Analyse du contenu expressif des gestes corporels

Arthur Truong

THESE DE DOCTORAT CONJOINT TELECOM SUDPARIS et L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique et Télécommunications

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Arthur TRUONG

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Analyse du contenu expressif des gestes corporels

Soutenue le 21 septembre 2016

Composition du jury :

Rapporteur :	**Monsieur le Professeur**	**Malik MALLEM**
Rapporteur :	**Monsieur le Professeur**	**Antoine MANZANERA**
Président du jury :	**Madame le Maître de Conférence, HDR**	**Catherine ACHARD**
Examinatrice :	**Madame la Chargée de Recherche INRIA, HDR**	**Anne VERROUST-BLONDET**
Examinateur :	**Monsieur le Docteur**	**Bertrand DELEZOÏDE**
Directeur de thèse :	**Monsieur le Professeur**	**Titus ZAHARIA**

N° NNT : 2016TELE0015

Analyse du contenu expressif des gestes corporels

Remerciements

En tout premier lieu, je souhaiterais remercier chaleureusement mon directeur de thèse, Titus Zaharia, pour son soutien, et pour l'intérêt porté à mes travaux de recherche durant presque quatre ans, un chemin semé d'embûches que ses conseils et encouragements m'ont permis de surmonter. Je veux témoigner du respect qui fut le sien à l'égard de mes divers choix et prises de position, notamment au cours des longs et difficiles premiers mois de débroussaillage du terrain. Plus spécifiquement, je le remercie pour la pleine confiance qu'il a bien voulu accorder à mon engagement sur des problématiques peu courantes à l'interface entre les sciences humaines et la technologie.

Je voudrais également exprimer ma gratitude envers les membres de mon jury de thèse, pour l'intérêt qu'ils ont prêté à mes recherches.

Mes chaleureux remerciements vont tout d'abord à Madame Catherine Achard, Maître de conférences HDR à l'Université Pierre et Marie Curie, pour l'intérêt qu'elle a manifesté pour mes travaux et pour m'avoir fait l'honneur de présider ce jury.

Je tiens à exprimer toute ma reconnaissance à Messieurs les Professeurs Malik Mallem, de l'Université d'Evry-Val-d'Essonne, et Antoine Manzanera, de l'ENSTA-ParisTech, qui ont accepté la tâche lourde et particulièrement prenante de rapporter sur mon travail de thèse.

Je tiens également à remercier Madame Anne Verroust-Blondet, chargée de recherche HDR à l'INRIA, pour son regard attentif et éclairé porté sur ce manuscrit.

Enfin, j'exprime mes remerciements les plus sincères à Monsieur Bertrand Delezoïde, du CEA-LIST, pour avoir accepté d'apporter sa vision sur ces travaux de recherche.

J'en profite également pour dire ma sympathie à l'ensemble du personnel du département ARTEMIS de l'Institut Mines-Télécom/Télécom SudParis, au sein duquel j'avais déjà effectué une partie de ma formation initiale d'ingénieur, et dont je savais par avance qu'il constituerait un environnement de recherche idéal en vue d'une thèse de doctorat. Merci donc aux professeurs pour ce climat de travail. Merci également à Madame Evelyne Taroni pour sa patience, son énergie et son efficacité sans reproche à gérer les aspects administratifs et nous permettre de travailler, voyager et vivre dans le laboratoire avec sérénité et tranquillité.

Je suis par ailleurs reconnaissant envers Raluca Diana Petre, Andrei Bursuc et Adriana Garboan, qui par leurs recommandations diverses m'ont mis le pied à l'étrier au tout début de ma thèse.

Aussi, je remercie tout particulièrement Hugo Boujut, dont les aides et conseils techniques multiples furent précieux et décisifs, durant nos deux ans et demi d'amicale collaboration.

En outre, je souhaiterais également remercier les musiciens, chefs d'orchestre et musicologues qui ont directement ou indirectement participé à l'élaboration de mon corpus de gestes de direction orchestrale, et sans lesquels tout un pan de mes expérimentations n'aurait jamais pu voir le jour. Merci aux chefs d'orchestre Bruno Poindefert, Gabriel Bourgoin, William Lesage, Christophe Dilys, Didier Horry, Luc Bonnaillie, ainsi qu'à Xavier Gagnepain, Jean-Marie Adrien, et à tous les élèves du Conservatoire National Supérieur de Musique et de Danse de la ville de Paris qui ont pris part à ce projet.

Merci également à Guillaume Devineau et Alexandre Brun, élèves pour la saison 2014-2015 dans la majeure High-Tech Imaging de Télécom SudParis, qui m'ont aidé à constituer le corpus de séquences

gestuelles qui a servi comme objet d'études et d'expérimentations pour les derniers développements de ma thèse.

Pour finir, je voudrais adresser un grand merci à mes proches, ma famille et mes amis, pour leurs encouragements et leur soutien permanent. Enfin et surtout, merci à ma bien-aimée Silvia Alvarez Baamonde pour son amour, sa tolérance et sa patience à l'égard des humeurs, stresses et doutes que n'ont pas manqué de me causer quatre années de thèse qui n'auraient jamais pu ni commencer ni aboutir sans elle.

Table des matières

Liste des figures

Liste des tableaux

Résumé

Les problématiques d'analyse et d'interprétation des contenus gestuels sont depuis quelques décennies l'objet d'un intérêt recrudescent, à mesure qu'émergent de nouvelles technologies de capture aussi précises que robustes. En effet, le geste est un domaine dont les enjeux et perspectives de développement ne sont pas moindres : nouvelles interfaces homme-machine (interfaces performatives, interfaces corporalisées, etc.), réalité augmentée, assistance à la médecine, rééducation des personnes handicapées, analyse de comportement de groupes, création artistique, analyse de l'affect et des émotions…

Aujourd'hui, les recherches portant sur le geste manquent cruellement de modèles mathématiques qui soient unifiés et génériques. Les modèles existants font trop souvent dépendre les descripteurs du mouvement corporel de la spécificité des contenus à analyser. En outre, le problème qui se pose aux spécialistes du geste est le suivant : il faut trouver un compromis entre d'une part, une formalisation excessivement conceptuelle, et d'autre part une description purement visuelle du mouvement, qui le cas échéant se révèle inapte à saisir le sens du mouvement *vécu*.

A ce titre, notre contribution est plurielle.

Nous reprenons tout d'abord les concepts développés par le chorégraphe Rudolf Laban pour l'analyse et l'enseignement de la danse classique contemporaine (*Laban Movement Analysis : LMA*), et proposons leur extension afin d'élaborer un modèle descriptif et générique du geste, basé sur ses éléments expressifs.

Nous présentons également deux corpus de gestes 3D que nous avons enregistrés à l'aide d'une Kinect, et que nous avons constitués pour tester la validité de notre modèle de l'expressivité du geste à caractériser des contenus variés. Le premier de ces corpus, *ORCHESTRE-3D*, se compose de gestes pré-segmentés de chefs d'orchestre enregistrés en répétition. Ce corpus a été annoté selon un lexique d'émotions musicales, dans le but d'étudier la relation entre le contenu émotionnel véhiculé par le chef d'orchestre et ses mouvements concrets. Le deuxième corpus, *HTI 2014-2015*, construit en collaboration avec les élèves de la filière High-Tech Imaging (HTI) de TELECOM SudParis pour la saison 2014-2015, propose des séquences d'actions variées de la vie quotidienne.

Dans une première approche de reconnaissance dite « globale », nous utilisons notre modèle pour définir un vecteur descripteur du geste qui se rapporte à l'entièreté de sa durée de vie. Ce descripteur global donne lieu à deux expériences de classification basées sur des méthodes d'apprentissage supervisé : la première a trait à la discrimination des actions d'un corpus de référence ; la deuxième a pour objet notre corpus *ORCHESTRE-3D*, et vise à reconnaître les émotions musicales que portent les gestes des chefs d'orchestre.

Dans une seconde approche dite « dynamique » ou « locale », les caractéristiques de notre modèle sont reprises pour constituer des descripteurs de trame gestuelle (*e.g.* pour chaque instant du geste). Ces descripteurs de trame sont utilisés pour créer des dictionnaires de poses-clés du mouvement. Ces poses permettent d'obtenir à tout instant une représentation simplifiée du geste, utilisable pour reconnaître des actions à la volée. Nous testons cette approche sur plusieurs bases de geste, dont notre propre corpus *HTI 2014-2015*.

Nous montrons que les résultats de classification obtenus sur nos propres bases de gestes, ainsi que sur d'autres corpus 3D de référence, valident notre modèle de geste et les utilisations que nous en faisons.

I. Introduction

L'analyse des comportements humains, et en particulier la compréhension des contenus gestuels, ont reçu ces dernières décennies une attention croissante. Si les progrès réalisés dans les domaines des technologies de capture et d'analyse de l'activité dans des contenus multimédias peuvent en partie expliquer un tel attrait, ils ne sauraient occulter la centralité du corps dans la constitution du savoir, pas plus qu'ils ne sauraient éclipser les liens entre la corporéité et l'élaboration de la subjectivité. Dans un premier temps et pour nous inscrire dans le fil directeur de ces problématiques, nous tentons de restituer la logique de l'émergence de la phénoménologie du corps-sujet à l'aune de la modernité, puis nous étudions sa saisie par l'univers des technologies du vingtième siècle, en lien avec l'art, l'esthétique, et la recherche d'interactions homme-machine (IHM) au plus près de l'expérience sensible.

I.1. Contexte général

L'interaction entre l'homme et son environnement met en jeu un certain nombre de signaux communicationnels (voix, posture, expressions faciales, mouvements corporels…), que ceux-ci soient l'objet d'actes délibérés ou non. Un geste est défini comme un ensemble de mouvements corporels qui contiennent de l'information [**1**]. La compréhension automatique des gestes reste un enjeu de taille pour de nombreuses applications : médecine assistée par ordinateur, vidéo surveillance, mesure d'activité au sein de groupes, jeux vidéos, création artistique ou plus généralement toute situation mettant en jeu des interfaces homme-machine (Figure I.1).

Figure I.1 Exemples de contextes mettant en jeu le traitement du geste : jeux vidéo, danse, direction musicale dansée.

L'apparition récente de technologies de capture capables de suivre en temps réel et avec une certaine fiabilité les positions 3D d'articulations de référence du corps (Kinect...) justifie que les applications logicielles exploitant les communications humaines aient fait l'objet de toujours plus de travaux au cours de la dernière décennie.

Dans ce travail de thèse, l'objectif général consiste à caractériser sémantiquement les mouvements corporels à l'aide de descripteurs construits à partir de signaux visuels. Aujourd'hui, les approches existantes utilisent bien souvent des représentations plus ou moins intuitives, qui resteraient à valider au moyen d'études perceptuelles sur le geste. En outre, il y a un manque de modèles gestuels unifiés et applicables à des situations variées. En effet, les types de contenus que nous souhaitons qualifier se veulent aussi divers que possible (actions spécifiques, caractérisation émotionnelle du geste selon un modèle de l'affect prédéfini, contenu expressif du mouvement…). Un modèle de description efficace devra donc nécessairement prendre en compte de multiples indices gestuels, dont notamment le dynamisme, l'expressivité, l'intentionnalité et la communication intersubjective.

Le corps en mouvement est aussi une préoccupation philosophique de longue date. Dans le paragraphe qui suit, nous rappelons brièvement quelques considérations d'ordre philosophique qui rendent intelligible le sens du corps et de son mouvement par rapport au geste, et qui ont guidé notre démarche tout au long de ce travail de thèse.

I.2. De la modernité au corps-sujet : quelques rappels de philosophie

L'analyse des gestes, la compréhension de leur sens, ainsi que les interrogations relatives aux contenus du mouvement corporel remontent à l'Antiquité. En effet, les rapports entre l'âme et le corps ont depuis très longtemps été un objet d'intérêt. Dans le cinquième livre de *La République* [**2**], Platon explique déjà que la musique doit être partie intégrante de l'éducation des jeunes hommes et femmes, dans la mesure où elle met en mouvement l'âme et le corps et participe de leur élaboration. Au dix-huitième siècle, dans *La Nouvelle Héloïse*, qui consacre l'apparition d'une littérature du sentiment et de la psyché à une époque où le rationalisme politique est conquérant, Jean-Jacques Rousseau évoque les mouvements de l'âme d'un de ses personnages, Saint-Preux, alors qu'il écoute de la musique. Il développera dans *Les rêveries du promeneur solitaire* [**3**], sa dernière œuvre, l'idée que le corps et ses mouvements jouent un rôle fondamental dans l'élaboration de la pensée, prenant ainsi de court les idées préexistantes qui ne voyaient de relation entre l'âme et de corps que dans le sens inverse. Au « je pense donc je suis » de Descartes [**4**], Rousseau répond « je bouge donc je suis ». Cette formule n'est pas sans évoquer celle de Baruch Spinoza selon laquelle *« l'âme est l'idée du corps »*, et la médiation que cette dernière implique entre deux substances que Descartes pensait séparées, à savoir la *res extensa* (*e.g.*, le corps) et la *res cogitans* (*e.g.*, la conscience). Dans *Spinoza avait raison; joie et tristesse, le cerveau des émotions* [**5**], Antonio Damasio défend la position du philosophe néerlandais en s'appuyant sur une conception matérielle de l'esprit qui trouve son fondement dans le développement récent des neurosciences : l'*homonculus*, qui est la carte du corps dans le cerveau, est affecté par les émotions. Le mode sensoriel et le mode émotionnel de la perception sont donc biologiquement liés. Les émotions ont trait au corps et précèdent les sentiments, qui eux relèvent de l'esprit et sont produits par l'expérience, la raison et l'imagination.

Cette idée selon laquelle l'âme est produite par le vécu, et donc en particulier le vécu corporel, nous intéresse, car un geste est aussi un acte d'exploration du monde. Il s'agit d'interroger les rapports qu'entretiennent la conscience, dont le siège est la subjectivité, et le monde extérieur, et de tenter de définir en quoi l'expérience, en particulier l'expérience corporelle, fonde notre rapport à l'altérité.

Pour la philosophie classique, seule l'intelligence nous découvre la réalité du monde. Dans la *Seconde Méditation métaphysique* [**4**] et le célèbre extrait consacré au *morceau de cire*, René Descartes explique qu'on ne peut concevoir la cire qu'avec l'intelligence et la pensée. Ainsi, user de ses sens, c'est penser – mais penser à travers les qualités qui tombent sous ces sens. Le rapport de la perception à la science est celui de l'apparence à la réalité. Pour René Descartes, il s'agit de se dresser à la fois contre les scholastiques de la Sorbonne, siège du dogmatisme théologique de son époque, mais également contre les sceptiques, qui doutent du dogme sans en retour proposer de méthodologie du savoir.

L'identification entre philosophie de la connaissance et philosophie de la conscience est un progrès historique fondamental, dont Emmanuel Kant définira les limites dans sa *Critique de la raison pure* [**6**]. Il y montre que la compréhension du monde se confronte à l'acte même de connaître, car cet acte de connaître modifie l'objet de la connaissance. Tout comme la révolution copernicienne a consisté à montrer que la Terre tourne autour du Soleil (et non l'inverse), la révolution épistémologique opérée par

Kant pose comme centre de la connaissance le sujet raisonnable et actif, et non plus un monde extérieur par rapport auquel celui-ci serait passif. Le sujet est posé comme au-delà, *transcendantal*. C'est le sens de la phrase suivante : *« Les objets se règlent sur notre connaissance »*, issue de la préface de ladite critique. Ainsi, le sujet de la connaissance ne peut avoir accès qu'au *phénomène* de la connaissance en lequel consiste l'acte de connaître, et non pas au *noumène*, autrement appelé *« chose en soi »*.

Dès la critique kantienne donc, la problématique de l'agencement du sujet et de l'objet, du moi et du monde, de l'intérieur et de l'extérieur, est posée. Le *moi transcendantal* est par ailleurs ce qui unie les différents états de notre *moi empirique* (cartésien), qui varie avec les expériences quotidiennes du désir, du doute, de la perception sensible, des sentiments, ou d'autres attitudes mentales ; il est ce qui ramène ces expériences subjectives à une unité nécessaire au cours du temps, et permet la représentation de cette unité.

C'est justement le problème du transcendantal qui intéresse Edmund Husserl dès le début du vingtième siècle, et particulièrement le *phénomène*. L'acte même en lequel consiste l'expérience subjective n'est pas dénué d'intérêt, quand bien même on considère comme Kant que la *chose en soi* ne peut pas être atteinte.

« Lorsque le monde général fait son « apparence » dans la conscience comme « le » monde […] il adopte une nouvelle dimension et devient « incomplètement intelligible, questionnable ». C'est là que réside le problème transcendantal : ce « faire son apparence », ce « être pour nous », qui ne peut conquérir sa signifiance que « subjectivement », qu'est-il ? » [**7**]

Pour Husserl, le problème transcendantal se formule ainsi : en tant qu'individu ou en tant qu'être social, l'homme appartient au monde, mais le monde n'est valide que dans sa conscience ; l'homme est sujet d'une vie psychique, et en même temps il est au-delà de lui-même :

« Le monde générique, dont l'être immanent est aussi obscur que la conscience par le biais de laquelle il existe, trouve cependant le moyen d'apparaître à nous en une variété d'aspects particuliers, et l'expérience nous enseigne que ce sont les aspects d'un seul monde, existant de façon autonome. » [**7**]

Par un tel constat, Husserl en appelle à un retour à l'*empirisme*, pour étudier le phénomène même de l'expérience. Il faut étudier la conscience de soi, mais aussi la conscience des autres soi, non pas pour accéder à une duplication de ce que nous trouvons comme notre conscience de soi, mais découvrir ce qui cimente la vie sociale : l'*intentionnalité*. Lorsque la conscience se tourne vers ses objets, le rôle échoit à l'intentionnalité d'unifier les impressions émanant des sens. Comme l'explique le philosophe Jean-Luc Petit dans [**8**], le corps peut dès lors devenir l'organe du « Je peux », d'une expérience subjective qui réunit les êtres humains au sein de diverses interactions où les anticipations jouent un rôle éminent. Le corps est l'organe du « vouloir connaître ».

De ce primat accordé à l'intersubjectivité, le philosophe romantique Theodor Lipps fait naître et explore la notion d'*empathie* qu'il définit comme un sentiment que n'importe qui peut avoir de sa propre activité interne alors qu'il est en train d'observer le geste d'autrui [**9**]. Selon Lipps, il est possible de « saisir » la vie extérieure depuis sa propre intériorité indépendamment du lieu où le geste en question est effectué – il donne l'exemple d'un public qui observe un funambule, et définit spontanément des critères de sécurité par transfert vers le corps du funambule ; on parle alors d' « *empathie kinesthétique* ». Le sujet sent sa propre activité corporelle, comme si c'était dans l'autre. *« Je sais bien ce que cet autre fait, parce que je peux faire la même chose. »* Pour le physiologiste et professeur au Collège de France Alain Berthoz, la *kinesthésie* ou *sens du mouvement* est même l'équivalent d'un sixième sens, qui présente la particularité de s'imprégner dans notre mémoire neuro-motrice de façon durable [**10**].

Suivant l'approche de Husserl, le philosophe français Maurice Merleau-Ponty propose dans ses cours au Collège de France [**11**] de nouvelles approches phénoménologiques pour la « connaissance du

monde », réunies sous le vocable de « *phénoménologie de la perception* » ; la perception doit dépasser le clivage classique entre idéalistes (ou subjectivistes, pour lesquels *le monde est le produit de l'esprit*) et matérialistes (ou objectivistes, pour qui *le monde est ce qui produit de l'esprit*), en investissant la relation entre expérience et conscience, entre action et perception (on parlera alors d'*expérience corporalisée*). L'expérience du corps brouille la distinction du sujet et de l'objet, parce que le corps n'est pas seulement le corps que l'on « a » mais également celui que l'on « est ». Le corps est donc « *corps-sujet* ».

Dans le troisième entretien de l'ouvrage intitulé *The World of Perception* [**12**], Merleau-Ponty développe le concept d' « *objet sensible* ». Il s'agit de rendre justice à l'idée qu'un objet puisse être une unité, un être, dont les différentes qualités sont les différentes manifestations. Pour le philosophe français, les couleurs sont des qualités sensibles, dont la particularité est de posséder une signification affective qui les met en correspondance avec des qualités que fournissent d'autres sens que la vue, comme par l'exemple l'ouïe : les couleurs sont capables d'afficher des atmosphères morales qui les rendent tristes ou gaies, et il en est de même pour les sons. Dès lors, la mise en correspondance entre une couleur et un son est possible ; ainsi, les aveugles arrivent-ils à se représenter les couleurs par l'analogie d'un son. L'expérience nous permet donc de dresser des rapports entre des qualités indépendantes. D'autre part, certaines qualités, auxquelles nous donne accès l'expérience, n'ont presque aucun sens en tant que tel, si ce n'est qu'elles provoquent des réactions de la part de notre corps. Merleau-Ponty donne l'exemple du *mielleux* : le miel que l'on manipule, se saisit lui-même des mains de celui qui voulait le saisir. Par un renversement des rôles, *« la main qui voulait attirer à elle le miel se retrouve engluée dans l'être extérieur »*. Le possédé finit par posséder le possédant. Le mielleux ne se comprend donc que par la confrontation entre le sujet incarné et l'objet extérieur qui la porte, une confrontation où le sujet *est fait par* l'objet. Et il en est de même d'autres qualités, comme le *sucré*. Ainsi, l'unité d'un objet sensible est-il réaffirmé par chacune de ses qualités : chacune d'elles est l'objet entier, et donc, chacune de ces qualités est toutes les autres. Une telle constatation donne pleinement son sens à la synesthésie – Paul Cézanne disait qu'il était possible de *« peindre l'odeur des arbres »*.

