Reconnaissance des gestes expressifs inspirée du modèle LMA pour une interaction naturelle homme-robot

Insaf Ajili

▶ **To cite this version:**

Insaf Ajili. Reconnaissance des gestes expressifs inspirée du modèle LMA pour une interaction naturelle homme-robot. Apprentissage [cs.LG]. Université Paris-Saclay; Université d'Evry-Val-d'Essonne, 2018. Français. NNT : 2018SACLE037 . tel-01994111

HAL Id: tel-01994111
https://hal.science/tel-01994111

Submitted on 25 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

NNT : 2018SACLE037

Thèse de doctorat

Reconnaissance des gestes expressifs inspirée du modèle LMA pour une interaction naturelle Homme-Robot

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université d'Evry-Val-d'Essonne

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Robotique

Thèse présentée et soutenue à Evry Courcouronnes, le 03/12/2018, par

INSAF AJILI

Composition du Jury :

Titus Zaharia Professeur, Télécom SudParis	Président
Catherine Achard Maitre de conférence Hdr, Sorbonne Université	Rapporteure
Fakhr-Eddine Ababsa Professeur, Arts et Métiers ParisTech	Rapporteur
Indira Thouvenin Professeur, Université de Technologie de Compiègne	Examinatrice
Malik Mallem Professeur, Université d'Evry Val d'Essonne	Directeur de thèse
Jean-Yves Didier Maitre de conférence, Université d'Evry Val d'Essonne	Co-directeur de thèse

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Monsieur Malik Mallem, pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour sa bienveillance, pour ses multiples conseils et pour sa disponibilité malgré ses nombreuses charges. Enfin, j'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de ce travail doctoral. Je tiens également à exprimer mes plus vifs remerciements à mon encadrant de thèse Monsieur Jean-Yves Didier qui m'a dirigé tout au long de ces trois années de thèse. J'aimerais également lui dire à quel point j'ai apprécié sa grande disponibilité et son respect sans faille des délais serrés de relecture des documents que je lui ai adressés. Je voudrais remercier, Madame Catherine Achard et Monsieur Fakhr-Eddine Ababsa d'avoir accepté d'être les rapporteurs de ma thèse et Madame Indira Thouvenin et Monsieur Titus Zaharia d'avoir examiné ce manuscrit. Je suis très sensible à l'intérêt que vous avez manifesté à l'égard de ce travail et je vous remercie de m'avoir consacré une partie de vos temps précieux. Une thèse c'est aussi un laboratoire où l'on passe de nombreuses heures et où il est important de se sentir bien. Alors, je remercie infiniment tous les membres du laboratoire IBISC pour leur générosité et leur bonne humeur.

D'un point de vue plus personnel, je remercie mes parents et ma soeur de m'avoir toujours soutenue dans mes études et dans mes choix. Malgré la distance, ils ont été toujours à l'écoute . Comment pourrais-je trouver les mots pour remercier mon mari Bilel de la patience, l'encouragement et l'attention qu'il m'a accordée durant cette thèse ? J'aurai difficilement pu terminer ma thèse dans de bonnes conditions sans son amour et son soutien. Finalement, je dédie cette thèse à mon petit garçon TIMO.

Résumé

Dans cette thèse, nous traitons le problème de la reconnaissance des gestes dans un contexte d'interaction homme-robot. De nouvelles contributions sont apportées à ce sujet. Notre système consiste à reconnaitre les gestes humains en se basant sur une méthode d'analyse de mouvement qui décrit le geste humain d'une manière précise. Dans le cadre de cette étude, un module de niveau supérieur est intégré afin de reconnaître les émotions de la personne à travers le mouvement de son corps. Trois approches sont réalisées : la première porte sur la reconnaissance des gestes dynamiques en appliquant le modèle de Markov caché (MMC) comme méthode de classification. Un descripteur de mouvement local est implémenté basé sur une méthode d'analyse de mouvement, nommée LMA (Laban Movement Analysis) qui permet de décrire le mouvement de la personne dans ses différents aspects. Notre système est invariant aux positions et orientations initiales des personnes. Un algorithme d'échantillonnage a été développé afin de réduire la taille de notre descripteur et aussi adapter les données aux modèles de Markov cachés. Une contribution est réalisée aux MMCs pour analyser le mouvement dans deux sens (son sens naturel et le sens inverse) et ainsi améliorer la classification des gestes similaires. Plusieurs expériences sont faites en utilisant des bases de données d'actions publiques, ainsi que notre base de données composée de gestes de contrôle. Dans la seconde approche, un système de reconnaissance des gestes expressifs est mis en place afin de reconnaitre les émotions des personnes à travers leurs gestes. Une deuxième contribution consiste en le choix d'un descripteur de mouvement global basé sur les caractéristiques locales proposées dans la première approche afin de décrire l'entièreté du geste. La composante Effort de LMA est quantifiée afin de décrire l'expressivité du geste avec ses 4 facteurs (espace, temps, poids et flux). La classification des gestes expressifs est réalisée avec 4 méthodes d'apprentissage automatique réputées (les forêts d'arbres décisionnels, le perceptron multicouches, les machines à vecteurs de support : un-contre-un et un-contre-tous). Une étude comparative est faite entre ces 4 méthodes afin de choisir la meilleure. L'approche est validée avec des bases publiques et notre propre base des gestes expressifs. La troisième approche consiste en une étude statistique basée sur la perception humaine afin d'évaluer le système de reconnaissance ainsi que le descripteur de mouvement proposé. Cela permet d'estimer

la capacité de notre système à pouvoir classifier et analyser les émotions comme un humain. Dans cette partie deux tâches sont réalisées avec les deux classifieurs (la méthode d'apprentissage RDF qui a donné les meilleurs résultats dans la deuxième approche et le classifieur humain) : la classification des émotions et l'étude de l'importance des caractéristiques de mouvement pour discriminer chaque émotion.

Mots clés : gestes corporels expressifs, émotions, modèle LMA, apprentissage supervisé, reconnaissance des gestes, interaction homme-robot.

Abstract

In this thesis, we deal with the problem of gesture recognition in a human-robot interaction context. New contributions are being made on this subject. Our system consists in recognizing human gestures based on a motion analysis method that describes movement in a precise way. As part of this study, a higher level module is integrated to recognize the emotions of the person through the movement of her body. Three approaches are carried out : the first deals with the recognition of dynamic gestures by applying the hidden Markov model (HMM) as a classification method. A local motion descriptor is implemented based on a motion analysis method, called LMA (Laban Movement Analysis), which describes the movement of the person in its different aspects. Our system is invariant to the initial positions and orientations of people. A sampling algorithm has been developed in order to reduce the size of our descriptor and also adapt the data to hidden Markov models. A contribution is made to HMMs to analyze the movement in two directions (its natural and opposite directions) and thus improve the classification of similar gestures. Several experiments are done using public action databases, as well as our database composed of control gestures. In the second approach, an expressive gestures recognition system is set up to recognize the emotions of people through their gestures. A second contribution consists of the choice of a global motion descriptor based on the local characteristics proposed in the first approach to describe the entire gesture. The LMA Effort component is quantified to describe the expressiveness of the gesture with its four factors (space, time, weight and flow). The classification of expressive gestures is carried out with four well-known machine learning methods (random decision forests, multilayer perceptron, support vector machines : one-against-one and one-against-all. A comparative study is made between these 4 methods in order to choose the best one. The approach is validated with public databases and our database of expressive gestures. The third approach is a statistical study based on human perception to evaluate the recognition system as well as the proposed motion descriptor. This allows us to estimate the ability of our system to classify and analyze emotions as a human. In this part, two tasks are carried out with the two classifiers (the RDF learning method that gave the best results in the second approach and the human classifier) : the classification of emotions and the

study of the importance of our motion features to discriminate each emotion.

Key words : expressive body gestures, emotions, LMA model, supervised learning, gestures recognition, human-robot interaction.

Table des matières

Table des figures **5**

Liste des tableaux **11**

1 Introduction générale **13**

1.1 Contexte général . . . 13

1.1.1 Reconnaissance des gestes . . . 14

1.1.2 Interaction sociale Homme-Robot . . . 14

1.2 Motivation . . . 15

1.3 Contributions et publications . . . 16

1.3.1 Contributions . . . 16

1.3.2 Liste des publications . . . 17

1.4 Organisation de la thèse . . . 18

2 État de l'art **21**

2.1 Le langage gestuel . . . 22

2.1.1 Définition d'un geste . . . 22

2.1.2 Système de reconnaissance des gestes . . . 23

2.2 Le langage émotionnel . . . 42

2.2.1 Définition d'une émotion . . . 42

2.2.2 Émotions exprimées par les paroles . . . 43

2.2.3 Émotions exprimées par le visage . . . 44

2.2.4 Émotions exprimées par le corps . . . 45

2.3 Modèle LMA . . . 47

2.4 Notre approche . . . 53

3 Reconnaissance des gestes dynamiques par les modèles de Markov cachés 57

3.1 Construction de CMKinect-10 . . . 58
3.1.1 Description de la base CMKinect-10 . . . 58
3.1.2 Acquisition des données . . . 59
3.1.3 Normalisation . . . 61
3.2 Descripteur local inspiré de LMA . . . 62
3.2.1 Composante Corps . . . 63
3.2.2 Composante Espace . . . 65
3.2.3 Composante Forme . . . 68
3.3 Application des MMCs . . . 72
3.3.1 Formalisme de MMC . . . 72
3.3.2 Topologies . . . 73
3.3.3 Échantillonnage . . . 74
3.3.4 Quantification vectorielle . . . 75
3.3.5 Entrainement et classification des gestes . . . 77
3.3.6 Contribution aux modèles de Markov cachés . . . 78
3.4 Évaluation expérimentale . . . 80
3.4.1 MSRC-12 . . . 81
3.4.2 MSR Action 3D . . . 84
3.4.3 UTKinect . . . 88
3.4.4 CMKinect-10 . . . 89
3.5 Bilan . . . 90

4 Reconnaissance des gestes expressifs 93

4.1 Construction de ECMXsens-5 . . . 94
4.1.1 Description de la base ECMXsens-5 . . . 94
4.2 Descripteur expressif . . . 98
4.2.1 Relation Effort-Forme . . . 98
4.2.2 Descripteur global . . . 99
4.2.3 Composante Effort . . . 101
4.3 Évaluation du descripteur . . . 107
4.3.1 MSRC12 . . . 108
4.3.2 MSR Action 3D . . . 114
4.3.3 UTKinect . . . 115

4.3.4 ECMXsens-5 . . . 116
4.4 Bilan . . . 118

5 Caractérisation des gestes expressifs et évaluation avec la perception humaine 121
5.1 Approche RDF . . . 122
5.1.1 Reconnaissance des gestes expressifs avec la méthode RDF . . . 122
5.1.2 Sélection de caractéristiques pertinentes avec la méthode RDF . . . 132
5.2 Approche humaine . . . 137
5.2.1 Reconnaissance des gestes expressifs avec la perception humaine . . . 137
5.2.2 Sélection des caractéristiques avec l'approche humaine . . . 142
5.3 Évaluation du système . . . 145
5.4 Bilan . . . 146

6 Conclusion générale et perspectives 147
6.1 Conclusion . . . 147
6.2 Perspectives . . . 148

Bibliographie 149

A Les modèles de Markov cachés 167
A.1 Définition d'un MMC . . . 167
A.1.1 Les hypothèses dans la théorie des MMC . . . 167
A.1.2 Les trois problèmes fondamentaux d'un MMC . . . 168

B Les machines à vecteurs de support 175

C Perceptron multicouches 181

D Les forêts d'arbres décisionnels 185

Table des figures

1.1 Les projets robotiques. 16

2.1 Capteur Stéréo. 24
2.2 Capteur temps de vol (TOF). 25
2.3 Capteur Kinect. 26
2.4 Détecteur de coin 3D de [Laptev and Lindeberg, 2003]. 27
2.5 Détecteur Cuboïde de [Dollar et al., 2005a]. 28
2.6 Principe de calcul du descripteur HOG3D [Klaser et al., 2008]. 28
2.7 Principe de construction du descripteur HOG/HOF : (a) Un cube autour du STIP est divisé en une grille de cellules, (b) un HOG et HOF sont calculés pour chaque cellule [Laptev et al., 2008a]. 29
2.8 Les différentes étapes pour calculer le descripteur HON4D. 30
2.9 Des descripteur HOG extraits à partir de la carte de profondeur de mouvement de chaque vue de projection sont combinés en tant que DMM-HOG, qui est utilisé pour représenter l'ensemble des séquences vidéo [Yang et al., 2012]. 31
2.10 Des exemples des AMI pour 5 actions sélectionnées de la base des données Weizmann [Blank et al., 2005]. 33
2.11 (a) Le référentiel des coordonnées de HOJ3D, (b) Système de coordonnées sphérique [Xia et al., 2012a]. 34
2.12 Reconnaissance des actions en utilisant une hiérarchie temporelle des descripteurs de covariance sur les positions 3D des articulations. 35
2.13 Les caractéristiques Eigenjoints développées par [Yang and Tian, 2012] 36
2.14 Illustration du mouvement du squelette pour des gestes de la base MSRC-12. 39
2.15 Exemples d'actions de la base MSR Action 3D. 41
2.16 Illustration du mouvement de la silhouette en 3D pour les gestes de la base MSR Action 3D. 41

2.17 Images extraites de la base UTkinect. 42
2.18 Modèle Circumplex de l'affect. 43
2.19 4 exemples d'un visage dégouté apparait dans 4 contextes différents : (a) dégouté, (b) en colère, (c) triste et (d) effrayé). 45
2.20 La plateforme Eyesweb. 46
2.21 Éditeur graphique d'Effort. 49
2.22 Variation dans le geste de pointage pour les cinq traits du modèle OCEAN. 50
2.23 Interface utilisateur graphique pour la récupération du mouvement. 51
2.24 La première rangée affiche quelques images d'une séquence de mouvement de danse et les autres correspondent aux mouvements les plus similaires. 51
2.25 La première rangée montre une danse contemporaine capturée. La deuxième rangée montre une danse «plus heureuse», tandis que la dernière correspond à une danse «plus triste». 52
2.26 Enfants avec TSA jouant avec le robot NAO. 55

3.1 Le processus de notre système de reconnaissance de gestes. 59
3.2 Suivi et visualisation du squelette sous l'interface RVIZ. 60
3.3 Des échantillons d'images de chaque geste de la base CMKinect-10. 60
3.4 Les systèmes des coordonnées liés au capteur Kinect et au squelette. 61
3.5 Normalisation des deux squelettes exécutants le même mouvement «Marcher» avec deux positions initiales différentes. 62
3.6 Les caractéristiques de la composante Corps. 64
3.7 Variation des caractéristiques θ_2^l (courbe bleue) et θ_2^r (courbe orange) dans le geste «Avancer» de la base CMKinect-10. 65
3.8 Variation des courbes de la distance entre les deux mains d_{Hs} dans le geste «faire un signe» (courbe orange) et le geste «s'arrêter» (courbe bleue). 66
3.9 Variation des courbes des caractéristiques $d_{h,lh}$ et $d_{h,rh}$ dans le geste «faire un signe» ($d_{h,rh}$ courbe bleue et $d_{h,lh}$ courbe orange) et le geste «s'arrêter» ($d_{h,rh}$ courbe violette et $d_{h,lh}$ courbe jaune). 66
3.10 Variation des valeurs des angles θ_4^l (courbe bleue) et θ_4^r (courbe orange) dans le geste «s'asseoir». 67
3.11 Variation de la direction du torse suivant les axes X (courbe bleue), Y (courbe orange) et Z (courbe jaune). 68

3.12 Variation du volume de l'enveloppe convexe du squelette dans le geste «démarrer la musique» de la base MSRC-12. 69
3.13 Courbures locales entre deux images successives. 70
3.14 Les trois plans : (A) horizontal, (B) frontal et (C) sagittal. 70
3.15 Variation des distances relatives à l'articulation du torse dans le geste «s'accroupir» de la base MSRC-12. 71
3.16 Relation de dépendance entre les états cachés et les observations. 73
3.17 Les différentes étapes appliquées à chaque séquence gestuelle avant son implication dans le modèle MMC. 75
3.18 Discrétisation des caractéristiques dans les gestes de la base CMKinect-10. 76
3.19 Un MMC de Bakis linéaire à 4 états. 77
3.20 Modélisation des gestes dans le sens direct et indirect. 80
3.21 Les taux de reconnaissance des gestes iconiques pour chaque valeur de K et S. . . . 82
3.22 Comparaison entre les taux de reconnaissance des gestes iconiques obtenu avec MMC basique et MMC modifié. 83
3.23 Les taux de reconnaissance des gestes métaphoriques pour chaque valeur de K et S. 83
3.24 Comparaison entre les taux de reconnaissance des gestes métaphoriques obtenu avec MMC basique et MMC modifié pour $(K = 40, S = 20)$. 84
3.25 Taux de reconnaissance pour les gestes de AS1 pour chaque valeur de K et S. 85
3.26 Taux de reconnaissance pour les gestes de AS2 pour chaque valeur de K et S. 86
3.27 Taux de reconnaissance pour les gestes de AS3 pour chaque valeur de K et S. 87
3.28 Comparaison entre les taux de reconnaissance des gestes AS1, AS2 et AS3 obtenus avec MMC basique et MMC modifié. 88
3.29 Taux de reconnaissance pour les gestes de la base UTkinect en variant le nombre des groupes K de 10 à 50 et fixant S à 5. 89
3.30 Taux de reconnaissance pour les gestes de contrôle de la base CMKinect-10 pour chaque valeur de K et S. 90
3.31 Comparaison entre les taux de reconnaissance des gestes de la base CMKinect-10 obtenu avec MMC basique et MMC modifié pour $(K = 40, S = 20)$. 90

4.1 La base de données ECMXsens-5, les gestes de haut vers le bas sont : danser, avancer, s'arrêter, faire un signe et pointer. 95
4.2 Le geste de danse effectué avec deux émotions différentes. 97
4.3 Le capteur MVN Awinda de Xsens. 98

4.4 Variation des distances entre les positions des articulations P_t^k à chaque trame et la position du torse J_1^s à l'instant initial (t=1) 101
4.5 Les facteurs de la composante Effort (Espace, Temps, Poids et Flux). 102
4.6 Variation de la vitesse (v_l) de la main gauche dans le geste "Avancer" avec l'état neutre (courbe bleue) et en colère (courbe orange). 105
4.7 Variation de la vitesse (v_r) de la main droite dans le geste "Pointer" avec l'état Neutre (courbe bleue) et de la Joie (courbe orange). 105
4.8 Variation de l'accélération (a_l) de la main gauche dans le geste "Avancer" avec l'état triste (courbe bleue) et en colère (courbe orange). 107
4.9 Les résultats de F-scores moyens de la méthode RDF dans la base MSRC-12. 110
4.10 Les résultats de F-scores moyens de la méthode SVM dans la base MSRC-12. 112
4.11 Les résultats de F-scores moyens de la méthode MLP dans la base MSRC-12. 113
4.12 Comparaison entre les 4 méthodes d'apprentissage dans la base MSRC-12. 113
4.13 Matrices de confusions de la base MSRC-12. 114
4.14 Les résultats de F-scores moyens des méthodes (a) RDF, (b) OAO, (c) OAA et (d) MLP dans la base MSR Action 3D. 114
4.15 Matrices de confusion de la base MSR 3D Action. 115
4.16 Comparaison entre les 4 méthodes d'apprentissage dans la base MSR Action3D. 116
4.17 Comparaison entre les 4 méthodes d'apprentissage dans la base UTkinect. 116
4.18 Matrice de confusion de la base UTKinect avec la méthode RDF. 117
4.19 Comparaison entre les 4 méthodes d'apprentissage dans la base ECMXsens-5. 117
4.20 La matrice de confusion de la base ECMXsens-5. 118

5.1 Motivation Statistique. 123
5.2 Motivation de Calcul. 124
5.3 Motivation Représentative. 124
5.4 Combinaison séquentielle. 125
5.5 Combinaison parallèle. 125
5.6 Combinaison hybride. 126
5.7 Principe de Bagging. 127
5.8 Principe de la sélection de caractéristiques. 128
5.9 Variation du taux d'erreur OOB ($errOOB$) en fonction de (T, c_{max}). 131

5.10 Matrice de confusion des gestes expressifs, 5 gestes (D danser, A avancer, S Faire un signe, Sa S'arrêter et P pointer) effectués avec 4 états (H heureux, C en colère, T triste et N neutre). . . . 131
5.11 Matrices de confusions entre les émotions exprimées (dans les lignes) et les émotions perçues (dans les colonnes) pour chaque geste avec la méthode RDF. . . . 132
5.12 Approche Filter. . . . 133
5.13 Approche Wrapper. . . . 134
5.14 Variation du taux d'erreur OOB en fonction des caractéristiques sélectionnées. . . . 137
5.15 Reproduction des gestes avec un avatar. . . . 140
5.16 Fiabilité inter-observateurs dans la perception des émotions à l'aide du coefficient de Cronbach. . . . 140
5.17 Matrices des confusions entre les émotions exprimées (les lignes) et les émotions perçues (les colonnes) pour chaque geste en se basant sur les scores donnés par les observateurs. . . . 141
5.18 Les scores moyens de la perception des émotions par 10 observateurs. . . . 142
5.19 Fiabilité inter-observateurs dans l'évaluation des facteurs des composantes Effort-Forme avec la mesure de coefficient de Cronbach. . . . 143

A.1 Algorithme de Forward. . . . 170
A.2 Algorithme de backward. . . . 171
A.3 L'algorithme de Baum-Welch. . . . 174

B.1 Hyperplan séparateur de la méthode SVM. . . . 176
B.2 Transformation de l'espace d'entrée en un espace de re-description. . . . 178
B.3 Classification multi-classes par la méthode Un-Contre-Tous. . . . 179
B.4 Classification multi-classes par la méthode Un-Contre-Un. . . . 180

C.1 Schéma d'un neurone artificiel. . . . 182

D.1 Un arbre de décision. . . . 186

Liste des tableaux

2.1 Les classes des gestes iconiques et métaphoriques de la base MSRC-12 avec leur nombre des répétitions (Nrep). 38
2.2 Les classes des gestes de la base MSR Action 3D et leurs répétitions (Nrep). 40
2.3 Les quatre composants de LMA avec leurs facteurs. 48

3.1 Comparaison avec les méthodes de l'état de l'art sur la base MSRC-12. 84
3.2 Comparaison avec les méthodes de l'état de l'art sur la base MSR Action 3D. 88

4.1 Les huit actions élémentaires d'Effort. 99
4.2 Ajustement des paramètres de RDF. 110
4.3 Ajustement des paramètres de SVM. 111
4.4 Ajustement des paramètres de MLP. 112

5.1 Résultats de la reconnaissance des émotions dans les différents gestes de notre base. . 132
5.2 Les caractéristiques pertinentes pour chaque émotion à travers les différents gestes. . 138
5.3 Coefficients de Pearson r pour la corrélation entre les scores donnés aux facteurs de Effort-Forme et ceux donnés aux émotions exprimées (**. la corrélation est significative au niveau 0.001). 144

Chapitre 1

Introduction générale

Sommaire

1.1 Contexte général 13

1.1.1 Reconnaissance des gestes 14

1.1.2 Interaction sociale Homme-Robot 14

1.2 Motivation 15

1.3 Contributions et publications 16

1.3.1 Contributions 16

1.3.2 Liste des publications 17

1.4 Organisation de la thèse 18

1.1 Contexte général

De nos jours, les machines et, en particulier, les ordinateurs et les robots sont devenus une partie importante de notre environnement. Ces dernières supportent nos communications, accomplissent des tâches difficiles pour nous et, au final, elles sont censées être nos assistants. A ce titre, les interactions avec ces machines apparaissent comme un problème clé, qui doit être traité soigneusement. En effet, une machine est acceptée et utile si (1) elle est facile à programmer et à contrôler, (2) réalise les résultats attendus. Les deux demandes sont fortement liées au naturel et à l'intuitivité de la communication entre cette machine et les humains. En s'inspirant de l'interaction homme-homme, le language corporel est toujours omniprésent et indispensable, particulièrement les gestes. Nous utilisons les gestes sans en avoir conscience, soit pour manipuler des objets, soit pour communiquer, transmettre des messages, exprimer des émotions, etc. Dans ce contexte, l'objectif de notre thèse consiste à développer une interaction naturelle entre l'homme et le robot NAO via un système de

reconnaissance de gestes.

1.1.1 Reconnaissance des gestes

L'analyse et la reconnaissance automatique du mouvement humain à partir d'une entrée visuelle est l'un des domaines de recherche en vision par ordinateur le plus traité. Ce n'est pas seulement dû à ses défis scientifiques passionnants, mais aussi en raison de son importance pratique. Des centaines d'applications potentielles existantes dans le besoin urgent de cette technologie incluent le contrôle, la navigation et la manipulation dans des environnements réels et virtuels, l'interaction homme-robot, les systèmes de télé présence, le diagnostic clinique et le suivi, l'assistance aux personnes âgées, l'apprentissage des applications et des jeux, les systèmes d'ingénierie et la conception assistée par ordinateur, l'annotation et l'indexage automatique des vidéos, l'identification médico-légale et la détection de mensonges, etc. Dans le cadre de notre projet où notre tâche consiste à doter le robot humanoïde NAO de la capacité de reconnaitre des mouvements des personnes, nous abordons le problème de la reconnaissance des actions humaines à partir de flux vidéo. Cette dernière a des atouts intéressants comme le caractère non-invasif et le faible coût. De plus, la vidéo a un contenu sémantiquement très riche et pertinent pour la reconnaissance des actions humaines. Dans ce contexte, nous abordons le sujet des gestes expressifs afin de rendre la communication plus efficace et de lui conférer un haut niveau d'impact.

1.1.2 Interaction sociale Homme-Robot

Les robots sont principalement utilisés dans l'industrie, mais avec le développement de la technologie, ils se sont de plus en plus introduits dans notre vie quotidienne. Ces robots, destinés à être directement mis en rapport avec les hommes, doivent pouvoir interagir de manière intuitive afin d'être acceptés par l'homme. Cela implique que le robot doit être capable de percevoir son environnement et d'agir en conséquence d'une manière humainement acceptable et accueillante. Traditionnellement, des robots autonomes ont été employés dans des zones où peu ou pas d'interaction est requis, tels que ; l'industrie automobile, l'inspection des puits de pétrole, les opérations de recherche et d'exploration d'environnements hostile à l'homme. Ces robots sont généralement télécommandés et supervisés par un opérateur humain. Des ajouts récents de robots de service dans les environnements domestiques, comme les robots aspirateurs, ont augmenté leur contact avec une personne commune, cependant, aucune interaction de haut niveau n'est impliquée avec l'homme. Les robots de service évoluent maintenant pour acquérir des rôles plus importants, tels que des assistants, des travailleurs hospitaliers ou des aides-soignants aux personnes âgés, où l'interaction sociale est une

partie importante de l'activité du robot. La présence de tels robots dans notre vie quotidienne est importante pour certaines personnes notamment celles à mobilité réduite qui nécessitent un assistant capable d'acquérir et d'interpréter l'information et ainsi de la transmettre. Aujourd'hui, nous trouvons un certain nombre de projets réalisés pour assurer une interaction sociale entre l'homme et le robot, nous pouvons citer celui de ARMAR (Stiefelhagen et al., 2004) réalisé dans l'Institut de Technologie de Karlsruhe. Ce projet consiste à une reconnaissance de la parole de l'utilisateur et une perception visuelle de ses gestes de pointage et de l'orientation de la tête. La combinaison des deux canaux vidéo et audio permet une analyse multimodale du discours qui mène à une interprétation de plus haut niveau, telle que l'analyse de l'intention de l'utilisateur dans une situation de dialogue avec le robot (Voir Figure1.1(a)). Nous citons aussi le projet HUMAVIPS réalisé à l'INRIA qui consiste à doter le robot NAO de capacités audiovisuelles (AV) : exploration, reconnaissance et interaction, de sorte qu'il présente un comportement adéquat lorsqu'il s'agit d'un groupe de personnes (Voir Figure 1.1(b)). Il y a aussi le projet ROMEO qui consiste à créer un robot humanoide compagnon et assistant personnel (Voir Figure 1.1(c)).

1.2 Motivation

Une hypothèse largement répandue dans le domaine de l'interaction homme-robot (IHR) est qu'une bonne interaction avec un robot doit refléter une communication la plus naturelle possible. Pour cela on cherche toujours à s'inspirer de la communication homme-homme qui demande comme modalités explicites la parole et le geste. Mais dans la communication entre les hommes il existe aussi un autre canal, implicite, au moyen duquel on transmet une information sur les individus, généralement sous forme d'émotions. Ce canal implicite est une caractéristique essentielle de la communication humaine. Par conséquent, si on souhaite obtenir une interaction réellement effective des robots avec les hommes, il faudra aussi aborder ce type de communication. Donc, l'objectif de notre travail consiste à développer un système de reconnaissance de gestes humains, tout en considérant ses mouvements et aussi ses humeurs. Bien que le contenu vidéo soit assez informatif, la tâche de reconnaissance de l'action humaine reste problématique. En effet, de grandes variations de style peuvent être observées dans la reproduction de la même action selon plusieurs facteurs. De plus, il faut prendre en compte l'ambiguïté inter-classes puisqu'il existe des actions similaires telles que les actions "courir lentement" (jogging) et "marcher". Dans notre travail, nous nous inspirons du modèle d'analyse de mouvement de Laban (LMA) pour la représentation des mouvements humains. Ce modèle a été utilisé dans plusieurs applications à des fins de recherche différentes, dans la danse [Aristidou et al., 2017b],

(a) ARMAR. (b) HUMAVIPS.

(c) ROMEO.

FIGURE 1.1 – Les projets robotiques.

la musique [Truong et al., 2016], la robotique [Masuda and Kato, 2010, Knight et al., 2016, Sharma et al., 2013, Knight and Simmons, 2014, Kim et al., 2012, Kim et al., 2013, Nishimura et al., 2012], etc. Dans notre travail, nous nous basons sur les composantes de ce modèle (Corps, Espace, Forme et Effort) pour décrire l'aspect quantitatif et qualitatif du geste.

1.3 Contributions et publications

1.3.1 Contributions

Notre travail de thèse vise à construire un système de reconnaissance des gestes qui soit adapté aux conditions réelles. Il comprend les contributions suivantes :

1. Nous avons développé un système de reconnaissance de gestes dédié au contrôle du robot NAO. Cela a demandé la création d'une base de données composée de gestes de contrôle. Le mouvement de la personne a été représenté par un descripteur local inspiré de la méthode

d'analyse de mouvement LMA. Nous avons appliqué les modèles de Markov cachés discrets pour l'entrainement et la classification des gestes. Un algorithme d'échantillonnage et une méthode de quantification sont implémentés pour adapter les données aux modèle MMC discret. Une contribution aux modèles MMCs est proposée pour résoudre les problèmes de classification touchant les mouvements similaires. Finalement, une évaluation du système avec des bases publiques et notre base de contrôle est réalisée [Ajili et al., 2017a, Ajili et al., 2018a].

2. Nous avons développé un système de reconnaissance de gestes expressifs. Ce dernier est constitué d'une base des données composée de 5 gestes expressifs, interprétés avec 4 émotions (joie, colère, tristesse et neutre). Le descripteur de mouvement local présenté dans le chapitre précédent a été utilisé et modifié en fonction des mesures globales pour décrire l'entièreté du geste expressif. Avec l'intégration de la composante Effort, responsable de la description de l'expressivité du mouvement, nous créons un descripteur de mouvement expressif global. Une étude comparative est réalisée entre 4 méthodes d'apprentissage automatique (forêts d'arbres décisionnels, perceptron multicouches, et deux approches de machines à vecteurs de support multi-classes (un-contre-un et un-contre-tous)). Le système est évalué avec des bases publiques et notre base de données expressive [Ajili et al., 2017b, Ajili et al., 2018a].

3. Nous avons proposé deux approches différentes (approche d'apprentissage automatique et approche humaine) pour la reconnaissance et l'analyse des gestes expressifs. Dans la première, la méthode des forêts d'arbres décisionnels a été appliquée pour l'entrainement et la classification des gestes expressifs. Un algorithme basé sur cette méthode a été développé afin d'estimer l'importance des différentes caractéristiques de notre descripteur envers la caractérisation de chaque émotion. Une corrélation a été établie entre le descripteur de mouvement et les émotions. Les mêmes tâches ont été répétées avec une approche statistique basée sur des scores donnés par des observateurs afin d'évaluer les émotions perçues et aussi l'importance de chaque indice de mouvement. En se référant aux résultats obtenus par l'approche humaine, nous avons conclu en la robustesse de notre système de reconnaissance et l'adéquation de notre descripteur de mouvement [Ajili et al., 2018b].

1.3.2 Liste des publications

— [Ajili et al., 2016] I. Ajili and M. Mallem and J. Y. Didier. Gesture Recognition for Robot Teleoperation. Journée AFRV, Réalité Virtuelle, Augmentée, Mixte et Interaction 3D, Brest, 2016.

— [Ajili et al., 2017b] I. Ajili and M. Mallem and J. Y. Didier. Robust human action recognition

system using Laban Movement Analysis. 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2017), Sep 2017, Marseille, France. (elec. proc.), 2017.

— [Ajili et al., 2017a] I. Ajili and M. Mallem and J. Y. Didier. Gesture recognition for humanoid robot teleoperation. 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbonne, 2017.

— [Ajili et al., 2018a] I. Ajili and M. Mallem and J. Y. Didier. An Efficient Motion Recognition System Based on LMA Technique and a Discrete Hidden Markov Model. 20th International Conference on Image Analysis and Processing, Paris, 2018.

— [Ajili et al., 2018b] I. Ajili and M. Mallem and J. Y. Didier. Relevant LMA Features for Human Motion Recognition. 20th International Conference on Image Analysis and Processing, Paris, 2018, **Best Paper Award**.

— I. Ajili and M. Mallem and J. Y. Didier. Expressive motions recognition and analysis with learning and statistical methods. Journal Multimedia Tools and Applications, 2018.

— I. Ajili and M. Mallem and J. Y. Didier. Human Motions and Emotions Recognition Using Laban Movement Analysis Technique. Journal The Visual Computer, 2018.

1.4 Organisation de la thèse

Cette thèse est organisée de la manière suivante :

— Le chapitre 1 présente le contexte général, la motivation de ce travail, et les contributions avec les différentes publications réalisées dans ces 3 années.

— Le chapitre 2 fournit un état de l'art divisé en 2 grandes parties : la première sur le langage gestuel où nous présentons les modules d'un système de reconnaissance de gestes en général et les descripteurs de mouvements proposés. La deuxième partie est sur le langage émotionnel, où nous définissons le thème "émotion" et les différentes modalités proposées dans l'état de l'art pour exprimer l'émotion, et une partie finale sur les gestes expressifs.

— Le chapitre 3 fournit une première implémentation du système de reconnaissance de gestes dédié à l'interaction homme-robot via des gestes. Les modèles de Markov cachés sont utilisés pour l'entrainement et la classification des gestes. Notre modèle est évalué sur notre base composée de gestes de contrôles et 3 bases publiques.

— Le chapitre 4 présente un système de reconnaissance de gestes expressifs. Un descripteur de mouvement expressif est construit. Une évaluation comparative entre 4 méthodes d'apprentissage est réalisée. Le modèle est évalué sur des bases publiques et notre base composée de

gestes expressifs.

— Le chapitre 5 présente deux approches différentes, la première est basée sur l'apprentissage automatique et la deuxième sur la perception humaine. La comparaison entre les deux approches en termes de classification des émotions et étude de l'importance des caractéristiques du descripteur proposé permet d'évaluer la fiabilité de notre système.

— Le chapitre 6 conclut cette thèse et fournit les différentes perspectives permettant de mettre en valeur ce système dans des applications réelles (interaction homme-robot).

Chapitre 2

État de l'art

Sommaire

2.1 Le langage gestuel 22

2.1.1 Définition d'un geste 22

2.1.2 Système de reconnaissance des gestes 23

2.2 Le langage émotionnel 42

2.2.1 Définition d'une émotion 42

2.2.2 Émotions exprimées par les paroles 43

2.2.3 Émotions exprimées par le visage 44

2.2.4 Émotions exprimées par le corps 45

2.3 Modèle LMA 47

2.4 Notre approche 53

Le problème de la reconnaissance des gestes consiste en général à décrire le contenu de la séquence dans le but de comprendre ce qui se passe dans le document vidéo. Cette information de haut niveau permet par la suite d'identifier une vidéo dans une base de données. La problématique consiste donc à extraire au cours du temps des caractéristiques relatives au mouvement appelées descripteurs de mouvement. Il existe deux types de descripteur, les descripteurs bas niveau basés sur le contenu brut de l'image (la couleur, les points d'intérêt, la texture, etc) et les descripteurs haut niveau orientés vers la sémantique du contenu de la scène pour décrire les messages transférés par les gestes ou l'état émotionnel exprimé par les personnes lors de la réalisation du mouvement. Ce chapitre contient trois parties principales. La première sur le langage gestuel définit le principe d'un système de reconnaissance de gestes avec ses différents modules. Nous présentons les différents descripteurs de mouvement utilisés dans le cadre de la reconnaissance des actions ainsi que les bases des actions publiques construites. La deuxième partie est sur le langage émotionnel où les différentes

modalités utilisées pour l'analyse et la reconnaissance des émotions sont définies. Par la suite, nous focalisons sur l'aspect des gestes expressifs qui sera notre sujet de thèse. Dans la troisième partie, nous introduisons le modèle d'analyse des mouvement (LMA) que nous avons utilisé dans notre travail pour interpréter et décrire les gestes expressifs.

2.1 Le langage gestuel

2.1.1 Définition d'un geste

Un geste peut être défini comme le mouvement élémentaire des parties du corps d'une personne. Il fait partie de la communication non verbale utilisée à la place ou en combinaison avec une communication verbale, pour exprimer un message particulier. L'interprétation des gestes est une tâche complexe, principalement en raison du fait que la signification gestuelle implique un contexte culturel. Par exemple, si un signe de tête indique généralement un accord, dans certains pays (comme en Grèce) un seul signe de tête indique un refus. Le pointage avec un doigt étendu est un geste commun aux États-Unis et en Europe, mais il est considéré comme un geste grossier et offensif en Asie. Chez les moines cisterciens, lorsqu'une personne approche son doigt à coté de son œil et le dirige par la suite à l'œil de son interlocuteur, ou s'il se touche le cœur puis dirige sa main en direction de l'autre, cela est un signe de suspicion. Contrairement, en Indie ces signes signifient la paix. Tout cela implique que l'interprétation sémantique d'un geste dépend strictement de la culture donnée. Le geste peut aussi être classifié en geste dynamique et statique. Ce dernier est également appelé posture, correspond à la configuration du corps ou d'une partie du corps à un instant fixe [Sharma and Verma, 2015] tandis que le geste dynamique [Barros et al., 2017] correspond à une succession continue des postures. Ceci n'est pas la façon unique de classifier les gestes, comme en témoigne la littérature :

- Dans la classification de Cadoz [Cadoz, 1994] les gestes sont divisés selon les fonctions gestuelles en 3 groupes :
  - Gestes épistémiques : gestes pour la perception avec le toucher.
  - Gestes ergotiques : gestes agissent sur le monde, peut manipuler, modifier ou transformer le monde physique.
  - Gestes sémiotiques : gestes d'expression pour communiquer des informations et émettre des messages à destination de l'environnement.
- Dans la classification de Kendon [Kendon, 2004] les gestes sont classifiés sous forme d'un continuum qui distingue entre des gestes accompagnant la parole (désignés sous le terme gesticulation) et ceux qui en sont indépendants (autonomes).

— Dans la classification de McNeill [McNeill, 1992], basée sur le continuum de Kendon, les gestes sont divisés en 4 groupes :
  — Gestes iconiques : gestes qui illustrent des concepts concrets.
  — Gestes métaphoriques : gestes représentent des concepts abstraits et des métaphores.
  — Gestes déictiques : gestes de pointage en direction d'un référent.
  — Battements : gestes rythmant la parole en accentuant les éléments importants, sans contenu sémantique.

À partir des définitions suivantes, nous plaçons nos travaux dans le cadre de la reconnaissance des gestes avec des contenus sémantiques. Spécifiquement nous allons parler des gestes expressifs, c'est à dire des gestes réalisés avec différentes émotions.

2.1.2 Système de reconnaissance des gestes

La reconnaissance des gestes humains est un processus qui consiste à identifier et interpréter automatiquement les mouvements humains. Trois étapes fondamentales dans un système de reconnaissance des gestes sont : l'**acquisition des données** qui consiste à extraire des informations numériques grâce à un système de capture de mouvement. L'**extraction des caractéristiques** pour convertir les données brutes en une représentation pertinente du mouvement. L'objectif de cette étape est d'extraire les caractéristiques utiles et compactes qui représentent les mouvements d'une manière le plus fiable possible. Ces caractéristiques constituent un vecteur descripteur spécifique à chaque geste. Finalement, ce vecteur est chargé dans le modèle d'**apprentissage** pour l'entrainement et/ou la classification suivant l'application demandée.

Les capteurs

Les premiers capteurs apparus dans le domaine de la vision ont été basés sur les images RGB 2D [Schuldt et al., 2004, Laptev et al., 2008b]. Cependant, elles ne fournissent uniquement que les informations d'apparence des objets de la scène. Avec ces informations limitées, il est extrêmement difficile, voire impossible, de résoudre certains problèmes tels que la séparation du premier plan et d'arrière-plan ayant des couleurs et des textures similaires. Par conséquent, les études récentes ont tendance à utiliser une nouvelle information qui est la profondeur. Cette information résout la question de l'inférence 3D. Pour cela, ces dernières années, ce domaine a connu une forte présence des capteurs 3D qui ont amélioré la performance de la reconnaissance des gestes. Il existe plusieurs techniques pour capturer la profondeur d'une scène. Certaines se basent sur le principe de la mise en correspondance qui consiste à trouver les points qui se correspondent entre les images droite et

gauche. Nous pouvons citer l'exemple des appareils stéréoscopique (Voir Figure 2.1). Leur popularité s'explique par l'analogie avec le système visuel humain et la disponibilité de caméras couleur à bas prix. [Song and Takatsuka, 2005] ont utilisé ce capteur pour le suivi des gestes. Ils ont détecté l'extrémité du doigt de l'utilisateur dans les deux images de cette paire stéréo. Dans ces images les deux points sur lesquels cette extrémité apparaît établissent une correspondance stéréo, qui est utilisée pour évaluer la position du doigt dans l'espace 3D. À son tour, cette position est utilisée par le système pour estimer la distance du doigt par rapport à la table augmentée et par conséquent, déterminer si l'utilisateur est en contact avec elle ou pas. Ces informations stéréoscopiques ont été aussi utilisées par [Igorevich et al., 2013] afin de suivre et reconnaitre les gestes des deux mains, tout en résolvant le problème du bruit horizontal généré par la caméra en raison de son hypersensibilité à la lumière. Cependant, les systèmes de vision stéréoscopiques ne peuvent calculer la profondeur de la scène que pour un ensemble restreint de pixels correspondant à des zones de la scène ayant une forte structure locale. Pour cela une nouvelle famille de caméra 3D est apparue, la caméra temps

FIGURE 2.1 – Capteur Stéréo.

de vol (TOF) présentée dans la Figure 2.2. Il s'agit d'un capteur actif qui fournit des images de profondeur en temps réel en s'appuyant sur la mesure du temps de vol. Son principe est similaire à celui d'un scanner par laser. Il consiste à illuminer la scène et les objets mesurés par un éclair de lumière, et calculer le temps que cet éclair prend pour effectuer le trajet entre l'objet et la caméra. Le temps de vol de cet éclair est directement proportionnel à la distance entre la caméra et l'objet mesuré. Cette mesure de temps de vol est effectuée indépendamment par chaque pixel de la caméra, permettant ainsi d'obtenir une image complète en 3D de l'objet mesuré. Dans ce cadre, [Holte et al., 2008] ont utilisé la caméra SwissRanger pour l'acquisition des gestes des bras. Le mouvement a été détecté par la différence entre deux gammes d'images et a été filtré par un filtre passe-bande. Leur descripteur de mouvement a été basé sur le contexte de forme afin d'assurer l'invariance de la rotation. La corrélation entre les différentes représentations des contextes de formes a été réalisée dans la phase de la reconnaissance des gestes. De même [Breuer et al., 2007] se sont basés sur la caméra TOF pour l'acquisition des gestes des mains. Les données sont transformées en des nuages de points après une phase de filtrage de bruit. Le principe d'analyse en composantes

principales a été appliqué pour avoir une première estimation de la position et l'orientation de la main. Un principe d'appariement a été réalisé afin de minimiser la distance entre le modèle et le nuage de point. [Droeschel et al., 2011] ont utilisé ce capteur pour la reconnaissance des gestes de pointage. Ils ont extrait trois informations pour la représentation des gestes (distance entre la tête et la main, l'angle entre le bras et l'axe vertical du corps et la vitesse de la main). Ils ont appliqué la méthode de la régression des processus gaussiens (GPR [Rasmussen and Nickisch, 2010]) pour modéliser une approximation d'une fonction qui associe les caractéristiques du corps extraites vers une direction de pointage. L'avantage de la caméra TOF, est la rapidité d'acquisition des images, vu que chaque pixel de la caméra livre indépendamment une mesure de la distance. En contre partie, ce type de caméra génère un nuage avec un grand nombre de points. Cela peut exiger un espace de stockage de grande taille sur les matériels utilisés. Par ailleurs, des limitations physiques liées à la taille du capteur peuvent entraîner des images de profondeur de faible résolution à une faible portée où à un bruit de mesure important. D'où l'apparition d'un deuxième capteur de profondeur qui est la kinect, cette caméra comme le TOF fournit la profondeur de l'image mais avec une résolution différente. En effet les caméras TOF ont encore une résolution limitée (200×200) tandis que la kinect présente une résolution VGA (640×480). Les données RGB-D réalisent une représentation

FIGURE 2.2 – Capteur temps de vol (TOF).

très utile d'une scène d'intérieur pour résoudre les problèmes fondamentaux en vision par ordinateur. Le capteur Kinect prend les avantages de l'image couleur qui fournit des informations d'apparence d'un objet et l'image de profondeur qui est immunisée contre les variations de couleur, d'éclairage, d'angle de rotation et d'échelle. Cependant, avec la sortie de ce capteur à faible coût, l'acquisition des données RGB-D devient moins chère et plus facile.

Capteur de profondeur Kinect

La Kinect est une évolution très compétente d'une caméra classique couplée à une Xbox 360 permettant à l'utilisateur de contrôler des jeux vidéo sans l'utilisation de manettes, mais seulement

avec les mouvements du corps. Elle a été conçue par Microsoft en septembre 2008. C'est un dispositif adapté à la reconstruction 3D de scènes en intérieur. Il intègre trois composants (Voir Figure 2.3) :

— Un microphone destiné à la commande vocale permettant la fonctionnalité de la reconnaissance vocale.

— Une caméra RGB fournissant une image couleur avec une résolution de 640×480 pixels à une fréquence moyenne de 30 trames par secondes.

— Un capteur de profondeur 3D : Un émetteur infra-rouge (laser) et une caméra infra-rouge permettant de calculer la profondeur d'un objet par rapport à la caméra.

FIGURE 2.3 – Capteur Kinect.

Outils spécifiques à la Kinect

OpenNI est un framework open source capable de s'interfacer avec différents types de matériel et de concevoir des applications pour des interactions naturelles. Ce Framework est accompagné d'un open source «libfreenect » pour la Kinect. Ce pilote fournit toutes les fonctions de base de la Kinect (profondeur, Infrarouge, RGB) et permet à ce capteur d'être utilisé avec Linux et sous ROS. Les algorithmes de traitement disponibles sous ce framework sont :

- Analyse du corps : récupération de la pose du squelette en déterminant la position et l'orientation de ses articulations.
- Analyse de la scène : extraction de l'avant-plan, segmentation du sol, etc.
- Détection des gestes et suivi des mouvements des mains.

Extraction des caractéristiques

La reconnaissance d'action humaine est un des sujets majeurs du domaine de la vision par ordinateur. Cela est dû à la grande variété d'applications potentielles, telles que la surveillance vidéo, l'analyse de contenu vidéo, l'entraînement sportif, la surveillance de la santé, l'analyse du comportement commercial, l'interaction Homme-Machine, la robotique, etc. Cependant, ce processus connait de véritables difficultés en raison de la forte variabilité des personnes en apparence et en mouvement. Par conséquent, il est essentiel d'extraire des représentations robustes à ces variations.

Plusieurs types de descripteurs ont été proposés dans la littérature de la vision par ordinateur. Nous avons choisi de les subdiviser en deux types : les descripteurs basés sur les caractéristiques locales (les points d'intérêt) et les descripteurs basés sur l'apparence (silhouette et squelette 3D).

Méthodes basées sur des caractéristiques locales Les premières approches dans l'état de l'art ont utilisé des méthodes connues dans le domaine de la reconnaissance d'image. Ces méthodes se basent sur des caractéristiques locales qui correspondent généralement à un ensemble de pixels présentant une singularité, que ce soit au niveau du gradient ou du contour. Ces méthodes utilisent des détecteurs de points d'intérêt dans des images afin de les représenter comme des collections d'éléments d'intérêt. Les points d'intérêt spatio-temporels (STIP) sont définis comme les points dans l'image où apparaissent un changement significatif dans le temps et dans l'espace. C'est une extension des points d'intérêts spatiaux (noté SIP pour "Spatial Interest Points"). Ils sont identifiés à partir des maxima locaux d'une fonction réponse qui caractérise le signal spatio-temporel. L'un des premiers travaux proposés pour extraire les points d'intérêts spatio-temporels (STIPs) est celui de [Laptev and Lindeberg, 2003]. Ils ont proposé le détecteur de point d'intérêt Harris3D (Voir Figure 2.4), qui présente une extension temporelle du détecteur de coin Harris2D [Harris and Stephens, 1988] afin de reconnaître les actions humaines. Il détecte les points dont le voisinage local est soumis à une variation spatiale et temporelle significative. Cependant, le nombre

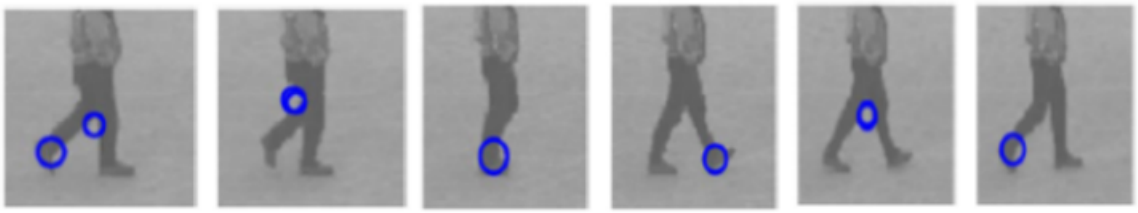


FIGURE 2.4 – Détecteur de coin 3D de [Laptev and Lindeberg, 2003].

des points d'intérêt vérifiant le critère Harris3D est relativement faible par rapport aux zones contenant des mouvements significatifs. [Dollar et al., 2005a] ont traité ce problème avec un nouveau détecteur spécialement conçu pour les mouvements périodiques locaux présents dans la vidéo (Voir Figure 2.5). Ils ont proposé des nouveaux points d'intérêt spatio-temporels plus denses. Leur détecteur est basé sur des filtres Gaussiens 2D appliqué à la dimension spatiale et des filtres de Gabor 1D appliqué à la dimension temporelle.

— Le filtre gaussien effectue une sélection d'échelle spatiale (σ) en lissant chaque image.

— Le filtre passe-bande Gabor donne des réponses élevées aux variations périodiques du signal.

Les auteurs ont concaténé les gradients calculés pour chaque pixel dans une région cuboïde en un

seul vecteur. Finalement, la méthode d'analyse en composantes principales (ACP) a été appliquée pour la projection des vecteurs sur un espace de dimension plus faible. Ce détecteur, nommé Cuboïde, a été appliqué initialement pour la reconnaissance des mouvements d'un animal et après pour les actions et les expressions faciales des personnes. [Klaser et al., 2008] ont proposé un nouveau

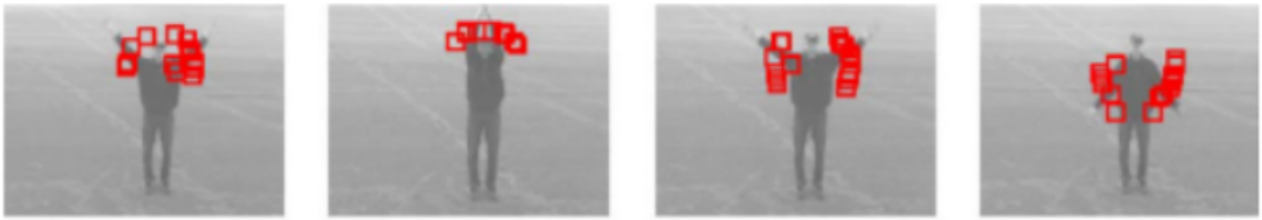


FIGURE 2.5 – Détecteur Cuboïde de [Dollar et al., 2005a].

descripteur local pour les séquences vidéo, le HOG3D. C'est une extension des histogrammes de gradients orientés (HOG) proposé par [Dalal and Triggs, 2005] dans le domaine spatio-temporel. Les gradients sont calculés à l'aide d'une représentation de vidéo intégrale. Des polyèdres réguliers sont utilisés pour quantifier uniformément l'orientation des gradients spatio-temporels. Leur descripteur combine les informations de forme et de mouvement en même temps. La Figure 2.6 montre le principe de calcul du descripteur HOG3D. [Laptev et al., 2008a] ont également intro-

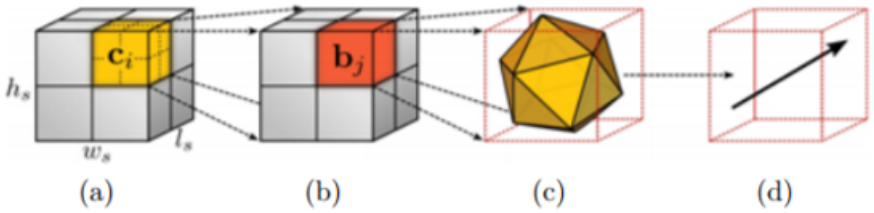


FIGURE 2.6 – Principe de calcul du descripteur HOG3D [Klaser et al., 2008].

duit les histogrammes de gradients orientés HOG combinés avec les histogrammes de flux optique (HOF) afin de caractériser le mouvement local et l'apparence. Pour caractériser le mouvement et l'apparence des caractéristiques locales, ils ont calculé les descripteurs d'histogramme des volumes spatio-temporels au voisinage des points détectés. Chaque volume est subdivisé en une grille de cellules $n_x \times n_y \times n_t$. Pour chaque cellule l'histogramme à 4 composantes d'orientations du gradient et à 5 composantes d'orientation de flux optique sont calculés. Les histogrammes normalisés sont concaténés dans les descripteurs finaux HOG et HOF. La Figure 2.7 montre le principe de construction du descripteur HOG/HOF. [Willems et al., 2008] ont étendu la caractéristique Hessienne 2D dans le domaine spatio-temporel en appliquant un filtre gaussien 3D. Ils ont calculé le déterminant de la matrice Hessienne afin de déterminer les points d'intérêts. [Chakraborty et al., 2012] ont pro-

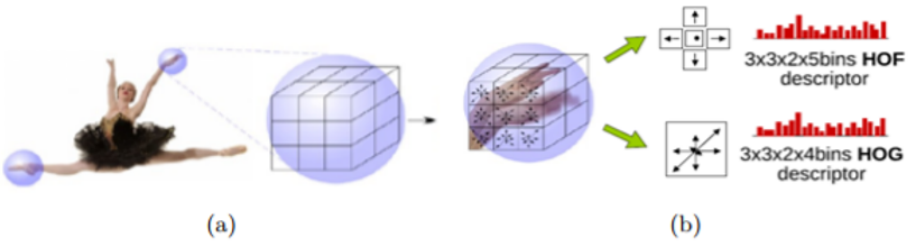


FIGURE 2.7 – Principe de construction du descripteur HOG/HOF : (a) Un cube autour du STIP est divisé en une grille de cellules, (b) un HOG et HOF sont calculés pour chaque cellule [Laptev et al., 2008a].

posé un détecteur sélectif de points d'intérêt spatio-temporels pour la reconnaissance des actions humaines. Des contraintes spatio-temporelles locales sont imposées pour obtenir un ensemble final de points d'intérêt plus robuste, tout en supprimant les points d'intérêt indésirables. La méthode des machines à vecteurs de support (SVM) a été utilisée pour la classification et la reconnaissance des actions. [Yan and Luo, 2012] ont proposé l'histogramme de l'algorithme de localisation de points d'intérêt (HIPL) en tant que complément du descripteur de sac des points d'intérêt (BIP) afin de capturer l'information spatiale de STIP. Dans leur approche, la méthode de boosting Adaboost et la représentation éparse (SR) ont été utilisées avec le classifieur de sortie pondéré (WOC) pour réaliser une meilleure classification de l'ensemble des caractéristiques. L'AdaBoost est un algorithme de boosting qui repose sur la sélection itérative de classifieur faible en fonction d'une distribution des exemples d'apprentissage. Chaque exemple est pondéré en fonction de sa difficulté avec le classifieur courant. Cet algorithme permet d'améliorer les performances du modèle de classification. La représentation éparse a été appliquée afin de gérer les grandes variations intra-classes des actions humaines. WOC est un framework de fusion des caractéristiques multiples qui consiste à exploiter le potentiel de chaque caractéristique et utiliser des poids pour combiner des classifieurs faibles entrainés par un type unique de caractéristique. Cependant, le modèle HIPL ne fournit aucune information temporelle des données dans la vidéo, comme la vitesse. Pour cela, ce système est incapable de discriminer correctement les classes d'action qui sont très proches les unes des autres, comme par exemple courir et faire du jogging. [Cao et al., 2010] ont traité le problème de la détection des actions à partir de vidéos encombrées. Pour la détection des actions, ils ont combiné des caractéristiques multiples, qui peuvent être basées sur les mouvements (par exemple, l'historique des mouvements, le flux optique) ou sur l'apparence (par exemple, bord, couleur). En plus de cela, ils ont utilisé des caractéristiques hétérogènes telles que le champ de mouvement filtré hiérarchique (HFMF) [Tian et al., 2012], des caractéristiques éparses [Dollar et al., 2005b], des histogrammes de gradient orienté et de flux optique (HOG/HOF) [Laptev et al., 2008a]. Ils ont employé les modèles de mélange gaussien (GMM)

pour modéliser et combiner ces caractéristiques hétérogènes. La probabilité d'un vecteur de caractéristiques donné est estimée en se basant sur le modèle GMM. [Oreifej and Liu, 2013] ont présenté un nouveau descripteur de reconnaissance d'activité à partir des vidéos acquises par un capteur de profondeur. Ce descripteur présente un histogramme capturant la distribution de l'orientation normale de la surface dans l'espace 4D, du temps, de la profondeur, et des coordonnées spatiales. Comme montré à la Figure 2.8, pour la construction de cet histogramme, ils ont créé des projecteurs 4D, qui quantifient l'espace 4D et représentent les directions possibles pour la normale 4D. Les projecteurs sont initialisés en utilisant les sommets d'un polychore régulier. Par conséquent, la quantification est affinée en utilisant une nouvelle mesure de densité discriminante, de sorte que des projecteurs supplémentaires sont induits dans les directions où les normales 4D sont plus denses et discriminantes. Finalement, la méthode SVM a été adoptée pour l'entrainement et la classification des actions. [Yang et al., 2012] ont utilisé les histogrammes de gradients orientés (HOG) calculés à

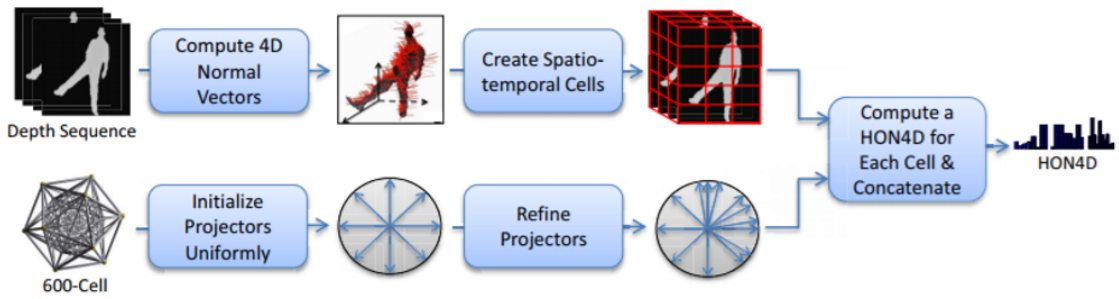


FIGURE 2.8 – Les différentes étapes pour calculer le descripteur HON4D.

partir des cartes de profondeur de mouvement (DMM), comme une représentation d'une séquence d'action. Ils ont projeté chaque carte de profondeur sur trois plans orthogonaux prédéfinis. Chaque carte projetée a été normalisée et une carte binaire a été générée présentant son énergie de mouvement en calculant la différence entre deux cartes consécutives. Les cartes binaires sont ensuite empilées pour obtenir le DMM pour chaque vue projective. L'histogramme des gradients orientés est ensuite appliqué à la carte DMM pour extraire les caractéristiques de chaque vue. La concaténation des descripteurs HOG à partir des trois vues forment l'ensemble des descripteurs DMM-HOG qui présentent les entrées d'un classificateur linéaire SVM pour la reconnaissance des actions. Une illustration des étapes de l'extraction de HOG à partir de DMM est présentée dans la Figure 2.9.

