés sur la fonction de coût afin de

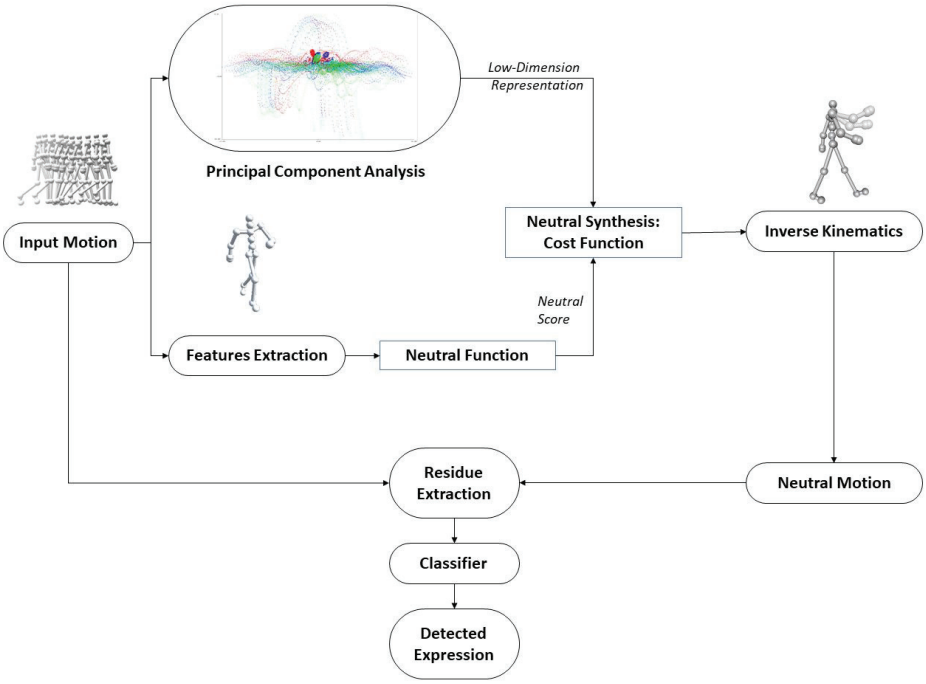


Figure 3.9 : Schéma global de notre méthode. A partir du mouvement expressif, nous synthétisons un mouvement neutre grâce à une fonction de neutralité qui donne un score de neutralité pour un mouvement donné. A partir du mouvement neutre synthétisé, nous extrayons le résidu formé entre le mouvement neutre produit et le mouvement original. Ce résidu est ensuite donné à un classifieur afin de reconnaître l'expression corporelle du mouvement en entrée.

synthétiser un mouvement neutre par optimisation, ici nous allons chercher à caractériser un mouvement neutre par un score de neutralité à travers une fonction de neutralité. Le but de cette fonction va être de nous fournir un score de neutralité sur un mouvement. Plus ce score sera bas, plus le mouvement sera neutre. Grâce à cette fonction, notre objectif va être de proposer une nouvelle fonction permettant la synthèse d'un mouvement neutre. Cette nouvelle fonction de coût est basée sur trois termes. Le premier terme utilise le score de neutralité d'un mouvement. Un terme de donnée est utilisé afin de ne pas s'éloigner du mouvement original. Le dernier terme permet d'ajouter des contraintes sur le mouvement synthétisé afin de respecter un mouvement humainement réaliste.

3.3.2 Synthèse du mouvement neutre : fonction de coût

Dans cette partie, nous allons présenter la nouvelle fonction de coût utilisée pour la synthèse d'un mouvement neutre. Cette fonction est décrite dans l'équation 3.3.

$$Cost(Motion) = \lambda Neutral(Motion) + \gamma Data(Motion) + \beta Penalty(Motion) \quad (3.3)$$

Le premier terme est le terme de neutralité, $Neutral$. Ce terme tel que nous l'avons défini est décrit en détail dans la Section 3.3.2. Il permet d'évaluer la neutralité d'un mouvement donné. Le second terme, $Data$, évalue la distance entre le mouvement neutre synthétisé et le mouvement expressif fournit en entrée du système. Ce terme est utilisé afin que le mouvement neutre réalise la même action que le mouvement original. Le dernier terme est un terme de régularisation, $Penalty$, qui ajoute une pénalité si le mouvement neutre synthétisé ne respecte pas les contraintes : problème de fluidité et les pieds qui glissent. L'espace de représentation des paramètres de cette fonction de coût est construit par deux étapes d'analyse en composante principale (ACP). Cette espace de représentation est décrit dans la Section 3.3.2. Nous utilisons cet espace car la dimension originale d'un mouvement est bien trop grande et complexe. Nous avons choisi d'utiliser une analyse en composante principale afin de résoudre ce problème ce qui nous permet d'avoir un espace compact et précis. Le choix de l'analyse en composante principale est décrit dans la Section 3.3.2.

Afin d'optimiser cette fonction de coût, nous utilisons la méthode de la stratégie d'évolution "CMA-ES" qui est un acronyme pour l'anglais *Covariance Matrix Adaptation Evolution Strategies*. Cette méthode d'optimisation a pour avantage d'être une méthode ne nécessitant pas de dérivées pour des problèmes d'optimisation non linéaire ou non convexe ce qui est notre cas. Il est probable que d'autres algorithmes d'optimisation aurait pu résoudre notre problème. CMA-ES est utilisé classiquement dans le domaine de l'optimisation de mouvement physique avec de très bons résultats. Nous avons fait le choix de ne plus utiliser PSO pour ce problème car l'espace de recherche des solutions est bien plus grand que notre ancien problème, voir Section 12 ce qui nécessitait un réglage de paramètres bien plus précis afin que PSO fonctionne de manière optimale contrairement à CMA-ES. Dans notre cas, nous cherchons à reconstruire un mouvement en entier, bien que représenté dans un espace réduit grâce à deux ACPs imbriqués, tandis que dans notre ancienne méthode nous cherchions des paramètres afin d'ajuster le mouvement fournit en entrée.

Représentation d'un mouvement dans un espace à dimension réduite

Dans cette section, nous présentons l'espace de paramètres utilisé pour la fonction de coût. Le but de ce changement d'espace est de pouvoir proposer un mouvement dans une représentation compacte mais toute fois précise. En effet, l'espace original d'un mouvement est bien trop complexe et à haute dimension pour pouvoir être utilisé directement dans le processus d'optimisation. Afin de résoudre ce problème, nous proposons d'utiliser deux analyses en composantes principales imbriquées : l'une travaillant sur les poses et l'autre travaillant sur l'aspect temporel. L'analyse en composante principale (ACP) cherche à réduire en dimension un espace à très haute dimension en un ensemble de vecteurs pertinents. La combinaison linéaire de ces vecteurs permet de retrouver l'espace initial. Il s'agit d'obtenir le résumé le plus pertinent de nos données originales. Pour cela, la matrice des variances-covariances va nous permettre de réaliser cette réduction de dimension en analysant la dispersion des données initiales. De cette matrice, on va extraire les facteurs que l'on recherche en déformant le moins possible la configuration globale des données originales. L'analyse en composantes principales est classique en analyse des données. Le livre de Jolliffe [72] présente tous les aspects et utilisations de façon exhaustive.

Dans nos travaux, nous allons utiliser deux phases d'analyses en composantes principales afin d'aboutir à notre espace de représentation d'un mouvement. Pour cela, nous allons calculer une première analyse en composantes principales sur toutes les postures de nos différentes bases de données sans se soucier du temps. A partir de cette analyse, nous allons pouvoir construire une seconde analyse en composantes principales sur un mouvement entier cette fois-ci. La figure 3.10 présente le pipeline de ce processus.

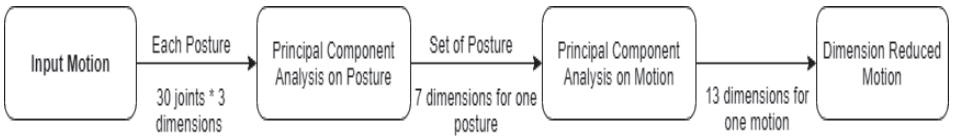


Figure 3.10 : Processus pour la réduction de dimension d'un mouvement.

L'espace original de nos mouvements est représenté par une série de posture. Le nombre de postures dépend du nombre de poses présentes dans le mouvement original que l'on va appeler n. Dans notre représentation, un squelette est composé de 30 articulations. Ainsi, une posture va être composé de 90 composants si l'on utilise la positon 3D de chacune des articulations ou de 120 composants lorsque l'on utilise la rotation, représentée en quaternion, de chacune des articulations. Dans ce travail, nous avons choisi d'utiliser la position 3D de chacune des articulations.

En appliquant l'ACP sur toutes les postures de chacune des bases de données utilisées, nous pouvons représenter une posture avec uniquement 7 composantes en conservant 95% de la variance originale. Nous avons choisi de conserver 95% de la variance originale car cela nous permet de perdre très peu d'informations tout en ayant un espace de représentation compact. A partir de la représentation réduite d'une posture, nous pouvons représenter un mouvement complet sous 7 composants * n le nombre de poses. Cependant, cette représentation n'est pas assez compacte pour l'utilisation souhaitée, c'est à dire l'optimisation de la fonction de coût afin de créer un mouvement complet. En effet, les mouvements que nous utilisons sont représentés par 60 poses par seconde et peuvent durer une douzaine de secondes au maximum ce qui représente 5040 composants. Afin de réussir à réduire le nombre de composants d'un mouvement, nous allons appliquer une seconde étape d'ACP sur les mouvements complets de toutes nos bases de données. Pour cela, nous appliquons le même processus utilisé pour les postures. Grâce à cette seconde ACP, nous arrivons à représenter un mouvement complet en 13 composants en conservant 95% de la variance originale. Ainsi, lorsque nous optimisons notre fonction de coût afin de synthétiser un mouvement neutre nous allons travailler dans notre espace de représentation obtenu par les deux ACPs. Nous allons donc avoir à faire deux projections inverses pour revenir dans notre espace original de mouvement. La figure 3.11 montre la projection de toutes les postures de la base de données SIGGRAPH dans l'espace obtenu par la première ACP. Afin d'illustrer l'ACP en deux dimensions, nous avons gardé les deux composantes principales de notre ACP. Malgré cette projection 2D, nous retrouvons les différents types de mouvements présents dans notre base de données. Ce qui est intéressant d'observer est que les différents types de mouvements sont organisés par clusters, ce qui va être utile pour la synthèse de notre mouvement neutre. En effet, l'une des problématiques pour la synthèse de notre mouvement neutre est tout d'abord de devoir retrouver le mouvement exécuté en entrée. Cette représentation des données va nous aider a réaliser cette tâche, ce qui fait que l'on peut se focaliser sur la neutralisation d'un mouvement expressif.

Formalisation d'un mouvement neutre

Afin de produire un mouvement neutre à partir d'un mouvement expressif, nous représentons le mouvement original dans l'espace réduit présenté dans la Section précédente. Par la suite, la phase d'optimisation va chercher à modifier les différents paramètres afin de minimiser la fonction de coût, permettant la synthèse du mouvement neutre, présentée dans la Section 3.3.2. Dans cette section, nous allons présenter le terme "Neutral" de cette fonction de coût. En effet, afin de pouvoir juger la neutralité d'un mouvement, nous avons formalisé différents descripteurs

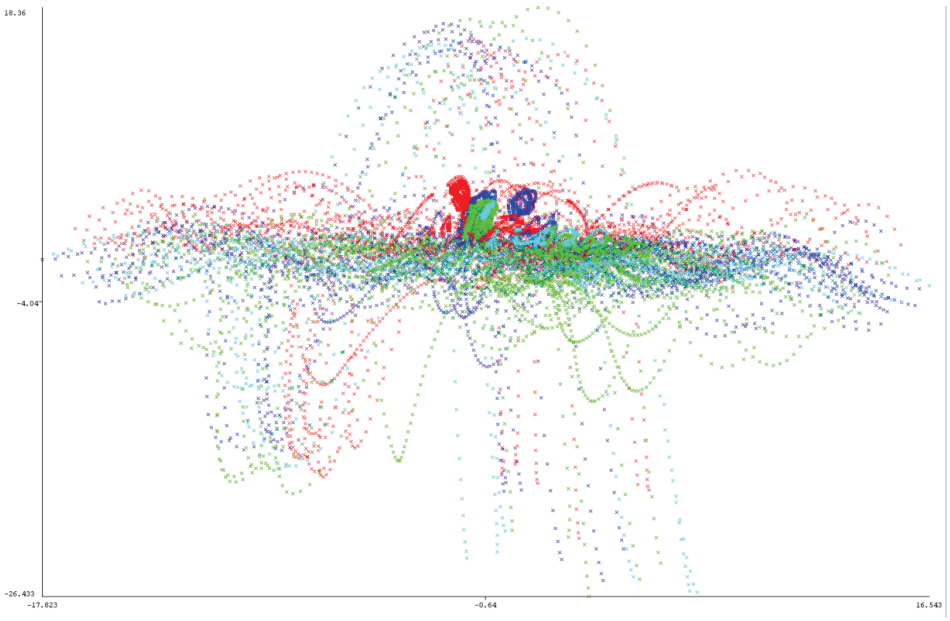


Figure 3.11 : Projection sur deux axes de l'ACP calculée sur toutes les postures de la base de données SIGGRAPH. Chaque courbe est un mouvement.

qualitatifs provenant du domaine de la psychologie [50]. Ces descripteurs décrivent qualitativement l'expression d'un mouvement. Nous les utilisons pour quantifier la neutralité d'un mouvement. Nous avons séparés les différents descripteurs en deux parties : les descripteurs de posture et les descripteurs temporels.