Les choses du monde ne sont donc pas de simples objets qui seraient extérieurs au sujet. Le sujet (l'esprit) et l'objet (du monde) se déterminent l'un l'autre réciproquement, dans la mesure où toute connaissance est une expérience intentionnelle du monde. C'est ainsi que Merleau-Ponty écrit : *« L'homme est investi dans les choses, et les choses sont investies en lui. »* Si Kant nous a bien avertis qu'il n'y a pas d'accès à un objet qui soit vierge de toute trace subjective, la phénoménologie, elle, rend justice à l'expérience humaine et sensible dans l'élaboration du savoir. Sans doute l'art y participe-t-il également. Sa confrontation avec la science est donc décisive à cet effet, comme nous allons le voir maintenant.

I.3. Du corps-sujet vers nouvelles technologies

La phénoménologie de la perception, du corps-sujet, connaît une descendance et des applications diverses. L'innovation en matière d'interfaces homme-machine ne concerne plus de simples enjeux ergonomiques, mais se tournera vers la conception même de l'interaction entre humain et interface homme-machine (IHM). Il s'agit d'étudier dans quelle mesure l'homme s'approprie une interface et développe un certain nombre de pratiques mettant en jeu la corporéité, l'esthétique, l'expressivité gestuelle, les affects, ou encore les émotions.

I.3.1. Perception, esthétique et interactions corporalisées

Dans [**13**], Billinghurst et Buxton font remarquer que les interactions par le geste mettent en jeu le mouvement corporel en tant qu'il est articulé, exprimé et reconnu par un tiers, et non en tant qu'acte dont les seules conséquences importeraient. Ils soulignent le fait que la plupart des IHM ne mettent en jeu que la reconnaissance de gestes dits « *symboliques* » ou « *déictiques* » selon la taxonomie de McNeill [**14**] (*cf.* section II.1.1). Il faut évaluer dans quelle mesure la technologie peut permettre de renseigner également sur les affects, les humeurs, les émotions… en un mot les *contenus* qui n'a pas directement trait au *but* poursuivi par l'utilisateur.

Dans *L'inscription corporelle de l'esprit* [**15**], Varela *et al.* développent la théorie de l'« *Enaction* ». Cette théorie découle d'une conception de l'évolution du vivant en rupture avec l'évolutionnisme darwinien, fondée sur l' « *autoconservation individuelle* ». L'être vivant est présenté comme un système « *autopoiétique* », c'est-à-dire :

« organisé comme un réseau de processus de production de composants qui régénèrent continuellement par leurs transformations et leurs interactions le réseau qui les a produits, et qui constituent le système en tant qu'unité concrète dans l'espace où il existe, en spécifiant le domaine topologique où il se réalise comme réseau » [**16**].

L'élaboration d'un organisme vivant se conçoit alors comme le résultat d'une relation non-dualiste entre l'intérieur et l'extérieur : l'intérieur s'autorégule et s'accommode avec l'extérieur. Perception et action sont donc intimement liées : la perception est une activité interprétative (les actions modifient la perception sensorielle) ; et l'interprétation est forgée par la perception (la perception oriente la formation des structures cognitives). La cognition est une action « *corporalisée* » : elle est déterminée par l'expérience corporelle et cette expérience est rendue possible grâce aux capacités sensori-motrices du corps : on parle d'une « *boucle sensori-motrice* ».

A cet égard, citons également les remarques de Shove et Repp [**17**] relatives à la perception musicale, selon lesquelles des relations aussi abstraites et métaphoriques que celles qui sont en jeu dans la perception d'une œuvre (structure, émotions…), en viennent à être représentées par le cerveau humain comme un simple signal visuel, dans l'espace, du fait de la constante interaction de l'appareil spatial du cerveau avec le monde physique concret.

Dans [**18**], Sha *et al.* présentent des dispositifs où ils invitent les utilisateurs à manipuler des textures en 2D régies par des modèles physiques thermodynamiques, en suivant leur intuition relativement à des matériaux physiques comme l'eau, l'encre ou la fumée. L'idée est de mettre en évidence le fait que la richesse d'utilisation des interfaces homme-machine provient de l'expérience corporelle ; « *l'intuition kinesthétique* » vient en complément des aspects cognitifs mis en jeu. Cette intuition kinesthétique n'est pas sans rappeler la notion d'*empathie kinesthétique* développée par Jean-Luc Petit en référence au panpsychisme de Theodor Lipps [**8**].

Dans [**19**], Schiphorst rend compte de l'*expérience soma-esthétique* des différentes qualités du toucher, et propose une étude pluridisciplinaire méliorative de l'expérience du corps vivant (soma) comme source d'appréciation sensori-esthétique. Richard Shusterman [**20**] fait de l'expérience soma-esthétique et de la conscience du corps un moyen en vue de la construction de soi. Selon lui, les représentations du corps en vigueur dans la société sont externes, et s'appuient trop souvent sur des critères de beauté stéréotypés, qu'il n'hésite pas à décrire comme oppressifs. Concevoir le corps comme intentionnalité consciente pouvant connaître et approfondir son expérience interne pour réguler ses stimulations sensorielles est un véritable enjeu, dans une société où les informations affleurent aux portes de notre perception aussi nombreuses qu'intenses.

Dans [**21**], Petersen *et al.* se proposent d'intégrer l'esthétique au sein des critères de fonctionnement d'une IHM, que sont l'ergonomie, l'efficacité, ou la clarté. Ils reprennent à leur compte le concept d'« *interaction esthétique* » défini par Richard Shusterman dans [**22**]. Ils défendent l'idée que la vision analytique des interactions entre les hommes et les machines tend généralement à placer l'interaction avec un objet, une œuvre, ou un dispositif, au-dessus voire en dehors de la vie quotidienne ; l'objet existe de lui-même et est simplement porteur d'attributs. Au contraire, une *approche pragmatique* de l'interaction avec la machine doit tenir compte d'éléments socioculturels que l'interaction rend visible, et se rapprocher davantage des besoins, des désirs, des peurs, des espoirs des individus, tenir compte du fait que l'appréhension, l'appropriation et l'usage des objets sont déterminés et structurés par des contextes socio-économiques et politiques. Les technologies affectives sont alors envisagées d'un point de vue *constructiviste* : il ne s'agit plus d'obtenir des retours relativement à la performance d'un produit, mais d'étudier la prise en main de l'utilisateur et l'expérience que cela dessine. Le design a alors pour fonction de donner aux individus l'opportunité de s'approprier les objets et de modifier leurs fonctionnalités premières, d'une part en accordant un rôle primordial à l'improvisation et à l'exploration, d'autre part en usant de nouvelles représentations physiques ou symboliques complexes des actions. L'utilisateur peut en quelque sorte « s'améliorer » dans son usage des artefacts – ceux-ci ne sont plus réductibles au rôle de simples outils – au sein d'un intrigue en constant remodelage. Petersen *et al.* proposent plusieurs expériences : l'une d'entre elles concerne l'écoute de la musique et propose de naviguer d'une piste à l'autre ou de gérer le volume à travers des mouvements corporels relatés par des capteurs cinétiques ou sonores ; une autre propose un environnement hybride où des jets de balles sur des objets virtuels permettent de déplacer des documents d'une surface à une autre (table, mur, sol…).

Figure I.2 *J'efface vos traces* : une oeuvre de l'artiste chinois Du Zhenjun (2001).

L'approche « expérience utilisateur » (« *user experience* ») intéresse également le monde de la création. Avec l'*art expérientiel*, l'artiste agit sur le corps du visiteur lui-même pour l'impliquer public dans la découverte de l'œuvre. *J'efface vos traces* est une œuvre de l'artiste Du Zhenjun (Figure I.2) parfaitement représentative d'un tel projet : la présence d'un visiteur y déclenche des vidéos d'hommes nus à même le sol qui effacent les traces virtuelles dudit visiteur. Le visiteur, prenant conscience de son incidence sur le contenu même de l'œuvre, en devient partie intégrante et improvisateur. Dans [**23**], Françoise Lejeune introduit même la notion de « *somagraphie* » pour désigner des *« scénographies plasticiennes spécifiquement destinées à l'éveil sensoriel et à l'augmentation de l'attention accordée […]*

au corps sentant ». De tels dispositifs visent à faire rompre le public d'avec ses automatismes moteurs. Lejeune parle même d'un « piège sensoriel », dont le but est de favoriser l'ouverture du corps pour y inscrire l'œuvre. Les tenants de cette forme d'art qualifiée d'« expérientielle » font leur la soma-esthétique dite « expérientielle », désignée par Richard Shusterman dans *Conscience du corps pour une soma-esthétique* [**24**] comme celle qui développe l'attention portée aux sensations du corps.

Dans [**25**], S. Alaoui présente un travail sur le mouvement dansé en s'appuyant sur les concepts proposés par Rudolf Laban, dite « *Laban Movement Analysis* » (*LMA* [**26**] [**27**]). Une étude approfondie de la LMA fera l'objet de la section II.3 de ce manuscrit. Les qualités de mouvement définies par Laban caractérisent le mouvement corporel tel qu'il est produit par sa dynamique et sont indépendantes de sa trajectoire dans l'espace. L'approche proposée se situe dans le droit fil de la « *danse augmentée* » (Figure I.3). L'auteure se propose de modéliser les dynamiques du mouvement dansé par des systèmes de « masses-ressorts » dans l'optique de l'élaboration d'une plateforme de réalité virtuelle augmentée. Par ailleurs danseuse, elle définit avec son équipe un certain nombre de gestes expressifs que sont l'inspiration (*Breathing*), le saut (*Jumping*), l'extension (*Expanding*), et la réduction (*Reducing*) chacun étant par la suite décliné en différentes nuances. Pour ces gestes, il s'agit d'estimer les paramètres d'une équation de mouvement modélisant le système masse-ressort, dans l'optique de fournir au danseur des retours visuels sur son activité au sein d'un dispositif interactif.

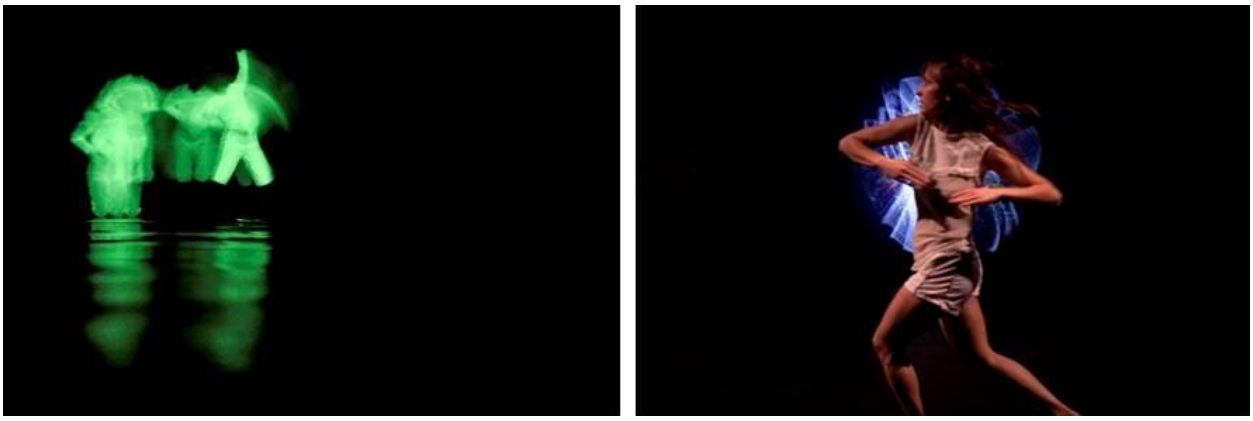


Figure I.3 Cycle *Danses Augmentées*, proposé par la chorégraphe et artiste plasticienne Mylène Benoit à la Gaîté lyrique en 2014 (a) ; extrait d'un spectacle proposé par la compagnie de danse Aux Pieds Levés et Un Des Sens à la bibliothèque municipale de Lyon en d'octobre 2013 à mars 2014 (b).

Les domaines de la danse et la musique, et plus généralement l'art, offrent d'ailleurs des pistes de développement intéressantes pour l'analyse des relations entre corporéité et subjectivité. L'exploration sensorielle y investit souvent les ambiguïtés de la relation du son au mouvement. Cela a fait l'objet de nombreuses études et expérimentations, dont certaines sont présentées dans la section suivante.

I.3.2. Synesthésie et interactions performatives

La phénoménologie de la perception offre un cadre théorique idéal pour des approches synesthésiques. Merleau-Ponty [**11**] décrit ce rapport synesthésique pour ce qui est du son et du mouvement :
« Le son d'un instrument à vent porte dans sa qualité la marque du souffle qui l'engendre et du rythme organique de ce souffle, comme le prouve l'impression d'étrangeté que l'on obtient en émettant à l'inverse des sons normalement enregistrés. Bien loin d'être un simple « déplacement », le mouvement est inscrit dans la texture des figures ou des qualités, il est comme un révélateur de leur être. Il y a, comme on l'a

dit, un espace et un mouvement « sensibles au cœur », prescrits par la dynamique interne du spectacle, et dont le changement de lieu est l'aboutissement ou l'enveloppe. »

A l'occasion de travaux sur la relation son-mouvement (Figure I.4), Robert Francès [**28**] formule l'idée que *« l'expressivité musicale est liée à l'acquisition progressive – à travers les messages sensoriels du corps et soulignés le plus souvent par la vue – des notions de mouvements, d'attitudes corporelles, d'effort, de relâchement, de vitesse, de rythme »*, et qualifie de tels éléments d'« *abstraits sentimentaux* ». Ainsi, l'expérience corporelle forge des « modèles temporels naturels » ayant des correspondances avec les « figures temporelles des musiciens ». C'est dans cet état d'esprit qu'ont été conceptualisées les « unités sémiotiques temporelles », en lien avec l'univers de la musique contemporaine [**29**]. Il s'agit d'éléments sonores musicaux supposés porter en eux même un contenu sémiotique indépendant du contexte de leur apparition, sous-tendant l'idée que le mouvement préside toujours à une impression musicale. Les classes formées par de tels éléments sonores empruntent leur vocabulaire métaphorique à l'analyse du mouvement : « *flottement* », « *freinage* », « *lourdeur* », « *qui tourne* », « *stationnaire* », « *suspension* », etc.

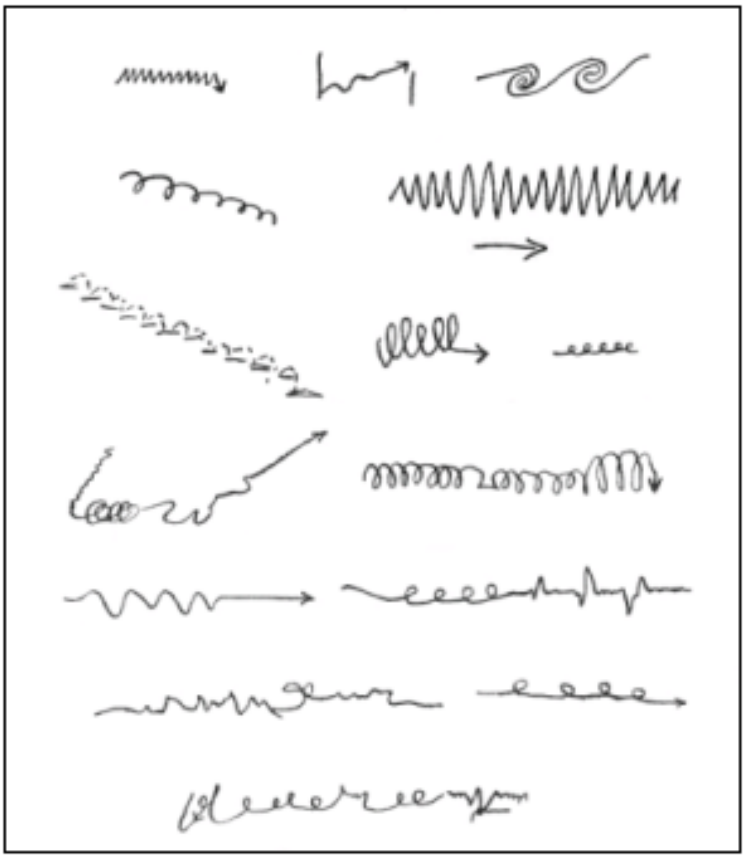


Figure I.4 Expérience menée par R. Francès [**28**] : dessins réalisés par des sujets à l'audition d'un fragment musical d'une œuvre de Debussy (*Images, Mouvement*, mesures 1 à 31).

Dans *Musical motion and performance : theoritical and empirical perspetives* [**17**], Patrick Shove et Bruno H. Repp évoquent, en s'appuyant sur les travaux du psychologue cognitiviste Roger Shepard, l'inscription dans l'espace de relations et de transformations abstraites ou métaphoriques, par l'appareil spatial du cerveau. Des mouvements mélodiques nous apparaissent dès lors, tels des « images auditives », bien qu'existant uniquement dans le temps et donnant d'ailleurs au temps son importance. Ils reprennent à leur compte l'affirmation du musicologue Gustav Becking selon laquelle la musique est habitée par un « flux dynamique rythmique », vertical, qui relie des points spécifiques dont les poids relatifs diffèrent.

Becking se propose de décrire les profils dynamiques correspondant à la musique de différents compositeurs en faisant l'acquisition de tels poids : cette acquisition est fondée sur l'accompagnement corporel des mouvements de la musique à l'aide d'une baguette du même type que celle d'un chef d'orchestre. Il parvient alors à caractériser différentes « attitudes philosophiques » vis-à-vis de la réalité physique, qui se traduisent par différents rapports à la gravité, en tant que celle-ci doit être surmontée, qu'il faut s'y adapter, ou qu'il faut la nier.

Comme l'écrit Claude Cadoz [**30**], le « *geste musical* » est autant un moyen d'action sur le monde physique qu'un moyen de communication informationnel. Ainsi, un geste musical comprend une dimension physique, régie par des éléments physiques moteurs, des aspects cinétiques, objectifs, des contraintes, et une dimension psychologique, intentionnelle, subjective. La trace du geste témoigne de son caractère contraint, limité. Elle montre aussi la possibilité d'un écart par rapport à une trace « normative », qui peut exprimer la liberté. Richard Shusterman exprime également cela à propos des artistes contemporains, et de la tendance très répandue de l' *« injection des modes de vie dans l'expression artistique »* [**20**] :

« Les codes font partie intégrante de la fonction communicative de l'art, sans laquelle il n'aurait pas de sens. […] La créativité réside dans l'interprétation et l'usage de ces codes tout autant que dans la rupture avec certains d'entre eux. »

Le musicien ou le danseur tiennent le premier rôle de cette créativité gestuelle. Mozart prétendait déjà qu'*« une symphonie est quelque part un opéra »* ; la composition d'œuvres musicales dessine alors un écart par rapport à la voix, qui fait figure de medium idéal au plus près de l'intériorité. L'instrument de musique est alors vécu comme un prolongement du corps, et les limites physiologiques que suscite son emploi fixent un univers de sons et de phrasés à explorer, d'où jailli l'intersubjectivité.

Figure I.5 *L'Intrus* (2004) : pièce conduite par Jean-Marie Adrien mettant en jeu la direction musicale dansée.

Dans [**31**], Jean-Marie Adrien introduit la « *Direction Musicale Dansée* » (D.M.D. : Figure I.5). Selon lui, c'est d'une intuition liée à la corrélation naturelle entre son et mouvement – du fait que le son est créé par une variation de vitesse et de pression – que provient l'empathie artistique. Cette empathie se manifeste par le lien qu'établit un chanteur dansant avec son public, ou encore lorsqu'un chef d'orchestre dirige ses musiciens. Dans ce dernier cas, cependant, la causalité et l'anticipation jouent un rôle

important. Le public d'un concert perçoit un tel contrat implicite entre le chef d'orchestre (siège du mouvement) et les musiciens (siège du son), et il se retrouve « pris » dans cette relation conventionnelle. Le son (du musicien) est un mouvement frustré, en tension, qui suscite sa propre libération, tandis que le mouvement (du chef d'orchestre) dessine des formes, suggère ce qu'il ne peut produire comme son, car il est emprisonné dans le corps. Ainsi, les deux entités que sont le musicien et le chef sont dans une lutte induite par la dualité son-mouvement. Il s'agit pour le chef d'orchestre de sculpter les motifs rythmiques des sons, notamment pour donner lieu à des contrastes (qu'il s'agisse de contrastes de nuance ou de tempo). En aucun cas il ne se contente de battre la mesure. Selon Jean-Luc Petit [**8**], le système de communication entre le chef et l'ensemble qu'il dirige est basé sur des « *noèmes* », c'est-à-dire des entités intentionnelles de conscience. De tels noèmes musicaux sont des entités objectives de la culture, et s'incarnent dans l'*empathie kinesthétique* évoquée à la section I.2. En règle générale, la gestuelle du chef d'orchestre est guidée par sa conscience musicale. Il se doit d'intégrer chaque élément de la musique à venir, pour l'incarner, en combinant cohérence et expressivité (le son conduit le mouvement), de façon à polir le son et obtenir la musique qu'il souhaite entendre (le mouvement guide alors le son). Alors seulement, le chef d'orchestre *est* la musique. Le mouvement est le son. C'est de l'idée de l'existence d'un matériau intelligible en amont de la musique que Jean-Marie Adrien appelle un *phrasé* [**32**], qu'est né le concept de « Direction musicale dansée », selon lequel le chef dirige et danse en même temps. Avec la « *Captation Gestuelle Causale* » (C.G.C. [**31**]), une limite supplémentaire est franchie : le son est produit directement à partir du geste capturé.

La relation son-mouvement est également étudiée par Caramiaux *et al.* dans [**33**], mais cette fois-ci à des fins d'indexation et de synthèse du son. Les auteurs s'intéressent à la corrélation entre la morphologie de descripteurs gestuels et celle de descripteurs sonores. Ils développent deux applications : un système permettant de sélectionner un son à partir du geste d'un utilisateur sensé représenter ce son abstrait, et une autre application visant à re-synthétiser un son préenregistré sur la base du geste de l'utilisateur.