Méthodes basées apparence D'autres chercheurs ont choisi d'incorporer des modèles de personnes tels que des silhouettes ou des squelettes pour la reconnaissance de l'action.

Méthodes basées silhouette : Ces méthodes utilisent des silhouettes comme entrée pour le système de reconnaissance d'actions humaines. Un volume spatio-temporel 3D (STV) est formé en

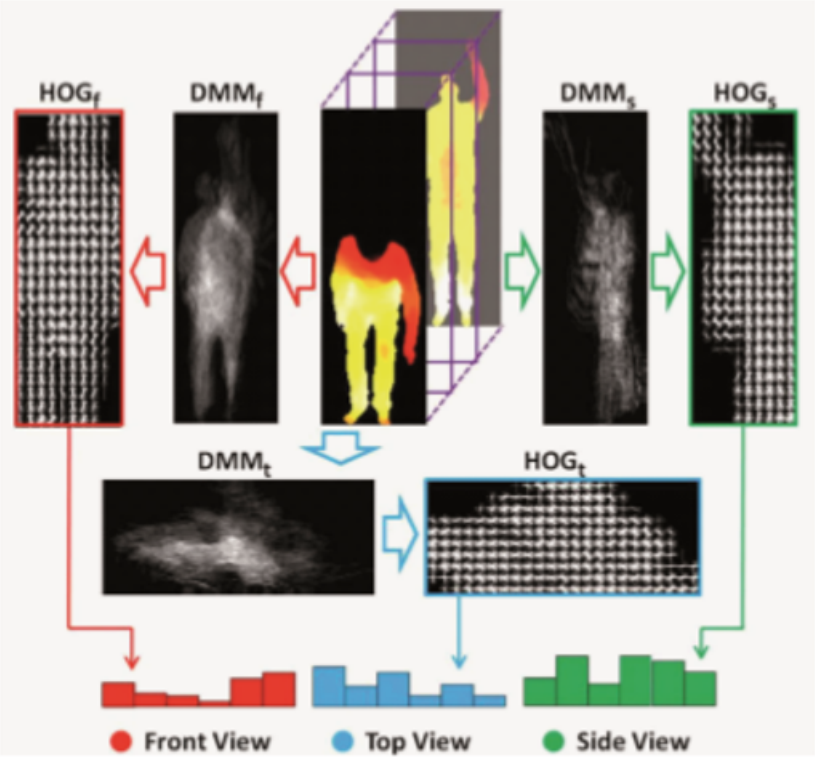


FIGURE 2.9 – Des descripteur HOG extraits à partir de la carte de profondeur de mouvement de chaque vue de projection sont combinés en tant que DMM-HOG, qui est utilisé pour représenter l'ensemble des séquences vidéo [Yang et al., 2012].

empilant des images au cours d'une séquence donnée. Une localisation précise, un alignement et éventuellement une soustraction d'un arrière-plan sont nécessaires. Weinland et al. [Weinland et al., 2006] ont introduit de nouveaux descripteurs de mouvement (MHV) basés sur des volumes d'historique de mouvement qui est une extension de MHI en 3D [Bobick and Davis, 2001]. Ils ont transformé les MHVs calculés en coordonnées cylindriques autour de l'axe vertical et ont extrait des caractéristiques invariantes dans l'espace de Fourier. Afin d'évaluer leur méthode, les auteurs ont construit une base de donnée, appelée IXMAS, composée des actions capturées à partir des points de vue différents. Les résultats sur cette base indiquent que cette représentation peut être utilisée pour apprendre et reconnaître les classes des actions indépendamment du sexe, de la taille du corps et du point de vue. Les même auteurs [Weinland and Boyer, 2008] ont proposé d'utiliser un ensemble de silhouette représentative comme des modèles, des exemples. Les actions sont représentées par des vecteurs de distances entre les exemples et les images dans la séquence de l'action. En outre, différentes méthodes de sélection des silhouettes de pose clés ont été discutées dans leur travail. [Ahmad and Lee, 2010] ont proposé une représentation spatio-temporelle de silhouette, appelée image d'énergie de silhouette (SEI) pour caractériser les propriétés de la forme et du mouvement pour la reconnaissance de l'ac-

tion humaine. Pour aborder la variabilité dans la reconnaissance des actions, ils ont proposé des modèles adaptables avec plusieurs paramètres, notamment l'anthropométrie de la personne, la vitesse de l'action, la phase (état initial et final d'une action), les observations de la caméra (distance par rapport à la caméra, mouvement oblique et rotation du corps humain) et les variations des vues. [Fang et al., 2009] ont d'abord réduit la dimension des silhouettes en des points de faible dimension en tant que description du mouvement spatial en utilisant l'approche de projection préservant la localité (LPP). Ce vecteur de mouvement obtenu après réduction de dimension a été pris pour décrire la structure du mouvement intrinsèque. Ensuite, trois informations temporelles différentes, le voisin temporel, la différence de mouvement et la trajectoire de mouvement, ont été appliquées aux descripteurs spatiaux pour obtenir les vecteurs caractéristiques, qui ont été les entrées du classifieur k plus proche voisins. Les même auteurs [Tseng et al., 2012] ont développé une autre approche basée sur la silhouette. Ils ont employé la méthode de projection adaptative préservant la localité (ALPP) afin de construire un sous-espace spatio-temporel discriminant. Ensuite, la méthode nommée Non-base Central-Difference Action Vector (NCDAV) a été utilisée pour extraire les données temporelles du sous-espace spatial réduit afin de caractériser l'information de mouvement dans un vecteur temporel. Ceci permet de résoudre le problème des chevauchements dans le sous-espace spatial résultant de l'ambiguïté de la forme du corps humain entre différentes classes d'action. Finalement, la méthode d'apprentissage de métrique de plus proche voisin à grande marge (LMNN) a été appliquée pour construire un sous-espace spatio-temporel discriminant où les vecteurs temporels appartenant à la même classe d'action sont regroupés ensemble alors que ceux associés aux classes différentes sont séparés par une marge. Dans l'étape de classification des actions, les auteurs ont fait appel à l'approche de k plus proches voisins. Cependant, leur solution dépend de la qualité des silhouettes extraites, ce qui rend l'étape de la reconnaissance plus sensible. [Guo et al., 2009] ont considéré une action comme une séquence temporelle des déformations de forme locales du centroïde d'objet silhouette. Chaque action est représentée par une matrice de covariance des vecteurs de caractéristiques géométriques normalisées en 13 dimensions qui capturent la forme du tunnel de silhouette. Le tunnel de silhouette d'une vidéo de test est divisé en des courts segments chevauchants et chaque segment est classé en utilisant un dictionnaire des matrices de covariance d'action étiquetée et la règle du plus proche voisin. [Kim et al., 2010] ont proposé le concept de l'image de mouvement accumulée (AMI) pour représenter les caractéristiques spatio-temporelles des actions. L'AMI a été présentée en fonction des différences d'images. Les actions humaines sont reconnues par le calcul des distances entre les matrices de classement de la vidéo d'action de requête et les matrices de classement de toutes les fenêtres locales dans les vidéos de référence. La Figure 2.10 montre un des exemples des AMI pour 5

actions sélectionnées de la base des données Weizmann [Blank et al., 2005]. [Shao and Chen, 2010]

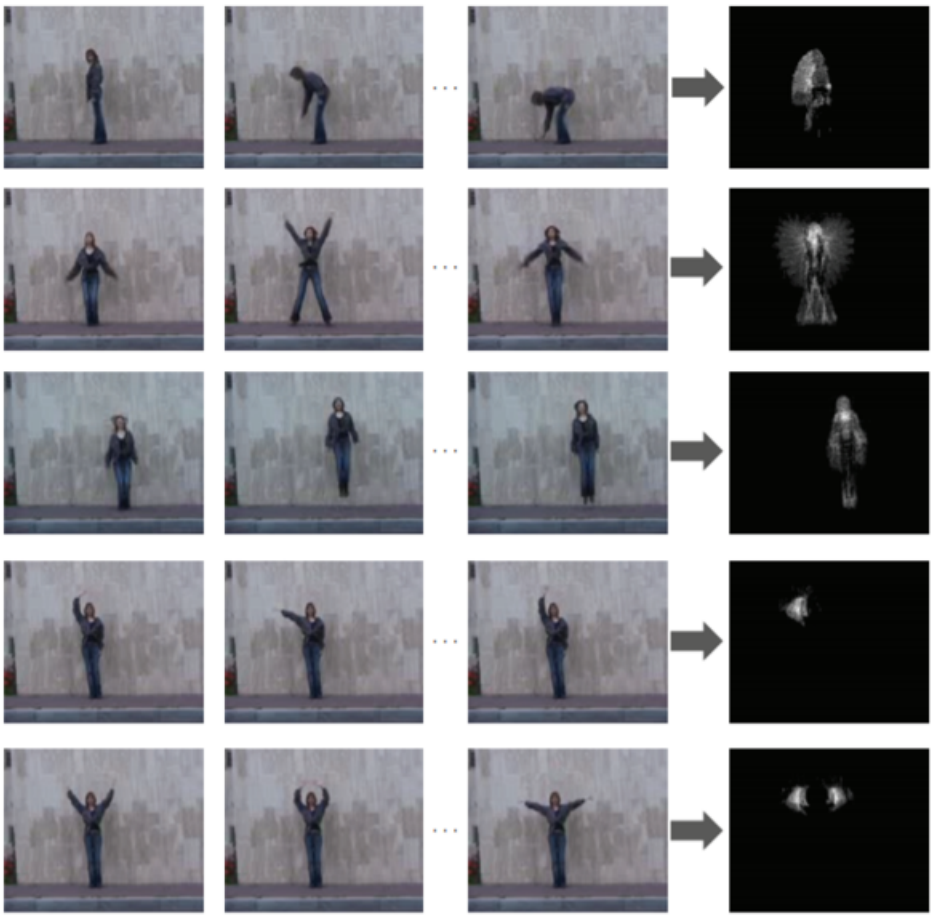


FIGURE 2.10 – Des exemples des AMI pour 5 actions sélectionnées de la base des données Weizmann [Blank et al., 2005].

se sont basés sur la silhouette pour la reconnaissance des actions. Ils ont utilisé l'histogramme des poses corporelles (HBP) pour décrire une séquence d'action humaine. Seulement les données brutes échantillonnées à partir les silhouettes associées aux séquences vidéo sont utilisées pour représenter des poses humaines. Ils ont aussi appliqué l'analyse discriminante par régression spectrale pour projeter chaque silhouette dans un espace de dimension inférieure. [Junejo et al., 2014] ont également proposé des descripteurs à base de silhouettes pour leur système de reconnaissance d'actions. Pour chaque séquence d'action, ils ont d'abord extrait le premier plan de l'image, puis localisé la silhouette dans chaque image. Ensuite, ils ont transformé chaque silhouette en une série temporelle. Finalement, ils ont calculé une approximation agrégée symbolique des séries temporelles, qui consiste à réduire leur taille et les quantifier en un ensemble de symboles. La classification des actions a été réalisée par l'algorithme des forêts aléatoires.

Méthodes basées sur les articulations de squelette : Lorsqu'on s'intéresse à la repré-

sentation du mouvement d'une personne, l'identification de l'acteur peut être un premier pas important. Cependant, la forme brute du corps, telle qu'une silhouette peut la présenter, n'est pas toujours discriminante. Certains chercheurs se sont inspirés des études du chercheur en psychologie Johanson [Johansson, 1973] et se sont focalisés sur les membres du corps de la personne pour la représentation de son mouvement. Ce chercheur a montré dans ses travaux qu'afin de reconnaitre les actions de la personne il suffit juste d'interpréter les trajectoires des membres de son corps. Ainsi, beaucoup de chercheurs ont considéré cette hypothèse dans leur recherches, et ont représenté le modèle humain par un squelette simplifié pour identifier les positions de ses membres. La première solution a été proposée par [Gavrila and Davis, 1995] qui ont récupéré la pose du corps en 3D à partir de plusieurs caméras pour le suivi et la reconnaissance du mouvement humain en 3D. Plus récemment, plusieurs chercheurs se sont appuyés sur le modèle squelette dans leurs travaux afin de construire leurs descripteurs de mouvement à partir des informations sur les articulations du squelette [Pham et al., 2018, Kong et al., 2016, Li et al., 2016, Ghorbel, 2017]. Dans ce contexte, [Xia et al., 2012a] ont présenté une nouvelle approche pour la reconnaissance des actions avec des histogrammes des positions des articulations en 3D (HOJ3D) comme une représentation compacte des postures. Dans cette présentation, l'espace 3D est partitionné en n-bins en utilisant un système de coordonnées sphériques modifié (Voir Figure2.11). Les vecteurs HOJ3D des séquences d'entrainement calculés sont d'abord reprojetés en utilisant la méthode d'analyse discriminante linéaire (ADL), puis partitionnés en k groupes avec la méthode des K-moyennes pour représenter les vecteurs des caractéristiques du mouvement. Finalement, le modèle de Markov caché (MMC) est appliqué pour la phase de classification des actions. [Jiang et al., 2014] ont proposé un modèle

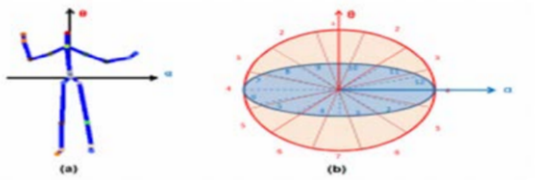


FIGURE 2.11 – (a) Le référentiel des coordonnées de HOJ3D, (b) Système de coordonnées sphérique [Xia et al., 2012a].

hiérarchique pour la reconnaissance des actions qui consiste à associer chaque action à un groupe en se basant sur les états de mouvement de chaque partie du corps. Ils ont divisé le corps en 4 segments (tête, abdomen, bras et jambes). Ensuite pour chaque groupe, un modèle des k plus proche voisins est entrainé. Il prend en entrée deux types de caractéristiques. La première présente le vecteur de mouvement de chaque articulation dans un intervalle de temps précis. La deuxième définit sa position relative par rapport à une articulation stable. Des sacs de mots sont utilisés pour repré-

senter l'ensemble des caractéristiques afin de réduire la taille du descripteur et rendre le système de reconnaissance plus rapide. Une approche de pondération adaptative est proposée afin d'ajuster le poids de chaque mot extrait de l'ensemble des caractéristiques spatio-temporelles et déterminer les mots clés pour la reconnaissance. [Zanfir et al., 2013] ont proposé une nouvelle représentation dynamique qui capture non seulement la pose du corps en 3D, mais également des propriétés différentielles comme la vitesse et l'accélération des articulations du corps humain. La concaténation de ces informations construit le descripteur "Moving Pose". Un algorithme est adopté pour calculer les images de poses mobiles les plus discriminantes. Un schéma de vote basé sur la méthode des k plus proches voisins est utilisé pour la classification des séquences de test. [Hussein et al., 2013a] ont introduit une approche pour la reconnaissance des actions basée sur la matrice de covariance des positions des articulations du squelette. Afin de définir la dépendance temporelle des positions des articulations, plusieurs matrices de covariances ont été déployées sur des sous-séquences de manière hiérarchique afin de déterminer l'ordre d'un mouvement au cours du temps (Voir Figure 2.12). Le classifieur SVM a été utilisé pour l'entrainement et la classification des actions. [Song et al., 2014]

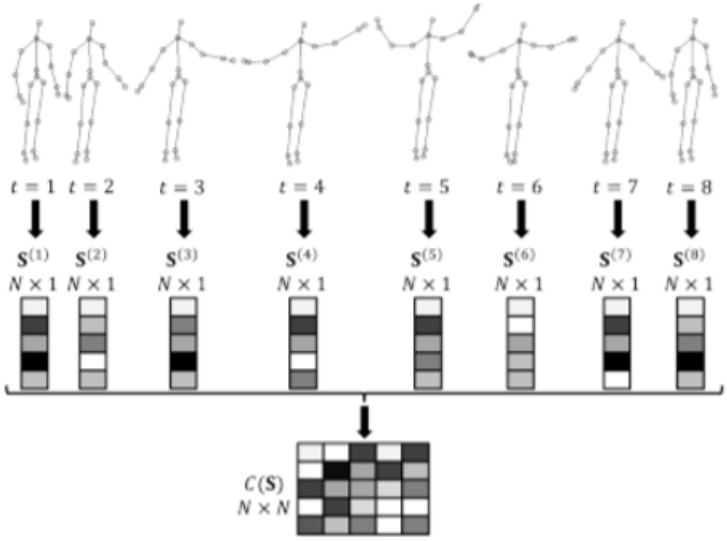


FIGURE 2.12 – Reconnaissance des actions en utilisant une hiérarchie temporelle des descripteurs de covariance sur les positions 3D des articulations.

ont utilisé le nuage de point 3D fourni par la caméra kinect pour représenter la surface externe du corps humain. Ils ont présenté cette surface en décrivant le déplacement relatif des surfaces voisines à partir d'un point de référence défini dans le nuage de point. Un nouveau système de coordonnées cylindrique a été défini pour rendre le système invariant à certaines transformations possibles y compris les translations et les rotations. De plus, trois schémas ont été proposés pour représenter les actions humaines en fonction du nouveau descripteur, y compris :

— le schéma à base de squelette qui définit la différence entre deux actions en calculant le déplacement entre deux postures.

— le schéma de points de référence aléatoires qui consiste à échantillonner un nombre approprié de points pour couvrir le corps tout en évitant les redondances possibles par les points proches.

— Le schéma spatio-temporel qui consiste à coder le descripteur dans le domaine spatio-temporel.

La méthode des k plus proches voisins et l'approche un-contre-tous de la méthode de machine à vecteurs de support (SVM) ont été utilisées pour la classification des actions. [Yang and Tian, 2012] ont proposé un nouveau descripteur en se basant sur les coordonnées 3D des articulations de squelette fournis par le capteur Kinect. Leur descripteur combine les aspects spatial et temporel. Afin de modéliser explicitement le déplacement, leur descripteur nommé EigenJoints, combine trois informations :

— La posture statique d'une pose.

— Le mouvement temporel d'une pose définit par la différence entre la pose actuelle et la pose précédente.

— Le décalage par rapport à une pose initiale, l'offset.

Après la normalisation de ces caractéristiques, la méthode analyse en composantes principales a été appliquée pour avoir un descripteur plus compact. Finalement, l'approche Bayesienne non paramétrique (NBNN) a été utilisée pour la classification des actions. Afin de supprimer les images confuses et réduire le coût de calcul dans la recherche des plus proches voisins, les auteurs ont proposé le concept de l'énergie de mouvement accumulée qui consiste à quantifier le caractère distinctif de chaque image et donc sélectionner les images informatives. La Figure 2.13 illustre le processus utilisé pour obtenir un descripteur EigenJoints. [Pedersoli et al., 2014] ont développé un frame-

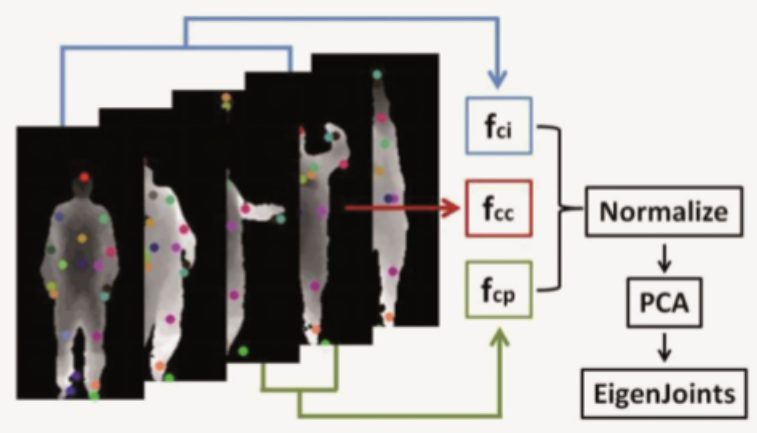


FIGURE 2.13 – Les caractéristiques Eigenjoints développées par [Yang and Tian, 2012]

work open source pour la reconnaissance des poses statiques des mains ainsi que les gestes dynamiques de ces dernières. L'étape de segmentation des mains repose uniquement sur les informations de profondeur fournies par le capteur kinect et réalisée par l'algorithme de segmentation Mean-Shift [Comaniciu and Meer, 2002]. Les caractéristiques de pose de la main sont basées sur le filtrage de Gabor. L'entrainement et la classification sont réalisés avec le modèle radial de la méthode SVM. Pour les gestes dynamiques, la trajectoire du centroïde de la main a été extraite et utilisée comme entrées pour les modèles de Markov cachés dans l'étape de classification. [Evangelidis et al., 2014] ont proposé un descripteur de squelette qui rend la représentation de la pose invariante. Ce descripteur est désigné sous le nom de "skeletal quad". Ils ont utilisé le modèle de mélange gaussien (GMM) et la représentation de noyau de kernel [Jaakkola and Haussler, 1998]. Le GMM est formé à partir de données d'entrainement, de sorte que la génération de tout ensemble de quads est codée par son vecteur Fisher. En outre, une représentation à plusieurs niveaux des vecteurs Fisher conduit à une description d'action qui intègre l'ordre d'exécution des sous-actions dans chaque séquence d'action. Ces vecteurs constituent les entrées d'un SVM linéaire multi-classes pour la classification des actions. [Vemulapalli et al., 2014] ont proposé une nouvelle représentation du squelette qui modélise explicitement les relations géométriques 3D entre diverses parties du corps en utilisant des rotations et des translations dans l'espace 3D. Mathématiquement, les rotations et les translations de corps rigides dans l'espace 3D sont membres du groupe euclidien spécial SE [Murray et al., 1994], qui est un groupe de Lie matriciel. Les actions humaines sont par la suite modélisées comme des courbes dans ce groupe de Lie. Pour faciliter la procédure, ils ont associé les courbes d'action du groupe de Lie à son algèbre de Lie, qui est un espace vectoriel. Finalement pour la classification des actions, une combinaison de la déformation temporelle dynamique, la représentation pyramidale temporelle de Fourier et le SVM linéaire a été réalisée. Donc, comme nous pouvons de voir différents descripteurs ont été proposés pour la représentation et la classification des mouvements. A coté de cela, un nombre croissant de bases des données RGB-D ont été construites pour être utilisées dans l'évaluation de ces algorithmes. L'utilisation des bases de données accessibles au public permet non seulement de gagner du temps et des ressources pour les chercheurs, mais permet également la comparaison équitable des différents systèmes. Chaque base de données s'appuie sur des critères spécifiques, comme le type de geste (métaphorique, iconiques, ...), la complexité et la similitude des gestes, le changement de point de vue, etc. Dans la partie suivante nous allons présenter quelques bases publiques utilisées dans notre thèse.

Les bases de données des actions humaines publiques

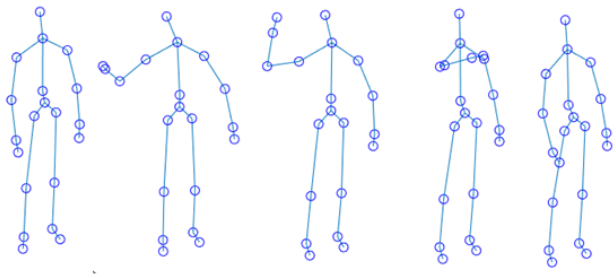
Dans notre travail, nous avons évalué notre système avec 3 bases de données publiques afin de s'assurer de la robustesse de notre approche, qui sont MSRC-12, MSR Action 3D et UTKinect.

MSRC12 : La base de données MSRC12 a été construite par [Fothergill et al., 2012], elle comprend 594 séquences (environ 6 heures et 40 minutes) associées à 12 classes de gestes. Les différentes séquences sont collectées auprès de 30 personnes et divisées en deux catégories : gestes iconiques (s'accroupir ou se cacher, tirer au pistolet, jeter un objet, changer d'arme, donner un coup de pied pour attaquer un ennemi et mettre des lunettes) et gestes métaphoriques (démarrer la musique / augmenter le volume, naviguer vers le menu suivant, terminer la musique, s'incliner pour mettre fin à la session musicale, protester contre la musique et fixer le tempo de la chanson). Chacune des séquences correspond à plusieurs répétitions successives du même geste. Donc une segmentation est nécessaire pour chaque séquence. Le nombre des répétitions (Nrep) est différent dans chaque classe de geste (voir Table 2.1). Au total, il y a 5653 instances de gestes. Les fichiers de mouvement contiennent les coordonnées des 20 articulations estimées à l'aide du capteur Kinect. Les poses sont capturées à une fréquence d'échantillonnage de 30 Hz. Trois modalités d'instructions ont été proposées pour les participants lors de l'enregistrement des gestes : i) Les descriptions de texte, ii) les séquences d'images et iii) les démos vidéo. Il existe également deux combinaisons de ces modalités (image+texte et vidéo+texte).

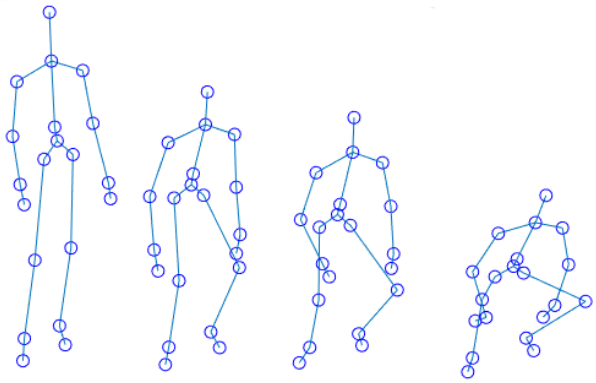
TABLE 2.1 – Les classes des gestes iconiques et métaphoriques de la base MSRC-12 avec leur nombre des répétitions (Nrep).

Gestes iconiques		Gestes métaphorique	
Classe	Nrep	Classe	Nrep
S'accroupir ou se cacher	450	Démarrer la musique	458
Tirer au pistolet	458	Naviguer vers le menu suivant	471
Jeter un objet	462	Terminer la musique	601
Changer d'arme	465	S'incliner	458
Mettre des lunettes	450	Protester contre la musique	458
Donner un coup de pied	454	Remonter le tempo de la chanson	468

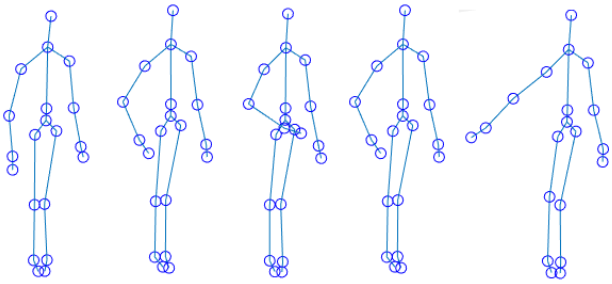
La Figure 2.14 illustre quatre gestes extraits de la base MSRC12, les deux premiers « jeter un objet » et « s'accroupir » font partie de la catégorie iconique et les deux autres « naviguer vers le menu suivant » et « fixer le tempo de la chanson » sont des gestes métaphoriques.



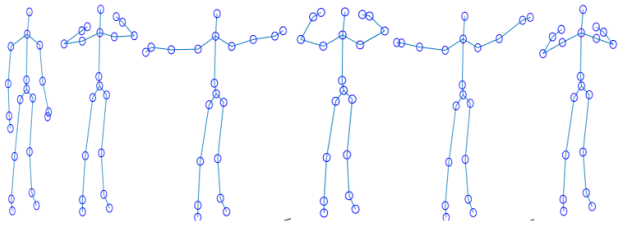
(a) Geste «jeter un objet».



(b) Geste «s'accroupir».



(c) Geste «naviguer vers le menu suivant».



(d) Geste «fixer le tempo de la chanson».

FIGURE 2.14 – Illustration du mouvement du squelette pour des gestes de la base MSRC-12.

MSR Action 3D La base de données MSR Action 3D a été introduite par [Li et al., 2010] et contient 20 actions différentes (Faire un signe vers le haut, faire un signe horizontal, coup de marteau, attraper d'une main, coup de poing vers l'avant, lancer au loin, dessiner un X, dessiner une coche, dessiner un cercle, taper les mains, faire un signe avec les deux mains, boxer sur le côté, se pencher,

coup de pied vers l'avant, coup de pied sur le côté, jogging, swing de tennis, service au tennis, swing de golf et Ramasser et jeter), réalisées par 10 personnes différents. Chaque action est répétée 2 ou 3 fois, au total il y a 567 séquences. Les positions en 3D de 20 articulations ont été capturées à l'aide d'un capteur de profondeur similaire à la caméra Kinect avec une résolution de 640×480 (15 images par seconde). Toutes les vidéos sont enregistrées à partir d'un point de vue fixe et tous les participants étaient en face de la caméra pendant la réalisation des actions. L'arrière-plan a été supprimé de la base de données par un post-traitement. Les données sont fournies en tant qu'échantillons segmentés. La base de données est divisée en 3 groupes, AS1, AS2 et AS3, chacun consistant en 8 actions comme indiqué dans la Table 2.2. Les sous-ensembles AS1 et AS2 étaient destinés à regrouper les actions similaires, tandis que AS3 était destiné à regrouper l'ensemble des actions complexes. La Figure 2.15

TABLE 2.2 – Les classes des gestes de la base MSR Action 3D et leurs répétitions (Nrep).

AS1		**AS2**		**AS3**	
Classe	Nrep	Classe	Nrep	Classe	Nrep
Faire un signe horizontal	27	Faire un signe vers le haut	27	Lancer au loin	26
Coup de marteau	27	Attraper d'une main	26	Coup de pied vers l'avant	30
Coup de poing vers l'avant	26	Dessiner un X	28	Coup de pied sur le coté	20
Lancer au loin	26	Dessiner une coche	30	Jogging	30
Taper les mains	30	Dessiner un cercle	30	Swing de tennis	30
Se pencher	30	Faire un signe avec les deux mains	30	Service au tennis	30
Service au tennis	30	Coup de pied vers l'avant	30	Swing de golf	30
Ramasser et jeter	30	Boxer sur le coté	30	Ramasser et jeter	30

présente quelques échantillons des actions de la base MSR Action 3D. Et la Figure 2.16 illustre les mouvements de la silhouette en 3D pour les gestes «dessiner une coche» et «service au tennis».

UTKinect La base de données UTKinect a été introduite par [Xia et al., 2012b]. Elle contient des vidéos de 10 types d'actions humaines (marcher, s'asseoir, se lever, ramasser, porter, jeter, pousser, tirer, faire un signe avec les deux mains et taper les mains). Chaque action est répétée 2 fois par 10 personnes différents. Les séquences sont enregistrées à l'aide du capteur Kinect avec une fréquence de 30 images par seconde. La résolution de la carte de profondeur est de 320×240 et la résolution de l'image RGB est de 640×480 pixels. Au total, cette base de données contient 200 échantillons d'action. La longueur des échantillons d'actions varie de 5 à 120 images. Quelques échantillons de la base de données sont illustrés dans la Figure 2.17. La difficulté de cette base réside dans la variation

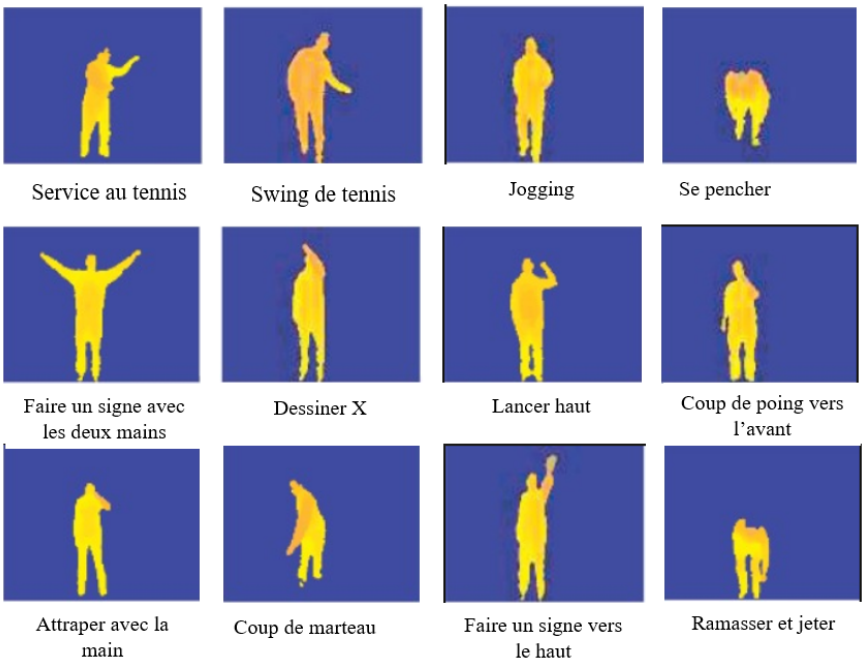
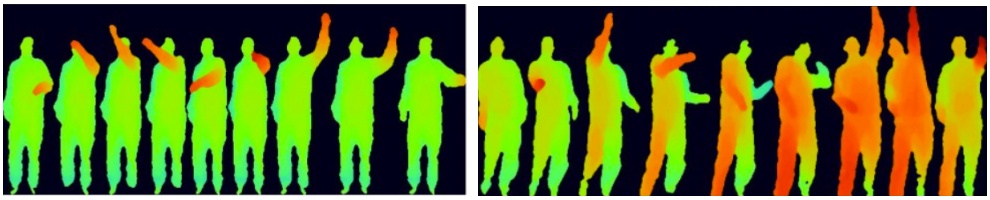


FIGURE 2.15 – Exemples d'actions de la base MSR Action 3D.



(a) «dessiner une coche» (b) «service au tennis»

FIGURE 2.16 – Illustration du mouvement de la silhouette en 3D pour les gestes de la base MSR Action 3D.

du point de vue des personnes. La grande différence entre les durées des vidéos présente une autre contrainte dans cet ensemble de données.

Jusqu'ici, nous considérions les gestes comme des éléments au service de l'interlocuteur et nous négligeons le côté qualitatif d'un mouvement qui présente une information très importante pour reconnaître les gestes de haut niveau et mettre en évidence les qualités expressives des mouvements. En 1921, Wilhelm Wundt [Wundt, 1973] un grand chercheur en psychologie a affirmé que lors d'une communication, les gestes ne trouvent pas leur raison d'être dans la motivation de communiquer un concept, mais bien dans l'émotion ressentie à propos du concept. Ce chercheur considère que le geste ne se réduit plus seulement à ce qu'il signifie dans le discours, il traduit aussi et surtout la personnalité et les sentiments de celui qui parle ; il constitue en lui-même un langage sur les intentions de la personne. Si nous prenons un exemple simple d'un même geste effectué avec deux états différents, par exemple un geste de pointage réalisé avec un état neutre et un état de colère.

FIGURE 2.17 – Images extraites de la base UTkinect.

La différence entre les deux mouvements effectués pour ces deux gestes est quasi-nulle en termes de structure du corps. Les deux reposent sur l'extension du bras mais si nous percevons la dynamique du mouvement, nous trouvons une différence dans l'intention, la rapidité, la force, etc. En outre, lorsque nous effectuons le même mouvement avec des humeurs différentes (heureux, en colère, triste, etc.), il faut un descripteur de plus haut niveau qui réussit à caractériser la qualité du mouvement.

2.2 Le langage émotionnel

2.2.1 Définition d'une émotion

Le mot "émotion" provient du mot français "émouvoir". Il est basé sur le latin emovere, dont e- (variante de ex-) signifie "hors de" et movere signifie "mouvement". Le terme lié "motivation" est également dérivé du mot movere. En général, on peut dire qu'une émotion est une réaction psychologique et physique à une situation. Elle a d'abord une manifestation interne et génère une réaction extérieure. Elle est générée par la confrontation à une situation ainsi qu'à l'interprétation de la réalité. Cependant, l'émotion reste toujours spécifique et propre à chaque individu [Picard, 2003]. A ce jour, il n'y a pas un accord sur ce qu'est une émotion [Scherer, 2005a, Frijda, 2007]. Dans une enquête récente, des experts de renommée internationale dans la recherche d'émotion ont été invités à donner une définition de l'émotion. Comme prévu, il n'y avait effectivement pas de consensus [Izard, 2007]. Par conséquent, plusieurs définitions et rôles ont été donnés à l'émotion (Francois et al., 2001 ; O'Regan, 2003). En 1879, Charles Darwin, fondateur de la théorie de l'évolution, la définit

comme une qualité innée, universelle et communicative, liée au passé de l'évolution de notre espèce. Cependant, il y a quand même un consensus en ce qui concerne le point de vue que les émotions ont plus d'une manifestation psychologique ou comportementale : en plus de sentiments subjectifs, elles contiennent également des tendances à l'action, l'éveil physiologique, l'évaluation cognitive et le comportement expressif [Niedenthal et al., 2006]. Les émotions sont généralement perçues comme l'un des nombreux différents phénomènes affectifs intrinsèques dans l'expérience humaine telle que l'humeur, la position interpersonnelle, l'attitude et les traits de personnalité [Scherer, 2005b]. Tous ces phénomènes affectifs ont le pouvoir de provoquer des changements dans la physiologie humaine. Ils peuvent être distingués suivant un certain nombre de dimensions y compris l'intensité, la durée et le degré de coordination entre différentes modalités. Le modèle de classification des émotions le plus connu est celui de James Russell [Russell, 1980], nommé le modèle Circumplex de l'affect (Voir Figure 2.18). Ce modèle permet de décrire l'émotion dans un espace affectif bidimensionnel représenté par deux axes : un axe horizontal correspond à la dimension de Valence qui permet de distinguer entre émotion positive et négative et un axe vertical correspond à la dimension Arousal qui définit l'intensité de l'émotion et différencie une émotion active d'une émotion passive. L'émotion peut être

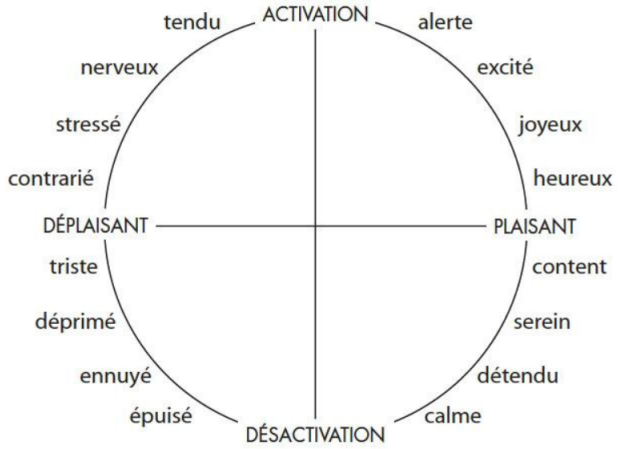


FIGURE 2.18 – Modèle Circumplex de l'affect.

exprimée par différentes modalités, les plus traitées sont : la parole, les expressions faciales et le corps.

2.2.2 Émotions exprimées par les paroles

De nombreuses recherches ont été consacrées à l'étude de la reconnaissance automatique des émotions à travers l'analyse de la parole humaine [Ververidis and Kotropoulos, 2006, Scherer, 2003, Sobol-Shikler and Robinson, 2010]. Certaines de ces recherches ont été appliquées à des centres d'appels, à des systèmes multi-agents ou à d'autres domaines

tels que [Beale and Peter, 2008, Yoon and Park, 2007, Lorini and Schwarzentruber, 2011, Van Deemter et al., 2008, Lorini and Schwarzentruber, 2011]. La reconnaissance de la parole nécessite l'extraction des caractéristiques pertinentes. Les caractéristiques les plus couramment utilisées sont de nature acoustique ou prosodique (énergie vocale, taux de parole, fréquence, durée, etc) et captent donc la qualité de la parole pour identifier l'émotion associée. Cependant, la plupart de ces recherches ne concernent que la reconnaissance dépendante du locuteur. La reconnaissance d'émotion indépendante du locuteur est une question difficile. Dans une enquête menée pour mesurer la performance humaine sur la reconnaissance des émotions, seules 60% des personnes peuvent déterminer correctement les émotions exprimées par des personnes inconnues [Schuller et al., 2006]. Certains auteurs ont montré que cette modalité nécessite l'adaptation du locuteur. Par exemple, [Vogt and André, 2006] ont démontré que l'ajout d'une étape de détection du genre dans le système de la reconnaissance des émotions conduit à de meilleures performances. D'autres auteurs ont souligné l'importance de la normalisation du locuteur pour la reconnaissance des émotions [Sethu et al., 2007, Vlasenko et al., 2007]. D'autres contraintes peuvent introduire dans un système de reconnaissance des émotions via les paroles, comme le cas d'une communication avec des personnes sourdes, d'une conversation à longue distance ou dans un environnement bruyant. Dans de telles situations, le système ne sera pas toujours valable pour transmettre les émotions à travers le canal vocal.

2.2.3 Émotions exprimées par le visage

Pour la communication non verbale, les expressions faciales ont été considérées comme la principale modalité utilisée pour transmettre les émotions [Russell, 1994]. Les expressions faciales d'une émotion sont une conséquence du mouvement des muscles sous la peau de notre visage [Duchenne de Boulogne, 1990]. Le mouvement de ces muscles provoque la déformation de la peau du visage d'une manière qu'un observateur externe peut l'utiliser pour interpréter l'émotion associée. Chaque muscle utilisé pour créer ces constructions faciales est appelé unité d'action (UA). [Ekman and Friesen, 1978] ont identifié les UA responsables pour générer les émotions les plus souvent observées dans la majorité des cultures : la colère, la tristesse, la peur, la surprise, le bonheur et le dégoût. Pour le codage de ces expressions faciales, ils ont développé le système FACS (facial action coding system), un système de codage des mouvements faciaux visibles, pour coder les mouvements du visage et fournir une indication sur le degré et l'intensité d'activation des muscles. Ce système a été largement exploité par plusieurs chercheurs pour la reconnaissance des émotions à travers les expressions faciales. En 1990, les même auteurs [Ekman et al., 1990] ont proposé le sourire

de Duchenne (D) comme une expression spontanée et authentique des émotion positives, comme le bonheur, le plaisir, etc. Il ont classé le sourire (D) comme une combinaison de deux muscles : le muscle zygomatique majeur (AU 12) qui tire les coins des lèvres vers le haut produisant ainsi une bouche souriante et l'orbicularis oculi (AU 42), un muscle situé autour des yeux, pour soulever les joues, rétrécir l'ouverture des yeux et former des rides autour de l'orbite.

Cependant, au cours des interactions sociales, certains chercheurs ont trouvé que la perception émotionnelle à travers le visage peut être influencée par certains facteurs, par exemple la longue distance où les expressions faciales deviennent pas trop claires, ou aussi le facteur de l'âge. Finalement, il y a aussi un autre facteur important qui est le contexte. Pour certains auteurs, la perception de l'émotion à travers les expressions faciales est biaisée dans la direction des expressions corporelles [Aviezer et al., 2008, de Gelder, 2006, H. Aviezer, 2012]. Un exemple est présenté dans la Figure 2.19 où l'expression faciale de l'émotion de "dégout" peut apparaitre différente suivant le contexte du corps : (a) le dégout, (b) la colère, (c) la tristesse et (d) la peur. Une autre modalité est

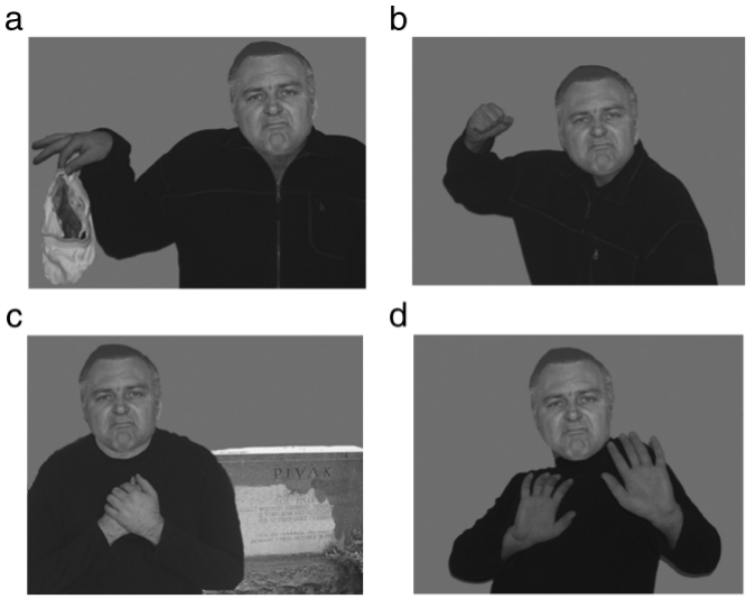


FIGURE 2.19 – 4 exemples d'un visage dégouté apparait dans 4 contextes différents : (a) dégouté, (b) en colère, (c) triste et (d) effrayé).

apparue dans ce domaine qui est le mouvement du corps. Le rôle des gestes dans la perception de l'émotion est devenu d'une grande importance.

2.2.4 Émotions exprimées par le corps

Les chercheurs en psychologie ont été les premiers à s'intéresser aux expressions corporelles selon la posture et le mouvement du corps. Ces derniers soutiennent l'idée que les expressions corpo-

relles "parlent" plus que les expressions faciales [Mehrabian and T. Friar, 1969]. On parle ici, des gestes expressifs, qui sont les gestes effectués dans un état émotionnel. Les gestes expressifs sont appliqués dans plusieurs domaines, notamment dans la danse, la musique, l'animation, etc. Camurri et al. [Camurri et al., 2004b] ont développé un logiciel nommé Eyesweb (Voir Figure 2.20) dans le laboratoire InfoMus de l'université de Gênes pour faciliter l'analyse en temps réel des gestes de danse expressifs. Ils ont identifié des indices des mouvements supposés importants pour la reconnaissance des émotions et ont étudié la façon dont ces indices peuvent être suivis par des techniques de reconnaissance automatique. Ils ont adopté une approche par couches [Leman et al., 2001] pour modéliser les mouvements à partir de mesures physiques de bas niveau (par exemple, position, vitesse, accélération des parties du corps) vers des descripteurs de caractéristiques de mouvement globales (par exemple, fluidité du mouvement, franchise, impulsivité). Ils ont finalement montré comment ces indices participent à transférer quatre émotions (la colère, la peur, le chagrin et la joie) dans une chorégraphie. Sur la base de ces indices de mouvement, les auteurs ont défini un classifieur automatique capable de distinguer les quatre états. Les gestes expressifs sont aussi employés dans l'animation

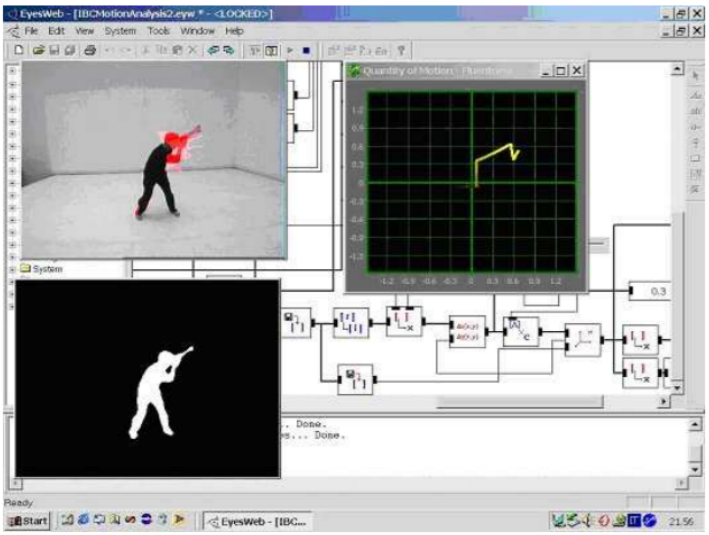


FIGURE 2.20 – La plateforme Eyesweb.

des agents conversationnels [Hartmann et al., 2006]. L'idée ici consiste à modifier les gestes en y ajoutant de l'expressivité afin d'augmenter la crédibilité des Agents et le naturel de leurs comportements. On les trouve dans le domaine du transfert de style qui consiste à transformer une séquence de mouvement en un nouveau style de mouvement tout en gardant son contenu original. Dans ce contexte, [Hsu et al., 2005] ont utilisé la méthode de la déformation temporelle pour transposer la séquence de mouvement d'entrée en une même séquence de mouvement expressive. [Xia et al., 2015] ont conçu un système d'animation pour ajouter différents styles à l'animation existante. Ils ont

construit une série de mélanges locaux de modèles auto-régressifs pour représenter les relations complexes entre les styles et transformer automatiquement une séquence des mouvements hétérogènes non étiquetée en différents styles. [Yumer and Mitra, 2016] ont proposé une approche pour le transfert de style basé sur l'analyse spectrale, qui gère des séquences de mouvement hétérogènes et aussi transfère le style entre des actions indépendantes.

La caractérisation des mouvements en général a dégagé deux niveaux de descripteurs : les descripteurs de bas niveau et ceux de haut niveau. Ces descripteurs se basent sur les informations liées aux articulations, telles que la position 3D, la vitesse, l'accélération, l'angle de rotation, etc. Les descripteurs de haut niveau dépendent du contexte de l'action et nécessitent un formalisme bien défini qui permet de décrire le mouvement tout en considérant son contexte. Dans ce cadre, un modèle est apparu nommé le modèle LMA (Laban Movement Analysis) développé par [Laban and Ullmann, 1971]. Ils s'agit d'une approche utilisée initialement dans le cadre d'étude du mouvement dansé avec une approche centrée sur la qualité du mouvement. Ce modèle permet de décrire le mouvement suivant quatre composantes : Corps, Espace, Forme et Effort. Ce formalisme permet ainsi de savoir comment décrire un mouvement d'une manière complète et avec le minimum nombre suffisent des caractéristiques.

2.3 Modèle LMA

Afin de réaliser une simulation satisfaisante pour le langage complexe du corps humain, une description aussi simple que possible mais complexe si nécessaire du mouvement humain est nécessaire et LMA remplit ces demandes. Cette méthode consiste à décrire, visualiser, interpréter et documenter le mouvement humain. Elle utilise une description multicouche du mouvement, en se concentrant sur ses quatre composantes (Corps, Espace, Forme et Effort). La composante Corps permet de décrire la locomotion humaine. L'Espace permet de décrire la trajectoire effectuée par les parties du corps lors de l'exécution d'un mouvement. La Forme traite le changement de la forme du corps lors d'un mouvement suivant trois facteurs : la mise en forme, le flux de forme et le mouvement directionnel. Finalement, la composante Effort décrit l'expressivité et la qualité du mouvement suivant quatre facteurs, la rectitude (espace), la rapidité (temps), la force (poids) et la fluidité (flux) du mouvement. La Table 2.3 illustre les différents facteurs de la méthode LMA.

— La composante **Corps** détermine ce qui est en mouvement, quelles parties sont connectées, quelles parties sont influencées par d'autres et l'ordre ou le séquençage des mouvements. En se focalisant sur la façon dont le corps est utilisé, il est possible de différencier entre les gestes qui impliquent des parties isolées du corps, les postures qui sont supportées par le corps et

TABLE 2.3 – Les quatre composants de LMA avec leurs facteurs.

LMA			
Corps	**Espace**	**Forme**	**Effort**
		Flux de forme Mvt directionnel Mise en forme	Espace Temps Poids Flux

les actions du corps entier comme les actions : sauter, courir, s'étirer ou se tordre. En ce qui concerne le séquençage du mouvement, l'observateur peut faire la distinction entre un séquençage simultané (deux parties ou plus à la fois), successif (parties de corps adjacentes l'une après l'autre), séquentiel (parties non adjacentes l'une après l'autre) ou unitaire (corps entier).

— La composante **Espace** décrit dans quel espace s'inscrit le mouvement. Laban définit la kinésphère comme un espace imaginaire personnel placé autour de la personne et accessible directement par ses membres jusqu'à l'extrémité des doigts et des pieds tendus dans toutes les directions.

— La composante **Forme** analyse la façon dont le corps change de forme pendant le mouvement. Elle a été développée dans les années 1950-1960 par Rudolf Laban et Warren Agneau pour observer et travailler sur les transformations structurelles du corps humain dans un espace tridimensionnel, lié à soi et à l'environnement. Elle traite les trois questions suivantes :

  — Quelles formes fait le corps ?

  — La forme change-t-elle par rapport à soi ou par rapport à l'environnement ?

  — Comment la forme change-t-elle ?

  Pour répondre à ces trois questions, la catégorie de Forme implique trois qualités distinctes de changement dans la forme de mouvement : flux de forme, mouvement directionnel, et mise en forme. Le *flux de forme* caractérise le changement de la forme du corps. Il peut être décrit par une attitude d'abandon interne qui est intimement liée à la respiration (Se remplir/Se vider). Le *mouvement directionnel* définit le chemin du mouvement dans l'espace, qui peut être rectiligne ou curviligne. Par exemple, pointer ou repousser un objet sont des mouvements linéaires. Par contre, balancer une raquette de tennis ou peindre une clôture présentent des mouvements courbés. La *mise en forme* représente la relation entre le corps en mouvement et l'espace 3D. Les changements de forme dans le mouvement peuvent être décrits en termes de trois dimensions : horizontal, vertical et sagittal. Chacune de ces dimensions est en effet associée à l'un des trois facteurs principaux (largeur, longueur et profondeur) ainsi que l'un des

trois plans (horizontal, vertical et sagittal) liés au corps humain. Les changements de forme dans la dimension horizontale se produisent principalement dans les directions latérales. Les changements dans la dimension verticale se manifestent principalement dans les directions en haut et en bas. Enfin, les changements dans la dimension sagittale sont plus évidents dans la profondeur du corps ou la direction avant-arrière.

— La composante Effort décrit l'expressivité du mouvement suivant ses quatre facteurs :
  — Espace : décrit la directivité du mouvement entre deux qualités (Direct et Indirect).
  — Temps : décrit la rapidité du mouvement entre deux qualités (Soudain et Soutenu).
  — Poids : décrit la force du mouvement entre deux qualités (Fort et Léger).
  — Flux : décrit la fluidité du mouvement entre deux qualités (Lié et Libre).

Le sous-domaine Effort-Forme a reçu un intérêt considérable dans sa façon de décrire les qualités du mouvement. Ce duo permet de décrire la qualité et le rythme du mouvement ainsi que l'expressivité du geste. La méthode LMA a été utilisée dans la littérature pour plusieurs fins, notamment :

— Animation gestuelle et synthèse des gestes expressifs : [Chi et al., 2000] ont développé un système de génération des gestes expressifs appelé modèle EMOTE (Expressive MOTion Engine). Ce système consiste à modifier un mouvement prédéfini afin de produire des gestes expressifs pour les agents virtuels. Ils ont utilisé les composantes Effort-Forme de LMA pour produire des mouvements expressifs de la partie supérieure du corps (torse et bras). Les paramètres de Forme sont appliqués aux bras et au torse, tandis que seuls les bras sont concernés par les qualités d'Effort. La Figure 2.21 représente l'éditeur graphique utilisé pour spécifier les paramètres d'Effort à travers une série d'images clés définies pour les bras. De

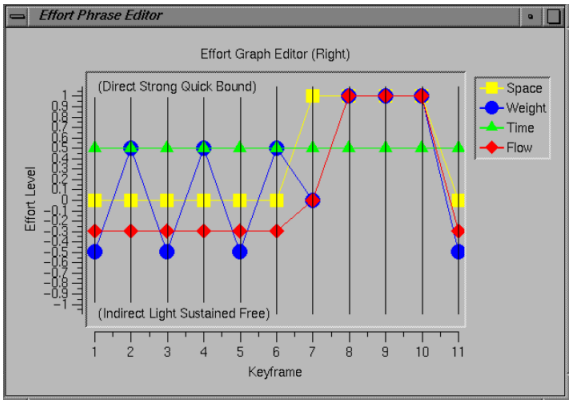


FIGURE 2.21 – Éditeur graphique d'Effort.

même, [Zhao and Badler, 2005] ont utilisé les résultats du modèle EMOTE afin de réaliser une animation gestuelle. Ils ont utilisé un réseau de neurones à une couche cachée avec l'approche de la rétro-propagation du gradient pour détecter les caractéristiques du mouvement

à partir des gestes réalisés et estimer les relations entre ces caractéristiques et les qualités d'Effort. La combinaison de leur système avec le modèle EMOTE permet d'automatiser les processus d'observation et d'analyse humains et de produire des gestes naturels pour des agents de communication à partir de la capture de mouvement 3D. [Durupinar et al., 2016] ont amélioré le modèle EMOTE proposé par [Chi et al., 2000] pour un objectif d'étude de la personnalité de l'Homme à travers le mouvement de son corps. Afin de caractériser les aspects dynamiques du mouvement, les auteurs ont dérivé des mesures physiques des facteurs de la composante Effort. Le modèle d'OCEAN a été adopté pour définir les 5 traits principaux de personnalité (Ouverture, Conscience, Extraversion, Agrément et Névrose). Une association est établie entre les paramètres de l'Effort et les facteurs d'OCEAN pour trouver la relation entre la personnalité et la composante Effort de LMA. Cette relation a été utilisée pour une généralisation de la représentation de la personnalité à travers les mouvements. Ils ont envisagé d'appliquer cette relation pour produire une variation stylisée du mouvement d'un agent virtuel en ajustant les paramètres de l'Effort. La Figure 2.22 montre un exemple du geste de pointage effectué par l'agent avatar suivant les 5 traits du modèle OCEAN.

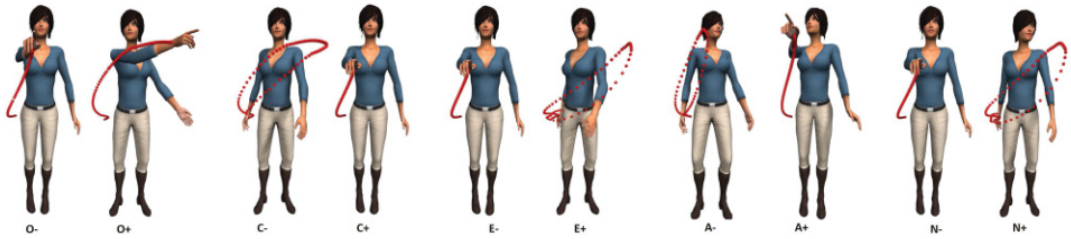


FIGURE 2.22 – Variation dans le geste de pointage pour les cinq traits du modèle OCEAN.

— Indexation et récupération du mouvement : Cette approche consiste en une récupération de mouvement basée sur son contenu à partir de l'ensemble de données. La méthode LMA est utilisée pour encoder les caractéristiques du mouvement. Ces caractéristiques fournissent un espace de recherche représentatif pour l'indexation des mouvements. Dans ce contexte, Kapadia et al. [Kapadia et al., 2013] ont proposé une représentation compacte qui capture suffisamment les caractéristiques du mouvement humain et fournit un moyen efficace pour l'indexation des mouvement indépendamment de la taille la base de données. Leurs caractéristiques se sont inspirées de trois facteurs de LMA (Corps, Effort et Forme). Ces caractéristiques sont ensuite combinées pour rechercher des mouvements complexes dans de grandes bases de données de mouvements. La Figure 2.23 illustre un prototype d'interface utilisateur graphique pour la recherche et la récupération de mouvement. Aristidou et al. [Aristidou and Chrysanthou, 2014] ont utilisé la méthode LMA pour extraire les caractéristiques indicatives pour des mouve-

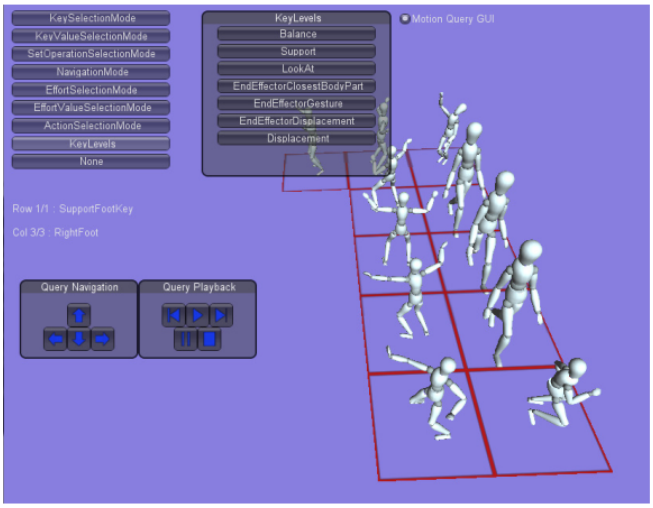


FIGURE 2.23 – Interface utilisateur graphique pour la récupération du mouvement.

ments de danse. Ils ont étudié la corrélation entre les différentes émotions en se basant sur les caractéristiques du mouvement. Les même auteurs [Aristidou et al., 2014a] se sont basés sur le même descripteur et ont proposé un algorithme de recherche qui consiste à mesurer la corrélation entre les différents mouvements en s'appuyant sur les caractéristiques de ce descripteur. Cela permet de trouver les similarités potentielles entre les différents clips de danse et ainsi récupérer les mouvements avec des qualités similaires, comme le montre la Figure 2.24.