Dans ce paragraphe, nous expliquons comment nous avons construit la fonction permettant la caractérisation de la neutralité d'un mouvement basé sur la formalisation de différents descripteurs présentés ci-dessous. La fonction de neutralité est définie comme une moyenne pondérée et est décrite dans l'équation 3.4 où α_i correspond au poids du descripteur courant, f_i. n est défini comme le nombre de descripteurs.

$$Neutral(Motion) = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i} \tag{3.4}$$

Afin de trouver les valeurs correctes pour chacun des poids présents dans la fonction de neutralité, nous avons utilisé les différentes bases de données décrites dans la Section 2.6. Nous avons optimisé la fonction afin d'obtenir un score tendant vers zéro lorsque que nous avons un mouvement neutre. Grâce à l'optimisation de cette fonction sur les différentes bases de données, nous avons trouvé les valeurs pertinentes pour chacun des poids. Nous allons maintenant présenter les différents descripteurs utilisés dans la fonction de neutralité.

Les decripteurs de posture sont les suivants.

— L'ouverture du corps $f_{bodyOpenness}$ calcule la distance entre les deux bras, la distance entre la main droite et le bassin et la distance entre la main gauche et le bassin. De même, ce descripteur calcule la distance entre les pieds.

— L'inclinaison sagittale du corps $f_{bodyLeaning}$ mesure l'angle entre le vecteur bassin / cou et l'axe up de notre monde. Ce descripteur donne une indication sur l'inclinaison du corps.

— La droiture du corps $f_{bodyStraightness}$ calcule l'angle de fléchissement de la tête, des genoux et du tronc.

Les decripteurs temporels sont les suivants.

— La puissance du mouvement $f_{mvtPower}$ représente la quantité de force déployée lors du mouvement.

— La fluidité du mouvement $f_{mvtFluidity}$ représente la variation ou non du mouvement global.

— La vitesse du mouvement $f_{mvtSpeed}$ calcule la vitesse à laquelle le mouvement est réalisé.

— La quantité de mouvements des bras $f_{quantityArmsMvt}$ mesure si les bras bougent ou non.

— La régularité du mouvement des bras $f_{regularityArmsMvt}$ cherche à détecter des patterns dans le mouvement des bras.

Descripteurs temporels Afin de calculer différents descripteurs temporels, nous avons besoin d'utiliser la transformée de Fourier. Pour cela, nous utilisons la transformée de Fourier rapide (FFT en anglais) qui est un formalisme classique afin d'extraire différents composants depuis une séquence temporelle. Dans ce paragraphe, nous détaillons comment nous l'utilisons afin d'extraire nos descripteurs. Tout d'abord, nous rappelons la formulation de la transformée de Fourier discrète (DFT). Soit x_n un signal temporel discret d'un des degrés de liberté d'une articulation de notre squelette. La transformée de Fourier discret x_k de x_n est donnée par l'équation suivante :

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{3.5}$$

avec N étant la longueur du signal et $i^2 = -1$. Le spectre unilatéral X_ω est donné par :

$$X_\omega = \frac{2}{N} X_k \quad k = 0, ..., N/2 \tag{3.6}$$

avec $\omega = (x_s/N)k$ est la fréquence transformée à partir des échantillons k dans l'espace spectral. Nous utilisons seulement le spectre unilatéral dans la gamme de fréquences positives ($\omega = 0 : x_s/2$). Ici, $x_n s$ est la fréquence d'échantillonage du signal temporel original x_n. De cette représentation spectrale, nous pouvons extraire l'amplitude et la phase du spectre. La magnitude définit l'existence et l'intensité d'un mouvement, tandis que la phase décrit le moment relatif. La transformée de Fourier rapide est calculée sur le mouvement des articulations, c'est-à-dire sur le signal de rotation de chaque articulation. Notre signal de rotation est représenté par des quaternions, nous calculons une transformée de Fourier par terme de nos quaternions.

Nous allons maintenant présenter la formalisation des différents descripteurs temporels que nous proposons. Tout d'abord, nous définissons la **puissance du mouvement** comme la somme du spectre de l'amplitude

$$f_{mvtPower} = \sum_{i=1}^{\theta} \sum_{j=1}^{N} R[\omega_{i,j}] \tag{3.7}$$

où θ est le nombre d'articulations présentes dans notre squelette et $R[\omega_{i,j}]$ est l'amplitude de l'articulation i à la fréquence j.

La transformée de Fourier nous donne une information sur la **régularité du mouvement des bras**. Nous l'avons défini comme étant la somme des différences entre l'amplitude du spectre de chacune des articulations des deux bras (épaules, coudes et mains).

$$f_{regularityArmsMvt} = \sum_{i=1}^{\theta_{arms}} \sum_{j=1}^{N} R[\omega_{i_{left},j}] - R[\omega_{i_{right},j}] \tag{3.8}$$

Notre troisième descripteur temporel est la **vitesse du mouvement**. Pour cela, nous calculons la moyenne de la vitesse de chacun des articulations du squelette. La vitesse d'une articulation est simplement définie par la différence entre la position d'une articulation à un temps t et au temps $t-1$.

$$f_{mvtSpeed} = \sum_{i=1}^{\theta} \sum_{t=1}^{N} \frac{P_i(t) - P_i(t-1)}{(t) - (t-1)} \tag{3.9}$$

où $P_i(t)$ est la position dans le monde de l'articulation i au temps t. Nous avons défini la **quantité de mouvement des bras** comme la distance parcourure cumuluée couverte par chacunes des articulations du bras lors du mouvement.

$$f_{quantityArmsMvt} = \sum_{i=1}^{\theta_{arms}} \sum_{t=0}^{N} P_i(t) \tag{3.10}$$

Notre dernier descripteur temporel est la **fluidité du mouvement**. Nous calculons la moyenne de l'accélération pour les deux mains et pieds. Ce terme nous donne une indication sur la fluidité du mouvement en entrée. En effet, la dérivée seconde nous permet de mesurer les taux de variations de l'articulation cible.

$$f_{mvtFluidity} = \sum_{i=1}^{\theta_{endEffectors}} \frac{\sum_{t=0}^{N} d^2 P_i(t)/dt^2}{N} \tag{3.11}$$

3.3.3 Extraction de l'expression corporelle via le résidu

Dans la Section 3.3.1 nous avons présenté comment obtenir un mouvement neutre correspondant au mouvement expressif en entrée. Nous cherchons, maintenant, à extraire l'expression corporelle du mouvement original. Pour cela, nous utilisons la même idée développée dans notre précédente méthode utilisant un mouvement neutre présenté dans la Section 3.2. Cette idée suppose que l'expression corporelle est présente dans le résidu formé par la soustraction du mouvement expressive au mouvement original. En effet, cette soustraction permet de séparer le mouvement de l'expression lorsque que le mouvement réalisé est le même. Afin de calculer ce résidu, nous défendons l'idée que la représentation fréquentielle convient très bien à ce calcul. Cette hypothèse est renforcée par notre ancienne méthode [26] et le travail de Yumer et Mitra [148] qui sont parvenus à transférer des expressions corporelles en se basant sur ce formalisme.

$$\begin{aligned} \textbf{Résidu} =& (|M_o(\omega, j, l) - M_s(\omega, j, l)|) \\ & j \in \theta \\ & l \in DOF \\ & \omega \in 1..N \quad \text{avec N la longueur de notre signal} \end{aligned} \tag{3.12}$$

La première étape consiste donc à calculer la représentation fréquentielle du mouvement original et du mouvement neutre synthétisé. Le résidu entre le mouvement neutre et le mouvement expressif est, ensuite, calculé pour chaque degrés de liberté de chaque articulations du squelette. Ce calcul est fait de manière indépendante entre chaque articulation et consiste en une soustraction entre l'amplitude du mouvement neutre et l'amplitude du mouvement expressif pour chaque signal de rotation des différentes articulations. De manière formelle, le résidu est décrit dans l'équation 3.12 où $M_s(\omega, j, l)$ (respectivement $M_o(\omega, j, l)$) est l'amplitude du spectre pour l'articulation j et pour le degré de liberté l à la fréquence ω lors du

mouvement neutre synthétisé par notre méthode (respectivement lors du mouvement original). La magnitude de notre spectre contient principalement l'information du mouvement qui est mélangée avec l'information de l'expression corporelle. En calculant ce résidu, nous arrivons à enlever l'information de mouvement dans l'amplitude de notre spectre et nous pouvons ainsi récupérer l'expression corporelle. L'avantage de cette méthode est donc de pouvoir séparer le mouvement réalisé de l'expression perçue. Le résidu forme le vecteur de descripteurs qui est utilisé en donnée d'entrées à un classifieur afin de pouvoir détecter l'expression corporelle d'un mouvement.

3.3.4 Résultats

La première étape avant de pouvoir exécuter notre méthode sur les différentes bases de données consiste à calculer les valeurs des poids pour chacun des descripteurs contenus dans le score de neutralité. Pour cela, nous avons utilisé les différentes bases de données contenants des mouvements neutres. En effet, nous avons optimisé les poids de façon à obtenir un score de neutralité faible lorsque nous cherchons à caractériser un mouvement neutre et un score pénalisant sur des mouvements expressifs. Une fois cette optimisation réalisée, nous avons pu fixer les valeurs des poids afin d'obtenir des scores de neutralité proches de la réalité terrain. Le tableau 3.6 présente le taux de reconnaissance obtenue par notre méthode sur différentes bases de données. La Figure 3.12 montre l'influence de la taille de l'ensemble d'apprentissage sur la performance de trois classifieurs. Nous avons calculé la performance de notre méthode avec un SVM (Support Vector Machine) avec un noyau χ^2, une Forêt d'arbres décisionnels (Random Forest en anglais) avec 100 arbres et une recherche naïve de plus proche voisin basée sur la distance Euclidienne. La Figure 3.12 a été produite sur la base de donnée Emilya. Pour chacun des classifiers, nous avons calculé le taux de classification en découpant nos données d'apprentissages en k sous ensembles de données en suivant une méthode de validation croisée (*k*-fold cross validation). Les résultats obtenus avec les forêts d'arbres décisionnels montrent que notre espace de descripteur est bien partionné. La performance de notre méthode a été évalué en utilisant un SVM en découpant nos données d'apprentissages en 10 sous ensembles.

Le tableau 3.13 présente la matrice de confusion des différentes expressions et mouvements de la base de données Emilya. Ce tableau montre que notre vecteur de descripteurs, basé sur le résidu, est correctement capable de discriminer différentes expressions à partir de mouvements hétérogènes. Nous pouvons voir à partir de ce tableau que l'expression neutre est la mieux détectée. Ce résultat valide notre approche consistant à synthétiser un mouvement neutre à partir d'un mouvement

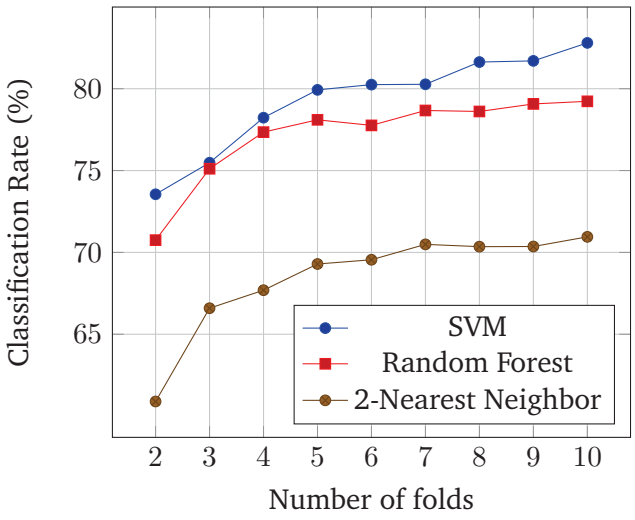


Figure 3.12 : L'évolution du taux de classification pour la base de données Emilya en faisant varier le nombre d'échantillons pour la validation croisée.

	Angry	Anxious	Joy	Neutral	Panic Fear	Proud	Sad
Angry	0.76	0.05	0.07	0.02	0.06	0.04	0
Anxious	0.04	0.79	0.04	0.02	0.07	0.04	0
Joy	0.05	0.03	0.77	0.02	0.08	0.05	0
Neutral	0.02	0.01	0.03	0.88	0.02	0.04	0
Panic Fear	0.03	0.04	0.05	0.03	0.84	0.03	0
Proud	0.03	0.03	0.07	0.04	0.03	0.80	0
Sad	0.03	0.06	0.10	0.04	0.05	0.08	0.64

Figure 3.13 : Matrice de confusion pour la base de données Emilya.

expressif. L'expression de tristesse est celle qui obtient le taux de reconnaissance le plus bas. Ce résultat s'explique par le fait que cette expression est sous représentée comparée aux autres. Le tableau 3.14 présente la même matrice de confusion lorsque l'on applique un filtre de ré-échantillonnage afin d'équilibrer la base de données. A partir de ce tableau, on peut observer que le taux de reconnaissance de l'expression tristesse est bien meilleur vu que la base de données est équilibrée. Selon ces résultats, nous pouvons valider l'efficacité de notre méthode pour la reconnaissance d'expressions corporelles à partir de mouvements hétérogènes.

Le tableau 3.6 présente la comparaison du taux de reconnaissance de notre méthode avec les méthodes de l'état de l'art sur les mêmes bases de données. En conclusion, notre méthode dépasse les méthodes les plus récentes en termes de taux de reconnaissance de l'expression corporelles. En effet, nous comparons notre méthode à des méthodes spécifiques développées pour une base de données unique

	Angry	Anxious	Joy	Neutral	Panic Fear	Proud	Sad
Angry	0.90	0.01	0.02	0.02	0.03	0.02	0
Anxious	0.02	0.90	0.03	0	0.04	0.01	0
Joy	0.02	0	0.90	0.01	0.03	0.03	0
Neutral	0.01	0.01	0.01	0.95	0.01	0.01	0
Panic Fear	0.01	0.02	0.02	0.01	0.92	0.02	0
Proud	0.02	0.02	0.04	0.02	0.01	0.89	0
Sad	0.02	0.03	0.06	0.02	0.03	0.04	0.80

Figure 3.14 : Matrice de confusion pour la base de données Emilya avec application d'un filtre de ré-échantillonnage afin d'équilibrer la base de données en terme de classes.

Table 3.6 : Comparaison de notre méthode avec les méthodes de l'état de l'art.