Pour illustrer une fois de plus, l'idée d'une créativité par exploration corporelle normée, nous terminerons avec l'exemple d'un peintre à l'exécution, où la créativité semble jaillir comme un écart en lequel consiste l'exploration du medium. C'est ce que le peintre et physicien Jacques Mandelbrojt exprime lorsqu'il affirme que *« dans une peinture, le trait reproduit moins l'objet que le mouvement physique ou spirituel par lequel le peintre en prend connaissance »* [**34**]. On retrouve ici l'idée propre à la phénoménologie selon laquelle une subjectivité peut réellement être forgée par son objet.

A la lumière de ces considérations de haut niveau, nous pouvons constater que le domaine du geste a vu fleurir de récentes contributions ces dernières décennies qui mêlent la psychologie, l'esthétique, l'art, ou encore les neurosciences. Soucieux de nous pencher à notre tour sur de tels rapports entre corporéité, subjectivité et interaction sociale, nous nous sommes proposé d'établir un modèle de description du mouvement corporel qui tente d'intégrer les caractéristiques intentionnelles et intersubjectives qui régissent la réalisation d'un geste quel qu'il soit.

I.4. Objectifs et contributions

I.4.1. Objectifs

Les recherches sur le geste manquent aujourd'hui encore de modèles mathématiques unifiés et génériques. Les approches existantes font trop souvent dépendre les descripteurs du mouvement corporel de la spécificité des contenus à analyser. Dans ce travail de thèse, nous nous proposons d'étudier

comment il est possible de caractériser des contenus haut-niveau comme des états affectifs, des émotions, mais également de simples actions, à partir de représentations intermédiaires du geste qui tiennent compte de ce que nous désignons comme la dimension intentionnelle et communicative du geste. Il s'agit de dégager des indices de l'expressivité et de l'exploration expérimentale de l'espace qu'opère le sujet, davantage que des profils précis ou des trajectoires fixées dans l'espace et le temps. Comme évoqué à plusieurs reprises, le geste est en lutte contre ses déterminations, et c'est bien le contenu de cette lutte qui nous intéresse [**17**] [**30**] [**31**] [**34**].

La musique ou la danse sont des cadres idéaux pour poser la problématique du geste. Elles réunissent des problématiques diverses : structuration, corporéité, exploration de l'espace, relation son-mouvement, communication, expressivité, émotions. Notre intérêt s'est porté en particulier sur la corrélation entre gestuelle et émotions musicales. Dans [**17**] est présenté une approche que développe le musicologue Eric Clarke relativement à la perception musicale. Selon lui, l'écoute se décompose en trois niveaux de perception de l'évènement musical. Un premier niveau, écologique, a trait à la performance, c'est-à-dire à l'acte de jouer de la musique et aux objets impliqués dans la production du son. Un second niveau concerne la structure de la musique (mélodie, harmonie, rythme, phrases, forme, etc.). La troisième couche traite de l'aspect expressif de la musique, des caractéristiques similaires aux mouvements de l'âme (et donc du corps) généralement qualifiés d'émotions, d'humeurs ou de sentiments. Une telle décomposition de la perception musicale, tournée vers des contenus toujours plus abstraits rend plausible l'idée d'une description d'échantillons musicaux à l'aide d'émotions plus ou moins quotidiennes. La légitimité d'une telle caractérisation est par ailleurs confirmée par le travail entrepris par Grewe *et al*. [**35**] qui demandent à des participants de caractériser des morceaux de musique classique à partir d'un lexique d'émotions. Par ailleurs, plusieurs chefs d'orchestre et musiciens, dont certains professionnels et intéressés par la transdisciplinarité entre la musique et les sciences sociales, nous ont confirmé la pertinence d'un tel projet. En effet, la caractérisation de gestes musicaux (gestes dansés, mouvements de chefs d'orchestre, mouvements de musiciens en concert) à l'aide de catégories référant à des émotions musicales, constituerait selon eux une avancée dans l'appréhension d'un langage émotionnel sous-jacent à la pratique de la musique, qui plus est si la possibilité s'offrait dans le futur de ramener la reconnaissance de telles émotions musicales à la perception d'indices gestuels précis.

Nos contributions sont résumées dans le paragraphe suivant.

I.4.2. Contributions

I.4.2.1. Descripteurs expressifs du geste

De façon heuristique, nous avons défini un ensemble de descripteurs expressifs du geste, dédiés à la quantification de concepts extraits de l'étude du mouvement que Rudolf Laban consacre aux gestes dansés. Cette analyse du mouvement expressif proposée par Laban (*Laban Movement Analysis* : LMA [**26**] [**27**]) fait l'objet de la section IV, où nous expliquons notamment pourquoi nous avons choisi d'étendre une telle caractérisation du mouvement dansé à la description générique du geste.

L'approche développée nous fournit donc des valeurs numériques que nous calculons à partir de positions 3D de référence du corps (les poses dans l'espace sont fournies à chaque instant par une Kinect), et ce dans le but de les exploiter directement en entrée d'algorithmes d'apprentissages supervisés pour la reconnaissance de contenus haut-niveau, comme des actions, des émotions, des états affectifs…. L'objectif est de caractériser ces contenus haut-niveaux sans chercher à évaluer explicitement les qualités de Laban sous-jacentes à partir desquelles les valeurs numériques ont été bâties.

I.4.2.2. Constitution de bases de données

Nous avons constitué deux bases de données. Une première, *ORCHESTRE-3D*, est dédiée à des gestes de chefs d'orchestre pré-segmentés, enregistrés lors de répétitions d'orchestre. Ces gestes ont été annotés par des praticiens de la musique à partir de catégories d'émotions musicales dont le lexique a été défini en concertation avec des musiciens, chefs d'orchestre, et musicologues, dans la perspective d'une appréhension d'un langage émotionnel sous-jacent à la direction orchestrale. Un deuxième corpus (*HTI 2014-2015*) dédié à des actions de la vie courante a été formalisé et constitué par des étudiants de la filière High-Tech Imaging (HTI) de TELECOM SudParis pour la saison 2014-2015, mettant en jeu 11 actions différentes dans des séquences gestuelles réalisées par une dizaine de participants. Ces corpus seront décrits en détail dans la section V.

I.4.2.3. Reconnaissance globale

Dans une première approche traitée à la section VI, dite « de reconnaissance globale », nos descripteurs de geste inspirés par l'analyse du mouvement de Laban sont utilisés dans la perspective d'une description globale du geste. L'analyse de chaque geste donne donc naissance à un vecteur descripteur dédié à l'entièreté de l'empan temporel du geste. Nous testons la pertinence de notre modèle global au travers de deux expériences : la première est dédiée à la reconnaissance d'actions et s'appuie sur le corpus *Microsoft Research Cambridge-12 Kinect gesture dataset* (*MSRC-12* [**36**]) ; la seconde traite de la reconnaissance de catégories émotionnelles et met en jeu notre base de chefs d'orchestre. Au cours de ces expériences, nous utilisons nos descripteurs de différents algorithmes d'apprentissage, essentiellement basés sur des séparateurs à vaste marge (*Support Vector Machine* ou *SVM*) et des forêts d'arbres décisionnels (*Random decision forest*).

I.4.2.4. Reconnaissance en temps réel

Nous adoptons ici un cadre de reconnaissance dynamique : les indices qui avaient servi à élaborer un descripteur global du geste, à la section précédente, sont désormais utilisés pour constituer un vecteur descripteur par trame. Ainsi, chaque instant du mouvement se retrouve décrit par une série de valeurs. A partir du nouvel espace vectoriel dans lequel chaque instant gestuel est représenté, nous construisons un dictionnaire de poses-clés, que nous extrayons depuis le corpus d'apprentissage, et qui nous servent de référence pour échantillonner le mouvement. Chaque instant gestuel peut dès lors se voir projeté sur un certain nombre de ces poses représentatives de la base de données, et acquérir ainsi une représentation réduite ; c'est l'objet de l'algorithme d'« affectation douce » (ou *soft assignment* dans la terminologie anglaise) que nous utilisons. Cette représentation des gestes par soft assignment est ensuite utilisée pour nourrir des chaînes de Markov cachées (*Hidden Markov Models* ou *HMM*) et assigner une action à chaque instant du geste. L'utilisation des bases de gestes *MSRC-12*, *UTKinect-HumanDetection* [**37**] et *MSR Action 3D* [**38**] nous permet de nous confronter à d'autres approches qui les utilisent.

Avant de nous positionner par rapport à l'état de l'art, et de présenter notre approche, nous avons consacré le chapitre suivant à une présentation des différents modèles de geste et d'émotions de la littérature.

II. Gestes, expressivité, émotions, musique

Dans ce chapitre, nous nous plaçons dans la perspective d'une modélisation du mouvement corporel. Nous soulignons qu'une telle modélisation nécessite la prise en compte de la relation entre corporéité et subjectivité. En premier lieu, nous présentons les taxonomies classiques du geste et soulignons leur incapacité à prendre en compte ses dimensions expressive et communicative. Nous nous penchons ensuite sur différentes modèles d'émotions provenant de sciences humaines et mettons en lumière certaines incitations à ramener les représentations de l'affect à l'expérience du corps et à la culture. Enfin, nous introduisons le modèle d'analyse du mouvement dansé proposé par Rudolf Laban, que nous avons adopté pour la suite de nos travaux. Ce modèle de l'expressivité du geste intègre des dimensions du mouvement autant corporelles que subjectives, et nous paraît être le cadre théorique le plus élaboré et le plus consistant à cet effet.

Dans cette seconde partie, nous souhaitons spécifier à quels impératifs la relation corps-sujet nous oblige à répondre dans la perspective d'une modélisation du geste. En tant que rapport conscient d'un individu à son propre corps, c'est en effet bien le geste qui médiatise la relation de ce corps à la subjectivité, et en particulier à la psyché. Deux objectifs émergent :

1. il faut se pencher sur la relation *corps-geste* et traiter de ses liens avec la subjectivité ;
2. puis évaluer un second rapport *geste-émotion* à l'aune de la corporéité.

Les relations *corps-geste* et *geste-émotion* sont généralement décrites sous l'égide de la détermination linguistique. Il s'agira pour nous de nous inscrire dans une critique des représentations conventionnelles du geste pour faire surgir la nécessité d'y prendre en compte la dimension affective et empathique, ainsi que de ramener les représentations émotionnelles génériques à l'expérience corporelle et plus généralement à la pratique et à la culture.

II.1. Du geste fonctionnel au geste expressif

De façon générique, Kurtenbach et Hulteen définissent le geste comme « *un mouvement du corps qui contient de l'information* » [**1**]. Dans un premier temps, citons les taxonomies de référence où les gestes sont classés d'un point de vue *fonctionnel*.

II.1.1. Analyses fonctionnelles du geste

Dans son analyse du geste instrumental des musiciens (relation musicien-instrument), Claude Cadoz [**39**] définit trois niveaux de gestualité.

- Un premier niveau a trait à l'action de modifier ou transformer l'environnement matériel. C'est la fonction ***ergodique*** du geste (exemple : appuyer sur un bouton pour lancer la musique).
- Le deuxième, dit ***épistémique***, désigne l'exploration empirique et perceptive par le toucher : informations relatives à la température, au type de surface, à la mollesse, à la forme, au poids…

Ces deux premiers niveaux définissent les gestes de *manipulation.*

- Enfin, la dernière fonction gestuelle concerne l'émission d'information à destination de l'extérieur, c'est-à-dire la *communication*. Elle est qualifiée de ***sémiotique***.

La classification d'Adam Kendon [**40**] est essentiellement dédiée à la *gestualité coverbale*, c'est-à-dire aux gestes d'accompagnement du discours, et au langage des signes. Kendon insère les phénomènes gestuels dans un continuum (Figure II.1) qui va des gestes les moins conventionnels, c'est-à-dire les plus *idiosyncrasiques* et spontanés – et pour la compréhension desquels l'accompagnement par la parole est essentiel – aux plus *linguistiques* – pour lesquels la parole est quasiment optionnelle voire inutile. Les gestes vont ainsi du ***complément spontané*** (par exemple de type rythmique – on imagine un conférencier accompagnant une énumération par des gestes circulaires du bras) à la substitution de la parole que constitue la ***langue des signes***, en passant par le geste ***langagier***, par le ***pantomime*** – geste d'imitation ou de simulation pour illustrer l'utilisation virtuelle d'objets –, et par l'***emblème*** – qui désigne un modèle sémantique partagé (par exemple : dire au revoir d'un signe de la main, lever le pouce en guise de félicitation).

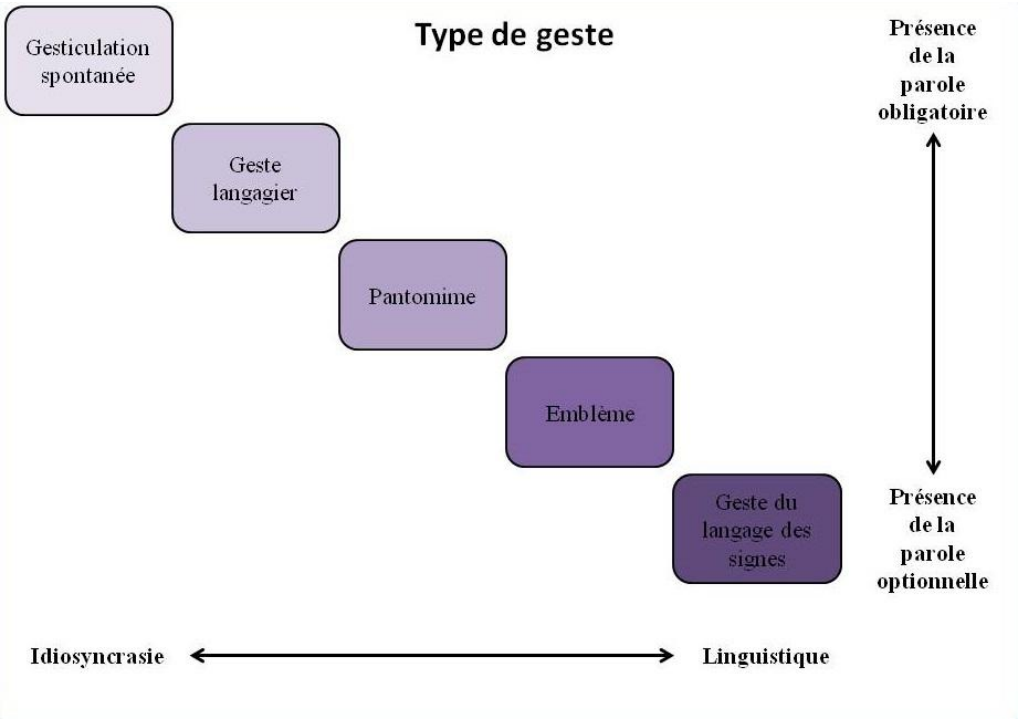


Figure II.1 Le continuum de Kendon [**40**].

Le travail de David McNeill [**41**], également dédié à la gestualité coverbale, s'inspire grandement du continuum de Kendon [**40**]. McNeill propose une taxonomie des gestes communicatifs des bras.

- Les gestes dits « ***imagés*** » relèvent du pur domaine de la représentation. Dans ce cadre, on retrouve :
    - Les gestes ***iconiques*** ou ***connotatifs***, qui sont ceux qui représentent un objet, par des évocations de forme, d'extension spatiale, de taille, d'orientation, ou une action concrète. Ils consistent à décrire un élément à partir de ses caractéristiques.
    - Les gestes ***métaphoriques*** représentent eux des idées abstraites (exemple : se prendre la tête à deux mains pour protester contre la musique).
- Les gestes ***non-imagés*** se subdivisent également en deux catégories.
    - Les gestes ***déictiques*** sont démonstratifs et désignent du doigt un point de l'espace en référence à des gens ou à des objets situés dans l'espace.
    - Les gestes de ***pulsation*** qui accompagnent rythmiquement la parole (pour accentuer de mots, des prises de parole…).
- Les ***emblèmes*** qui correspondent au type du même nom dans la taxonomie de Kendon (Figure II.1).

Le caractère prétendument exhaustif des taxonomies du geste rend possible des passerelles entre elles. Ainsi les propositions de Kendon et McNeill, ayant trait à la gestualité coverbale, classifient des gestes que l'on identifie avec Cadoz comme *sémiotiques*. Cette catégorie est très foisonnante, dans la mesure où elle a pour enjeux les diverses dimensions de la communication. Les gestes de *pulsation* de Kendon sont un exemple de *gesticulation spontanée* telle que définit par McNeill. Les gestes iconiques de ce dernier sont partie intégrante des gestes *langagiers* que propose Kendon.

Remarquons que les approches évoquées exploitent le concept d' « *image-schéma* », développé par Johnson dans [**42**], qui interprète les formes élaborées au cours de la production gestuelle comme des

mimiques se référant à des notions partagées dans la culture. Selon lui, la plus grande part de la communication relative à des concepts abstraits est rendue possible par l'utilisation de métaphores et de métonymies.

Dans le droit fil de cette perspective, Boyes Braem and Bräm [**43**] proposent une classification des gestes effectués par les chefs d'orchestre en se focalisant sur les mouvements de la main « non-dominante » – celle qui ne tient pas la baguette, et n'est par conséquent pas dédiée à la pulsation. Cette main est communément utilisée pour suggérer aux musiciens des dynamiques particulières, des articulations entre les notes, des couleurs de sons particulières, l'entrée de musiciens ou d'autres évènements spontanés (comme un piano subito ou un accent fortement marqué). Ils tentent d'établir des ponts entre des images explicites formées par la main et les indications musicales qu'elles peuvent signifier. Par exemple, les mouvements dénotant le regroupement d'objets sont associés à une demande d'homogénéité de son, ceux qui réfèrent à la portée d'un objet traduisent une demande de soutien du son. Pour les auteurs, ces relations geste-son peuvent être décrites à l'aide de positions ou de mouvements de la main assez génériques.

Les classifications proposées ci-dessus ont donc trait à la *fonction* des gestes, ou au *but* poursuivi lors de leur exécution. Elles ne réfèrent à aucun *contenu* concret du mouvement. Or c'est bien en la qualification d'un tel contenu que consiste notre objectif d'analyse sémantique de la gestualité. Dans le paragraphe qui suit, nous montrons qu'une description générique du geste ne peut faire l'impasse sur les indices de l'« expressivité ».

II.1.2. Vers le geste expressif

Dans [**44**] Dominique Boutet propose une catégorisation des gestes à l'aune d'une physiologie articulaire basée sur des *degrés de liberté*. Il insiste sur le caractère compositionnel et non-linéaire de la production gestuelle. En dehors des emblèmes (*cf.* taxonomie de McNeill [**14**]), aucun inventaire lexical ne fournit selon lui les notions suffisantes pour étudier la production gestuelle et son caractère multilinéaire. Pour pallier ce manque, Boutet se propose d'aborder le geste d'un point de vue physiologique et morphologique. Dans un travail à portée diverse (langue des signes, gestualité co-verbale) sur la structuration articulaire du geste, Boutet insiste sur le caractère contraint du geste, sur la dépendance existant entre les articulations et les relations qu'entretiennent leurs divers degrés de libertés : parfois, certains degrés de liberté perdent leur indépendance sur d'autres. Il s'agit de phénomènes dits de *co-articulation* :

« Les mouvements s'opèrent naturellement par des contractions musculaires qui, faisant bouger le squelette, provoquent également des mouvements induits issus de forces cinétiques et inertielles qui sont transmises le long de segments. Ainsi, des forces sont transférées sur le même segment, ou sur un autre adjacent selon un vecteur somme (celui des forces en présence) qui tient compte aussi du cadre géométrique de motilité qu'imposent les degrés de libertés concernés. Si les forces ont des directions calculées par les mathématiques, les mouvements sont déterminés par la physiologie articulaire, leurs directions par ce que permettent les articulations ; de la sorte les transferts involontaires de mouvement obéissent à des règles issues des rapports géométriques parfois changeants qu'entretiennent les degrés de liberté entre eux. »

Dans [**45**], ce même auteur montre que la caractérisation d'unités gestuelles à partir de leur structuration physiologique permet d'isoler des gestes très proches par leur forme.

La conception du geste développée par Boutet dans [**44**] s'inscrit dans une critique des approches du mouvement corporel qui réduisent le corps à un simple support d'expression linguistique, et par conséquent font du geste une manifestation corporelle d'un concept fini et prédéterminé. Dans la mesure où les mouvements de certaines parties du corps induisent la dynamique d'autres, on peut dire que les unités gestuelles ont de véritables contours, et ces contours les *fondent*.

L'idée fondamental qui se dégage de cette analyse peut être résumée comme suit : **le corps n'est pas un simple support pour les gestes, c'est un *substrat* : il *produit* et *informe* les gestes en même temps qu'il les *révèle*, les porte et est utilisé par eux.**

Cette considération en appelle une autre : s'il est impossible au geste de s'incorporer sans se heurter aux limites du corps, il ne peut pas être la pleine traduction d'un concept initial qui se suffirait à lui-même et serait fixé dans la langue, puisqu'il est en même temps *informé* par le corps. Le geste « *prend connaissance* » de l'extérieur, que cet extérieur désigne un objet ou le corps lui-même [**34**]. Un tel constat est une invitation à rompre avec la position de surplomb que tient le langage dans l'explication de la structuration des gestes, pour donner la primauté à la dimension élaboratrice du corps.

Cette élaboration du geste *par* le corps est guidée par une expérience subjective qui présente des dimensions inconscientes, affectives et émotionnelles – le corps est le siège des émotions, et nous avons vu avec Damasio [**5**] que la carte du corps dans le cerveau est affectée par les émotions. Il faut d'ailleurs préciser que la dimension *sémiotique* à laquelle Claude Cadoz [**39**] fait écho dans sa classification prend en compte autant la communication stricte, que l'expressivité du geste et les indices pluriels de l'intentionnalité.