FIGURE 2.24 – La première rangée affiche quelques images d'une séquence de mouvement de danse et les autres correspondent aux mouvements les plus similaires.

— Transfert des styles : Les méthodes de transfert de style examinent la question du transfert du style d'un mouvement d'une personne à un autre. [Torresani et al., 2006] ont entrainé un mappage non linéaire entre les paramètres d'animation et les styles de mouvement dans

l'espace perceptuel. Ce mapping peut ensuite être utilisé pour synthétiser des variations stylistiques à partir d'exemples générés artificiellement en utilisant les facteurs d'Effort de LMA. [Aristidou et al., 2017b] ont employé la technique LMA pour synthétiser les mouvements humains à partir des données de capture de mouvement existantes. Ils ont extrait les caractéristiques quantitatives et qualitatives du mouvement basées sur les composantes de LMA. Ils ont appliqué le modèle de régression RBF pour associer les caractéristiques de mouvement à leurs coordonnées d'émotion sur le modèle du circumplex de Russell. Cela permet de styliser les mouvements avec des émotions en modifiant les caractéristiques sélectionnées. La Figure 2.25 montre un exemple de changement de style d'une danse contemporaine en une danse "plus heureuse" et "plus triste". Les même auteurs [Aristidou et al., 2017a] ont étendu

FIGURE 2.25 – La première rangée montre une danse contemporaine capturée. La deuxième rangée montre une danse «plus heureuse», tandis que la dernière correspond à une danse «plus triste».

leur framework basé LMA en ajoutant d'autres caractéristiques qui prennent en considération les modes d'interaction avec soi-même, les autres et l'environnement, visant à améliorer la cohérence stylistique par rapport à la composante spatiale qui n'a pas été complètement étudiée dans [Aristidou et al., 2017b]. Ils ont d'abord extrait les caractéristiques du mouvement inspirées des composantes de LMA et trouvé leurs corrélations stylistiques. Ils ont construit un graphe de mouvement en se basant sur les corrélations entre les postures afin de trouver les transitions potentielles entre les vidéos. Ces corrélations sont utilisées pour élaguer les transitions du graphe de mouvement qui ne sont pas cohérentes d'un point de vue stylistique, conduisant à un graphe de mouvement basé sur LMA. Ce nouveau graphe de mouvement peut être utilisé pour synthétiser des animations de danse plausibles.

— Quantification du contenu expressif des gestes par rapport à l'émotion : [Samadani et al., 2013] ont adopté la méthode de LMA pour l'analyse du mouvement des mains et des bras. Ils ont quantifié les facteurs de l'Effort (espace, temps, poids et

flux) ainsi que le facteur du mouvement directionnel de la Forme en se basant sur des caractéristiques de mouvement mesurées, comme la vitesse, l'accélération, l'à-coup, etc. Dans leur base de données, six mouvements des mains et des bras ont été réalisés par un acteur professionnel pour transmettre six émotions (colère, bonheur, tristesse, peur, dégoût et surprise). Par la suite, leur base de données a été annotée par un analyste de mouvements certifié (CMA) pour étudier la corrélation statistique entre les annotations de CMA et les facteurs quantifiés de LMA. [Truong et al., 2016] ont utilisé la méthode LMA pour classifier les gestes des chefs d'orchestre et aussi analyser le contenu expressif de ses gestes. Dans leur base de données, les auteurs ont enregistré 8 sessions différentes. Par la suite, ils ont segmenté chaque session manuellement afin de créer des échantillons des gestes. Ces gestes sont par la suite annotés par des experts en sélectionnant des émotions parmi les catégories proposées dans leur ensemble de données. [Aristidou et al., 2015b] ont utilisé la méthode LMA pour encoder les caractéristiques physiques et aussi stylistiques du mouvement. Ils ont développé une plateforme d'apprentissage de danse folklorique pour aider les débutants à apprendre cette danse en suivant un avatar 3D. Des danseurs professionnels ont été invités à construire leur base de données. Chaque utilisateur imite donc la danse de l'avatar. Son mouvement est analysé et comparé à celui du modèle. Cette comparaison s'appuie sur les qualités extraites de l'utilisateur et de l'avatar inspirées de la méthode LMA. Finalement, une évaluation de la performance de l'utilisateur est déduite.

2.4 Positionnement de notre approche dans le domaine IHR

Comme nous pouvons le constater, la plupart des travaux sur les gestes expressifs se sont souvent tournés vers la danse ou la musique. Ces séquences de danse ont tendance à être continues, sans la restriction artificielle de devoir commencer et finir avec une pose neutre. Pour cela, la plupart des travaux se sont concentrés sur l'analyse des gestes expressifs sans passer par l'entrainement et la classification. D'autres limites dans ces travaux sont apparues au niveau de la construction de leurs bases de mouvements qui nécessitent toujours des cours d'entrainement et aussi des experts du domaine. Dans le cadre d'une Interaction Homme-Robot (IHR), nous avons trouvé quelques articles qui ont impliqué le modèle de LMA à des fins différentes. Certains ont utilisé ce modèle pour caractériser les trajectoires des robots, à l'instar de [Knight and Simmons, 2014] qui ont généré des trajectoires de mouvement expressif du robot en se basant sur la composante Effort de LMA. Ils ont extrait 3 mesures (position x,y et orientation θ) et ont créé un vecteur descripteur de mouvement basé sur ces caractéristiques pour quantifier chaque facteur de la composante

Effort. Finalement, ils ont étudié la relation entre 6 émotions et les paramètres de l'Effort. Les même auteurs [Knight et al., 2016] ont quantifié le facteur Espace de la composante Effort pour étudier l'interprétation des gens sur les attitudes du robot par rapport à son point d'arrivée (hésitation, direct, perdu la trace de but) à travers les caractéristiques de la trajectoire de son mouvement. [Sharma et al., 2013] se sont basés sur la composante Effort de LMA pour caractériser la trajectoire du mouvement d'un robot volant. Ils ont recruté un artiste entrainé à Laban pour créer un ensemble de mouvements pour chaque combinaison de paramètres de l'Effort (espace, temps, poids et flux). Les 4 paramètres avec leurs deux qualités extrêmes donnent alors 16 combinaisons. Puis, ils ont adopté le modèle de Circumplex pour associer l'affect perçu à partir de mouvements robotiques sur les deux dimensions : valence et excitation. Un autre type de recherche dans le domaine IHR basé sur le modèle LMA consiste à générer des gestes expressifs au robot humanoïde en variant les paramètres des facteurs de LMA. Nous pouvons citer [Kim et al., 2012] qui ont proposé un modèle computationnel des facteurs de poids et de temps et l'ont appliqué à des plates-formes robotisées pour développer une méthode de diversification des mouvements gestuels du robot Darwin-OP. Aussi, [Masuda and Kato, 2010, Masuda et al., 2010] ont utilisé les qualités de Laban pour donner des gestes expressifs (plaisir, colère, tristesse et relaxation) au robot humanoïde KHR-2HV. Leur méthode consiste à ajouter une émotion cible à des mouvements corporels arbitraires d'un robot de forme humaine tout en modifiant les paramètres de Laban. D'autres chercheurs ont considéré que l'interprétation et la reconnaissance du mouvement de la personne par un robot rend l'interaction Homme-Robot plus naturelle. Ils ont donc appliqué le modèle LMA pour la caractérisation des gestes de l'utilisateur, tels que [Kim et al., 2013] qui ont représenté les mouvements émotionnels du corps humain avec les trois facteurs de la composante Effort. Ils ont quantifié les trois facteurs (espace, poids et temps) pour caractériser deux mouvement émotionnels (se réjouit et se plaint). Finalement, nous citons le travail de [Lourens et al., 2010, Barakova and Lourens, 2010] qui ont considéré que la perception et l'interprétation du comportement non verbal par un robot est importante pour une interaction naturelle Homme-Robot. Deux caractéristiques (l'accélération et la fréquence) sont extraites pour quantifier la composante Effort afin de caractériser le geste "faire un signe avec une main" effectué avec 4 émotions (joie, colère, tristesse et politesse). Leur objectif est de développer un robot NAO capable de reconnaître et imiter les mouvements humains et ainsi assurer une interaction naturelle avec des enfants autistes dans le cadre d'un jeu collectif. Les même auteurs [Kim et al., 2014] ont développé récemment un framework pour l'entraînement assisté par robot des enfants atteints de troubles du spectre autistique (TSA) (Voir Figure 2.26). Leur framework est créé à l'aide du logiciel Choregraphe développé par la société Aldebaran, qui permet de

créer les comportement du robot NAO avec un langage graphique.

FIGURE 2.26 – Enfants avec TSA jouant avec le robot NAO.

L'objectif de notre travail de thèse est proche de ce dernier, nous envisageons à développer un système de reconnaissance des gestes expressifs afin d'assurer une interaction naturelle entre l'homme et le robot NAO. L'idée ici est d'avoir un robot capable de reconnaitre le mouvement de la personne et aussi son état d'une manière automatique. [Lourens et al., 2010, Barakova and Lourens, 2010] ont étudié la variation des caractéristiques dans le même mouvement effectué avec 4 émotions. Dans notre cas, nous développons un système de reconnaissance des gestes automatique qui permet de classifier les gestes. Nous considérons plusieurs gestes effectués avec 4 émotions. Puis, nous quantifions toutes les composantes de LMA (corps, espace, forme et effort) afin de décrire l'aspect quantitatif et qualitatif du mouvement. Notre descripteur de mouvement est capable de différencier des mouvements mais aussi les émotions exprimées par un même mouvement. Nous avons construit une base de données des gestes expressifs accessible au public, facile à utiliser et à enrichir par n'importe quelle personne. Cette base est composée de 5 gestes expressifs. Chaque geste est effectué avec différentes émotions. Nous évaluons notre descripteur de mouvement sur des bases publiques ainsi que sur notre base construite. Un algorithme de sélection des caractéristiques est développé pour étudier l'importance de chaque paramètre de mouvement pour discriminer chaque émotion. Une deuxième évaluation de notre système est réalisée avec une approche humaine basée sur les avis des hommes dans la perception des émotions et dans l'estimation du descripteur proposé. Finalement, pour déduire la robustesse et l'adéquation de notre système, nous comparons les résultats de notre système de reconnaissance automatique avec les résultats issus de l'approche humaine.

Chapitre 3

Reconnaissance des gestes dynamiques par les modèles de Markov cachés

Sommaire

3.1 Construction de CMKinect-10 58
3.1.1 Description de la base CMKinect-10 58
3.1.2 Acquisition des données 59
3.1.3 Normalisation 61
3.2 Descripteur local inspiré de LMA 62
3.2.1 Composante Corps 63
3.2.2 Composante Espace 65
3.2.3 Composante Forme 68
3.3 Application des MMCs 72
3.3.1 Formalisme de MMC 72
3.3.2 Topologies 73
3.3.3 Échantillonnage 74
3.3.4 Quantification vectorielle 75
3.3.5 Entrainement et classification des gestes 77
3.3.6 Contribution aux modèles de Markov cachés 78
3.4 Évaluation expérimentale 80
3.4.1 MSRC-12 81
3.4.2 MSR Action 3D 84
3.4.3 UTKinect 88
3.4.4 CMKinect-10 89
3.5 Bilan 90

Dans ce chapitre nous présentons notre système de reconnaissance des gestes dynamiques avec ses différentes étapes, notamment, l'acquisition des données avec le capteur Kinect, la normalisation des données, l'extraction des caractéristiques et la classification des gestes. Nous introduisons la base de données que nous avons construite pour une application robotique afin de contrôler le robot NAO via les gestes humains. Il s'agit de 10 gestes de contrôle effectués par 20 personnes et capturés avec la caméra Kinect. Chaque geste est désormais représenté à chaque instant par un vecteur des caractéristiques inspiré de la méthode d'analyse de mouvement de Laban, nommée LMA (Laban Movement Analysis). Le geste est alors représenté par une séquence d'observations composée par des caractéristiques locales visant à décrire le mouvement à chaque instant. L'ensemble des séquences sera ainsi les entrées du Modèle de Markov Caché (MMC) pour l'entrainement et la classification des gestes. Une contribution est réalisée à la méthode MMC qui permet de mieux distinguer des mouvements similaires. Notre système est évalué sur des bases de données publiques, ainsi que notre base, nommée CMKinect-10.

Ce chapitre est organisé de la manière suivante : dans la section 3.1, nous présentons notre base de données construite pour notre première application robotique pour le contrôle du robot NAO. La section 3.2 concerne l'étape de la description du mouvement en se basant sur les qualités de la méthode LMA. Nous quantifions les 3 composantes de LMA (Corps, Espace et Forme) pour construire un descripteur de mouvement adapté au type de notre application. La dernière étape du processus de reconnaissance de gestes concernant l'entrainement et la classification des gestes avec le modèle de Markov caché est détaillé dans la section 3.3. Dans la section 3.4, nous présentons les différentes expérimentations faites sur ce système. Nous concluons dans la section 3.5 .

3.1 Construction de la base de gestes de contrôle CMKinect-10

3.1.1 Description de la base CMKinect-10

Nous avons construit une base de données composée de 10 gestes de contrôle. Cette base sera par la suite utilisée dans notre système de reconnaissance de gestes pour le contrôle du NAO. La Figure 3.1 donne une vue globale sur les différentes étapes de notre processus de reconnaissance de gestes. Commençant par l'acquisition des données qui a été faite avec le capteur Kinect sous ROS (Robot Operating system), un framework complet pour robots. Le module suivant est le pré-traitement des données qui transforme les données dans une forme appropriée pour qu'elles soient exploitables par notre système, y inclus, la normalisation des données. Cette étape est cruciale dans

la présentation des descripteurs car ces derniers sont très sensibles aux problèmes des variabilités qui peuvent se produire entre les différents individus. Par exemple, deux mouvements identiques peuvent donner lieu à deux descripteurs différents s'ils sont exécutés dans deux positions différentes. Après la normalisation, vient l'étape la plus importante dans notre système, qui est l'extraction des caractéristiques. Il s'agit de déterminer les descripteurs qui représentent le mieux les mouvements. Un descripteur forme un vecteur des caractéristiques qui renseigne sur l'état du système. On le nomme donc explicitement le descripteur de mouvement. Notre descripteur de mouvement a été inspiré de la méthode d'analyse de mouvement LMA proposé par [Laban, 1994] pour pouvoir caractériser l'aspect quantitatif et qualitatif du mouvement. Finalement, ce vecteur sera l'entrée du modèle de Markov caché pour l'entrainement et la classification des gestes.

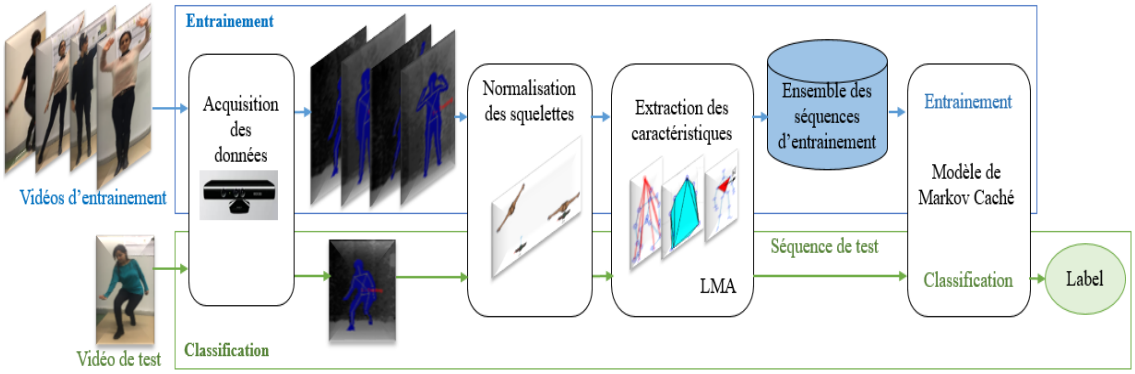


FIGURE 3.1 – Le processus de notre système de reconnaissance de gestes.

3.1.2 Acquisition des données

CMKinect-10 (10 Control Motions) est une base de données composée de 10 gestes de contrôle (danser, se présenter, s'asseoir, diminuer la vitesse, augmenter la vitesse, s'arrêter, tourner à gauche, tourner à droite, faire un signe avec les deux mains et avancer). 20 personnes (10 hommes et 10 femmes) de l'Université d'Evry Val d'Essonne, âgées de 27 à 36 ans (moyenne= 29,85 ans, écart-type= 2,47) ont participé à la construction de cette base. Nous avons demandé à chaque personne de répéter chaque geste 10 fois. Au total, nous avons eu 2000 séquences (20 personnes× 10 gestes× 10 répétitions). L'acquisition des données a été faite sous ROS [1] avec le module OpenNI pour la détection et le suivi du squelette via le Middleware NITE. Cela permet de publier des informations liées à la position et l'orientation de 15 articulations suivantes (tête, cou, torse, épaule gauche, coude gauche, main gauche, épaule droite, coude droit, main droite, hanche gauche, genou gauche, pied gauche, hanche droite, genou droit et pied droit) capturées à une résolution de 640×480 à 30 trames

1. http ://wiki.ros.org/fr

par seconde. La Figure 3.2 montre le squelette affiché lors de l'activation du processus de tracking sous une interface graphique sous ROS, nommée l'interface RVIZ. Nous présentons dans la Figure 3.3

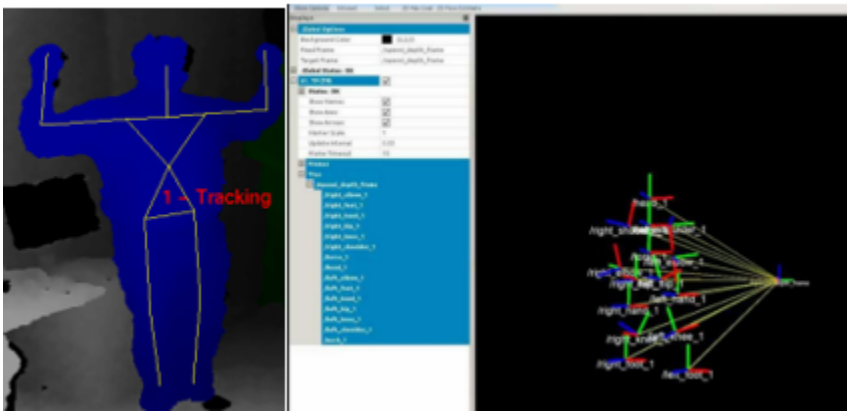


FIGURE 3.2 – Suivi et visualisation du squelette sous l'interface RVIZ.

quelques images capturées dans chaque geste de la base CMKinect-10.

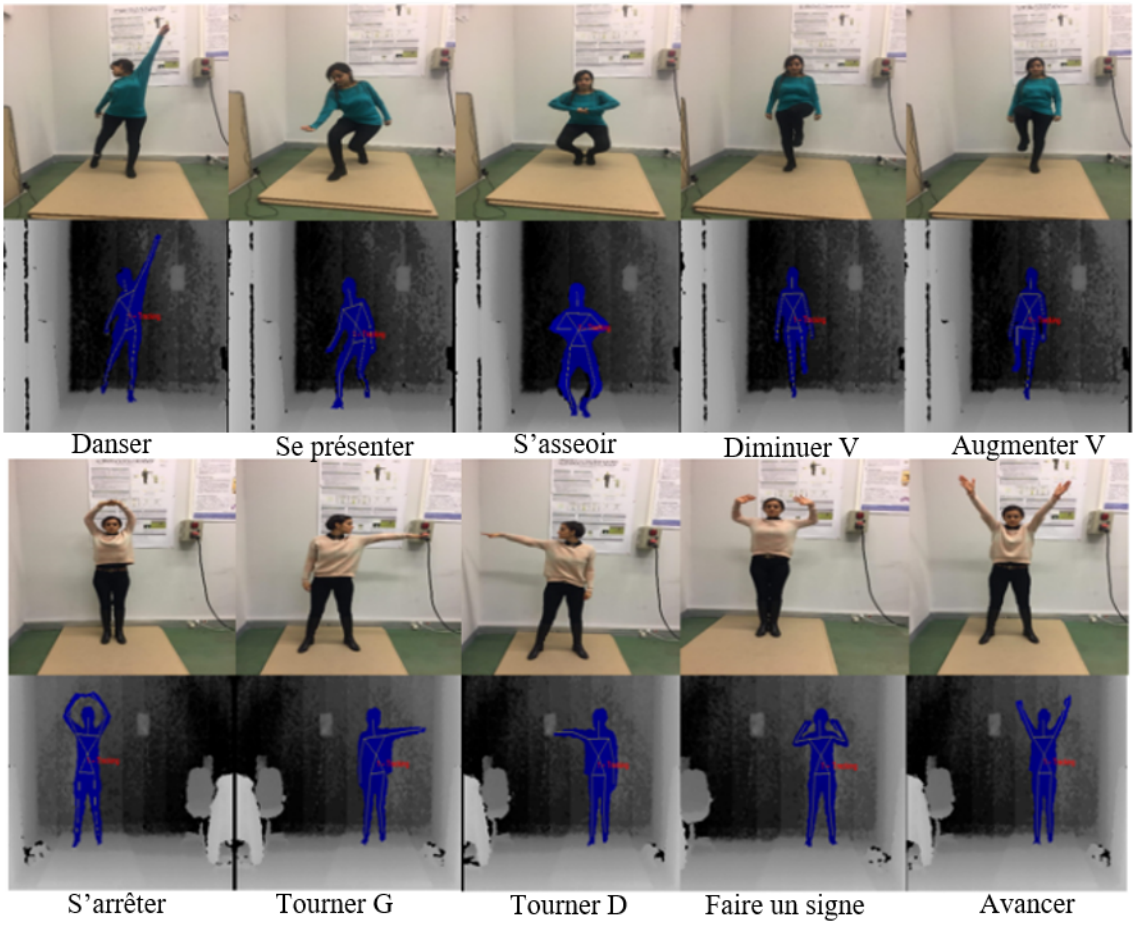


FIGURE 3.3 – Des échantillons d'images de chaque geste de la base CMKinect-10.

3.1.3 Normalisation

Les premières difficultés rencontrées dans les algorithmes de reconnaissance de gestes sont les variations de pose due à la variabilité liée aux positions et aux orientations des articulations. Un même geste effectué par deux personnes se mettant à deux positions initiales différentes sera interprété différemment. Afin de résoudre ce problème, nous avons besoin de transformer le système de coordonnées global lié au capteur Kinect avec l'origine O situé au centre du capteur dans un système local lié au centre du squelette (Voir Figure 3.4). Dans ce cas, toutes informations issues du squelette sera définies par rapport à son système local. Étant donné une séquence de mouvement

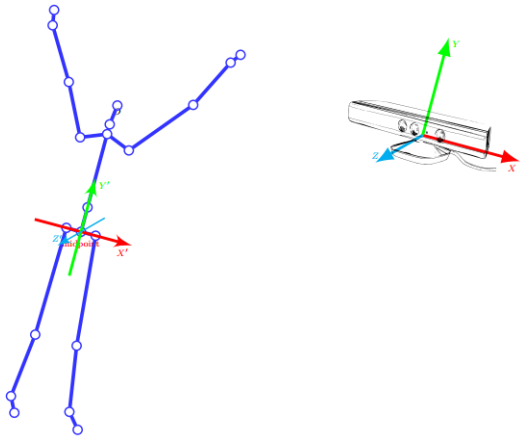


FIGURE 3.4 – Les systèmes des coordonnées liés au capteur Kinect et au squelette.

$S = \{P_t^j\}\,;\, j \in \{1,\ldots,N\},\, t \in \{1,\ldots,T\}$, où P_t^j correspond à la position 3D de l'articulation j capturée à la trame t et T correspond à la longueur de la séquence du geste. Nous définissons un système de coordonnées local lié au centre de la hanche du squelette, représenté par la base B', équipé de trois vecteurs unitaires : le vecteur $\vec{n_{lhi}}$ relié entre le centre de la hanche et l'articulation de la hanche gauche, le vecteur de la colonne vertébrale $\vec{n_s}$ et leur produit vectoriel $\vec{n_c} = \vec{n_{lhi}} \wedge \vec{n_s}$. Pour chaque séquence, nous appliquons d'abord une translation pour déplacer le centre du système lié au squelette au centre de la Kinect et après une rotation pour aligner les deux systèmes de coordonnées. La position de chaque articulation transformée est calculée suivant les transformations suivantes :

$$[P_t^j]_{B'} = R_{B\leftarrow B'}^{-1}([P_t^j]_B - [P_1^c]_B) \tag{3.1}$$

$$R_{B\leftarrow B'} = [\frac{\vec{n_{lhi}}}{\|\vec{n_{lhi}}\|} \frac{\vec{n_s}}{\|\vec{n_s}\|} \frac{\vec{n_c}}{\|\vec{n_c}\|}] \tag{3.2}$$

$$\vec{n_{lhi}} = [P_1^{lhi}]_B - [P_1^c]_B \tag{3.3}$$

$$\vec{n_s} = [P_1^s]_B - [P_1^c]_B \tag{3.4}$$

$$[P_1^c]_B = \frac{([P_1^{lhi}]_B + [P_1^{rhi}]_B)}{2} \tag{3.5}$$

P_1^c, P_1^{lhi} et P_1^{rhi} correspondent respectivement, aux positions des articulations de centre de la hanche, hanche gauche et hanche droite capturées à l'instant initial ($t = 1$). Puisque les vecteurs de rotation sont des vecteurs unitaires orthogonaux, donc $R^{-1} = R^T$.

$$[P_t^j]_{B'} = R^T_{B \leftarrow B'}([P_t^j]_B - [P_1^c]_B) \tag{3.6}$$

Grâce à ces transformations, nous aurons un système de reconnaissance plus robuste face aux positions et aux orientations initiales des personnes. Le résultat de cette normalisation est illustré à la Figure 3.5.

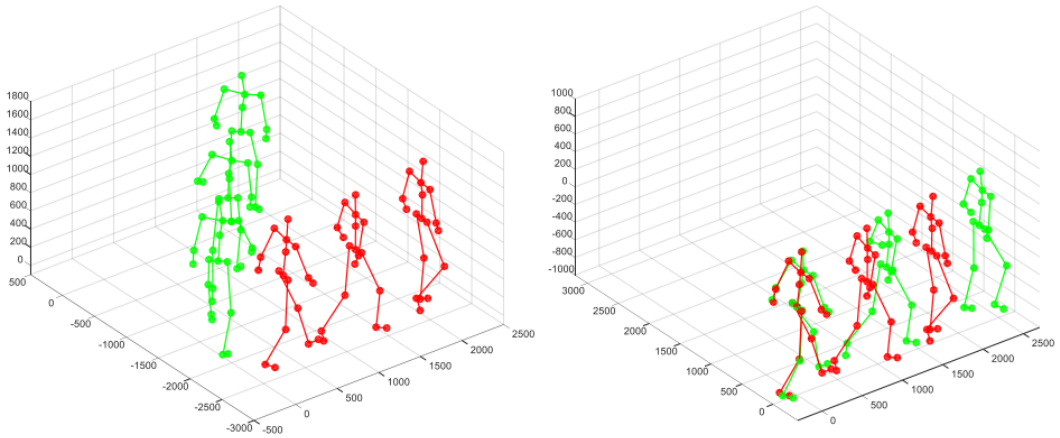


FIGURE 3.5 – Normalisation des deux squelettes exécutants le même mouvement «Marcher» avec deux positions initiales différentes.

3.2 Descripteur local inspiré de LMA

Pour notre première application qui consiste à contrôler un robot NAO avec les gestes humains, nous cherchons à construire un descripteur de mouvement à la fois robuste et indépendant de cer-

taines contraintes qui peuvent influencer notre système de reconnaissance de gestes, comme le sexe ou l'âge de l'utilisateur. En effet, un jeune enfant, un homme adulte ou une femme âgée produiront différemment le même geste suivant leurs rythmes ce qui risque d'avoir des représentations du mouvement différentes. La vitesse à laquelle un geste est effectué varie également en fonction du contexte et de l'état de la personne. Un même geste peut être effectué à une vitesse différente en fonction de l'humeur actuelle de la personne ou l'intention spécifique qu'il veut communiquer. Donc, pour pouvoir contrôler le robot sans être influencé par ces facteurs, notre descripteur est basé sur trois qualités de LMA : Corps, Espace et Forme. La composante Effort permet de décrire l'intention, le rythme et l'expressivité du mouvement, alors que ce n'est pas le but de notre première application.

3.2.1 Composante Corps

La composante corps décrit les caractéristiques structurelles et physiques du corps humain durant son mouvement. Dans l'état de l'art, peu de chercheurs ont utilisé les deux composantes Corps et Espace, la plupart se sont focalisés sur l'Effort et la Forme pour caractériser la qualité du mouvement et décrire l'expressivité du geste [Samadani et al., 2013, Zacharatos et al., 2013, Kim et al., 2013, Suzuki et al., 2000, Masuda et al., 2010, Masuda and Kato, 2010, Nishimura et al., 2012]. Pour notre cas, cette composante est indispensable car nous en avons besoin pour décrire les caractéristiques structurelles du corps et faire la distinction entre les gestes différents. Pour le même principe, quelques chercheurs n'ont pas ignoré ce facteur dans leur application, commençant par [Aristidou and Chrysanthou, 2014, Aristidou et al., 2014b, Aristidou et al., 2015b] qui ont adopté l'approche LMA pour l'analyse des gestes de danse, où ils ont quantifié les quatre composantes de LMA. Pour le Corps, ils ont extrait les 9 caractéristiques suivantes :

- Les distances : entre les pieds et les genoux, entre les mains et les épaules, entre les deux mains, entre les deux pieds, entre la tête et les mains, entre le bassin et le centroide du squelette.
- Les hauteurs de bassin et de centroide du squelette par rapport au sol.
- La différence entre les deux distance suivantes : la distance entre les hanches et le sol et la distance entre les pieds et les hanches.

[Kapadia et al., 2013] dans leur méthode qui consiste à indexer un mouvement dans une base de données très large, ont reposé sur les composantes de LMA afin de pouvoir récupérer les mouvements complexes. Ils ont représenté le facteur Corps avec les caractéristiques suivantes :

- Une valeur booléenne indiquant la présence ou l'absence du mouvement en comparant le déplacement d'un segment du corps entre deux trames successives avec un seuil prédéfini.

— Les déplacement et l'orientation d'un effecteur terminal par rapport à sa racine.

— L'indice du segment du corps le plus proche des effecteurs terminaux ainsi que la distance entre les deux.

— La position du centre de masse du squelette et son déplacement par rapport à sa position de repos.

— Une valeur booléenne pour indiquer la position relative du centre de masse par rapport au polygone support du squelette corporel.

— Un indice sur le segment du corps utilisé pour soutenir le poids corporel et en contact avec le sol.

De même [Truong and Zaharia, 2016, Truong et al., 2016] se sont basés sur la technique de LMA pour analyser les gestes d'un chef d'orchestre. Il ont quantifié cette composante avec trois mesures différentes : les distances entre les mains et les épaules (parties gauche et droite) et la dissymétrie spatiale du corps. La caractéristique de la symétrie du corps a été aussi utilisée dans plusieurs travaux [Glowinski et al., 2011, Wang et al., 2015, Garber-Barron and Si, 2012] afin de décrire la symétrie du corps lors de la représentation des gestes expressifs. Afin d'exprimer la connectivité du corps et de trouver la relation entre les parties du corps, nous avons mesuré 13 caractéristiques. Nous avons considéré deux parties, la partie supérieure et la partie inférieure (Voir Figure 3.6). Pour la première partie, l'extension des différentes articulations est caractérisée par les angles suivants : entre les mains et les épaules (θ_1^l, θ_1^r), entre les coudes et les hanches (θ_2^l, θ_2^r), entre les coudes et les épaules de la partie symétrique (θ_3^l, θ_3^r). Également l'angle formé par les deux mains par rapport au centre des épaules (θ_{Hs}) est calculé. Nous avons aussi calculé les distances entre les deux mains (d_{Hs}) ainsi que les distances entre la tête et les mains gauche et droite (d_{hlh}, d_{hrh}). La Figure 3.7 montre

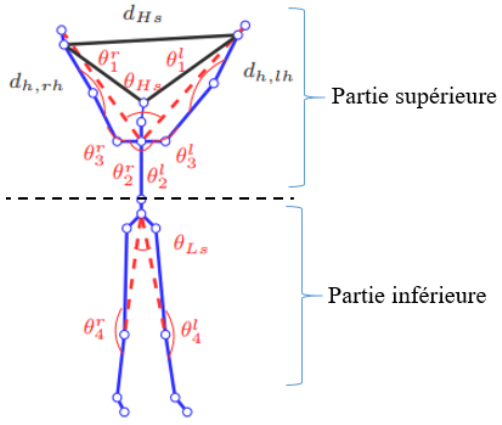


FIGURE 3.6 – Les caractéristiques de la composante Corps.

la variation des deux caractéristiques θ_2^l et θ_2^r dans le geste "Avancer" de la base CMKinect-10. A

l'instant initial les deux angles sont inférieurs à 30°, les deux bras s'élèvent vers le haut (les valeurs de θ_2^l et θ_2^r augmentent) et restent ouvertes dans la même position jusqu'à la fin du geste (θ_2^l et θ_2^r se stabilisent à 120°). Les trois distances $(d_{h,lh}, d_{h,rh}, d_{Hs})$ permettent de différencier les mouvements

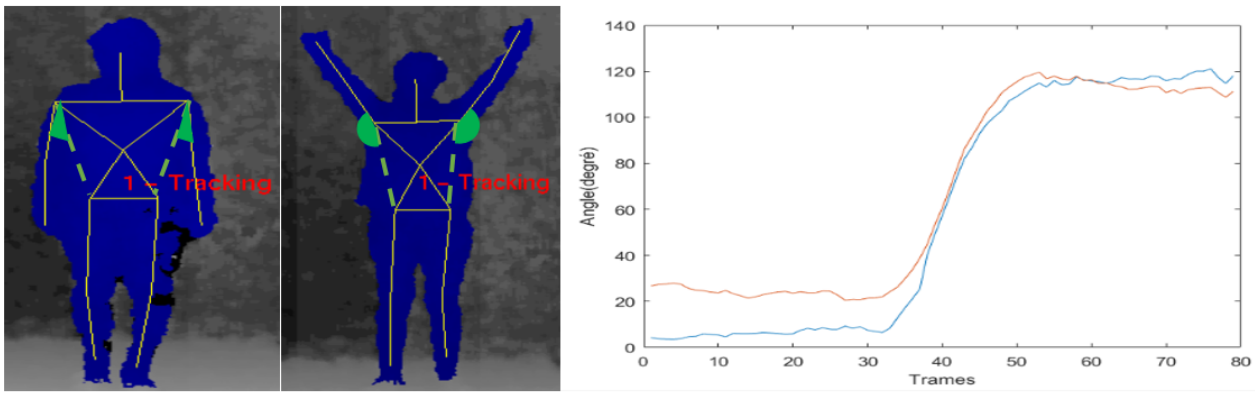


FIGURE 3.7 – Variation des caractéristiques θ_2^l (courbe bleue) et θ_2^r (courbe orange) dans le geste «Avancer» de la base CMKinect-10.

des mains similaires comme les deux gestes «Faire un signe avec les deux mains» et «s'arrêter» de notre base CMKinect-10 (Voir Figure 3.8). Dans les deux gestes, les deux mains se touchent au-dessus de la tête. Sauf que dans le geste "s'arrêter" les deux mains restent en contact dans la même position jusqu'à la fin du geste. Par contre dans le deuxième geste, les deux mains se rapprochent et reviennent à la fin du geste à leurs positions initiales. Avec la caractéristique d_{Hs} nous pouvons faire la distinction entre ces deux gestes. Pour le geste «faire un signe», comme présenté dans la Figure 3.8, la variable d_{Hs} (courbe orange) commence par une valeur constante puis diminue et revient à la fin à sa valeur initiale. Pour le geste «s'arrêter», la courbe bleue de la variable d_{Hs} tend vers 0 à la fin du geste. Aussi pour les deux caractéristiques $d_{h,lh}$ et $d_{h,rh}$ (Voir Figure 3.9), dans le geste «faire un signe» (courbes bleue et orange), elles diminuent et puis reviennent à leurs valeurs initiales alors que dans le geste «s'arrêter» (courbes violette et jaune) elles diminuent et restent constantes à la fin du geste. Pour la partie inférieure du corps, l'extension des genoux est décrite par les angles calculés entre les pieds et les hanches (θ_4^l, θ_4^r). Ces deux caractéristiques permettent de caractériser des actions comme «s'asseoir» présentée dans la Figure 3.10. Aussi, nous avons calculé l'angle formé par les deux genoux par rapport au centre de la hanche (θ_{Ls}). Cette caractéristique permet de décrire l'écartement des jambes.

3.2.2 Composante Espace

La composante Espace décrit la localisation, les directions et les chemins d'un mouvement. [Truong and Zaharia, 2016, Truong et al., 2016] ont quantifié la qualité d'Espace avec deux ca-

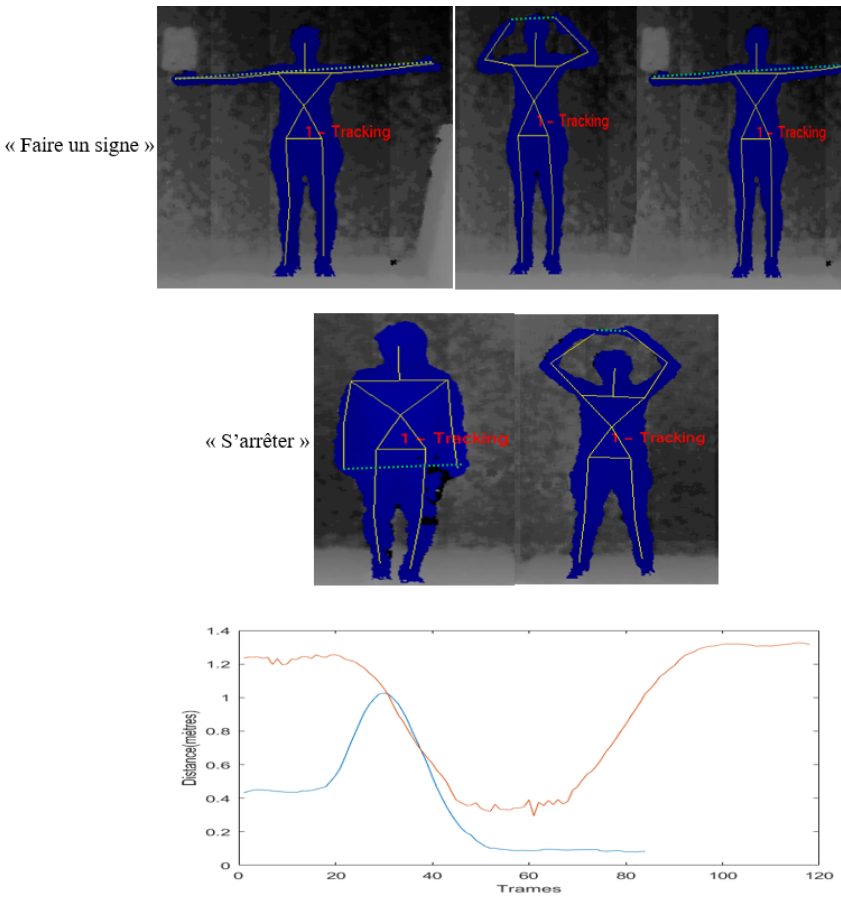


FIGURE 3.8 – Variation des courbes de la distance entre les deux mains d_{Hs} dans le geste «faire un signe» (courbe orange) et le geste «s'arrêter» (courbe bleue).

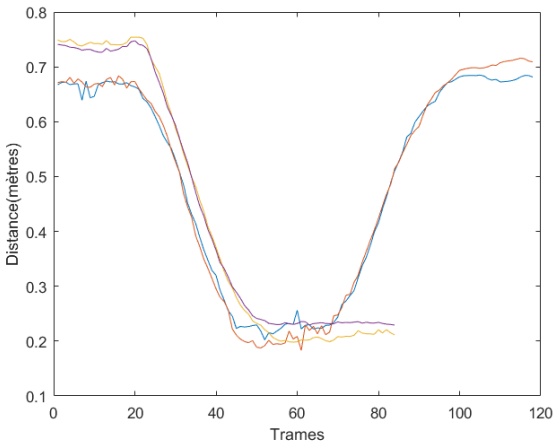


FIGURE 3.9 – Variation des courbes des caractéristiques $d_{h,lh}$ et $d_{h,rh}$ dans le geste «faire un signe» ($d_{h,rh}$ courbe bleue et $d_{h,lh}$ courbe orange) et le geste «s'arrêter» ($d_{h,rh}$ courbe violette et $d_{h,lh}$ courbe jaune).

ractéristiques, la position de la tête pour caractériser le mouvement avant-arrière de la tête et l'angle d'inclinaison vers l'avant défini comme l'angle entre l'axe verticale Y et l'axe reliant le centre de la

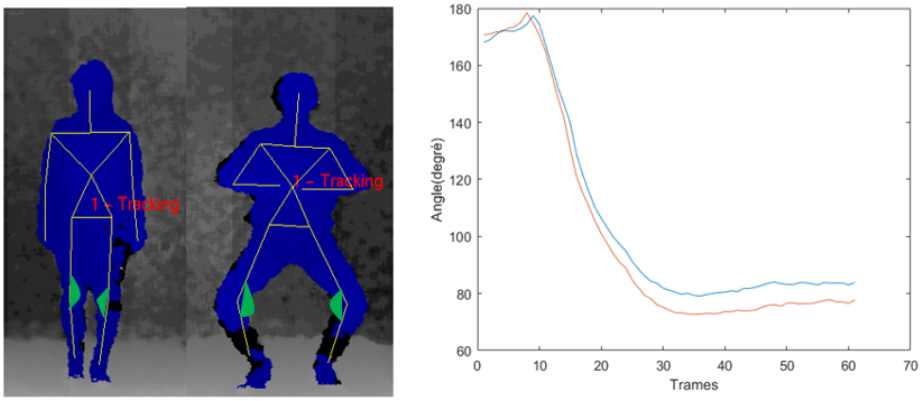


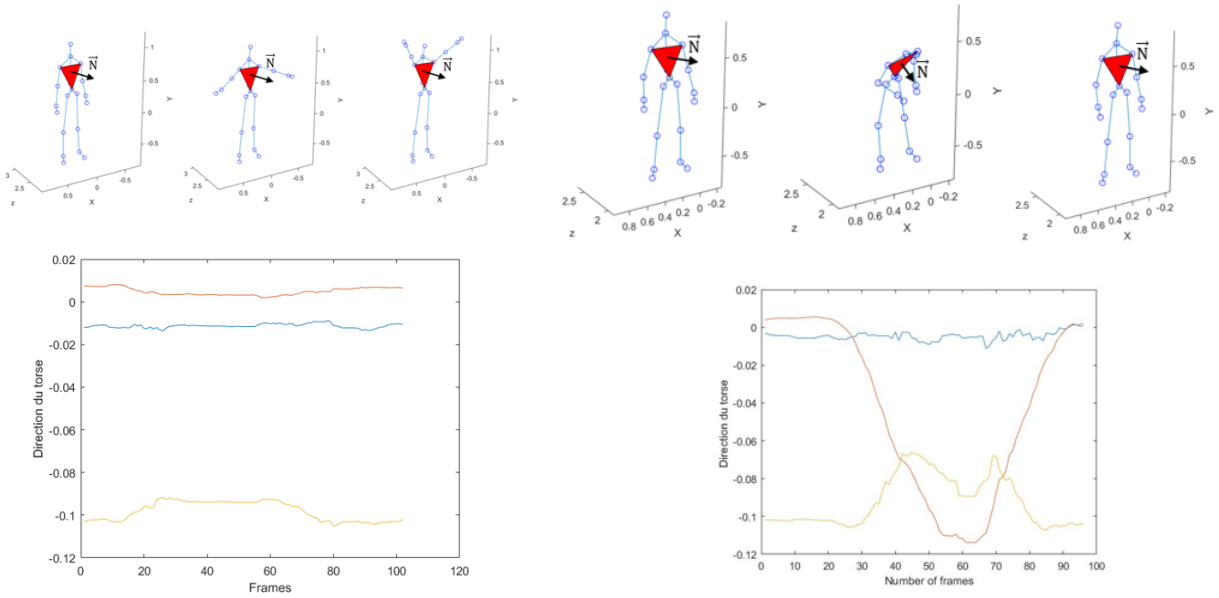
FIGURE 3.10 – Variation des valeurs des angles θ_4^l (courbe bleue) et θ_4^r (courbe orange) dans le geste «s'asseoir».

hanche et la tête. [Aristidou and Chrysanthou, 2014, Aristidou et al., 2014b, Aristidou et al., 2015b] ont aussi utilisé deux caractéristiques différentes pour caractériser l'Espace qui sont la distance totale parcourue sur une période de temps où ils l'ont utilisée pour l'évaluation de trois durées différentes de 30, 15 et 5 secondes et la zone couverte pour la même période. Afin de quantifier cette composante, nous avons décrit la direction du corps de la personne dans son environnement. Donc, nous avons identifié la direction du torse en calculant le vecteur normal $\vec{N}$ du triangle formé par les trois articulations suivantes : l'épaule gauche, l'épaule droite et le centre de la hanche.

$$\vec{N} = \frac{\vec{n_{shl}} \wedge \vec{n_{shr}}}{\|\vec{n_{shl}} \wedge \vec{n_{shr}}\|}$$

$\vec{n_{shl}}$ est le vecteur entre le centre de la hanche et l'épaule gauche et $\vec{n_{shr}}$ est le vecteur entre le centre de la hanche et l'épaule droite. Cette équation donne le vecteur normal au torse $\vec{N}(n_x, n_y, n_z)$ suivant les trois axes X, Y et Z à chaque instant. Cette caractéristique est importante pour décrire la posture de la personne lors de l'exécution de l'action s'il est bien vertical ou incliné. Nous illustrons la direction du torse pour les deux gestes de la base MSRC-12 [Fothergill et al., 2012] «démarrer la musique» et «s'incliner» avec les trois composantes du vecteur $\vec{N}$ (n_x courbe bleue, n_y courbe orange, n_z courbe jaune) présentées dans la Figure 3.11. Pour le geste «démarrer la musique» le torse est toujours vertical donc nous remarquons bien que les trois composantes du vecteur normal $\vec{N}$ sont toujours presque constantes (Voir Figure 3.11(a)). Pour le geste «s'incliner», au début du geste le torse est dans sa position verticale, la courbe de la composante n_y est constante et diminue après pour atteindre un palier qui correspond au penchement vers l'avant de la personne puis un retour à la position initiale qui mène à un retour de la courbe à son niveau initial. Nous remarquons aussi une variation moins forte dans la courbe jaune qui représente la composante n_z due à l'avancement du torse un peu suivant l'axe Z lors du penchement. Par contre, la direction du torse suivant l'axe

horizontale (X) est presque constante (Voir Figure 3.11(b)).



(a) Geste «démarrer la musique».

(b) Geste «s'incliner».

FIGURE 3.11 – Variation de la direction du torse suivant les axes X (courbe bleue), Y (courbe orange) et Z (courbe jaune).

3.2.3 Composante Forme

La composante Forme implique trois qualités distinctes de changement dans la forme du mouvement : le flux de forme, le mouvement directionnel et la mise en forme.

Le flux de forme

Ce facteur reflète la relation du corps avec lui-même. Les changements peuvent être perçus comme le volume croissant ou décroissant de la forme du corps Certains auteurs ont représenté la composante de la Forme avec seulement ce facteur et ont ignoré les autres, comme [Aristidou and Chrysanthou, 2014, Aristidou et al., 2014b, Aristidou et al., 2015b] et [Kapadia et al., 2013] qui ont utilisé le volume limitant du squelette comme mesure pour ce facteur. Ils ont divisé le squelette en 4 parties (supérieure, inférieure, droite et gauche) et ont calculé leurs volumes correspondants et le volume limitant de toutes les articulations. De plus, ils ont ajouté deux mesures, l'hauteur du torse (la distance entre la tête et le centre de la hanche) et le niveau des mains (le niveau supérieur au dessus de la tête, le niveau intermédiaire entre la tête et le point milieu entre la tête et le centre de la hanche, et le niveau bas sous le point milieu). [Glowinski et al., 2011] ont calculé l'aire d'un triangle reliant les trois articulations suivante :

la tête, la main droite et la main gauche. [Truong et al., 2016] ont quantifié le flux de forme par un indice relatant la contraction du corps afin de caractériser l'extension des membres par rapport au centre du corps. Pour ce facteur, nous construisons l'enveloppe convexe du squelette en se basant sur l'algorithme Quickhull [Barber et al., 1996]. L'enveloppe convexe de l'ensemble des articulations du squelette est le plus petit ensemble convexe qui les contient tous. Nous calculons le volume de cette enveloppe afin de caractériser la déformation de la boite englobante du squelette au cours du temps. La Figure 3.12 illustre la déformation de l'enveloppe convexe du squelette en 3D au cours du temps pour le geste «démarrer la musique» de la base MSRC-12. Au début du geste, la personne est dans sa position initiale, puis elle lève ses deux mains donc le volume de l'enveloppe augmente. Avec cette caractéristique nous pouvons avoir une idée sur l'extension du corps tout au long du mouvement.

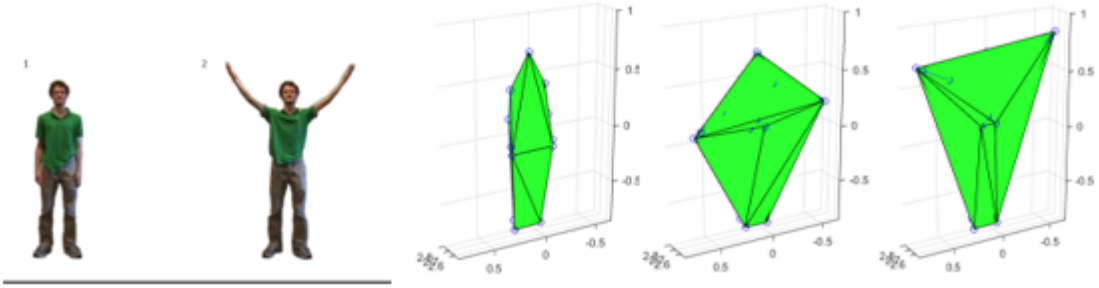


FIGURE 3.12 – Variation du volume de l'enveloppe convexe du squelette dans le geste «démarrer la musique» de la base MSRC-12.

Le mouvement directionnel

Il décrit la trajectoire du mouvement et différencie les deux caractéristiques suivantes : rectiligne et curviligne. Nous décrivons l'allure des trajectoires des extrémités supérieures du corps, les mains et la tête, en considérant leurs courbures locales. Nous calculons donc le changement angulaire $\phi_{P_t^k}$ de l'articulation k entre deux images successives (t et $t+1$) comme représenté dans la Figure 3.13, de la manière suivante :

$$\phi_{P_t^k} = \arccos\left(\frac{\overrightarrow{P_{t-1}^k P_t^k}}{\left\|\overrightarrow{P_{t-1}^k P_t^k}\right\|} \cdot \frac{\overrightarrow{P_t^k P_{t+1}^k}}{\left\|\overrightarrow{P_t^k P_{t+1}^k}\right\|}\right) \tag{3.7}$$

La mise en forme

Tandis que le flux de forme concerne principalement la détection des changements de forme du corps en lui-même, le facteur «mise en forme» décrit ce changement par rapport à l'espace 3D. Ce facteur décrit les changements de forme du mouvement suivant les trois plans présentés dans la

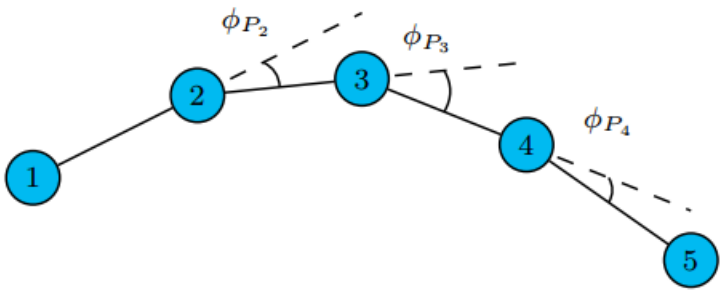


FIGURE 3.13 – Courbures locales entre deux images successives.

Figure 3.14 :

— Le plan horizontal est un plan qui divise le corps en moitiés supérieure et inférieure. Il est couramment utilisé pour se référer à la dimension droite/gauche afin de décrire les mouvements d'ouverture et de fermeture.

— Le plan frontal est un plan qui divise le corps en deux moitiés avant et arrière. Il caractérise les mouvements ascendants et descendants.

— Le plan sagittal est un plan qui divise le corps en deux moitiés droite et gauche. Il comprend les mouvements vers l'avant et vers l'arrière.

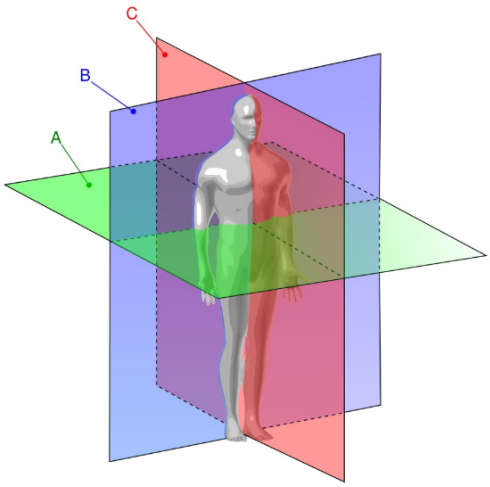


FIGURE 3.14 – Les trois plans : (A) horizontal, (B) frontal et (C) sagittal.

[Truong et al., 2016] ont quantifié ce facteur à l'aide de trois caractéristiques qui correspondent aux amplitudes corporelles selon les directions perpendiculaires aux plans vertical, horizontal et sagittal. [Durupinar et al., 2016] ont décrit la composante Forme avec seulement le facteur de la mise en forme. Ils ont introduit un coefficient pour décrire le changement de forme suivant les trois dimensions. Les coefficients de chaque dimension de forme prennent des valeurs dans l'intervalle $[-1, 1]$, où -1 signifie une posture contractée et 1 une posture croissante. Afin de décrire le changement

de forme suivant les trois plans, nous avons calculé les distances (d) entre l'articulation du torse à l'instant initial (P_1^s) et les articulations du squelette à chaque instant, projetés sur chaque plan : frontal $\{x, y\}$, sagittal $\{y, z\}$ et horizontal $\{z, x\}$. Donc finalement, 27 caractéristiques sont extraites pour la quantification du facteur de la mise en forme.

$$d = \sqrt{\sum_e ([P_t^k]_e - [P_1^s]_e)^2}$$

$[P_t^k]_e$ correspond à la position de l'articulation k à l'instant t sur le plan e. $k \in \{$la tête, les bras (mains et coudes) et les jambes (pieds et genoux)$\}$. e désigne la dimension et appartient à l'un des ensembles suivants $\{x, y\}$, $\{y, z\}$ et $\{z, x\}$ pour chaque projection considérée. La Figure 3.15 montre la variation des distances entre la position du torse capturée à l'instant initial et les différentes articulations au cours du temps dans le geste «s'accroupir» de la base MSRC-12. Une fois que toutes

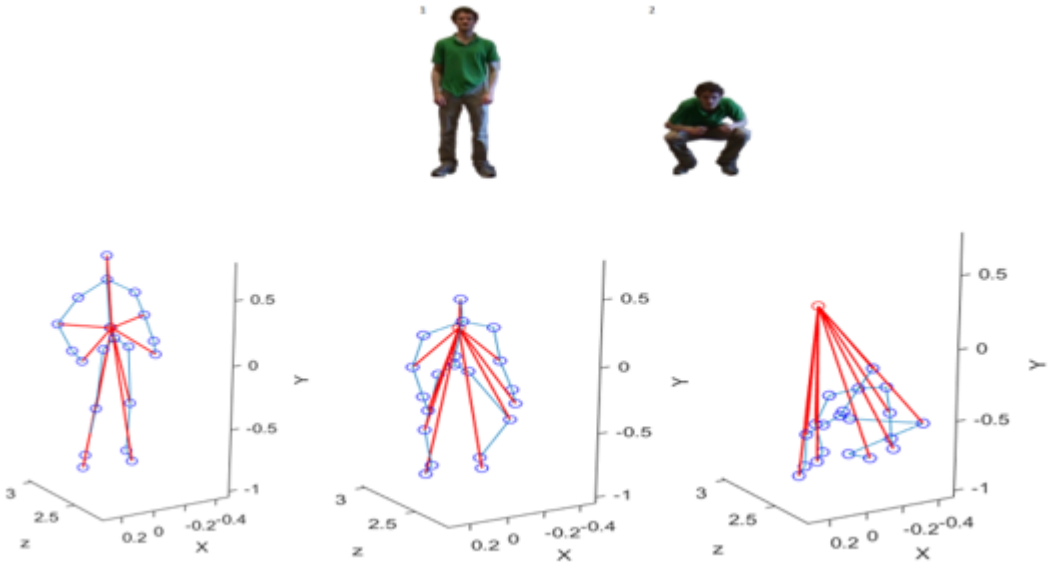


FIGURE 3.15 – Variation des distances relatives à l'articulation du torse dans le geste «s'accroupir» de la base MSRC-12.

les composantes de LMA sont quantifiées, un vecteur descripteur final composé de 47 caractéristiques est obtenu. Donc finalement, chaque geste sera représenté sous forme d'un ensemble des vecteurs descripteurs de taille 47 qui seront les entrées du modèle de Markov caché pour l'entrainement et la classification des gestes.

3.3 Application des modèles de Markov cachés

Les caractéristiques présentées ci-dessus permettent de définir un vecteur descripteur pour chaque instant d'un geste analysé. Ces descripteurs sont ensuite injectés dans un algorithme d'apprentissage pour l'entrainement et la classification des gestes. Dans l'étape d'entrainement, l'objectif est de construire des règles de décisions à partir de l'ensemble des données d'entrainement. Dans cette phase, nous entrainons des modèles à partir des exemples d'apprentissage associés à des classes prédéfinies en ajustant leurs paramètres internes. Ici, nous parlons de l'apprentissage supervisé. Dans l'étape de classification, il s'agit de comparer un exemple inconnu aux références issues de la phase d'entrainement afin de prédire sa classe. La performance d'un système de classification dépend principalement du choix de la méthode d'apprentissage. Pour notre cas, nous cherchons à reconnaitre des gestes dynamiques, nous avons un mouvement qui évolue dans le temps. Dans le cadre de la reconnaissance des séries temporelles, les modèles de Markov cachés ont déjà démontré leur capacité à conserver l'identité spatio-temporelle et à encoder la séquentialité des données. Ils se sont alors naturellement imposés dans le domaine de la reconnaissance des gestes [Chen et al., 2003, Lee and Kim, 1999, Gedat et al., 2017, Truong and Zaharia, 2016]. Les MMCs sont des modèles stochastiques qui ont été largement utilisés pour encoder des séries temporelles. En effet, leur nature Markovienne, qui lie les observations passées aux observations futures, en fait des méthodes adaptées à la modélisation de données séquentielles. Ces modèles sont riches en structures mathématiques et par conséquent peuvent être utilisés dans un large domaine d'applications, notamment, la reconnaissance vocale [Levinson et al., 1983, Rabiner and Juang, 1986, Chow et al., 1987, Lee, 1988], la reconnaissance de forme [Fielding and Ruck, 1995], la reconnaissance de l'écriture manuscrite cursive [Procter et al., 2000, Bunke et al., 1995], la modélisation de séquences biologiques [BOULARD H., 2003], etc.

3.3.1 Formalisme de MMC

Un MMC est caractérisé par des probabilités de transition entre des états cachés et de génération des symboles observés par chaque état. Un mouvement de longueur T est décrit par une séquence d'observations $O = (o_1, o_2, \ldots, o_T)$, où chaque observation o_i présente le vecteur descripteur de mouvement à l'instant i. Cette séquence d'observation peut être modélisée par un MMC σ avec les éléments suivants :

— N est le nombre des états, $S = \{s_1, s_2, \ldots, s_N\}$ est l'ensemble des états cachés, on note q_t l'état à l'instant t ($q_t \in S$).

— M est le nombre des symboles, $V = \{v_1, v_2, \ldots, v_M\}$ est l'ensemble des M observations discrètes, on note o_t le symbole observé à l'instant t ($o_t \in V$).