Base de données	**Résultats de l'état de l'art**	**Nos résultats**
Emilya	–	**82.2**%
UCLIC	**87.30**% [34]	74%
MPI	50%[26]	%
SIGGRAPH	98% [26]	**98.8**%

ne contenant souvent qu'un type de mouvement alors que notre méthode se veut générique. Notre méthode générique surpasse l'état de l'art des bases de données sur SIGGRAPH, MPI et Emilya. Pour la base de données UCLIC, nous avons comparé notre méthode à une approche spécifique [**dewan2017**]. Cette méthode utilise une formalisation plus élaborée de la théorie de Laban combinée à une fenêtre temporelle. Ils ont concentré leur méthode sur la base de données UCLIC et ont obtenu un très bon taux de reconnaissance, 87.30% par rapport à notre résultat 74%. Pour la base de données MPI, notre méthode permet d'obtenir un taux de reconnaissance de 78,6%, ce qui est mieux que l'état de l'art. Ce qui est intéressant à noter sur cette base de données est que l'état de l'art provenait de notre ancienne méthode basée sur la synthèse, non optimale, d'un mouvement neutre, c'est-à-dire robotique et contenant des artefacts. Avec ce résultat, nous confirmons l'idée que plus le mouvement neutre synthétisé est réaliste, meilleur est le taux de reconnaissance. Cette hypothèse est aussi confirmée par les résultats obtenus sur la base de données UCLIC. En effet, notre ancienne méthode avec un mouvement neutre grossier obtenait un taux de reconnaissance de 71% tandis que notre méthode actuelle obtient un taux de reconnaissance de 74%. Enfin, à notre connaissance, aucune méthode dans la littérature sur la reconnaissance des expressions corporelles n'a testé la base de données Emilya, notre méthode obtient un taux de reconnaissance de 82%. C'est un excellent résultat car la base de données Emilya contient beau-

coup de mouvements et d'expressions différentes et sa taille en fait une référence importante. En effet, la variété et le nombre de mouvements dans cette base de données constituent un défi de taille. Cela montre la généricité de notre méthode, même dans des cas complexes.

En conclusion, nous avons évalué notre méthode sur différentes bases de données publiques et nous pouvons confirmer la robustesse de notre approche en matière de reconnaissance d'expressions corporelles. En effet, par rapport à l'état de l'art, nous avons proposé une méthode générique pour la reconnaissance de l'expression corporelle sur des mouvements hétérogènes. Contrairement aux méthodes de l'état de l'art qui se sont souvent focalisées sur une base de données unique. Nous pensons qu'une méthode combinant des connaissances de vision par ordinateur et de synthèse d'animations est meilleur que les méthodes basées sur l'extraction de descripteurs afin de proposer une reconnaissance générique d'expressions corporelles. Les méthodes basées sur les descripteurs sont adaptées uniquement si nous savons sur quels types de mouvements nous travaillons. Cette hypothèse est confirmée par le résultat de la base de données UCLIC dans laquelle la méthode spécifique obtient un très bon taux de reconnaissance par rapport à notre méthode.

4 Reconnaissance d'expressions faciales de visage d'enfants

Contents

4.1 Base de données de visages d'enfants

Dans cette section nous allons présenter la base de données de visages d'enfants que nous avons construite et partagée à la communauté. Le but de cette base de données est d'aider les différents chercheurs sur le comportement humain. Cette base de données constituera une excellente ressource pour la communauté de la vision par ordinateur, notamment pour l'analyse comparative de différentes méthodes. L'objectif de cette base de données est de capturer différentes expressions spontanées de visages d'enfants.

4.1.1 Introduction

Comme énoncé dans l'état de l'art, la plupart des humains expriment leurs émotions par le canal facial, aussi connu sous le nom d'expressions faciales[88]. Les humains ont la capacité de reconnaître l'expression faciale en temps réel, mais pour les machines, c'est toujours une tâche difficile. En effet, la variabilité de la pose, de l'illumination et de la façon dont les gens montrent les expressions entre les différentes cultures sont quelques-un des paramètres rendant cette tâche extrêmement difficile [80].

Un autre problème qui entrave le développement d'un tel système pour les applications du monde réel est le manque de bases de données avec des expressions spontanées [134]. Il existe un certain nombre de bases de données d'expressions faciales réalisés par différents acteurs jouant les six expressions de bases [45], c'est-à-dire le bonheur, la colère, le dégoût, la peur, la surprise et la tristesse mais il n'y a très peu d'équivalent pour les expressions spontanées / naturelles. De plus, il a été prouvé que les expressions faciales spontanées diffèrent considérablement des expressions jouées par des acteurs [10].

Enfin, le dernier problème avec la plupart des bases de données accessibles au public est l'absence d'enfants dans les vidéos ou dans les images. La communauté de la vision par ordinateur a déployé beaucoup d'efforts pour créer des bases de données de vidéos ou d'images, mais presque toutes contiennent des visages d'adultes [97, 112]. En excluant les stimuli des enfants dans les bases de données accessibles au public, la communauté de la vision par ordinateur s'est non seulement limitée à des applications destinées uniquement aux adultes, mais a également produit une étude limitée pour l'interprétation des expressions [95].

4.1.2 Présentation des bases de données existantes

A notre connaissance, il n'existe que trois bases de données publiques qui contiennent des stimuli émotionnels de visages d'enfants. Ces bases de données sont les suivantes :

1. la base de données NIMH Child Emotional Faces Picture Set (NIMH-ChEFS). [40]
2. la base de données Dartmouth Database of Children's Faces. [28]
3. la base de données Child Affective Facial Expression (CAFE). [95]

Figure 4.1 : Les six expressions universelles : la première ligne présente des exemples d'images de la base de donnée à partir de la base de données Dartmouth [28]. La seconde ligne montre des images de la base de données Child Affective Facial Expression (CAFE) [95], tandis que la dernière ligne présente des images des vidéos le **notre base de données, LIRIS-CSE**. La première colonne correspond à l'expression "bonheur", la seconde colonne correspond à l'expression "surprise", la troisième colonne correspond à l'expression "tristesse", la quatrième colonne correspond à l'expression "colère", la cinquième colonne correspond à l'expression "peur" et la dernière colonne correspond à l'expression "dégoût". Les expressions proposées dans notre base de données sont spontanées et naturelles et peuvent être facilement différenciées des expressions actées / exagérées que l'on retrouvent dans les deux autres bases de données

La base de données NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) [40] contient 482 images émotionnelles contenant des expression de "peur", "colère", "heureux" et "triste" avec deux conditions : regard direct et regard détourné. Les enfants qui ont posé pour cette base de données étaient âgés de 10 à 17 ans. La base de données est validée par 20 évaluateurs adultes.

La base de données Dartmouth Database of children Faces [28] contient des images émotionnelles (des six émotions de base) de 40 garçons et de 40 filles blancs âgés entre 6 et 16 ans. Toutes les images de la base de données ont été évaluées par au moins 20 évaluateurs humains pour l'identification et l'intensité de l'expression faciale. L'expression du bonheur a été identifiée avec le plus de précision, tandis que la peur a été identifiée avec le moins de précisions par les évaluateurs humains. Ces derniers ont correctement classé 94.3% des visages heureux tandis que l'expression de la peur a été correctement identifiée dans 49.08% des images. En moyenne, les évaluateurs humains ont correctement identifié l'expression dans 79.7% des images. Référez-vous à la Figure 4.1 pour des exemples d'images de la base de données.

La base de données CAFE (Child Affective Facial Expression) [95] est composée de 1192 images émotionnelles (six émotions de base et le neutre) d'enfants âgés de 2 à 8 ans. Les enfants qui ont posés pour cette base de données étaient d'origines ethniques et raciales diverses. Référez-vous à la Figure 4.1 pour des exemples d'images de la base de données.

Les bases de données d'expressions des visages d'enfants sont diverses en termes de pose, d'angles de caméra et d'éclairage mais présentent les inconvénients suivants :

1. Les bases de données mentionnées ci-dessus contiennent des expressions actées et comme mentionné précédemment, les expressions faciales spontanées ou naturelles diffèrent des expressions actées car elles expriment des émotions réelles alors que les expressions actées sont souvent fausses et dissimulent des sentiments intérieurs [10, 51].
2. Toutes ces bases de données ne contiennent que des images statiques / photos d'identité avec une expression à intensité maximale. Selon une étude menée par le psychologue Bassili [12], il a été conclu que le mouvement des muscles faciaux est fondamental pour la reconnaissance des expressions faciales. Il a également conclu que l'être humain peut reconnaître les expressions d'un clip vidéo de façon robuste comparé à la vision simple d'une photo fixe.

En général, pour évaluer et comparer différents algorithmes d'analyse d'expressions faciale, des bases de données normalisées sont nécessaires pour permettre une comparaison significative. En l'absence de tests comparatifs sur ces bases de données normalisées, il est difficile de trouver les forces et faiblesses relatives des

différents algorithmes de reconnaissance d'expressions faciale. Ainsi, il est très important de développer une base de données d'expressions faciale naturelles de visages d'enfants contenant des clips vidéo / images dynamiques. Cela permettra à la communauté de la vision par ordinateur de construire un système robuste pour la reconnaissance de l'expression faciale naturelle de visage d'enfants.

4.1.3 Nouveauté de la base de données proposée (LIRIS-CSE)

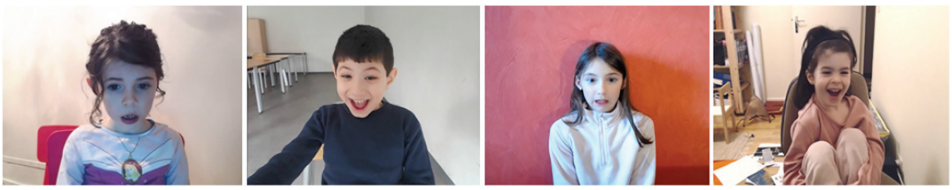


Figure 4.2 : Exemple de variations des conditions d'éclairage et du fond. Les images des colonnes 1, 3 et 4 ont été enregistrées à la maison, tandis que les images de la colonne 2 ont été enregistrées en laboratoire ou en classe.

Figure 4.3 : Exemple de transition d'expression. La première ligne montre un exemple d'expression de "Dégoût". La seconde ligne indique l'expression de la "Tristesse", tandis que la troisième ligne correspond à l'expression du "Bonheur".

Pour surmonter les inconvénients susmentionnés des bases de données sur l'expression faciale des enfants, nous présentons une nouvelle base de données émotionnelles qui contient des extraits de films et des images dynamiques de 12 enfants d'origines ethniques diverses. Cette base de données unique contient des expressions faciales spontanées / naturelles d'enfants dans divers contextes (voir Figure4.2 pour voir les variations dans des scénarios d'enregistrement) montrant

six expressions émotionnelles universelles[42, 43] (“bonheur”, “tristesse”, “colère”,“surprise”,“dégoût” et “peur”). Les enfants ont été enregistrés dans un environnement sans contrainte (pas de restriction sur le mouvement de la tête, pas de restriction sur le mouvement des mains, position assise libre, pas de restriction d’aucune sorte) pendant qu’ils regardent des stimuli spécialement conçus ou sélectionnés. Cet environnement sans contrainte nous a permis d’enregistrer l’expression spontanée / naturelle des enfants au fur et à mesure qu’ils s’expriment à travers le visionnage d’une vidéo conçue spécialement pour cette expérimentation. La base de données a été validée par 22 évaluateurs humains. Les détails des paramètres d’enregistrement sont présentés dans le tableau 4.2.

La spontanéité des expressions enregistrées peut être facilement observée dans la Figure 4.1. Les expressions dans notre base de données sont spontanées et naturelles et peuvent facilement être différenciées des expressions actées / exagérées des deux autres bases de données. La figure 4.3 montre le mouvement des muscles faciaux / transition pour différentes expressions spontanées.

Participants

Au total, 12 enfants (cinq garçon et sept filles) de diverses origines ethniques âgés de 6 à 12 ans, dont l’âge moyen est de 7.3 ans, ont participé à notre séance d’enregistrement de la base de données. 60% des enregistrements ont été effectués en classe ou en laboratoire et 40% des clips de la base de données ont été enregistrés à la maison. L’enregistrement des enfants dans deux environnements distincts nous permet d’obtenir des conditions d’arrière-plan et d’éclairage différents dans la base de données enregistrée. Reportez-vous à la figure 4.2 pour des images d’exemple avec des arrière-plans et des conditions d’éclairage différents.

Stimulateurs d’émotions

Afin d’obtenir des expressions faciales spontanées de visage d’enfants, nous avons sélectionné des clips de dessins animés / films d’animation ou de petits clips vidéo d’enfants réalisant différentes actions. Les raisons pour lesquelles nous avons choisi les vidéos afin d’induire des expressions chez les enfants sont les suivantes :

1. Toutes les vidéos sélectionnées pour induire des expressions contiennent également du son. Les stimuli vidéo avec l’audio permettent une expérience immersive, c’est donc un puissant inducteur d’expressions [93].

2. Les stimuli vidéo offrent une expérience plus engageante que les images statiques, limitant les mouvements indésirables de la tête.

3. Les stimuli vidéo peuvent évoquer des expressions sur une plus longue durée. Cela nous a aidé à enregistrer et à repérer les expressions faciales des enfants.

Le tableau 4.1 présente la liste des stimuli choisis comme inducteurs d'émotions. La durée totale des stimuli sélectionnés est de 17 minutes et 35 secondes. L'une des considérations pour ne pas choisir plus de stimuli est de prévenir la perte d'intérêt ou le désengagement des enfants avec le temps [79].