La direction orchestrale souligne particulièrement bien l'idée d'une élaboration subjective et émotionnelle (c'est-à-dire médiatisée par les émotions) du geste. Selon le violoncelliste et chef d'orchestre Xavier Gagnepain, avec lequel nous nous sommes entretenus en janvier 2013 à la suite d'une de ses répétitions avec un ensemble de jeunes instrumentistes en formation préprofessionnelle à Boulogne-Billancourt, la direction orchestrale procède de ce qu'il nomme « *le regard vers l'intérieur* ». Lorsque le chef d'orchestre voit venir le mineur, le majeur, il le voit venir par référence à une expérience passée, expérience qu'il va « *devenir* » pour le transmettre à autrui. A la section I.3.2, nous avions déjà évoqué le fait que le chef d'orchestre se doit d'intégrer chaque élément de la musique à venir, pour *être* la musique [**31**]. Le chef se chante quelque chose en essayant de s'en faire une image fidèle pour avoir une belle production. De ce « *chant intérieur* » surgit toute *une expressivité*. Lui incombe la tâche de préparer l'*intégration* de tous les éléments de la musique à venir – la note doit être teintée de l'harmonie, du timbre, de l'orchestration… –, y compris ceux qui peuvent être antagonistes (deux attaques différentes par deux instruments par exemple), et de peaufiner en temps réel son incarnation selon la convenance de la musique. Pour cette raison, les chefs d'orchestre sont accoutumés à un travail personnel de perfection de leur gestuelle et à une véritable *exploration émotionnelle* au cours de laquelle « ils s'imaginent dirigeant l'orchestre ». Ils travaillent le rapport à leur corps et construisent leur expressivité dans un va-et-vient entre l'extériorisation de leurs émotions dans le geste, et le remodelage constant de celui-ci.

L'émotion, qui n'est pas *en soi* un but pour un corps dirigeant de la musique, devient un but *pour soi* produit par l'expérimentation gestuelle. Réciproquement, cette émotion médiatise la production du geste par le corps.

Pour appréhender le geste, il s'agit donc de ramener la psyché dans le corps en mouvement, en examinant le lien entre gestualité et états affectifs. Afin d'approfondir ces aspects, intéressons-nous à présent aux divers modèles d'émotions proposés dans la littérature.

II.2. Modèles d'émotions

Le psycho-cognitiviste Klaus Scherer [**46**] regroupe différentes notions comme ayant trait aux émotions :

- les *émotions utilitaires* facilitent notre adaptation ou notre réaction à un évènement déterminant (exemple : la peur, la souffrance) ;
- les *préférences* sont des jugements relatifs à l'évaluation ou à la comparaison d'objets ;
- les *attitudes* sont des jugements, prédispositions ou croyances, relativement stables ;
- les *humeurs* sont des états affectifs diffus, dont la cause est difficilement identifiable ;
- les *dispositions affectives* ou *traits émotionnels* correspondent à des traits de personnalité, à des humeurs récurrentes (exemples : nerveux, jaloux, irritable) ;
- les *positions relationnelles* sont spontanées ou stratégiques (exemples : amical, froid, avenant) ;
- les *émotions esthétiques* sont désintéressées, non-utilitaires (exemples : fasciné, transporté, ému).

Dans [**47**], Eva Hudlika énumère ce qui selon elle constitue les modalités de l'émotion.

- modalité comportementale, expressive
- modalité somatique, physiologique
- modalité cognitive, interprétative
- modalité expérimentale, subjective (idiosyncrasique).

Pour aborder la complexité d'états affectifs divers, présentons les divers modèles émotionnels reconnus.

II.2.1. Catégories émotionnelles

Paul Ekman définit dans [**48**] les caractéristiques communes à ce qu'il appelle des *émotions* « *élémentaires* » ou « *de base* ». Ces caractéristiques sont énumérées ci-dessous :

1. une mise en jeu de signaux universels distinctifs, notamment faciaux – qui ont donné notamment naissance au système de codage des actions faciales FACS (*Facial Action Coding System*) [**49**] ;
2. la présence d'expressions comparables chez d'autres primates ; Charles Darwin considérait un tel élément crucial [**50**], dans la mesure où il était constitutif de sa théorie de l'évolution ;
3. une physiologie distinctive ;
4. des éléments d'amorçage universels : si l'on considère que les émotions ont changé selon les principes de l'adaptation évolutive et de la constitution du genre humain, il est cohérent d'imaginer qu'une même émotion apparaisse dans des contextes présentant des similarités ;
5. une cohérence notable entre la réponse expressive et la réponse nerveuse au cours de l'émotion ;
6. au début du phénomène émotionnel, une mobilisation rapide de l'organisme ;
7. une durée brève ;
8. un mécanisme d'évaluation automatique ;
9. une spontanéité de l'occurrence qui ne laisse pas de place au choix.

Les éléments 1, 3 et 4 permettent de distinguer une émotion d'une autre. Les autres sont sensés permettre de différencier les émotions d'autres phénomènes dits « affectifs ». Ekman définit alors un certain nombre d'émotions élémentaires que sont la *colère*, la *peur*, la *tristesse*, la *joie*, le *dégoût*, et la *surprise*, auxquelles il rajoute le *mépris*, la *honte*, la *culpabilité*, l'*embarras*, le *respect* (mêlé

d'admiration) en montrant que de telles dispositions affectives peuvent également satisfaire les critères énumérés. Il distingue ces émotions « *de base* » d'émotions « *secondaires* » comme l'*amour* ou la *nostalgie*.

Dans [**51**], Beller propose un modèle hybride qui rend compte de la plurivocité de l'état émotionnel, en permettant de sélectionner plusieurs étiquettes pour le décrire, ainsi que l'intensité des émotions, en proposant pour chaque catégorie émotionnelle concernée une valeur sur une échelle d'intensité de 1 à 5. L'émotion est donc représentée par un vecteur lexical. Une telle approche se rapproche de celle du psychologue américain Robert Plutchik, qui propose une représentation des émotions au sein d'une structure tridimensionnelle permettant de rendre de compte de différentes intensités au sein d'un même champ émotionnel (Figure II.2.a). L'idée d'« intensité émotionnelle » que met en œuvre le *circumplex* de Plutchik (Figure II.2.a) suggère une dimensionnalité, voire une continuité. Cela ouvre notamment la porte aux représentations dimensionnelles de l'émotion, rappelées dans le paragraphe suivant.

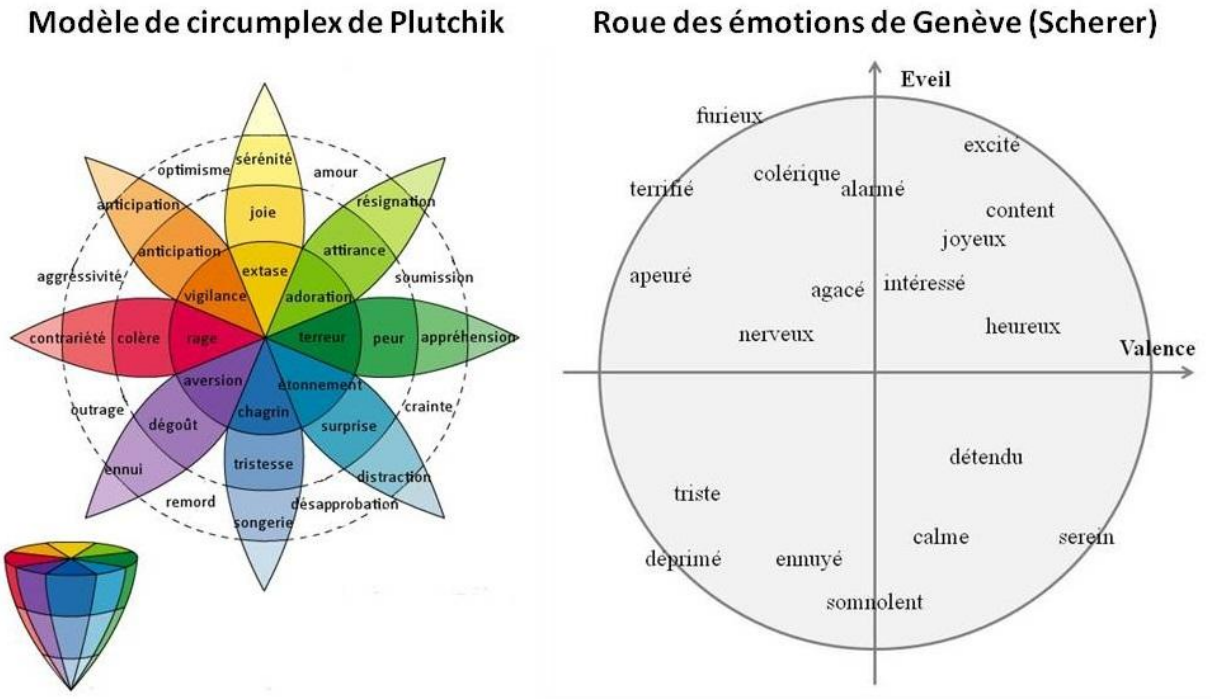


Figure II.2 Modèle de circumplex de Plutchik déplié en 2 dimensions (a) ; Roue des émotions de Genève (Scherer [**46**]) (b).

II.2.2. Représentations dimensionnelles des émotions

Les représentations dimensionnelles des émotions consistent à décrire les états affectifs comme des vecteurs dans un espace. La plupart des espaces considérés comprennent 2 ou 3 dimensions. La plus populaire est la représentation 2D introduite pour la première fois par Wilhelm Wundt [**52**] où les deux dimensions correspondent à :

- la *valence* (qui spécifie le caractère *positif*/*négatif*, *plaisant*/*déplaisant*) ;
- l'*éveil* (qui différencie les états d'*activation* des états de *repos*, les états d'*activité* des états de *passivité*, ou encore les états d'*excitation* des états *calmes*).

Remarquons que la notion de « *valence* », qui qualifie l'émotion selon son caractère positif ou négatif est déjà très importante pour Darwin [**50**], puisqu'elle renvoie à des réactions de survie.

Dans ce cadre, mentionnons l'outil « *FeelTrace* » de Schröder [**53**] qui permet d'annoter en continu des scènes ou des enregistrements en déplaçant un curseur dans l'espace bidimensionnel que constituent la valence et l'éveil. Notons également que le psychologue Wilhelm Wundt suggérait déjà au dix-neuvième siècle d'ajouter une troisième composante complémentaire à cette représentation, dite de *tension* et discriminant les états de tension des états de *relaxation* ou de *détente*.

La roue de Genève de Klaus Scherer (Figure II.2.b) propose une inscription des catégories émotionnelles dans l'espace valence-éveil.

Par ailleurs, il existe d'autres représentations dans lesquelles Scherer intègre deux autres dimensions dont les continuums représentatifs sont : *contrôle*/*soumission* et *propension*/*obstruction* [**46**].

La dimension de *contrôle* est parfois qualifiée de *domination*, comme dans le modèle de Mehrabian, où elle est associée aux dimensions de valence et à d'éveil pour constituer le modèle « *Pleasure-Arousal-Dominance* » (P.A.D.) [**54**].

De tels modèles dimensionnels présentent l'avantage de permettre l'évaluation des émotions selon différents niveaux d'intensité, et avec des possibilités de composition que ne permettent que rarement les catégories émotionnelles. Ils offrent d'ailleurs la possibilité de se focaliser sur des dimensions singulières de la réaction émotionnelle. Néanmoins, l'usage d'un nombre restreint de dimensions (comme dans le cas de l'espace valence-éveil) avec un échantillonnage grossier sur chaque dimension peut s'avérer insuffisant à discriminer des émotions « proches ». Ainsi, dans un espace valence-éveil échantillonné en trois valeurs {-1, 0, 1}, la *peur* et la *colère* seront toutes deux représentées par le vecteur (-1, 1). Pour ces raisons, dans [**55**], Devillers *et al.* proposent un nouveau protocole d'annotation combinant des représentations par catégories et par dimensions.

Dans [**56**], Hudlicka *et al.* soulignent les bénéfices et limitations des représentations continues des émotions en calcul affectif. Ils proposent de nouveaux arguments pour montrer l'importance de la dimension de *domination* du modèle *P.A.D.* [**54**]. Notamment, ils démontrent qu'elle permet de représenter les réactions d'approche, d'évitement, ou tout ce qui a trait à la *prise en main*. Dans cette optique, il est nécessaire de mettre en valeur des mécanismes sous-jacents de calcul affectif cognitif et notamment la « direction du but » et le « taux de sécurité ». Les émotions sont ainsi vues comme des méthodes globales perceptuelles qui modulent la cognition et d'autres comportements, ce qui nous amène aux théories cognitives.

II.2.3. Emotions et évaluation cognitive

Nous avons déjà évoqué la conception darwinienne des émotions comme des réactions prototypiques dont les fonctions sont liées à la survie [**50**]. En 1884, William James [**57**] met en valeur le lien entre le réflexe émotionnel et la perception de cette réaction qui selon lui constitue à proprement parler l'émotion. Il introduit déjà l'idée de processus d'évaluation cognitive.

La direction orchestrale et plus généralement le geste artistique nous invitent en outre à poser la problématique des émotions d'un point de vue « utilitaire » voir « stratégique ». Une distinction nette apparaît à cet égard entre les émotions *déclenchées* ou *suscitées*, qui sont le fruit d'une évaluation subjective, et les émotions *exprimées* qui obéissent, entre autres, à des règles conventionnelles.

Pour K. Scherer, la perception et l'évaluation cognitive d'un évènement déterminent le type et l'intensité de l'émotion ressentie par une personne [**46**]. Les théories cognitivistes font des émotions le fruit d'une évaluation cognitive d'un évènement selon un certain nombre de variables dites d'« *évaluation* ».

II.2.3.1. Le modèle « *OCC* »

En 1988, Ortony, Clore et Collins [**58**] segmentent les émotions en trois catégories :

1. celles qui relèvent de l'interprétation d'un *évènement* ;
2. celles qui sont liées à l'*action* d'individus ou de soi-même ;
3. celles qui résultent de l'évaluation d'un *objet*.

Par conséquent, le modèle *OCC* définit trois classes d'émotions (selon qu'elles sont liées à des évènements, des actions, ou des objets), chacune réunissant des émotions types par une même configuration du monde. Une *variable centrale* permet de déterminer le type et l'intensité de l'émotion créée selon les *buts*, les *principes* et les *préférences* de l'agent. Par exemple, la perception d'un évènement qui satisfait les buts d'un individu peut déclencher de la joie. L'action d'une personne qui entre en contradiction avec les principes moraux d'un individu pourra provoquer chez ce dernier de la colère. La présentation à un individu d'un met qu'il ne lui plait pas suscitera sans doute du dégoût. Par ailleurs, des *variables d'intensité* modélisent pour Ortony *et al.* modélisent l'influence de l'intensité réactionnelle. Les auteurs du modèle ainsi spécifient les états mentaux de catégories émotionnelles.

Dans [**59**], Valitutti et Strapparava tentent d'intégrer des éléments du modèle *OCC* dans leur perspective d'indexation de contenus émotionnels présents dans les contenus textuels. Outre cette intégration, ils insistent sur la notion de *hiérarchie affective*, qui permet la structuration des classes et sous-classes d'émotions de façon arborescente (exemple : la joie est ainsi présentée comme une instance d'une joie plus générique, elle-même comprise parmi la classe des émotions « positives »). Ils retiennent également la qualification émotionnelle en termes de *valence*.

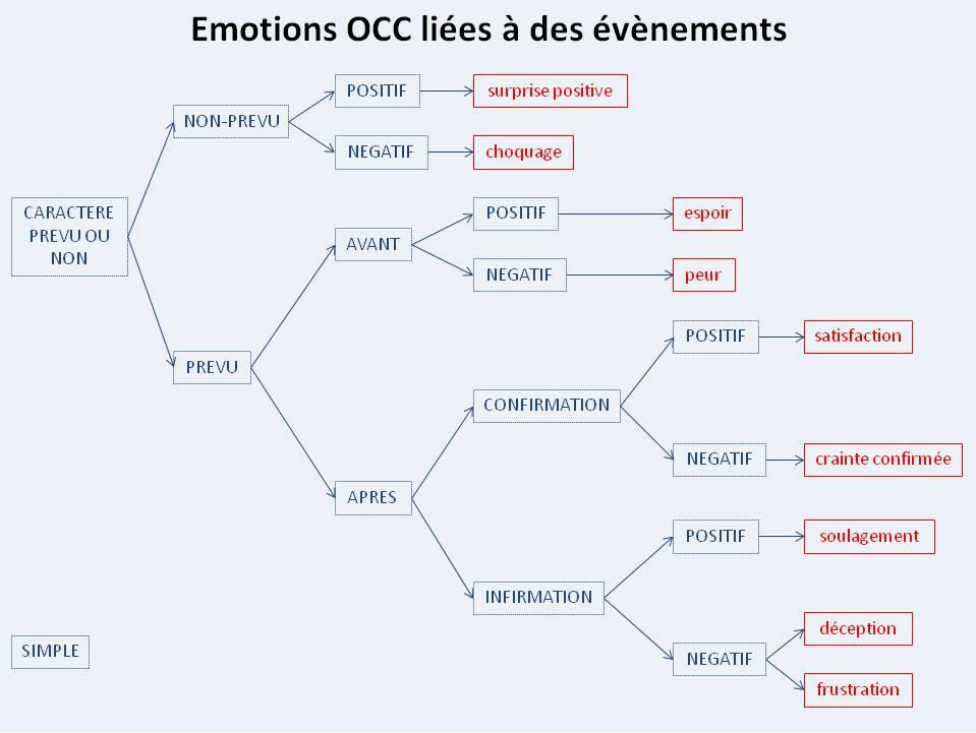


Figure II.3 Hiérarchie des émotions relevant de l'interprétation d'un évènement selon Valitutti et Strapparava [**59**].

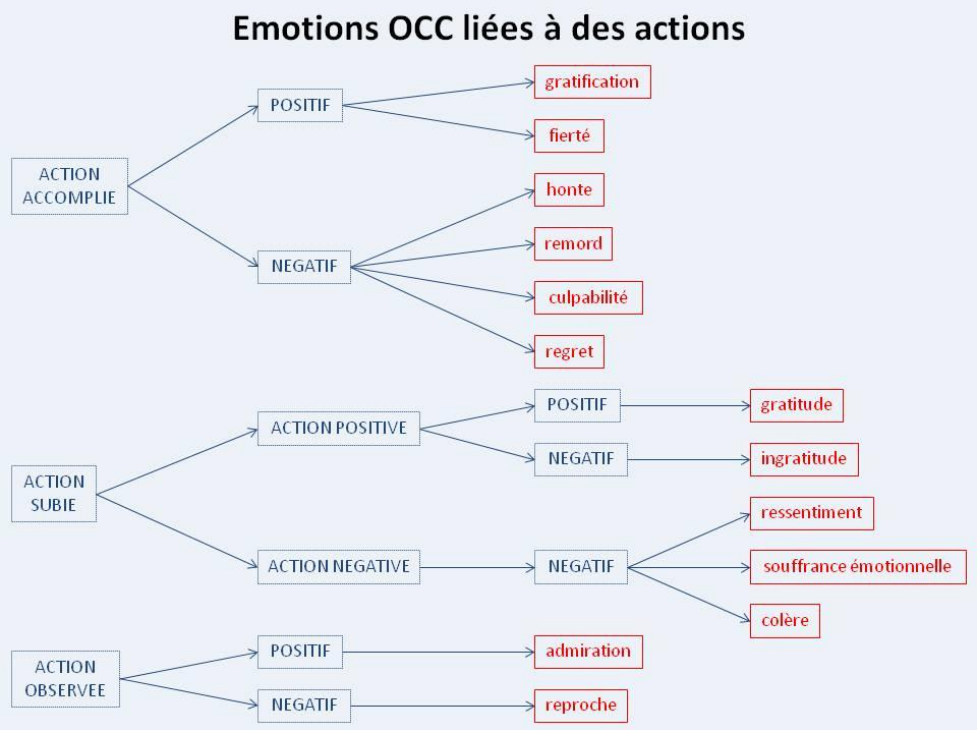


Figure II.4 Hiérarchie des émotions liées à une action de l'individu, sur l'individu, ou observée par lui, selon Valitutti et Strapparava [**59**].

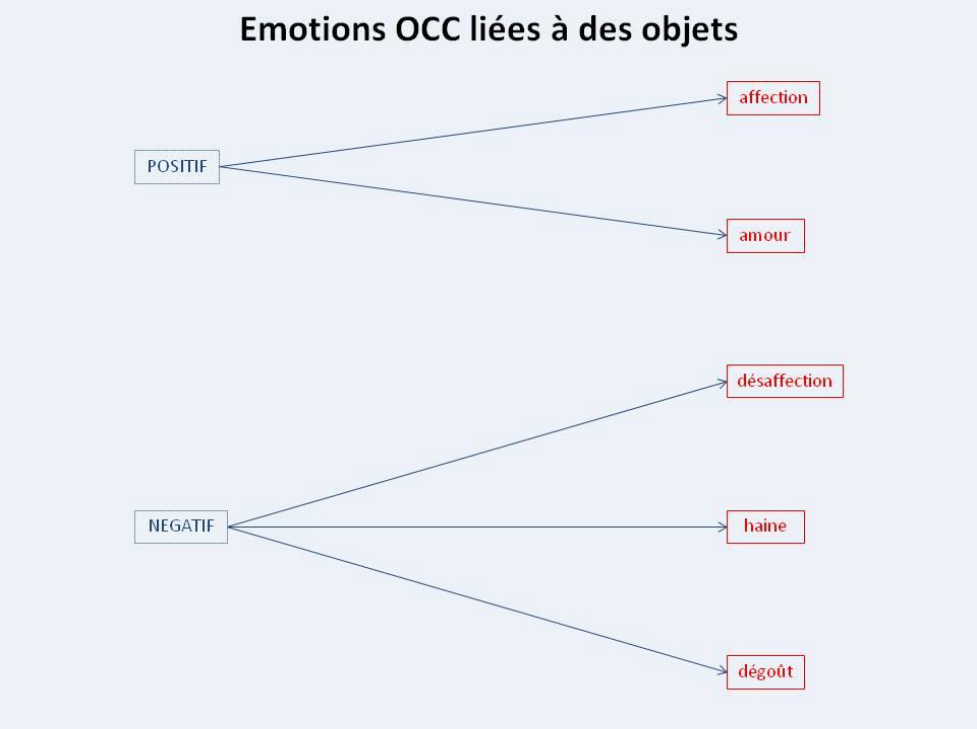


Figure II.5 Hiérarchie des émotions résultant de l'évalutation d'un objet selon Valitutti et Strapparava [**59**].