— $A = \{a_{ij}\}$ est la matrice des probabilités de transition de taille $[N \times N]$, a_{ij} présente la probabilité de transiter de l'état s_i à l'état s_j.

$a_{ij} = P(q_{t+1} = s_j | q_t = s_i),\ \forall 1 \leq i, j \leq N\,; \quad a_{ij} \geq 0$ et $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$

— $B = \{b_j(k)\}$ est la matrice des probabilités d'émission de taille $[N \times M]$, $b_j(k)$ est la probabilité d'émission du symbole v_k par l'état s_j.

$b_j(k) = P(o_t = v_k | q_t = s_j),\ \ 1 \leq j \leq N$ et $k \leq 1 \leq M\,; \quad b_j(k) \geq 0\ \forall j, k\ \ et\ \ \sum_{k=1}^{M} b_j(k) = 1$

— $\pi = \{\pi_i\}$ la matrice des probabilités initiales $[1 \times N]$, $1 \leq i \leq N$, π_i est la probabilité d'être à l'état i à l'instant initial.

$\pi_i = P(q_1 = s_i),\ 1 \leq i \leq N\,; \quad \sum_{i=1}^{N} \pi_i = 1$

La Figure 3.16 présente les propriétés d'un MMC, les transitions entre les états et les générations des observations par chaque état caché.

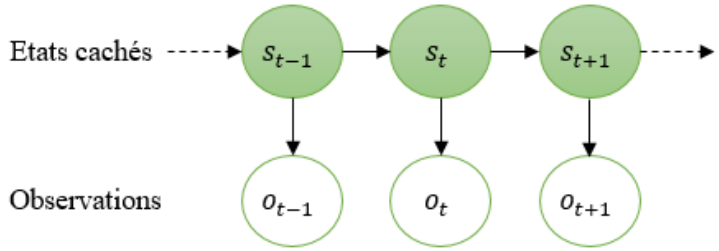


FIGURE 3.16 – Relation de dépendance entre les états cachés et les observations.

3.3.2 Topologies

En pratique, on utilise deux types de modèles de Markov cachés, le modèle ergodique et le modèle gauche-droite.

— Le modèle ergodique est un modèle entièrement connecté où toutes les transitions d'un état vers les autres sont possibles. Autrement dit, tous les a_{ij} sont strictement positifs.

$$a_{i,j} > 0\ \ \forall i \in [1, N],\ \ \forall j \in [1, M] \tag{3.8}$$

— Le modèle gauche-droite, appelé aussi modèle de Bakis, est une topologie qui n'autorise aucune transition d'un état vers un autre d'indice inférieur : les états qui se succèdent ont donc des indices égaux ou supérieurs aux précédents (pas de retour en arrière).

$$a_{i,j} = 0\ \ \forall j < i \tag{3.9}$$

Ces modèles peuvent être utilisés pour modéliser des processus possédant des propriétés variantes dans le temps, tel que les signaux de parole [Fink, 2014], le diagnostic et le pronostic de défauts [Miao and Makis, 2007], l'usure de roulements [Nelwamondo et al., 2005], etc.

D'une manière générale, un MMC est désigné par sa notation compacte $\sigma(A, B, \pi)$. Selon les types d'observations (primitives) extraites, un MMC peut être discret ou continu. Dans le cas discret, les primitives sont quantifiées et converties en des symboles d'observations. Quant au cas continu, les primitives sont utilisées directement comme des observations sans aucune transformation. Dans notre cas, nous choisissons de modéliser un geste par un MMC discret vu son avantage de rapidité [Rigoll et al., 1996].

Dans la modélisation des mouvements, chaque geste est décrit par l'ensemble des caractéristiques extraites du mouvement. Une observation d'un mouvement est alors décrite par une séquence d'observations de longueur T, $O = (o_1, o_2, \ldots, o_T)$, où chaque observation $o_t = (f_1, f_2, \ldots, f_d)(t)$, présente le vecteur descripteur composé de d caractéristiques, décrivant la configuration posturale à l'instant t. Deux étapes préalables à la phase de la reconnaissance des gestes sont nécessaires afin d'adapter les données aux types d'entrées acceptés par le MMC discret : l'échantillonnage et la quantification. La Figure 3.17 explique les différentes étapes de l'adaptation de la séquence de geste avant son implication dans le modèle MMC discret. La sortie du module «extraction des caractéristiques» correspond à une séquence de geste S_i présentée sous forme d'une matrice de taille $[N_i \times d]$ contenant des vecteurs descripteurs f_t capturés à chaque instant t, $t \in [1, \ldots, N_i]$ où N_i correspond à la taille de la séquence S_i et d est le nombre des caractéristiques dans chaque vecteur descripteur f_t. Nous appliquons par la suite un algorithme d'échantillonnage qui permet d'échantillonner les différentes séquences des gestes et donner un ensemble des gestes de taille fixe T. L'étape suivante, est la quantification des données avec l'algorithme des K-moyennes qui consiste à partitionner l'ensemble des vecteurs descripteurs en K groupes et ainsi transformer chaque vecteur descripteur en une variable discrète $o \in \{1, \ldots, K\}$. Donc, chaque séquence de geste sera présentée sous forme d'un vecteur contenant T symboles qui sera implémenté dans le modèle MMC pour l'entrainement et la classification.

3.3.3 Échantillonnage

L'algorithme d'échantillonnage consiste à réduire la taille de l'ensemble des données et transformer des séquences gestuelles de différentes longueurs en une taille fixe. Cet algorithme prend comme entrées une séquence de mouvement et la taille désirée T et donne en sortie la même séquence réduite en une taille fixe T. Une séquence gestuelle $s_{in} = \{f_1, f_2, \ldots, f_N\}$ est définie comme une matrice

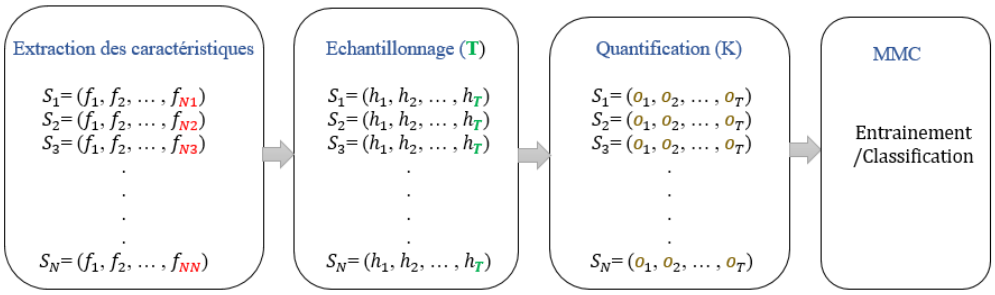


FIGURE 3.17 – Les différentes étapes appliquées à chaque séquence gestuelle avant son implication dans le modèle MMC.

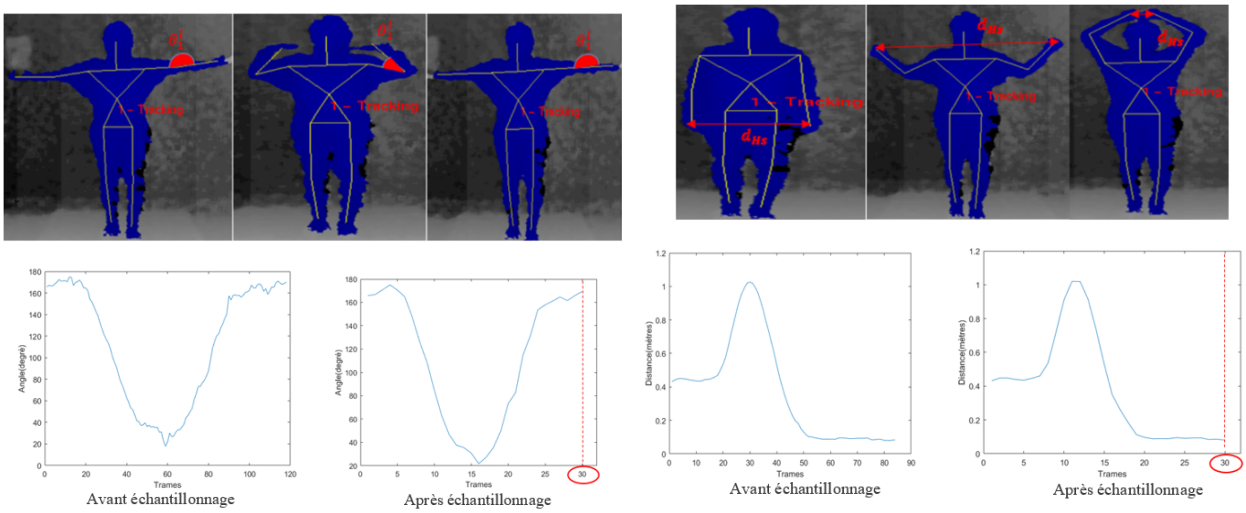
de taille $[N \times d]$ composée de N vecteurs descripteurs f_k $(k \in \{1, \dots, N\})$ avec d caractéristiques. La sortie de notre algorithme de discrétisation est une séquence gestuelle S_{out} présentée sous forme d'une matrice de taille $[T \times d]$ qui contient T vecteurs descripteurs h_k $(k \in \{1, \dots, T\})$ avec d caractéristiques. Notre algorithme comprend réellement deux étapes : la première consiste à filtrer le bruit dans les données et la deuxième à échantillonner les séquences en un nombre fixe de trames en représentant fidèlement les données. Le programme parcourt d'abord tous les échantillons capturés par la kinect $\{s_{in}\}$ dans le but de filtrer le bruit. A chaque fois, la distance entre deux points successifs (deux vecteurs descripteurs) est mesurée et comparée à un certain seuil ϵ. La valeur de ϵ est définie en calculant la variation entre la position maximale et minimale d'un squelette au repos donnée par le capteur kinect. Nous ne gardons que les points dont la distance entre eux est supérieure à ϵ. A la fin de la boucle, nous aurons un ensemble de N' points. Soit g la matrice de N' vecteurs descripteurs, la deuxième étape consiste à réduire le nombre des vecteurs en T. Cela se fait en mesurant le rapport entre N' et T pour avoir le pas désiré entre deux points (D'). Ainsi, nous sélectionnons à chaque fois le point d'indice $(1 + i \times D')$, $i \in \{0, \dots, T-1\}$. Nous obtenons à la fin l'ensemble $\{s_{out}\}$ contenant T vecteurs descripteurs. La Figure 3.18(a) montre un exemple d'échantillonnage de la caractéristique θ_1^l, l'angle entre la main gauche et l'épaule gauche dans le geste «Faire un signe» de la base CMKinect-10. Nous avons une séquence de mouvement de 118 trames. Nous appliquons notre algorithme d'échantillonnage tout en fixant T à 30 trames, nous aurons la même allure de la courbe de θ_1^l, avec une taille de 30 trames. De même, la Figure 3.18(b) présente la courbe de la caractéristique d_{Hs}, la distance entre les deux mains dans le geste «s'arrêter» d'une longueur de 87 trames. Avec l'algorithme proposé nous discrétisons la courbe de d_{Hs} en 30 trames.

3.3.4 Quantification vectorielle

Après l'étape d'échantillonnage, nous appliquons l'algorithme de K-moyennes pour regrouper les données d'apprentissage en des classes discrètes. Cet algorithme consiste donc à partitionner

Algorithm 1 Algorithme d'échantillonnage.

Inputs : $s_{in} = \{f_1, f_2, \ldots, f_N\}, T$
Outputs : $s_{out} = \{h_1, h_2, \ldots, h_T\}$
Initialization :
$g_1 = f_1$
$j \leftarrow 2$
$k \leftarrow 1$
for $i \leftarrow 2$ to N **do**
$D = \sqrt{\sum_{l=1}^{d}(f_{i,l} - f_{k,l})^2}$
•D est la distance entre les deux vecteurs descripteurs f_i et f_k.
if $D \geq \epsilon$ **then**
$g_j = f_i$
$j \leftarrow j + 1$
$k = i$
end if
end for
Calculer le pas $D' = \frac{N'}{T}$ •g est une matrice composée de N' vecteurs descripteurs obtenus après le filtrage du bruit.
for $i \leftarrow 0$ to $T - 1$ **do**
$h_{i+1} = g_{1+i*D'}$
end for



(a) θ_1^l (geste «Faire un signe») (b) d_{Hs} (geste «s'arrêter»)

FIGURE 3.18 – Discrétisation des caractéristiques dans les gestes de la base CMKinect-10.

l'ensemble des vecteurs descripteurs $\{f_1, f_2, \ldots, f_N\}$ en K groupes $\{G_1, G_2, \ldots, G_K\}$ $(K \leq N)$ tout en minimisant la distance entre les vecteurs à l'intérieur de chaque partition :

$$\underset{G}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{f_j \in G_i} \|f_j - b_i\|^2 \tag{3.10}$$

où b_i est le centroide du groupe G_i. A ce stade, nous transformons chaque vecteur descripteur en une observation discrète. Donc, pour une matrice d'apprentissage de taille $[N \times d]$ nous obtiendrons un vecteur d'observations O de taille N, $O = \{o_1, o_2, \ldots, o_N\}$, avec o_i est un symbole discret $\in \{G_1, G_2, \ldots, G_K\}$ correspond au groupe du vecteur des caractéristiques f_i. Nous obtenons alors pour chaque séquence gestuelle, N symboles discrets qui seront injectés dans le modèle MMC pour l'entrainement et la classification.

3.3.5 Entrainement et classification des gestes

Pour notre système de reconnaissance de gestes, nous modélisons chaque geste par un modèle MMC. Trois phases importantes dans le processus de la reconnaissance des gestes avec les MMCs sont :

Phase d'initialisation : Un modèle MMC est initialisé suivant plusieurs facteurs, la topologie, le nombre des états (N), le nombre des symboles (M) et la taille de la séquence d'observation (T). Nous avons choisi la structure la plus simple des MMC de Bakis, qui est le modèle linéaire, où seules les auto-transitions et les transitions aux états suivants sont autorisées. Un exemple de modèle linéaire d'un MMC à 4 états et 5 symboles est présenté dans la Figure 3.19. Il s'agit de 4 états cachés, numérotés $s_1, \ldots, s_4$. Chaque état possède une auto-transition avec la probabilité a_{ii} et une transition vers l'état suivant avec la probabilité $a_{i(i+1)} = 1 - a_{ii}$. Soit un ensemble d'observations discrètes de taille 5 $(o_1, o_2, \ldots, o_5)$, où chaque état s_i émet un nombre d'observations. Par exemple l'état s_3 émet 3 symboles o_2, o_3, et o_5 avec les probabilités respectives $b_3(o_2)$, $b_3(o_3)$ et $b_3(o_5)$. Chaque

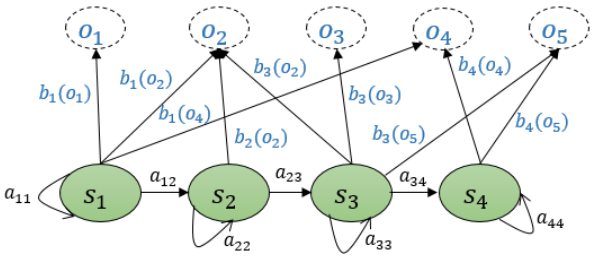


FIGURE 3.19 – Un MMC de Bakis linéaire à 4 états.

modèle initial MMC du geste g noté σ_g est alors définit par les paramètres suivants :

— Probabilités initiales : $\pi_i[1 \times N]$

$$\pi_i = \begin{cases} 1 & \text{si} \quad i = 1. \\ 0 & \text{sinon.} \end{cases}$$

— Matrice de transition : $A[N \times N]$

$$A(i,j) = \begin{cases} a_{ii} & \text{si} \quad i = j. \\ 1 - a_{ii} & \text{si} \quad j = i+1 \text{ ,} \forall i \ \sum_j a_{ij} = 1 \\ 0 & \text{sinon.} \end{cases}$$

— Matrice d'émission : $B[N \times M]$

$B(i,j) = b_{ij} \ \forall i,j$

Phase d'entrainement : L'entrainement d'un MMC consiste à ajuster ses paramètres (probabilités de l'état initial, de transition et d'émission) de façon à maximiser au mieux la probabilité de la génération des gestes d'entrainement. Donc, en utilisant ces gestes avec le modèle initialisé σ_g, nous pouvons affiner les paramètres de ce modèle par l'algorithme de Baum-Welch présenté dans l'Annexe A. Cet algorithme permet de déterminer les nouveaux paramètres de σ_g qui collent au mieux aux données d'apprentissage et donc un MMC optimal $\bar{\sigma_g}$ sera construit pour chaque geste. Soit une séquence d'observation $O = (o_1, o_2, \ldots, o_T)$ et un modèle initial MMC σ_g, le modèle qui explique le mieux la séquence O est donné par la formule suivante :

$$\bar{\sigma_g} = argmax_{\sigma_g}(P(O|\sigma_g)) \tag{3.11}$$

Phase de classification : Cette phase consiste à classifier un geste inconnu représenté par la séquence d'observation $O = (o_1, o_2, \ldots, o_T)$ en le comparant avec les modèles entrainés $\bar{\sigma_g}$ sur l'ensemble d'entrainement. Donc, nous devons choisir parmi les modèles $\bar{\sigma_g}$, $1 \leq g \leq G$, celui qui correspond au mieux à la séquence d'observation O. C'est le problème d'évaluation d'un MMC présenté dans l'Annexe A. Donc, nous calculons la probabilité de la séquence d'observation O étant donné les paramètres du modèle MMC pour chaque geste $P(O|\bar{\sigma_g})$ avec l'algorithme de Forward. Le geste correspondant à la probabilité maximale est choisi comme la décision finale.

$$g* = \underset{1 \leq g \leq G}{\text{argmax}} \, P(O|\bar{\sigma_g}) \tag{3.12}$$

3.3.6 Contribution aux modèles de Markov cachés

Dans notre système nous proposons une nouvelle méthode de classification en se basant sur l'algorithme de Forward. L'objectif est de traiter les confusions qui peuvent se produire dans certains cas, notamment, quand il s'agit de deux gestes similaires. La Figure 3.20 montre un exemple des deux gestes qui se ressemblent beaucoup surtout dans les premières trames. Nous notons g_1 et g_2 respectivement, les classes du premier geste et deuxième geste. Ici, les deux gestes partagent le même mouvement de la main gauche au début du geste. Notre idée consiste à classifier les gestes en

considérant les deux sens de mouvement (sens direct et sens indirect). Soient $\sigma^d_{g_1}$, $\sigma^d_{g_2}$ respectivement les modèles MMCs du premier et deuxième geste considérés dans le sens direct. $\sigma^i_{g_1}$ et $\sigma^i_{g_2}$ sont les modèles des même gestes considérés dans le sens indirect. Soit s_1 et s_2 les espaces d'état, nous pouvons dire que l'espace d'état s_1 est un sous-ensemble de s_2. Par conséquent, les distributions des probabilités initiales ainsi que les probabilités de transition de $\sigma^d_{g_1}$ et $\sigma^d_{g_2}$ sont presque identiques. Ce qui n'est pas le cas pour $\sigma^i_{g_1}$ et $\sigma^i_{g_2}$ vu que les MMCs dépendent du temps. Pour cela, pour un échantillon de test E_{test} de g_1 produisant une séquence d'observation O^d_{test} dans le sens direct, nous avons l'égalité suivante :

$$P(O^d_{test}|\sigma^d_{g_1}) \approx P(O^d_{test}|\sigma^d_{g_2}) \tag{3.13}$$

Donc l'utilisation d'un seul MMC qui considère la séquence d'observation dans le sens direct peut mal classer E_{test} dans la classe g_2. D'autre part, si nous considérons la séquence d'observation O^i_{test} dans le sens inverse (indirect) nous obtenons les relation suivantes :

$$P(O^i_{test}|\sigma^i_{g_1}) \approx P(O^d_{test}|\sigma^d_{g_1}) \tag{3.14}$$

$$P(O^i_{test}|\sigma^i_{g_1}) > P(O^i_{test}|\sigma^i_{g_2}) \tag{3.15}$$

A partir des équations 3.13, 3.14 et 3.15, nous tirons la relation suivante :

$$\min\{P(O^d_{test}|\sigma^d_{g_1}), P(O^i_{test}|\sigma^i_{g_1})\} > \min\{P(O^d_{test}|\sigma^d_{g_2}), P(O^i_{test}|\sigma^i_{g_2})\} \tag{3.16}$$

Ce qui permet de classifier l'échantillon E_{test} dans la classe g_1. En conclusion, lors de la phase de classification, nous construisons deux MMCs σ^d_g et σ^i_g pour chaque classe g. Le premier modèle (σ^d_g) correspond à l'ensemble des séquences d'observations présentées dans le sens direct, et le modèle σ^i_g est associé à l'ensemble des séquences d'observations considérées dans le sens indirect. La classe d'un exemple de test inconnu avec sa séquence d'observation O^d_{test} dans le sens direct et O^i_{test} dans le sens indirect est obtenue par la formule suivante :

$$g* = \underset{1\leq g\leq G}{\operatorname{argmax}} \min\{P(O^d_{test}|\sigma^d_g), P(O^i_{test}|\sigma^i_g)\} \tag{3.17}$$

Les deux probabilités $P(O^d_{test}|\sigma^d_g)$ et $P(O^i_{test}|\sigma^i_g)$ sont calculées par l'algorithme de Forward.

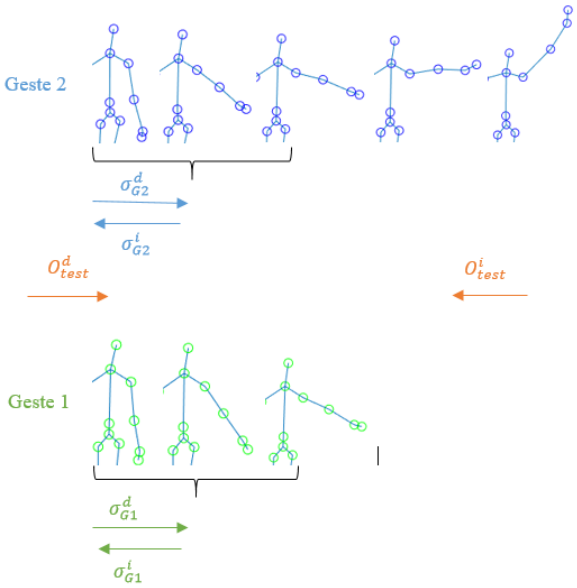


FIGURE 3.20 – Modélisation des gestes dans le sens direct et indirect.

3.4 Évaluation expérimentale

Notre système est évalué sur quatre bases de données : 3 bases publiques (MSRC12 [Fothergill et al., 2012], MSR Action 3D [Li et al., 2010] et UTkinect [Xia et al., 2012b]) et notre base dédiée à une application robotique composée de 10 gestes de contrôle, CMKinect-10. Nous avons divisé la base de données en deux parties où nous prenons la moitié des personnes pour le test et l'autre moitié pour l'entrainement. Le processus de la reconnaissance des gestes comprend plusieurs étapes :

— Acquisition des données : pour la construction de notre base, nous avons utilisé la kinect pour le tracking du squelette à une fréquence de 30 trames par seconde. Le programme est implémenté sous ROS (Robot Operating System). Nous avons enregistré les positions ainsi que les orientations des 15 articulations (tête, cou, torse, épaule gauche, coude gauche, main gauche, épaule droite, coude droit, main droite, hanche gauche, genou gauche, pied gauche, hanche droite, genou droit et pied droit).

— Extraction des caractéristiques : nous convertissons les données brutes provenant de la kinect (sous forme des positions en 3D et des rotations) en un descripteur de mouvement robuste en se basant sur les trois composantes de LMA (Corps, Espace et Forme). Finalement, nous obtenons un descripteur qui contient 47 caractéristiques. Donc, chaque séquence gestuelle est présentée sous forme d'une matrice de taille $[N \times d]$, avec N est le nombre des trames et d est le nombre des caractéristiques, qui est dans notre cas 47.

— Echantillonnage : nous appliquons l'algorithme d'échantillonnage proposé sur toutes les séquences gestuelles pour obtenir des séquences de même taille T. Cet algorithme accepte deux entrées, qui sont l'ensemble des gestes et la taille de la séquence désirée T. La sortie de cet algorithme sera des séquences de geste représentées par des matrices de taille fixe $[T \times d]$.

— Quantification : le modèle MMC discret accepte des entrées discrètes, pour cela nous appliquons l'algorithme des k-moyennes. Ceci permet de convertir une séquence de geste de taille $[T \times d]$ en une séquence d'observation discrète de taille $[1 \times T]$. Les entrées de cet algorithme sont l'ensemble des gestes échantillonnés et le nombre de groupes désiré (K).

— Apprentissage : pour le modèle MMC, une étape d'initialisation est nécessaire afin de définir ses paramètres (A, B, π). Cet algorithme a comme entrées, les données d'entrainement et le modèle initial et permet d'affiner les paramètres de ce modèle pour une vraisemblance maximale.

En résumant, les différents paramètres à ajuster dans chaque module de notre système de reconnaissance sont les suivants :

— **Module d'échantillonnage :** nous fixons la valeur de T pour avoir des séquences des gestes d'une même taille.

— **Module de quantification :**nous varions la valeur du paramètre K qui définit le nombre de groupes dans l'algorithme des K-moyennes de 10 jusqu'à 50 groupes avec un pas de 10.

— **Module de reconnaissance MMC :** nous varions d'abord le nombre des états S de 5 jusqu'à 25 états avec un pas de 5. Et nous initialisons aléatoirement les paramètres du modèle MMC (A, B, π).

3.4.1 MSRC-12

La base de données MSRC12 est présentée dans le chapitre 2, elle contient 12 classes de gestes. Nous divisons l'ensemble de données en deux parties suivant le type de geste :

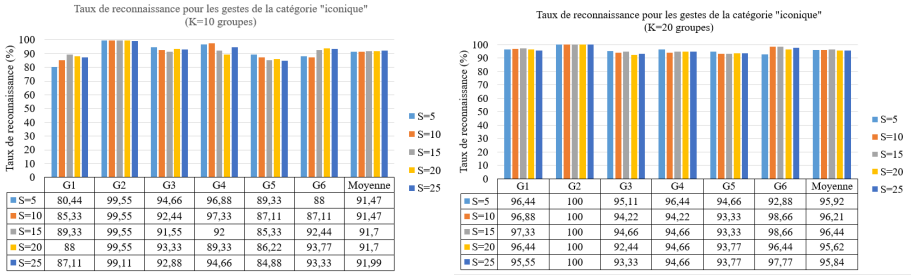
— Gestes métaphoriques : G1 s'accroupir, G2 tirer au pistolet, G3 jeter un objet, G4 changer d'arme, G5 mettre des lunettes, G6 donner un coup de pied.

— Gestes iconiques : G1 démarrer la musique, G2 naviguer vers le menu suivant, G3 terminer la musique, G4 s'incliner, G5 protester contre la musique, G6 fixer le tempo de la chanson.

D'abord, nous fixons T à 50 trames, par la suite à chaque fois nous fixons le nombre des groupes K et nous varions le nombre des états S. La Figure 3.21 résume les résultats des taux de reconnaissance obtenus pour chaque valeur de K et S pour les deux catégories iconique et métaphorique.

Evaluation des gestes iconiques : Pour $K = 10$ groupes, nous obtenons un taux de reconnaissance

moyen important autour de 91%. Nous remarquons que la variation de nombre des états n'a que peu d'influence sur le taux de reconnaissance. Même chose pour $K = 20, 30$ et 40, les résultats sont similaires en variant les valeurs de S. En fixant K à 20 et à 30, nous marquons un taux de reconnaissance moyen qui varie entre 95% et 96%. Pour $K = 40$ groupes, nous obtenons un taux de reconnaissance moyen supérieur à 96% pour toutes les valeurs de S. Le meilleur taux de reconnaissance est de 96.81%, obtenu pour $K = 40$ groupes et $S = 5$ et $S = 10$. Par la suite, nous appliquons notre nouvelle méthode de classification expliquée dans la section 3.3.6. Nous gardons le même partitionnement des données et nous prenons les meilleurs des cas dans la classification des gestes iconiques : $(K = 40, S = 5)$ et $(K = 40, S = 10)$. La Figure 3.22 présente une comparaison

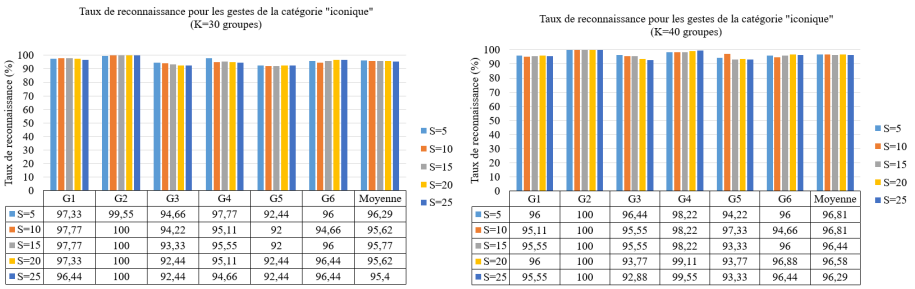


	G1	G2	G3	G4	G5	G6	Moyenne
S=5	80,44	99,55	94,66	96,88	89,33	88	91,47
S=10	85,33	99,55	92,44	97,33	87,11	87,11	91,47
S=15	89,33	99,55	91,55	92	85,33	92,44	91,7
S=20	88	99,55	93,33	89,33	86,22	93,77	91,7
S=25	87,11	99,11	92,88	94,66	84,88	93,33	91,99

(a) K=10 groupes

	G1	G2	G3	G4	G5	G6	Moyenne
S=5	96,44	100	95,11	96,44	94,66	92,88	95,92
S=10	96,88	100	94,22	94,22	93,33	98,66	96,21
S=15	97,33	100	94,66	94,66	93,33	98,66	96,44
S=20	96,44	100	92,44	94,66	93,77	96,44	95,62
S=25	95,55	100	93,33	94,66	93,77	97,77	95,84

(b) K=20 groupes



	G1	G2	G3	G4	G5	G6	Moyenne
S=5	97,33	99,55	94,66	97,77	92,44	96	96,29
S=10	97,77	100	94,22	95,11	92	94,66	95,62
S=15	97,77	100	93,33	95,55	92	96	95,77
S=20	97,33	100	92,44	95,11	92,44	96,44	95,62
S=25	96,44	100	92,44	94,66	92,44	96,44	95,4

(c) K=30 groupes

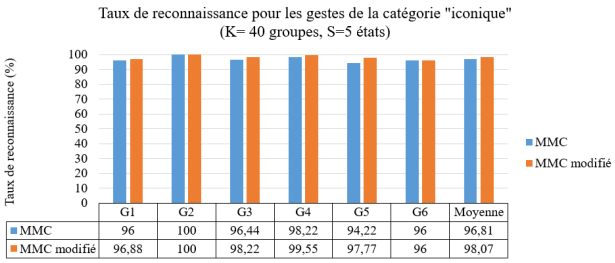
	G1	G2	G3	G4	G5	G6	Moyenne
S=5	96	100	96,44	98,22	94,22	96	96,81
S=10	95,11	100	95,55	98,22	97,33	94,66	96,81
S=15	95,55	100	95,55	98,22	93,33	96	96,44
S=20	96	100	93,77	99,11	93,77	96,88	96,58
S=25	95,55	100	92,88	99,55	93,33	96,44	96,29

(d) K=40 groupes

FIGURE 3.21 – Les taux de reconnaissance des gestes iconiques pour chaque valeur de K et S.

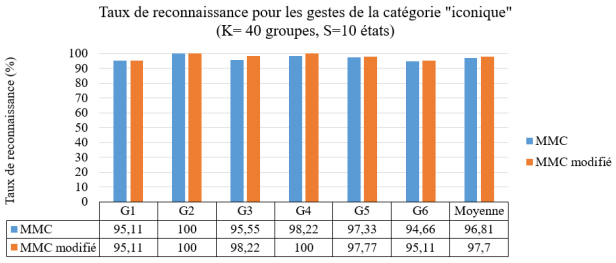
entre les résultats obtenus avec le MMC basique et le MMC modifié. Nous marquons une amélioration considérable dans le taux de reconnaissance moyen qui a augmenté de 1.26% pour $(K = 40, S = 5)$ et de 0.89% pour $(K = 40, S = 10)$.

Evaluation des gestes métaphoriques : La même chose pour les gestes métaphoriques, nous fixons à chaque fois le nombre des groupes K et nous varions le nombre des états S. La Figure 3.23 présente les résultats obtenus pour chaque test. Le meilleur résultat de 83.18% est obtenu pour $K = 40$ groupes et $S = 20$ états. Nous considérons le modèle où $(K = 40, S = 20)$, nous parcourons les différentes séquences des gestes dans les deux sens, direct et indirect, et nous appliquons les formules de la nouvelle approche de classification. Comme le montre la Figure 3.24, notre méthode a



	G1	G2	G3	G4	G5	G6	Moyenne
MMC	96	100	96,44	98,22	94,22	96	96,81
MMC modifié	96,88	100	98,22	99,55	97,77	96	98,07

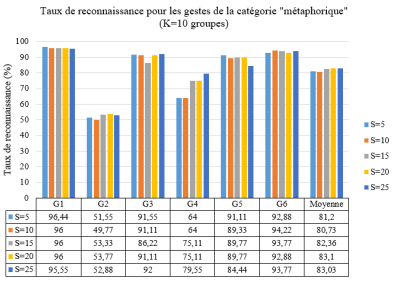
(a) $(K = 40, S = 5)$



	G1	G2	G3	G4	G5	G6	Moyenne
MMC	95,11	100	95,55	98,22	97,33	94,66	96,81
MMC modifié	95,11	100	98,22	100	97,77	95,11	97,7

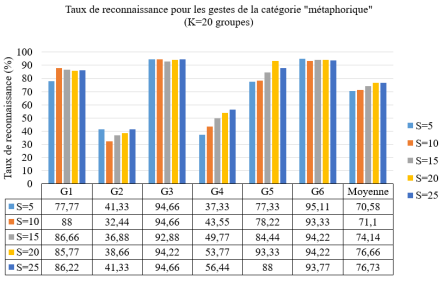
(b) $(K = 40, S = 10)$

FIGURE 3.22 – Comparaison entre les taux de reconnaissance des gestes iconiques obtenu avec MMC basique et MMC modifié.



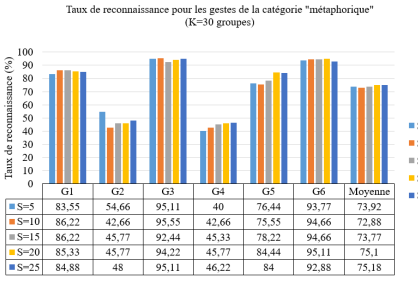
	G1	G2	G3	G4	G5	G6	Moyenne
S=5	96,44	51,55	91,55	64	91,11	92,88	81,2
S=10	96	49,77	91,11	64	89,33	94,22	80,73
S=15	96	53,33	86,22	75,11	89,77	93,77	82,36
S=20	96	53,77	91,11	75,11	89,77	92,88	83,1
S=25	95,55	52,88	92	79,55	84,44	93,77	83,03

(a) K=10 groupes



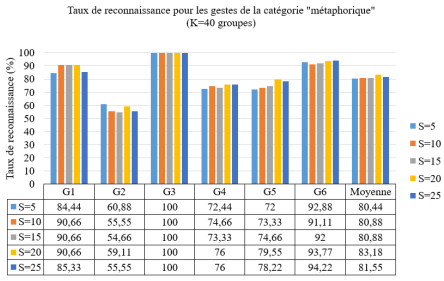
	G1	G2	G3	G4	G5	G6	Moyenne
S=5	77,77	41,33	94,66	37,33	77,33	95,11	70,58
S=10	88	32,44	94,66	43,55	78,22	93,33	71,1
S=15	86,66	36,88	92,88	49,77	84,44	94,22	74,14
S=20	85,77	38,66	94,22	53,77	93,33	94,22	76,66
S=25	86,22	41,33	94,66	56,44	88	93,77	76,73

(b) K=20 groupes



	G1	G2	G3	G4	G5	G6	Moyenne
S=5	83,55	54,66	95,11	40	76,44	93,77	73,92
S=10	86,22	42,66	95,55	42,66	75,55	94,66	72,88
S=15	86,22	45,77	92,44	45,33	78,22	94,66	73,77
S=20	85,33	45,77	94,22	45,77	84,44	95,11	75,1
S=25	84,88	48	95,11	46,22	84	92,88	75,18

(c) K=30 groupes



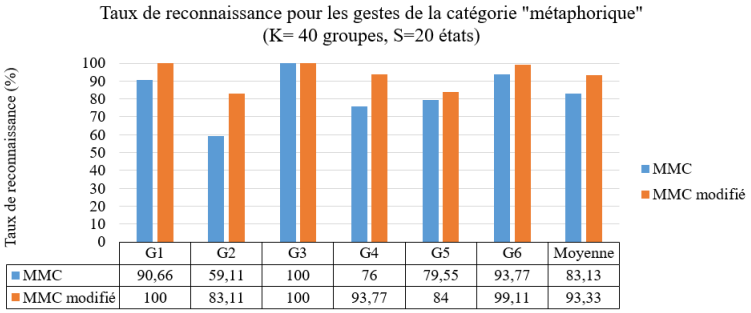
	G1	G2	G3	G4	G5	G6	Moyenne
S=5	84,44	60,88	100	72,44	72	92,88	80,44
S=10	90,66	55,55	100	74,66	73,33	91,11	80,88
S=15	90,66	54,66	100	73,33	74,66	92	80,88
S=20	90,66	59,11	100	76	79,55	93,77	83,18
S=25	85,33	55,55	100	76	78,22	94,22	81,55

(d) K=40 groupes

FIGURE 3.23 – Les taux de reconnaissance des gestes métaphoriques pour chaque valeur de K et S.

contribué à une amélioration dans l'ensemble des gestes métaphoriques. Avec cette nouvelle méthode, nous observons une augmentation importante du taux de reconnaissance moyen de 10.2%.

Notre méthode s'est montrée bien compétitive avec des méthodes de l'état de l'art évaluées sur la base MSRC-12 [Fothergill et al., 2012]. Certains ont divisé l'ensemble de MSRC-12 en deux catégories iconique et métaphorique ce qui est notre cas. Comme le montre la Table 3.1, notre méthode se place au premier rang dans les méthodes évaluées sur la base MSRC-12 avec un taux de reconnaissance de 98.07% pour les gestes iconiques et 93.33% pour les gestes métaphoriques. Pour une comparaison fidèle, nous prenons le résultat de Truong et al. [Truong and Zaharia, 2016] qui ont utilisé la même méthode d'apprentissage MMC. Notre méthode surpasse leurs résultats de 9.47% pour les gestes iconiques et 18.13% pour les gestes métaphoriques. D'autres travaux ont traité



	G1	G2	G3	G4	G5	G6	Moyenne
MMC	90,66	59,11	100	76	79,55	93,77	83,13
MMC modifié	100	83,11	100	93,77	84	99,11	93,33

FIGURE 3.24 – Comparaison entre les taux de reconnaissance des gestes métaphoriques obtenu avec MMC basique et MMC modifié pour $(K = 40, S = 20)$.

l'ensemble entier des gestes dans un seul groupe, comme [Hussein et al., 2013b] où ils ont marqué un taux de reconnaissance moyen de 91.7%. [Zhao et al., 2013] ont suivi le même protocole expérimental utilisé par [Fothergill et al., 2012]. Ils ont divisé l'ensemble des gestes suivant la modalité utilisée (texte, image, vidéo, image+text, vidéo+text). Ils ont obtenu un meilleur résultat de 73% avec la modalité (image+texte). A partir de ces études, nous pouvons confirmer la robustesse de notre descripteur de mouvement ainsi que notre système de reconnaissance dans la classification des gestes métaphoriques et iconiques.

TABLE 3.1 – Comparaison avec les méthodes de l'état de l'art sur la base MSRC-12.

Méthodes	**Iconique**	**Métaphorique**
[Lehrmann et al., 2014]	90.90	-
[Song et al., 2013]	79.77	81
[Truong and Zaharia, 2016]	88.6	75.2
Notre méthode (MMC)	**96.81**	**83.13**
Notre méthode (MMC modifié)	**98.07**	**93.33**

3.4.2 MSR Action 3D

Nous évaluons notre méthode sur la base de MSR Action 3D présentée aussi dans le chapitre 2. Nous avons divisé l'ensemble des données en 3 groupes :

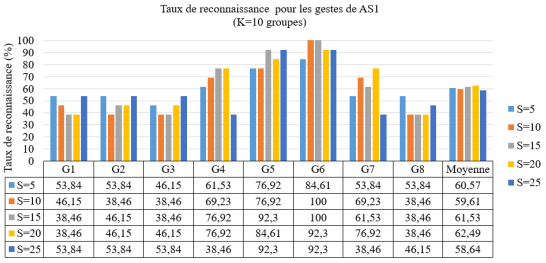
— Groupe AS1 : G1 faire un signe horizontal, G2 coup de marteau, G3 coup de poing vers l'avant, G4 lancer au loin, G5 taper les mains, G6 se pencher, G7 service au tennis, G8 ramasser et jeter.

— Groupe AS2 : G1 faire un signe vers le haut, G2 attraper d'une main, G3 dessiner un X, G4 dessiner une coche, G5 dessiner un cercle, G6 faire un signe avec les deux mains, G7 coup de

pied vers l'avant, G8 boxer sur le côté.

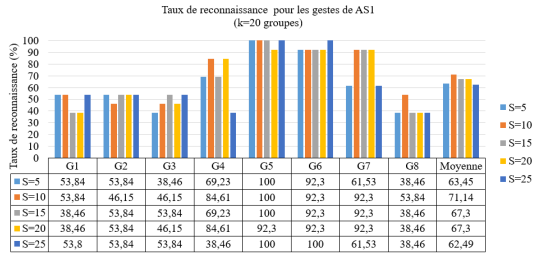
— Groupe AS3 : G1 lancer au loin, G2 coup de pied vers l'avant, G3 coup de pied sur le côté, G4 jogging, G5 swing de tennis, G6 service au tennis, G7 swing de golf, G8 ramasser et jeter.

Les ensembles AS1 et AS2 contiennent des gestes similaires, tandis que le groupe AS3 comprend des gestes complexes. Pour l'évaluation du groupe AS1, nous varions K de 10 à 50 groupes (avec un pas de 10) et S de 5 à 25 états (avec un pas de 5). Pour l'ensemble des gestes du groupe AS1, nous remarquons que les gestes «G5 taper les mains» et « G6 se pencher» enregistrent toujours des meilleurs résultats (Voir Figure 3.25). Ils ont atteint un taux de reconnaissance de 100% pour certains cas. Par contre, les gestes «G3 coup de poing vers l'avant» et «G8 ramasser et jeter» ont marqué des taux de reconnaissance faibles. Le meilleur taux de reconnaissance du groupe AS1 est de 83.65%, obtenu pour $K = 40$ groupes et $S = 5$ états.



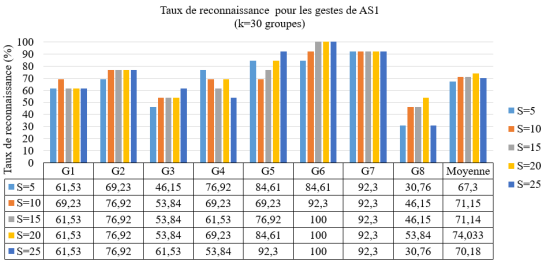
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	53,84	53,84	46,15	61,53	76,92	84,61	53,84	53,84	60,57
S=10	46,15	38,46	38,46	69,23	76,92	100	69,23	38,46	59,61
S=15	38,46	46,15	38,46	76,92	92,3	100	61,53	38,46	61,53
S=20	38,46	46,15	46,15	76,92	84,61	92,3	76,92	38,46	62,49
S=25	53,84	53,84	53,84	38,46	92,3	92,3	38,46	46,15	58,64

(a) K=10 groupes



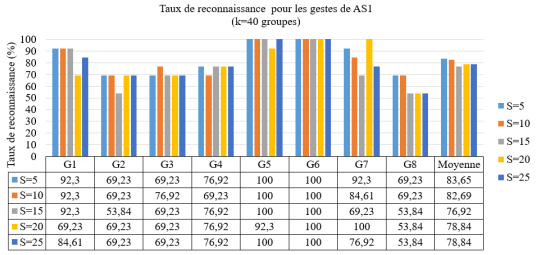
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	53,84	53,84	38,46	69,23	100	92,3	61,53	38,46	63,45
S=10	53,84	46,15	46,15	84,61	100	92,3	92,3	53,84	71,14
S=15	38,46	53,84	53,84	69,23	100	92,3	92,3	38,46	67,3
S=20	38,46	53,84	46,15	84,61	92,3	92,3	92,3	38,46	67,3
S=25	53,8	53,84	53,84	38,46	100	100	61,53	38,46	62,49

(b) K=20 groupes



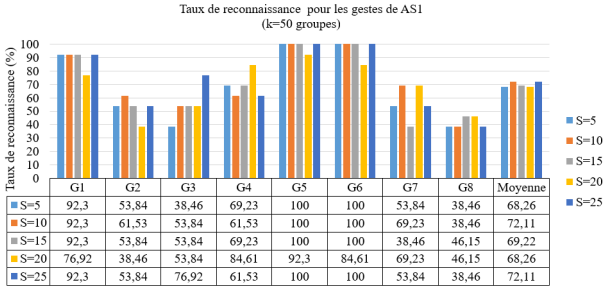
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	61,53	69,23	46,15	76,92	84,61	84,61	92,3	30,76	67,3
S=10	69,23	76,92	53,84	69,23	69,23	92,3	92,3	46,15	71,15
S=15	61,53	76,92	53,84	61,53	76,92	100	92,3	46,15	71,14
S=20	61,53	76,92	53,84	69,23	84,61	100	92,3	53,84	74,033
S=25	61,53	76,92	61,53	53,84	92,3	100	92,3	30,76	70,18

(c) K=30 groupes



	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	92,3	69,23	69,23	76,92	100	100	92,3	69,23	83,65
S=10	92,3	69,23	76,92	69,23	100	100	84,61	69,23	82,69
S=15	92,3	53,84	69,23	76,92	100	100	69,23	53,84	76,92
S=20	69,23	69,23	69,23	76,92	92,3	100	100	53,84	78,84
S=25	84,61	69,23	69,23	76,92	100	100	76,92	53,84	78,84

(d) K=40 groupes



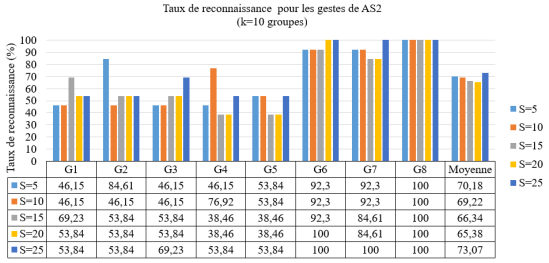
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	92,3	53,84	38,46	69,23	100	100	53,84	38,46	68,26
S=10	92,3	61,53	53,84	61,53	100	100	69,23	38,46	72,11
S=15	92,3	53,84	53,84	69,23	100	100	38,46	46,15	69,22
S=20	76,92	38,46	53,84	84,61	92,3	84,61	69,23	46,15	68,26
S=25	92,3	53,84	76,92	61,53	100	100	53,84	38,46	72,11

(e) K=50 groupes

FIGURE 3.25 – Taux de reconnaissance pour les gestes de AS1 pour chaque valeur de K et S.

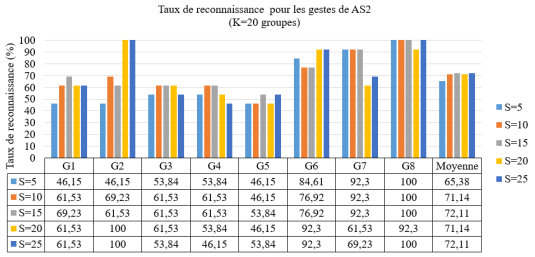
Pour le groupe AS2 (Voir Figure 3.26), le geste «G8 boxer sur le côté» a été reconnu à 100% dans

presque toutes les différentes valeurs de K et S. Aussi, les gestes «G6 Faire un signe avec les deux mains» et «G7 coup de pied vers l'avant» ont enregistré des taux de reconnaissances importants qui ont atteint 100% dans certains cas. En contrepartie, les gestes «G3 dessiner un X», « G4 dessiner une coche» et «G5 dessiner un cercle» ont marqué des taux de reconnaissance moins importants. Cela peut être expliqué par la forte similitude des mouvements dans les trois gestes. Le meilleur taux de reconnaissance pour ce groupe a été de 80.76%, obtenu pour $K = 40$ groupes et $S = 15$ états. Finalement, pour le troisième groupe de AS3, nous marquons des taux de reconnaissances



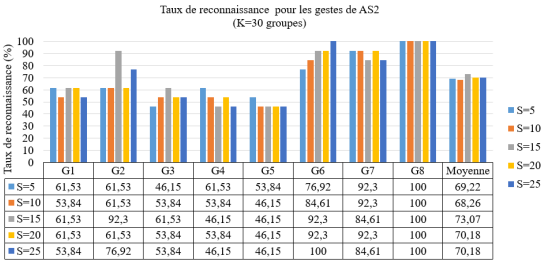
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	46,15	84,61	46,15	46,15	53,84	92,3	92,3	100	70,18
S=10	46,15	46,15	46,15	76,92	53,84	92,3	92,3	100	69,22
S=15	69,23	53,84	53,84	38,46	38,46	92,3	84,61	100	66,34
S=20	53,84	53,84	53,84	38,46	38,46	100	84,61	100	65,38
S=25	53,84	53,84	69,23	53,84	53,84	100	100	100	73,07

(a) K=10 groupes



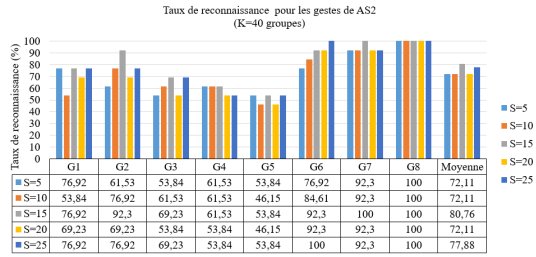
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	46,15	46,15	53,84	53,84	46,15	84,61	92,3	100	65,38
S=10	61,53	69,23	61,53	61,53	46,15	76,92	92,3	100	71,14
S=15	69,23	61,53	61,53	61,53	53,84	76,92	92,3	100	72,11
S=20	61,53	100	61,53	53,84	46,15	92,3	61,53	92,3	71,14
S=25	61,53	100	53,84	46,15	53,84	92,3	69,23	100	72,11

(b) K=20 groupes



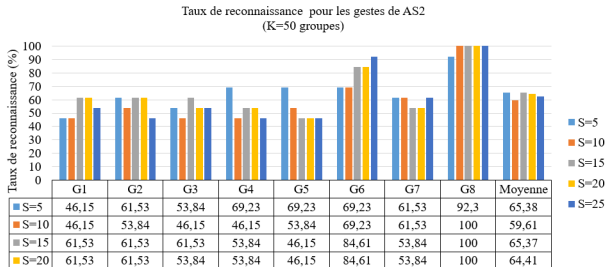
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	61,53	61,53	46,15	61,53	53,84	76,92	92,3	100	69,22
S=10	53,84	61,53	53,84	53,84	46,15	84,61	92,3	100	68,26
S=15	61,53	92,3	61,53	46,15	46,15	92,3	84,61	100	73,07
S=20	61,53	61,53	53,84	53,84	46,15	92,3	92,3	100	70,18
S=25	53,84	76,92	53,84	46,15	46,15	100	84,61	100	70,18

(c) K=30 groupes



	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	76,92	61,53	53,84	61,53	53,84	76,92	92,3	100	72,11
S=10	53,84	76,92	61,53	61,53	46,15	84,61	92,3	100	72,11
S=15	76,92	92,3	69,23	61,53	53,84	92,3	100	100	80,76
S=20	69,23	69,23	53,84	53,84	46,15	92,3	92,3	100	72,11
S=25	76,92	76,92	69,23	53,84	53,84	100	92,3	100	77,88

(d) K=40 groupes

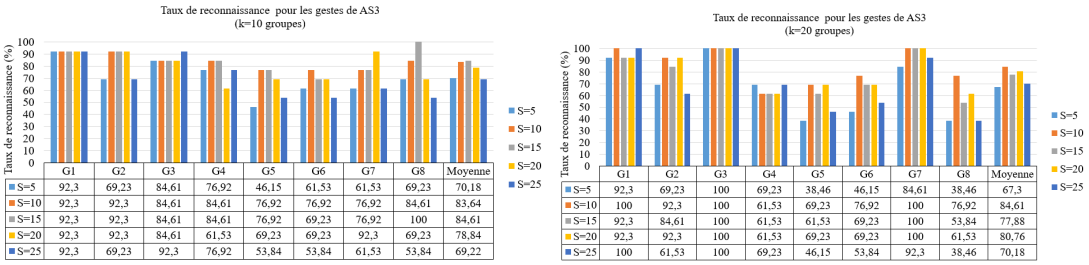


	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	46,15	61,53	53,84	69,23	69,23	69,23	61,53	92,3	65,38
S=10	46,15	53,84	46,15	46,15	53,84	69,23	61,53	100	59,61
S=15	61,53	61,53	61,53	53,84	46,15	84,61	53,84	100	65,37
S=20	61,53	61,53	53,84	53,84	46,15	84,61	53,84	100	64,41
S=25	53,84	46,15	53,84	46,15	46,15	92,3	61,53	100	62,49

(e) K=50 groupes

FIGURE 3.26 – Taux de reconnaissance pour les gestes de AS2 pour chaque valeur de K et S.

plus importants que ceux obtenus dans les deux groupes AS1 et AS2. La Figure 3.27 montre que certains gestes ont parfois atteint un taux de reconnaissance de 100%, comme les gestes «G1 lancer au loin», «G1=2 coup de pied vers l'avant», «G3 coup de pied sur le côté», «G6 service au tennis» et «G7 swing de golf». Le meilleur taux de reconnaissance de 84.61% a été obtenu dans les deux cas suivant : $(K = 10, S = 15)$ et $(K = 20, S = 10)$. Le deuxième test consiste à prendre les meilleurs

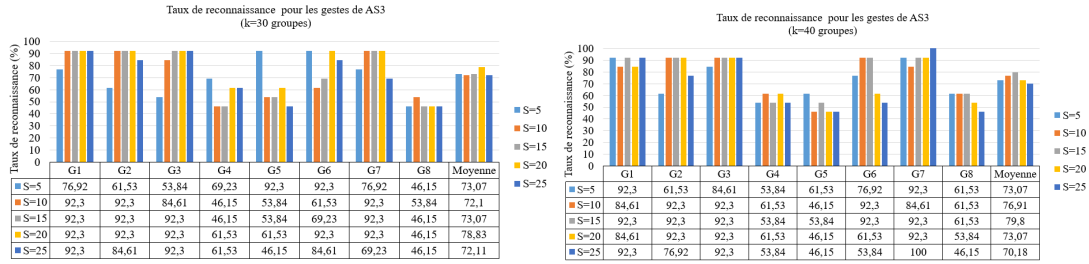


	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	92,3	69,23	84,61	76,92	46,15	61,53	61,53	69,23	70,18
S=10	92,3	92,3	84,61	84,61	76,92	76,92	76,92	84,61	83,64
S=15	92,3	92,3	84,61	84,61	76,92	69,23	76,92	100	84,61
S=20	92,3	92,3	84,61	61,53	69,23	69,23	92,3	69,23	78,84
S=25	92,3	69,23	92,3	76,92	53,84	53,84	61,53	53,84	69,22

(a) K=10 groupes

	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	92,3	69,23	100	69,23	38,46	46,15	84,61	38,46	67,3
S=10	100	92,3	100	61,53	69,23	76,92	100	76,92	84,61
S=15	92,3	84,61	100	61,53	61,53	69,23	100	53,84	77,88
S=20	92,3	92,3	100	61,53	69,23	69,23	100	61,53	80,76
S=25	100	61,53	100	69,23	46,15	53,84	92,3	38,46	70,18

(b) K=20 groupes

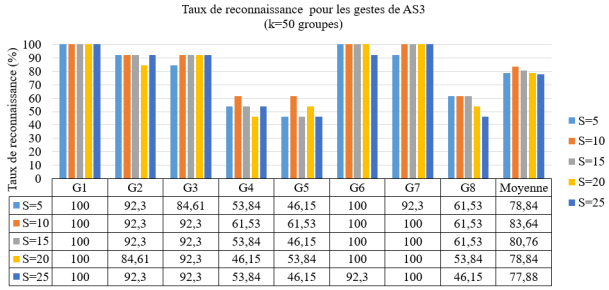


	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	76,92	61,53	53,84	69,23	92,3	92,3	76,92	46,15	73,07
S=10	92,3	92,3	84,61	46,15	53,84	61,53	92,3	53,84	72,1
S=15	92,3	92,3	92,3	46,15	53,84	69,23	92,3	46,15	73,07
S=20	92,3	92,3	92,3	61,53	61,53	92,3	92,3	46,15	78,83
S=25	92,3	84,61	92,3	61,53	46,15	84,61	69,23	46,15	72,11

(c) K=30 groupes

	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	92,3	61,53	84,61	53,84	61,53	76,92	92,3	61,53	73,07
S=10	84,61	92,3	92,3	61,53	46,15	92,3	84,61	61,53	76,91
S=15	92,3	92,3	92,3	53,84	53,84	92,3	92,3	61,53	79,8
S=20	84,61	92,3	92,3	61,53	46,15	61,53	92,3	53,84	73,07
S=25	92,3	76,92	92,3	53,84	46,15	53,84	100	46,15	70,18

(d) K=40 groupes

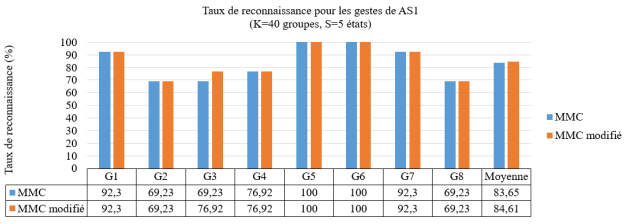


	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
S=5	100	92,3	84,61	53,84	46,15	100	92,3	61,53	78,84
S=10	100	92,3	92,3	61,53	61,53	100	100	61,53	83,64
S=15	100	92,3	92,3	53,84	46,15	100	100	61,53	80,76
S=20	100	84,61	92,3	46,15	53,84	100	100	53,84	78,84
S=25	100	92,3	92,3	53,84	46,15	92,3	100	46,15	77,88

(e) K=50 groupes

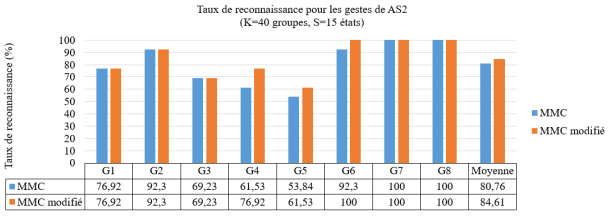
FIGURE 3.27 – Taux de reconnaissance pour les gestes de AS3 pour chaque valeur de K et S.

résultats obtenus et appliquer notre méthode de classification dans l'ensemble des gestes. Nous pouvons observer dans la Figure 3.28 l'augmentation des taux de reconnaissance de 0.96%, 3.85% et 5.69% respectivement, dans les trois groupes AS1, AS2 et AS3. Nous avons comparé notre méthode avec les méthodes de l'état de l'art sur la base MSR Action 3D. Comme le montre la Table 3.2, notre méthode surpasse moyennement celles de [Li et al., 2010, Alwani et al., 2014, Xia et al., 2012b, Soh and Demiris, 2012, Yang and Tian, 2014]. Nous achevons un taux de reconnaissance moyen de 83% proche des résultats [Chaaraoui et al., 2012, Yang and Tian, 2014, Negin et al., 2015]. Après l'application de notre méthode de classification, nous achevons un taux de reconnaissance plus important de 86.50%. Ce qui positionne notre résultat au deuxième rang après le travail de [Chaaraoui et al., 2014] qui ont achevé un taux de reconnaissance moyen de 93.23%. Une comparaison fiable se fait avec le travail de [Xia et al., 2012b] qui ont appliqué le MMC discret pour l'entrainement et la classification des gestes. Nous surpassons leur résultats de 7.53%.