Configuration de l'enregistrement

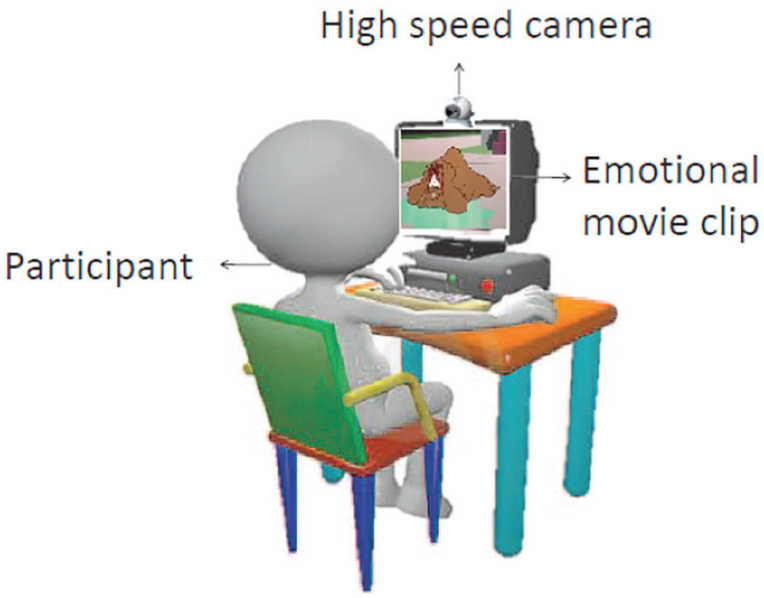


Figure 4.4 : Configuration de l'enregistrement de la base de données. Les enfants regardent les vidéos des différents stimuli tandis que la caméra enregistre les expressions faciales. Figure inspirée de [93].

Inspirés par les travaux de Li et al. [93], nous avons installé une caméra sur le dessus de l'ordinateur portable possédant une sortie audio. La caméra est placée à une distance de 50 cm du siège. Comme expliqué ci-dessus, la sortie audio a amélioré l'expérience des enfants, nous aidant ainsi à induire des expressions fortes. La configuration de l'enregistrement est illustrée dans la Figure 4.4. Comme nous l'avons mentionné à la Section 4.1.3, les enfants ont été enregistrés dans deux environnements différents, soit l'environnement de la classe ou du laboratoire, soit l'environnement familial. Les détails des paramètres d'enregistrement sont présentés dans le tableau 4.2.

Sr.No :	Expression	Source	Nom du clip	Durée
1	Dégoût	YouTube	Babies Eating Lemons for the First Time Compilation 2013	42 Sec
2	Dégoût	YouTube	On a Plane with Mr Bean (Kid puke)	50 Sec
3	Peur and surprise	YouTube	Ghoul Friend - A Mickey Mouse Cartoon - Disney Shows	50 Sec
4	Peur	YouTube	Mickey Mouse - The Mad Doctor - 1933	57 Sec
5	Peur & surprise	Film	"How To Train Your Dragon" (Monster dragon suddenly appears and kills small dragon)	121 Sec
6	Peur	Film	"How To Train Your Dragon" (Monster dragon throwing fire)	65 Sec
7	Peur	YouTube	Les Trois Petits Cochons	104 Sec
8	Joie	YouTube	Best Babies Laughing Video Compilation 2014 (three clips)	59 Sec
9	Joie	YouTube	Tom And Jerry Cartoon Trap Happy	81 Sec
10	Joie, surprise & fear	YouTube	Donald Duck- Lion Around 1950	40 Sec
11	Surprise heureuse	YouTube	Bip Bip et Coyote - Tired and feathered	44 Sec
12	Surprise heureuse	YouTube	Donald Duck - Happy Camping	53 Sec
13	Tristesse	YouTube	Fox and the Hound - Sad scene	57 Sec
14	Tristesse	YouTube	Crying Anime Crying Spree 3	14 Sec
15	Tristesse	YouTube	Bulldog and Kitten Snuggling	29 Sec
16	Surprise	Film	Ice Age- Scrat's Continental Crack-Up	32 Sec
17	Surprise & joie	Film	Ice Age (4-5) Movie CLIP - Ice Slide (2002)	111 Sec
18	Joie	YouTube	bikes funny (3)	03 Sec
19	Joie	YouTube	bikes funny	06 Sec
20	Joie	YouTube	The Pink Panther in 'Pink Blue Plate	37 Sec
			Temps total des stimuli = 17 minutes and 35 secondes	

Table 4.1 : Stimuli utilisé afin de produire des expressions spontanées.

Segmentation des vidéos

Après l'enregistrement des différentes vidéos pour chaque enfant, nous avons soigneusement examiné le clip et supprimé chacune des partie enregistrée inutile, généralement au début et à la fin de l'enregistrement vidéo. Comme la vidéo que les enfants regardaient contenait différents clips cherchant à induire différentes expressions faciales (voir la Section 4.1.3 pour le détail des stimuli visuels et auditifs),

Sujet	Enregistrement Environnement	FPS	Vidéo Résolution
S1-S7	Salle de classe	25	800 * 600
S8-S10	Maison	25	720 * 480
S11-S12	Maison	25	1920 * 1080

Table 4.2 : Présentation des différents paramètres des vidéos de la base de données

Figure 4.5 : **Expression mélangée**. Exemples d'images qui montrent l'existence de plus d'une expressions dans un même clip de la base de données. La première ligne présente les images d'un clip qui montre l'occurrence de la "tristesse" et de la "colère" . De même, la seconde ligne montre l'existence de l'expression "Surprise" et du "bonheur".

notre vidéo enregistrée contenait donc tout le spectre des différentes expressions. C'est pourquoi nous avons ensuite manuellement segmenté différents clips vidéo contenant une seule expression faciale. Référez-vous à la Figure 4.3 pour voir les résultats après le processus de segmentation. On peut observer à partir de la figure ci-dessus que chaque clip vidéo contient une expression neutre au début, puis montre le début d'une expression, et se termine lorsque l'expression est visible à son maximum. Le nombre total de clips vidéo contenant chacun une expression spécifique est de 208 dans cette base de données.

Cependant, après le processus de segmentation, nous obtenons 17 clips vidéos contenant deux étiquettes, par exemple "surpris de manière heureux", "peur / surprise", etc. Ceci est dû au fait que pour les enfants, différentes expressions peuvent coexister [129, 143] ou un stimulus visuel était si immersif que la transition d'une expression à une autre expression n'est pas prononcée. Référez-vous à la Figure 4.5 pour voir des exemples d'images de clips segmentés contenant des expressions mélangées.

Validation de la base de données

La base de données a été validée par 22 évaluateurs humains âgés de 18 à 40 ans, l'âge moyen étant de 26.8 ans. La moitié des évaluateurs de la base de données se situent dans la tranche d'âge de [18 - 25] ans et le reste dans la tranche d'âge de [26 - 40] ans. Les évaluateurs se situant dans la première tranche d'âge étaient des étudiants et l'autre tranche des professeurs d'université. Les évaluateurs ont été informés de l'expérience avant de commencer à valider la base de données.

Pour la validation de la base de données, nous avons construit un logiciel qui joue des vidéos segmentées (dans un ordre aléatoire) et permet l'enregistrement du choix de l'évaluateur humain pour choisir l'expression faciale. La capture d'écran du logiciel est présenté dans la Figure 4.6. Si nécessaire, l'évaluateur peut lire la vidéo plusieurs fois avant de choisir l'expression faciale. Nous avons également offert le choix à l'évaluateur de mettre une option indécise. Cela nous a aidés à recueillir des réponses authentiques et à éliminer les biais, car l'évaluateur peut choisir cette option lorsque la vidéo diffusée ne montrant aucune expression visible.

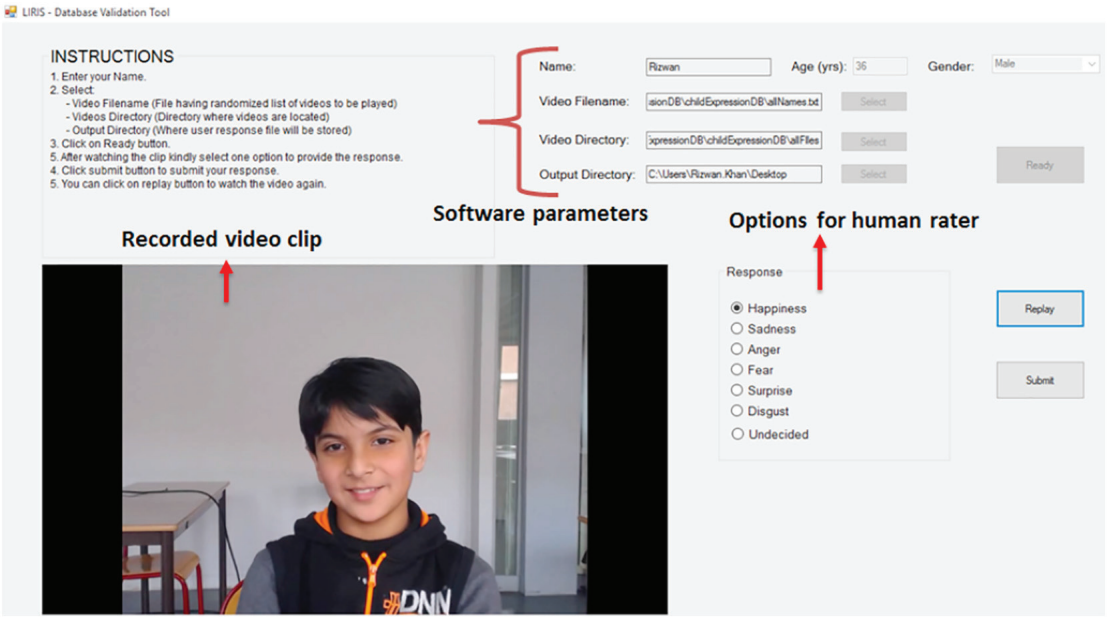


Figure 4.6 : Outil de validation. Capture d'écran de l'outil de validation utilisé pour collecter les expressions faciales des évaluateurs humains pour chaque clip vidéo.

En résumé, les instructions données aux évaluateurs humains étaient les suivantes :

1. Regardez attentivement chaque vidéo segmentée et sélectionnez l'expression faciale correspondante.

2. Si la vidéo présentée ne montrait aucune expression visible ou si vous avez un doute, sélectionnez l'option "indécise". Chaque vidéo peut être lue plusieurs fois sans limite.
3. Une fois que l'expression faciale est validée pour une vidéo lue, vous ne pourrez plus modifier votre choix.

Analyse des données de validation

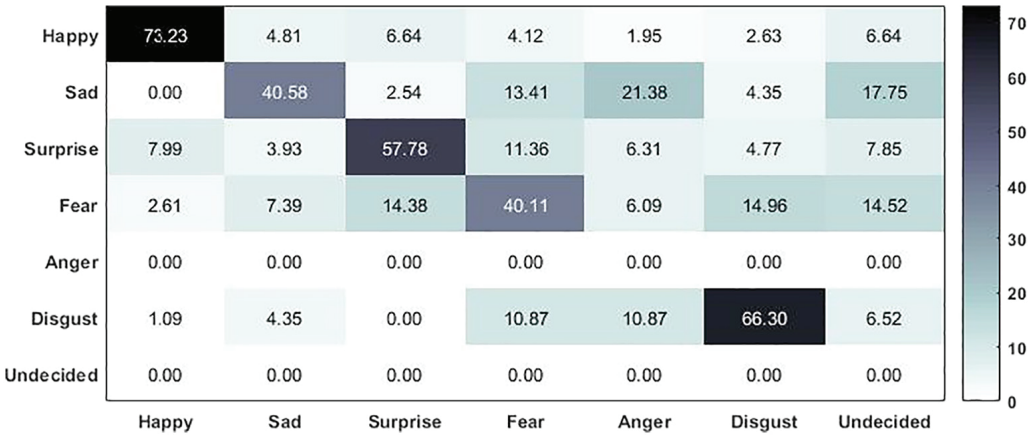


Figure 4.7 : Matrice de confusion. Lignes : expressions souhaitées lors de la création des clips vidéos (moyenne). Colonnes : expressions choisies par les évaluateurs humains (moyenne). La diagonale représente l'accord entre les expressions souhaitées et l'étiquette donnée par les évaluateurs humains.

Après la collecte des données de validation, nous avons effectué une analyse statistique des données recueillies et calculé la matrice de confusion. Référez-vous à la Figure 4.7 pour voir la matrice de confusion calculée. Les lignes de la matrice de confusion montrent les expressions souhaitées lors de la création des différents stimuli tandis que les colonnes indiquent l'étiquette de l'expression donnée par les évaluateurs humains. Les valeurs de la diagonale représentent l'accord entre les expressions souhaitées et les étiquettes données par les évaluateurs humains.

D'après les résultats calculés, l'expression du "bonheur" a été correctement repérée par les évaluateurs, avec une précision moyenne de 73.23%. D'autres part, l'expression de la "peur" a été la moins bien identifiée par les évaluateurs, avec une précision moyenne de 35.65%. Ces résultats concordent avec ceux de [28, 42].

L'expression de la "peur", qui est la moins identifiée, est souvent mélangée de façon perceptible aux expressions de "surprise" et de "dégoût". Comme nous l'avons mentionné plus haut, cela est dû au fait que pour les enfants, différentes expressions coexistent (expressions mélangées) [129, 143] ou un stimulus visuel était si immersif que le changement d'une expression à une autre expression n'était pas prononcé.

Référez-vous à la Figure 4.5 pour voir des exemples d'images de segments vidéo segmentés qui montrent des expressions mélangées.

La précision moyenne globale des évaluateurs humains est de 54,7%. Selon une étude publiée par Matsumoto et al. [102], les humains peuvent habituellement repérer correctement les expressions 50% du temps et l'expression la plus facile à identifier pour les humains est "heureux" et "surprise". Ces résultats concordent bien avec les résultats que nous avons obtenus des évaluateurs humains, car l'expression de "heureux" a été correctement identifiée, alors que l'exactitude moyenne des évaluateurs humains est également d'environ 50% (54,7% pour être exact).

4.1.4 La reconnaissance automatique de l'expression faciale, une approche basée sur l'apprentissage profond par transfert

Dans le but de proposer un apprentissage machine de référence sur notre base de données (LIRIS-CSE), nous avons effectués différentes expériences basées sur le concept de l'apprentissage par transfert. En effet, un algorithme classique de classification va faire des prédictions sur des données similaires à ses données d'entraînements. Les données d'entraînements et des tests sont globalement tirées de la même distribution. Au contraire de l'apprentissage par transfert qui permet aux distributions utilisées lors de l'entraînement et des tests d'être différentes. Nous avons utilisé l'approche de l'apprentissage par transfert dans notre expérience en raison du facteur suivant. Nous voulions bénéficier d'un modèle d'apprentissage en profondeur qui a atteint une grande précision sur les tâches de reconnaissance travaillant sur une image en entrée, c'est-à-dire le défi de reconnaissance visuelle à grande échelle ImageNet [33], et qui est disponible pour la recherche.