Par ailleurs, les auteurs enrichissent leur indexation émotionnelle avec un aspect « *causal/statif* » : un adjectif est causal s'il spécifie une émotion causée par l'entité que désigne le nom modifié (ex : un homme « énervant »), il est statif s'il décrit une émotion ressentie par l'entité que désigne le nom modifié (ex : un homme « heureux »). La place du *sujet* y est également importante. Par exemple, le score final

d'une rencontre sportive implique la fierté du vainqueur et la déception du perdant. Enfin, la *dimension temporelle* est prise en compte : une victoire peut être célébrée (au présent) ou espérée (dans le futur).

Les Figure II.3, Figure II.4 et Figure II.5 montrent les hiérarchies qui résultent du modèle d'indexation de l'affect de Valitutti et Strapparava, selon chacune des trois classes d'émotions définies dans l'approche *OCC* [**58**].

II.2.3.2. Le modèle des processus composants (« *Component Process Model* » : *CPM*)

Sander *et al.* s'inscrivent également dans le cadre d'une théorie « évaluative » des émotions. Selon eux, les processus émotionnels sont suscités et gouvernés dynamiquement au fur et à mesure que l'individu évalue des objets, des comportements, des évènements ou des situations, de façon continue et récursive, en relation avec ses valeurs, buts et bien-être général [**60**].

L'émotion se présente sous forme de « *processus composants* ». La notion de « composant » réfère à des *sous-systèmes organismiques* : des réseaux de neurones impliqués dans les fonctions cognitives. Ces sous-systèmes sont au nombre de cinq (*cf.* Figure II.6 sur la colonne de gauche) et gèrent : 1) le traitement de l'information (cognition), 2) le soutien physiologique autonome, 3) la motivation (tendances à l'action), 4) l'action (motricité), 5) et le contrôle (sentiment subjectif). L'émotion est alors définie comme *« un épisode de changements interdépendants et synchronisés dans les états de tout ou majeure partie des cinq sous-systèmes organismiques, en réponse à l'évaluation de stimuli externes ou internes à l'organisme comme pertinents pour les préoccupations centrales de celui-ci »*.

L'émotion est donc un processus qui consiste en l'évaluation séquentielle de stimuli ayant trait à différents éléments. On parle d' « *objectifs d'évaluations* » (« *appraisal objectives* »). Ces objectifs sont :

1. la *pertinence* de l'évènement, en termes de nouveauté (soudaineté, familiarité, prévisibilité), de valence (l'évènement présente-t-il un caractère désirable ?), de but, de besoin ;
2. les *implications sociales* de l'évènement : à quelle cause l'attribue-t-on ? à quels résultats peut-on s'attendre ? quel est le niveau de contradiction entre ce qu'on observe et ce à quoi l'on s'attendait ? à quel but ou besoin les comportements qui guident l'évènement répondent-ils ? l'évènement implique-t-il quelque chose d'urgent ?
3. le potentiel de *contrôle* de l'individu sur la situation, sa capacité à faire face, à s'ajuster ;
4. la *signification normative*, d'un point de vue interne (propre éthique, autocensure, morale commune intégrée…) et externe (normes sociales, codes sociaux…).

L'évaluation séquentielle des stimuli et les fonctions cognitives – que sont l'attention, la mémoire, la motivation, le raisonnement, la constitution du soi, et qui impliquent les cinq composants organismiques comme nous l'avons expliqué plus haut, ont une influence réciproque. Par exemple, si la majorité des stimuli sont généralement comparés à des schémas déjà présents en mémoire, certains stimuli pertinents seront en revanche conservés en mémoire comme schémas émotionnels. De même, alors que la majorité des conséquences prévues d'un évènement perçu sont comparées avec les motivations de l'individu, il est des résultats particuliers d'évaluation qui modifient les motivations et produisent leurs propres tendances à l'action. Les effets réciproques entre évaluations et fonctions cognitives sont symbolisés Figure II.6 par les flèches sur le haut du schéma.

Une telle modélisation rend difficile la prise en compte de la *subjectivité des sentiments* (idiosyncrasie), car celle-ci est supposée assurer une fonction de contrôle, essentielle à la régulation. Pour remédier à cette difficulté, les auteurs intègrent cette subjectivité dans la représentation des réponses conduites par les évaluations de stimuli : plus les évènements sont évalués comme pertinents (ce qui se traduit par un certain degré de synchronisation des composants), plus ils « génèrent de la conscience » et

donc de sentiments associés : lorsqu'un évènement a lieu, le cerveau intègre les changements des composants dans des représentations ; une partie de ces représentations devient consciente et est utilisée à des fins de régulation (représentation personnelle, normes sociales) : c'est ici que la subjectivité prend place. Une partie de ces représentations conscientes peut être exprimée verbalement. Une difficulté persiste néanmoins du fait que la langue ne dit jamais ce que l'on veut dire consciemment, et qu'elle exprime également ce qui ne touche pas au conscient.

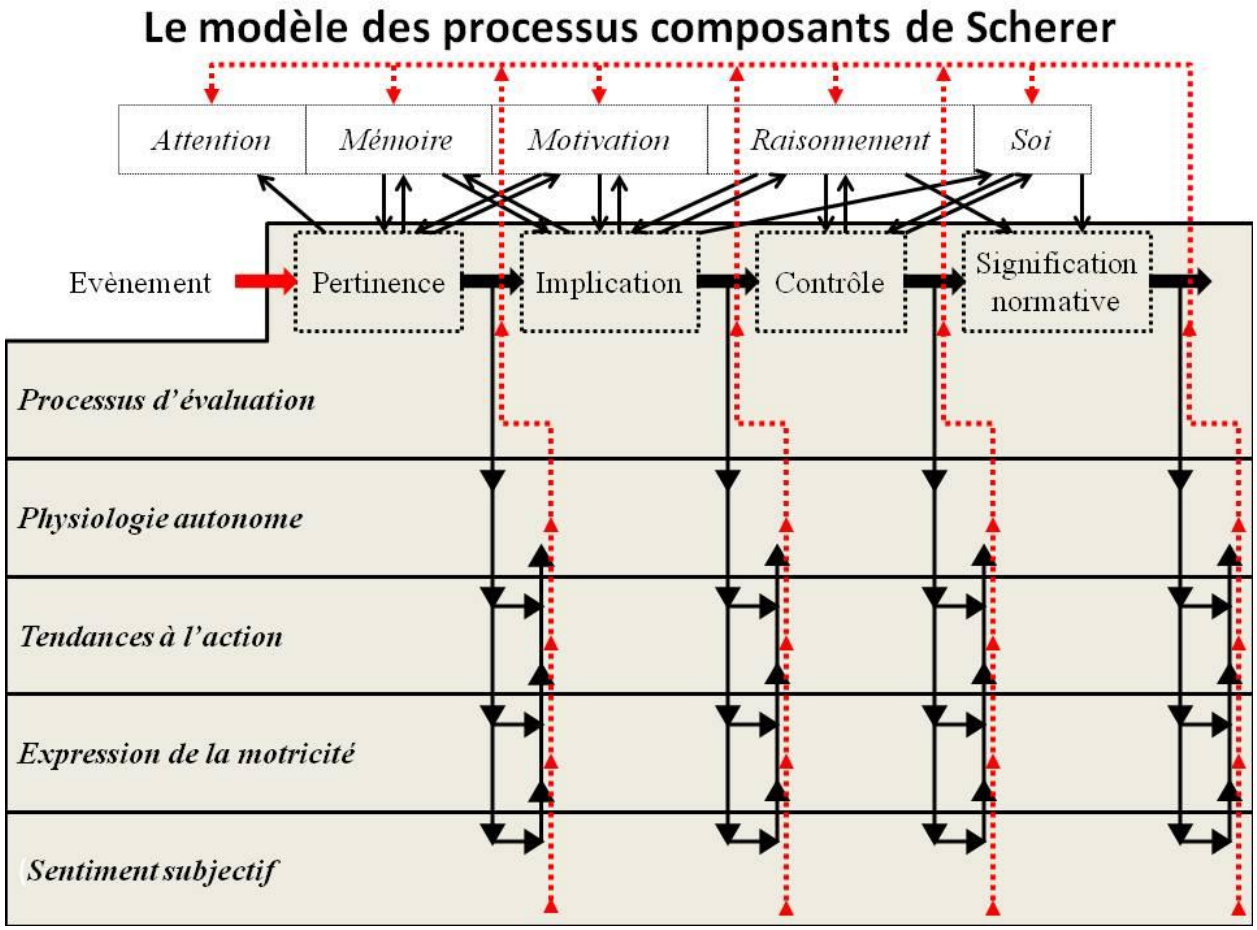


Figure II.6 Illustration du modèle des processus composants des émotions proposé par Klaus R. Scherer [**60**].

Les émotions sont-elles les variables de fonction d'effet ? Ou bien sont-ce les émotions qui donnent leurs valeurs à de telles variables ? Les approches cognitives des émotions tentent ainsi de rendre compte d'une telle réciprocité, et soulignent leur caractère processuel, non-linéaire, et contrôlé. Elles dressent des liens entre évaluations de stimuli et fonctions cognitives, et pour cela, elles sont une source d'inspiration en vue de l'incorporation de contenu émotionnel dans l'appareil cognitif d'agents rationnels conversationnels [**61**] [**62**] [**63**] [**64**] [**65**] [**66**] [**67**] (*cf*. paragraphe III.3).

En dépit du déterminisme revendiqué dans de telles approches, Eva Hudlika souligne dans [**47**] certaines opacités auxquelles la psychologie cognitiviste ou les modèles d'évaluation cognitive ne répondent pas forcément. Quelle place accorder au contexte culturel et aux pratiques dans les appariements stimuli-émotions ? Ces appariements présentent-ils une variabilité interculturelle ou intra-culturelle? Comment en rendre compte lors de la génération d'affect sur des agents conversationnels ou des avatars ? Si nous avons déjà différencié le réflexe émotionnel de la perception de la réaction

émotionnelle en tant que telle, ou l'émotion déclenchée de son expression – prise dans une symbolique, des codes, des conventions –, il nous faut par ailleurs nous pencher sur les effets des émotions sur la personne en termes d'élaboration consciente de soi, par un retour sur l'expérience vécue.

II.2.4. Emotions et constructivisme

Tout comme le geste ne peut-être envisagé comme un contenu autonome, linguistique, dans l'élaboration duquel le corps jouerait un simple rôle de support, les émotions sont à envisager à l'aune d'un rapport à l'intention et à la pratique ; elles ne doivent pas être laissées prisonnières de déterminations strictement conceptuelles.

Dans une étude des émotions en Grèce égéenne, Papataxiarchis [**68**], souligne que les langues européennes, dans la mesure où elles sont tournées vers une grammaire de *nominalisation*, ont tendance à restreindre les désignations émotionnelles à des rapports mot-référent. Elles produisent une délimitation terminologique qui consiste à postuler des objets pré-culturels (disons, dans le cas de l'émotion, des catégories affectives), sans que de telles notions puissent rendre compte de la subtilité des usages concrets de la langue. Il s'agit donc d'interroger la structure de la signification selon d'autres niveaux de langage, comme la prosodie ou la syntaxe, en vue d'appréhender ce que Crapanzano [**69**] appelle les « *foyers d'intérêt indigènes* » de la langue.

Dans *Language and the Politics of Emotion* [**70**], Lutz et Abu-Lughod adressent une critique à l'anthropologie classique et universaliste des émotions. Selon elles, cette dernière chosifie les émotions au lieu de rendre disponibles leurs pleines significations sociale et culturelle, leur rôle *constitutif de* la culture (et non seulement *conceptuel en son sein*).

« Intimement liées à la parole sur soi mais considérées comme éléments inférieurs de celui-ci du fait qu'elles trouvent leur place dans le corps, les émotions […] sont représentées comme la dimension de l'expérience humaine la moins contrôlée, la moins construite, la moins apprise. […] [On doit considérer les émotions comme des constructions qui, tout à la fois, répondent à certaines conditions socio-culturelles et jouent un rôle déterminant dans la formation […] de ce contexte. »

Les émotions rendent donc compte d'un contexte tout en contribuant à sa définition. Elles sont à la fois produit *de* et matériau *pour* l'expérience. Sous ce rapport, des *discours émotionnels* sont importants. Ils recouvrent :

- les discours provoquant des émotions ;
- les discours exprimant des émotions ;
- les discours à propos des émotions.

Se pencher sur ces discours, c'est étudier le *caractère interlocutoire de l'émotion*. L'expression de l'émotion peut déterminer elle-même le contexte dans lequel elle surgit, ainsi que l'évaluation par les acteurs de l'échange. C'est pour révéler cette dimension interlocutoire et intersubjective qu'Olivier Roueff, dans son texte consacré aux liens entre musique et émotions [**71**], s'attaque aux approches conventionnelles de la musique en sciences humaines. Il affirme que la musique s'y retrouve souvent démystifiée. L'approche structuraliste la réduit à des objets structurés, aux contours prédéterminés et suppose que « *les musiciens […] ont une connaissance complète, partagée et purement mentale d'un système de règles dont chacune de leurs prestations serait la simple exécution.* » La sociologie durkheimienne, elle, fait de la musique un sacré, une divinisation de la société, où chaque individu est supposé retrouver son appartenance au groupe. Pour Roueff, davantage que de repérer des « homologies de structure », selon lesquelles a) le morceau de musique instancierait le récit général, et b) ce dernier se

rendrait malléable et extensible de façon à prendre en compte l'œuvre particulière, il s'agit d' « *interroger les pratiques sociales et les discours qui les commentent* » et en se positionnant au plus près de l'expérience :
« [l]es émotions relèvent, au même titre que la mémoire ou la symbolisation, des capacités générales de la personne humaine que chaque situation historique spécifie et agence dans ses contours propres. »

Il faut rendre notamment compte du phénomène d'individualisation des pratiques musicales à l'ère des nouvelles technologies et de la reproductibilité technique des œuvres, où le rapport à celles-ci se veut sans cesse plus personnel, esthétique et expérientiel (*cf.* notion « *art expérientiel* » évoquée à la section I.3.1). L'individu se voit donner la possibilité d'un contrôle de ses propres émotions : il peut provoquer ses expériences émotionnelles et redéfinir ses propres attentes vis-à-vis de la musique. Il y a donc une pratique de la pratique. **L'émotion est donc aussi une expérimentation. Elle est *produite par* le geste et *matériau pour* le geste.**

En particulier, la création artistique, et le geste intentionnel qui est y associé, produisent de l'émotion, tout comme l'émotion est un outil expressif pour le geste artistique. L'expérience de l'art a ceci de singulier qu'elle ramène la psyché dans la *matérialité* de la réalisation. Ainsi, le corps du musicien, c'est-à-dire le lieu du geste musical, est un lieu paradigmatique où **émotion et geste entrent dans une économie médiatisée par le corps**. Le travail de l'artiste dans la construction de son propre style dévoile un rapport de lutte. Dans [**72**], l'ethnologue Annie Paradis décrit l'élaboration du style du chanteur comme une exploration faite de « *vacillement, de brisure, joie et douleur mêlées* », où de façon décousue se mêlent des « forçages », parfois contre-nature, à visée exploratrice. Ce sont ces dispositions émotionnelles paroxystiques, ces *climax*, qu'il s'agit ensuite de mettre en ordre pour bâtir un style. Il s'agit donc d'agencer ses *émotions* pour faire surgir de l'*expressivité*. Interrogée après une leçon de chant, une élève formule la problématique en ces termes : *« Comment accorder sa propre émotion au geste théâtral qui donne l'illusion de l'émotion ? »* Les émotions jouent également leur rôle lors de la performance elle-même. Fort du travail de recherche effectué en répétition, l'artiste n'hésite pas à s'abandonner au plaisir de sa propre création, et peut alors *évaluer* l'impact de ce « laisser-aller » sur le contenu de sa performance et sur le public.

L'artiste joue donc bien avec ses émotions. Elles peuvent devenir un moyen de *contrôle* sur soi, des éléments à agencer de façon presque stratégique, pour faire du corps un moyen expressif. C'est ce qui fait dire à Olivier Roueff que les émotions sont « *le matériau constitutif de l'expérience sociale, et le produit d'une transgression contrôlée des clivages entre image et son, soi et autrui, âme et corps* ».

Le corps est le lieu de l'histoire. Il est contextuel. Il est matière pour le « vouloir explorer gestuel » du sujet. En retour, ce sujet s'incorpore. Le geste du sujet investit la contrainte extérieure et donc produit le corps.

Traiter le geste donc, c'est être à la frontière entre corps et psyché. Le principe que nous retirons de cette étude peut se résumer dans les deux axiomes suivants :

1. Le corps en mouvement est sujet, intériorité, émotion. L'information du corps qui préside à l'élaboration du geste est guidée par une expérience subjective. Les émotions s'extériorisent dans le geste et participent du remodelage de celui-ci. Ainsi, le corps produit les émotions comme les émotions le produisent.
2. Le geste émotionnel, c'est-à-dire l'émotion qui surgit, et qui de par son caractère expérimental détermine elle-même son contexte d'apparition, est corps, extériorité, autre, contrainte. L'intentionnalité du geste ramène la psyché dans la matérialité de la réalisation. L'artiste agence ses émotions pour façonner son corps.

La Figure II.7 illustre les diverses interactions entre le corps, le geste et le sujet.

Bien qu'inspiré par des préoccupations qui concernent l'étude de l'art du mouvement dans le cadre de la danse expressive, le modèle de Laban propose une caractérisation de l'expressivité gestuelle, c'est-à-dire de *l'émotion dans le corps*. Ce modèle nous semble parfaitement adapté pour l'étude du geste en général.

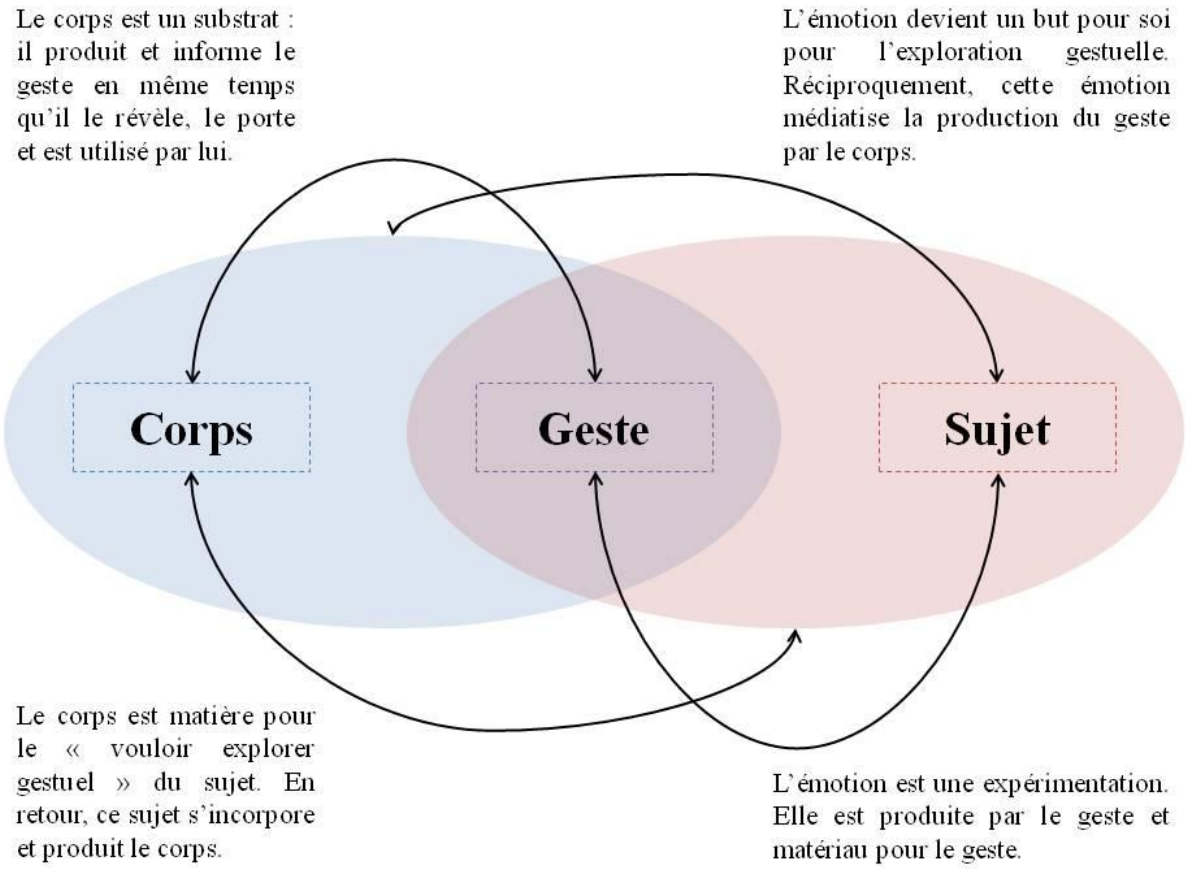


Figure II.7 Interactions entre corps, geste et sujet. Le corps en mouvement *produit* et *est produit par* la subjectivité. Le corps *produit* et *est produit par* le geste émotionnel.

II.3. Le modèle de Laban

II.3.1. Présentation du modèle

Rudolf Laban (Figure II.8.b) était un danseur, chorégraphe, et théoricien de la danse hongrois, qui a développé une méthode d'analyse du mouvement dansé, dite « *Laban Movement Analysis* » (*LMA* [**26**]). Son approche avait en premier lieu des visées pédagogiques et était dédiée à la danse moderne et contemporaine. En effet, la formation à ces dernières requérait que l'on intègre en tant que concepts des références au mouvement souvent métaphoriques voire instinctives.