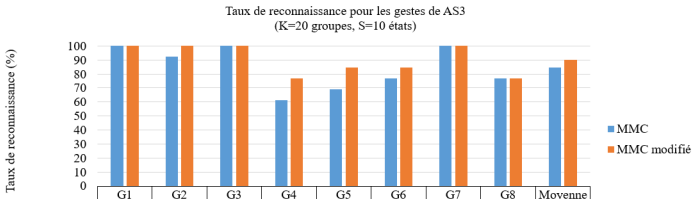
	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
MMC	92,3	69,23	69,23	76,92	100	100	92,3	69,23	83,65
MMC modifié	92,3	69,23	76,92	76,92	100	100	92,3	69,23	84,61

(a) AS1 pour $(K = 40, S = 5)$



	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
MMC	76,92	92,3	69,23	61,53	53,84	92,3	100	100	80,76
MMC modifié	76,92	92,3	69,23	76,92	61,53	100	100	100	84,61

(b) AS2 pour $(K = 40, S = 15)$



	G1	G2	G3	G4	G5	G6	G7	G8	Moyenne
MMC	100	92,3	100	61,53	69,23	76,92	100	76,92	84,61
MMC modifié	100	100	100	76,92	84,61	84,61	100	76,92	90,3

(c) AS3 pour $(K = 20, S = 10)$

FIGURE 3.28 – Comparaison entre les taux de reconnaissance des gestes AS1, AS2 et AS3 obtenus avec MMC basique et MMC modifié.

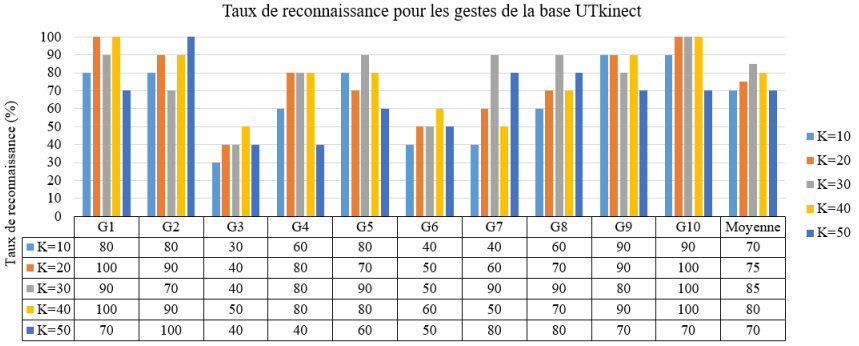
TABLE 3.2 – Comparaison avec les méthodes de l'état de l'art sur la base MSR Action 3D.

Méthodes	**AS1**	**AS2**	**AS3**	**Moyenne**
[Li et al., 2010]	72.9	71.9	79.2	74.66
[Alwani et al., 2014]	86.30	65.40	77.70	76.46
[Xia et al., 2012b]	87.48	85.48	63.46	78.97
[Soh and Demiris, 2012]	80.6	74.9	87.1	80.87
[Yang and Tian, 2014]	74.5	76.1	96.4	82.33
Notre méthode (MMC)	**83.65**	**80.76**	**84.61**	**83**
[Chaaraoui et al., 2012]	87.90	74.12	89.21	83.74
[Negin et al., 2015]	82.66	83.33	87.17	84.38
[Ghorbel et al., 2015]	83.08	79.46	93.69	85.41
Notre méthode (MMC modifié)	**84.61**	**84.61**	**90.3**	**86.50**
[Chaaraoui et al., 2014]	91.59	90.83	97.28	93.23

3.4.3 UTKinect

Nous évaluons notre système de reconnaissance de gestes sur la base UTkinect [Xia et al., 2012b] composée de 10 actions (G1 porter, G2 taper les mains, G3 ramasser, G4 tirer, G5 pousser, G6 s'asseoir, G7 se lever, G8 jeter, G9 marcher et G10 faire un signe avec les deux mains). Nous fixons la taille de la séquence T à la valeur minimale commune entre les différentes actions qui est 5 trames. Nous varions le nombre des groupes K de 10 à 50 groupes. La Figure 3.29 montre les différents taux de reconnaissance obtenus pour chaque valeur de K. Le meilleur résultat est de 85% obtenu pour $K = 30$ groupes. Nous remarquons que l'action «G10 faire un signe avec les deux mains» a atteint un

taux de reconnaissance de 100%. Par contre, les actions «G3 ramasser» et «G6 s'asseoir» ont marqué des taux de reconnaissance moins importants respectivement de 40% et 50%. Avec notre nouvelle méthode de classification, nous obtenons les mêmes résultats dans tous les gestes. Nous n'avons pas marqué une amélioration au niveau des taux de reconnaissance, ce qui peut être expliqué par la courte taille des séquences des mouvements. En effet, dans cette base nous avons fixé T à 5 trames, cela correspond à la taille de la plus courte séquence des mouvements dans la base. Dans tel cas, l'algorithme d'apprentissage peut donner des résultats moins performants.

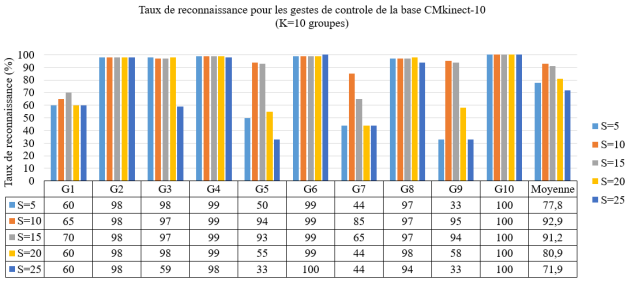


	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
K=10	80	80	30	60	80	40	40	60	90	90	70
K=20	100	90	40	80	70	50	60	70	90	100	75
K=30	90	70	40	80	90	50	90	90	80	100	85
K=40	100	90	50	80	80	60	50	70	90	100	80
K=50	70	100	40	40	60	50	80	80	70	70	70

FIGURE 3.29 – Taux de reconnaissance pour les gestes de la base UTkinect en variant le nombre des groupes K de 10 à 50 et fixant S à 5.

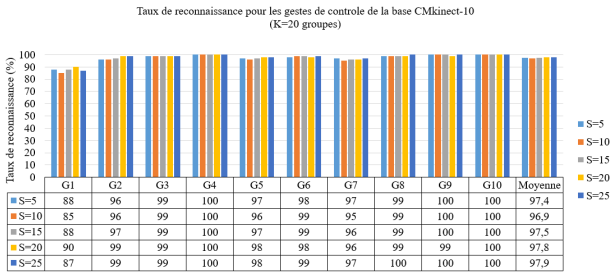
3.4.4 CMKinect-10

Nous avons également évalué notre système sur notre base de données CMKinect-10 dédiée à la télé-opération robotique composée de dix gestes de contrôle (G1 danser, G2 se présenter, G3 diminuer la vitesse, G4 avancer, G5 augmenter la vitesse, G6 s'asseoir, G7 s'arrêter, G8 tourner à gauche, G9 tourner à droite, G10 faire un signe avec les deux mains). D'abord, nous fixons le paramètre T pour l'algorithme d'échantillonnage à 30 trames. De même, nous varions les deux paramètres K (de 10 à 40 groupes) et S (de 5 à 25 états). Pour $k = 20$, 30 et 40, nous avons obtenu des taux de reconnaissance élevés pour tous les gestes supérieurs à 94% (Voir Figure 3.30). Le geste G10 «faire un signe avec les deux mains» a toujours été reconnu à 100% pour les différentes valeurs de K et S. Le meilleur résultat de 99% est obtenu pour $K = 40$ et $S = 20$. Ici encore, nous notons une amélioration grâce à notre méthode de classification proposée qui a donné un taux de reconnaissance de 99.7% au lieu de 99% (Voir Figure 3.31).



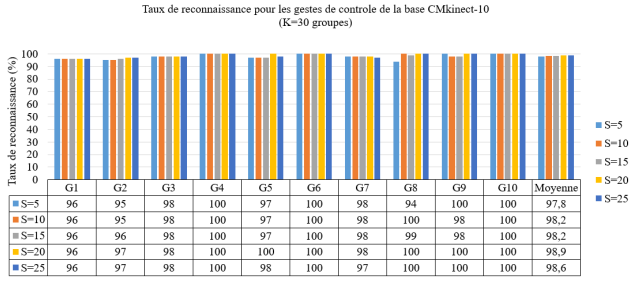
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
S=5	60	98	98	99	50	99	44	97	33	100	77,8
S=10	65	98	97	99	94	99	85	97	95	100	92,9
S=15	70	98	97	99	93	99	65	97	94	100	91,2
S=20	60	98	98	99	55	99	44	98	58	100	80,9
S=25	60	98	59	98	33	100	44	94	33	100	71,9

(a) K=10 groupes



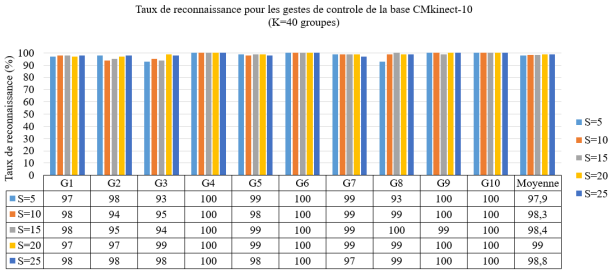
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
S=5	88	96	99	100	97	98	97	99	100	100	97,4
S=10	85	96	99	100	96	99	95	99	100	100	96,9
S=15	88	97	99	100	97	99	96	99	100	100	97,5
S=20	90	99	99	100	98	98	96	99	99	100	97,8
S=25	87	99	99	100	98	99	97	100	100	100	97,9

(b) K=20 groupes



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
S=5	96	95	98	100	97	100	98	94	100	100	97,8
S=10	96	95	98	100	97	100	98	100	98	100	98,2
S=15	96	96	98	100	97	100	98	99	98	100	98,2
S=20	96	97	98	100	100	100	98	100	100	100	98,9
S=25	96	97	98	100	98	100	97	100	100	100	98,6

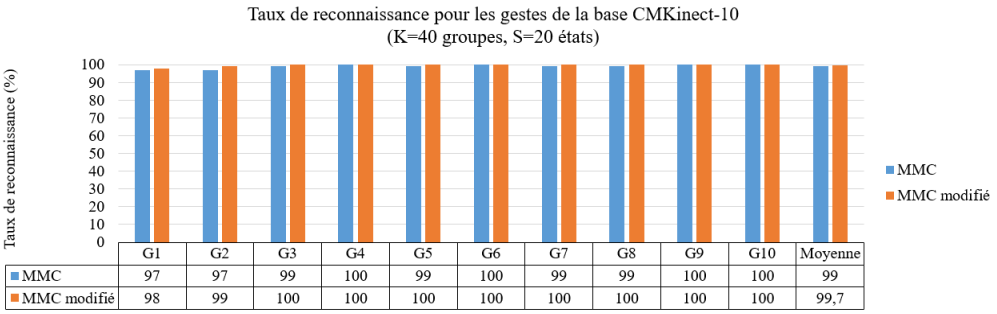
(c) K=30 groupes



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
S=5	97	98	93	100	99	100	99	93	100	100	97,9
S=10	98	94	95	100	98	100	99	99	100	100	98,3
S=15	98	95	94	100	99	100	99	100	99	100	98,4
S=20	97	97	99	100	99	100	99	99	100	100	99
S=25	98	98	98	100	98	100	97	99	100	100	98,8

(d) K=40 groupes

FIGURE 3.30 – Taux de reconnaissance pour les gestes de contrôle de la base CMKinect-10 pour chaque valeur de K et S.



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Moyenne
MMC	97	97	99	100	99	100	99	99	100	100	99
MMC modifié	98	99	100	100	100	100	100	100	100	100	99,7

FIGURE 3.31 – Comparaison entre les taux de reconnaissance des gestes de la base CMKinect-10 obtenu avec MMC basique et MMC modifié pour $(K = 40, S = 20)$.

3.5 Bilan

Dans ce chapitre, nous avons développé un système de reconnaissance de gestes dynamiques basé sur les modèles de Markov cachés discrets. Nous avons représenté les mouvements de la personne avec un descripteur robuste inspiré des facteurs de la méthode de Laban. Notre première application consiste à contrôler le robot avec les gestes humains, donc nous avons décrit les gestes avec les trois composantes de LMA « Corps, Espace et Forme ». Notre application est indépendante de l'état de la personne et donc du rythme de mouvement. Pour cela la composante d'Effort qui décrit l'expressivité du geste a été ignorée. Nous avons proposé une méthode de classification basée sur l'algorithme de Forward qui permet de modéliser chaque geste dans deux sens (direct et indirect). Cela contribue

à la différenciation entre les mouvements similaires et donc l'amélioration de la performance de notre système de reconnaissance. Finalement, notre descripteur de geste ainsi que notre méthode de classification ont été évalués sur 4 bases d'actions, 3 bases publiques (MSRC-12, MSR Action 3D, UTKinect) et notre base (CMKinect-10). Dans le chapitre suivant, nous allons améliorer notre application pour rendre l'interface Homme-Robot plus naturelle. La dimension d'expressivité sera intégrée grâce à la composante Effort de LMA afin de reconnaitre les actions ainsi que les émotions de la personne. Des méthodes de classification globales comme RDF, SVM, etc seront appliquées afin de caractériser l'entièreté du geste et éviter quelques problèmes de classification qui peut être produits avec une méthode locale. En effet, si la séquence du geste est courte et que, par conséquent, seules quelques données sont disponibles, l'algorithme d'apprentissage Espérance-Maximisation (EM) peut renvoyer des estimations peu fiables. De plus, pour analyser l'expressivité et le rythme d'un geste il faut étudier le geste entier.

Chapitre 4

Reconnaissance des gestes expressifs

Sommaire

4.1 Construction de ECMXsens-5 94

4.1.1 Description de la base ECMXsens-5 94

4.2 Descripteur expressif 98

4.2.1 Relation Effort-Forme 98

4.2.2 Descripteur global 99

4.2.3 Composante Effort 101

4.3 Évaluation du descripteur 107

4.3.1 MSRC12 108

4.3.2 MSR Action 3D 114

4.3.3 UTKinect 115

4.3.4 ECMXsens-5 116

4.4 Bilan 118

Dans ce chapitre, nous développons un système de reconnaissance des gestes expressifs. Nous intégrons ici le terme d'expressivité pour améliorer notre application robotique «Interaction Homme-Robot» et la rendre plus naturelle. Notre système devra être capable de reconnaitre le geste de la personne et aussi son état émotionnel à travers son mouvement. Nous construisons une base de données avec le capteur de mouvement MVN Awinda de Xsens, composée de 5 gestes interprétés avec 4 émotions différentes (joie, colère, tristesse et neutre). Nous nous basons sur le descripteur présenté dans le chapitre 3 et nous ajoutons la composante d'Effort pour décrire l'expressivité du geste. Considérant le geste entier, nous quantifions les facteurs de LMA avec des mesures globales afin de représenter le mouvement entier et réaliser la reconnaissance globale du geste. Pour les phases d'entrainement et de classification, nous utilisons la bibliothèque d'apprentissage automa-

tique «Scikit-learn». Nous choisissons quatre méthodes d'apprentissage parmi les plus réputées (les forêts d'arbres décisionnels, le perceptron multicouches, les machines à vecteurs de support : Un-Contre-Un et Un-Contre-Tous). Un ajustement des différents paramètres des modèles est réalisé afin d'avoir une meilleure performance du système. Une comparaison entre les différents algorithmes de classification est faite afin de choisir le meilleur. Nous évaluons notre approche avec les mêmes bases publiques utilisées dans le chapitre 3, ainsi que notre base de gestes expressifs.
Ce chapitre est organisé de la manière suivante : la section 4.1 présente la mise en place de la base de données des gestes expressifs. Dans la section 4.2, nous exposons notre descripteur de mouvement global inspiré du descripteur local présenté dans le troisième chapitre. La section 4.3 concerne la partie de classification des gestes expressifs avec une étude comparative entre les 4 méthodes d'apprentissage afin de sélectionner la meilleure pour les prochaines études. Nous concluons dans la section 4.4.

4.1 Construction de la base de données des gestes expressifs

4.1.1 Description de la base ECMXsens-5

Notre base de données ECMXsens-5 (5 Expressive Control Motions) est composée de cinq mouvements expressifs (danser, avancer, faire un signe, pointer et s'arrêter), illustrés dans la Figure 4.1. Nous avons capturé des mouvements neutres, ainsi que des mouvements effectués avec des émotions : joie, colère et tristesse. Chaque geste expressif est répété 5 fois.

Les participants

11 personnes (cinq hommes et six femmes) de l'Université d'Evry Val d'Essonne, âgées de 27 à 36 ans (moyenne= 29,85 ans, écart-type= 2,47) ont participé à la construction de notre base. Les sessions de capture de mouvement ont été enregistrées avec la permission des participants. Notre corpus est bien anonymisé, collecté avec le consentement des différents participants. Tous les acteurs ont reçu une compensation monétaire pour leur participation, et ont été informé que les données collectées allaient être anonymisées, partagées et analysées strictement dans des fins de recherche. Afin d'éviter toute exagération de l'expression des émotions, nous avons choisi des gens qui ne sont pas acteurs professionnels, ce qui permet d'avoir une application adaptable à tout le monde.

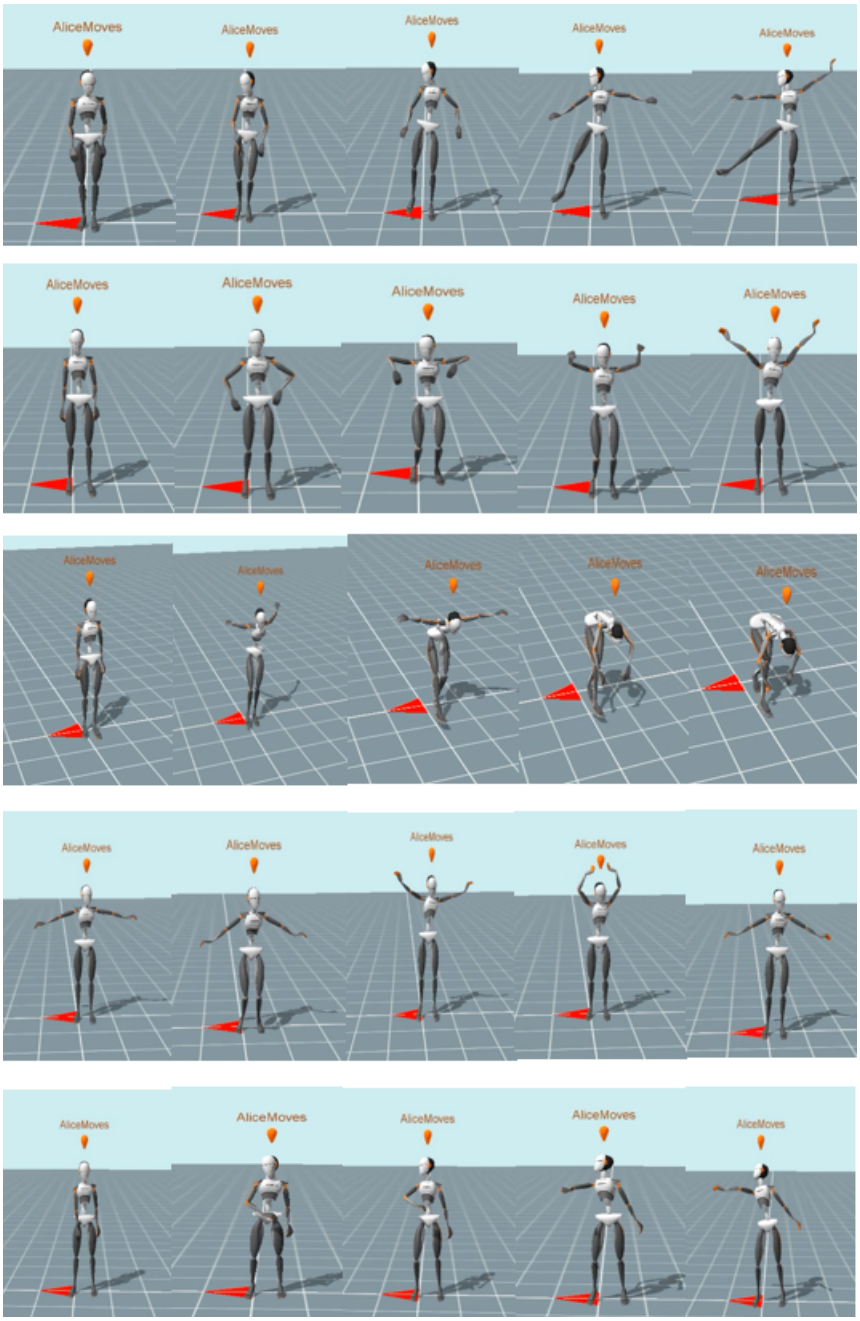


FIGURE 4.1 – La base de données ECMXsens-5, les gestes de haut vers le bas sont : danser, avancer, s'arrêter, faire un signe et pointer.

Les scénarios

Pour aider les participants à bien exprimer leurs émotions, nous proposons des scénarios présentant des situations émotionnelles et nous avons diffusé de la musique. Chaque participant a été invité à lire le scénario proposé, prendre le temps pour s'imprégner de la situation et, dès qu'il se sent prêt, effectue le geste demandé 5 fois. Des exemples de scénarios proposés pour les émotions de

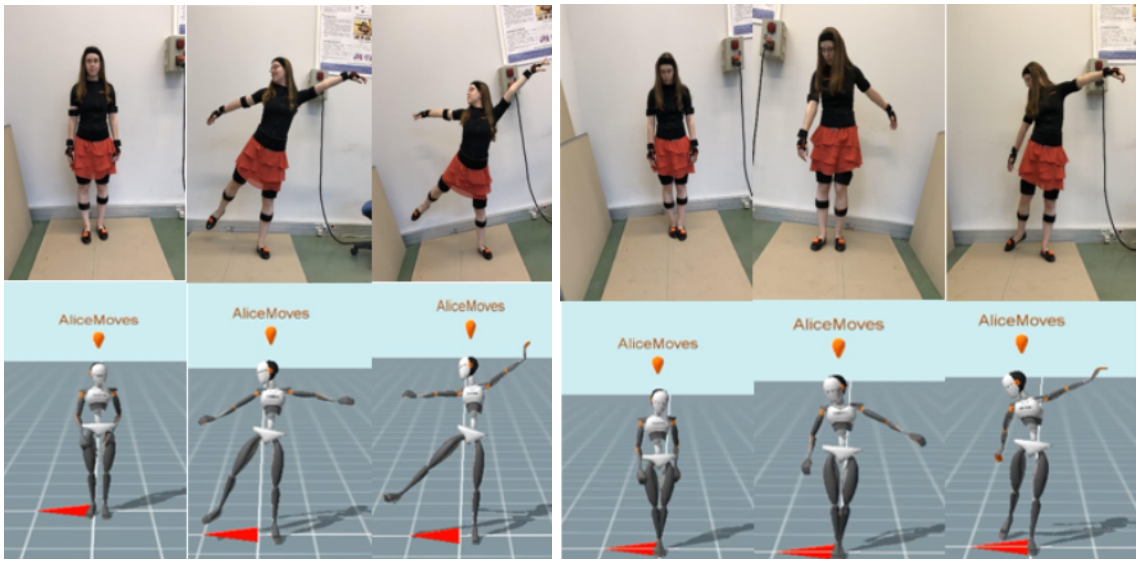
la joie, de la tristesse et de la colère sont présentés ci-dessous :

- Emotion de la joie
  - Vous avez réussi tous vos cours à l'université et le jour même où vous avez obtenu vos résultats, vous recevez un appel téléphonique d'une entreprise pour une embauche avec un très bon salaire.
  - Vous êtes dans une bonne relation. L'anniversaire de votre partenaire est proche et vous prévoyez faire une fête, mais vous n'avez pas assez d'argent, tout à coup vous recevez un email qui vous informe que vous avez gagné dans une loterie.
- Emotion de la tristesse
  - Pour un festival populaire, vous avez réservé votre billet et votre hôtel depuis 6 mois mais à cause d'un empêchement vous ratez votre vol et donc vous ne pouvez pas arriver à votre Festival préféré.
  - Vous avez perdu quelqu'un de très proche dans un accident.
- Emotion de la colère
  - Vous avez un entretien d'embauche que vous attendiez depuis longtemps. Aujourd'hui est le passage à l'heure d'été et vous oubliez de modifier votre horloge. Donc, vous vous réveillez tard, vous vous habillez rapidement et montez dans votre voiture que vous avez trouvée en panne, donc vous n'avez pas d'autre choix que les transports publiques. Enfin, dans la gare de métro, vous trouvez que les chauffeurs sont en grève.
  - Vous avez votre propre restaurant qui perd plus d'argent qu'il n'en gagne. Vous décidez de vérifier la cause, alors vous vous présentez en tant que nouvel employé auprès de vos employés. La première semaine, vous remarquez que le restaurant est ouvert à une heure tardive et, par conséquent, il y a toujours un long délai pour les commandes des clients. Vous découvrez un service et un comportement désagréable des serveurs envers les clients. En fin de compte, vous êtes très en colère contre l'inefficacité de vos employés.

  Pendant les sessions d'enregistrement, l'ordre des scénarios, des émotions et des mouvements a été randomisé d'un participant à l'autre.

Matériels

Pour l'acquisition des gestes expressifs, nous avons utilisé un produit de "motion capture" sans caméra fourni par le groupe Xsens, qui est le système MVN Awinda [Roetenberg et al., 2013]. Il s'agit d'une combinaison avec des capteurs attachés par des straps pour le suivi de 17 articulations (tête, sternum, bassin, épaules gauche/droite, avant-bras gauche/droit, arrière-bras gauche/droit, mains



(a) La joie (b) La tristesse

FIGURE 4.2 – Le geste de danse effectué avec deux émotions différentes.

gauche/droite, l'avant jambes gauche/droite, l'arrière jambes gauche/droite, pieds gauche/droit). Ce système est composé de 17 centrales inertielles utilisées pour capturer les mouvements de 23 segments du corps (Voir Figure 4.3). Chaque centrale inertielle est composée d'un accéléromètre, un gyroscope, un magnétomètre et un baromètre. Des informations de positions et d'orientations issues des capteurs en temps réel, à 60 trames par secondes sont enregistrées. De plus, ce système combine deux algorithmes, le filtre de kalman XKF3-hm et l'algorithme SDI (Strap-Down Integration), qui permettent de fournir des informations précises sur les positions et les orientations des articulations. Ce système possède deux avantages majeurs par rapport au capteur Kinect. Il nécessite pas de condition d'éclairage particulière pour son fonctionnement. En plus, avec ce système il n'y a pas des limitations qui peuvent être obtenues par occlusion avec des objets entourant ou des personnes interagissant avec l'acteur. De plus en mettant la combinaison sans caméra, les acteurs ne sont pas contraints à un volume de mesure spécifique et leurs mouvements peuvent être mesurés dans un environnement familier tout en effectuant leur tâches, comme dans la vie quotidienne.

Format des fichiers de capture de mouvement

Pour chaque participant, nous avons collecté des fichiers **MVNX** (Moven Open XML format). C'est un format XML qui peut être importé vers d'autres logiciels, y compris MATLAB et Excel. Ce format contient plusieurs informations, y compris les données de capteur, la cinématique des segments et les angles des articulations, ainsi que les informations sur le sujet nécessaires pour recréer une visualisation 3D d'un avatar.

FIGURE 4.3 – Le capteur MVN Awinda de Xsens.

4.2 Descripteur global expressif inspiré de LMA

4.2.1 Relation Effort-Forme

Laban a élaboré un modèle intéressant pour décrire et analyser la dynamique et les styles de mouvement, nommé le modèle Effort-Forme [Groff, 1995]. Ce modèle reflète l'état interne de la personne lors de l'exécution du mouvement en reliant un ensemble de propriétés physiques du mouvement avec des qualités expressives, comme le poids, le temps, etc [Bartenieff et al., 1984]. Il a été adopté dans plusieurs contextes, notamment pour l'analyse des mouvements expressifs dans la danse [Bartenieff et al., 1984, Aristidou et al., 2015a, Aristidou et al., 2014a, Guest, 2013, Hachimura et al., 2005, Aristidou et al., 2017a, Camurri et al., 2004a, Camurri et al., 2000], dans la musique [Camurri et al., 2004a, Broughton and Stevens, 2009, Camurri et al., 2000], dans la médecine [Foroud and Whishaw, 2006], etc. De plus, ce modèle a suscité un intérêt dans la modélisation computationnelle des agents virtuels pour une animation plus réaliste et expressive [Chi et al., 2000, Kapadia et al., 2013]. L'Effort concerne les rythmes dynamiques observables d'effort physique et la phraséologie du mouvement corporel [Maletic, 1987]. L'Effort reflète l'attitude intérieure à l'égard de l'utilisation de l'énergie suivant ses quatre facteurs : espace, temps, poids et flux. Ces 4 facteurs, une fois arrangés d'une manière spécifique, créent les huit actions d'Effort de base (Voir Table 4.1). Chaque facteur étant un continuum entre deux éléments d'Effort, soit l'espace (indirect/direct), le temps (soudain/ soutenu), le poids (léger/fort), soit le Flux (libre/lié). La relation Effort/Forme permet de focaliser l'attention sur deux aspects de mouvements corporels : d'une

part, comment l'énergie cinétique est dépensée dans l'espace, la force et le temps dans le comportement fonctionnel et expressif. D'autre part, la forme du mouvement, ou comment le corps change et se déplace dans l'espace.

TABLE 4.1 – Les huit actions élémentaires d'Effort.

	Espace	**Temps**	**Poids**	**Flux**
Coup de poing	Direct	Soudain	Fort	Lié
Presser	Direct	Soutenu	Fort	Lié
Couper	Indirect	Soudain	Fort	Lié
Tordre	Indirect	Soutenu	Fort	Lié
Toucher	Direct	Soudain	Léger	Libre
Glisser	Direct	Soutenu	Léger	Libre
Feuilleter	Indirect	Soudain	Léger	Libre
Flotter	Indirect	Soutenu	Léger	Libre

4.2.2 Descripteur global

Afin de décrire la totalité du geste, nous utilisons des mesures globales en se basant sur les composantes de LMA présentées dans le chapitre 3. Nous quantifions la composante Effort pour capturer l'aspect qualitatif et expressif du geste. Nous commençons par la composante du Corps, qui est définie par les angles entre les différentes articulations à chaque instant t. La représentation globale associée à cette qualité est composée de 3 mesures : la moyenne (M_c), l'écart-type (E_c) et la plage (P_c).

$$M_{c^k} = \frac{1}{T}\sum_{t=1}^{T} c_t^k \tag{4.1}$$

$$E_{c^k} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(c_t^k - M_{c^k})^2} \tag{4.2}$$

$$P_{c^k} = \underset{1\leq t\leq T}{max}\, c_t^k - \underset{1\leq t\leq T}{min}\, c_t^k \tag{4.3}$$

c_t^k correspond à la caractéristique locale définie dans la composante du Corps dans le chapitre précédent. $k \in \{1, \ldots, l\}$ calculée à l'instant t et l est le nombre des caractéristiques (13 caractéristiques). Pour la composante Espace, nous calculons la longueur (L^k) des trajectoires générées par les trois articulations suivantes, $k \in$ (main droite, main gauche et tête). Connaître la longueur des trajectoires des articulations est important car cela nous donne plus d'informations sur le type du mouvement.

$$L^k = \sum_{t=1}^{T-1} ||P_{t+1}^k - P_t^k|| \tag{4.4}$$

P_t^k correspond à la position en 3D de l'articulation k capturée à l'instant t. Nous présentons la composante Forme avec 9 caractéristiques, 3 pour le flux de forme, 3 pour la mise en forme et 3 pour le facteur du mouvement directionnel. Dans le chapitre 3, nous avons quantifié le flux de forme avec le volume de l'enveloppe convexe du squelette ($datav_t$) à chaque instant t. Nous mesurons donc la moyenne (M_{datav}), l'écart type (E_{datav}) et la plage (P_{datav}) de cette caractéristique.

$$M_{datav} = \frac{1}{T}\sum_{t=1}^{T} datav_t \tag{4.5}$$

$$E_{datav} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(datav_t - M_{datav})^2} \tag{4.6}$$

$$P_{datav} = \max_{1\leq t\leq T} datav_t - \min_{1\leq t\leq T} datav_t \tag{4.7}$$

Pour la mise en forme, nous proposons une quantification globale des caractéristiques locales présentées dans le chapitre 3. Nous décrivons l'expansion du corps suivant les trois plans en calculant le mouvement moyen de toutes les articulations du squelette (N articulations) par rapport à l'articulation du torse (P_1^s) capturée à l'instant initial ($t = 1$).

— Sur le plan horizontal, nous calculons le mouvement moyen d'ouverture/fermeture, relatif à la position de l'articulation du torse à l'instant initial.

$$D_H = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N}\sqrt{([P_t^k]_X - [P_1^s]_X)^2} \tag{4.8}$$

— Sur le plan frontal, nous calculons le mouvement moyen d'ascendant/descendant, relatif à la position de l'articulation du torse à l'instant initial.

$$D_F = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N}\sqrt{([P_t^k]_Y - [P_1^s]_Y)^2} \tag{4.9}$$

— Sur le plan sagittal, nous calculons le mouvement moyen d'avancement/recul, relatif à la position de l'articulation du torse à l'instant initial.

$$D_S = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N}\sqrt{([P_t^k]_Z - [P_1^s]_Z)^2} \tag{4.10}$$

T présente la longueur de la séquence du geste et $[P_t^k]_{X,Y,Z}$ correspond à la position de l'articulation k à l'instant t. La Figure 4.4 illustre la variation des distances à chaque trame entre les différentes articulations par rapport à l'articulation du torse capturée à l'instant initial dans le geste «remonter

la musique» de la base MSRC-12.

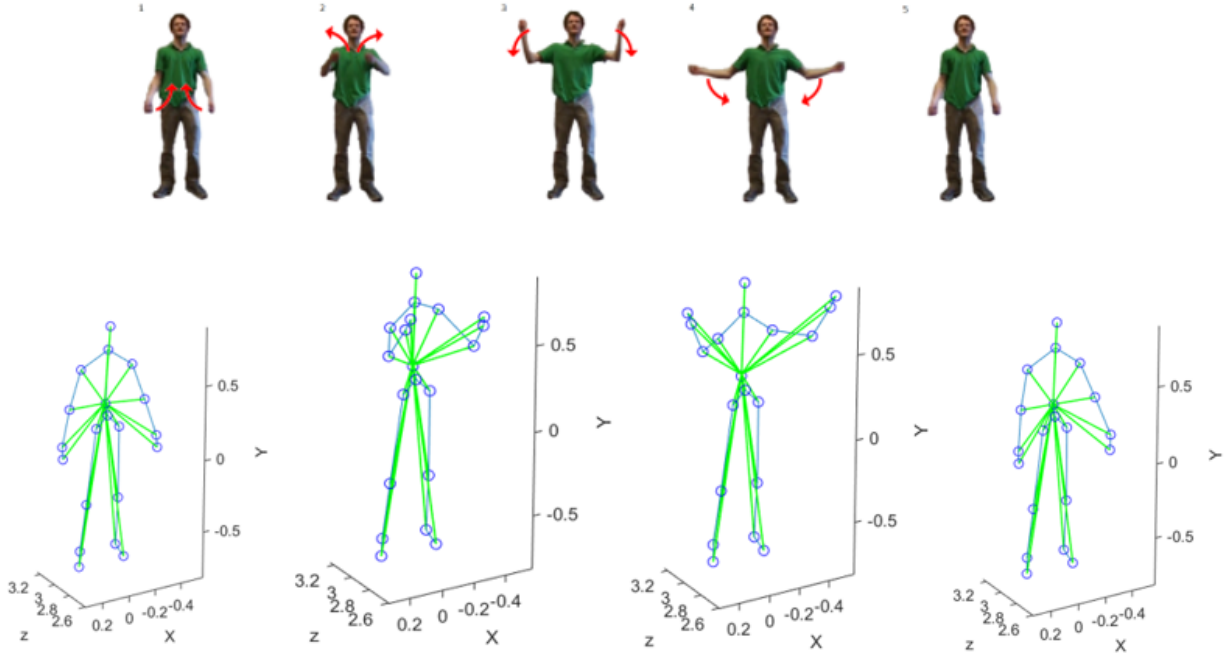


FIGURE 4.4 – Variation des distances entre les positions des articulations P_t^k à chaque trame et la position du torse J_1^s à l'instant initial (t=1) .

Le facteur du mouvement directionnel est décrit par la caractéristique de la courbure des trajectoires des extrémités supérieures du squelette (les mains et la tête). Donc à partir de l'angle de la courbure locale ϕ définie dans le descripteur local dans le chapitre 3 nous dérivons la caractéristique C qui présente la courbure globale du mouvement avec la formule suivante :

$$\phi_{P_t^k} = arcos(\frac{P_{t-1}^k \vec{P_t^k}}{||P_{t-1}^k \vec{P_t^k}||} . \frac{P_t^k \vec{P_{t+1}^k}}{||P_t^k \vec{P_{t+1}^k}||}) \tag{4.11}$$

$$C^k = \sum_{t=2}^{T-1} \phi_{P_t^k} \tag{4.12}$$

$\phi_{P_t^k}$ est la courbure locale de l'articulation k à l'instant t. Dans ce cas si nous avons des trajectoires linéaires, l'indice de courbure (C) sera proche de 0. Par contre, si les mouvements sont courbés, la valeur de C augmente.

4.2.3 Composante Effort

L'Effort est la texture, la couleur, les émotions et l'attitude intérieure qui est exprimée par le mouvement. Il est souvent décrit comme la dynamique du mouvement, l'utilisation qualitative de l'énergie. Par exemple, si on regarde les deux actions "pousser un objet lourd" et "fermer une porte",

les deux sont très similaires en termes d'organisation du corps. En effet, les deux actions s'appuient sur l'extension du bras. Par contre, les attentions portées à la force du mouvement, au contrôle du mouvement et à la durée du mouvement sont très différentes. Laban a proposé que la dynamique du mouvement humain soit résumé par une combinaison des facteurs suivants, dont chacun a deux polarités opposées (Figure 4.5) :

— Espace (indirect/direct) : le "où" du mouvement ; attention/pensée.
— Temps (soutenu/soudain) : le "quand" du mouvement ; intuition.
— Poids (léger/fort) : le "quoi" du mouvement ; sensation.
— Flux (libre/lié) : le "comment" du mouvement ; sentiment.

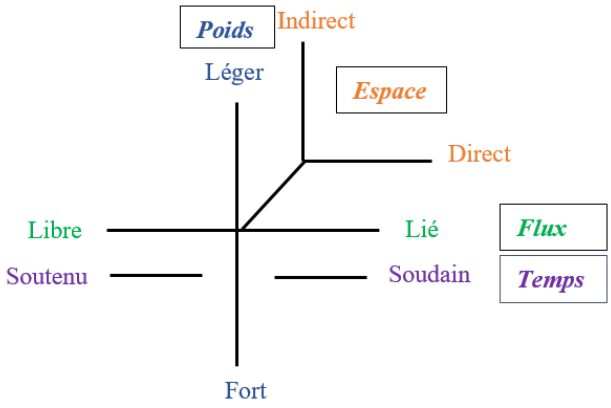


FIGURE 4.5 – Les facteurs de la composante Effort (Espace, Temps, Poids et Flux).

Dans notre cas, pour caractériser les gestes expressifs, Nous avons supposé que la partie supérieure du corps est la partie la plus expressive et qui bouge le plus dans une application de contrôle. Donc, nous nous sommes concentrés davantage sur la partie supérieure plus précisément les 4 articulations suivantes : la tête, la main droite, la main gauche et le torse.

Espace

L'espace exprime la qualité de l'attention que la personne porte sur les environs, et différencie un mouvement **Direct** (lorsque l'action est directe, l'attention est concentrée sur un seul point dans l'espace, ciblée et spécifique) d'un mouvement **Indirect** (en accordant une attention aux directions multiples dans l'espace, multi focalisé et flexible). [Masuda and Kato, 2010, Masuda et al., 2010] ont quantifié ce facteur avec le calcul de la direction de la tête. De même [Aristidou et al., 2017a] ont considéré que le mouvement est direct si le squelette se déplace dans la même direction que l'orientation de la tête, sinon il est classé comme indirect. Pour cela ils ont caractérisé le facteur Espace par la mesure de l'orientation de la tête, en calculant l'angle entre l'orientation de la tête

et la trajectoire du corps de l'artiste qui est définie par la trajectoire de l'articulation du centre de la hanche. Nous caractérisons cette qualité avec l'indice (S^k) de la rectitude des trajectoires des articulations (k) du haut du corps. Cet indice est exprimé par le rapport entre la distance euclidienne des deux positions de la première et la dernière trame (D^k) et la somme des distances entre des trames successives (L^k), calculé avec l'équation suivante :

$$S^k = \frac{D^k}{L^k} = \frac{||P_T^k - P_1^k||}{\sum_{t=1}^{T-1} ||P_{t+1}^k - P_t^k||} \tag{4.13}$$

Dans un mouvement direct (rectiligne), nous obtenons un indice de rectitude proche de 1 et dans le cas d'un mouvement indirect, la valeur de S^k sera proche de 0.

Temps

La catégorie temporelle décrit le rythme du mouvement relativement à son urgence et donc distingue un mouvement **Soudain** (rapide, urgent, inattendu, surprenant) d'un mouvement **Soutenu** (stable, continu). Pour [Samadani et al., 2013] le facteur du temps est déterminé par les accélérations accumulées au cours du temps au niveau des parties du corps. De leur part [Masuda and Kato, 2010, Masuda et al., 2010] ont caractérisé le facteur du temps par la vitesse angulaire de toutes les articulations. [Truong et al., 2016] ont utilisé 8 caractéristiques pour quantifier le temps, la durée totale du geste, le pourcentage des périodes de pause relativement à la séquence entière du geste. De plus, ils ont pris les deux séries qui correspondent respectivement aux durées des pauses et aux durées des périodes d'activité et ont calculé pour chacune les trois paramètres suivantes :la moyenne, l'écart-type et la valeur maximale. [Kapadia et al., 2013] ont supposé qu'un mouvement soudain est caractérisé par des pics d'accélération et un mouvement soutenu par une vitesse uniforme (pas d'accélération). Donc pour quantifier le facteur du temps, ils ont mesuré l'accélération nette accumulée sur la durée d'un intervalle de mouvement. [Aristidou et al., 2017a] ont introduit 5 caractéristiques qui correspondent aux vitesses et accélérations des articulations mains, pieds et pelvis. Pour notre cas, comme nous l'avons mentionné au début de la section, nous nous intéressons sur la partie supérieure du corps. Nous décrivons la façon dont le mouvement évolue dans le temps en mesurant sa vitesse. Trois mesures sont utilisées pour caractériser le facteur du Temps : la moyenne (M_v), l'écart-type (E_v) et

la plage (P_v) de la vitesse.

$$v_t^k = \frac{P_t^k - P_{t-1}^k}{t - (t-1)} \tag{4.14}$$

$$M_{v^k} = \frac{1}{T}\sum_{t=1}^{T} v_t^k \tag{4.15}$$

$$E_{v^k} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(v_t^k - M_{v^k})^2} \tag{4.16}$$

$$P_{v^k} = \underset{1\leq t\leq T}{max}\, v_t^k - \underset{1\leq t\leq T}{min}\, v_t^k \tag{4.17}$$

P_t^k et v_t^k représentent respectivement la position en 3D et la vitesse de l'articulation k à l'instant t. Nous prenons le cas des gestes (Avancer et Pointer) de notre base ECMXsens-5, réalisés par la même personne avec deux états différents respectivement (en colère/neutre et heureux/neutre). La Figure 4.6 présente la variation de la vitesse de la main gauche à chaque instant lors de l'exécution du premier geste "Avancer" avec les deux états (neutre et colère). Nous remarquons que le rythme du mouvement avec l'état neutre (courbe bleue) est presque stable, soutenu. Par contre, la courbe orange qui présente le même geste effectué avec l'état de la colère marque un pic qui dépasse une vitesse de $4m/s^2$. De même, dans la Figure 4.7 nous présentons la variation de la vitesse de la main gauche lors de l'exécution du geste de pointage avec l'état neutre (courbe bleue) et l'état de la joie (courbe orange). Nous voyons bien qu'il y a une différence remarquable entre les deux vitesses pour le même geste ce qui caractérise des rythmes différents des mouvements dans chaque état. Cela confirme l'importance de la caractéristique de la vitesse à identifier la différence entre les rythmes des mouvements.

Poids

Le poids qualifie la force ou la puissance avec laquelle un mouvement est effectué et caractérise un mouvement **Fort** d'un mouvement **Léger**. [Samadani et al., 2013] ont caractérisé ce facteur avec la somme maximale des énergies cinétiques des parties mobiles du corps. Plus la valeur maximale est élevée plus le poids est classifié comme Fort. [Masuda and Kato, 2010] ont estimé le facteur du poids avec la mesure de l'accélération angulaire des articulations. [Truong et al., 2016] ont utilisé 6 caractéristiques pour ce facteur qui sont les vitesses et les accélérations verticales des deux mains et du centre des hanches. [Kapadia et al., 2013] et [Aristidou et al., 2017a] ont utilisé la même caractéristique pour la quantification du poids qui est la décélération du mouvement. Tous les deux ont considéré qu'un mouvement fort est caractérisé par une grande décélération du mouvement, tandis

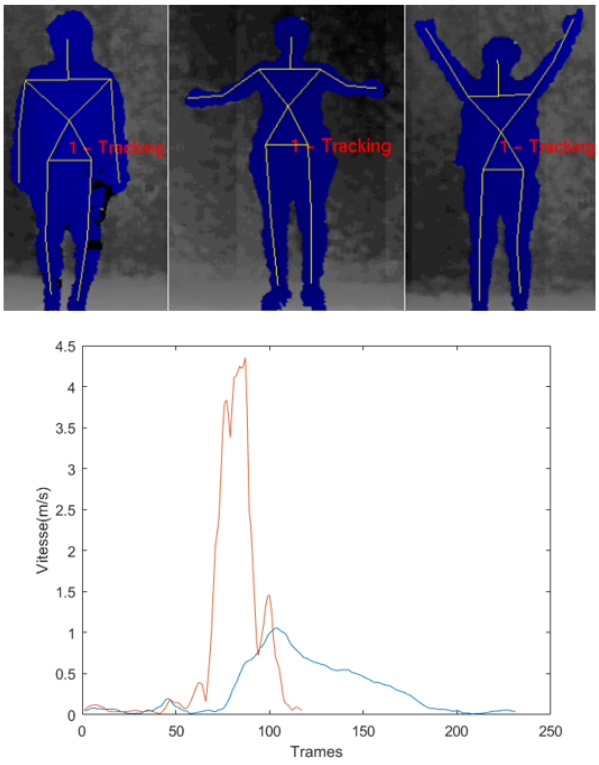


FIGURE 4.6 – Variation de la vitesse (v_l) de la main gauche dans le geste "Avancer" avec l'état neutre (courbe bleue) et en colère (courbe orange).

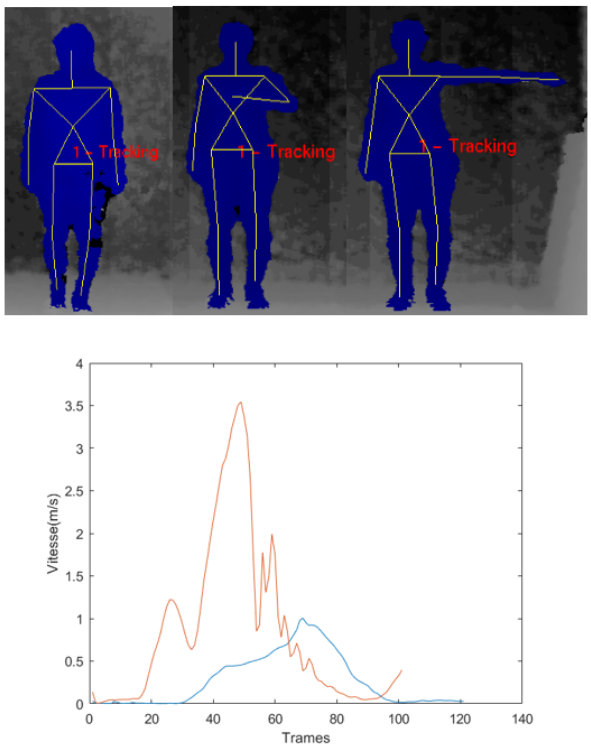


FIGURE 4.7 – Variation de la vitesse (v_r) de la main droite dans le geste "Pointer" avec l'état Neutre (courbe bleue) et de la Joie (courbe orange).

qu'un mouvement fluide est représentatif d'une décélération faible ou nulle. Pour notre cas, nous considérons qu'un mouvement fort (par exemple pousser un objet très lourd) présente une haute résistance et des frictions élevées qui se refléteraient dans des accélérations croissantes. Donc pour quantifier le facteur de poids, nous retenons les mesures des accélérations des articulations de la partie supérieure du corps (main gauche, main droite, tête et torse). Ainsi, nous calculons la moyenne (M_a), l'écart-type (E_a) et la plage (P_a) pour chaque mesure.

$$a_t^k = \frac{v_t^k - v_{t-1}^k}{t - (t-1)} \tag{4.18}$$

$$M_{a^k} = \frac{1}{T}\sum_{t=1}^{T} a_t^k \tag{4.19}$$

$$E_{a^k} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(a_t^k - M_{a^k})^2} \tag{4.20}$$

$$P_{a^k} = \underset{1\leq t\leq T}{max}\, a_t^k - \underset{1\leq t\leq T}{min}\, a_t^k \tag{4.21}$$

a_t^k correspond à l'accélération de l'articulation k à l'instant t. Donc, si nous prenons le cas des émotions de la colère et de la tristesse, la force exercée dans la réalisation du mouvement avec un état de colère va être plus importante que celle avec l'état de tristesse et donc une accélération plus élevée dans l'émotion de la colère. Nous reprenons le même exemple du geste "Avancer" avec les états triste et en colère, et nous illustrons les variations de l'accélération de la main gauche à chaque instant. La Figure 4.8, montre une accélération faible et presque nulle de la courbe bleue qui correspond à l'émotion de la tristesse et une accélération plus importante avec un pic de $60m/s^2$ correspondant à l'émotion de la colère (courbe orange). Cela confirme la pertinence de notre caractéristique pour distinguer des mouvements léger (accélération faible) d'un mouvement fort (accélération élevée).

Flux

L'axe de flux traite essentiellement le degré perçu de contrôle en mouvement. C'est un facteur de continuité, de progression, et permet de caractériser un mouvement lié ou contrôlé d'un mouvement libre ou imparable. Selon [Chi et al., 2000], les mouvements liés expriment le fait que le mouvement pourrait s'arrêter rapidement si les conditions changent, alors que les mouvements libres présentent une continuité qui faciliterait la transition. Pour ce facteur, la plupart des auteurs [Samadani et al., 2013, Masuda and Kato, 2010, Truong et al., 2016, Kapadia et al., 2013, Aristidou et al., 2017a] ont adopté la même caractéristique qui est l'à-coups, et ont considéré qu'un mouvement lié présente des à-coups très forts, tandis qu'un mouvement libre présente des à-coups

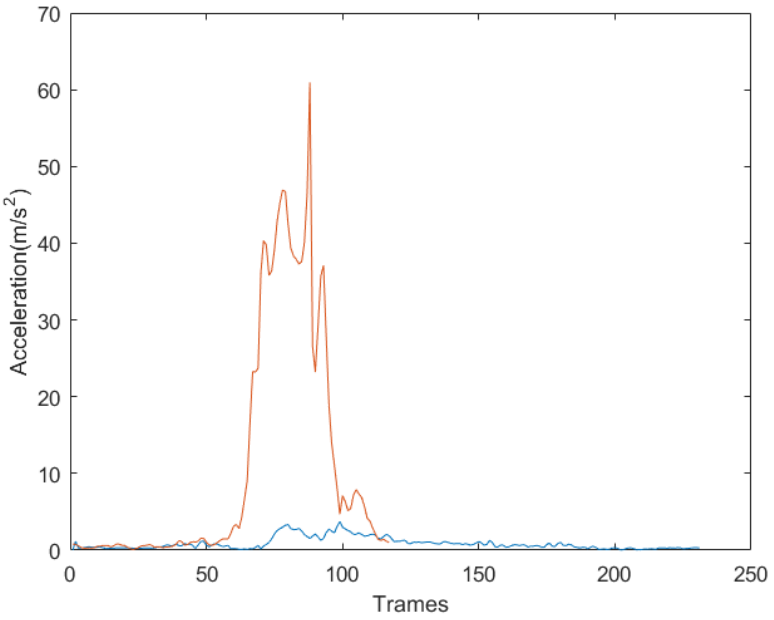


FIGURE 4.8 – Variation de l'accélération (a_l) de la main gauche dans le geste "Avancer" avec l'état triste (courbe bleue) et en colère (courbe orange).

faibles. Dans notre cas, nous caractérisons l'étendue global du mouvement en calculant les plages des mouvements des rotations suivantes : le tangage qui caractérise le degré d'inclinaison dans l'axe transversal, vers l'avant et vers l'arrière et le lacet qui présente le mouvement de rotation autour de l'axe vertical. Donc, nous calculons la plage de lacet P^k_{lacet} et de tangage $P^k_{tangage}$ des mouvements des articulations suivantes, $k \in$(tête, main droite et main gauche). Dan un mouvement libre, nous obtiendrons une plage plus grande que dans un mouvement lié. Nous avons pris un exemple de geste "Avancer" effectué avec deux état différents (neutre et heureux), nous avons calculé la plage des deux angles tangage et lacet dans le mouvement de l'articulation de la tête. Nous avons trouvé une plage de tangage de 33.78° et de lacet de 33.29° dans l'état heureux et une plage de tangage de 17.85° et de lacet de 8.79° dans l'état neutre. Cela confirme que le mouvement le plus libre se caractérise par une plage de tangage et lacet plus élevée.

4.3 Évaluation comparative du descripteur entre les quatre méthodes d'apprentissage

Dans cette section nous évaluons la robustesse de notre descripteur dans la reconnaissance des actions et aussi des gestes expressifs. Nous réalisons une classification globale des mouvements à travers 4 méthodes d'apprentissage développées dans la bibliothèque Scikit-learn [Pedregosa et al., 2011] :

les forêts d'arbres décisionnels (RDF), le perceptron multicouches (MLP), les deux approches de la méthode des machines à vecteurs de support multiclasses (Un-Contre-Un (OAO) et Un-Contre-Tous (OAA)). Nous détaillons en annexe le principe de chaque méthode. Nous testons notre système sur les trois bases publiques (MSRC-12, MSR Action 3D et UTKinect) et sur notre base de gestes expressifs ECMXsens-5. Pour la validation de notre méthode nous utilisons la technique de «validation croisée de k groupes» pour $k = 2$, 5 et 10. Donc nous divisons l'ensemble des données en k groupes, puis nous sélectionnons un des k groupes comme ensemble de validation et les $(k-1)$ groupes restants constitueront l'ensemble d'entrainement. Nous calculons le score de performance, puis nous répétons l'opération en sélectionnant un autre groupe de validation parmi les $(k-1)$ groupes qui n'ont pas encore été utilisés pour la validation du modèle. La procédure se répète ainsi k fois et la moyenne des k scores est enfin calculée pour estimer le score final de la reconnaissance. Comme mesure de la performance de notre méthode nous utilisons le F-score, qui combine les mesures de précision et de rappel avec la formule suivante :

$$F - score = 2 * \frac{precision.rappel}{precision + rappel} \tag{4.22}$$

$$\text{Précision} = \frac{nombre\ de\ vrais\ positifs}{nombre\ de\ vrais\ positifs\ +\ nombre\ de\ faux\ positifs}$$

$$\text{Rappel} = \frac{nombre\ de\ vrais\ positifs}{nombre\ de\ vrais\ positifs\ +\ nombre\ de\ faux\ negatifs}$$

4.3.1 MSRC12

Comme dans le chapitre 3, nous avons évalué notre système de reconnaissance de gestes avec le descripteur global sur les trois catégories de la base MSRC-12 : iconique, métaphorique et tous les gestes avec les quatre méthode d'apprentissages (RDF, MLP, SVM (OAO) et SVM (OAA)).

Méthode des forêts d'arbres décisionnels (RDF)

Le RDF est un algorithme de classification qui consiste en un ensemble d'arbres de décision indépendants (Voir Annexe D). Dans l'étape d'entrainement, chaque arbre est construit en utilisant des échantillons bootstrap. Environ 1/3 des données ne sont pas utilisées pour la construction de la forêt et peuvent être utilisées pour les tests. Ce sont les échantillons OOB (Out Of Bag) qui sont utilisés pour estimer l'erreur de prédiction du modèle. Au cours de la construction des arbres, les nœuds sont subdivisés progressivement en nœuds fils et l'arbre est ensuite développé jusqu'à ce qu'une profondeur maximale soit atteinte. Il existe différentes fonctions pour mesurer la qualité

d'une division, comme le gain d'information [Quinlan, 1992] ou l'impureté de Gini [Breiman, 1984]. Pour la méthode de validation croisée de k groupes, nous fixons d'abord k à 5 groupes. Donc, à chaque fois, nous considérons 4 groupes pour l'entrainement et 1 groupe pour le test. Nous répétons la procédure 5 fois et nous calculons la moyenne des résultats obtenus pour chaque test. La librairie Scikit-learn propose différents paramètres d'optimisation pour la méthode RDF. La première étape consiste donc à ajuster ces paramètres afin d'avoir une méthode de classification performante. Nous avons d'abord utilisé l'impureté de Gini comme critère de division et nous avons varié le nombre des arbres T de 10 à 200. La profondeur maximale de l'arbre est fixé avec sa valeur par défaut, c'est à dire jusqu'à ce que toutes les feuilles deviennent pures ou qu'elles contiennent un nombre d'échantillons inférieur au nombre minimum d'échantillons requis pour diviser un nœud interne. Les meilleurs résultats ont été obtenus pour $T = 100$ avec un F-score moyen de 0.94. Nous avons aussi testé notre modèle avec le gain d'information et nous avons eu le même résultat obtenu avec le critère de Gini. Les autres paramètres sont ajustés avec leurs valeurs par défaut. Les différentes valeurs optimales des paramètres de RDF sont résumées dans la Table 4.2.

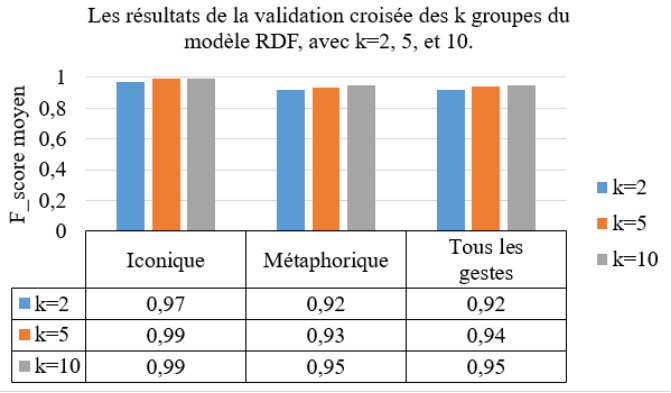
Les paramètres de RDF sont :

- T : le nombre des arbres.
- $critere$: le critère de répartition, gini ou entropie.
- c_{max} : le nombre des caractéristiques à considérer lors de la recherche de la meilleure répartition, $c_{max} < d$ (d est le nombre des caractéristiques total) :
  - Auto : $\sqrt{d}$.
  - Sqrt : $\sqrt{d}$.
  - Log2 : $\log 2(d)$.
  - None : d.
- p_{max} : la profondeur maximale de l'arbre.
- e_{min} : le nombre minimal d'échantillons requis pour diviser un nœud interne, par défaut c'est 2.
- ef_{min} : le nombre minimum d'échantillons requis pour être sur un nœud feuille, par défaut c'est 1.

Nous répétons le même test pour $K = 2$ et $K = 10$. La Figure 4.9 résume les résultats obtenus pour les différentes valeurs de K. Le meilleur résultat est obtenu pour $K = 10$ groupes, où le F-score moyen pour les gestes iconiques est de 0.99, pour les gestes métaphoriques est de 0.95 et pour l'ensemble des gestes est de 0.95.

TABLE 4.2 – Ajustement des paramètres de RDF.

Paramètres	**Valeurs**
T	100
$critere$	Gini
c_{max}	$\sqrt{85}$
p_{max}	None
e_{min}	2
ef_{min}	1



	Iconique	Métaphorique	Tous les gestes
k=2	0,97	0,92	0,92
k=5	0,99	0,93	0,94
k=10	0,99	0,95	0,95

FIGURE 4.9 – Les résultats de F-scores moyens de la méthode RDF dans la base MSRC-12.

Les machines à vecteurs de support (SVM)

Les SVMs ont été conçus à l'origine pour les classifications binaires. Cette méthode recherche simplement l'hyperplan séparateur avec la plus grande marge qui construit un hyperplan qui sépare le mieux possible les classes. Nous avons détaillé le principe de cette méthode en Annexe B. Les implémentations souvent suggérées pour la classification SVM multi-classes sont la méthode Un-Contre-Tous (OAA) et Un-Contre-Un (OAO).

— L'approche Un-Contre-Tous : entraine n SVM binaires où n représente le nombre des classes. Chaque SVM biclasse est entrainé en utilisant les élément d'une classe contre toutes les autres.