L'architecture de CNN, telle que vue dans la Figure 4.8, a été proposée pour la première fois par LeCun [92]. Il s'agit d'une architecture multi-étape ou multicouche. Cela signifie essentiellement qu'il y a plusieurs étapes dans le CNN pour l'extraction de descripteurs. Chaque étape du réseau dispose d'une entrée et d'une sortie composée de tableaux appelés cartes de caractéristiques. Chaque carte de caractéristiques de sortie est constituée de motifs ou d'entités extraits à l'emplacement de la carte d'entités d'entrée. Chaque étape est composée de trois couches, après quoi la classification a lieu [58, 149, 2]. Ces couches sont les suivantes.

1. Convolution layer : cette couche utilise des filtres, qui sont convolué avec l'image, produisant des cartes d'activation ou de descripteurs.

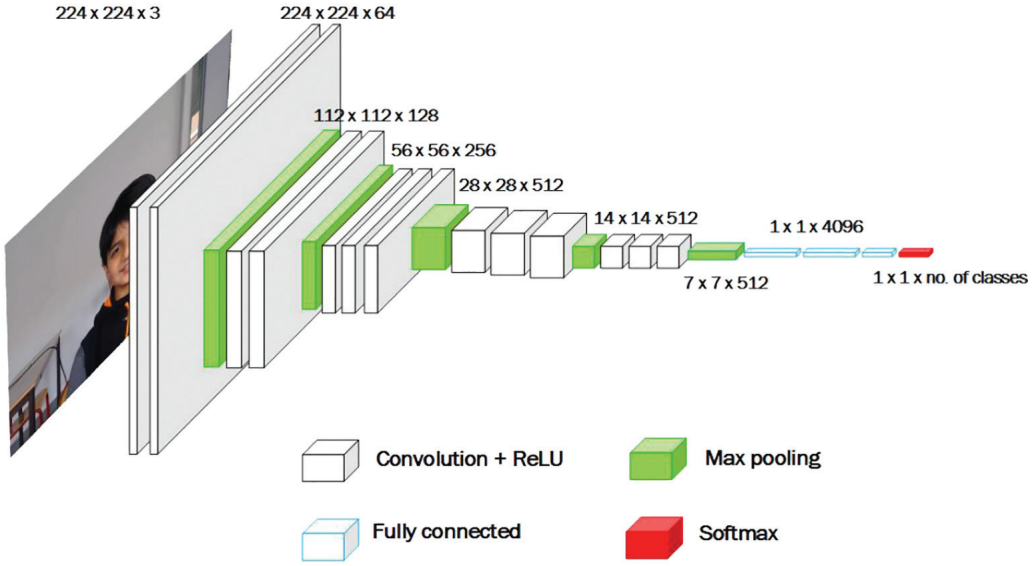


Figure 4.8 : Une vue d'ensemble de l'architecture de CNN. Un CNN comprend une couche d'entrée, plusieurs couches alternées de convolution et de mise en commun maximale, une couche entièrement connectée et une couche de classification.[2]

2. Feature Pooling layer : cette couche est insérée pour réduire la taille de la représentation de l'image, afin de rendre le calcul efficace. Le nombre de paramètres est également réduit, ce qui permet de contrôler le sur-apprentissage.
3. Couche de classification des éléments : c'est la couche entièrement connectée. Cette couche calcule la probabilité / score des classes apprises à partir des caractéristiques extraites de la couche de convolution dans les étapes précédentes.

4.1.5 Résultats

Un réseau CNN a besoin d'une grande base de données pour apprendre un concept [154], ce qui le rend peu adapté pour différentes applications dans lesquelles nous n'avons pas assez de données suffisantes. Ce problème peut être abordé en utilisant les techniques d'apprentissages par transfert [110]. L'apprentissage par transfert est une approche d'apprentissage machine qui met l'accent sur la capacité d'appliquer les connaissances pertinentes des expériences d'apprentissage antérieures à un problème différent mais connexe.

Nous avons utilisé une approche d'apprentissage par transfert pour construire une méthode de reconnaissance d'expressions faciales car la taille de notre base de données n'est pas suffisante pour former de manière robuste toutes les couches du CNN.

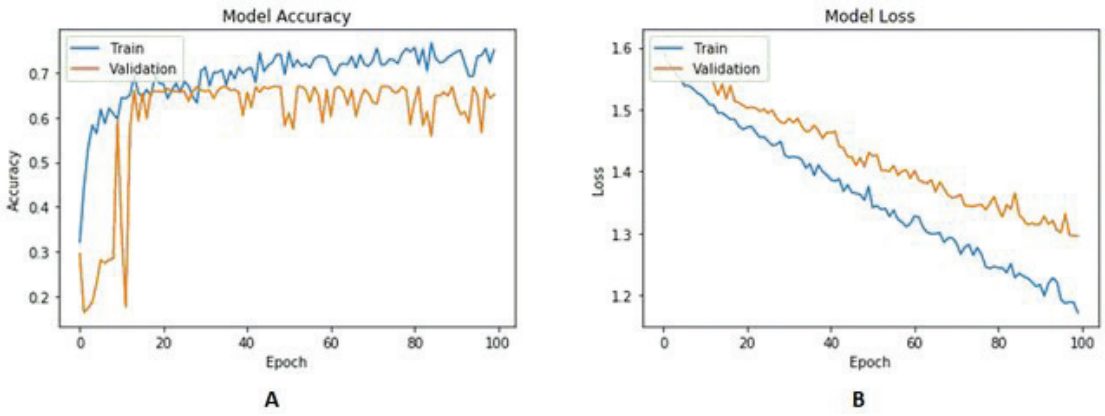


Figure 4.9 : Apprentissage du modèle CNN : (A) Précision de l'apprentissage par rapport à la précision de la validation (B) Perte d'apprentissage par rapport à la perte de validation.

Ainsi, nous avons utilisé un modèle VGG pré-entraîné sur des images génériques. VGG est un réseau convolutionnel profond entraîné pour la reconnaissance d'objets [153]. Il est développé et formé par le Groupe de Géométrie Visuelle (VGG) de l'Université d'Oxford et s'est révélé très performant sur l'ensemble de données ImageNet [33] pour la reconnaissance d'objets.

Nous avons remplacé la dernière couche entièrement connectée du modèle VGG préformé par une couche dense à cinq sorties. Le nombre de sorties de la dernière couche dense correspond au nombre de classes à reconnaître, dans notre expérience nous avons appris le concept de cinq classes soit cinq expressions à reconnaître (sur six expressions universelles, l'expression de la "colère" n'étant pas incluse dans cette expérience car il y a peu de clips pour "colère", voir section 4.1.3). Nous avons entrainé cette dernière couche dense avec les images de notre base de données en utilisant la fonction d'activation softmax et une optimisation par descente de gradient stochastique [81].

Notre base de données proposée est constituée de vidéos, mais pour cette expérience, nous avons extrait des images des vidéos et nous les avons fournies à l'architecture ConvNet décrite ci-dessus. Nous avons utilisé 80% des images pour l'apprentissage et 20% des images pour le processus de validation. Avec les paramètres mentionnés ci-dessus, le réseau CNN proposé a atteint un pourcentage de reconnaissance d'expression moyen de 75.5% sur notre base de données proposée (cinq expressions). Les courbes de précision et de perte du modèle sont illustrées à la figure 4.9.

Pour conclure cette section, nous avons présenté une nouvelle base de données pour la reconnaissance d'expressions spontanées de visages d'enfants (LIRIS-CSE). La base de données contient six expressions spontanées universelles exprimées par

12 enfants d'origines ethniques diverses âgés de 6 à 12 ans, l'âge moyen étant de 7,3 ans. La base de données LIRIS-CSE contient 208 petits morceaux vidéo contenant une expression neutre au début. Ensuite, ces segments vidéo montrent l'apparition d'une expression et se terminent lorsque l'expression est visible à son apogée, ainsi que quelques images après le pic de l'expression. La base de données a été validée par 22 évaluateurs humains âgés de 18 à 40 ans.

A notre connaissance, cette base de données est la première du genre car elle enregistre et présente six expressions spontanées universelles d'enfants. Auparavant, il y avait peu de bases de données d'expressions d'enfants et toutes montrent des expressions posées ou exagérées qui sont différentes des expressions spontanées ou naturelles. Cette base de données constituera donc un jalon important pour les chercheurs sur le comportement humain. Cette base de données constituera une excellente ressource pour la communauté des visionnaires pour l'analyse comparative et la comparaison des résultats. Actuellement, nous avons déjà plus de 55 universités / chercheurs qui se sont enregistrés afin de télécharger et utiliser la base de données. La base de données est disponible à l'adresse suivante : `https://childrenfacialexpression.projet.liris.cnrs.fr/`.

Pour l'évaluation comparative de la reconnaissance automatique d'expressions faciales, nous avons fourni des résultats à l'aide d'un réseau CNN avec une approche d'apprentissage par transfert. L'approche proposée a obtenu une précision d'expression moyenne de 72,45% sur notre base de données, LIRIS-CSE (cinq expressions).

Conclusion et perspectives

5

Contents

5.1 Contributions

Expressions faciales

Nos travaux nous ont menés à la création et au partage d'une nouvelle base de données d'expressions faciales spontanées de visages d'enfants. Cette contribution marque un pas vers la résolution d'un verrou que nous avons présenté dans l'état de l'art 1.3. En effet, cette base de données contient des vidéos d'expressions spontanées / naturelles de visages d'enfants capturées dans un environnement naturel. A notre connaissance, cette base de données est la première du genre puisqu'elle enregistre et montre les six expressions basiques et spontanées (plutôt cinq comme l'expression de "colère" est enregistrée seulement une fois). Auparavant, les seules bases de données existantes contenaient des images d'expressions d'enfants posées ou exagérées qui sont différentes des expressions spontanées ou naturelles. Cette base de données constituera donc un jalon important pour les chercheurs sur le comportement humain. Ce sera une excellente ressource pour la communauté de la vision par ordinateur en matière d'analyse d'expressions faciales notamment pour étudier les différences entre des expressions faciales d'adultes et d'enfants. Pour l'évaluation de la reconnaissance automatique d'expressions, nous avons également fourni des résultats à l'aide d'une architecture de réseau neuronal convolutionnel (CNN) avec transfert d'apprentissage. L'approche proposée a permis d'obtenir une précision d'expression moyenne de 75% dans notre base de données proposée. A noter que cette base de données rendu disponible en janvier 2019 aux chercheurs sur demande, compte déjà plus de 50 chercheurs enregistrés sur notre site web : `https://childrenfacialexpression.projet.liris.cnrs.fr/`.

Expressions corporelles

Dans ce travail de recherche, nous avons proposé différentes méthodes pour la reconnaissance d'expressions corporelles en se basant sur le mouvement du squelette. Tout d'abord, nous avons construit un ensemble de descripteurs inspiré du domaine de la psychologie sur le système visuel humain. Cet ensemble de descripteurs nous a permis d'obtenir de bons résultats de classification sur différentes bases de données. Cependant, grâce à ces premiers travaux, nous avons aussi compris le principal défaut des approches basées descripteurs pour la reconnaissance d'expressions corporelles. En effet, cette famille de méthode permet d'obtenir de très bons résultats lorsque le mouvement en entrée est connu et spécifique. Lorsque l'on cherche à extraire l'expression corporelles sur des mouvements hétérogènes

(marche, course, saut, coup de poing, etc.), les approches basées descripteurs n'arrivent plus à séparer correctement le mouvement réalisé de l'expression perçue. En effet, certains mouvements vont uniquement concerner le haut du corps tandis que d'autres le bas du corps. Afin que les approches basées descripteurs fonctionnent pour des mouvements hétérogènes, il serait nécessaire de construire différents ensembles de descripteurs pour chaque catégorie de mouvement. Cependant, cette solution rajouterait une étape de classification sur le geste réalisé en entrée. Afin de résoudre ce problème, nous avons décidé de proposer de nouvelles approches combinant le domaine de la vision par ordinateur, de la psychologie ainsi que la synthèse d'animations.

En s'inspirant de la synthèse d'animations, nous avons proposé une nouvelle approche basée sur l'extraction d'un résidu formé par la différence entre le mouvement expressif et un mouvement neutre synthétisé à partir du mouvement expressif. Ce point est l'hypothèse et la contribution majeure de nos travaux. Séparer le mouvement réalisé de l'expression perçue permet de ne concentrer que l'information pertinente pour la classification. Le verrou majeur que nous avons levé est l'obtention automatique d'un mouvement neutre à partir d'un mouvement expressif. En effet, la soustraction entre un mouvement expressif et le même mouvement neutre nous permet d'obtenir le résidu contenant l'expression contenue dans le mouvement en entrée. Cette approche nous a conduit a proposer deux méthodes pour la reconnaissance d'expressions corporelles. Dans la première méthode, nous avons cherché à formaliser et obtenir automatiquement un mouvement neutre par l'optimisation d'une fonction de coût. Cette dernière est basée sur la notion d'énergie utilisée pour réaliser un mouvement neutre comparé à un mouvement expressif. En combinant l'optimisation de cette fonction de coût à un algorithme de cinématique inverse, nous arrivons à produire des mouvements neutres correspondant aux différents mouvements fournis en entrée. Cependant, les mouvements neutres synthétisés contiennent des artefacts qui les rendent peu naturels. Malgré ce défaut, notre méthode a surclassé l'état de l'art en reconnaissance d'expressions corporelles sur quatre bases de données différentes. Notre seconde méthode, basée sur la synthèse d'un mouvement neutre, va chercher à résoudre ce problème afin de produire des animations neutres humainement plus réaliste. Pour cela, nous avons formalisé la neutralité d'un mouvement par une nouvelle fonction de coût. Cette fonction de coût inclus un terme de neutralité, un terme d'attache aux données et un terme de contraintes. Son optimisation permet de synthétiser un mouvement neutre à partir de n'importe quel mouvement donné. L'expression corporelle est obtenue en calculant la différence, dans le domaine spectral, entre le mouvement d'entrée et le mouvement neutre synthétique. Le résultat obtenu sur les quatre bases de données évaluées montre la généricité de notre approche. En effet, nous avons évalué notre approche sur des ensembles de données hétérogènes avec différents mouvements

et expressions. Nous obtenons de meilleurs taux de reconnaissance de l'expression corporelle sur trois bases de données par rapport à l'état de l'art. Pour les différentes méthodes de reconnaissance d'expressions corporelles, il est important de souligner que les méthodes de l'état de l'art se sont focalisées à chaque fois sur une base de données spécifique. Ce point montre que notre approche est invariante par rapport au mouvement d'entrée. Nous pensons que cette approche est meilleure que les méthodes basées sur l'extraction de descripteurs qui fonctionnent bien si le mouvement d'entrée est spécifique, mais sont limitées lorsqu'elles doivent reconnaître l'expression corporelle à partir de mouvements hétérogènes.