A partir d'une liste de ce qu'il désigne comme des « *qualités de mouvement* », Laban propose de définir le geste dansé à partir d'une « palette » d'éléments constitutifs de son *caractère intentionnel*.

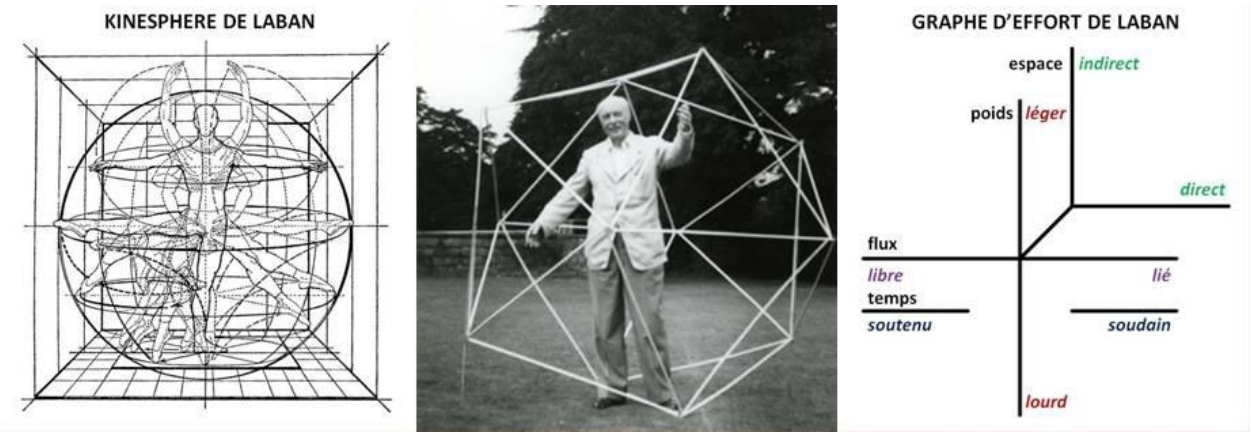


Figure II.8 Kinésphère de Laban (a) ; Rudolf Laban (b) ; Graphe d'effort de Laban (c).

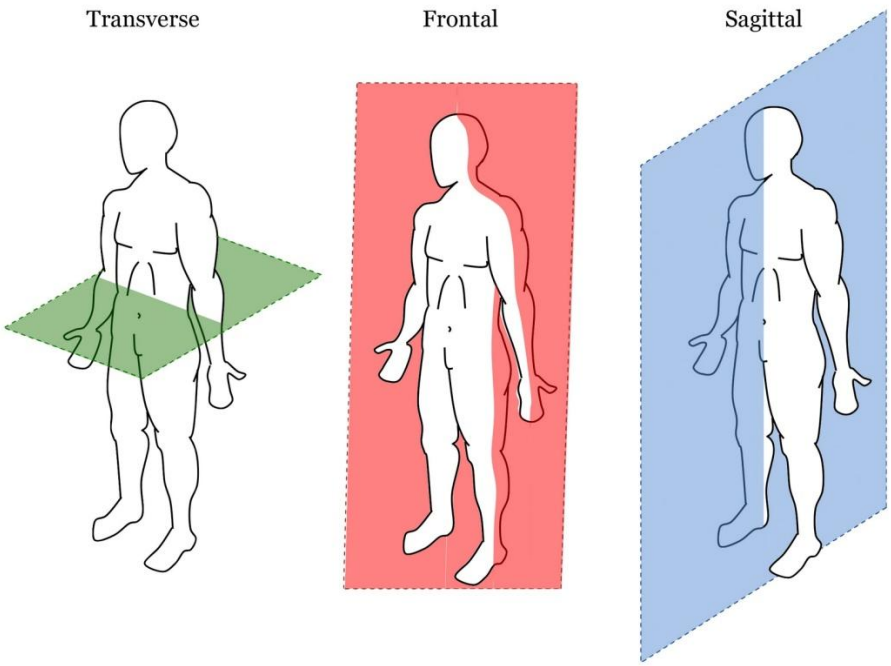


Figure II.9 Plan horizontal ou transverse (a) ; plan vertical ou frontal (b) ; plan sagittal (c).

Davantage que les éléments purement objectifs du geste que sont les déplacements des différentes parties du corps, c'est sa *production* qui intéresse Laban. Cette production témoigne du rapport de l'exécutant à son corps et d'une intentionnalité gestuelle qui fait du geste un phénomène de langage, au sens où Jacques Lacan en parle dans son célèbre article *Au-delà du « Principe de réalité »* [**73**].

Chaque qualité de mouvement traite d'un aspect particulier du mouvement corporel, indépendamment de la trajectoire précise du corps dans l'espace.

Par la suite, le modèle LMA a été étendu à l'étude du geste en général, dans de nombreux travaux de recherche [**74**] [**75**] [**76**] [**77**] [**78**] [**79**] [**80**] [**81**] [**82**], qui témoignent de sa pertinence (*cf.* section III.4).

Les qualités de mouvement définies par Laban sont détaillées ci-dessous :

- La qualité de ***Relation*** décrit les échanges et interactions entre des individus dans le cas de performances collectives.
- La composante de ***Corps*** décrit l'usage des parties du corps, leur coordination au sein du mouvement. De façon générale, il s'agit d'étudier le corps en déplacement.
- La composante ***Espace*** réfère à l'exploration des environs et au chemin du mouvement. Sa caractérisation requiert souvent l'usage du concept de « *kinésphère* », sorte d'espace abstrait dans lequel le corps s'insère et explore les possibilités de son déploiement total (Figure II.8.a).

Ces deux dernières qualités traitent de l'aspect structurel du mouvement. Elles répondent à la question : « *Quel* mouvement est effectivement réalisé ? »

- La composante de ***Forme*** caractérise le changement de forme du corps au cours du mouvement accompli. Elle se subdivise en trois sous-qualités ;
    - la sous-qualité de ***Flux de forme*** traite souci du participant à propos des relations changeantes entre les différentes parties de son corps ;
    - la sous-qualité de ***Mouvement directionnel*** décrit la direction éventuelle du mouvement vers un point particulier de l'espace, l'intention de relier l'action à un point de l'environnement (par exemple : toucher un objet, saluer d'une poignée de main) ; elle est la sous-composante qui traduit le *but* ;
    - la sous-qualité de ***Mise en forme*** caractérise les changements de la forme du corps, l'attitude moulante, sculptante de celui-ci lorsqu'il interagit avec l'environnement (par exemple lors du passage dans une foule), et s'évalue dans des directions de référence : les oppositions élévation/abaissement, avancement/recul, élargissement/rétrécissement définissent généralement des mises en forme le long de directions perpendiculaires aux plans respectivement horizontal (ou transverse), vertical (ou frontal) et sagittal (Figure II.9) ; la *Mise en forme* est une sous-composante orientée « *processus* ».
- Enfin, la composante ***Effort*** (Figure II.8.c) décrit comment l'individu concentre son énergie pour effectuer le mouvement. Elle a trait à l'attitude intérieure de l'individu envers quatre facteurs moteurs qui constituent quatre sous-qualités d'effort. Chacune de ces sous-qualités s'évalue au sein d'un continuum entre deux extrêmes, l'un consistant à se livrer pleinement dans le sens de la qualité *(indulging in the quality*), l'autre incarnant une lutte contre *(fighting against the quality*) :
    - la sous-qualité d'***Espace*** (à ne pas confondre la qualité d'*Espace*) traite de l'attention que le sujet porte sur les environs, et discerne un mouvement *direct* ou *rectiligne* d'un mouvement *indirect* ou *flexible* ;
    - la sous-qualité de ***Temps*** qualifie le mouvement selon qu'il traduit une urgence plus ou moins forte, et différencie un mouvement *soudain* d'un mouvement *soutenu* ou *continu* ;

- la sous-qualité de ***Flux*** renseigne sur le caractère tendu ou non du geste, elle traduit l'attitude de *contrôle*, et distingue un mouvement *libre* d'un mouvement *lié*, *contraint* ou *contrôlé* ;
- la sous-qualité de ***Poids*** caractérise l'impact du mouvement et différencie un mouvement *lourd* ou *ferme* d'un mouvement *léger*.

Les qualités de *Forme* et d'*Effort* décrivent l'aspect qualitatif du mouvement. Elles répondent à la question : « *Comment* le mouvement est-il réalisé ? »

Ce n'est qu'en 1994, bien après la naissance de la LMA, que Rudolf Laban conceptualise sa « théorie de l'effort » [**26**], et ce faisant la définition d'un espace pluridimensionnel de sous-composantes d'*Effort* énumérées ci-dessus. Cette théorie, qui vise à caractériser la concentration de l'énergie et le rapport du sujet du mouvement à son intériorité, est régulièrement utilisée seule pour caractériser le geste [**82**] [**83**] [**80**] [**25**] [**78**] [**81**]. Pour illustrer son caractère « autosuffisant », nous donnons dans le Tableau II.1 des exemples de descriptions de mouvements, selon certaines sous-dimensions d'*Effort*.

Tableau II.1 Exemples de mouvements décrits selon des caractéristiques d'Effort de Laban.

Mouvement	**Espace**	**Temps**	**Poids**
Donner un coup de poing	Direct	Soudain	Lourd
Tamponner une feuille	Direct	Soudain	Léger
Presser un bouton	Direct	Soutenu	Lourd
Planer	Direct	Soutenu	Léger
Sabrer	Indirect	Soudain	Lourd
Feuilleter un dossier	Indirect	Soudain	Léger
Tordre un objet	Indirect	Soutenu	Lourd
Flotter sur l'eau	Indirect	Soutenu	Léger

II.3.2. Discussion

La plupart des concepts du modèle LMA sont abstraits et intuitifs, dans la mesure où ils traitent de l'intentionnalité et de l'exploration corporelle. Ce niveau d'abstraction rend difficile l'idée d'une quantification des qualités ou sous-qualités de mouvement en une formalisation mathématique rigoureuse. Dans le chapitre suivant, nous nous pencherons sur des travaux dédiés à de telles quantifications des qualités de Laban, ainsi qu'à la qualification de gestes selon les qualités ou sous-qualités de Laban à l'aide de systèmes d'apprentissage impliquant des descripteurs mi-niveau.

Par ailleurs, soulignons ici qu'il existe un certain degré de recouvrement entre différents concepts de la LMA. On citera par exemple la proximité qui existe entre d'un côté la qualité de *Corps*, relative à l'usage des parties du corps, ainsi que leur coordination au sein du mouvement et généralement relié à la « *kinésphère* » de Laban (Figure II.8.a), et de l'autre la sous-qualité de *Flux de forme*, qui traite de l'évolution dynamique des relations entre les différentes parties du corps et est souvent mise en rapport avec l'agrandissement ou la réduction de la forme du corps (notamment du torse).

De même, la sous-composante d'*Espace* de la qualité d'*Effort*, qui traite de l'attention que le sujet porte sur les environs, et qui distingue les mouvements en ligne droite des mouvements plus hésitants ou indécis, n'est pas aisément différenciable de la notion de *Mouvement directionnel* (sous-qualité de *Forme*) qui caractérise la direction éventuelle du mouvement vers un point particulier de l'espace.

Aussi les quatre facteurs d'effort sont-ils corrélés les uns aux autres. Dans [**80**], à partir de gestes annotés selon la LMA par des spécialistes, Zhao et Badler entrainent quatre réseaux de neurones (un pour chaque sous-qualité de l'*Effort*) pour caractériser un geste selon que le performeur se livre pleinement dans le sens de la qualité en question *(indulging in the quality*), ou lutte contre *(fighting against the quality*). Le classifieur « *Temps* » est par exemple entrainé pour étiqueter un geste selon qu'il est *soudain* ou *soutenu* (ou « *neutre* », dans le cas où la discrimination ne se fait pas avec netteté). En comparant les résultats issus de la classification et ceux de la vérité terrain définie par les annotateurs, les auteurs proposent une recension des erreurs de classification qu'opèrent les réseaux de neurones.

- Ils constatent notamment que des confusions ont lieu entre des mouvements *libres* et des mouvements *liés* et *soudains*, ainsi qu'entre des mouvements *liés* et des mouvements *libres* et *soutenus*. Les recouvrements *libre*/*soudain* et *lié*/*soutenu* illustrent ainsi la corrélation qui existe entre les sous-qualités de ***Flux*** et de ***Temps***.
- De même, ils soulignent les confusions entre caractérisations du mouvement selon qu'il est *lourd* et selon qu'il est *léger* et *soudain*, ou entre caractérisations du mouvement selon qu'il est *léger* et selon qu'il est *lourd* et *soutenu*. Ici, c'est la corrélation entre ***Poids*** et ***Temps*** qui est pointée, de par les appariements *lourd*/*soudain* et *léger*/*soutenu*.
- Enfin, la triple corrélation entre les sous-composantes d'***Espace***, de ***Poids*** et de ***Temps*** est illustrée par les divergences entre classification et vérité terrain, d'un côté à propos des mouvements *directs* et des mouvements *indirects*, *lourds* et *soudains*, de l'autre à propos des mouvements *indirects* et des mouvements *directs*, *légers* et *soutenus*. Des appariements *direct*/*lourd-soudain* et *indirect*/*léger-soutenu* existent.

De telles observations témoignent de la compacité du concept d'*Effort*.

Enfin, notons que certaines qualités semblent plus ou moins impliquées dans la description du mouvement selon le type de geste considéré. Ainsi, le concept de *Relation* est tout à fait pertinent dans le cas de l'étude d'activités de groupes, mais relativement peu intéressant pour ce qui est de performances individuelles, à moins d'en étudier l'enjeu communicationnel vis-à-vis d'un public ou de tierces personnes que l'acteur du geste est amené à se représenter, même métaphoriquement. En tout état de cause, bien qu'elle puisse se révéler pertinente pour l'étude de scènes ou d'interactions artistiques – relation entre un chef d'orchestre et ses musiciens [**84**], interactions et leadership au sein d'un quatuor à corde [**85**] [**86**] ou au sein d'un ensemble plus large [**87**]) – nous n'avons pas traité la qualité de Relation dans notre travail.

Dans la section suivante, nous passons en revue les récents travaux en matière d'analyse automatique du geste qui tentent d'élaborer un modèle de description mathématique du geste capable de prendre en compte ses dimensions intersubjective et intentionnelle. En dépit de la difficulté que constitue la caractérisation précise de chacun des concepts employés dans LMA, certains résultats obtenus par des approches de l'état de l'art qui y sont dédiées confirment la légitimité de notre démarche.

III. Etat de l'art

Dans ce chapitre, nous dressons un état de l'art de divers travaux ayant trait à l'analyse du mouvement corporel. Dans un premier temps, nous nous penchons sur des méthodes dédiées à la reconnaissance d'actions dans des contenus vidéo 2D et 3D. Ensuite, nous évoquons brièvement certains travaux de recherche dédiés à l'analyse de l'activité ou de la créativité de groupes sociaux. Une troisième partie est consacrée à l'analyse de contenus émotionnels dans des vidéos 2D et 3D, ainsi qu'à la synthèse d'indices affectifs sur des avatars. Enfin, nous présentons des approches dédiées à l'analyse de l'expressivité du geste, à des fins de caractérisation expressive, ou en vue de l'élaboration de descripteurs mi-niveau pour la reconnaissance d'autres types de contenus sémantiques (actions, émotions).

Comme nous l'avons vu dès le premier chapitre, le corps est le lieu par le biais duquel se forme la conscience de soi autant qu'il est un substrat pour la communication et l'empathie intersubjective. Le contenu d'un geste ne peut par conséquent en aucun cas se référer à l'unique coordination des membres du corps en vue de la réalisation d'une action précise, mais comprend également des éléments intentionnels et communicationnels. Ces aspects ont été au centre de nombreuses recherches dans le domaine du traitement et de l'analyse automatique du geste, qui seront présentées et discutées dans ce chapitre, afin de répondre aux questions suivantes :

- Les éléments descriptifs du geste sont-ils locaux (les descripteurs du mouvement sont alors dédiés à un ensemble de trames correspondant au voisinage temporel qui délimite un état du geste) ou globaux (descripteurs qualifiant le geste sur l'entièreté de son empan temporel) ?
- Quelle est la nature des informations quantifiées par les descripteurs en vue de l'analyse du geste ?
- Quelles informations sémantiques sont reconnues et analysées ?

Nous nous consacrons en premier lieu aux travaux dédiés à l'analyse purement visuelle et structurelle du mouvement, avant d'aborder la prise en compte des aspects d'intersubjectivité.

III.1. Analyse de mouvement et reconnaissance d'actions

Le domaine d'analyse et de reconnaissance d'actions nécessite la disponibilité de corpus de test avec vérité terrain, afin de permettre la mise en place de méthodes d'apprentissage supervisés mais aussi pour pouvoir évaluer objectivement les performances des algorithmes et comparer les méthodes de l'état de l'art. Les premières bases de données, apparues bien avant l'émergence des technologies de capture 3D, concernent des séquences gestuelles sous forme de vidéos 2D.

III.1.1. Bases de données de vidéos 2D

Parmi les bases de données 2D les plus populaires et largement utilisés par les méthodes de l'état de l'art pour effectuer des benchmarks, citons :

- Le corpus **KTH** [**88**], qui inclut 6 types d'actions humaines réalisées par vingt-cinq personnes dans quatre scénarios différentes : *walking*, *jogging*, *running*, *boxing*, *hand waving* et *hand clapping*. Il consiste en 600 vidéos enregistrées à partir d'une caméra statique et mettant en jeu des arrières plans en intérieur/extérieur.
- Le corpus **Weizmann** [**89**] comprend 10 types d'actions humaines : *walk*, *run*, *jump*, *gallop sideways*, *bend*, *one-hand wave*, *two-hands wave*, *jump in place*, *jumping jack* et *skip*. Il se compose de 90 vidéos enregistrées depuis une caméra statique.
- Le corpus **HDM05-MoCap** [**90**] contient plus de 3 heures de données aux formats C3D ou ASF/AMC. Il met en jeu plus de 70 classes de mouvement réalisées chacune 10 à 50 fois par 5 acteurs selon des scénarios pré-définis.
- Le corpus **Naval Air Training and Operating Procedures Standardization (NATOPS) aircraft handling signals dataset** [**91**] utilise le vocabulaire gestuel officiel utilisé sur les porte-avions de la marine américaine et consiste en un ensemble de gestes de mains et de corps : *have command*, *all clear*, *not clear*, *spread wings*, *fold wings*, *lock wings*, etc.

- Le corpus **INRIA Xmas Motion Acquisition Sequences dataset** (IXMAS [**92**]) est une base de vidéos multi-vues dédié à la reconnaissance d'actions humaines, consistant en 14 gestes effectués 3 fois par 12 acteurs : *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *point*, *pick up*, *throw (over head)* et *throw (from bottom up)*.
- Le corpus **HOHA** [**93**] propose des échantillons vidéo extraits de 32 films et annotés selon 8 classes d'actions : *answer the phone*, *get out of the car*, *handshake*, *hug a person*, *kiss*, *sit down*, *sit up* et *stand up*.
- Le corpus **HMDB51** [**94**] propose 51 catégories d'actions et au moins 101 vidéos pour chaque catégorie, pour un total de 6 766 vidéos de provenances diverses (films, YouTube, etc.). Ces vidéos mettent en jeu des difficultés diverses : mouvement de la caméra, divers points de vue, différentes qualité de vidéo, problèmes d'occlusion.
- Le corpus **ADL** [**95**] propose des scènes de la vie quotidienne qui correspondent à 18 actions : *combing hair*, *make up*, *brushing teeth*, *dental floss*, *washing hand/face*, *drying hand/face*, *laundry*, *vacuuming*, *washing dishes*, *moving dishes*, *making tea*, *making coffee*, *drinking water/bottle*, *drinking water/tap*, *making cold food/snack*, *watching TV*, *using computer* et *using cell*. Sa durée totale est de 10 heures, enregistrées par 20 personnes avec une caméra portable GoPro.
- Le corpus **UCF Sports Action** [**96**] présente 10 classes d'actions en rapport avec le sport : *diving*, *golf swinging*, *kicking*, *lifting*, *riding horse*, *running*, *skate boarding*, *swing-bench*, swing-side et *walking*. Ces actions sont réparties dans 182 vidéos enregistrées depuis une caméra statique. Ces vidéos proviennent de programmes télévision dédiés au sport.

L'obtention d'indices pertinents du mouvement pour l'analyse et la reconnaissance des geste nécessite dans le cas de ces différentes bases de vidéos 2D la mise en œuvre de méthodes de traitement, de segmentation et d'extraction de caractéristiques élaborées, ce qui représente une difficulté supplémentaire à gérer.

Une première famille d'approches s'appuie sur l'analyse et la caractérisation du geste en termes de silhouettes, de postures ou de poses.

- Une silhouette est définie comme un profil de forme dénotant l'attitude globale d'un individu à un instant précis. Le plus souvent, la silhouette représente la région 2D, ou dans certains cas 2D+*t* de support de la forme du corps. On parle aussi de posture pour caractériser ces formes du corps.
- La notion de pose renvoie à la position 3D de l'ensemble des articulations 3D du corps.

III.1.2. Silhouettes, postures, poses

Dans [**97**], Chen *et al.* extraient des postures corporelles à partir d'images 2D, en utilisant les angles d'une « étoile » définie par les positions de la tête et des quatre membres. Un geste est alors défini comme une séquence d'étoiles successives. Une telle séquence définit un vecteur caractéristique, qu'il s'agit de transformer en une suite de symboles, de sorte à pouvoir l'utiliser en entrée de modèles de Markov cachées (*Hidden Markov Models* : *HMM* [**98**]) à observations discrètes. Pour ce faire, un lexique de postures est constitué qui contient les étoiles les plus représentatives de chaque type d'action, grâce auquel chaque étoile d'une séquence gestuelle peut se voir attribuer un symbole.

L'approche est testée sur une base de dix actions, réparties dans une centaine de clips vidéo. Les résultats atteignant 98% de taux de reconnaissance.