— L'approche Un-contre-Un : consiste à utiliser un classifieur pour chaque paire de classes. Elle entraine donc $n(n-1)/2$ fonctions de décisions.

Ces méthodes sont définies par divers paramètres tels que les fonctions du noyau y compris la fonction linéaire, fonction de base radiale gaussienne, sigmoïde et polynomiale avec les paramètres des noyaux et le paramètre de régularisation C. Nous avons d'abord utilisé la méthode de validation croisée de 5 groupes. Nous avons commencé notre test avec la sélection du noyau linéaire puisque c'est le seul noyau sans paramètres. Une série de tests a été réalisée afin d'évaluer les performances du classificateur tout en variant la valeur de C. Ce paramètre ajuste le compromis entre la minimisation de l'erreur des données d'apprentissage et la maximisation de la marge. Pour une valeur élevée de

C, une grande pénalité est attribuée aux erreurs tandis qu'une valeur faible augmente la marge et permet ensuite d'ignorer les points proches de la limite. Nous avons varié la valeur de C de 0.01 jusqu'à 10. Le meilleur résultat de F-score est de 0.91 obtenu pour $C = 1.0$. Une fois que la valeur de C est sélectionnée, nous procédons à l'évaluation des autres types de noyau, la fonction de base radiale (RBF), la sigmoïde et la polynomiale. Contrairement au premier type de noyau utilisé (linéaire), dans les autre types la performance de la classification avec SVM dépend à la fois du paramètre de pénalité et d'autres paramètres liés au noyau y compris le coefficient du noyau (γ) et le degré (d) pour le noyau polynomial. Les tests ont été menés avec des valeurs de $\gamma \in \{1, \ldots, 10\}$, et de d allant de 1 à 5. Le meilleur résultat de reconnaissance a été obtenu avec la fonction du noyau polynomial pour $d = 2$, $\gamma = 4$ et $C = 1$ avec un F-score de 0.92. La Table 4.3 résume les valeurs des différents paramètres optimaux de la méthode SVM.

Les paramètres de SVM sont :

- *noyau* : le type de noyau : linéaire, polynomial, fonction de base radiale gaussienne (RBF), sigmoïde, par défaut c'est RBF.
- C : le paramètre de pénalité, par défaut c'est 1.0.
- d : le degré de la fonction polynomiale, par défaut c'est 3.
- γ : le coefficient des noyaux (polynomial, RBF et sigmoïde).

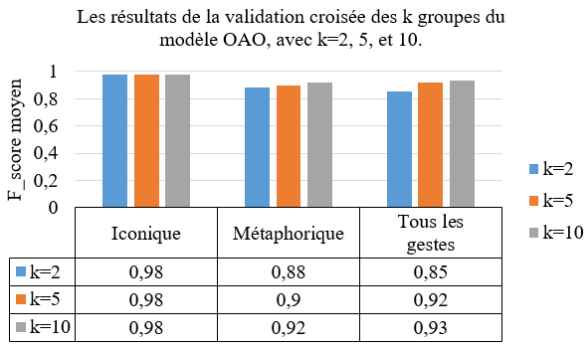
TABLE 4.3 – Ajustement des paramètres de SVM.

Paramètres	**Valeurs**
noyau	polynomial
d	2
C	1
γ	4

Nous évaluons notre système avec la méthode de validation croisée pour $K = 2$, 5 et 10. Le meilleur résultat est obtenu pour $K = 10$ dans les trois catégories (iconique, métaphorique et tous les gestes) comme indiqué dans la Figure 4.10.

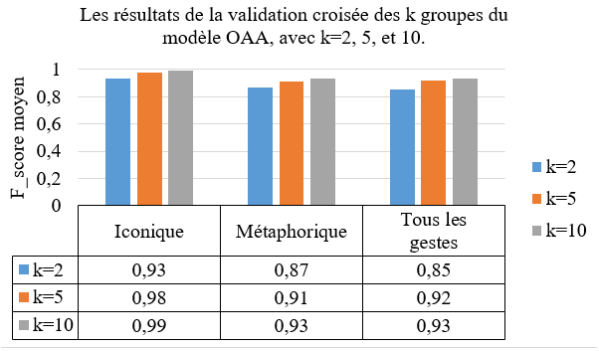
Le perceptron multicouches (MLP)

Le MLP est un réseau neuronal feed-forward, constitué d'un certain nombre de neurones organisés en plusieurs couches (une couche d'entrée et une couche de sortie avec une ou plusieurs couches cachées). La couche d'entrée reçoit un vecteur d'activation externe et passe à travers les connexions pondérées aux neurones dans la première couche cachée. Celles-ci calculent leurs activations et les transmettent aux neurones dans la couche suivante. Pour trouver le nombre optimal des couches cachées, nous avons varié leur nombre de 1 à 3. Pour chaque couche nous avons varié le nombre de



	Iconique	Métaphorique	Tous les gestes
k=2	0,98	0,88	0,85
k=5	0,98	0,9	0,92
k=10	0,98	0,92	0,93

(a) OAO



	Iconique	Métaphorique	Tous les gestes
k=2	0,93	0,87	0,85
k=5	0,98	0,91	0,92
k=10	0,99	0,93	0,93

(b) OAA

FIGURE 4.10 – Les résultats de F-scores moyens de la méthode SVM dans la base MSRC-12.

neurones (S) de 10 à 100 avec un pas de 10. Pour la fonction d'activation (A_f), nous avons comparé les trois fonctions suivantes : sigmoïde logistique, tangente hyperbolique et la fonction d'unité linéaire rectifiée. Pour l'étape d'entrainement, MLP utilise une technique d'apprentissage supervisée appelée rétropropagation (Voir Annexe C). Dans cette approche deux paramètres sont importants et doivent être ajustés, le taux d'apprentissage (ρ) qui contrôle la taille de pas dans la mise à jour des poids qui devrait être compris entre 0 et 1 et le facteur d'inertie (α) pour accélérer la convergence du réseau tout en évitant l'instabilité. Les meilleurs résultats ont été obtenus avec une seule couche contenant 60 neurones et la fonction d'activation sigmoïde logistique (Voir Table 4.4). Pour l'algorithme de rétropropagation, les paramètres optimaux ont été obtenus en réglant α à 0.9 et ρ à 0.001 ce qui donne une moyenne de F-scores de 0.92.

Les paramètres de MLP sont :

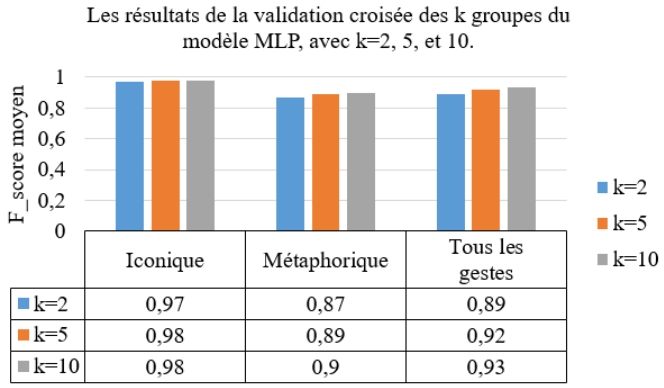
- S : le nombre de neurones dans une couche cachée.
- *noyau* : la fonction d'activation : identité, sigmoïde logistique, tangente hyperbolique, unité linéaire rectifiée.
- ρ : le taux d'apprentissage pour contrôler l'incrémentation dans la mise à jour des poids.
- α : le facteur d'inertie pour la mise à jour de la descente de gradient, compris entre 0 et 1.

TABLE 4.4 – Ajustement des paramètres de MLP.

Paramètres	**Valeurs**
A_f	sigmoïde logistique
S	60
ρ	0.001
α	0.9

De même, nous appliquons la méthode de validation pour $K = 2$, 5 et 10 (Voir la Figure 4.11) et le meilleur résultat est obtenu pour $K = 10$ groupes avec des F-scores moyens de 0.98 (catégorie

iconique), 0.9 (catégorie métaphorique) et 0.93 (tous les gestes). Finalement, nous comparons entre



	Iconique	Métaphorique	Tous les gestes
k=2	0,97	0,87	0,89
k=5	0,98	0,89	0,92
k=10	0,98	0,9	0,93

FIGURE 4.11 – Les résultats de F-scores moyens de la méthode MLP dans la base MSRC-12.

les quatre méthodes d'apprentissage (RDF, OAO, OAA et LMP), nous trouvons que la méthode RDF produit les meilleurs résultats. Elle a marqué des taux de reconnaissance plus élevé que les autres dans les trois catégories de la base MSRC-12 (iconique 0.99, métaphorique 0.95 et tous les gestes 0.95). La Figure 4.12 résume tous les résultats obtenus avec ces quatre méthodes. Les matrices

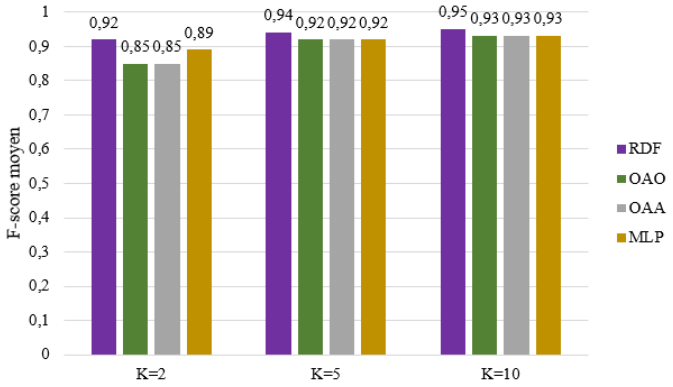
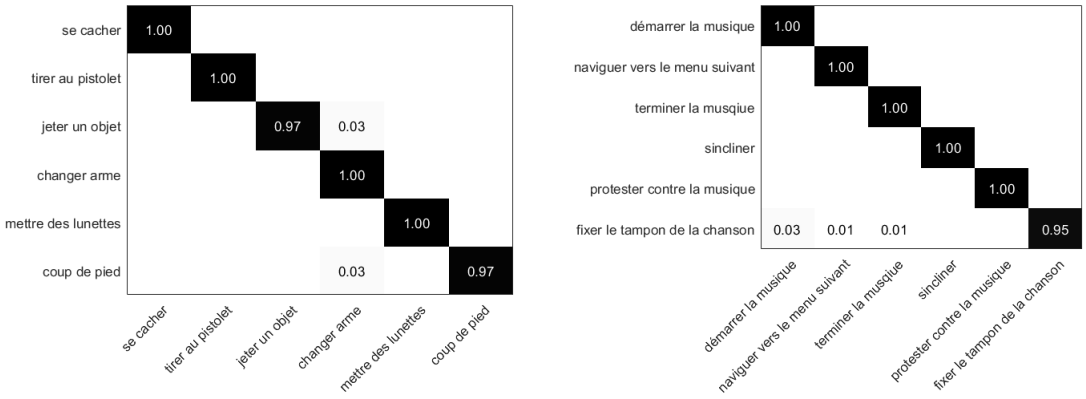


FIGURE 4.12 – Comparaison entre les 4 méthodes d'apprentissage dans la base MSRC-12.

de confusion de gestes iconiques et métaphoriques avec la méthode RDF sont présentées dans la Figure 5.11. Nous remarquons que notre système a réussi à faire la distinction entre les différents gestes dans les deux catégories. Quelques confusions sont apparues entre certains gestes, comme "jeter un objet" et "changer d'arme". Aussi, dans la classe du geste "coup de pied", il y a 3% des données qui sont confondues avec le geste "changer d'arme". Pour la catégorie iconique, dans la classe du geste "fixer le tempo de la chanson", 95% sont classés correctement, 2% sont confondus en moitié avec les deux classes "terminer la musique" et "naviguer vers le menu suivant", et 3% avec la classe de geste "démarrer la musique".



(a) Gestes iconiques (b) Gestes métaphoriques

FIGURE 4.13 – Matrices de confusions de la base MSRC-12.

4.3.2 MSR Action 3D

Nous avons évalué l'efficacité de notre descripteur global dans les cas d'une similarité interclasse élevée. Nous choisissons donc l'ensemble des données de la base MSR Action 3D, qui est composé de 557 instances d'action pré-segmentées de 20 actions. Comme dans le chapitre 3, nous avons divisé l'ensemble de données en trois sous-ensembles AS1, AS2 et AS3. Nous utilisons la méthode de validation de k groupes, pour $k = 2, 5$ et 10. La Figure 4.14 présente les résultats de la reconnaissance des gestes de la base MSR Action 3D en utilisant les méthodes, RDF, OAO, OAA et MLP. Nous

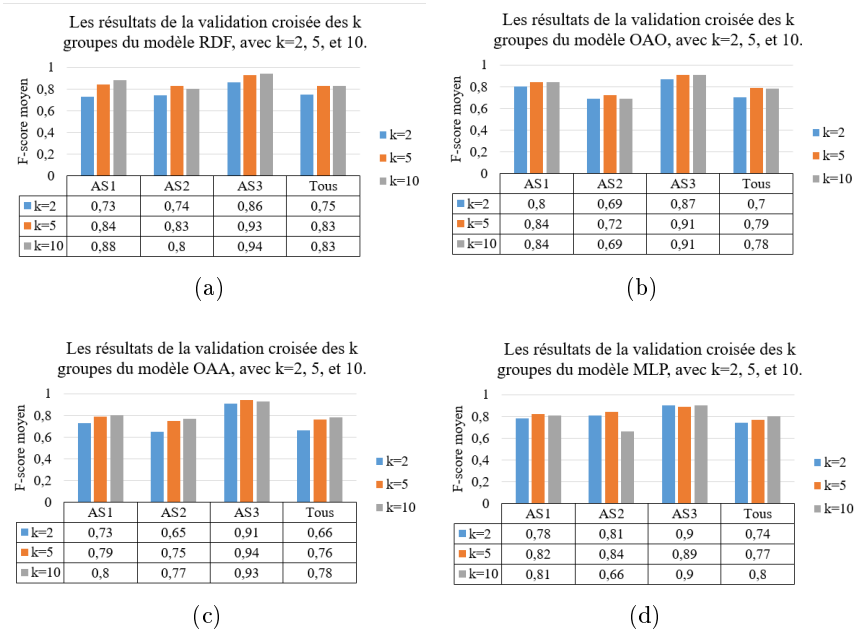
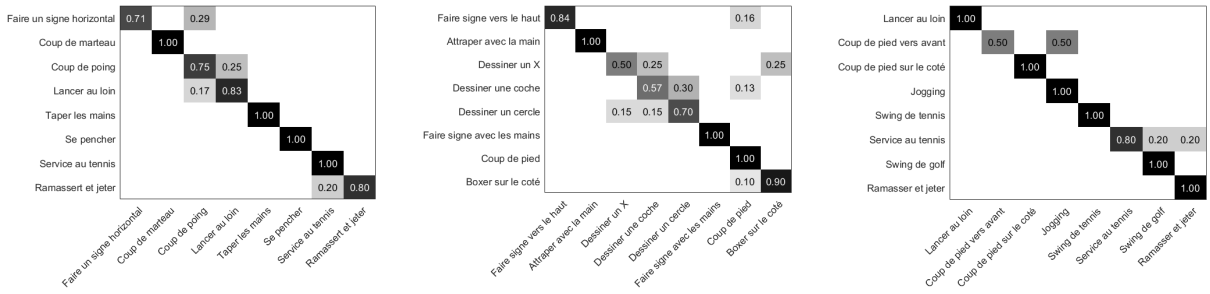


FIGURE 4.14 – Les résultats de F-scores moyens des méthodes (a) RDF, (b) OAO, (c) OAA et (d) MLP dans la base MSR Action 3D.

remarquons que AS3 marque toujours le meilleur résultat de reconnaissance, tandis que le groupe de AS2 a enregistré des résultats moins importants. Nous rappelons que le sous ensemble AS2 contient des gestes très similaires. Avec la méthode RDF, nous obtenons les meilleurs résultats des F-scores moyens en appliquant la méthode de validation croisée de 10 groupes avec les valeurs suivantes : 0.88 (pour AS1), 0.8 (pour AS2) et 0.94 (pour AS3). Las Figure 4.15 montre les matrices de confusion des sous ensembles (a) AS1, (b) AS2 et (c) AS3 avec la méthode RDF. AS1 et AS2 contiennent les actions similaires ce qui explique les différentes confusions trouvées entre certains gestes dans ces groupes. Dans le groupe AS1, des confusions sont présentées entre les deux actions "coup de poing" et "lancer au loin". 29% des données de la classe de l'action "faire un signe horizontal" sont indûment rangées dans la classe de l'action "coup de poing" et 20% des actions de la classe "ramasser et jeter" sont confondues avec la classe des actions "service au tennis". Dans le groupe AS2, nous remarquons des confusions surtout entre les trois actions suivantes "dessiner une coche", "dessiner un X" et "dessiner un cercle", ceci s'explique notamment par la forte similarité entre les mouvements dans ces trois actions. Notre méthode a correctement classifié la plupart des actions du groupe AS3, à part quelques exceptions comme l'action "coup de pied vers l'avant" qui a été confondue avec l'action "jogging" et l'action "service au tennis" avec les deux classes "swing de golf" et "ramasser et jeter". En comparant entre les 4 méthodes d'apprentissage (voir Figure 4.16), nous remarquons



(a) AS1 (b) AS2 (c) AS3

FIGURE 4.15 – Matrices de confusion de la base MSR 3D Action.

que dans les trois répartitions de groupes de validation la méthode de RDF dépasse toujours un peu les autres méthodes.

4.3.3 UTKinect

Notre descripteur a été évalué aussi sur la base UTKinect qui contient 10 activités effectuées par 10 personnes dans des vues différentes. Avec cette base nous évaluons la performance de notre système de la reconnaissance des actions et aussi l'étape de la normalisation des personnes. Pour

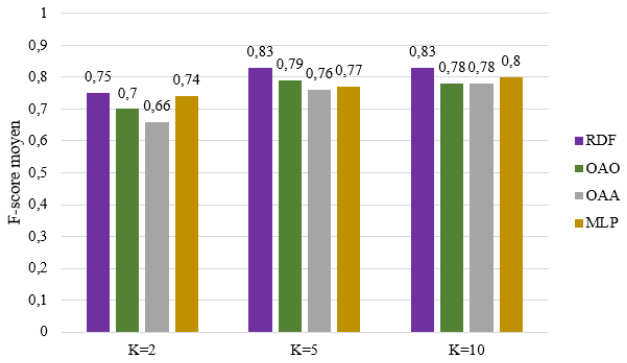


FIGURE 4.16 – Comparaison entre les 4 méthodes d'apprentissage dans la base MSR Action3D.

le deuxième objectif, nos appliquons notre programme de normalisation de squelette qui permet de rendre notre application invariante aux positions et aux orientations initiales des personnes. De même nous utilisons les 4 méthodes RDF, OAO, OAA et MLP pour l'entrainement et la classification des actions. La Figure 4.17 présente les différents résultats obtenus pour chaque méthode de classification en variant le paramètre k dans la méthode de validation croisée. Le meilleur F-score moyen de 0.92 est obtenu avec la méthode RDF pour une validation croisée de 10 groupes. La matrice de confusion

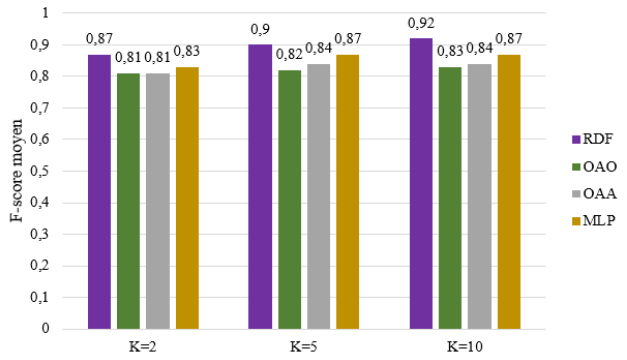


FIGURE 4.17 – Comparaison entre les 4 méthodes d'apprentissage dans la base UTkinect.

présentée dans la Figure 4.18 montre que notre méthode a réussi à distinguer entre les différentes actions de la base UTkinect, avec une confusion apparue entre les actions "pousser" et "tirer" et entre les actions "se lever" et "s'asseoir".

4.3.4 ECMXsens-5

Après avoir évalué notre descripteur sur des bases publiques, nous avons testé notre système sur notre base de gestes expressifs. De même, une évaluation comparative est faite entre les quatre méthodes d'apprentissage (RDF, OAO, OAA et MLP) avec la méthode de validation croisée de K groupes (k = 2, 5 et 10). Notre base ECMXsens-5 se compose de 5 gestes exprimés avec 4 émotions, donc nous avons 20 classes à reconnaitre : **Danser** Heureux (DH)/en Co-

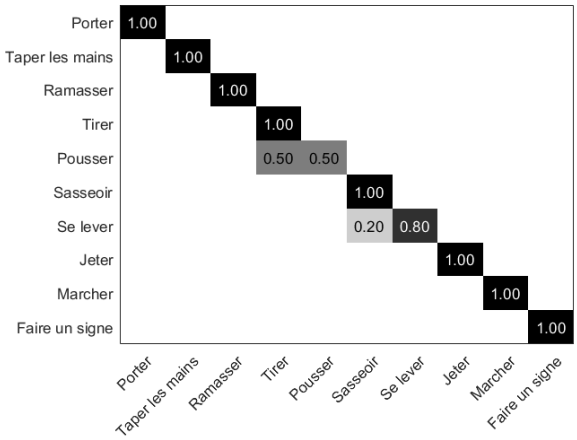


FIGURE 4.18 – Matrice de confusion de la base UTKinect avec la méthode RDF.

lère (DC)/Triste(DT)/Neutre (DN), **Avancer** Heureux (AH)/en Colère (AC)/Triste (AT)/Neutre (AN), **Faire un signe** Heureux (SH)/en Colère (SC)/Triste (ST)/Neutre (SN), **S'arrêter** Heureux (SaH)/en Colère (SaC)/Triste (SaT)/Neutre (SaN), **Pointer** Heureux (PH)/en Colère (PC)/Triste (PT)/Neutre (PN). Nous avons utilisé le capteur de Xsens pour le tracking de squelette. L'acquisition des données est cadencée à 60 trames par secondes, où les positions en 3D de chaque articulation sont enregistrées. La Figure 4.19 présente les différents résultats obtenus avec chaque méthode de classification utilisée. La méthode RDF marque encore un meilleur F-score moyen de 0.80 dans la classification de la base entière. La matrice de confusion de notre base est présentée dans la Fi-

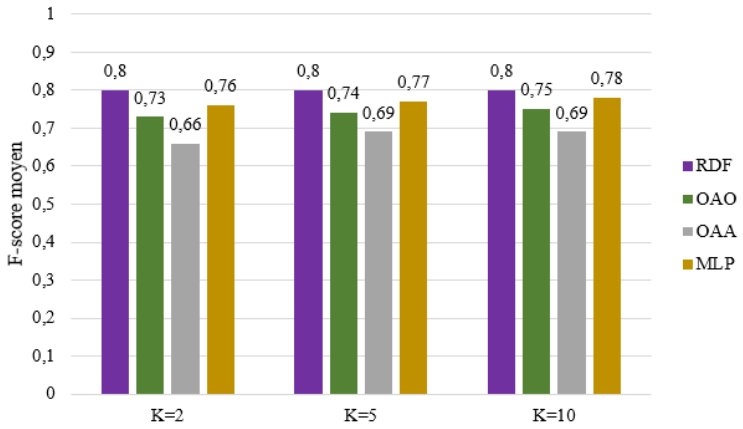


FIGURE 4.19 – Comparaison entre les 4 méthodes d'apprentissage dans la base ECMXsens-5.

gure 4.20. La diagonale qui présente les gestes reconnus correctement contient toujours les valeurs les plus élevées. Nous remarquons parfois des confusions dans les même gestes entre les émotions heureux et en colère ou entre les états triste et neutre. Après ces différents test, nous pouvons nous assurer de la robustesse de notre descripteur proposé basé sur les facteurs de LMA dans la caractérisation des actions et aussi des gestes expressifs. Les caractéristiques du couple Effort-Forme ont

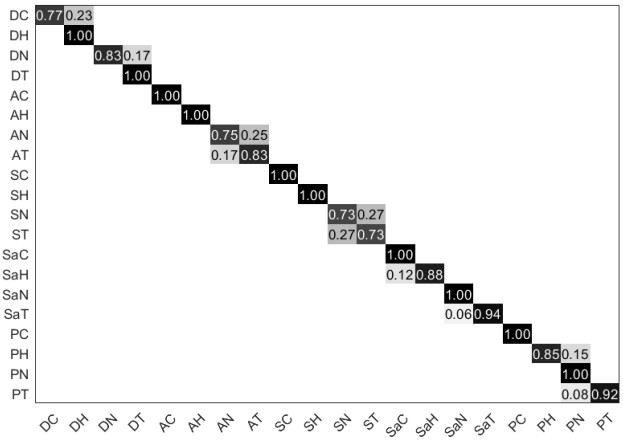


FIGURE 4.20 – La matrice de confusion de la base ECMXsens-5.

permis de distinguer l'aspect qualitatif du geste ainsi que le caractère expressif dans le mouvement. Les résultats obtenus dans la comparaison entre les quatre méthodes d'apprentissage concluent à la supériorité de la méthode des forêts d'arbres décisionnels sur les autres. Cette méthode fait partie de la famille des méthodes d'ensemble qui a connu beaucoup d'avantages dans le processus de classification.

4.4 Bilan

Dans ce chapitre nous avons construit une base de gestes expressifs composée de 5 gestes de contrôle réalisés avec 4 états différents (heureux, en colère, triste et neutre). Nous envisageons de transposer ce système à notre application robotique pour réaliser une interaction Homme-Robot la plus naturelle possible. Nous avons construit un descripteur global qui décrit l'entièreté du geste tout en caractérisant l'aspect quantitatif et qualitatif du mouvement. Les quatre composantes de LMA (Corps, Espace, Forme et Effort) sont quantifiées afin d'obtenir un descripteur de mouvement assez robuste et performant pour décrire les caractéristique physiques structurelles du mouvement, le rythme ainsi que l'expressivité dégagée derrière chaque geste. Nous avons aussi utilisé des méthodes réputées dans le domaine de l'apprentissage automatique (RDF, OAO, OAA et MLP). En utilisant la bibliothèque Scikit-learn nous ajustons les différents paramètres de chaque méthode afin d'affiner au mieux les modèles. En comparant les différents résultats obtenus par chaque méthode, nous concluons à la supériorité de la méthode RDF dans la classification des mouvements. Cela peut être expliqué par les différents avantages de cette méthode, le plus connu et important étant la résolution du problème de sur-apprentissage connu dans le domaine de l'apprentissage automatique. Dans le chapitre suivant, nous allons exploiter les caractéristiques de cette méthode pour classifier les

émotions dans chaque type de geste et aussi étudier l'importance des caractéristiques du mouvement du corps pour l'expression de chaque émotion dans notre base. Afin d'évaluer la performance de notre système de reconnaissance automatique, nous allons nous référer à une approche humaine qui consiste à analyser et reconnaitre les émotions en se basant sur la perception humaine.

Chapitre 5

Caractérisation des gestes expressifs et évaluation avec la perception humaine

Sommaire

5.1 Approche RDF 122

5.1.1 Reconnaissance des gestes expressifs avec la méthode RDF 122

5.1.2 Sélection de caractéristiques pertinentes avec la méthode RDF 132

5.2 Approche humaine 137

5.2.1 Reconnaissance des gestes expressifs avec la perception humaine 137

5.2.2 Sélection des caractéristiques avec l'approche humaine 142

5.3 Évaluation du système 145

5.4 Bilan 146

Dans ce chapitre, nous évaluons la performance de notre système de reconnaissance de gestes expressifs sur notre base de gestes expressifs ECMXsens-5 en le comparant avec la perception humaine. Le but de cette étude est d'utiliser l'évaluation humaine comme point de référence pour évaluer à la fois l'exactitude du classifieur et l'adéquation du modèle de mouvement proposé dans cette thèse. Deux tâches importantes sont réalisées : la classification des gestes expressifs et l'étude des caractéristiques discriminantes pour la description de chaque émotion avec la méthode RDF qui a donné les meilleurs résultats dans le chapitre précédent. A cet effet, nous élaborons un questionnaire d'évaluation basé sur des échelles de scores afin d'évaluer la perception humaine pour chaque émotion pour la classification et l'étude de l'importance des caractéristiques.

Ce chapitre est organisé de la manière suivante : la section 5.1 concerne la caractérisation des gestes expressifs avec la méthode d'apprentissage, RDF. Cette section est divisée en deux parties, la première concerne l'étape de classification, et la deuxième est l'étude de l'importance des caractéristiques

de notre descripteur de mouvement envers chaque émotion. La section 5.2 concerne la caractérisation des gestes expressifs avec l'approche humaine. De même, cette section est divisée en deux parties, la première considère l'homme en tant que classifieur des émotions et la deuxième en tant qu'évaluateur des caractéristiques de mouvement. Dans la section 5.3, nous résumons les résultats provenant des deux approches pour évaluer la performance de notre système de reconnaissance en se comparant à l'approche humaine. Nous concluons dans la section 5.4.

5.1 Caractérisation des gestes expressifs avec l'approche RDF

5.1.1 Reconnaissance des gestes expressifs avec la méthode RDF

Généralement, de faibles changements dans la base d'apprentissage peuvent entraîner de grands changements dans la construction du classifieur. Cela nécessite donc une combinaison de classifieurs complémentaires. La méthode RDF repose sur le principe de construire des ensembles de classifieurs divers par une règle de combinaison. Ce qui explique sa supériorité sur les autres méthodes dans les parties expérimentales précédentes. De nombreuses recherches sont faites pour étudier les avantages et les motivations dans l'utilisation de cette approche, nommée "ensemble de classifieurs".

Combinaison de classifieurs

Combiner plusieurs classifieurs revient à utiliser plusieurs classifieurs et les combiner afin d'obtenir un classifieur qui surpasse chacun d'entre eux. Ce type d'approche est intuitif puisqu'il imite notre façon de chercher plusieurs opinions avant de prendre une décision cruciale [Rokach, 2010]. Le domaine de recherche des "systèmes de classification multiple" (MCS) est devenu très populaire dans le domaine de l'apprentissage automatique. Plusieurs articles sur la construction d'ensembles de classifieurs ont été publiés [Dietterich, 2000, Fumera and Roli, 2005, TUMER and GHOSH, 1996] et ont montré l'efficacité de ce type de modèle à améliorer la capacité de généralisation d'un seul classifieur, et ainsi réduire la variance et améliorer la précision. L'amélioration des performances dans les systèmes de classification multiple repose sur le concept de la diversité, qui stipule qu'un bon ensemble est celui dans lequel les exemples mal classés sont différents d'un classifieur individuel à l'autre. Par conséquent, diverses stratégies sont utilisées pour obtenir un groupe de classifieurs basé sur la diversité. Cette diversité est réalisée de plusieurs manières, par exemple, sous-rééchantillonnage de données d'apprentissage, sélection de sous-ensembles d'objets, etc. Selon [Fumera and Roli, 2005], il y a trois motivations principales pour combiner les classifieurs :

— **Raison Statistique** : un algorithme d'apprentissage peut être vu comme une recherche dans

un espace H de classifieurs pour identifier le meilleur classifieur. Le problème statistique se pose lorsque l'ensemble de données d'entraînement est trop petit par rapport à la taille de l'espace des classifieurs. Dans ce cas, sans données suffisantes, l'algorithme d'apprentissage peut trouver de nombreux classifieurs différents dans H qui donnent tous la même précision sur les données d'entraînement. Cependant, ils peuvent avoir des performances de généralisation différentes. Cela peut donc augmenter le risque de sélectionner un mauvais classifieur, ie un classifieur avec une mauvaise capacité de généralisation. Donc, la solution est de construire un ensemble de tous ces classifieurs performants en combinant leurs sorties. Dietterich donne une illustration de cette situation dans la Figure 5.1. La courbe extérieure désigne l'espace des classifieurs H. La courbe interne désigne l'ensemble de classifieurs qui sont performants en apprentissage. Le point marqué f présente le vrai classifieur. La combinaison des classifieurs précis permet de donner une bonne approximation de f.

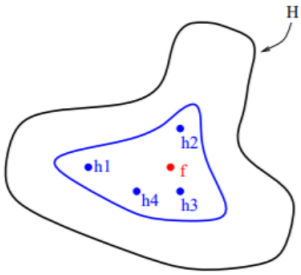


FIGURE 5.1 – Motivation Statistique.

— **Raison de Calcul** : plusieurs modèles d'apprentissage sont basés sur des techniques d'optimisation locales, ce qui rend le modèle sensible aux optima locaux. Dietterich a cité deux exemples, l'algorithme des réseaux de neurones qui utilise la descente de gradient afin de minimiser une fonction d'erreur sur les données d'entraînement et les algorithmes des arbres de décisions qui emploient une règle de fractionnement d'une manière récursive pour développer l'arbre de décision. Dans le cas où il y a suffisamment de données d'entrainement (pour que le problème statistique soit absent), il peut encore être très difficile sur le plan informatique pour l'algorithme d'apprentissage de trouver la meilleure hypothèse. Un ensemble construit en exécutant la recherche locale à partir de nombreux points de départ différents peut produire une meilleure approximation de la vraie fonction inconnue, comme le montre la Figure 5.2.

— **Raison Représentative :** il est possible que l'espace de classifieurs H considéré pour le problème ne contienne pas le classifieur optimal f. Cependant, la combinaison de plusieurs classifieurs peut étendre l'espace des classifieurs H. De cette manière, le vrai classifieur f peut être approximé en dehors de l'espace des classifieurs. La Figure 5.3 donne une illustration de

FIGURE 5.2 – Motivation de Calcul.

cette situation, où le classifieur optimal f est en dehors de l'espace de classifieurs H.

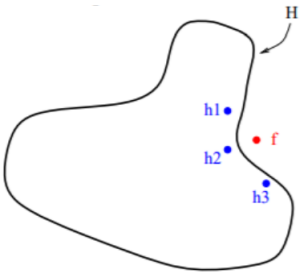


FIGURE 5.3 – Motivation Représentative.

Ces trois raisons montrent l'avantage de la combinaison de plusieurs classifieurs et les limites des classifieurs individuels. Il existe trois approches de combinaisons possibles : l'approche séquentielle, l'approche parallèle et l'approche hybride.

— **Approche séquentielle** : dans cette architecture les classifieurs sont entrainés séquentiellement, de sorte que chaque classifieur utilise les résultats dérivés du classifieur précédent (voir Figure 5.4). Par conséquent, à chaque étape, il n'y a qu'un seul classifieur en activité. Il y a deux approches pour cette combinaison séquentielle, l'approche de réduction de l'ensemble de classes et l'approche de réévaluation. Dans la première approche, le nombre de classes possibles est réduit d'une manière permanente, tandis que la seconde approche nécessite une réévaluation des modèles, qui sont rejetés dans l'étape précédente. La décision d'un classifieur dans une combinaison en série est rejetée si son niveau de confiance est inférieur à un seuil prédéfini. Une telle approche peut être considérée comme un filtrage progressif des décisions. Généralement, cela réduira le taux d'erreur global de la chaîne. Néanmoins, une combinaison de ce type est particulièrement sensible à l'ordre dans lequel les classifieurs sont placés. En effet, même s'ils n'ont pas besoin d'être les plus efficaces, les premiers classifieurs de la chaîne doivent être robustes.

— **Approche Parallèle** : dans cette architecture l'ensemble de classifieurs sont entrainés en parallèle indépendamment les uns des autres, et leurs résultats sont combinés ensuite pour

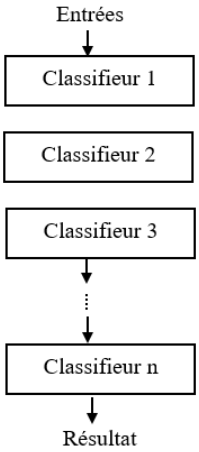


FIGURE 5.4 – Combinaison séquentielle.

donner la décision finale (Voir Figure 5.5). Dans les études de la recherche de la combinaison des classifieurs, l'approche parallèle a gagné beaucoup d'attention car elle a l'avantage d'être simple et facile à utiliser. Contrairement à l'approche séquentielle, l'organisation parallèle des classifieurs exige que les classifieurs individuels produisent simultanément leurs sorties. Toutes ces sorties sont alors fusionnées avec un opérateur de combinaison, comme un simple vote majoritaire, pour produire une décision finale. Dans cette approche l'ordre dans lequel les classifieurs sont placés n'a pas d'influence.

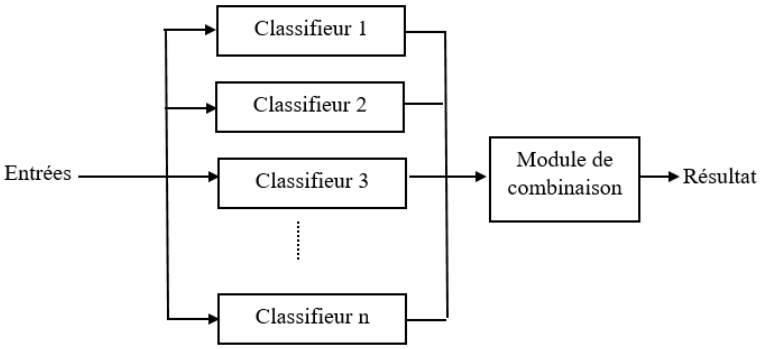


FIGURE 5.5 – Combinaison parallèle.

— **Approche Hybride** : l'idée de l'approche hybride consiste à combiner les deux approches ci-dessus afin de conserver les avantages des deux (Voir Figure 5.6).

De nombreux travaux de recherche encouragent l'adaptation de la combinaison des classifieurs pour améliorer la performance d'un modèle, ou réduire la probabilité de sélectionner un classifieur faible [TUMER and GHOSH, 1996, Dietterich, 2000]. Cependant, parmi ces trois approches, celle qui a suscité un grand intérêt chez la communauté scientifique est la combinaison parallèle. La motivation principale de cette combinaison est d'exploiter l'indépendance entre les classifieurs ce qui

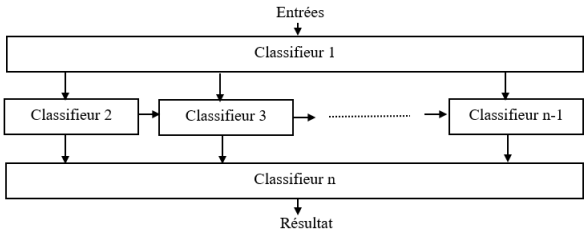


FIGURE 5.6 – Combinaison hybride.

mène à la réduction de l'erreur en faisant la moyenne. Dans la catégorie parallèle deux approches sont proposées : les méthodes qui utilisent des classifieurs hétérogènes, c'est-à-dire des classifieurs de types différents, conduisant à des ensembles hétérogènes. On peut citer le travail de [Kittler et al., 1998] qui ont combiné trois classifieurs hétérogènes, les réseaux de neurones, le classifieur bayésien et les modèles de Markov cachés pour une application de reconnaissance d'écriture manuscrite. Il existe également des méthodes qui utilisent des classifieurs homogènes, c'est-à-dire des classifieurs de même types, conduisant à des ensembles homogènes. Cela consiste donc à appliquer la même méthode d'apprentissage pour tous les classifieurs mais en produisant des différences dans leurs sorties. Cela revient par exemple à sélectionner des sous ensembles de données d'entrainement différents ou des sous espaces de caractéristiques différents pour ces données. On peut citer un exemple qui appartient à cette famille, c'est celui de la méthode des forêts d'arbres décisionnels [Breiman, 2001]. Cette méthode utilise un ensemble de classifieurs homogènes qui sont les arbres de décisions et applique l'algorithme de Bagging afin de créer des différences au niveau des prédictions de chaque arbre, ce qui rend ce modèle plus généralisé et donc plus performant.

Forêts d'arbres décisionnels

La méthode de forêts d'arbres décisionnels consiste en un ensemble d'arbres de décision construits parallèlement, où chaque arbre se développe aléatoirement et indépendamment des autres. Le caractère aléatoire est important lors de la construction d'un arbre. Ceci assure une variété dans la forêt ou, en d'autres termes, les arbres deviennent moins corrélés entre eux. L'algorithme RDF s'appuie sur l'utilisation de deux principes de randomisation : (i) Bagging et (ii) sélection de caractéristiques pour diviser chaque nœud d'un arbre.

Le principe de la méthode Bagging : Il s'agit d'une méthode introduite par Breiman (1996). Le mot bagging est la combinaison des mots Bootstrap et Aggregating. Le principe de Bagging consiste à construire plusieurs sous ensembles bootstrap à partir de l'ensemble d'apprentissage D_n (voir Figure 5.7). Chaque échantillon bootstrap D_i, $i = 1 \ldots, m$ est obtenu par un tirage aléatoire et avec remise de n observations dans D_n. A partir de chaque boostrap D_i un classifieur $\hat{g}(., D_i)$ est induit. Fi-

nalement, la collection de prédicteurs est alors agrégée en faisant simplement un vote majoritaire. Le modèle classifieur résultant réduit la variance des classifieurs individuels [Bauer and Kohavi, 1999].

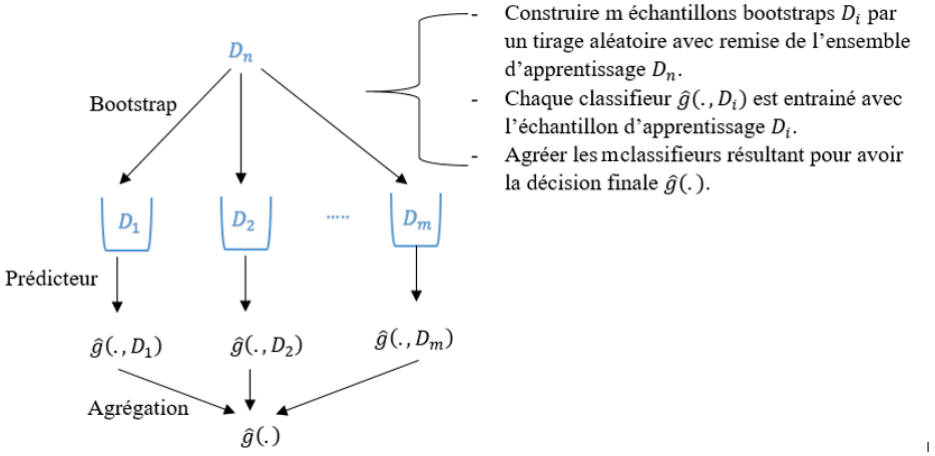


FIGURE 5.7 – Principe de Bagging.

Sélection aléatoire des caractéristiques : Pour la sélection d'un sous ensemble de variables, l'idée est de tirer à chaque nœud k caractéristiques de l'ensemble de p caractéristiques disponibles d'une manière aléatoire et sans remise, $k << p$. A chaque nœud on sélectionne la meilleure découpe sur la base de k variables choisies, sachant que k est le même pour tous les nœuds de tous les arbres de la forêt. Pour choisir la variable de séparation dans un nœud, les algorithmes testent les différentes variables d'entrée possibles et sélectionnent celle qui optimise le noeud par rapport à une mesure de pureté tel que le gain d'information [Quinlan, 1992] ou l'indice de Gini [Breiman, 1984]. La sélection aléatoire d'un nombre réduit de variables à chaque étape de construction d'un arbre, augmente significativement la variabilité en mettant en avant nécessairement d'autres variables. Chaque modèle de base est évidemment moins performant mais l'agrégation conduit finalement à un classifieur plus robuste et plus précis. Cela permet aussi de réduire davantage le temps du calcul. La Figure 5.8 illustre le principe de la sélection aléatoire des caractéristiques. Donc on peut résumer les étapes de l'algorithme des forêts d'arbres décisionnels comme suit :

1. Pour $i = 1, \ldots, m$
   1.1. Tirer aléatoirement un bootstrap de D_n
   1.2. Construire un arbre T_i pour chaque boostrap D_i.
      i. Tirer aléatoirement k caractéristiques parmi les p, avec $k << p$
      ii. Parmi les k caractéristiques, choisir celle qui donne la meilleure division du noeud.
      iii. Diviser le noeud en deux nœuds fils

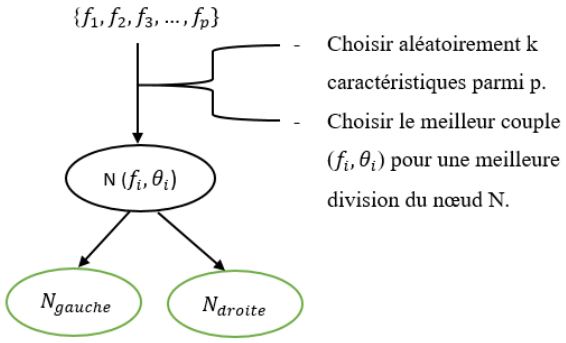


FIGURE 5.8 – Principe de la sélection de caractéristiques.

1.3. Calculer le classifieur sur cet échantillon : $\hat{g}(., D_i)$

2. Sortie $\hat{g}(x) = \underset{c}{argmax} \sum_{i=1}^{m} I_{\hat{g}(.,D_i)=c}$

Erreur OOB : Dans la méthode RDF, il n'y a pas besoin d'une validation croisée ou d'un ensemble de test séparé pour obtenir une estimation de l'erreur de prédiction. Il est estimé en interne comme suit : chaque arbre est construit à l'aide d'un échantillon bootstrap aléatoire différent de données d'origine. Chaque bootstrap laisse de coté un certain nombre d'observations ($\approx 1/3$), appelés observations OOB (Out Of Bag) qui sont utilisées pour estimer la performance du modèle en calculant l'erreur Out-of-bag (errOOB). Donc, pour un arbre donné T_b dans la forêt, seulement 2/3 des données d'entrainement sont utilisées dans sa construction et le 1/3 restant présentent les échantillons OOBs associés. Pour chaque observation (x_i, y_i), on sélectionne les échantillons bootstrap D_i ne contenant pas (x_i, y_i). Pour ces échantillons, on dit que l'observation (x_i, y_i) est un échantillon OOB. On prédit cette observation avec tous les arbres construits sur ces échantillons bootstrap. On agrège leurs prédictions par un vote majoritaire et on note $\hat{y}_i$ la prédiction OOB de x_i. Par la suite, on peut calculer le taux d'erreur moyen OOB de la forêt avec la formule suivante :

$$err_{OOB} = \frac{1}{n} \sum_{i=1}^{n} I_{y_i \neq \hat{y}_i} \tag{5.1}$$

L'avantage de l'erreur OOB est que l'ensemble original et entier des données peut être utilisé pour construire le classifieur RDF. Contrairement aux méthodes de validation croisée dans lesquelles un sous-ensemble des échantillons est utilisé pour la construction d'un RDF, la procédure OOB permet d'utiliser tous les échantillons pour la construction du classifieur. Cela donne des classifieurs RDF qui ont une précision plus importante que celle obtenue à partir de la validation croisée. Un autre avantage de l'utilisation de l'erreur OOB est le temps de calcul, en particulier lorsqu'il s'agit d'un nombre volumineux de données, où la construction d'un seul RDF peut durer plusieurs jours, voire plusieurs

semaines. Avec la procédure de OOB, un seul RDF doit être construit, contrairement à la validation croisée de k groupes où k RDFs doivent être construits [Bylander, 2002, Zhang et al., 2010]. Les échantillons OOBs serviront donc à l'évaluation interne d'une forêt et aussi à l'estimation de l'importance des variables.

Importance des variables : Suivant le principe de RDF, l'importance d'une variable f est la différence entre la précision de prédiction (c.-à-d. le nombre d'observations correctement classées) avant et après la permutation de cette variable dans les échantillons OOBs, moyennée sur tous les arbres. Une diminution élevée dans la précision de la prédiction dénote l'importance de cette caractéristique. Le calcul de l'erreur de prédiction des échantillons OOBs permet d'estimer l'importance des variables de la manière suivante : l'erreur que chaque arbre commet sur son échantillon OOB associé est calculé. Dans tous les échantillons OOBs, les valeurs de la variable f sont aléatoirement permutées. L'erreur que chaque arbre commet sur son échantillon OOB permuté est recalculé et par la suite comparé avec l'erreur OOB d'origine (avant permutation). Si le taux d'erreur de OOB après la permutation est plus grand que celui avec les observations OOB d'origine, la variable f est considérée importante. Nous résumons les étapes du calcul de l'importance d'une variable f par l'approche RDF comme suit :

1. Pour chaque arbre $t = 1, \ldots, T$ dans la forêt, calculer err_{OOBt}, le taux d'erreur moyen sur toutes les observations OOB dans l'arbre t.

$$err_{OOBt} = \frac{1}{Card(OOBt)} \sum_{x_i \in OOBt} I_{y_i \neq y_{it}} \tag{5.2}$$

   $OOBt$ contient les observations qui n'apparaissent pas dans l'échantillon bootstrap utilisé pour construire l'arbre t, $Card(OOBt)$ dénote sa cardinalité. y_i et $\hat{y}_{it}$ présentent respectivement, le vrai label et la prédiction de la ième observation par l'arbre t .

2. Permuter aléatoirement les valeurs de la caractéristique f dans l'échantillon $OOBt$. Ceci donne un échantillon perturbé, noté $OOBt^f$. Calculer enfin $errOOBt^f$, le taux d'erreur moyen sur $OOBt^f$.

$$errOOBt^f = \frac{1}{Card(OOBt^f)} \sum_{x_i \in OOBt^f} I_{y_i \neq \hat{y}_{it}} \tag{5.3}$$

3. L'importance d'une caractéristique f par un arbre t est calculée comme suit :

$$I^t(f) = errOOBt^f - errOOB^t \tag{5.4}$$

   Notez que $I^t(f) = 0$, si la caractéristique f n'est pas dans l'arbre t. Le score d'une caractéris-

tique f est alors calculé par la moyenne des importances sur tous les arbres.

$$I(f) = \frac{1}{T}\sum_{t=1}^{T} I^t(f) \tag{5.5}$$

où T est le nombre des arbres.

Ainsi, plus les permutations aléatoires de la caractéristique f conduisent à une forte augmentation de l'erreur, plus la caractéristique est considérée importante.

Résultats expérimentaux de la classification des gestes expressifs avec la méthode RDF

Dans la première partie expérimentale de cette étude, nous évaluons la performance de notre descripteur global sur la caractérisation des différents gestes expressifs de notre base ECMXsens-5. Donc, nous considérons l'ensemble entier des données, 1100 séquences (11 participants ×5 gestes×4 émotions×5 répétitions). La première étape consiste à ajuster les paramètres de la méthode RDF. Suivant le même procédé qu'au chapitre 4, nous ajustons les deux paramètres les plus importants dans la méthode de RDF, qui sont : le nombre des arbres T et le nombre des caractéristiques sélectionnées pour chaque division c_{max}. Nous utilisons deux méthodes de validation différentes : la validation croisée de 3 groupes et l'estimation du taux d'erreur moyen OOB. Nous gardons la même mesure de performance, le F-score.

— Avec la méthode de validation croisée : pour le paramètre T, nous varions sa valeur de 10 jusqu'à 300 arbres. Pour le paramètre c_{max}, nous testons 3 valeurs : 85, $\sqrt{85}$ et $log_2(85)$. Le meilleur F-score moyen est de 0.83, obtenu pour $c_{max} = log_2(85)$ et $T = 150$.

— Avec la mesure d'erreur OOB : nous mesurons le taux d'erreur OOB ($errOOB$) pour chaque valeur possible du couple (T, c_{max}). La Figure 5.9 montre les résultats obtenus du taux d'erreur OOB pour chaque variation de T et c_{max}. Les courbes vert, bleue et orange correspondent respectivement, aux valeurs de c_{max} égales à 85, $\sqrt{85}$ et $log_2(85)$. Nous remarquons, que la courbe orange ($c_{max} = log_2(85)$) est la plus basse, c'est la courbe qui affiche les valeurs les plus faibles de $errOOB$. Nous remarquons aussi que les deux courbes bleue ($c_{max} = \sqrt{85}$) et orange ($c_{max} = log_2(85)$) sont très proches avec une supériorité légère de la courbe orange. Nous avons identifié la valeur de T où le taux d'erreur OOB se stabilise avec une valeur minimale (autour de $0,06$). Nous avons trouvé une valeur d'environ 150 arbres, ce qui confirme le résultat obtenu avec la méthode de validation croisée de 3 groupes.

Nous avons aussi présenté la matrice de confusion de tous les gestes expressifs dans la Figure 5.10. Dans ce cas, nous avons 20 classes (5 gestes×4 émotions). Comme nous pouvons le voir, les va-

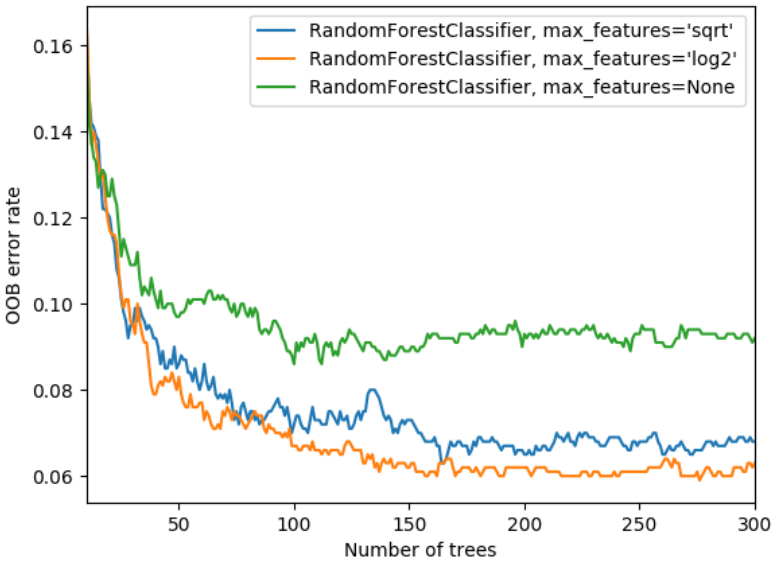


FIGURE 5.9 – Variation du taux d'erreur OOB ($errOOB$) en fonction de (T, c_{max}).

leurs les plus élevées sont marquées dans la diagonale de la matrice, ce qui confirme la performance de nos résultats. Suivant la matrice de confusion, il n'y a pas de confusion entre les gestes, mais il y a quelques confusions dans le même geste quand il est réalisé avec des émotions différentes. Par exemple, dans les gestes "avancer", "faire un signe" et "pointer" nous trouvons une confusion entre l'état "triste" et l'état "neutre". Généralement, nous pouvons dire que notre descripteur de mouvement parvient à caractériser les aspects quantitatif et aussi qualitatif du mouvement. Pour

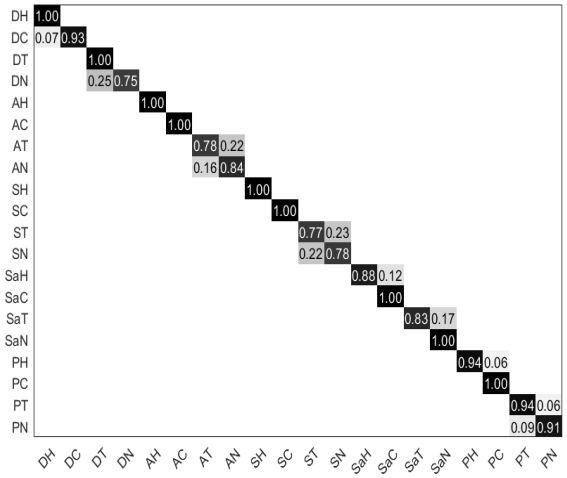


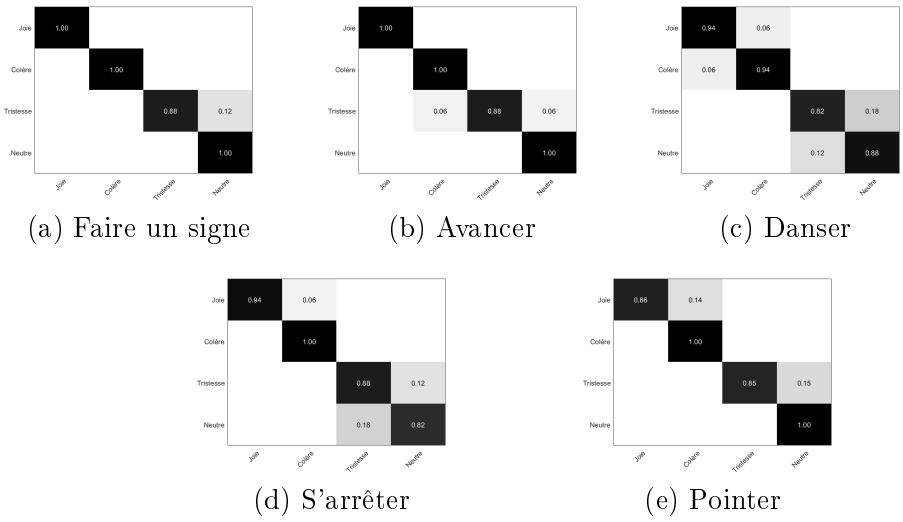
FIGURE 5.10 – Matrice de confusion des gestes expressifs, 5 gestes (D danser, A avancer, S Faire un signe, Sa S'arrêter et P pointer) effectués avec 4 états (H heureux, C en colère, T triste et N neutre).

le deuxième test, nous voulons classifier les émotions suivant le type de geste. Nous avons appliqué

Gestes	Faire un signe	Avancer	Danser	S'arreter	Pointer	Tous les gestes
F-score	0.93	0.80	0.75	0.84	0.91	0.83

TABLE 5.1 – Résultats de la reconnaissance des émotions dans les différents gestes de notre base.

la méthode RDF avec ses valeurs ajustées. La base de données est divisée en 5 groupes par classe de geste. Chaque groupe est composé de 220 séquences (1 geste×4 émotions×11 personnes×5 répétitions). La même méthode de validation croisée de 3 groupes est appliquée. La Table 5.1 montre les résultats obtenus dans la reconnaissance des émotions dans chaque groupe de geste considéré (danser, avancer, faire un signe, pointer et s'arrêter). Les matrices des confusions des différentes émotions sont présentées dans la Figure 5.11. Le résultats montrent que notre descripteur a réussi à faire la distinction entre les différentes émotions dans chaque type de geste. Certaines émotions ont été reconnues avec succès, par exemple les émotions de la joie et de la colère dans les gestes "faire un signe" et "avancer" ont été reconnues à 100%. Cependant, certaines confusions ont été obtenues dans le même mouvement, en particulier entre les émotions de la tristesse et neutre, comme dans les gestes "danser" et "s'arrêter". Selon ces résultats, nous pouvons approuver l'efficacité de notre descripteur de mouvement dans la reconnaissance des mouvements et des émotions humaines.



(a) Faire un signe (b) Avancer (c) Danser

(d) S'arrêter (e) Pointer

FIGURE 5.11 – Matrices de confusions entre les émotions exprimées (dans les lignes) et les émotions perçues (dans les colonnes) pour chaque geste avec la méthode RDF.

5.1.2 Sélection de caractéristiques pertinentes avec la méthode RDF

Méthodes de sélection de caractéristiques

La sélection de caractéristiques est un sujet important dans plusieurs domaines, notamment, dans la reconnaissance des formes, l'analyse exploratoire de données, en particulier pour les ensembles

de données de grande dimension. La sélection de caractéristiques (également appelée sélection de sous-ensembles) est un processus couramment utilisé dans l'apprentissage automatique, qui consiste à sélectionner un sous ensemble de caractéristiques les plus discriminantes parmi l'ensemble de caractéristiques disponibles. Le meilleur sous-ensemble contient les caractéristiques les plus pertinentes qui contribuent le plus à la précision. Cela permet d'identifier et supprimer autant que possible les informations non pertinentes et redondantes. Les algorithmes de sélection de caractéristiques peuvent être divisés en trois approches : Filtre, Wrapper et Embedded.

— Le modèle **Filtre** sélectionne les caractéristiques en fonction d'une mesure de performance. Le processus de sélection est indépendant du processus de classification. Ce modèle applique une mesure statistique pour attribuer un score à chaque caractéristique et fournir un classement. À partir de la liste classée, les caractéristiques avec les scores les plus élevés sont sélectionnées et les caractéristiques à faible score sont supprimées. Ces caractéristiques peuvent être choisies manuellement ou en définissant un seuil. Ce type d'approche repose donc uniquement sur les propriétés des données. Il est indépendant de tout algorithme d'apprentissage automatique particulier. Des exemples de cette méthode sont le gain d'information [Hoque et al., 2014], le rapport de gain [Witten et al., 2011], le test de Chi carré [Witten et al., 2011], les coefficients de corrélation [Yu and Liu, 2003], etc. Ces méthodes sont rapides (s'exécutent en une seule étape, voir Figure 5.12) et indépendantes du classifieur. En revanche, leur inconvénient majeur est qu'elles ignorent l'impact des sous ensembles choisis sur les performances de l'algorithme d'apprentissage.