5.2 Perspectives

Pour la suite, nous planifions de compléter la base de données d'expressions faciales de visages d'enfants. En effet, elle est encore incomplète avec seulement 12 enfants alors que nous disposons des vidéos de 18 enfants. Notre but est d'arriver à plusieurs dizaines d'enfants afin d'avoir une base de données plus large qui nous permettra d'analyser les différences d'expressions faciales entre un adulte et un enfant. Le but est de pouvoir fournir à la communauté de la vision par ordinateur une base de travail afin d'avancer dans la compréhension des expressions faciales. De même, une base de données plus complète permettra de proposer de nouvelles approches de reconnaissances d'expressions faciales de visages d'enfants. Notre méthode de classification actuelle est basée sur une approche d'apprentissage par transfert vu que la base de données est encore relativement petite.

Pour les expressions corporelles, nous avons comme objectif de combiner toutes les bases de données que nous avons utilisés dans ce manuscrit afin de pouvoir avoir accès à une base de données extrêmement complète en terme d'expressions de mouvements. Grâce à cette base de données, nous pourrons explorer les approches d'apprentissage profonds pour la reconnaissance d'expressions corporelles qui nécessitent souvent une grande bases de données complète. L'objectif d'utiliser ce genre d'approche est de pouvoir comprendre un peu plus où se situe l'information d'expressions corporelles dans un mouvement. En effet, de nombreux travaux dans l'état de l'art travaillent sur des bases de données non partagées, ce qui complique la tâche d'évaluation et de comparaison des différentes méthodes existantes. Construire une base de données complètes qui servira de référence permettra de faire avancer la communauté de vision par ordinateur. Il serait intéressant d'essayer de concevoir un réseau qui produirait l'expression neutre automatiquement. Nous pensons que l'utilisation d'un auto-encodeur ou d'un réseau antagonistes généra-

tifs (Generative Adversarial Networks ou GAN) serait approprié pour ce type de tâche.

Notre objectif à long terme est de proposer des approches pour la synthèse d'animations expressives, nous souhaitons mettre à profit notre méthode de reconnaissance d'expressions corporelles basée sur la synthèse d'un mouvement neutre dans ce but. Pour cela, nous planifions de faire une évaluation utilisateur afin d'évaluer la qualité de notre mouvement neutre produit de manière à l'utiliser en synthèse d'animations. En effet, les méthodes existantes en synthèse d'animations ont besoin d'un premier mouvement pour extraire l'expression que l'on cherche à transférer sur un second mouvement neutre. Afin d'avoir une méthode synthétisant n'importe quelle expression à partir de n'importe quel mouvement en entrée facilitera la tâche de nombreuses applications : jeux vidéos, films, etc. De même qu'énoncé ci-dessus, nous avons aussi comme projet d'utiliser des réseaux antagonistes génératifs (GAN : generative adversarial networks) afin de générer des animations expressives réalistes. Des premiers travaux ont été réalisés par Holden et al. [64, 66] et obtiennent des résultats très prometteurs. Notre idée est de combiner les approches fréquentielles, que nous avons utilisées pour la reconnaissance, avec ce type de réseaux afin de faciliter l'apprentissage de ces réseaux.

Le dernier point pour les perspectives consiste à proposer une méthode multimodale de reconnaissance d'expressions. En effet, en combinant les connaissances acquises pour la reconnaissance d'expressions faciales et corporelles nous avons pour objectif de construire un système complet de détection d'expressions. Il existe quelques bases de données combinant ces données afin de travailler sur ce problème. Cependant, la plupart des bases de données s'intéressent uniquement au haut du corps (visage, épaules et bras de temps en temps). La base de données réalisée par Sapinski et al. [125] contient le corps en entier ainsi que de l'audio. Il serait intéressant de tester nos différentes approches sur cette base de données. Le principal problème de ce domaine est de savoir correctement combiner les informations expressives du corps, du visage et de la parole. En effet, que se passe-t-il lorsque le corps indique une expression et le visage une expression différente. Cette question n'est pas encore totalement résolue dans le domaine de la psychologie, mais différentes études s'y intéressent, tout comme le domaine de la vision par ordinateur.

Bibliographie

[1]N. AIFANTI, C. PAPACHRISTOU et A. DELOPOULOS. “The MUG facial expression database”. In : *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. Avr. 2010, p. 1-4 (cf. p. 33).

[2]Md Zahangir ALOM, Tarek M TAHA, Christopher YAKOPCIC et al. “The history began from alexnet : A comprehensive survey on deep learning approaches”. In : *arXiv preprint arXiv :1803.01164* (2018) (cf. p. 94, 95).

[3]Kenji AMAYA, Armin BRUDERLIN et Tom CALVERT. “Emotion from motion”. In : *Graphics interface*. T. 96. 00306. Toronto, Canada, 1996, p. 222-229 (cf. p. 46).

[4]Andreas ARISTIDOU, Yiorgos CHRYSANTHOU et Joan LASENBY. “Extending FABRIK with model constraints”. In : *Computer Animation and Virtual Worlds* 27.1 (2016), p. 35-57 (cf. p. 62).

[5]Andreas ARISTIDOU et Joan LASENBY. “FABRIK : A fast, iterative solver for the Inverse Kinematics problem”. en. In : *Graphical Models* 73.5 (sept. 2011). 00104, p. 243-260 (cf. p. 62).

[6]Joel ARONOFF, Barbara A. WOIKE et Lester M. HYMAN. “Which are the stimuli in facial displays of anger and happiness ? : Configurational bases of emotion recognition.” In : *Journal of Personality and Social Psychology* 62.6 (1992), p. 1050-1066 (cf. p. 36).

[7]A. P. ATKINSON, M. L. TUNSTALL et W. H. DITTRICH. “Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures”. In : *Cognition* 104.1 (juil. 2007), p. 59-72 (cf. p. 36).

[8]A BALESTRINO, Giuseppe DE MARIA et L SCIAVICCO. “Robust control of robotic manipulators”. In : *IFAC Proceedings Volumes* 17.2 (1984), p. 2435-2440 (cf. p. 62).

[9]T. BANZIGER, M. MORTILLARO et K. R. SCHERER. “Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception”. In : *Emotion* 12.5 (oct. 2012), p. 1161-1179 (cf. p. 37).

[10]Marian Stewart BARTLETT, Gwen C LITTLEWORT, TJ SEJNOWSKI et JR MOVELLAN. “A prototype for automatic recognition of spontaneous facial actions”. In : *Advances in neural information processing systems*. 2003, p. 1295-1302 (cf. p. 84, 86).

[11]Shishir BASHYAL et Ganesh K. VENAYAGAMOORTHY. “Recognition of facial expressions using Gabor wavelets and learning vector quantization”. In : *Engineering Applications of Artificial Intelligence* 21.7 (2008), p. 1056-1064 (cf. p. 29).

[12]John N BASSILI. “Emotion recognition : The role of facial movement and the relative importance of upper and lower areas of the face.” In : *Journal of personality and social psychology* 37.11 (1979), p. 2049 (cf. p. 86).

[13]Daniel BERNHARDT et Peter ROBINSON. “Detecting affect from non-stylised body motions”. In : *Affective Computing and Intelligent Interaction*. 00129. Springer, 2007, p. 59-70 (cf. p. 53, 59, 68).

[14]Nadia BIANCHI-BERTHOUZE et Andrea KLEINSMITH. “A categorical approach to affective gesture recognition”. In : *Connection Science* 15.4 (2003), p. 259-269. eprint : https://doi.org/10.1080/09540090310001658793 (cf. p. 42).

[15]Suparna BISWAS et Jaya SIL. “An Efficient Expression Recognition Method Using Contourlet Transform”. In : *Proceedings of the 2Nd International Conference on Perception and Machine Intelligence*. PerMIn ’15. Kolkata, West Bengal, India : ACM, 2015, p. 167-174 (cf. p. 30).

[16]Matthew BRAND et Aaron HERTZMANN. “Style machines”. In : *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 00740. ACM Press/Addison-Wesley Publishing Co., 2000, p. 183-192 (cf. p. 44).

[17]Armin BRUDERLIN et Lance WILLIAMS. “Motion signal processing”. In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00955. ACM, 1995, p. 97-104 (cf. p. 45).

[18]John T CACIOPPO et Gary G BERNTSON. “Relationship between attitudes and evaluative space : A critical review, with emphasis on the separability of positive and negative substrates.” In : *Psychological bulletin* 115.3 (1994), p. 401 (cf. p. 25).

[19]A. CAMURRI, N. DAEL, D. GLOWINSKI et al. “Toward a Minimal Representation of Affective Gestures”. In : *IEEE Transactions on Affective Computing* 2.02 (avr. 2011), p. 106-118 (cf. p. 36).

[20]Antonio CAMURRI, Barbara MAZZARINO, Matteo RICCHETTI, Renee TIMMERS et Gualtiero VOLPE. “Multimodal analysis of expressive gesture in music and dance performances”. In : *International gesture workshop*. Springer. 2003, p. 20-39 (cf. p. 42).

[21]Antonio CAMURRI, Riccardo TROCCA et Gualtiero VOLPE. “Interactive Systems Design : A KANSEI-based Approach”. In : *Proceedings of the 2002 Conference on New Interfaces for Musical Expression*. NIME ’02. Dublin, Ireland : National University of Singapore, 2002, p. 1-8 (cf. p. 42).

[22]Chih-Chung CHANG et Chih-Jen LIN. “LIBSVM : A library for support vector machines”. In : *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27 :1-27 :27 (cf. p. 58).

[23]M. J. COSSETIN, J. C. NIEVOLA et A. L. KOERICH. “Facial expression recognition using a pairwise feature selection and classification approach”. In : *2016 International Joint Conference on Neural Networks (IJCNN)*. Juil. 2016, p. 5149-5155 (cf. p. 29, 30).

[24]Mark COULSON. “Attributing Emotion to Static Body Postures : Recognition Accuracy, Confusions, and Viewpoint Dependence”. In : *Journal of Nonverbal Behavior* 28.2 (juin 2004), p. 117-139 (cf. p. 40).

[25]Arthur CRENN, Rizwan Ahmed KHAN, Alexandre MEYER et Saida BOUAKAZ. “Body expression recognition from animated 3D skeleton”. In : 00000. IEEE, déc. 2016, p. 1-7 (cf. p. 68).

[26]Arthur CRENN, Alexandre MEYER, Rizwan Ahmed KHAN, Hubert KONIK et Saïda BOUAKAZ. “Toward an Efficient Body Expression Recognition Based on the Synthesis of a Neutral Movement”. In : *19th ACM International Conference on Multimodal Interaction*. T. 17. ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow, United Kingdom, nov. 2017 (cf. p. 77, 80).

[27]M. DAHMANE et J. MEUNIER. “Prototype-Based Modeling forFacial Expression Analysis”. In : *IEEE Transactions on Multimedia* 16.6 (oct. 2014), p. 1574-1584 (cf. p. 31).

[28]Kirsten A DALRYMPLE, Jesse GOMEZ et Brad DUCHAINE. “The Dartmouth Database of Children’s Faces : acquisition and validation of a new face stimulus set”. In : *PloS one* 8.11 (2013), e79131 (cf. p. 85, 86, 93).

[29]Charles DARWIN. *L’expression des émotions chez l’homme et les animaux*. 1877 (cf. p. 24).

[30]Fernando DE LA TORRE, Wen-Sheng CHU, Xuehan XIONG et al. “IntraFace”. In : mai 2015 (cf. p. 25).

[31]P. Ravindra DE SILVA et Nadia BIANCHI-BERTHOUZE. “Modeling human affective postures : an information theoretic characterization of posture features”. In : *Computer Animation and Virtual Worlds* 15.3-4 (2004), p. 269-276. eprint : `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.29` (cf. p. 37).

[32]Özlem DEMIR, Mehdi LABAIED, Chris MERRITT, Ken STUART et Rommie E. AMARO. “Computer-Aided Discovery of Trypanosoma brucei RNA-Editing Terminal Uridylyl Transferase 2 Inhibitors”. In : *Chemical Biology & Drug Design* 84.2 (2014), p. 131-139 (cf. p. 32).

[33]Jia DENG, Wei DONG, Richard SOCHER et al. “Imagenet : A large-scale hierarchical image database”. In : *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, p. 248-255 (cf. p. 94, 96).

[34]Swati DEWAN, Shubham AGARWAL et Navjyoti SINGH. “Laban movement analysis to classify emotions from motion”. In : *Tenth International Conference on Machine Vision (ICMV 2017)*. T. 10696. International Society for Optics et Photonics, avr. 2018, 106962Q (cf. p. 43, 80).

[35]A. DHALL, R. GOECKE, S. LUCEY et T. GEDEON. “Collecting Large, Richly Annotated Facial-Expression Databases from Movies”. In : *IEEE MultiMedia* 19.3 (juil. 2012), p. 34-41 (cf. p. 34).