Dans [**99**], Wang *et al*. utilisent la transformée de Radon sur des silhouettes extraites de séquences vidéos pour nourrir des HMM [**98**] et reconnaître cinq activités humaines (*rushing*, *carrying a bag*, *suddenly bending down when walking*, *walking normally* et *jumping*), avec de forts taux de reconnaissance (jusqu'à 98%). L'analyse en composantes principales (A.C.P.) est effectuée pour réduire la dimensionnalité de l'espace de description.

L'approche de Huang et Xu [**100**] exploite différentes sections planes d'une silhouette à un instant donné, ainsi que la projection de ces sections sur les repères attachés à deux caméras orthogonales. Ce vecteur d'enveloppe de forme est utilisé en entrée de HMM [**98**] pour reconnaître neuf actions réalisées par sept acteurs. Les taux de reconnaissance rapportés sont supérieurs à 95% et 83% dans les cas respectifs de stratégie de reconnaissance dépendante et indépendante du sujet. Une réduction dimensionnelle par ACP est également opérée au préalable.

Junejo *et al.* [**101**] proposent également des descripteurs à base de silhouettes à des fins de reconnaissance d'actions. Pour chaque trame, après avoir effectué une extraction du premier plan de l'image, ils y localisent la silhouette-objet. La silhouette est convertie en une série temporelle, qui constitue une représentation 1D de l'action à un instant particulier. Ensuite, ils calculent une « *approximation agrégée symbolique* » (*Symbolic Aggregate Approximation* : *SAX*) des séries temporelles, qui consiste à réduire la taille de chaque série à un petit nombre de segments et à échantillonner les valeurs selon un nombre donné de symboles. Ces représentations de silhouettes sont ensuite utilisées pour des objectifs de classification supervisée à l'aide de forêts aléatoires *(Random Forests* [**102**]). Sur la base de test Weizmann [**89**] (*cf*. section III.1.1), la méthode permet d'atteindre des taux de précision de 89%.

Dans [**91**], Song *et al.* se focalisent sur des poses 3D du haut du corps, estimées avec une méthode d'inférence bayésienne multi-hypothèses. Les poses 3D obtenues sont ensuite utilisées pour la détection des mains et l'identification main gauche/droite. Des caractéristiques dérivées d'histogrammes de gradient [**103**] sont calculées pour les poses des mains et utilisées comme descripteurs en entrée d'un classificateur SVM (*Support Vector Machines* [**104**]).

L'évaluation de la méthode effectuée sur le corpus de gestes NATOPS [**91**] (section III.1.1) conduit à des performances très élevées : les F-scores obtenus pour chacune des quatre catégories gestuelles considérées sont respectivement de 94%, 99%, 94% et 89%.

Singh et Nevatia [**105**] proposent l'usage d'une combinaison de réseaux Bayésiens dynamiques (*Dynamic Bayesian Action Networks* : *DBAN*) avec des modèles intermédiaires 2D des parties du corps, pour l'estimation des poses et la reconnaissance d'actions. Des actions composites sont décomposées en séquences de primitives décrivant les variations entre les poses-clés 3D de trames consécutives. Les poses 3D sont projetées sur des modèles graphiques 2D représentant un squelette de dix articulations de référence du corps sous forme de « structures picturales », de manière à ce qu'un état de DBAN à un instant t soit décrit par :

- une action composite ce_t ;
- une action primitive pe_t ;
- le temps écoulé depuis le début l'action primitive d_t ;
- une pose 2D p_t.

La méthode est testée sur une base de données de gestes de la main de 500 échantillons, avec des taux de reconnaissance autour de 85-90% pour chacune des douze classes d'action considérées.

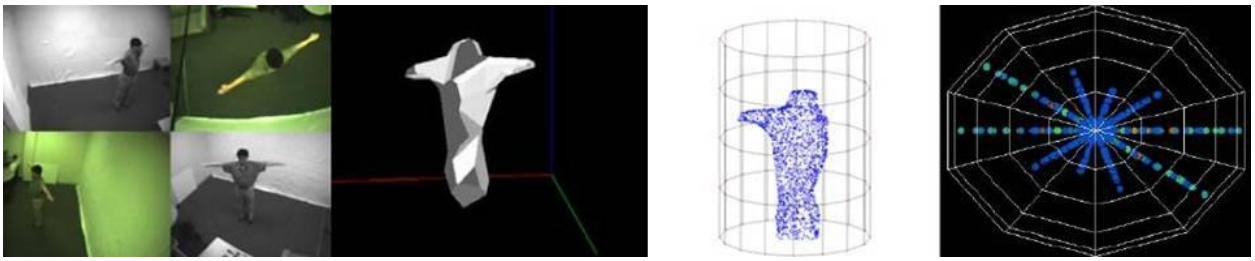


Figure III.1 Approche proposée par Chu *et al.* (source : [**106**]) : différentes vues du corps (a) ; enveloppe visuelle 3D reconstruite à partir des vues (b) ; cylindre et points de références Q_j choisis (c) ; distribution des points P_i (d).

En [**106**], Chu *et al.* introduisent une approche fondée sur la décomposition de postures en combinaisons d'atomes primaires et secondaires à des fins de reconnaissance de gestes et de postures. Etant donné une série d'images de silhouettes en 2D correspondant à différentes vues (Figure III.1.a), une forme 3D approximative correspondant à l'enveloppe visuelle (EV) du corps est construite (Figure III.1.b). Cette EV est échantillonnée en un ensemble de points P_i (en bleu sur la Figure III.1.c), afin d'obtenir une représentation simplifiée de la silhouette. Un cylindre englobant de l'EV, centré sur le centroïde de la silhouette (*e.g.*, sur le centroïde des points P_i), est ensuite défini et échantillonné à son tour en un ensemble de points de référence Q_j (chaque point Q_j correspond à l'intersection entre un cercle et une ligne verticale sur la Figure III.1.c). A chaque point Q_j est attaché un système de coordonnées sphériques. Dans chacun de ces systèmes $S_j = (r_j, \theta_j, \varphi_j)$ sont calculées les coordonnées de chaque sommet P_i de la forme. De cette façon, la forme initiale peut être représentée dans chaque référentiel attaché à un point Q_j. Si l'on considère l'échantillonnage des dimensions dans lesquelles les coordonnées sphériques (c'est-à-dire la distance ou le rayon r_j, l'angle de longitude θ_j, et l'angle d'élévation φ_j) évoluent, il est possible de définir une parcellisation de l'espace autour du point Q_j (Figure III.1.c). Chaque parcelle d'espace résultant de cet échantillonnage du système de coordonnées sphériques attaché au point Q_j peut ainsi se voir attribuer une valeur correspondant au nombre de sommets P_i de la silhouette qui y appartiennent (Figure III.1.d). Ce faisant, il possible de construire pour chaque Q_j un histogramme représentant les « remplissages » de telles parcelles. Les histogrammes correspondant à chaque point de référence Q_j sont sommés et normalisés relativement à la plus grande valeur rencontrée. Ces histogrammes normalisés constituent le descripteur de forme pour la posture donnée.

La représentation obtenue jouit d'une intéressante propriété d'additivité, qui suggère que des sous-ensembles de postures suffisent à représenter des postures complexes ou pour reconnaitre des gestes. Trente postures différentes et arbitraires sont collectées. Les descripteurs de forme sont calculés pour chacune d'entre elles. Pour chaque paire de postures (P_m, P_n) une procédure fondée sur l'algorithme « *Matching pursuit* » [**107**] est appliquée pour décomposer P_n en utilisant P_m comme atome. Les poids associés à la décomposition $w_{m,n}$ sont conservés dans une matrice symétrique. Une décomposition en valeurs singulières (*Singular Value Decomposition* : *SVD*) est ensuite appliquée à la matrice ainsi obtenue. Cinq atomes correspondant aux plus grandes valeurs singulières sont extraits. Ces cinq atomes sont utilisés dans le processus de décomposition en postures : pour chacun d'eux, les deux atomes différents de la position de repos et avec les plus grands poids dans la décomposition sont sélectionnés comme première et seconde posture-atome. Ainsi, un dictionnaire de quinze poses est obtenu : les cinq

atomes initiaux et les deux postures résultantes pour chacun d'entre eux obtenues par l'application du processus de décomposition.

La méthode de reconnaissance de postures utilise des HMM à états-duaux [**98**], où une observation correspond à une paire d'atomes primaire/secondaire. Dans une première expérience, l'objectif est à la discrimination entre différentes postures. Les auteurs utilisent des séquences vidéo combinant cinq postures différentes, et obtiennent des taux de reconnaissance supérieurs à 81% pour chacune des cinq postures et supérieurs à 70% dans le cas de postures composites, pour un taux moyen avoisinant les 90%.

Dans [**38**], Li *et al.* utilisent à chaque trame une carte de profondeur acquise avec une caméra Kinect, représentant la surface 3D de la pose corporelle, pour en extraire des points 3D de référence. Ces points sont calculés à l'aide des trois projections orthogonales sur les plans cartésiens de la carte de profondeur. Sur chaque projection, les contours de silhouette sont extraits, échantillonnés et utilisés pour reconstituer des points 3D.

Les auteurs modélisent les dynamiques du mouvement corporel à l'aide d'un graphe d'action dont les nœuds correspondent à des postures saillantes que partagent les actions.

Un tel modèle est décrit à l'aide :

- d'une liste de postures $\{\omega_m\}_{m\in\{1,M\}}$;
- d'une liste d'actions à analyser $\{A_l\}_{l\in\{1,L\}}$;
- de probabilités de transition $\{a_{i,j,k}\}_{i\in\{1,M\},j\in\{1,M\}}^{k\in\{1,L\}}$ modélisant chacune le passage d'une posture *i* à une autre posture *j* pour une action donnée A_l ;
- de probabilités de transition $\{b_{k,l}\}_{k\in\{1,L\},l\in\{1,L\}}$ modélisant le passage d'une action *k* à une autre *l* ;
- de probabilités d'émission d'observations – lots de points 3D – $\{p(x/\omega_m)\}_{m\in\{1,M\}}$ pour chaque posture, modélisées par des distributions gaussiennes.

Il est alors possible de reconnaître l'action la plus probable sous-jacente à une série d'observations. Pour cela, les auteurs adoptent un schéma de décodage bi-gramme avec maximum de vraisemblance (*bi-gram with maximum likelihood decoding scheme* : *BMLD*).

L'approche est testée sur le corpus MSR Action 3D [**38**] (*cf.* section V.1). Sur les différents sous-ensembles de gestes proposés, les taux de reconnaissance sont à supérieurs à 71.9% dans le cas de test « cross subject » (les gestes d'une moitié d'individus sont utilisés pour l'entrainement, les gestes de l'autre moitié pour le test), et de 89% dans le cas de test où les réalisations de chaque sujet sont présentes dans les données d'entrainement comme de test.

Cette première série d'approches basées sur des images 2D ou de profondeur montre l'importance accordée à un sous-échantillonnage du mouvement en postures ou en poses de référence. De façon générale, ces approches définissent implicitement le geste comme une séquence articulée, traduisible en série de symboles utilisable dans un cadre de reconnaissance dynamique du geste. L'inconvénient majeur qu'elles présentent réside dans l'absence de prise en compte de la dimension temporelle de la réalisation du mouvement corporel Celui-ci y est réduit à une succession de configurations spatiales qui manque les indices dynamiques du mouvement ainsi que son caractère plus ou moins instantané.

Une deuxième famille d'approches d'analyse de gestes se focalise sur des aspects plus locaux du mouvement. En général, ces méthodes s'appuient sur des représentations spatio-temporelles et tendent à repérer différents motifs de saillance. Wang *et al.* présentent dans [**108**] un recensement des caractéristiques spatio-temporelles habituellement extraites des vidéos 2D en vue de la construction de

descripteurs du mouvement dédiés à la reconnaissance d'actions. Les approches les plus représentatives de l'état de l'art sont présentées dans le paragraphe suivant.

III.1.3. Modèles locaux et saillance spatio-temporelle

Dans [**109**], Nguyen et Manzanera utilisent les trajectoires rendues en temps réel par un algorithme de tracking semi-dense pour en extraire les points correspondant à des maxima locaux de courbure, c'est-à-dire d'accélération radiale. Ces points présentent l'avantage d'être robustes aux transformations géométriques et aux variations d'apparence. Sur chaque trajectoire suivie, chaque point d'intérêt définit donc autour de lui un tronçon de trajectoire. Sur ces diverses portions, 14 descripteurs sont calculés qui ont trait à la géométrie et à la dynamique du mouvement. L'opération est répétée avec différents points correspondant à divers niveaux de lissage de la trajectoire (*e.g.*, différentes échelles). Les descripteurs sont alors utilisés avec une procédure de clustering classique (*e.g.*, k-means [**110**]) pour sélectionner des mots-clés représentant chacun un élément caractéristique du mouvement. Chaque action peut alors se voir attribuer un histogramme de mots visuels, dénotant le pouvoir représentatif des différents éléments du dictionnaire ainsi constitué. Cet histogramme résultant est utilisé en entrée de SVM classiques [**104**] dédiés à la reconnaissance d'actions.

Le protocole est testé sur les différentes actions du KTH dataset [**88**]**,** avec un taux de reconnaissance moyen de 95%.

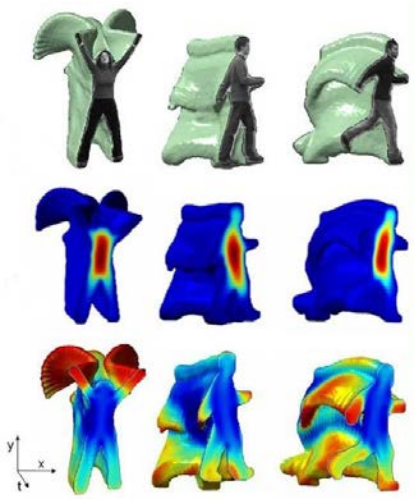


Figure III.2 Approche de Blank *et al.* (source : [**89**]) : exemples formes spatio-temporelles pour trois actions (a) ; solution de l'équation de Poisson, des grandes valeurs (rouge) aux petites (bleu) (b) ; représentation de la saillance spatio-temporelle locale (grandes valeurs en rouge, petites valeurs en bleu (c).

Citons ensuite la méthode proposée par Blank *et al.* [**89**] qui prend pour point de départ la représentation d'une action en tant que forme spatio-temporelle laissée par le corps lors de son passage dans l'espace à chaque instant du geste (Figure III.2.a). Pour chaque point à l'intérieur de cette forme, une valeur scalaire est associée. Elle correspond au temps moyen nécessaire à une particule en déplacement aléatoire, qui démarre son mouvement au point considéré, pour atteindre la frontière de la forme. Les auteurs montrent que le calcul d'une telle valeur revient à résoudre l'équation de Poisson (Figure III.2.b). La solution de cette équation est alors utilisée pour identifier la saillance spatio-temporelle des parties en

mouvement (Figure III.2.c). Ces mesures de saillance sont enfin intégrées dans un vecteur global qui sert de descripteur pour l'action.

L'approche est testée sur le corpus Weizmann [**89**] (*cf.* section III.1.1) avec des performances en termes de taux de reconnaissance proches de 100% pour la majorité des dix catégories de geste. Une seule catégorie (*jump*) conduit à un score légèrement inférieur à 90%.

Dans le même ordre que l'approche précédente, nous pouvons également citer le travail d'Achard *et al.* [**111**], qui se proposent de caractériser diverses actions à partir de micro mouvements, c'est-à-dire d'indices spatio-temporels calculés au sein de micro-volumes spatio-temporels (*e.g.*, dans une fenêtre d'images). Ils affirment qu'une telle représentation du mouvement préserve la dynamique de l'action au sein des différentes fenêtres temporelles. Chaque micro-volume spatio-temporel présente des silhouettes binaires successives qui ont été extraites à partir de la détection initiale de pixels en mouvement. Le descripteur de trame (*e.g.*, pour ce micro-volume centré autour de l'instant *t*) se compose de divers moments de la silhouette variante, sous réserve de diverses normalisations spatiales et temporelles. Les descripteurs successifs pour une vidéo sont utilisés en entrée de HMM [**98**], qui sont idéaux pour la reconnaissance de contenus temporels pouvant varier en durée.

Les auteurs testent leur approche sur une base comprenant huit actions, avec différentes tailles de fenêtre pour les micro-volumes spatio-temporels. Le meilleur taux de reconnaissance de 89% est atteint pour une taille de fenêtre temporelle de sept trames et un nombre d'états de HMM égal à trois.

Dans [**112**], la saillance d'un point sur l'image d'une vidéo est évaluée relativement à l'information (au sens de l'entropie) contenue dans un voisinage spatio-temporel. Une technique de déformation temporelle est appliquée pour normaliser 152 échantillons d'exercices de gymnastique effectués par des participants amateurs. Des modèles *RVM (Relevance Vector Machines)* [**113**] sont ici utilisés pour inférer des catégories d'action. Les RVM utilisent l'inférence bayésienne en introduisant un préalable dans les poids du modèle, régi par une liste d'hyper-paramètres – un pour chaque poids. Les valeurs les plus probables de ces hyper-paramètres sont estimées itérativement à partir des données. Au contraire des SVM, les poids non nuls des RVM ne sont pas associés à des exemples proches de la frontière de décision, mais apparaissent plutôt comme représentant des exemples de classes prototypiques. Ces exemples sont appelés « vecteurs de pertinence » (*relevance vectors*), et dans le cas présent, sont interprétés comme des exécutions représentatives d'une action humaine. Les RVM utilisent moins de noyaux que les SVM classiques [**104**]. Par ailleurs, les prédictions des RVM sont probabilistes, au contraire du déterminisme inhérent aux décisions rendues par les SVM.

Dans [**92**], Weinland *et al.* introduisent la notion de « *volumes d'histoire du mouvement* » (*motion History Volumes* : *MHV*) pour la reconnaissance d'actions dans le cas de vidéos et pour des mouvements corporels libres en termes de lieu, d'orientation et de longueur. Les MHV étendent les « images d'histoire du mouvement » (*Motion History Images* : *MHI*) [**114**] de l'univers 2D au contexte 3D et mettent en jeu notamment les amplitudes de la transformée de Fourier et les coordonnées cylindriques, centrées sur le corps. Une description du mouvement corporel invariante à la translation et aux rotations autour de l'axe vertical est ainsi obtenue.

Deux stratégies de classification des actions sont proposées, qui utilisent respectivement l'analyse en composantes principales et l'analyse du discriminant linéaire. L'approche est testée sur la base de gestes IXMAS [**92**] (*cf.* section III.1.1) avec des scores avoisinant les 100% de reconnaissance.

Kellokumpu *et al.* [**115**] établissent leur description des séquences vidéos à partir de représentations spatio-temporelles de celles-ci. Ils proposent d'évaluer la dynamique d'une séquence vidéo à partir de modèles binaires locaux *LBP-TOP* (*Local Binary Patterns for Three Orthogonal Planes*) selon trois plans orthogonaux (Figure III.3) définis par le volume spatio-temporel que constituent les dimensions *x*, *y* et *t*. Sur chaque plan *xt* et *yt*, les modèles binaires locaux LBP peuvent être calculés pour chaque pixel. Dans leur proposition, les auteurs préfèrent un modèle d'échantillonnage elliptique autour du pixel d'intérêt au modèle d'échantillonnage circulaire qui était préconisé dans l'approche LBP originale [**116**]. La description LBP-TOP consiste à calculer ces caractéristiques LBP sur chaque plan et à les concaténer dans des histogrammes. Ces histogrammes sont utilisés en entrée de HMM [**98**] dont les états cachés prennent leurs valeurs au sein des classes d'activité d'intérêt.

L'approche, testée sur les dix activités proposées par le corpus Weizmann [**89**] (*cf.* section III.1.1), atteint des taux de reconnaissance par classe supérieurs à 95%.

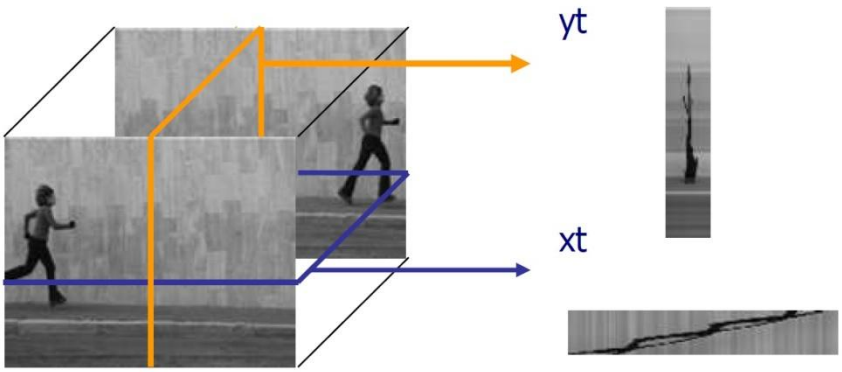


Figure III.3 Plan orthogonaux *yt* et *xt* utilisés par Kellokumpu *et al.* (source : [**115**]) pour le calcul des LBP.

Dans [**117**], Rapantzikos *et al.* proposent une représentation volumétrique multi-échelle des vidéos. Pour chaque pixel $q = (x, y, t)$ un volume spatio-temporel $V(q)$ de voisinage est tout d'abord construit. Ce voisinage est ensuite caractérisé en termes d'intensité, de couleur et d'orientation (évaluée à l'aide d'une dérivation de filtres gaussiens 3D séparables et orientables), et ce sur plusieurs échelles, correspondant aux différents niveaux d'une pyramide de gaussiennes. Un indicateur de saillance est ensuite calculé. Les maxima locaux des cartes de saillance obtenues sont retenus comme points caractéristiques du mouvement.

La méthode des *k* plus proches voisins (*k-Nearest Neighbors algorithm* : *k-NN*) est utilisée à des fins de classification d'actions. Un taux de reconnaissance moyen de 88.3% est obtenu sur le base de test KTH [**88**] (*cf.* section III.1.1).