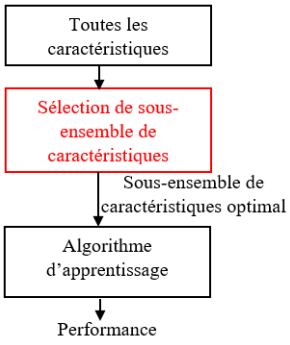


FIGURE 5.12 – Approche Filter.

— Le modèle **Wrapper** nécessite un algorithme d'apprentissage prédéterminé pour utiliser sa performance en tant que critère d'évaluation de l'ensemble des caractéristiques sélectionnées. Ce type d'approche génère des sous ensembles de caractéristiques et les évalue grâce à un algorithme de classification. Cette évaluation est répétée pour chaque sous-ensemble, et un appel d'algorithme de classification est fait à chaque évaluation. La génération de

chaque sous-ensemble dépend de la stratégie de recherche choisie (Voir Figure 5.13). Deux méthodes sont proposées dans l'approche de wrapper, la méthode de sélection backward, un algorithme de recherche introduit par [Marill and Green, 1963] et l'algorithme de sélection forward par [Whitney, 1971]. Dans le cas de la recherche backward, la méthode commence avec l'ensemble de toutes les caractéristiques et supprime les caractéristiques l'une après l'autre. A chaque étape cette méthode supprime la caractéristique qui a l'erreur la plus élevée jusqu'à ce que toute autre suppression augmente considérablement l'erreur. Dans cette recherche "descendante", les caractéristiques ignorées ne peuvent pas être sélectionnées à nouveau. Dans le cas de la recherche forward, la méthode commence sans variables et ajoute les caractéristiques l'une après l'autre. A chaque étape, elle ajoute la caractéristique qui a l'erreur minimale jusqu'à ce que tout autre ajout ne signifie aucune diminution d'erreur. Dans cette recherche "ascendante", les caractéristiques sélectionnées ne peuvent pas être rejetées plus tard. Les méthodes wrappers résolvent essentiellement le "vrai" problème (optimisation des performances du classifieur), mais ils sont également plus coûteux en termes de calcul par rapport aux méthodes de filtrage en raison des étapes d'apprentissage répétées et de validation croisée.

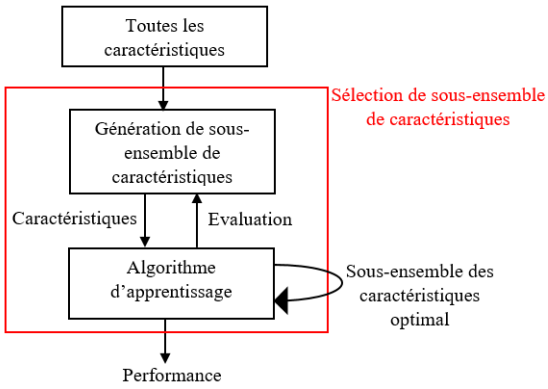


FIGURE 5.13 – Approche Wrapper.

— Le modèle **Embedded** : dans cette approche, la sélection de caractéristiques fait partie intégrante du modèle de classification. Ce méthodes évaluent les caractéristiques qui contribuent le mieux à la précision du modèle pendant la création du modèle. Nous pouvons citer l'exemple de la méthode RDF expliqué précédemment qui permet de mesurer l'importance des variables durant la création des arbres.

Notre méthode de sélection de caractéristiques basée sur RDF

Dans notre travail, l'objectif principal consiste à sélectionner le sous-ensemble optimal des caractéristiques pour la caractérisation de l'expression émotionnelle du corps dans les gestes de contrôle de notre base. Comme nous l'avons vu, la méthode RDF a montré sa performance dans l'étape d'apprentissage et aussi sa capacité de mesurer la pertinence des caractéristiques en se basant sur les mesures des taux d'erreur des échantillons OOBs. Donc, notre idée consiste à exploiter les avantages de cette méthode et la considérer comme un modèle embedded pour la sélection de caractéristiques. Elle permet de mesurer l'importance des caractéristiques au cours de la création des arbres. Donc, notre algorithme de sélection de caractéristiques importantes est considéré de la manière suivante : nous entrainons le modèle RDF avec toutes les caractéristiques et nous enregistrons le taux d'erreur OOB correspondant. Par la suite, nous mesurons l'importance de chaque caractéristique et nous les trions suivant l'ordre d'importance décroissant. Nous supprimons récursivement les caractéristiques les moins importantes. Ici, dans cette étape les caractéristiques les moins importantes et aussi celles qui sont redondantes sont supprimées en appliquant le test de Tukey HSD (Honestly Significant Difference). Donc, à la suite de chaque tri, nous supprimons l'ensemble qui ne contribue pas à un changement significatif de l'erreur OOB suivant le test de Tukey ($\alpha = 0.05$). Ceci rend notre système plus performant et plus rapide. Le processus s'arrête lorsque le nombre de caractéristiques restant dans l'ensemble est égal à 1. La sortie de cet algorithme donne une courbe qui décrit la diminution de l'erreur du système (taux d'erreur OOB) en fonction des sous ensembles de caractéristiques sélectionnés. Le sous-ensemble qui correspond à la valeur minimale de l'erreur OOB est considéré comme le sous-ensemble optimal. L'algorithme 2 résume les différentes étapes de notre algorithme de sélection de caractéristiques proposé :

Résultats des caractéristiques pertinentes avec RDF

Dans cette partie nous réalisons deux expérimentations, d'abord nous évaluons la pertinence des caractéristiques de notre descripteur de mouvement dans la caractérisation de l'ensemble entier des gestes expressifs de notre base. Dans la deuxième expérimentation, nous étudions l'importance des caractéristiques envers chaque émotion. Pour cela nous appliquons notre algorithme de sélection dans chaque geste de la base. Afin de définir les caractéristiques les plus pertinentes à chaque émotion, nous prenons à chaque fois comme référence l'état neutre.

La première étude consiste à considérer toute la base, donc le modèle RDF est entrainé sur tous les gestes d'apprentissage de la base. Nous appliquons l'algorithme 2 de sélection de caractéristiques

Algorithm 2 La méthode de la sélection de caractéristiques importantes.

Inputs : $v_0 = \{f_i\}, i = 1, \ldots, p$ •v_0 l'ensemble entier des caractéristiques.
Outputs : $v^* = \{f_j\}, j = 1, \ldots, p^*$ •v^* le sous-ensemble de caractéristiques les plus pertinentes.
$k = 0$
while $Card(v_k) \geq 1$ **do**
  Calculer et enregistrer le taux d'erreur de OOB : $E_k(v_k)$
  for $i = 1$ **to** p **do**
    Calculer $I(f_i)$ • l'importance de chaque caractéristique dans v_k.
  end for
  Trier $\{f_i\}$ dans l'ordre décroissant suivant les valeurs de $I(f_i)$
  $f_{min} = \underset{i}{\text{argmin}}\{I(f_i)\}$
  Appliquer le test de Tukey et sélectionner l'ensemble de caractéristiques $\{f_t\}$ qui ne contribue pas à un changement significatif de E_k
  $R = f_{min} \bigcup \{f_t\}$
  $v_{k+1} = v_k \backslash R$
  $k = k + 1$
end while
$v^* = \underset{l}{\text{argmin}}\{E_l(v_l))\}$ •v^* est le sous-ensemble optimal des caractéristiques avec l'erreur OOB minimal.

sur nos données. Les résultats sont présentés dans la Figure 5.14 qui affiche la courbe de l'erreur OOB en fonction des caractéristiques sélectionnées. La valeur minimale du taux d'erreur de OOB est de 0.04 obtenu avec l'ensemble entier des caractéristiques (85 caractéristiques). Cela confirme l'importance de la combinaison de toutes les composantes de LMA pour la caractérisation de notre base de gestes expressifs.

Dans la deuxième expérimentation, nous divisons la base de données en 5 groupes par classe de geste. Donc, à chaque fois nous considérons le même geste réalisé avec plusieurs émotions. Nous évaluons l'importance de chaque caractéristique pour la caractérisation de chaque émotion en comparant avec l'état neutre. Par conséquent, le modèle RDF est entrainé à chaque fois sur deux classes de gestes, un geste réalisé avec une émotion (la joie, la colère, ou la tristesse) et le même geste avec l'état neutre. Dans cette étude nous nous sommes davantage intéressés à l'aspect qualitatif du mouvement, vu qu'il s'agit du même geste effectué avec plusieurs émotions, donc nous aurons forcément la même importance entre les caractéristiques quantitatives. Pour cela, nous gardons que les facteurs des deux composantes Forme et Effort qui décrivent les différentes qualités expressives perçues dans les mouvements. La pertinence de chaque caractéristique est calculée avec le même algorithme de sélection utilisé avant. L'objectif de cette étude est de trouver les caractéristiques les plus pertinentes pour caractériser chaque émotion suivant le type de geste. La Table 5.2 résume les résultats trouvés pour chaque geste et chaque émotion. L'application de notre algorithme de sélection

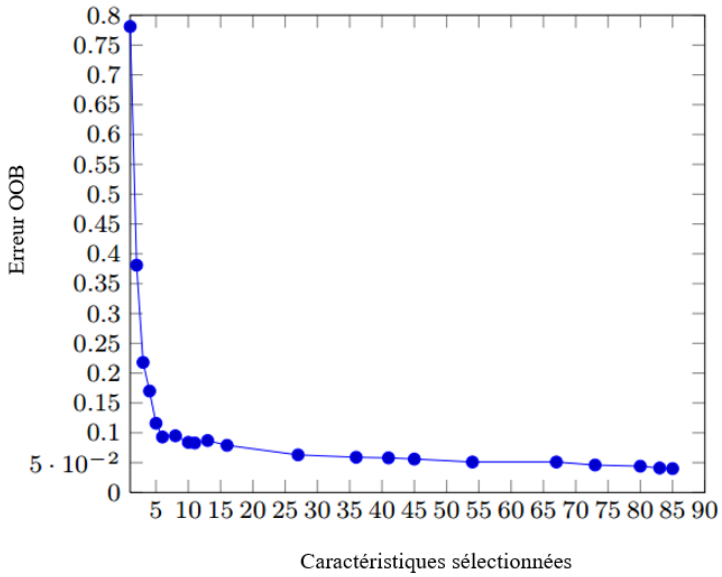


FIGURE 5.14 – Variation du taux d'erreur OOB en fonction des caractéristiques sélectionnées.

de caractéristiques montre que les facteurs de la composante Effort, particulièrement les qualités de temps et de poids impliquent une discrimination forte entre l'état neutre et les émotions de la joie et de la colère. Donc, on peut dire que les émotions de la joie et de la colère sont caractérisées par la rapidité et la force du mouvement. Tandis que l'émotion de la tristesse est caractérisée par les deux composantes de LMA (Effort et Forme), à l'exception du facteur du mouvement directionnel.

5.2 Caractérisation des gestes expressifs avec l'approche humaine

Après l'évaluation de la base expressive avec la méthode d'apprentissage, nous passons à l'expérimentation suivante. Nous répétons les même étapes mais cette fois en se basant sur l'approche humaine afin de pouvoir par la suite évaluer la précision de notre système de reconnaissance de gestes en se comparant aux résultats obtenus avec l'approche humaine. Deux études sont réalisées : la première concerne la classification des gestes expressifs et la deuxième présente l'évaluation de l'importance des caractéristiques de notre descripteur de mouvement.

5.2.1 Reconnaissance des gestes expressifs avec la perception humaine

Évaluation des émotions

L'évaluation des émotions avec la perception humaine peut être sous deux formes :

— Auto-évaluation : une auto-évaluation est un test, une mesure ou une enquête qui repose sur le propre rapport du participant sur ses sentiments, ses attitudes, ou ses croyances. Les auto-évaluations sont couramment utilisées dans les études psychologiques, en grande partie parce

TABLE 5.2 – Les caractéristiques pertinentes pour chaque émotion à travers les différents gestes.

Gestes expressifs		**Descripteur**					
		Forme		**Effort**			
		Flux de forme	Mise en forme	Temps	Poids	Espace	Flux
Heureux	Avancer			E_{vl} P_{vl} E_{vr} P_{vr}	M_{al} E_{al} M_{ar} E_{ar} P_{ar} E_{ah} P_{ah}		
	Danser				M_{al} E_{al} E_{ar} M_{ah} E_{ah}		
	Faire un signe			E_{vl} P_{vr}	E_{al} E_{ar}		
	S'arrêter			M_{vl} E_{vl} P_{vl} M_{vr} E_{vr} P_{vr} E_{vh} P_{vh}	M_{al} E_{al} P_{al} M_{ar} E_{ar} P_{ar} M_{ah} E_{ah} P_{ah} M_{as} E_{as} P_{as}		
	Pointer			P_{vl} E_{vr} P_{vr} E_{vs} P_{vs}	M_{al} M_{ar} E_{ar} P_{ar} M_{ah} E_{ah} P_{ah} M_{as} E_{as} P_{as}		
En colère	Avancer			P_{vl},	E_{al} E_{ar} P_{ar}		
	Danser				E_{al} P_{al} E_{ah}		
	Faire un signe			P_{vr}	M_{al} E_{al} P_{al} M_{ar} E_{ar} P_{ar}		
	S'arrêter				M_{ah} E_{ah} M_{as}		
	Pointer				E_{ar},P_{ar}		
Triste	Avancer	M_{datav} E_{datav} P_{datav}	D_H D_F	M_{vl} E_{vl} M_{vr} E_{vr}	M_{al} M_{ar} M_{ah} M_{as}	S^l S^r	$P^r_{tangage}$ P^r_{lacet} $P^h_{tangage}$ P^h_{lacet}
	Danser	E_{datav}		E_{vr},P_{vr}	M_{al} E_{al} P_{al} E_{ar} M_{as} E_{as}	S^r	$P^l_{tangage}$ P^r_{lacet} $P^h_{tangage}$
	Faire un signe		D_H D_S	M_{vl} E_{vl} P_{vl} M_{vr} E_{vr} P_{vr},	M_{al} E_{al} M_{ar} E_{ar}		
	S'arrêter	E_{datav} P_{datav}	D_F	M_{vl} E_{vl} P_{vl} M_{vr} E_{vr} P_{vr} M_{vh} E_{vh} E_{vs} P_{vs}	M_{al} E_{al} $M_{ar}E_{ar}$ M_{as} E_{as}	S^s	$P^r_{tangage}$
	Pointer	M_{datav} P_{datav}		M_{vr} E_{vr} P_{vr}	M_{ar} E_{ar} Pa^r	S^r	$P^h_{tangage}$ P^h_{lacet}

que beaucoup d'informations précieuses et diagnostiques sur une personne sont révélées à un chercheur ou un clinicien basé sur le rapport d'une personne sur elle-même. L'un des outils d'auto-évaluation le plus couramment utilisé est l'Inventaire multiphasique de personnalité du Minnesota (MMPI) pour les tests de personnalité. Il s'agit d'un auto-questionnaire de personnalité à visée diagnostique, descriptive et thérapeutique. Cet outil a l'avantage d'être moins couteux en temps que les méthodes d'observation. Il peut atteindre beaucoup plus de sujets de test de sorte qu'un chercheur peut obtenir des résultats en quelques jours sans devoir observer une population au cours de périodes qui peuvent durer plus longtemps. Toutefois, la collecte d'informations via un rapport d'auto-évaluation a ses limites. En effet, les résultats sont souvent biaisés lorsque les personnes rapportent leurs propres expériences par plusieurs facteurs. Par exemple, elles peuvent donner la réponse la plus acceptable socialement plutôt que d'être véridiques. Elles peuvent aussi être incapables de s'évaluer correctement. Le libellé des questions peut être source de confusion ou avoir des significations différentes pour différents sujets. Parfois dans le domaine de la médecine, ces études peuvent avoir dans certains

cas des problèmes de validité. Les patients peuvent exagérer les symptômes afin de rendre leur situation pire, ou sous-estimer la gravité ou la fréquence des symptômes afin de minimiser leurs problèmes.

— Évaluation des observateurs : il s'agit d'une évaluation réalisée par un ensemble d'observateurs sur les comportements, les symptômes, ou les attitudes des autres personnes. Le format de réponse des observateurs a souvent été basé sur l'option du choix forcé, l'attribution d'une seule étiquette à un comportement corporel exprimé est nécessaire et l'étiquette la plus fréquente est utilisée. Les évaluations basées sur des observateurs ont montré des avantages par rapport aux auto-évaluations surtout dans le domaine psychologique dans l'évaluation de la personnalité [Connelly and Ones, 2010, Poropat, 2014]. Il y a plusieurs raisons pour cet avantage, nous pouvons citer la fiabilité dans les évaluations (c.-à-d. Cronbach alpha plus élevé [Balsis et al., 2015]). De plus, cette fiabilité est une condition nécessaire à la validité [Hundleby, 1968].

Dans notre cas, nous avons choisi le deuxième type d'évaluation, au vu des limites de la première méthode. Nous nous intéressons à la fiabilité du test plus qu'à sa durée.

Participants

Nous faisons appel à 10 observateurs (5 hommes et 5 femmes) de l'Université d'Evry Val d'Essonne dont leurs âges varient entre 28 et 37 ans (moyenne= 30.9 ans, écart-type= 3.16). Chaque participant est invité à regarder les vidéos enregistrées et évaluer l'émotion exprimée dans chaque geste en utilisant l'échelle de Likert à 5 éléments (de 1= fortement en désaccord, 3=neutre, à 5= fortement d'accord). Pour réaliser une évaluation fiable, tous les gestes expressifs enregistrés dans les vidéos sont reproduits par un avatar virtuel (voir Figure 5.15). Cela aide les observateurs à évaluer les émotions sans être influencés par certains facteurs comme les expressions faciales, le genre, le sexe, etc.

Fiabilité inter-observateurs dans l'évaluation des émotions

Après l'évaluation par les différents observateurs, une étape très importante doit être réalisée pour la validation de la partie expérimentale, qui est l'estimation du degré d'homogénéité et de cohésion entre les observateurs. Cette étude permet d'identifier les parties qui contribuent peu à l'évaluation. L'indice de fiabilité s'exprime alors par la mesure de la cohérence entre les différents observateurs. La mesure est dite fiable s'il existe un degré d'accord suffisamment élevé entre les différentes évaluations et n'est pas fiable dans le cas contraire. Plusieurs indicateurs statistiques

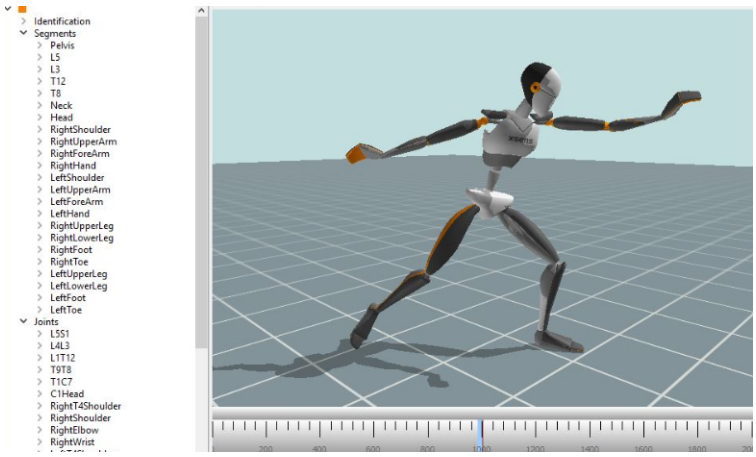


FIGURE 5.15 – Reproduction des gestes avec un avatar.

permettent d'évaluer l'accord inter-éléments, le plus utilisé est le coefficient de Cronbach, donné par la formule suivante :

$$\alpha = \frac{n}{n-1}(1 - \frac{\sum_{i=1}^{n} V_i}{V_T}) \tag{5.6}$$

avec n est le nombre des éléments, V_i se réfère à la variance des scores sur chaque élément et V_T représente la variance totale de l'ensemble des scores. Le Cronbach alpha s'interprète comme un coefficient de corrélation classique, plus il est proche de 1, plus le score est fiable. Nous calculons ce coefficient pour mesurer la corrélation entre les différents scores donnés par les observateurs pour chaque émotion. Comme le montre la Figure 5.16, le coefficient de Cronbach est toujours supérieur à 0.8 : pour heureux (0.897), en colère (0.894), triste (0.879) et neutre (0.814). Cela signifie, selon [Tavakol and Dennick, 2011], qu'il y a une grande cohérence entre les différentes évaluations faites par les observateurs dans la perception des émotions.

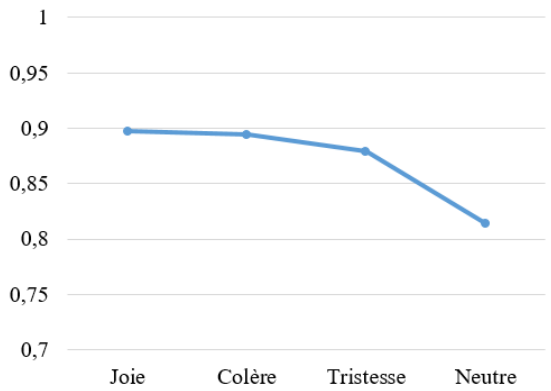
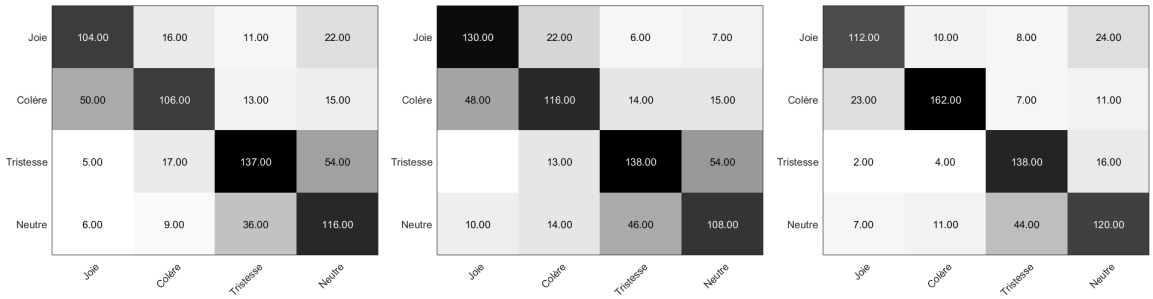


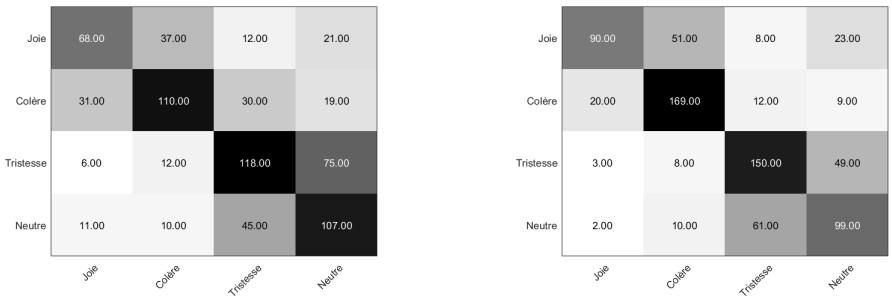
FIGURE 5.16 – Fiabilité inter-observateurs dans la perception des émotions à l'aide du coefficient de Cronbach.

Résultats

Pour la classification des émotions avec l'approche humaine, nous prenons les évaluations résultantes des observateurs, et nous considérons une émotion reconnue si le score donné est supérieur à 3 (état neutre). Afin de classifier les émotions, nous prenons à chaque fois un geste et pour chaque émotion nous calculons le nombre de fois que le score est supérieur à 3. Les résultats sont présentés dans les matrices de confusion dans la Figure 5.17. Les lignes correspondent aux émotions exprimées et les colonnes aux émotions perçues (évaluées) par les observateurs. Les cellules diagonales correspondent au nombre de fois que les émotions rapportées par les observateurs (score > 3) sont reconnues. Les cellules hors diagonales contiennent la fréquence des mauvaises classifications. Comme nous pouvons le remarquer, les valeurs les plus élevées sont dans la diagonale dans tous les gestes, avec des confusions plus importantes entre les émotions de la joie et de la colère et entre l'émotion de la tristesse et l'état neutre. Dans une deuxième expérimentation, nous considérons tous les gestes



(a) Danser (b) Avancer (c) Faire un signe



(d) S'arrêter (e) Pointer

FIGURE 5.17 – Matrices des confusions entre les émotions exprimées (les lignes) et les émotions perçues (les colonnes) pour chaque geste en se basant sur les scores donnés par les observateurs.

expressifs et nous calculons la moyenne des scores des observateurs dans chaque émotion exprimée. Comme nous pouvons le voir sur la Figure 5.18, l'émotion de la joie a été reconnue avec succès par les observateurs avec quelques confusions apparues avec l'émotion de la colère. L'émotion de la joie est faiblement confondue avec celle de la tristesse. Pour l'émotion de la colère, le score moyen le plus

élevé a été obtenu dans la perception de l'émotion de la colère suivie de l'émotion de la joie. Les états neutre et de la tristesse ont été bien reconnus par les observateurs dans les différents gestes avec une confusion bidirectionnelle entre les deux émotions par certains observateurs.

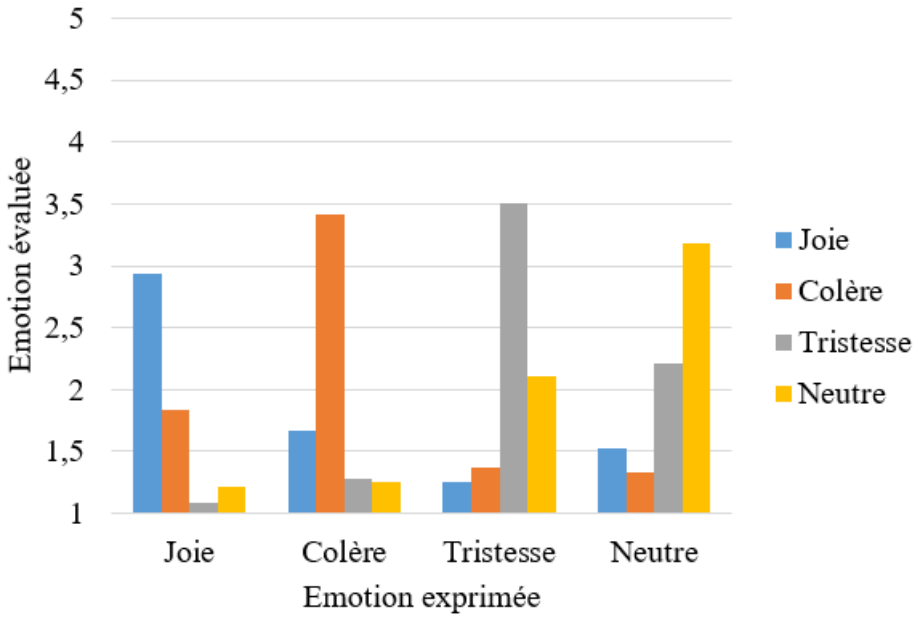


FIGURE 5.18 – Les scores moyens de la perception des émotions par 10 observateurs.

5.2.2 Sélection des caractéristiques avec l'approche humaine

Pour la deuxième expérimentation concernant l'évaluation de notre descripteur avec l'approche humaine, nous demandons à un autre groupe de 10 observateurs de l'Université d'Evry Val d'Essonne, leurs âges varient entre 28 et 31 ans, à regarder les même vidéos et évaluer les composantes de LMA. De même, dans cette partie, nous nous sommes concentrés sur les facteurs des deux composantes Effort-Forme (flux de forme, mise en forme, mouvement directionnel, espace, temps, poids et flux), parce qu'elles sont les seules responsables à la spécification des mouvements humains expressifs. Nous utilisons l'échelle de Likert à 7 éléments comme suit :

- Flux de forme : volume de l'enveloppe convexe du squelette (de 1=très petit à 7=très grand)
- Mise en forme : distance entre le centre de squelette et les extrémités (de 1=très contracté à 7=très étendu)
- Mouvement directionnel : la voie des articulations du corps (de 1=très curviligne à 7=très rectiligne)
- Temps : la vitesse du mouvement (de 1=très soutenu à 7=très soudain)
- Poids : la force du mouvement (de 1=très léger à 7=très fort)
- Espace : la rectitude du mouvement (de 1=très indirect à 7=très direct)

— Flux : la souplesse du mouvement (de 1=très libre à 7=très lié)

Par exemple, pour évaluer le facteur de flux de forme il faut percevoir le développement du volume de l'enveloppe convexe du squelette. Si l'observateur perçoit que le volume du squelette a fortement augmenté tout au long du mouvement, il attribue au facteur de flux de forme un score de 7 et s'il perçoit une forte diminution du volume, il donne le score de 1.

Fiabilité inter-observateurs dans l'évaluation des caractéristiques

De même, cette évaluation nécessite la mesure de la fiabilité entre les scores donnés par les observateurs dans l'évaluation des caractéristiques. Suivant les résultats présentés dans la Figure 5.19, le coefficient de Cronbach est supérieur à 0.7 (niveau accepté) dans toutes les évaluations. Cela confirme la cohérence entre les observateurs lors de l'évaluation des caractéristiques dans notre base expressive. Le coefficient de fiabilité le plus élevé est de $0,958$ obtenu dans l'évaluation des facteurs de temps et de poids. Ce résultat montre une forte homogénéité entre les observateurs dans l'estimation de ces deux facteurs. Cependant, il existe une corrélation moins importante mais acceptable dans l'évaluation des facteurs d'espace, flux et mouvement directionnel, respectivement avec des coefficients de Cronbach de $0,754$, $0,763$ et $0,703$. Pour les facteurs de mise en forme et flux de forme, il y a un bon accord entre les scores attribués par les observateurs, respectivement avec des coefficients de Cronbach de $0,830$ et $0,839$.

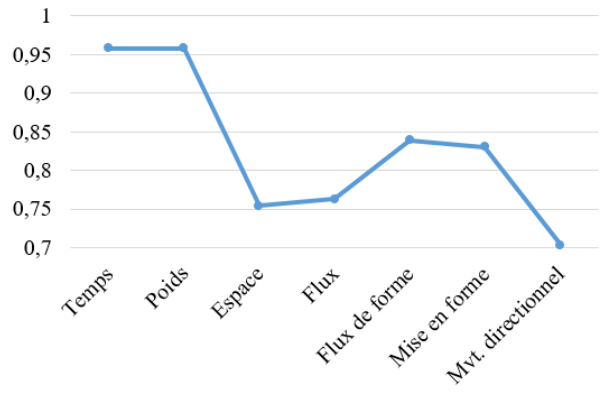


FIGURE 5.19 – Fiabilité inter-observateurs dans l'évaluation des facteurs des composantes Effort-Forme avec la mesure de coefficient de Cronbach.

Résultats

Pour étudier l'importance des caractéristiques dans la caractérisation des émotions humaines, nous mesurons la corrélation entre les évaluations faites dans la perception des émotions et dans la caractérisation des facteurs de LMA. Cela nous aide à définir la relation entre les caractéristiques de

TABLE 5.3 – Coefficients de Pearson r pour la corrélation entre les scores donnés aux facteurs de Effort-Forme et ceux donnés aux émotions exprimées (**. la corrélation est significative au niveau 0.001).

		Heureux	**en Colère**	**Triste**	**Neutre**
Forme	**Flux de forme**	**0.541****	0.132	**-0.524****	-0.316**
	Mise en forme	**0.542****	0.194**	**-0.613****	-0.328**
	Mvt Directionnel	0.269**	**0.568****	**-0.505****	-0.397**
Effort	**Temps**	**0.555****	**0.622****	**-0.795****	**-0.554****
	Poids	**0.543****	**0.640****	**-0.780****	**-0.594****
	Espace	0.326**	**0.682****	**-0.566****	-0.487**
	Flux	-0.316**	**-0.665****	**0.559****	0.497**

descripteur et les 4 états (heureux, en colère, triste et neutre). Nous collectons les scores donnés aux facteurs d'Effort-Forme et à la perception des émotions et nous calculons le coefficient de corrélation de Pearson. Les coefficients de Pearson sont utilisés dans les statistiques pour mesurer la relation entre deux variables, appelée r Pearson, avec une valeur comprise entre -1 pour une corrélation parfaitement négative et $+1$ pour une corrélation parfaitement positive, 0 si les deux variables ne représentent aucune corrélation. La corrélation de Pearson entre les variables X et Y est calculée par la formule suivante :

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5.7}$$

La Table 5.3 résume les résultats des coefficients de Pearson obtenus dans cette étude : Pour l'émotion de la joie, nous trouvons une corrélation positive avec les facteurs de la Forme (Flux de forme $r = 0.541, p < 0.001$; Mise en forme $r = 0.542, p < 0.001$; Mvt directionnel $r = 0.269, p < 0.001$) et les facteurs d'Effort (Temps $r = 0.555, p < 0.001$; Poids $r = 0.543, p < 0.001$). Donc l'émotion de la joie est significativement caractérisée par une augmentation de la Forme et une extension des membres du corps. Cette émotion est également associée à la rapidité et la force du mouvement. [Masuda and Kato, 2010] ont également confirmé dans leur expérience la même relation entre l'émotion de la joie et les qualités d'Effort.

Pour l'émotion de la colère, nous trouvons une forte corrélation positive avec les trois facteurs d'Effort suivants (Temps $r = 0.622, p < 0.001$, Poids $r = 0.640, p < 0.001$, Espace $r = 0.682, p < 0.001$) et une corrélation positive avec le facteur du mouvement directionnel ($r = 0,568, p < 0,001$). Le facteur de flux est négativement associé à l'émotion de la colère ($r = -0,665, p < 0,001$). Donc, l'émotion de la colère est fortement caractérisée par un mouvement rapide, fort, direct, rectiligne et libre. Certaines qualités sont cohérentes avec celles trouvées par [Masuda and Kato, 2010]. Les auteurs ont trouvé une corrélation positive entre l'émotion de la colère et les qualités de la rapidité

et la force du mouvement. Pour les émotions de la joie et de la colère, nous avons trouvé une relation similaire avec les facteurs des composantes Effort-Forme, mais avec une importance de corrélation différente. Dans l'évaluation des qualités d'Effort, l'émotion de la colère a été évaluée comme significativement plus rapide, plus forte et plus libre que l'émotion de la joie. Cependant, dans l'évaluation des facteurs de Forme, l'émotion de la joie a été évaluée par une forme plus développée.

L'émotion de la tristesse est négativement corrélée avec tous les facteurs d'Effort-Forme, à l'exception du facteur de flux : facteurs de forme (flux de forme $r = -0.524, p < 0.001$; mise en forme $r = -0.613, p < 0.001$; mvt directionnel $r = -0.505, p < 0.001$), facteurs d'effort (temps $r = -0.795, p < 0.001$; poids $r = -0.780, p < 0.001$; espace $r = -0.566, p < 0.001$; Flux $r = 0.559, p < 0.001$). Selon l'évaluation des facteurs de Forme, l'émotion de la tristesse était significativement caractérisée par une forme rétrécie, des extrémités du corps contractées et un mouvement courbé. Selon l'évaluation des facteurs d'Effort, l'émotion de la tristesse est caractérisée par un mouvement léger, lié, soutenu et indirect.

L'état neutre : a eu une relation avec les qualités d'Effort-Forme similaire à celle obtenue avec l'émotion de la tristesse mais avec une corrélation moins forte. De même, [Masuda and Kato, 2010] ont trouvé que les deux émotions (relâché et tristesse) sont en corrélation avec les mêmes qualités d'Effort (lenteur et faiblesse).

5.3 Évaluation du système

Le but de cette étude est de caractériser et de reconnaître les émotions humaines exprimées à travers le mouvement du corps d'abord en se basant sur la méthode d'apprentissage automatique et par la suite sur l'approche humaine afin d'évaluer la performance de notre système de reconnaissance. Nous avons obtenu les conclusions suivantes :

— Dans l'étude de la reconnaissance des émotions, les classifieurs humains et RDF ont réussi à faire la distinction entre les 4 états, avec quelques confusions entre les émotions de la joie et de la colère et entre les deux états neutre et triste. Si nous comparons les deux résultats, nous trouvons que la méthode d'apprentissage était plus précise que les observateurs. Cela peut être dû au nombre limité d'observateurs ayant participé à l'évaluation des différentes séquences. En outre, cela peut s'expliquer par le type de geste choisi. Nous essayons de reconnaître les émotions à travers des gestes limités. Par exemple, en effectuant le geste de pointage avec l'émotion de la tristesse ou avec l'état neutre, nous aurons presque le même mouvement avec un rythme un peu stable. Donc, visuellement, ça sera difficile pour les observateurs de

distinguer les deux types de mouvement.

— Dans l'étude de l'importance des caractéristiques avec l'approche humaine, chaque observateur a évalué l'importance de chaque facteur de LMA dans la caractérisation de chaque émotion. Avec le classifieur RDF, l'algorithme consiste à étudier l'importance des caractéristiques dans la caractérisation des émotions par rapport à l'état neutre. De plus, nous appliquons le test de Tukey pour supprimer les caractéristiques redondantes. Comme présenté dans la Table 5.2, l'émotion de la tristesse était caractérisée par les facteurs des deux composantes Effort-Forme, à l'exception du facteur de mouvement directionnel. Le même résultat est obtenu dans l'évaluation des observateurs dans la Table 5.3. Cependant, pour les émotions de la joie et de la colère, le classifieur RDF a trouvé que les deux caractéristiques les plus discriminantes dans la caractérisation de ces deux émotions sont les facteurs de temps et de poids. Ces résultats confirment que notre méthode proposée permet de caractériser les émotions et de définir les caractéristiques importantes tout en optimisant notre descripteur de mouvement en gardant que les caractéristiques à la fois pertinentes et non redondantes.

5.4 Bilan

Dans ce chapitre nous avons proposé deux approches différentes, la première se repose sur le classifieur RDF et la deuxième sur la perception humaine. Deux expérimentations ont été réalisées avec les deux approches : la première pour la classification des gestes expressifs ainsi que les émotions par classe de geste et la deuxième pour l'identification des caractéristiques du mouvement les plus pertinentes. L'approche humaine repose sur une étude statistique basée sur les avis des participants avec des scores indiquant le niveau de la perception humaine pour chaque émotion. A la fin, nous évaluons la robustesse de notre système de reconnaissance de gestes en se référant à l'approche humaine. Nous trouvons que dans la phase de classification notre système a surpassé les résultats donnés par les observateurs. Dans l'étude de l'importance des caractéristiques, notre système a réussi à caractériser les gestes expressifs avec un descripteur à la fois optimal et pertinent.

Chapitre 6

Conclusion générale et perspectives

Sommaire

6.1 Conclusion 147

6.2 Perspectives 148

6.1 Conclusion

Dans ce travail, nous traitons le problème de la reconnaissance des gestes en développant un système robuste et performant qui considère le geste et aussi l'émotion de la personne. Trois approches sont réalisées dans ce travail :

— La première consiste à la reconnaissance des gestes dynamiques avec les modèles de Markov cachés discrets. Un descripteur de mouvement local est implémenté pour la représentation des mouvement. Nous avons réussi à minimiser la taille du descripteur avec notre algorithme d'échantillonnage proposé et l'adapter aux entrées du modèle MMC discret. Une contribution est réalisée au modèle MMC pour améliorer les taux de reconnaissance dans des conditions de similitudes entre les mouvements, où la séquence des gestes est traitée dans deux sens (le sens naturel et le sens inverse). L'évaluation du système est faite sur trois bases d'actions publiques et notre base de gestes de contrôles. Les taux de reconnaissance obtenus varient entre 80% et 99.7%.

— La deuxième approche était de développer un système de reconnaissance des gestes expressifs avec des méthodes d'apprentissage globales. Un descripteur global est crée afin de décrire le mouvement entier et son expressivité. Un ajustement des différents paramètres des méthodes d'apprentissage est réalisé suivi par une étude comparative entre ces méthodes afin de sélectionner la meilleure. L'évaluation du système est faite sur des bases publiques et notre base

composée de gestes expressifs. Cette approche nous a permis de choisir la méthode RDF pour les prochains études.

— La troisième approche consiste à évaluer notre système établi en se référant à l'approche humaine. Il s'agit d'une étude statistique qui repose sur les avis d'un ensemble d'observateurs dans la perception des émotions et aussi l'évaluation du descripteur de mouvement proposé. Un algorithme de sélection de caractéristiques est mis en place pour étudier l'importance de chacune envers l'expression de chaque émotion. Finalement, les résultats issus de la méthode d'apprentissage automatique sont comparés à ceux de l'approche humaine pour conclure sur la fiabilité de notre système. Notre système de reconnaissance des gestes a réussi à classifier les émotions comme un humain et sélectionner des caractéristiques pertinentes communes avec celles choisies par l'approche humaine tout en optimisant la taille de notre descripteur de mouvement.

Notre système traite des données dans une complexité de calcul faible grâce aux algorithmes proposées (échantillonnage, quantification, sélection de caractéristiques) et atteint une bonne précision de prédiction. Nos résultats sont comparables aux systèmes spécialisés de l'état de l'art.

6.2 Perspectives

Comme perspectives, il serait intéressant d'explorer notre système de reconnaissance de gestes dans une application robotique comme la notre qui consiste à contrôler le robot NAO via les gestes. En appliquant notre système, le robot sera capable de reconnaitre les gestes de la personne et aussi son émotion à travers son mouvement. Ainsi, il pourra faire les tâches associées aux gestes tout en interagissant avec la personne suivant son humeur, ce qui rend l'interaction entre les deux parties plus naturelle. Cependant, cet objectif nécessite quelques améliorations dans notre système :

— Dans la partie reconnaissance des gestes, notre système fonctionne en mode "offline". Donc nous envisageons de rendre le fonctionnement en mode "online". Cela nécessite la phase de la détection de début et fin du geste.

— Dans la partie émotionnelle, quelques perspectives sont envisagées telles que : l'amélioration de notre base expressive pour reconnaitre un nombre plus grand d'émotions. Aussi, à partir de la relation trouvée entre le descripteur de mouvement et les émotions, nous envisageons à développer une interface graphique qui permet de régler les émotions en modifiant les paramètres de mouvement et appliquer ce système sur un avatar animé. Par exemple pour améliorer l'émotion de la joie on peut augmenter ou diminuer la caractéristique responsable à la caractérisation de cette émotion.

Bibliographie

[Ahmad and Lee, 2010] Ahmad, M. and Lee, S.-W. (2010). Variable silhouette energy image representations for recognizing human actions. *Image and vision computing*, 28(5) :814–824.

[Ajili et al., 2016] Ajili, I., Mallem, M., and Didier, J.-Y. (2016). Gesture recognition for robot teleoperation.

[Ajili et al., 2017a] Ajili, I., Mallem, M., and Didier, J.-Y. (2017a). Gesture recognition for humanoid robot teleoperation. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*, pages 1115–1120. IEEE.

[Ajili et al., 2017b] Ajili, I., Mallem, M., and Didier, J.-Y. (2017b). Robust human action recognition system using laban movement analysis. *Procedia Computer Science*, 112 :554–563.

[Ajili et al., 2018a] Ajili, I., Mallem, M., and Didier, J.-Y. (2018a). An efficient motion recognition system based on lma technique and a discrete hidden markov model. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 12(9) :707 – 713.

[Ajili et al., 2018b] Ajili, I., Mallem, M., and Didier, J.-Y. (2018b). Relevant lma features for human motion recognition. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 12(9) :792 – 796.

[Alwani et al., 2014] Alwani, A. A., Chahir, Y., Goumidi, D. E., Molina, M., and Jouen, F. (2014). 3d-posture recognition using joint angle representation. In Laurent, A., Strauss, O., Bouchon-Meunier, B., and Yager, R. R., editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 106–115, Cham. Springer International Publishing.

[Aristidou et al., 2015a] Aristidou, A., Charalambous, P., and Chrysanthou, Y. (2015a). Emotion Analysis and Classification : Understanding the Performers' Emotions Using the LMA Entities. *Computer Graphics Forum*.

[Aristidou and Chrysanthou, 2014] Aristidou, A. and Chrysanthou, Y. (2014). Feature extraction for human motion indexing of acted dance performances. In *2014 International Conference on Computer Graphics Theory and Applications (GRAPP)*, pages 1–11.

[Aristidou et al., 2015b] Aristidou, A., Stavrakis, E., Charalambous, P., Chrysanthou, Y., and Himona, S. L. (2015b). Folk dance evaluation using laban movement analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 8(4) :20.

[Aristidou et al., 2014a] Aristidou, A., Stavrakis, E., and Chrysanthou, Y. (2014a). Lma-based motion retrieval for folk dance cultural heritage. In Ioannides, M., Magnenat-Thalmann, N., Fink, E., Žarnić, R., Yen, A.-Y., and Quak, E., editors, *Digital Heritage. Progress in Cultural Heritage : Documentation, Preservation, and Protection*, pages 207–216, Cham. Springer International Publishing.

[Aristidou et al., 2014b] Aristidou, A., Stavrakis, E., and Chrysanthou, Y. (2014b). *LMA-Based Motion Retrieval for Folk Dance Cultural Heritage*, pages 207–216. Springer International Publishing, Cham.

[Aristidou et al., 2017a] Aristidou, A., Stavrakis, E., Papaefthimiou, M., Papagiannakis, G., and Chrysanthou, Y. (2017a). Style-based motion analysis for dance composition. *The Visual Computer*.

[Aristidou et al., 2017b] Aristidou, A., Zeng, Q., Stavrakis, E., Yin, K., Cohen-Or, D., Chrysanthou, Y., and Chen, B. (2017b). Emotion control of unstructured dance movements. In *Symposium on Computer Animation*.

[Aviezer et al., 2008] Aviezer, H., Hassin, R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., and Bentin, S. (2008). Angry, disgusted, or afraid ? studies on the malleability of emotion perception. *Psychological science*, 19 7 :724–32.

[Balsis et al., 2015] Balsis, S., Cooper, L. D., and Oltmanns, T. F. (2015). Are informant reports of personality more internally consistent than self reports of personality ? *Assessment*, 22(4) :399–404. PMID : 25376588.

[Barakova and Lourens, 2010] Barakova, E. I. and Lourens, T. (2010). Expressing and interpreting emotional movements in social games with robots. *Personal and ubiquitous computing*, 14(5) :457–467.

[Barber et al., 1996] Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4) :469–483.

[Barros et al., 2017] Barros, P., Maciel-Junior, N. T., Fernandes, B. J., Bezerra, B. L., and Fernandes, S. M. (2017). A dynamic gesture recognition and prediction system using the convexity approach. *Computer Vision and Image Understanding*, 155 :139–149.

[Bartenieff et al., 1984] Bartenieff, I., Hackney, P., Jones, B. T., Van Zile, J., and Wolz, C. (1984). The potential of movement analysis as a research tool : a preliminary analysis. *Dance Research Journal*, 16(1) :3–26.

[Bauer and Kohavi, 1999] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms : Bagging, boosting, and variants. *Machine Learning*, 36(1) :105–139.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1) :164–171.

[Beale and Peter, 2008] Beale, R. and Peter, C. (2008). *Affect and emotion in human-computer interaction*. Springer.

[Blank et al., 2005] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *null*, pages 1395–1402. IEEE.

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3) :257–267.

[BOULARD H., 2003] BOULARD H., B. S. (2003). *Hidden Markov Model*. MIT Press, England London.

[Breiman, 1984] Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont, Calif.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1) :5–32.

[Breuer et al., 2007] Breuer, P., Eckes, C., and Müller, S. (2007). Hand gesture recognition with a novel ir time-of-flight range camera–a pilot study. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 247–260. Springer.

[Broughton and Stevens, 2009] Broughton, M. and Stevens, C. (2009). Music, movement and marimba : an investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychology of Music*, 37(2) :137–153.

[Bunke et al., 1995] Bunke, H., Roth, M., and Schukat-Talamazzini, E. (1995). Off-line cursive handwriting recognition using hidden markov models. *Pattern Recognition*, 28(9) :1399 – 1413.

[Bylander, 2002] Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48(1) :287–297.

[Cadoz, 1994] Cadoz, C. (1994). Le geste canal de communication homme/machine : la communication" instrumentale". *Technique et science informatiques*, 13(1) :31–61.

[Camurri et al., 2000] Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R., and Volpe, G. (2000). Eyesweb : Toward gesture and affect recognition in interactive dance and music systems. *Comput. Music J.*, 24(1) :57–69.

[Camurri et al., 2004a] Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., and Volpe, G. (2004a). Multimodal analysis of expressive gesture in music and dance performances. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction*, pages 20–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Camurri et al., 2004b] Camurri, A., Mazzarino, B., and Volpe, G. (2004b). Analysis of expressive gesture : The eyesweb expressive gesture processing library. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction*, pages 460–467, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Cao et al., 2010] Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z., and Huang, T. S. (2010). Action detection using multiple spatial-temporal interest point features. In *2010 IEEE International Conference on Multimedia and Expo*, pages 340–345.

[Chaaraoui et al., 2012] Chaaraoui, A. A., Climent-Pérez, P., and Flórez-Revuelta, F. (2012). An efficient approach for multi-view human action recognition based on bag-of-key-poses. In Salah, A. A., Ruiz-del Solar, J., Meriçli, Ç., and Oudeyer, P.-Y., editors, *Human Behavior Understanding*, pages 29–40, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Chaaraoui et al., 2014] Chaaraoui, A. A., Padilla-López, J. R., Climent-Pérez, P., and Flórez-Revuelta, F. (2014). Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert Systems with Applications*, 41(3) :786 – 794. Methods and Applications of Artificial and Computational Intelligence.

[Chakraborty et al., 2012] Chakraborty, B., Holte, M. B., Moeslund, T. B., and Gonzílez, J. (2012). Selective spatio-temporal interest points. *Comput. Vis. Image Underst.*, 116(3) :396–410.

[Chen et al., 2003] Chen, F.-S., Fu, C.-M., and Huang, C.-L. (2003). Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing*, 21(8) :745 – 758.

[Chi et al., 2000] Chi, D., Costa, M., Zhao, L., and Badler, N. (2000). The emote model for effort and shape. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 173–182, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Chow et al., 1987] Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, G., Makhoul, J., Price, P., Roucos, S., and Schwartz, R. (1987). Byblos : The bbn continuous speech recogni-

tion system. In *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 89–92.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 :603–619.

[Connelly and Ones, 2010] Connelly, B. and Ones, D. S. (2010). An other perspective on personality : meta-analytic integration of observers' accuracy and predictive validity. *Psychological bulletin*, 136 6 :1092–122.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

[de Gelder, 2006] de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3) :242–249.

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Dollar et al., 2005a] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005a). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA. IEEE Computer Society.

[Dollar et al., 2005b] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005b). Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72.

[Droeschel et al., 2011] Droeschel, D., Stückler, J., and Behnke, S. (2011). Learning to interpret pointing gestures with a time-of-flight camera. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 481–488. ACM.

[Duchenne de Boulogne, 1990] Duchenne de Boulogne, G. B. (1990). *The Mechanism of Human Facial Expression*. Studies in Emotion and Social Interaction. Cambridge University Press.

[Durupinar et al., 2016] Durupinar, F., Kapadia, M., Deutsch, S., Neff, M., and Badler, N. I. (2016). Perform : Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Trans. Graph.*, 36(4).

[Ekman et al., 1990] Ekman, P., Dav, R. J., and Friesen, W. V. (1990). The duchenne smile : emotional expression and brain physiology ii. *Journal of Personality and Social Psychology*, pages 342–353.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). *Facial Action Coding System : A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.

[Evangelidis et al., 2014] Evangelidis, G., Singh, G., and Horaud, R. (2014). Skeletal quads : Human action recognition using joint quadruples. In *2014 22nd International Conference on Pattern Recognition*, pages 4513–4518.

[Fang et al., 2009] Fang, C.-H., Chen, J.-C., Tseng, C.-C., and Lien, J.-J. J. (2009). Human action recognition using spatio-temporal classification. In *Asian Conference on Computer Vision*, pages 98–109. Springer.

[Fielding and Ruck, 1995] Fielding, K. H. and Ruck, D. W. (1995). Spatio-temporal pattern recognition using hidden markov models. *IEEE Transactions on Aerospace and Electronic Systems*, 31(4) :1292–1300.

[Fink, 2014] Fink, G. A. (2014). *Markov Models for Pattern Recognition : From Theory to Applications*. Springer Publishing Company, Incorporated, 2nd edition.

[Foroud and Whishaw, 2006] Foroud, A. and Whishaw, I. Q. (2006). Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke : A laban movement analysis in two case studies. *Journal of Neuroscience Methods*, 158(1) :137 – 149.

[Fothergill et al., 2012] Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1737–1746, New York, NY, USA. ACM.

[Frijda, 2007] Frijda, N. H. (2007). What might emotions be ? comments on the comments. *Social Science Information*, 46(3) :433–443.

[Fumera and Roli, 2005] Fumera, G. and Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6) :942–956.

[Garber-Barron and Si, 2012] Garber-Barron, M. and Si, M. (2012). Using body movement and posture for emotion detection in non-acted scenarios. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8.

[Gavrila and Davis, 1995] Gavrila, D. M. and Davis, L. S. (1995). Towards 3-d model-based tracking and recognition of human movement : a multi-view approach. In *In International Workshop on Automatic Face- and Gesture-Recognition. IEEE Computer Society*, pages 272–277.

[Gedat et al., 2017] Gedat, E., Fechner, P., Fiebelkorn, R., and Vandenhouten, R. (2017). Human action recognition with hidden markov models and neural network derived poses. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000157–000162.

[Ghorbel, 2017] Ghorbel, E. (2017). *Fast and accurate human action recognition using RGB-D cameras*. PhD thesis, Normandie Université.

[Ghorbel et al., 2015] Ghorbel, E., Boutteau, R., Boonaert, J., Savatier, X., and Lecoeuche, S. (2015). 3d real-time human action recognition using a spline interpolation approach. In *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 61–66.

[Glowinski et al., 2011] Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., and Scherer, K. (2011). Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, 2(2) :106–118.

[Groff, 1995] Groff, E. (1995). Laban movement analysis : Charting the ineffable domain of human movement. *Journal of Physical Education, Recreation & Dance*, 66(2) :27–30.

[Guest, 2013] Guest, A. (2013). *Labanotation : The System of Analyzing and Recording Movement*. Taylor & Francis.

[Guo et al., 2009] Guo, K., Ishwar, P., and Konrad, J. (2009). Action recognition in video by covariance matching of silhouette tunnels. In *XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 299–306. IEEE.

[H. Aviezer, 2012] H. Aviezer, Y. Trope, A. T. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111) :1225–9.

[Hachimura et al., 2005] Hachimura, K., Takashina, K., and Yoshimura, M. (2005). Analysis and evaluation of dancing movement based on lma. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 294–299.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151.

[Hartmann et al., 2006] Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In *Proceedings of the 6th Inter-*

national Conference on Gesture in Human-Computer Interaction and Simulation, GW'05, pages 188–199, Berlin, Heidelberg. Springer-Verlag.

[Holte et al., 2008] Holte, M. B., Moeslund, T. B., and Fihl, P. (2008). View invariant gesture recognition using the csem swissranger sr-2 camera. *International Journal of Intelligent Systems Technologies and Applications*, 5(3-4) :295–303.

[Hoque et al., 2014] Hoque, N., Bhattacharyya, D., and Kalita, J. (2014). Mifs-nd : A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14) :6371 – 6385.

[Hsu et al., 2005] Hsu, E., Pulli, K., and Popović, J. (2005). Style translation for human motion. *ACM Trans. Graph.*, 24(3) :1082–1089.

[Hundleby, 1968] Hundleby, J. D. (1968). Reviews : Nunnally, jum. psychometric theory. new york : Mcgraw-hill, 1967. 640 + xiii pp. $12.95. *American Educational Research Journal*, 5(3) :431–433.

[Hussein et al., 2013a] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. (2013a). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, volume 13, pages 2466–2472.

[Hussein et al., 2013b] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. (2013b). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2466–2472. AAAI Press.

[Igorevich et al., 2013] Igorevich, R. R., Park, P., Choi, J., and Min, D. (2013). Two hand gesture recognition using stereo camera. *International Journal of Computer and Electrical Engineering*, 5(1) :69–72.

[Izard, 2007] Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 2 3 :260–80.

[Jaakkola and Haussler, 1998] Jaakkola, T. S. and Haussler, D. (1998). Exploiting generative models in discriminative classifiers. In *NIPS*.

[Jiang et al., 2014] Jiang, X., Zhong, F., Peng, Q., and Qin, X. (2014). Online robust action recognition based on a hierarchical model. *The Visual Computer*, 30(9) :1021–1033.

[Johansson, 1973] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2) :201–211.

[Junejo et al., 2014] Junejo, I. N., Junejo, K. N., and Al Aghbari, Z. (2014). Silhouette-based human action recognition using sax-shapes. *The Visual Computer*, 30(3) :259–269.

[Kapadia et al., 2013] Kapadia, M., Chiang, I.-k., Thomas, T., Badler, N. I., and Kider, Jr., J. T. (2013). Efficient motion retrieval in large motion databases. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '13, pages 19–28, New York, NY, USA. ACM.

[Kendon, 2004] Kendon, A. (2004). *Gesture : Visible action as utterance.* Cambridge University Press.

[Kim et al., 2012] Kim, J., Seo, J.-H., and Kwon, D.-S. (2012). Application of effort parameter to robot gesture motion. In *2012 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 80–82.

[Kim et al., 2014] Kim, M.-G., Barakova, E., and Lourens, T. (2014). Rapid prototyping framework for robot-assisted training of autistic children. In *Robot and Human Interactive Communication, 2014 RO-MAN : The 23rd IEEE International Symposium on*, pages 353–358. IEEE.

[Kim et al., 2010] Kim, W., Lee, J., Kim, M., Oh, D., and Kim, C. (2010). Human action recognition using ordinal measure of accumulated motion. *EURASIP Journal on Advances in Signal Processing*, 2010 :2.

[Kim et al., 2013] Kim, W. H., Park, J. W., Lee, W. H., Chung, M. J., and Lee, H. S. (2013). Lma based emotional motion representation using rgb-d camera. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 163–164.

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3) :226–239.

[Klaser et al., 2008] Klaser, A., Marszalek, M., and Schmid, C. (2008). A Spatio-Temporal Descriptor Based on 3D-Gradients. In Everingham, M., Needham, C., and Fraile, R., editors, *BMVC 2008 - 19th British Machine Vision Conference*, pages 275 :1–10, Leeds, United Kingdom. British Machine Vision Association.

[Knight and Simmons, 2014] Knight, H. and Simmons, R. (2014). Expressive motion with x, y and theta : Laban effort features for mobile robots. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 267–273.

[Knight et al., 2016] Knight, H., Thielstrom, R., and Simmons, R. (2016). Expressive path shape (swagger) : Simple features that illustrate a robot's attitude toward its goal in real time. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1475–1482.

[Kong et al., 2016] Kong, Y., Satarboroujeni, B., and Fu, Y. (2016). Learning hierarchical 3d kernel descriptors for rgb-d action recognition. *Computer Vision and Image Understanding*, 144 :14–23.

[Laban, 1994] Laban, R. (1994). *La Maitrise du mouvement.* Actes Sud Beaux Arts.

[Laban and Ullmann, 1971] Laban, R. and Ullmann, L. (1971). The mastery of movement.

[Laptev and Lindeberg, 2003] Laptev and Lindeberg (2003). Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 432–439 vol.1.

[Laptev et al., 2008a] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008a). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

[Laptev et al., 2008b] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008b). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

[Lee and Kim, 1999] Lee, H.-K. and Kim, J. (1999). An hmm-based threshold model approach for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10) :961 –973.

[Lee, 1988] Lee, K.-F. (1988). On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, 7(4) :375 – 379. Word Recognition in Large Vocabularies.

[Lehrmann et al., 2014] Lehrmann, A. M., Gehler, P. V., and Nowozin, S. (2014). Efficient nonlinear markov models for human motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321.

[Leman et al., 2001] Leman, M., CAMURRI, A., and DE POLI, G. (2001). Megase : a multisensory expressive gesture applications system environment for artistic performances. In *Proceedings of CAST01 : living in mixed reality - Conference on Communication of Art, Science and Technology,*.

[Levinson et al., 1983] Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4) :1035–1074.

[Li et al., 2016] Li, M., Leung, H., and Shum, H. P. (2016). Human action recognition via skeletal and depth based feature fusion. In *Proceedings of the 9th International Conference on Motion in Games*, pages 123–132. ACM.

[Li et al., 2010] Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14.

[Lorini and Schwarzentruber, 2011] Lorini, E. and Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3) :814.

[Lourens et al., 2010] Lourens, T., van Berkel, R., and Barakova, E. (2010). Communicating emotions and mental states to robots in a real time parallel framework using laban movement analysis. *Robotics and Autonomous Systems*, 58(12) :1256 – 1265. Intelligent Robotics and Neuroscience.

[Maletic, 1987] Maletic, V. (1987). *Body, space, expression : the development of Rudolf Laban's movement and dance concepts / Vera Maletic.* Mouton de Gruyter Berlin ; New York.

[Marill and Green, 1963] Marill, T. and Green, D. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1) :11–17.