[36]A. DHALL, R. GOECKE, S. LUCEY et T. GEDEON. “Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark”. In : *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. Nov. 2011, p. 2106-2112 (cf. p. 34).

[37]Abhinav DHALL, Akshay ASTHANA, Roland GOECKE et Tom GEDEON. “Emotion recognition using PHOG and LPQ features”. In : *Face and Gesture 2011*. IEEE. 2011, p. 878-883 (cf. p. 32).

[38]Abhinav DHALL, Lucey Tom GEDEON, Tr-cs– Daniel FRAMPTON et al. *Acted Facial Expressions In The Wild Database*. 2009 (cf. p. 34).

[39]Russell EBERHART et James KENNEDY. “Particle swarm optimization”. In : *Proceedings of the IEEE international conference on neural networks*. T. 4. Citeseer. 1995, p. 1942-1948 (cf. p. 66).

[40]Helen Link EGGER, Daniel S PINE, Eric NELSON et al. “The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) : a new set of children’s facial emotion stimuli”. In : *International journal of methods in psychiatric research* 20.3 (2011), p. 145-156 (cf. p. 85).

[41]P. EKMAN, W. V. FRIESEN, M. O’SULLIVAN et al. “Universals and cultural differences in the judgments of facial expressions of emotion”. In : *J Pers Soc Psychol* 53.4 (oct. 1987), p. 712-717 (cf. p. 23).

[42]P EKMAN et WV FRIESEN. “Pictures of facial affect consulting psychologists press”. In : *Palo Alto, CA* (1976) (cf. p. 88, 93).

[43]Paul EKMAN. “Facial expression and emotion.” In : *American psychologist* 48.4 (1993), p. 384 (cf. p. 88).

[44]Paul EKMAN et Wallace V FRIESEN. “Measuring facial movement”. In : *Environmental psychology and nonverbal behavior* 1.1 (1976), p. 56-75 (cf. p. 24).

[45]Paul EKMAN, Wallace V FRIESEN, Maureen O’SULLIVAN et al. “Universals and cultural differences in the judgments of facial expressions of emotion.” In : *Journal of personality and social psychology* 53.4 (1987), p. 712 (cf. p. 84).

[46]Rosenberg EKMAN. *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997 (cf. p. 33).

[47]Ildar FARKHATDINOV. “Modelling verticality estimation during locomotion”. Theses. Université Pierre et Marie Curie - Paris VI, juin 2013 (cf. p. 31).

[48]Klaus FÖRGER et Tapio TAKALA. “Animating with style : defining expressive semantics of motion”. In : *The Visual Computer* (2015). 00000, p. 1-13 (cf. p. 53).

[49]Nesrine FOURATI et Catherine PELACHAUD. “Emilya : Emotional body expression in daily actions database.” In : *LREC*. 2014, p. 3486-3493 (cf. p. 48-50).

[50]Nesrine FOURATI et Catherine PELACHAUD. “Perception of emotions and body movement in the Emilya database”. In : *IEEE Transactions on Affective Computing* (2016) (cf. p. 74).

[51]Quan GAN, Siqi NIE, Shangfei WANG et Qiang JI. “Differentiating between posed and spontaneous expressions with latent regression bayesian network”. In : *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (cf. p. 86).

[52]Leon A GATYS, Alexander S ECKER et Matthias BETHGE. “A neural algorithm of artistic style”. In : *arXiv preprint arXiv :1508.06576* (2015) (cf. p. 47).

[53]A.S. GEORGHIADES, P.N. BELHUMEUR et D.J. KRIEGMAN. “From Few to Many : Illumination Cone Models for Face Recognition under Variable Lighting and Pose”. In : *IEEE Trans. Pattern Anal. Mach. Intelligence* 23.6 (2001), p. 643-660 (cf. p. 33).

[54]M. A. GIESE et T. POGGIO. “Neural mechanisms for the recognition of biological movements”. In : *Nat. Rev. Neurosci.* 4.3 (mar. 2003), p. 179-192 (cf. p. 35).

[55]Virginie GOUTTE et Anne-Marie ERGIS. “Traitement des émotions dans les pathologies neurodégénératives : une revue de la littérature”. In : *Revue de neuropsychologie* 3.3 (2011), p. 161-175 (cf. p. 22).

[56]Keith GROCHOW, Steven L. MARTIN, Aaron HERTZMANN et Zoran POPOVIĆ. “Style-based inverse kinematics”. In : *ACM Transactions on Graphics (TOG)*. T. 23. 00748. ACM, 2004, p. 522-531 (cf. p. 44).

[57]M. Melissa GROSS, Elizabeth A. CRANE et Barbara L. FREDRICKSON. “Methodology for Assessing Bodily Expression of Emotion”. In : *Journal of Nonverbal Behavior* 34.4 (déc. 2010), p. 223-248 (cf. p. 36, 41).

[58]Isma HADJI et Richard P WILDES. “What do we understand about convolutional networks ?” In : *arXiv preprint arXiv :1803.08834* (2018) (cf. p. 94).

[59]S. L. HAPPY et A. ROUTRAY. “Automatic facial expression recognition using features of salient facial patches”. In : *IEEE Transactions on Affective Computing* 6.1 (jan. 2015), p. 1-12 (cf. p. 29).

[60]Jinni A. HARRIGAN et Robert ROSENTHAL. “Physicians’ H]ead and Body Positions as Determinants of Perceived Rapport”. In : *Journal of Applied Social Psychology* 13.6 (1983), p. 496-509. eprint : `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1983.tb02332.x` (cf. p. 36).

[61]G.P. HEGDE, M. SEETHA et Nagaratna HEGDE. “Kernel Locality Preserving Symmetrical Weighted Fisher Discriminant Analysis based subspace approach for expression recognition”. In : *Engineering Science and Technology, an International Journal* 19.3 (2016), p. 1321-1333 (cf. p. 29).

[62]Andres HERNANDEZ-MATAMOROS, Andrea BONARINI, Enrique ESCAMILLA-HERNANDEZ, Mariko NAKANO-MIYATAKE et Hector PEREZ-MEANA. “Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach”. In : *Knowledge-Based Systems* 110 (2016), p. 1-14 (cf. p. 29).

[63]M. HIRAI et K. HIRAKI. “The relative importance of spatial versus temporal structure in the perception of biological motion : an event-related potential study”. In : *Cognition* 99.1 (fév. 2006), p. 15-29 (cf. p. 36).

[64]Daniel HOLDEN, Ikhsanul HABIBIE, Ikuo KUSAJIMA et Taku KOMURA. “Fast Neural Style Transfer for Motion Data”. In : *IEEE Computer Graphics and Applications* 37.4 (2017), p. 42-49 (cf. p. 103).

[65]Daniel HOLDEN, Jun SAITO et Taku KOMURA. “A deep learning framework for character motion synthesis and editing”. In : *ACM Transactions on Graphics (TOG)* 35.4 (2016), p. 138 (cf. p. 46, 48).

[66]Daniel HOLDEN, Jun SAITO et Taku KOMURA. “A deep learning framework for character motion synthesis and editing”. en. In : *ACM Transactions on Graphics* 35.4 (juil. 2016). 00022, p. 1-11 (cf. p. 103).

[67]Daniel HOLDEN, Jun SAITO, Taku KOMURA et Thomas JOYCE. “Learning motion manifolds with convolutional autoencoders”. In : *SIGGRAPH Asia 2015 Technical Briefs*. ACM. 2015, p. 18 (cf. p. 46).

[68]Eugene HSU, Kari PULLI et Jovan POPOVIĆ. “Style translation for human motion”. In : *ACM Transactions on Graphics (TOG)* 24.3 (2005). 00302, p. 1082-1089 (cf. p. 44, 45).

[69]William JAMES. “What is an Emotion?” In : *Mind* 9.34 (1884), p. 188-205 (cf. p. 22).

[70]William T. JAMES. “A Study of the Expression of Bodily Posture”. In : *The Journal of General Psychology* 7.2 (1932), p. 405-437. eprint : `https://doi.org/10.1080/00221309.1932.9918475` (cf. p. 36).

[71]Yi JI et Khalid IDRISSI. “Automatic facial expression recognition based on spatio-temporal descriptors”. In : *Pattern Recognition Letters* 33.10 (2012), p. 1373-1380 (cf. p. 30).

[72]Ian JOLLIFFE. *Principal component analysis*. Springer, 2011 (cf. p. 72).

[73]Miyuki KAMACHI, Michael LYONS et Jiro GYOBA. “The japanese female facial expression (jaffe) database”. In : *Availble : http ://www. kasrl. org/jaffe. html* (jan. 1997) (cf. p. 33).

[74]T. KANADE, J. F. COHN et YINGLI TIAN. “Comprehensive database for facial expression analysis”. In : *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. Mar. 2000, p. 46-53 (cf. p. 33).

[75]Asha KAPUR, Ajay KAPUR, Naznin VIRJI-BABUL, George TZANETAKIS et Peter F. DRIESSEN. “Gesture-Based Affective Computing on Motion Capture Data”. In : *Affective Computing and Intelligent Interaction*. Sous la dir. de Jianhua TAO, Tieniu TAN et Rosalind W. PICARD. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, p. 1-7 (cf. p. 42).

[76]M. KARG, K. KUHNLENZ et M. BUSS. “Recognition of Affect Based on Gait Patterns”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4 (août 2010), p. 1050-1061 (cf. p. 42).

[77]Michelle KARG, Kolja KÜHNLENZ et Martin BUSS. “Recognition of Affect Based on Gait Patterns”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4 (août 2010). 00080, p. 1050-1061 (cf. p. 18, 53).

[78]Rizwan Ahmed KHAN. “Détection des émotions à partir de vidéos dans un environnement non contrôlé”. 2013LYO10227. Thèse de doct. 2013 (cf. p. 28).

[79]Rizwan Ahmed KHAN, Alexandre MEYER, Hubert KONIK et Saida BOUAKAZ. “Exploring human visual system : study to aid the development of automatic facial expression recognition framework”. In : *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, p. 49-54 (cf. p. 89).

[80]Rizwan Ahmed KHAN, Alexandre MEYER, Hubert KONIK et Saida BOUAKAZ. “Framework for reliable, real-time facial expression recognition for low resolution images”. In : *Pattern Recognition Letters* 34.10 (2013), p. 1159-1168 (cf. p. 84).

[81]Diederik P KINGMA et Jimmy BA. “Adam : A method for stochastic optimization”. In : *arXiv preprint arXiv :1412.6980* (2014) (cf. p. 96).

[82]Paul R. KLEINGINNA et Anne M. KLEINGINNA. “A categorized list of emotion definitions, with suggestions for a consensual definition”. In : *Motivation and Emotion* 5.4 (déc. 1981), p. 345-379 (cf. p. 22).

[83]A. KLEINSMITH, N. BIANCHI-BERTHOUZE et A. STEED. “Automatic Recognition of Non-Acted Affective Postures”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.4 (août 2011). 00125, p. 1027-1038 (cf. p. 41, 42).

[84]Andrea KLEINSMITH et Nadia BIANCHI-BERTHOUZE. “Recognizing Affective Dimensions from Body Posture”. In : *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*. ACII ’07. Lisbon, Portugal : Springer-Verlag, 2007, p. 48-58 (cf. p. 37).

[85]Andrea KLEINSMITH et Nadia BIANCHI-BERTHOUZE. “Recognizing Affective Dimensions from Body Posture”. In : *Affective Computing and Intelligent Interaction*. Sous la dir. d’Ana C. R. PAIVA, Rui PRADA et Rosalind W. PICARD. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 48-58 (cf. p. 38, 42).

[86]Andrea KLEINSMITH, P. Ravindra DE SILVA et Nadia BIANCHI-BERTHOUZE. “Cross-cultural differences in recognizing affect from body posture”. In : *Interacting with Computers* 18.6 (2006). 00163, p. 1371-1389 (cf. p. 38, 40, 43, 48-50).

[87]Andrea KLEINSMITH, Tsuyoshi FUSHIMI et Nadia BIANCHI-BERTHOUZE. “An incremental and interactive affective posture recognition system”. In : *International Workshop on Adapting the Interaction Style to Affective Factors*. 00038. 2005, p. 378-387 (cf. p. 53, 59).

[88]Byoung KO. “A brief review of facial emotion recognition based on visual information”. In : *sensors* 18.2 (2018), p. 401 (cf. p. 84).

[89]Irene KOTSIA, Stefanos ZAFEIRIOU et Ioannis PITAS. “Texture and shape information fusion for facial expression and facial action unit recognition”. In : *Pattern Recognition* 41.3 (2008), p. 833-851 (cf. p. 32).

[90]S. KUMAR, M. K. BHUYAN et B. K. CHAKRABORTY. “Extraction of informative regions of a face for facial expression recognition”. In : *IET Computer Vision* 10.6 (2016), p. 567-576 (cf. p. 30).

[91]Neil D. LAWRENCE. “Gaussian process latent variable models for visualisation of high dimensional data”. In : *Advances in neural information processing systems* 16.3 (2004). 00629, p. 329-336 (cf. p. 44).

[92]Yann LECUN, Koray KAVUKCUOGLU et Clément FARABET. “Convolutional networks and applications in vision”. In : *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE. 2010, p. 253-256 (cf. p. 94).

[93]Xiaobai LI, Tomas PFISTER, Xiaohua HUANG, Guoying ZHAO et Matti PIETIKÄINEN. “A spontaneous micro-expression database : Inducement, collection and baseline”. In : *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE. 2013, p. 1-6 (cf. p. 88, 89).

[94]Ce LIU, Antonio TORRALBA, William T. FREEMAN, Frédo DURAND et Edward H. ADELSON. “Motion magnification”. In : *ACM transactions on graphics (TOG)* 24.3 (2005). 00182, p. 519-526 (cf. p. 45).

[95]Vanessa LOBUE et Cat THRASHER. “The Child Affective Facial Expression (CAFE) set : Validity and reliability from untrained adults”. In : *Frontiers in psychology* 5 (2015), p. 1532 (cf. p. 84-86).

[96]P. LUCEY, J. F. COHN, T. KANADE et al. “The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression”. In : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. Juin 2010, p. 94-101 (cf. p. 33).