[Masuda and Kato, 2010] Masuda, M. and Kato, S. (2010). Motion rendering system for emotion expression of human form robots based on laban movement analysis. In *19th International Symposium in Robot and Human Interactive Communication*, pages 324–329.

[Masuda et al., 2010] Masuda, M., Kato, S., and Itoh, H. (2010). *A Laban-Based Approach to Emotional Motion Rendering for Human-Robot Interaction*, pages 372–380. Springer Berlin Heidelberg, Berlin, Heidelberg.

[McNeill, 1992] McNeill, D. (1992). *Hand and mind : What gestures reveal about thought.* University of Chicago press.

[Mehrabian and T. Friar, 1969] Mehrabian, A. and T. Friar, J. (1969). Enconding of attitude by a seated communicator via posture and position cues. 33 :330–336.

[Miao and Makis, 2007] Miao, Q. and Makis, V. (2007). Condition monitoring and classification of rotating machinery using wavelets and hidden markov models. *Mechanical Systems and Signal Processing*, 21(2) :840 – 855.

[Murray et al., 1994] Murray, R. M., Sastry, S. S., and Zexiang, L. (1994). *A Mathematical Introduction to Robotic Manipulation.* CRC Press, Inc., Boca Raton, FL, USA, 1st edition.

[Negin et al., 2015] Negin, F., Akgül, C. B., Yüksel, K. A., and Erçil, A. (2015). An rdf-based action recognition framework with feature selection capability, considering therapy exercises utilizing depth cameras.

[Nelwamondo et al., 2005] Nelwamondo, F. V., Marwala, T., and Mahola, U. (2005). Early classifications of bearing faults using hidden markov models , gaussian mixture models , mel-frequency cepstral coefficients and fractals.

[Niedenthal et al., 2006] Niedenthal, P. M., Krauth-Gruber, S., and Ric, F. (2006). Psychology of emotion : Interpersonal, experimental and cognitive approaches.

[Nishimura et al., 2012] Nishimura, K., Kubota, N., and Woo, J. (2012). Design support system for emotional expression of robot partners using interactive evolutionary computation. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–7.

[Oreifej and Liu, 2013] Oreifej, O. and Liu, Z. (2013). Hon4d : Histogram of oriented 4d normals for activity recognition from depth sequences. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723.

[Pedersoli et al., 2014] Pedersoli, F., Benini, S., Adami, N., and Leonardi, R. (2014). Xkin : An open source framework for hand pose and gesture recognition using kinect. *Vis. Comput.*, 30(10) :1107–1122.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

[Pham et al., 2018] Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., and Velastin, S. A. (2018). Exploiting deep residual networks for human action recognition from skeletal data. *Computer Vision and Image Understanding*.

[Picard, 2003] Picard, R. W. (2003). Affective computing : challenges. *International Journal of Human-Computer Studies*, 59(1) :55 – 64. Applications of Affective Computing in Human-Computer Interaction.

[Poropat, 2014] Poropat, A. E. (2014). Other-rated personality and academic performance : Evidence and implications. *Learning and Individual Differences*, 34 :24 – 32.

[Procter et al., 2000] Procter, S., Illingworth, J., and Mokhtarian, F. (2000). Cursive handwriting recognition using hidden markov models and a lexicon-driven level building algorithm. *IEE Proceedings - Vision, Image and Signal Processing*, 147(4) :332–339.

[Quinlan, 1992] Quinlan, J. R. (1992). Learning with continuous classes. pages 343–348. World Scientific.

[Rabiner and Juang, 1986] Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1) :4–16.

[Rasmussen and Nickisch, 2010] Rasmussen, C. E. and Nickisch, H. (2010). Gaussian processes for machine learning (gpml) toolbox. *Journal of machine learning research*, 11(Nov) :3011–3015.

[Rigoll et al., 1996] Rigoll, G., Kosmala, A., Rattland, J., and Neukirchen, C. (1996). A comparison between continuous and discrete density hidden markov models for cursive handwriting recogni-

tion. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 2, pages 205–209 vol.2.

[Roetenberg et al., 2013] Roetenberg, D., Luinge, H., and Slycke, P. J. (2013). Xsens mvn : Full 6 dof human motion tracking using miniature inertial sensors.

[Rokach, 2010] Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

[Russell, 1980] Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6) :1161–1178.

[Russell, 1994] Russell, J. A. (1994). Is there universal recognition of emotion from facial expression ? a review of the cross-cultural studies. *Psychological Bulletin*, 115 :102–141.

[Samadani et al., 2013] Samadani, A. A., Burton, S., Gorbet, R., and Kulic, D. (2013). Laban effort and shape analysis of affective hand and arm movements. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 343–348.

[Scherer, 2003] Scherer, K. R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech communication*, 40(1-2) :227–256.

[Scherer, 2005a] Scherer, K. R. (2005a). What are emotions ? and how can they be measured ? *Social Science Information*, 44(4) :695–729.

[Scherer, 2005b] Scherer, K. R. (2005b). What are emotions ? and how can they be measured ? *Social Science Information*, 44(4) :695–729.

[Scholkopf and Smola, 2001] Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

[Schuldt et al., 2004] Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions : a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE.

[Schuller et al., 2006] Schuller, B., Reiter, S., and Rigoll, G. (2006). Evolutionary feature generation in speech emotion recognition. In *2006 IEEE International Conference on Multimedia and Expo*, pages 5–8.

[Sethu et al., 2007] Sethu, V., Ambikairajah, E., and Epps, J. (2007). Speaker normalisation for speech-based emotion detection. In *Digital Signal Processing, 2007 15th International Conference on*, pages 611–614. IEEE.

[Shao and Chen, 2010] Shao, L. and Chen, X. (2010). Histogram of body poses and spectral regression discriminant analysis for human action categorization. In *BMVC*, pages 1–11.

[Sharma et al., 2013] Sharma, M., Hildebrandt, D., Newman, G., Young, J. E., and Eskicioglu, R. (2013). Communicating affect via flight path exploring use of the laban effort system for designing affective locomotion paths. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 293–300.

[Sharma and Verma, 2015] Sharma, R. P. and Verma, G. K. (2015). Human computer interaction using hand gesture. *Procedia Computer Science*, 54 :721–727.

[Sobol-Shikler and Robinson, 2010] Sobol-Shikler, T. and Robinson, P. (2010). Classification of complex information : Inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7) :1284–1297.

[Soh and Demiris, 2012] Soh, H. and Demiris, Y. (2012). Iterative temporal learning and prediction with the sparse online echo state gaussian process. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Song and Takatsuka, 2005] Song, L. and Takatsuka, M. (2005). Real-time 3d finger pointing for an augmented desk. In *Proceedings of the Sixth Australasian Conference on User Interface - Volume 40*, AUIC '05, pages 99–108, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

[Song et al., 2013] Song, Y., Morency, L. P., and Davis, R. (2013). Distribution-sensitive learning for imbalanced datasets. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

[Song et al., 2014] Song, Y., Tang, J., Liu, F., and Yan, S. (2014). Body surface context : A new robust feature for action recognition from depth videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6) :952–964.

[Suzuki et al., 2000] Suzuki, R., Iwadate, Y., Inoue, M., and Woo, W. (2000). Midas : Mic interactive dance system. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 2, pages 751–756 vol.2.

[Tavakol and Dennick, 2011] Tavakol, M. and Dennick, R. (2011). Making sense of cronbach's alpha. *Int J Med Educ*, 2 :53–55.

[Tian et al., 2012] Tian, Y., Cao, L., Liu, Z., and Zhang, Z. (2012). Hierarchical filtered motion for action recognition in crowded videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3) :313–323.

[Torresani et al., 2006] Torresani, L., Hackney, P., and Bregler, C. (2006). Learning motion style synthesis from perceptual observations. In *Proceedings of the 19th International Conference on*

Neural Information Processing Systems, NIPS'06, pages 1393–1400, Cambridge, MA, USA. MIT Press.

[Truong et al., 2016] Truong, A., Boujut, H., and Zaharia, T. (2016). Laban descriptors for gesture recognition and emotional analysis. *The Visual Computer*, 32(1) :83–98.

[Truong and Zaharia, 2016] Truong, A. and Zaharia, T. (2016). Dynamic gesture recognition with laban movement analysis and hidden markov models. In *Proceedings of the 33rd Computer Graphics International*, CGI '16, pages 21–24, New York, NY, USA. ACM.

[Tseng et al., 2012] Tseng, C.-C., Chen, J.-C., Fang, C.-H., and Lien, J.-J. J. (2012). Human action recognition based on graph-embedded spatio-temporal subspace. *Pattern Recognition*, 45(10) :3611–3624.

[TUMER and GHOSH, 1996] TUMER, K. and GHOSH, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4) :385–404.

[Van Deemter et al., 2008] Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., and Baumann, S. (2008). Fully generated scripted dialogue for embodied agents. *Artificial Intelligence*, 172(10) :1219–1244.

[Vemulapalli et al., 2014] Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

[Ververidis and Kotropoulos, 2006] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition : Resources, features, and methods. *Speech communication*, 48(9) :1162–1181.

[Vlasenko et al., 2007] Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007). Frame vs. turn-level : emotion recognition from speech considering static and dynamic processing. In *International Conference on Affective Computing and Intelligent Interaction*, pages 139–147. Springer.

[Vogt and André, 2006] Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa.*

[Wang et al., 2015] Wang, W., Enescu, V., and Sahli, H. (2015). Adaptive real-time emotion recognition from body movements. *ACM Trans. Interact. Intell. Syst.*, 5(4) :18 :1–18 :21.

[Weinland and Boyer, 2008] Weinland, D. and Boyer, E. (2008). Action recognition using exemplar-based embedding. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7.

[Weinland et al., 2006] Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.*, 104(2) :249–257.

[Whitney, 1971] Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, C-20(9) :1100–1103.

[Willems et al., 2008] Willems, G., Tuytelaars, T., and Van Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In Forsyth, D., Torr, P., and Zisserman, A., editors, *Computer Vision – ECCV 2008*, pages 650–663, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edition.

[Wundt, 1973] Wundt, W. (1973). *The language of gestures*. Approaches to semiotics : Paperback series. Mouton.

[Xia et al., 2012a] Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012a). View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 20–27. IEEE.

[Xia et al., 2012b] Xia, L., Chen, C. C., and Aggarwal, J. K. (2012b). View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27.

[Xia et al., 2015] Xia, S., Wang, C., Chai, J., and Hodgins, J. (2015). Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4) :119.

[Yan and Luo, 2012] Yan, X. and Luo, Y. (2012). Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier. *Neurocomputing*, 87 :51–61.

[Yang and Tian, 2014] Yang, X. and Tian, Y. (2014). Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1) :2 – 11. Visual Understanding and Applications with RGB-D Cameras.

[Yang and Tian, 2012] Yang, X. and Tian, Y. L. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 14–19. IEEE.

[Yang et al., 2012] Yang, X., Zhang, C., and Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060. ACM.

[Yoon and Park, 2007] Yoon, W.-J. and Park, K.-S. (2007). A study of emotion recognition and its applications. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 455–462. Springer.

[Yu and Liu, 2003] Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data : A fast correlation-based filter solution. In Fawcett, T. and Mishra, N., editors, *Proceedings, Twentieth International Conference on Machine Learning*, volume 2, pages 856–863.

[Yumer and Mitra, 2016] Yumer, M. E. and Mitra, N. J. (2016). Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)*, 35(4) :137.

[Zacharatos et al., 2013] Zacharatos, H., Gatzoulis, C., Chrysanthou, Y., and Aristidou, A. (2013). Emotion recognition for exergames using laban movement analysis. In *Proceedings of Motion on Games*, MIG '13, pages 39 :61–39 :66, New York, NY, USA. ACM.

[Zanfir et al., 2013] Zanfir, M., Leordeanu, M., and Sminchisescu, C. (2013). The moving pose : An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *2013 IEEE International Conference on Computer Vision*, pages 2752–2759.

[Zhang et al., 2010] Zhang, G.-Y., Zhang, C.-X., and Zhang, J.-S. (2010). Out-of-bag estimation of the optimal hyperparameter in subbag ensemble method. *Communications in Statistics - Simulation and Computation*, 39(10) :1877–1892.

[Zhao and Badler, 2005] Zhao, L. and Badler, N. I. (2005). Acquiring and validating motion qualities from live limb gestures. *Graph. Models*, 67(1) :1–16.

[Zhao et al., 2013] Zhao, X., Li, X., Pang, C., Zhu, X., and Sheng, Q. Z. (2013). Online human gesture recognition from motion data streams. In *ACM Multimedia*.

Annexe A

Les modèles de Markov cachés

A.1 Définition d'un MMC

Les MMC sont des outils permettant la modélisation des phénomènes stochastiques. Ils sont des modèles puissants pour la modélisation de données séquentielles ou de séries temporelles, et ont été utilisés avec succès dans de nombreuses applications en reconnaissance de la parole, en biologie, en compression des données et dans d'autres domaines de l'intelligence artificielle et de la reconnaissance de formes. Un MMC est un modèle statistique basé sur un modèle de Markov, composé d'un ensemble d'états qui transitent entre-deux. Ces états sont cachés et chaque état émet des "observations" qui sont observables. Une structure temporelle est encodée à l'intérieur du processus caché, où chaque état est conditionné par le précédent. Chaque état a une distribution de probabilité sur les observations possibles. Par conséquent, la séquence d'observations générée par un MMC donne des informations sur la séquence d'états.

A.1.1 Les hypothèses dans la théorie des MMC

1. **L'hypothèse de Markov :**

   Le processus de Markov est un processus stochastique qui satisfait à la condition de Markov dans laquelle son comportement futur ne dépend que de son état actuel et non pas du passé. Il est également appelé un système sans mémoire. Soit un processus de Markov à N états discrets, $s_1, s_2, \ldots, s_N$, et soit q_t l'état du processus à l'instant t. La probabilité que le processus soit dans l'état s_i à l'instant t est noté par $P(q_t = s_i)$. La condition de Markov implique une hypothèse d'indépendance d'état. En termes simples, l'état actuel dépend uniquement de l'état précédent. Il peut être formellement énoncé comme suit :

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_{i1}, \ldots, q_1 = s_{ik}) = P(q_t = s_j | q_{t-1} = s_i) \quad \forall i, i_1, \ldots, i_k, j, t \tag{A.1}$$

1.1. **L'hypothèse de stationnarité :** Cette hypothèse suppose que les probabilités de transition d'état sont indépendantes de l'instant actuel auquel les transitions ont lieu. Mathématiquement :

$$P(q_{t_1+1} = sj | q_{t_1} = s_i) = P(q_{t_2+1} = s_j | q_{t_2} = s_i) \ \forall \ t_1 \ et \ t_2 \tag{A.2}$$

1.2. **L'hypothèse d'indépendance de sortie :**

C'est l'hypothèse que la sortie actuelle (observation) est statistiquement indépendante des sorties précédentes. Nous pouvons formuler cette hypothèse mathématiquement, en considérant une suite d'observations, $O = (o_1, o_2, \ldots, o_T)$. Suivant cette hypothèse :

$$P(O | q_1, q_2, \ldots, q_T, \sigma) = \Pi_{t=1}^{T} P(o_t | q_t, \sigma) \tag{A.3}$$

A.1.2 Les trois problèmes fondamentaux d'un MMC

Pour que le modèle soit utile dans des applications réelles, il y a trois problèmes d'intérêt qui doivent être résolus. Ces problèmes sont les suivants :

— **Problème d'évaluation :**

Étant donné un MMC σ et une séquence d'observations $O = \{o_1, o_2, \ldots, o_T\}$, quelle est la probabilité que les observations soient générées par le modèle σ, $P(O|\sigma)$?

— **Problème de décodage :**

Étant donné un modèle σ et une séquence d'observations $O = \{o_1, o_2, \ldots, o_T\}$, quelle est la séquence d'état la plus probable dans le modèle qui a produit les observations ?

— **Problème d'apprentissage**

Étant donné un modèle σ et une séquence d'observations $O = \{o_1, o_2, \ldots, o_T\}$, comment devrions-nous ajuster les paramètres de ce modèle (A, B, π) afin de maximiser $P(O|\sigma)$?

Le premier problème peut être résolu avec les algorithmes itératifs forward et backward, le deuxième problème avec l'algorithme de Viterbi, également un algorithme itératif pour développer le meilleur chemin en considérant séquentiellement chaque symbole observé. Le problème d'apprentissage se

résout avec l'algorithme de Baum-Welch, un dérivé de l'algorithme espérance-maximisation, qui utilise les probabilités de forward et backward pour mettre à jour les paramètres d'une manière itérative. Ci-dessous, nous donnons les formules utilisées pour chaque étape de calcul. Pour notre application, nous avons traité que les deux problèmes d'évaluation et d'apprentissage. Nous avons utilisé la solution au problème d'évaluation pour la phase de classification, et celle au problème d'apprentissage pour la phase d'entrainement.

Problème d'évaluation : Algorithme de Forward

Le problème d'évaluation du modèle σ consiste à calculer la probabilité de la séquence d'observation $O = \{o_1, o_2, \ldots, o_T\}$ générée par ce modèle, c'est à dire $P(O|\sigma)$. Nous pouvons calculer cette quantité en utilisant des arguments probabilistes simples. Supposons que O est généré par la séquence d'états $Q = \{q_1, q_2, \ldots, q_T\}$. Nous calculons $P(O|\sigma)$ en sommant la loi conjointe $P(O, Q|\sigma)$ sur toutes les états possibles Q.

$$P(O|\sigma) = \sum_{Q} P(O, Q|\sigma) \tag{A.4}$$

La loi conjointe $P(O, Q|\sigma)$ qui est la probabilité que O et Q se produisent en même temps se décompose par la règle de Bayes comme suit :

$$P(O, Q|\sigma) = P(O|Q, \sigma)P(Q|\sigma) \tag{A.5}$$

L'hypothèse d'indépendance de l'État (Équation A.1) donne la formule suivante :

$$P(Q|\sigma) = \pi_{q_1} \Pi_{t=1}^{T-1} a_{q_t q_{t+1}} = \pi_{q_1}.a_{q_1 q_2} \ldots a_{q_{T-1} q_T} \tag{A.6}$$

En outre, l'hypothèse d'indépendance de sortie (Équation A.3) donne la formule suivante :

$$P(O|Q, \sigma) = \Pi_{t=1}^{T} P(o_t|q_t, \sigma) = b_{q_1}(o_1).b_{q_2}(o_2) \ldots b_{q_T}(o_T) \tag{A.7}$$

Par conséquent,

$$P(O|\sigma) = \sum_{Q} \pi_{q_1} b_{q_1}(O_1) \cdot a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{A.8}$$

Le calcul direct de $P(O|\sigma)$ dans l'Équation A.8 est un peu compliqué et implique des calculs de l'ordre de $2\ TN^T$. Ce calcul devient irréalisable si le nombre d'états possibles, N, ou la longueur de la séquence d'observation T augmente. Par exemple, pour $T = 100$ et $N = 5$ états, il y a 200×5^{100}

opérations requises. Donc, nous avons besoin d'un moyen plus efficace pour calculer $P(O|\sigma)$. Baum et al. [Baum et al., 1970] ont proposé une solution avec un algorithme efficace appelé l'algorithme de forward-backward. Nous définissons d'abord la variable forward :

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = s_i|\sigma), \;\; 1 \leq i \leq N \tag{A.9}$$

La variable $\alpha_t(i)$ correspond à la probabilité d'observer la séquence partielle $O = (o_1, o_2, \ldots, o_t)$ (jusqu'au temps t) et d'être à l'état s_i à l'instant t.

— Initialisation :

$$\alpha_1(i) = \pi_i b_i(o_1) \;\; 1 \leq i \leq N \tag{A.10}$$

— Récursivité (Voir Figure A.1) :

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \, b_j(o_{t+1}), \;\; 1 \leq t \leq T-1 \tag{A.11}$$

— Terminaison :

$$P(O|\sigma) = \sum_{i=1}^{N} \alpha_T(i) \tag{A.12}$$

Les équations A.9, A.11, et A.16 indiquent comment calculer $P(O|\sigma)$ en évaluant d'abord récursivement les variables forward,$\alpha_t(i)$, de $t = 1$ à $t = T$, puis en sommant toutes les variables en avant à l'instant T.

Pour T observations et N états, le nombre de calcul impliqué devient de l'ordre TN^2 au lieu de $2TN^T$. Par conséquent, l'algorithme forward est plus efficace pour la résolution du problème d'évaluation.

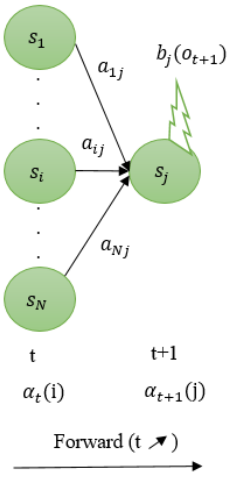


FIGURE A.1 – Algorithme de Forward.

De manière analogue, soit la variable backward :

$$\beta_t(i) = P(o_{t+1}, \ldots, o_T | q_t = s_i, \sigma) \tag{A.13}$$

La variable $\beta_t(i)$ correspond à la probabilité d'observer la séquence partielle ultérieure $O = (o_{t+1}, o_{t+2}, \ldots, o_T)$ (jusqu'au temps T) en partant de l'état s_i à l'instant t. $\beta_t(i)$ peut également être calculé d'une manière récursive :

— Initialisation :

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \tag{A.14}$$

— Récursivité (Voir Figure A.2) :

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}), \;\; 1 \leq i \leq N, \;\; 1 \leq t \leq T-1 \tag{A.15}$$

— Terminaison :

$$P(O|\sigma) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \tag{A.16}$$

De la même manière que l'algorithme forward, l'algorithme backward implique un calcul de l'ordre de TN^2 au lieu de $2TN^T$.

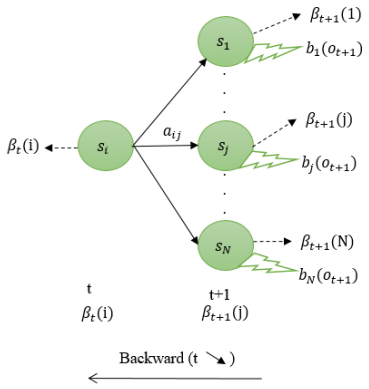


FIGURE A.2 – Algorithme de backward.

L'algorithme de forward et de backward seront combinés par la suite pour résoudre le problème de l'apprentissage avec l'algorithme de Baum-Welch.

Algorithme de Baum-Welch

Etant donné un modèle initial σ et une séquence d'observation $O = (o_1, o_2, \ldots, o_T)$, l'algorithme de Baum-Welch [Baum et al., 1970] consiste à trouver un modèle $\bar{\sigma}$ qui explique mieux la séquence

O.

$$\bar{\sigma} = argmax_{\sigma} P(O|\sigma) \tag{A.17}$$

L'algorithme de Baum-Welch permet de réestimer itérativement les différents paramètres du modèle σ. C'est un cas particulier d'une généralisation de l'algorithme espérance-maximisation (EM), qui permet de trouver des estimations de maximum de vraisemblance et des estimations pour les paramètres (probabilités de transition et d'émission) d'un MMC, étant donné seulement les données d'entraînement d'observation. L'algorithme d'espérance-maximisation comporte une étape d'évaluation de l'espérance (E), qui calcule l'espérance de la vraisemblance et une étape de maximisation (M), qui estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape (E). Les paramètres trouvés dans l'étape (M) sont ensuite utilisés pour commencer une autre étape (E), ainsi de suite. Dans l'algorithme de Baum-welch, outre que les variables «forward» et «backward», deux autres variables sont introduites, la première est $\epsilon_t(i,j) = P(q_t = s_i, q_{t+1} = s_j|O,\sigma)$ qui présente la probabilité de passer de l'état s_i à l'état s_j à l'instant t quand σ génère une série d'observations $O = (o_1, o_2, ..., o_T)$ et qui peut être aussi écrit comme suit :

$$\epsilon_t(i,j) = \frac{P(q_t = s_i, q_{t+1} = s_j, O|\sigma)}{P(O|\sigma)} \tag{A.18}$$

Cette variable peut être exprimée en fonction des variables forward et backward :

$$\epsilon_t(i,j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(O_{t+1})}{P(O|\sigma)} \tag{A.19}$$

$$\epsilon_t(i,j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(O_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(O_{t+1})} \tag{A.20}$$

La deuxième variable est $\gamma_t(i) = P(q_t = s_i|O,\sigma)$ qui présente la probabilité d'être dans l'état s_i à l'instant t, étant donné la séquence O et le modèle σ. Avec les variables forward et backward, cela peut être exprimé par :

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{A.21}$$

La relation entre $\gamma_t(i)$ et $\epsilon_t(i,j)$ est exprimée donc par la formule suivante :

$$\gamma_t(i) = \sum_{j=1}^{N} \epsilon_t(i,j), \quad 1 \leq i \leq N \tag{A.22}$$

Maintenant, il est possible de décrire le processus d'apprentissage de Baum-Welch, où les paramètres du MMC sont mis à jour de manière à maximiser la quantité $P(O|\sigma)$. Soit un modèle initial $\sigma = (A, B, \pi)$, les variables α et β sont définies respectivement par les équations de récursivité A.11 et A.15, et les variables ϵ et γ par les équations A.20 et A.22. L'étape suivante consiste à mettre à jour les paramètres MMC suivant les équations A.23, A.24, A.25 appelées formules de ré-estimation :

$$\bar{\pi_i} = \gamma_1(i), \;\; 1 \leq i \leq N \tag{A.23}$$

$$\bar{a_{ij}} = \frac{\sum_{t=1}^{T-1} \epsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \;\; 1 \leq i \leq N, \;\; 1 \leq j \leq N \tag{A.24}$$

$$\bar{b_j}(k) = \frac{\sum_{t=1, o_t = v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \;\; 1 \leq j \leq N, \;\; 1 \leq k \leq M \tag{A.25}$$

$\bar{a_{i,j}}$ peut être considéré comme le rapport entre le nombre de transitions attendues de l'état s_i à s_j et le nombre de transitions attendues de l'état s_i. De même, $b_j\bar{(k)}$ correspond au rapport entre le nombre de fois où le MMC s'est trouvé dans l'état s_j en observant v_k et le nombre de fois où le MMC s'est trouvé dans l'état s_j. $\bar{\pi_i}$ correspond au nombre de fois que le MMC s'est trouvé dans l'état s_i à l'instant $t = 1$. Finalement, nous pouvons résumer les étapes de l'algorithme Baum-Welch avec le diagramme présenté dans la Figure A.3 :

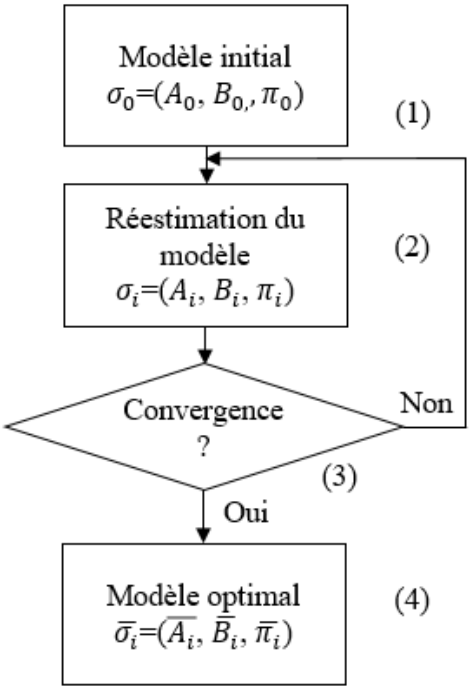


FIGURE A.3 – L'algorithme de Baum-Welch.

— La première étape (1) consiste à choisir un ensemble initial de paramètres σ_0.

— La deuxième étape est la réestimation itérative des paramètres du MMC, tout en vérifiant pour chaque itération i si

$$P(O|\sigma_{i+1}) \geq P(O|\sigma_i) \tag{A.26}$$

Le critère d'arrêt de cet algorithme est si :

$$P(O|\sigma_{i+1}) \approx P(O|\sigma_i) \tag{A.27}$$

— On passe alors à l'étape 3, après la convergence du système (Equation A.27 est validé). On confirme que σ_{i+1} est un modèle optimal permet de maximiser $P(O|\sigma)$.

— Finalement (étape 4), $\bar{\sigma} = \text{argmax}_\sigma\, P(O|\sigma)$.

Annexe B

Les machines à vecteurs de support

Les machines à vecteurs de support ont été introduites à la fin des années 70 par Vapnik, et initialement formulées comme des méthodes supervisées adoptées pour leur capacité à travailler avec des données de grandes dimension. Étant donné un ensemble d'entrainement $D_l = (x_i, y_i)$, $i = 1, \ldots, l$ qui constitue un ensemble de données linéairement séparables, où x_i représente une observation et y_i la décision associée appartient à $\{-1, +1\}$. Le but des SVM est de construire un hyperplan qui sépare le mieux possible les deux classes D_+ et D_- qui correspondent respectivement aux points x_i tels que $y_i = +1$ et $y_i = -1$. Cette méthode recherche simplement l'hyperplan séparateur avec la plus grande marge. Supposons que toutes les données d'apprentissage satisfont aux contraintes suivantes :

$$w.x_i + b \geq +1 \quad pour \quad y_i = +1 \tag{B.1}$$

$$w.x_i + b \leq -1 \quad pour \quad y_i = -1 \tag{B.2}$$

Les équations B.1 et B.2 sont combinées dans l'ensemble d'inégalités suivant :

$$y_i(w.x_i + b) - 1 \geq 0 \quad \forall i \tag{B.3}$$

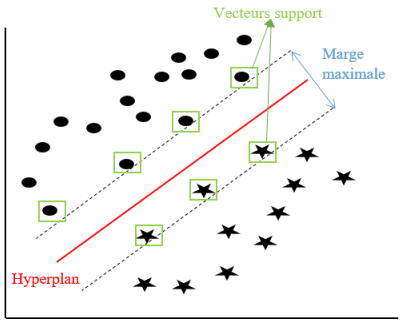


FIGURE B.1 – Hyperplan séparateur de la méthode SVM.

Suivant les deux équations B.1 et B.2, la marge est simplement égale à $2/\|w\|$. Comme l'objectif est de maximiser cette marge, cela revient alors à minimiser $\|w\|$. L'hyperplan à marge maximale est la solution du problème primal d'optimisation suivant portant sur les paramètres w et b.

$$\begin{cases} \min \frac{1}{2}\left\|w^2\right\| \quad w \in R^d, b \in R \\ \textit{Sous les contraintes} \quad y_i(w.x_i + b) \geq 1 \quad \forall \quad i = 1, \ldots, l. \end{cases} \tag{B.4}$$

Ce problème peut être résolu en utilisant les multiplicateurs de Lagrange $(\alpha_i)_{1 \leq i \leq l}$ associés aux contraintes du problème B.4. Le coefficient 1/2 qui apparaît ici est rajouté pour simplifier les calculs de dérivée qui vont suivre. Le dual lagrangien de ce problème devient :

$$L_p = \frac{1}{2}\left\|w^2\right\| - \sum_{i=1}^{l} \alpha_i [y_i(w.x_i + b) - 1] \tag{B.5}$$

Le lagrangien doit être minimisé par rapport aux variables dites primales w et b et maximisé par rapport aux multiplicateurs α_i : ce sont les conditions de Karush-Kuhn-Tucker (KKT) [Scholkopf and Smola, 2001].

$$\frac{\partial L}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{i=1}^{l} \alpha_i x_i y_i \tag{B.6}$$

$$\frac{\partial L}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{B.7}$$

Par substitution de B.6 et B.7 dans l'équation du lagrangien B.5 on obtient le problème dual :

$$\begin{cases} \max\limits_{\alpha_i} \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i.x_j \\ \textit{Sous les contraintes}, \quad \alpha_i \geq 0, \forall i \quad \textit{et} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \tag{B.8}$$

La solution du problème dual donne les multiplicateurs de Lagrange optimaux α_i^*. En pratique, seuls les points x_i qui sont sur les hyperplans frontière $|x_i.w + b| = 1$ interviennent dans la fonction de décision, car les α_i sont non nuls seulement pour ces points. Ainsi, l'hyperplan optimal ne dépend que des k vecteurs support ($k \leq l$). Une fois que nous avons trouvé les multiplicateurs de Lagrange optimaux α_i^*, nous obtenons le w^* du séparateur optimal avec la marge maximale :

$$w^* = \sum_{i=1}^{k} \alpha_i^* y_i x_i \tag{B.9}$$

Le paramètre b peut être déterminé en utilisant n'importe quel vecteur support $(x_i, y_i)_{i \in k}$ dans l'équation $|x_i.w^* + b^*| = 1$. La fonction de décision permettant de classer une nouvelle observation x est définie par le signe de :

$$f(x) = \sum_{i=1}^{k} \alpha_i^*.y_i x_i.x + b^* \tag{B.10}$$

En pratique, il est souvent préférable de tolérer certaines erreurs, au bénéfice d'une marge plus grande car ces erreurs peuvent être dues à des outliers (des observations aberrantes) de la classe qui leur est associée. Nous parlons alors de classificateur à marge souple [Scholkopf and Smola, 2001]. Un premier remède consiste à rendre les contraintes de l'équation B.4 moins rigides en introduisant des variables d'écart positives $\epsilon_i \geq 0$ pour que les contraintes deviennent :

$$y_i(w.x_i + b) \geq 1 - \epsilon_i, \quad \forall i$$

Ainsi l'hyperplan optimal peut être considéré comme la solution du problème d'optimisation convexe suivant :

$$\begin{cases} \min \frac{1}{2} \left\| w^2 \right\| + C\Sigma_{i=1}^{l} \epsilon_i \quad w \in R^d, b \in R \\ Sous \quad les \quad contraintes \quad y_i(w.x_i + b) \geq 1 - \epsilon_i \quad \forall \quad i = 1, \ldots, l. \end{cases} \tag{B.11}$$

Le terme $C\Sigma_{i=1}^{l} \epsilon_i$ représente une mesure de la quantité mal classée. C est un paramètre de pénalité, permettant de contrôler le compromis entre le fait de maximiser la marge et minimiser les erreurs de classification commises sur l'ensemble d'apprentissage. Plus il est grand plus une pénalité attribuée aux erreurs est élevée.

La méthode SVM est aussi utilisée dans le cas où les ensembles d'entraînement sont non linéairement séparables. Une fonction noyau $K(x_i, x_j)$ est appliquée afin de projeter l'espace d'entrée non linéaire dans un espace de dimension plus élevée (Figure B.2). Il s'agit d'une fonction continue, symétrique, semi-définie positive basée sur une transformation non linéaire de l'espace d'entrée X en un espace de re-description $\Phi(X)$.

$$K(x_i, x_j) = < \Phi(x_i), \Phi(x_j) > \tag{B.12}$$

x_i et x_j sont des vecteurs dans l'espace d'entrée, c'est-à-dire des vecteurs de caractéristiques calculées à partir des échantillons d'apprentissage.

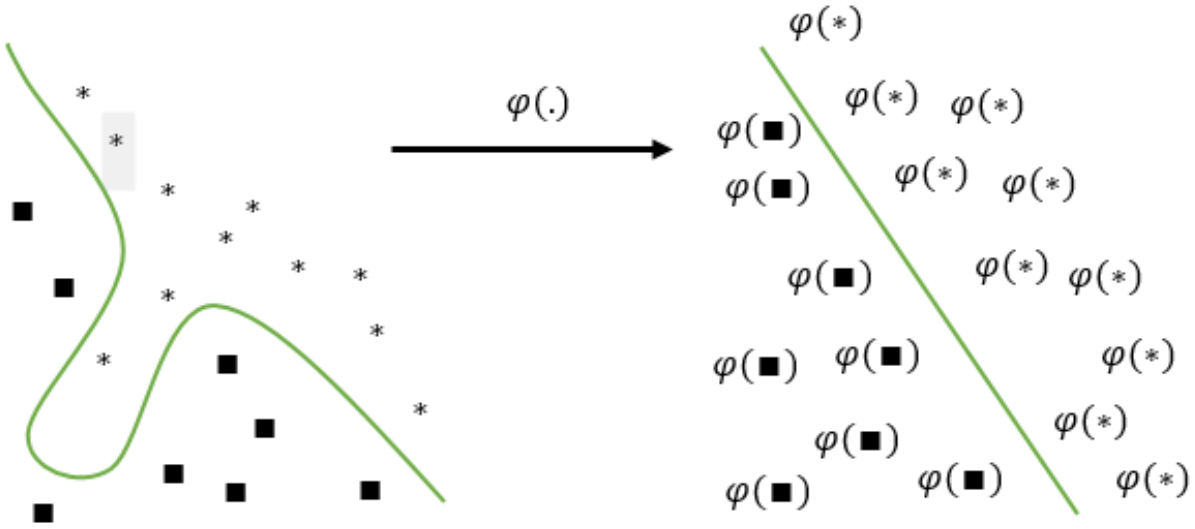


FIGURE B.2 – Transformation de l'espace d'entrée en un espace de re-description.

Les noyaux les plus fréquemment rencontrés sont :

— Linéaire : $K(x_i, x_j) =< x_i, x_j >$

— RBF : $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$

— Polynomial : $K(x_i, x_j) = \gamma < x_i, x_j >^d$

— Sigmoïdal : $K(x_i, x_j) = \tanh(\gamma < x_i, x_j >)$

d est le degré de polynôme. Le paramètre γ définit la portée de l'influence d'un seul exemple d'entraînement. Pour la fonction RBF, un γ faible signifie une gaussienne avec une grande variance, donc l'influence de x_j est plus grande, c'est-à-dire que si x_j est un vecteur support, une valeur faible de γ implique que la classe de ce vecteur support aura une influence sur la décision de la classe du vecteur x_i même si la distance entre eux est grande. Si la valeur de γ est élevée, alors la variance est faible, ce qui implique que le vecteur support n'a pas une influence répandue. Donc, un γ élevé conduit à des modèles à biais élevé et à faible variance, et inversement. Le paramètre C échange la mauvaise classification des exemples d'entraînement contre la simplicité de la surface de décision. Un C faible rend la surface de décision lisse, alors qu'un C élevé vise à classer correctement tous les exemples d'entraînement en donnant au modèle la liberté de sélectionner plus d'échantillons comme vecteurs de support. D'un autre coté Guermeur et al.[1] ont traité le problème de classification multi classes en proposant des extensions de SVM, les plus utilisées sont les suivantes :

— Le SVM Un-Contre-Tous forme k SVM binaires où k représente le nombre des classes. Le

1. SVM Multiclasses, Théorie et Applications. Y. Guermeur (2007). HDR, Université Nancy 1).

ième SVM est entraîné avec tous les échantillons appartenant à la ième classe comme des échantillons positifs, et considère les autres exemples comme des échantillons négatifs. Ainsi k fonctions de décision sont générées (Figure B.3). Pour classifier un exemple de test, la décision s'obtient avec le principe de "winner-takes-all". Cette approche présente donc k classifieurs, et la décision correspond au classifieur ayant renvoyé la valeur la plus élevée. La classe de l'entrée $x = \text{argmax}_{i=1,\ldots,k}\, w_i.x + b$

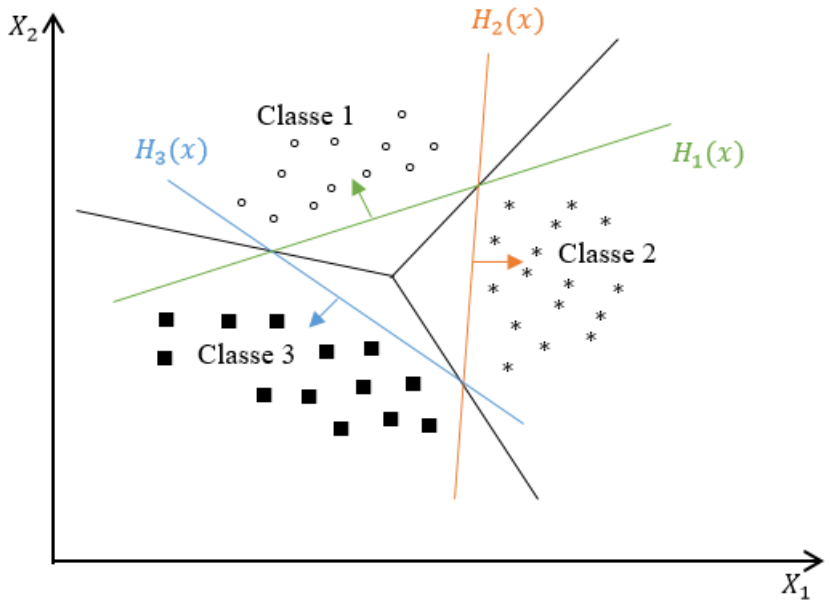


FIGURE B.3 – Classification multi-classes par la méthode Un-Contre-Tous.

— Le SVM Un-Contre-Un entraine un SVM pour chaque paire de classes i et j. Comme le montre la Figure B.4, le rôle du classifieur d'indice (i, j), avec $(1 \leq i < j \leq k)$, est de caractériser la catégorie d'indice i de celle d'indice j. Par conséquent, pour K classes cette méthode entraîne $k(k-1)/2$ SVM binaires. Pour classifier un exemple de test, la décision s'obtient avec le principe de "max-wins voting". Si signe $(w_{ij}\Delta x + b_{ij})$ montre que x appartient à la ième classe, alors le vote pour la ième classe est incrémenté. Sinon, celui de la jème classe est incrémenté. Finalement, la décision correspond à la classe qui a reçu le plus grand nombre de votes.

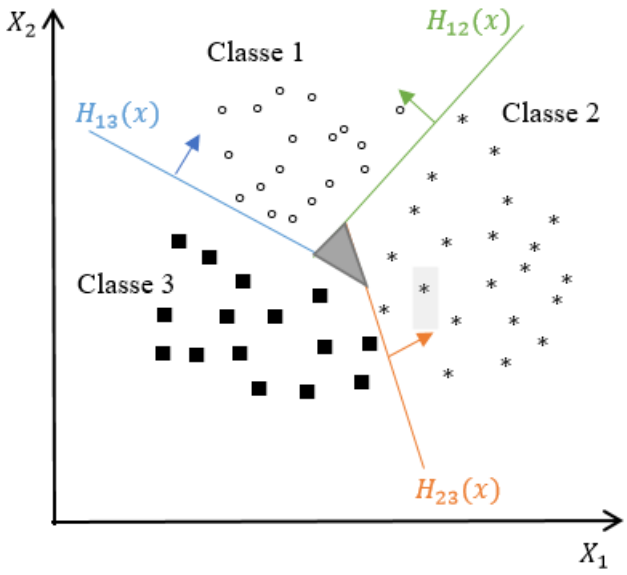


FIGURE B.4 – Classification multi-classes par la méthode Un-Contre-Un.

Annexe C

Perceptron multicouches

Le perceptron multicouche (MLP) est le réseau le plus ancien et le plus utilisé des réseaux de neurones artificiels. Un MLP est composé des éléments simples appelés neurones. Il est constitué d'une couche d'entrée, de plusieurs couches cachées et d'une couche de sortie. Au sein du réseau, les neurones sont structurés par couches. Il n'y a pas de connexion entre les neurones d'une même couche. Chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. On appelle couche d'entrée l'ensemble des neurones d'entrée qui reçoivent les informations de l'extérieur du réseau et renvoient les résultats intermédiaires aux neurones de la couche suivante. La couche de sortie représente l'ensemble des neurones de sortie qui calcule les résultats finaux et fournit la réponse du classifieur. Les couches intermédiaires n'ont pas de contact avec l'extérieur et sont donc nommées couches cachées. Les neurones de la couche d'entrée sont associés avec des poids synaptiques (w) qui représentent la force de connexion avec un neurone amont, et des biais (b). La couche de sortie peut avoir un neurone (pour obtenir une fonction discriminante à deux classes) ou plusieurs neurones (correspondant chacun à une classe). Le signal d'entrée, simplement propagé à travers les neurones de la couche d'entrée, est utilisé pour stimuler les couches suivantes de neurones cachées et de sortie. Le modèle mathématique d'un neurone artificiel j est illustré dans la Figure C.1. Un neurone est essentiellement constitué d'un intégrateur qui effectue la somme pondérée de ses entrées x_i complété par un biais b_j qui représente le seuil d'activation du neurone j.

$$s_j = \Sigma_i w_{ij} x_i + b_j \tag{C.1}$$

s_j et w_ij représentent respectivement la valeur d'activation et les poids synaptiques du neurone j. Après avoir calculé s_j, une fonction d'activation f_j appliquée à la somme pondérée de ses entrées est nécessaire pour déterminer l'état d'activation du neurone j donné par la relation : $y_j = f_j(s_j)$.

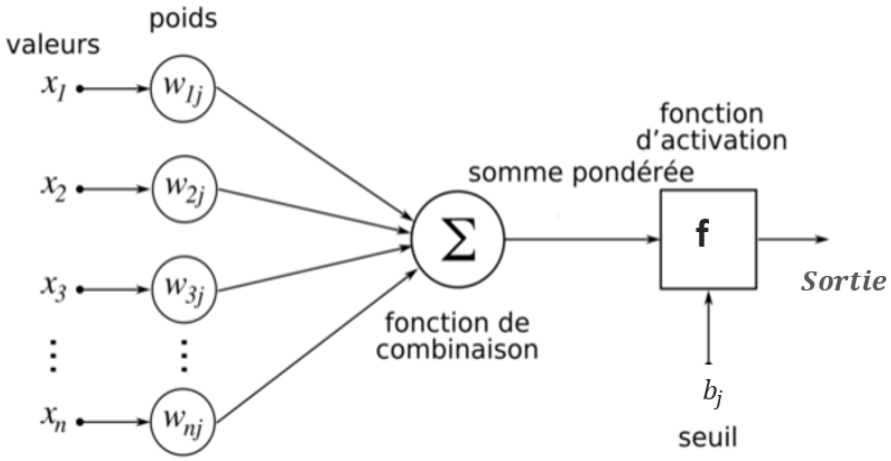


FIGURE C.1 – Schéma d'un neurone artificiel.

Fonctions d'activation :

Plusieurs formes de cette fonction d'activation peuvent être appliquée, on peut citer :

— L'identité : $f(x) = x$

— Logistique (ou marche douce ou sigmoïde) : $f(x) = 1/(1 + e - x)$

— Tangente Hyperbolique (TanH) : $f(x) = \tanh(x) = 2/(1 + e^{-2x}) - 1$

— Unité de Rectification Linéaire (ReLU) : $f(x) = \begin{cases} 0 & for \quad x < 0 \\ x & for \quad x \geq 0 \end{cases}$

Apprentissage avec l'algorithme de la rétro-propagation La technique d'apprentissage la plus connue dans les réseaux MLP est l'algorithme de rétro-propagation du gradient (backpropagation). Cet algorithme ajuste les poids d'un réseau dont l'architecture est fixée par l'opérateur, à chaque fois qu'un exemple est présenté. Cet ajustement est fait afin de minimiser l'erreur entre la sortie obtenue par le réseau et la sortie désirée. Les opérations de la rétro-propagation dans les réseaux de neurones peuvent être divisées en deux étapes :

— Propagation des entrées vers l'avant produisant une sortie

— Rétropropagation d'erreur de la couche de sortie à la couche d'entrée : Si la sortie du MLP est différente de la sortie désirée alors l'erreur est propagée de la couche de sortie vers la couche d'entrée tout en ajustant les poids.

Dans l'étape de propagation vers l'avant, une donnée d'entrainement x est appliqué à la couche d'entrée et son effet se propage de couche en couche à travers le réseau jusqu'à ce qu'une sortie y soit produite. Le réseau est initialisé avec des poids choisis au hasard. En notant w_{ij} le poids de la connexion menant de l'entrée d'indice i à la sortie d'indice j, on peut calculer pour une entrée x sa

sortie correspondante :

$$y_j = f(b_j + \sum_i w_{ij} x_i) \tag{C.2}$$

f est la fonction d'activation, puisque l'une des exigences de l'algorithme de rétro-propagation est que la fonction d'activation est différentiable, une fonction d'activation typique est l'équation sigmoïde. Donc f est la fonction sigmoide de paramètre $\lambda = 1$.

$$y_j = f(u_j) = \frac{1}{e^{-u_j}} \tag{C.3}$$

$$u_j = b_j + \sum_i w_{ij} x_i \tag{C.4}$$

La valeur de la sortie y obtenue par le réseau est ensuite comparée à la sortie désirée d, et un signal d'erreur E est calculé pour chaque neurone j dans la couche de sortie.

$$E = \frac{1}{2} \sum_{j=1}^{n} (d_j - y_j)^2 \tag{C.5}$$

n est le nombre de neurones dans la couche de sortie. Le calcul des erreurs est effectué à partir de la couche de sortie et se propager vers l'arrière dans les couches cachées jusqu'à atteindre la couche d'entrée. La couche d'entrée est exclue du calcul d'erreur. Les poids w_{ij} dans le réseau sont les seuls paramètres qui peuvent être modifiés afin de rendre l'erreur E aussi faible que possible. La minimisation de E se fait avec la méthode de la descente du gradient basée sur la formule itérative suivante :

$$w_{ij}(t) = w_{ij}(t-1) + \Delta w_{ij}(t) \tag{C.6}$$

$w_{ij}(t)$ est le poids de la connexion de neurone i au neurone j à l'itération t et $\Delta w_{ij}(t)$ correspond à l'ajustement du poids. Suivant la régle de delta :

$$\Delta w_{ij}(t) = -\rho \frac{\delta E(w)}{\delta w_{ij}(t)} \tag{C.7}$$

$$= \rho \sigma_j x_i \tag{C.8}$$

ρ est le taux d'apprentissage, $0 < \rho < 1$, et σ_j représente l'erreur de gradient locale faite par le neurone j.

$$\sigma_j = \begin{cases} f'(u_j)(d_j - y_j) \quad si \quad j \in couche \quad de \quad sortie \\ f'(u_j) \sum_{k \in dest(j)} w_{kj} \sigma_k, \quad si \quad j \in couche \quad cachee \end{cases}$$

$dest(j)$ est l'ensemble des neurones auxquels j se connecte, $f'(u_j) = f(u_j)(1 - f(u_j))$.

La convergence est souvent plus rapide en ajoutant un terme dynamique $0 < \alpha < 1$:

$$w_{ij}(t+1) = w_{ij}(t) + \rho\sigma_j x_i + \alpha(w_{ij}(t) - w_{ij}(t-1)) \tag{C.9}$$

La procédure se termine quand un nombre maximum d'itérations est atteint ou jusqu'à ce que la valeur de la fonction d'erreurs soit inférieure à un certain seuil. Les deux paramètres importants dans l'algorithme de rétro-propagation sont le taux d'apprentissage ρ et α un facteur d'inertie. Le facteur ρ influe sur la vitesse de convergence du réseau. Un taux d'apprentissage trop faible produit un ralentissement dans la convergence et un taux trop élevé produit la divergence. L'ajustement du paramètre α permet d'éviter les effets d'oscillations ou bien de rester coincé dans un minimum local.

Annexe D

Les forêts d'arbres décisionnels

Phase d'entrainement

Les forêts d'arbres décisionnels ou forêts aléatoires sont des techniques d'apprentissage automatique proposées en 2001 par Leo Breiman [Breiman, 2001]. Cet algorithme effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. Un arbre de décision binaire générique peut être décrit récursivement par un nœud n, appelé soit un nœud terminal, soit un nœud divisé, comme indiqué dans la Figure D.1. Si le nœud n est terminal (représenté en gris sur la Figure D.1), alors on lui associe une distribution terminale, qui consiste soit en une probabilité d'assignation $P(c|x)$ pour chaque classe $c \in \{1, \ldots, C\}$. D'un autre côté, si le nœud n n'est pas terminal, il est appelé un nœud de division (nœuds verts sur la Figure D.1) et contient une fonction de division paramétrique associé à un seuil pour le diviser en deux nœuds fils. Donc à chaque pas du partitionnement, on découpe une partie de l'espace en deux sous parties. On associe alors naturellement un arbre binaire à la partition construite. Les nœuds de l'arbre sont associés aux éléments de la partition, et ainsi de suite jusqu'à atteindre la taille maximale de l'arbre. Donc commençant par la racine de l'arbre qui contient l'espace d'entrée tout entier E, la règle de partitionnement consiste à construire deux sous-parties E_L et E_R comme suit : $E_L = x \in E|x_i < \theta_i$ et $E_R = x \in E|x_i \geq \theta_i$. Le couple (i, θ_i) est choisi de sorte que chaque nœud fils soit le plus homogène possible.

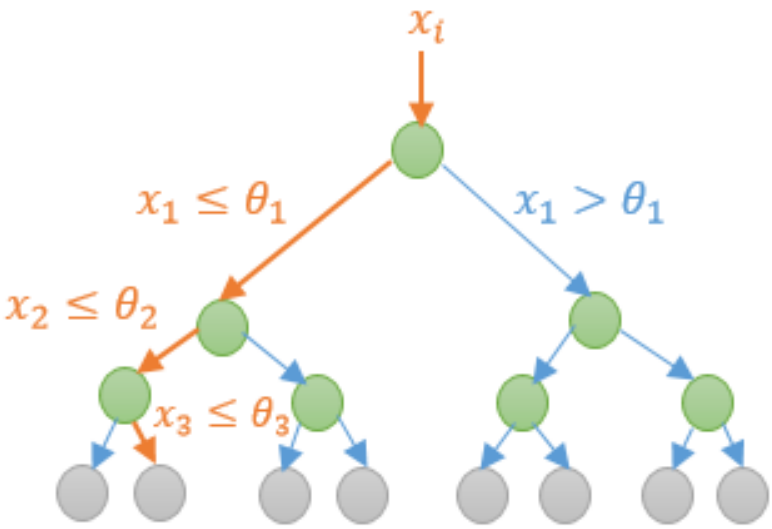


FIGURE D.1 – Un arbre de décision.

L'homogénéité d'un nœud N est généralement basée sur une mesure d'impureté, qui dépend du degré d'homogénéité des échantillons appartenant au nœud N. L'impureté est minimale lorsque les échantillons qui composent le noeud N appartiennent à la même classe, et elle est maximale lorsque les classes sont uniformément distribuées dans un nœud avec la même proportion. Les deux mesures standard d'impureté d'un nœud i sont :

— Le gain d'information qui consiste à mesurer la quantité d'information nécessaire pour déterminer la classe d'un échantillon. Il est exprimé en fonction de l'entropie de Shannon :

$$I(i) = \sum_{k=1}^{K} \frac{n_{ki}}{N_i} \log 2 \frac{n_{ki}}{N_i} \tag{D.1}$$

K est le nombre des classes, n_{ki} représente l'effectif de la classe k dans le nœud i, N_i correspond au nombre des échantillons dans le nœud de division i.

— L'indice de Gini d'un noeud i est calculé de la manière suivante :

$$I(i) = 1 - \sum_{k=1}^{K} (\frac{n_{ki}}{N_i})^2 \tag{D.2}$$

Le changement d'impureté dans le nœud de division i :

$$\Delta I(i) = I(i) - \frac{G_i}{N_i} I(g_i) - \frac{D_i}{N_i} I(d_i) \tag{D.3}$$

D_i, G_i correspondent respectivement aux nombres des échantillons dans le nœud fils à droite et à

gauche, avec l'indice d_i et g_i. Formellement, le critère ΔI cherche à maximiser la différence d'impureté entre les échantillons du nœud i et les échantillons des deux nœuds fils. Nous choisissons donc la caractéristique et le seuil qui minimise l'impureté du noeud par rapport à la classe cible.

Phase de prédiction

Considérons un ensemble de données d'apprentissage E constitué de n échantillons, utilisé pour dériver des règles de prédiction en appliquant l'algorithme RF avec un nombre d'arbres T. Idéalement, la performance de ces règles de prédiction est estimée en se basant sur une base de test indépendante, notée D_{test}, constituée d'échantillons de test n_{test} . Considérant le ième échantillon de l'ensemble de données de test ($i = 1, \ldots, n_{test}$), nous notons sa réponse réelle, y_i, qui présente une étiquette binaire 0 contre 1 (dans le cas d'une classification binaire). La valeur prédite de la sortie produite par l'arbre t (avec $t = 1, \ldots, T$) est noté $\hat{y}_{it}$, où $\hat{y}_i$ est la valeur prédite de la sortie par toute la forêt aléatoire.

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} I(\hat{y}_{it} = 1) \tag{D.4}$$

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Reconnaissance des gestes expressifs inspirée du modèle LMA pour une interaction naturelle Homme-Robot

Mots clés : gestes corporels expressifs, émotions, modèle LMA, apprentissage supervisé, reconnaissance des gestes, interaction homme-robot

Résumé : Dans cette thèse, nous traitons le problème de la reconnaissance des gestes dans un contexte d'interaction homme-robot. De nouvelles contributions sont apportées à ce sujet. Notre système consiste à reconnaitre les gestes humains en se basant sur une méthode d'analyse de mouvement qui décrit le geste humain d'une manière précise. Dans le cadre de cette étude, un module de niveau supérieur est intégré afin de reconnaître les émotions de la personne à travers le mouvement de son corps. Trois approches sont réalisées: la première porte sur la reconnaissance des gestes dynamiques en appliquant le modèle de Markov caché (MMC) comme méthode de classification. Un descripteur de mouvement local est implémenté basé sur une méthode d'analyse de mouvement, nommée LMA (Laban Movement Analysis) qui permet de décrire le mouvement de la personne dans ses différents aspects. Notre système est invariant aux positions et orientations initiales des personnes. Un algorithme d'échantillonnage a été développé afin de réduire la taille de notre descripteur et aussi adapter les données aux modèles de Markov cachés. Une contribution est réalisée aux MMCs pour analyser le mouvement dans deux sens (son sens naturel et le sens inverse) et ainsi améliorer la classification des gestes similaires. Plusieurs expériences sont faites en utilisant des bases de données d'actions publiques, ainsi que notre base de données composée de gestes de contrôle. Dans la seconde approche, un système de reconnaissance des gestes expressifs est mis en place afin de reconnaitre les émotions des personnes à travers leurs gestes. Une deuxième contribution consiste en le choix d'un descripteur de mouvement global basé sur les caractéristiques locales proposées dans la première approche afin de décrire l'entièreté du geste. La composante Effort de LMA est quantifiée afin de décrire l'expressivité du geste avec ses 4 facteurs (espace, temps, poids et flux). La classification des gestes expressifs est réalisée avec 4 méthodes d'apprentissage automatique réputées (les forêts d'arbres décisionnels, le perceptron multicouches, les machines à vecteurs de support: un-contre-un et un-contre-tous). Une étude comparative est faite entre ces 4 méthodes afin de choisir la meilleure. L'approche est validée avec des bases publiques et notre propre base des gestes expressifs. La troisième approche consiste en une étude statistique basée sur la perception humaine afin d'évaluer le système de reconnaissance ainsi que le descripteur de mouvement proposé. Cela permet d'estimer la capacité de notre système à pouvoir classifier et analyser les émotions comme un humain. Dans cette partie deux tâches sont réalisées avec les deux classifieurs (la méthode d'apprentissage RDF qui a donné les meilleurs résultats dans la deuxième approche et le classifieur humain): la classification des émotions et l'étude de l'importance des caractéristiques de mouvement pour discriminer chaque émotion.

Title : Expressive gestures recognition inspired by the LMA model for natural Human-Robot interaction

Keywords : expressive body gestures, emotions, LMA model, supervised learning, gestures recognition, human-robot interaction

Abstract : In this thesis, we deal with the problem of gesture recognition in a human-robot interaction context. New contributions are being made on this subject. Our system consists in recognizing human gestures based on a motion analysis method that describes movement in a precise way. As part of this study, a higher level module is integrated to recognize the emotions of the person through the movement of her body. Three approaches are carried out: the first deals with the recognition of dynamic gestures by applying the hidden Markov model (HMM) as a classification method. A local motion descriptor is implemented based on a motion analysis method, called LMA (Laban Movement Analysis), which describes the movement of the person in its different aspects. Our system is invariant to the initial positions and orientations of people. A sampling algorithm has been developed in order to reduce the size of our descriptor and also adapt the data to hidden Markov models. A contribution is made to HMMs to analyze the movement in two directions (its natural and opposite directions) and thus improve the classification of similar gestures. Several experiments are done using public action databases, as well as our database composed of control gestures. In the second approach, an expressive gestures recognition system is set up to recognize the emotions of people through their gestures. A second contribution consists of the choice of a global motion descriptor based on the local characteristics proposed in the first approach to describe the entire gesture. The LMA Effort component is quantified to describe the expressiveness of the gesture with its four factors (space, time, weight and flow). The classification of expressive gestures is carried out with four well-known machine learning methods (random decision forests, multilayer perceptron, support vector machines: one-against-one and one-against-all. A comparative study is made between these 4 methods in order to choose the best one. The approach is validated with public databases and our database of expressive gestures. The third approach is a statistical study based on human perception to evaluate the recognition system as well as the proposed motion descriptor. This allows us to estimate the ability of our system to classify and analyze emotions as a human. In this part, two tasks are carried out with the two classifiers (the RDF learning method that gave the best results in the second approach and the human classifier): the classification of emotions and the study of the importance of our motion features to discriminate each emotion.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France