[97]P LUCEY, JF COHN, T KANADE, J SARAGIH et Z AMBADAR. “A complete facial expression dataset for action unit and emotion-specied expression”. In : *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2010, p. 94-101 (cf. p. 84).

[98]Wanli MA, Shihong XIA, Jessica K. HODGINS et al. “Modeling style and variation in human motion”. In : *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 00035. Eurographics Association, 2010, p. 21-30 (cf. p. 44, 45).

[99]Yingliang MA, Helena M. PATERSON et Frank E. POLLICK. “A motion capture library for the study of identity, gender, and emotion perception from biological motion”. In : *Behavior research methods* 38.1 (2006). 00130, p. 134-141 (cf. p. 48-50, 57).

[100]Hela MAHERSIA et Kamel HAMROUNI. “Using multiple steerable filters and Bayesian regularization for facial expression recognition”. In : *Engineering Applications of Artificial Intelligence* 38 (2015), p. 190-202 (cf. p. 32).

[101]A. M. MARTINEZ. “The AR face database”. In : *CVC Technical Report24* (1998) (cf. p. 33).

[102]David MATSUMOTO et Hyi Sung HWANG. “Reading facial expressions of emotion”. In : *Psychological Science Agenda* 25.5 (2011) (cf. p. 94).

[103]P. MCLEOD. “Preserved and Impaired Detection of Structure From Motion by a 'Motion-blind" Patient”. In : *Visual Cognition* 3.4 (1996), p. 363-392. eprint : `https://doi.org/10.1080/135062896395634` (cf. p. 36).

[104]A. MEHRABIAN. “Inference of attitudes from the posture, orientation, and distance of a communicator”. In : *J Consult Clin Psychol* 32.3 (juin 1968), p. 296-308 (cf. p. 36).

[105]Marco de MEIJER. “The contribution of general features of body movement to the attribution of emotions”. In : *Journal of Nonverbal Behavior* 13.4 (déc. 1989), p. 247-268 (cf. p. 37, 42).

[106]Florian MUELLER et Nadia BIANCHI-BERTHOUZE. “Evaluating exertion games”. In : *Game User Experience Evaluation*. Springer, 2015, p. 239-262 (cf. p. 16).

[107]Sungkyu NOH, Hanhoon PARK, Yoonjong JIN et Jong-Il PARK. “Feature-Adaptive Motion Energy Analysis for Facial Expression Recognition”. In : *Advances in Visual Computing*. Sous la dir. de George BEBIS, Richard BOYLE, Bahram PARVIN et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 452-463 (cf. p. 30).

[108]Lars OMLOR et Martin A. GIESE. “Extraction of spatio-temporal primitives of emotional body expressions”. In : *Neurocomputing* 70.10 (2007). Computational Neuroscience : Trends in Research 2007, p. 1938-1942 (cf. p. 36).

[109]Ebenezer OWUSU, Yongzhao ZHAN et Qi Rong MAO. “A neural-AdaBoost based facial expression recognition system”. In : *Expert Systems with Applications* 41.7 (2014), p. 3383-3390 (cf. p. 29).

[110]Sinno Jialin PAN et Qiang YANG. “A survey on transfer learning”. In : *IEEE Transactions on knowledge and data engineering* 22.10 (2009), p. 1345-1359 (cf. p. 95).

[111]M. PANTIC, M. VALSTAR, R. RADEMAKER et L. MAAT. “Web-based database for facial expression analysis”. In : *2005 IEEE International Conference on Multimedia and Expo*. Juil. 2005 (cf. p. 33).

[112]Maja PANTIC, Michel VALSTAR, Ron RADEMAKER et Ludo MAAT. “Web-based database for facial expression analysis”. In : *2005 IEEE international conference on multimedia and Expo*. IEEE. 2005, 5-pp (cf. p. 84).

[113]Hanhoon PARK, Jong-Il PARK, Un-Mi KIM et Woontack WOO. “Emotion Recognition from Dance Image Sequences Using Contour Approximation”. In : *Structural, Syntactic, and Statistical Pattern Recognition*. Sous la dir. d’Ana FRED, Terry M. CAELLI, Robert P. W. DUIN, Aurélio C. CAMPILHO et Dick de RIDDER. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 547-555 (cf. p. 42).

[114]Hanhoon PARK, Jong-Il PARK, Un-Mi KIM et Woontack WOO. “Emotion recognition from dance image sequences using contour approximation”. In : *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer. 2004, p. 547-555 (cf. p. 18).

[115]M. V. PEELEN, A. J. WIGGETT et P. E. DOWNING. “Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion”. In : *Neuron* 49.6 (mar. 2006), p. 815-822 (cf. p. 36).

[116]Robert PLUTCHIK. “The nature of emotions : Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”. In : *American scientist* 89.4 (2001), p. 344-350 (cf. p. 26).

[117]Frank E POLLICK, Vaia LESTOU, Jungwon RYU et Sung-Bae CHO. “Estimating the efficiency of recognizing gender and affect from biological motion”. In : *Vision research* 42.20 (2002), p. 2345-2355 (cf. p. 18).

[118]Ahmad POURSABERI, Hossain AHMADI NOUBARI, Marina GAVRILOVA et Svetlana YANUSHKEVICH. “Gauss–Laguerre wavelet textural feature fusion with geometrical information for facial expression identification”. In : *EURASIP Journal on Image and Video Processing* 2012 (sept. 2012) (cf. p. 30).

[119]I.Michael REVINA et W.R. Sam EMMANUEL. “A Survey on Human Face Expression Recognition Techniques”. In : *Journal of King Saud University - Computer and Information Sciences* (2018) (cf. p. 33).

[120]C. L. ROETHER, L. OMLOR, A. CHRISTENSEN et M. A. GIESE. “Critical features for the perception of emotion from gait”. In : *J Vis* 9.6 (juin 2009), p. 1-32 (cf. p. 36).

[121]Claire L. ROETHER, Lars OMLOR, Andrea CHRISTENSEN et Martin A. GIESE. “Critical features for the perception of emotion from gait”. In : *Journal of Vision* 9.6 (juin 2009), p. 15-15. eprint : `https://jov.arvojournals.org/arvo/content\_public/journal/jov/933553/jov-9-6-15.pdf` (cf. p. 40).

[122]J. A. RUSSELL. “Core affect and the psychological construction of emotion”. In : *Psychol Rev* 110.1 (jan. 2003), p. 145-172 (cf. p. 41).

[123]James A RUSSELL. “A circumplex model of affect.” In : *Journal of personality and social psychology* 39.6 (1980), p. 1161 (cf. p. 25).

[124]F. Z. SALMAM, A. MADANI et M. KISSI. “Facial Expression Recognition Using Decision Trees”. In : *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. Mar. 2016, p. 125-130 (cf. p. 30).

[125]Tomasz SAPINSKI, Dorota KAMINSKA, Adam PELIKANT et al. “Multimodal Database of Emotional Speech, Video and Gestures”. In : *Pattern Recognition and Information Forensics - ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers*. 2018, p. 153-163 (cf. p. 103).

[126]Klaus R. SCHERER, Tanja WRANIK, Janique SANGSUE, Véronique TRAN et Ursula SCHERER. “Emotions in everyday life : probability of occurrence, risk factors, appraisal and reaction patterns”. In : *Social Science Information* 43.4 (2004), p. 499-570. eprint : `https://doi.org/10.1177/0539018404047701` (cf. p. 22).

[127]M. H. SIDDIQI, R. ALI, A. M. KHAN, Y. PARK et S. LEE. “Human Facial Expression Recognition Using Stepwise Linear Discriminant Analysis and Hidden Conditional Random Fields”. In : *IEEE Transactions on Image Processing* 24.4 (avr. 2015), p. 1386-1398 (cf. p. 31).

[128]M. SONG, D. TAO, Z. LIU, X. LI et M. ZHOU. “Image Ratio Features for Facial Expression Recognition Application”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.3 (juin 2010), p. 779-788 (cf. p. 30).

[129]Margaret Wolan SULLIVAN et Michael LEWIS. “Emotional expressions of young infants and children : A practitioner’s primer”. In : *Infants & Young Children* 16.2 (2003), p. 120-142 (cf. p. 91, 93).

[130]Fredric W. TAYLOR. “Epilogue”. In : *The Scientific Exploration of Venus*. Cambridge University Press, 2014, p. 277-278 (cf. p. 31).

[131]Arthur TRUONG, Hugo BOUJUT et Titus ZAHARIA. “Laban descriptors for gesture recognition and emotional analysis”. en. In : *The Visual Computer* 32.1 (jan. 2016). 00000, p. 83-98 (cf. p. 43, 53).

[132]Munetoshi UNUMA, Ken ANJYO et Ryozo TAKEUCHI. “Fourier principles for emotion-based human figure animation”. In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00582. ACM, 1995, p. 91-96 (cf. p. 45).

[133]L. M. VAINA, M. LEMAY, D. C. BIENFANG, A. Y. CHOI et K. NAKAYAMA. “Intact "biological motion" and "structure from motion" perception in a patient with impaired motion mechanisms : a case study”. In : *Vis. Neurosci.* 5.4 (oct. 1990), p. 353-369 (cf. p. 35).

[134]Michel VALSTAR et Maja PANTIC. “Induced disgust, happiness and surprise : an addition to the mmi facial expression database”. In : *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC) : Corpora for Research on Emotion and Affect*. Paris, France. 2010, p. 65 (cf. p. 84).

[135]Ekaterina VOLKOVA, Stephan de la ROSA, Heinrich H. BÜLTHOFF et Betty MOHLER. “The MPI Emotional Body Expressions Database for Narrative Scenarios”. en. In : *PLoS ONE* 9.12 (déc. 2014). Sous la dir. de Marko NARDINI. 00000, e113647 (cf. p. 48-50).

[136]Rudolf VON LABAN, Jacqueline CHALLET-HAAS et Marion HANSEN. *La maıtrise du mouvement*. Actes sud, 1994 (cf. p. 38).

[137]Harald G WALLBOTT. “Bodily expression of emotion”. In : *European journal of social psychology* 28.6 (1998), p. 879-896 (cf. p. 40).

[138]Harald G. WALLBOTT. “Bodily expression of emotion”. In : *European Journal of Social Psychology* 28.6 (1998), p. 879-896 (cf. p. 37, 40).

[139]Jack M. WANG, David J. FLEET et Aaron HERTZMANN. “Multifactor Gaussian process models for style-content separation”. In : *Proceedings of the 24th international conference on Machine learning*. 00000. ACM, 2007, p. 975-982 (cf. p. 44, 45).

[140]L-CT WANG et Chih-Cheng CHEN. “A combined optimization method for solving the inverse kinematics problems of mechanical manipulators”. In : *IEEE Transactions on Robotics and Automation* 7.4 (1991), p. 489-499 (cf. p. 62).

[141]Weiyi WANG, Valentin ENESCU et Hichem SAHLI. “Adaptive Real-Time Emotion Recognition from Body Movements”. en. In : *ACM Transactions on Interactive Intelligent Systems* 5.4 (déc. 2015). 00000, p. 1-21 (cf. p. 68).

[142] Chris WELMAN. “Inverse kinematics and geometric constraints for articulated figure manipulation”. Thèse de doct. Theses (School of Computing Science)/Simon Fraser University, 1993 (cf. p. 62).

[143] Sherri C WIDEN et James A RUSSELL. “Children’s recognition of disgust in others.” In : *Psychological Bulletin* 139.2 (2013), p. 271 (cf. p. 91, 93).

[144] Andrew WITKIN et Zoran POPOVIC. “Motion warping”. In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00000. ACM, 1995, p. 105-108 (cf. p. 45).

[145] William A WOLOVICH et H ELLIOTT. “A computational technique for inverse kinematics”. In : *The 23rd IEEE Conference on Decision and Control*. IEEE. 1984, p. 1359-1363 (cf. p. 62).

[146] Shihong XIA, Congyi WANG, Jinxiang CHAI et Jessica HODGINS. “Realtime style transfer for unlabeled heterogeneous human motion”. In : *ACM Transactions on Graphics (TOG)* 34.4 (2015). 00000, p. 119 (cf. p. 45, 48, 50, 57, 66).

[147] YONGSHENG GAO, M. K. H. LEUNG, SIU CHEUNG HUI et M. W. TANANDA. “Facial expression recognition from line-based caricatures”. In : *IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans* 33.3 (mai 2003), p. 407-412 (cf. p. 30).

[148] M. Ersin YUMER et Niloy J. MITRA. “Spectral style transfer for human motion between independent actions”. en. In : *ACM Transactions on Graphics* 35.4 (juil. 2016). 00000, p. 1-8 (cf. p. 46, 60, 66, 77).

[149] Matthew D ZEILER et Rob FERGUS. “Visualizing and understanding convolutional networks”. In : *European conference on computer vision*. Springer. 2014, p. 818-833 (cf. p. 94).

[150] L. ZHANG et D. TJONDRONEGORO. “Facial Expression Recognition Using Facial Movement Features”. In : *IEEE Transactions on Affective Computing* 2.4 (oct. 2011), p. 219-229 (cf. p. 32).

[151] Ligang ZHANG, Dian TJONDRONEGORO et Vinod CHANDRAN. “Random Gabor based templates for facial expression recognition in images with facial occlusion”. In : *Neurocomputing* 145 (2014), p. 451-464 (cf. p. 29).

[152] Guoying ZHAO et Matti PIETIKÄINEN. “Boosted multi-resolution spatiotemporal descriptors for facial expression recognition”. In : *Pattern Recognition Letters* 30.12 (2009). Image/video-based Pattern Analysis and HCI Applications, p. 1117-1127 (cf. p. 29).

[153] Bolei ZHOU, Agata LAPEDRIZA, Aditya KHOSLA, Aude OLIVA et Antonio TORRALBA. “Places : A 10 million image database for scene recognition”. In : *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), p. 1452-1464 (cf. p. 96).

[154] Fengwei ZHOU, Bin WU et Zhenguo LI. “Deep meta-learning : Learning to learn in the concept space”. In : *arXiv preprint arXiv :1802.03596* (2018) (cf. p. 95).