Laptev et Lindelberg [**118**] proposent d'extraire un ensemble de points d'intérêt spatio-temporels, appelés STIP (*Spatio-Temporal Interest Points*) de vidéos à l'aide d'une extension 3D du détecteur de Harris [**119**]. Un descripteur dédié des STIP est ensuite proposée dans [**93**]. Pour chaque STIP, un volume de voisinage est défini (dont les dimensions sont fonctions de l'échelle de détection utilisée). Dans chaque cuboïde résultant, des histogrammes de gradients orientés (*histogram of oriented gradient* : *HOG*) et de flot optique (*histogram of optical flow* : *HOF*) sont calculés. Ces histogrammes sont concaténés pour constituer le descripteur vidéo. Les descripteurs sont ensuite utilisés pour construire un dictionnaire de

clés (*spatiotemporal bag-of-features* : *BoF*), obtenu avec l'algorithme k-means [**110**]. Après assignation de chaque caractéristique à une clé, il est possible d'utiliser la représentation finale du mouvement dans un cadre d'apprentissage supervisé à base de SVM [**104**] pour classer les séquences vidéo au sein de différentes catégories d'actions.

Les résultats obtenus sur le KTH dataset [**88**] (*cf.* section III.1.1) s'élèvent à 91.8% de taux de reconnaissance.

Dans [**103**], Kaâniche et Brémond exploitent des histogrammes de gradient orientés pour construire des signatures du mouvement local appelées LMS (*Local Motion Signature)*) dédiées à la description de vidéos. L'algorithme k-means [**110**] est appliqué pour déterminer un dictionnaire initial de signatures. L'algorithme de maximisation de l'information mutuelle M.M.I. (*Maximization of Mutual Information* [**120**]) est ensuite appliqué afin de réduire la taille du dictionnaire de mots. Cela permet d'obtenir une représentation par mots vidéos, utilisée pour des objectifs de reconnaissance d'actions à l'aide d'un algorithme des *k* plus proches voisins.

L'approche a été évaluée sur trois corpus de gestes : l'un émane d'eux-mêmes, les deux autres sont les bases de test KTH [**88**] et IXMAS [**92**] (*cf.* section III.1.1). Les F-scores obtenus sur les bases KTH et IXMAS sont remarquables, et respectivement égaux à 97.15% et 88.07%.

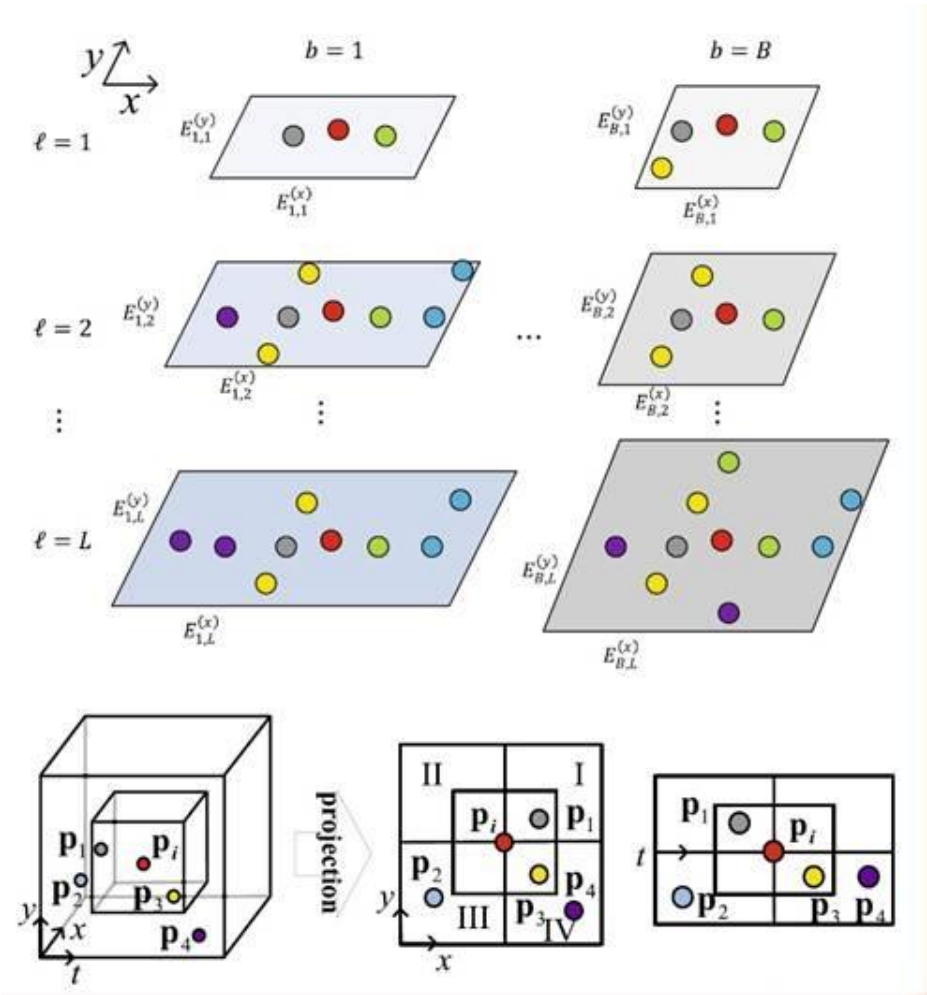


Figure III.4 Illustration de l'approche de Li *et al.* (source : [**121**]) – en haut : échelonnage de différents cuböides selon des couches pyramidales (a) ; en bas à gauche : distribution spatio-temporelle de STIP autour d'un point p_i dans un cuboïde *(b, l)* (b) ; en bas à droite : projections spatiale et temporelle du cuboïde et illustration de la subdivision de l'espace en N=4 sous-parties (c).

Dans [**121**], Li *et al*. proposent aussi une méthode pour la reconnaissance d'actions par mots vidéos. Pour chaque trame vidéo, les points d'intérêt spatio-temporels (STIP) sont tout d'abord détectés. Pour chaque point p_i, sa caractéristique spatio-temporelle locale f_i est définie à partir d'un histogramme de gradient et d'un histogramme de flux dans son voisinage. Les caractéristiques spatio-temporelles locales sont groupées avec l'algorithme k-means [**110**] pour constituer un dictionnaire $D = \{d_1, d_2, \dots, d_C\}$ de mots visuels. Chaque point STIP se voit attribuer un vecteur $v_i = [p(d_1/f_i), p(d_2/f_i), \dots, p(d_C/f_i)]$ de coefficients de codage, où $p(d_j/f_i)$ correspond à la probabilité pour la caractéristique spatio-temporelle f_i d'appartenir au mot visuel d_j. Les probabilités $p(d_j/f_i)$ sont construites de manière à prendre en compte les erreurs d'approximation d'une caractéristique f_i par un mot du dictionnaire d_j.

Ainsi, un point STIP est représenté par une signature $p_i = \{x_i, y_i, t_i, f_i, v_i)$ où x_i, y_i et t_i représentent les coordonnées spatio-temporelles du point considéré. Dans le voisinage de chaque STIP, une série de *B* cuboïdes 3D est définie. Ces cuboïdes dits « de contexte » sont chacun échelonnés en *L* différents cuboïdes de forme similaire de façon à prendre en compte un voisinage plus ou moins grand. Ainsi, un cuboïde *b* se présente sur plusieurs couches pyramidales à différentes échelles (Figure III.4.a). Chaque cuboïde contextuel *(b, l)* est projeté sur les trois sous-espaces *xt*, *yt* et *xy* (Figure III.4.b et Figure III.4.c). Chacune de ces projections se retrouve subdivisée en *N* sous-parties, au sein desquelles sont calculés des histogrammes de densité en fonction des STIP qui s'y trouvent projetés. Au final, un histogramme de densité globale qui résume les informations de ces histogrammes de contexte calculées pour les projections *xt*, *yt* et *xy* des cuboïdes $\{(b,l)\}_{b\in\{1,B\},l\in\{1,L\}}$ est proposé comme caractérisation du contexte d'un point STIP p_i. La concaténation de tels histogrammes fournit une représentation de l'intégralité de la vidéo. Li *et al.* utilisent ces représentations pour nourrir des SVM multi-classes non linéaires [**104**] à des fins de reconnaissance d'action.

L'approche est testée sur quatre bases de gestes, incluant KTH [**88**], UCF sports action [**96**], HMDB51 [**94**] et ADL [**95**] (*cf.* section III.1.1), pour des résultats en termes de taux de reconnaissance globalement supérieurs à 92%.

Dans [**122**], Niebles *et al.* proposent également une méthode de reconnaissance d'actions fondée sur des mots visuels, mais requérant l'usage d'une analyse sémantique sous-jacente probabiliste (*probabilistic Latent Semantic Analysis* : *pLSA*). Tout comme l'approche précédente, cette méthode s'appuie sur l'extraction de points d'intérêt spatio-temporels, effectuée ici à l'aide de filtres linéaires séparables à un seul niveau d'échelle. Les gradients selon les directions *x*, *y* et *t* sont calculés pour différentes échelles dans les zones d'intérêt détectées, et forment un vecteur descriptif dont la dimensionnalité est réduite par une ACP. Les descripteurs sont calculés sur l'intégralité du corpus d'apprentissage, et regroupés par l'emploi de l'algorithme k-means [**110**] dans des clusters pour constituer le vocabulaire.

Pour un corpus de séquences *M* vidéos $\{d_0, d_1, \dots, d_M\}$ où sont répertoriées K actions $\{z_0, z_1, \dots, z_K\}$ et un dictionnaire de *V* mots $\{w_0, w_1, \dots, w_V\}$, le cadre de la pLSA est paramétré de la façon suivante :

- $m(w_i, d_j)$ correspond au nombre d'occurrences du mot w_i dans la vidéo d_j ;
- $\{P(z_k|\, d_j)\}_{k=1}^{K}$ dénote les coefficients de mélange d'actions pour la vidéo d_j ;
- $\{P(w_i|\, z_k)\}_{i=1}^{V}$ est le vecteur d'aspects de l'action z_k ;
- $\{P(w_i|\, d_j)\}_{i=1}^{V}$ est la distribution des mots spécifique à la vidéo d_j, de telle sorte que :

$$P(w_i|\, d_j) = \sum_{k=1}^{K} P(z_k|\, d_j)\, P(w_i|\, z_k) \qquad \text{(III.1)}$$

L'estimation des paramètres du modèle revient à maximiser la probabilité qu'un mot w_i apparaisse dans une vidéo d_j d'autant plus qu'il apparait factuellement, c'est-à-dire de maximiser la fonction objective suivante :

$$\prod_{i\in\{1,V\},j\in\{1,M\}} P(w_i|\, d_j)^{m(w_i,d_j)} \qquad \text{(III.2)}$$

Lors de l'étape de test, l'action z_k reconnue est celle qui maximise $P(z_k|\, d_{test})$

L'approche est testée sur les bases de test KTH [**88**] et Weizmann [**89**] (*cf.* section III.1.1) avec des taux de reconnaissance respectifs de 83% pour 1500 mots et de 90% pour 1200 mots.

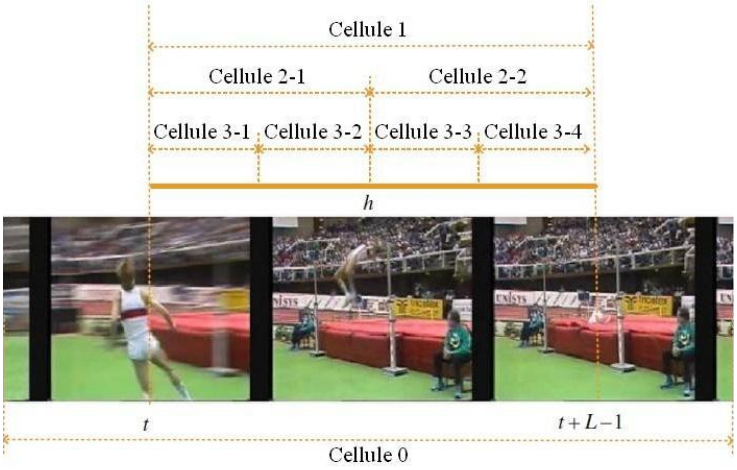


Figure III.5 Illustration du modèle par pyramide de cellules temporelles présenté par Wu *et al.* (source : [**123**]) pour un sous-segment vidéo, c'est-à-dire une hypothèse h, de taille fixée L.

Wu *et al.* [**123**] proposent une méthode de reconnaissance des gestes fondée sur des sacs de mots qui est directement liée à une segmentation de la vidéo. Pour une vidéo donnée, ils considèrent tous les sous-segments possibles d'une taille *L* fixée. La structure temporelle de tels sous-segments est modélisée par une pyramide de cellules temporelles à trois niveaux. Le nombre total est de sept cellules, auxquelles se rajoute une cellule de niveau « zéro » correspondant à l'intégralité de la vidéo (Figure III.5). Un tel sous-segment prend le nom d'« hypothèse ».

Chaque cellule d'hypothèse est décrite dans sa globalité à l'aide de cinq descripteurs :

- un descripteur de forme de trajectoire ;
- des histogrammes de gradient spatial ;
- des histogrammes de flot optique ;
- deux histogrammes de frontière du mouvement, selon la hauteur et selon la largeur.

Pour chaque hypothèse (*e.g.*, sous-segment) pour chacune des huit cellules de l'hypothèse et pour chacun des cinq types de descripteurs calculés dans la cellule considérée, un sac de *V* mots (*bag-of-words*) est calculé à l'aide de l'algorithme k-means [**110**]. Pour chaque descripteur de la cellule, une fonction de cartographie (*feature mapping*) est calculée. Elle fait ressortir les caractéristiques d'intérêt sous forme d'un vecteur de dimension *3V*. Une cellule a alors pour représentant *5*3V = 15V* valeurs. Le vecteur caractéristique pour l'hypothèse correspond à la concaténation des descripteurs des huit cellules pyramidales.

Pour chaque classe d'action *i*, *M* modèles temporels sont employés, chacun possédant une taille L_m différente, et par conséquent un nombre d'hypothèses H_m qui en dépend. Le vecteur caractéristique pour un modèle et une hypothèse donnés est utilisé en tant qu'observation dans le cadre d'une méthode de reconnaissance d'actions fondée sur des SVM [**104**].

L'approche est testée sur deux bases de données, dont la base HMDB51 [**94**] (*cf.* section III.1.1), et atteigne des taux de reconnaissance d'environ 84%.

Dans [**124**], Ballas *et al.* se proposent également d'utiliser l'information spatio-temporelle contenue dans des vidéos pour en déterminer le contenu. L'hypothèse de départ est que l'information permettant de discriminer différentes actions n'est pas uniformément distribuée dans le domaine spatio-temporel de la vidéo. Ainsi, des actions aussi proches que *jouer au football* et *courir* se différencient davantage par la présence d'une balle aux pieds de l'acteur de la scène que par des différences en termes de structuration du mouvement. Par ailleurs, les régions d'intérêt que des algorithmes permettent d'extraire à chaque instant d'une action donnée peuvent fortement varier en taille et en position à mesure que celle-ci est exécutée.

Les divers indices de saillance spatio-temporelle (ici, détections d'angles, de lumière et de vitesse) mettent généralement en valeur des régions spatio-temporelles différentes dont il s'agit d'agréger l'information. Cette agrégation a lieu comme suit : les informations de saillance sont tout d'abord extraites de la vidéo ; pour chacune de ces informations de saillance, les caractéristiques locales sont triées de celle qui présente la plus forte saillance à celle qui présente la plus faible ; ces caractéristiques locales servent alors à segmenter des sous-régions de saillance fixes, définies à différentes échelles, et qui forment ce que les auteurs appellent des « primitives de structure ». Ces primitives définissent alors la signature d'une vidéo pour l'étape d'entrainement. L'apprentissage consiste alors à déterminer la combinaison de primitives de structure la plus optimale à caractériser chaque action. Cela est réalisé à l'aide de « Weighted SVM », capables de pondérer par des poids spécifiques différents groupes de caractéristiques.

L'approche de Ballas *et al.*est testée sur plusieurs corpus 2D, dont les KTH [**88**] et HMDB [**94**] datasets (*cf.* section III.1.1), avec des taux de précision atteints respectivement égaux à 94.6 et 51.8%

Un cadre théorique dédié à la reconnaissance des gestes des mains est présenté dans [**125**]. Alon *et al.* y proposent une approche exploitant une segmentation du mouvement en sous-mouvements.

La main est tout d'abord segmentée à partir des flux vidéo en utilisant soit des modèles de peau, soit des gants de couleurs. Pour une trame donnée *j* sera donc extraite une série de candidats à la position de la main : $Q_j = (Q_{j,1}, Q_{j,2}, \dots, Q_{j,K})$, chaque candidat *k* se présentant sous la forme : $Q_{j,k} = (x_{j,k}, y_{j,k}, u_{j,k}, v_{j,k})$, où $(x_{j,k}, y_{j,k})$ dénote la position sur l'image et $(u_{j,k}, v_{j,k})$ le flot optique. A une sous-séquence vidéo de *n* trames correspondra donc une séquence de lots de positions candidates : $Q = (Q_1, Q_2, \dots, Q_n)$.

Un modèle de geste *g* est quant à lui représenté par une succession de *m+1* états : $M^g = (M_0^g, M_1^g, \dots, M_m^g)$, où chaque état *i* du geste *g* est modélisé par une distribution gaussienne : $M_i^g \sim (\mu_i^g, \Sigma_i^g)$.

L'objectif de l'approche est d'évaluer la proximité entre d'une part, une sous-séquence de séries de candidats à la position de la main $Q = (Q_1, Q_2, \dots, Q_n)$ extraites d'un segment vidéo, et d'autre part un modèle de geste $M^g = (M_0^g, M_1^g, \dots, M_m^g)$.

Une telle correspondance se mesure à l'aide d'un chemin de déformation spatio-temporelle (*warping path*) qui se présente sous forme d'une série $W = (w_1, w_2, \dots, w_T)$, où l'indice $w_t = (i_t, j_t, k_t)$ est un triplet indiquant selon quel degré il est possible de faire correspondre l'état i_t du modèle M^g, c'est-à-dire l'état

$M_{i_t}^g$, avec le k_t-ième candidat à la position de la main à l'instant j_t, c'est-à-dire le candidat Q_{j_t,k_t}. Les auteurs expriment ce degré de correspondance comme la distance de Mahalanobis de la position candidate au geste g :

$$dist\left(M_{i_t}^g, Q_{j_t,k_t}\right) = \left(Q_{j_t,k_t} - \mu_{i_t}^g\right)^T . \left(\Sigma_{i_t}^g\right)^{-1} . \left(Q_{j_t,k_t} - \mu_{i_t}^g\right) \quad \text{(III.3)}$$

La déformation W^* qui minimise la somme :

$$\sum_{t=1}^{T} dist\left(M_{i_t}^g, Q_{j_t,k_t}\right) \quad \text{(III.4)}$$

est celle qui permet d'associer à la sous-séquence $Q = (Q_1, Q_2, \ldots, Q_n)$ le modèle de geste adéquat g^*. Sous réserve que cette sous-séquence Q corresponde bien à la réalisation du geste g^*, la déformation W^* fournit donc, outre le geste g^*, sa meilleure segmentation spatiale du geste (*e.g.*, les positions candidates idéales de la main à chaque instant), ainsi que sa meilleure segmentation temporelle.

Alon *et al.* testent leur approche sur deux corpus, composés respectivement de gestes symbolisant des chiffres et d'occurrences de la langue des signes des sourds et malentendants (ASL - *American Sign Language*).

Dans [**126**], Martinez *et al.* proposent une représentation du mouvement par mixtures de gaussiennes à des fins de reconnaissance de geste du langage des signes. A partir du flux vidéo peuvent être calculés deux types de trajectoires du corps : dense et semi-dense, la première étant basée sur le flot optique et la seconde sur la saillance de points. A chaque trajectoire de point suivie au cours du temps se trouve assigné un certain nombre de caractéristiques cinématiques : direction et module de la vitesse, et courbure. Ces caractéristiques sont modélisées par des mixtures de distributions gaussiennes dont les paramètres sont estimés récursivement à chaque instant *t* du geste. L'usage d'un algorithme k-means classique [**110**] permet d'extraire de ces paramètres de distributions gaussiennes des « mots-cinématiques » représentatifs du mouvement corporel. Des dictionnaires de mots-clés sont définis dans différentes sous-régions d'une partition pyramidale de chaque frame qui satisfait le caractère multi-échelle de la représentation du mouvement. Le descripteur d'un échantillon vidéo se compose au final d'un histogramme comptant le nombre de fois où chaque centroïde cinématique (*e.g.*, mot-clé) est le plus proche des caractéristiques calculées pour ledit échantillon. Cet histogramme sert à entrainer des SVM [**104**] qui lors de l'étape de reconnaissance sont utilisés comme classifieurs.

Une première expérience permet aux auteurs de tester leur méthode sur un corpus de signes représentants des dates (4 jours et 5 mois). Ils obtiennent des taux de reconnaissance s'élevant respectivement à 81.5% dans le cas de gestes signifiant des jours, et à 80% dans le cas de gestes signifiant des mois. Par ailleurs, ils testent la capacité de leur représentation à discriminer les différents gestes du KTH dataset [**88**] (*cf.* section III.1.1), pour lequel ils obtiennent un taux de précision égal à 90.3%.

Dans [**127**], Pedersoli *et al.* proposent une nouvelle méthode pour la reconnaissance des poses des mains et des gestes effectués avec les mains, fondés sur les cartes de profondeur fournies par une caméra Kinect. A partir de telles cartes, un premier algorithme de segmentation d'image par *mean-shift* [**128**] et une procédure dédiée de détection de la paume fournissent la région support de la main. Un système de reconnaissance de la pose de la main, à base de SVM [**104**], est entrainé avec des descripteurs issus d'un filtrage de Gabor.

Les résultats obtenus en termes de taux de reconnaissance sur le corpus de langue des signes américaine (*American Sign Language* : *ASL*) atteignent 97% pour certaines lettres. La reconnaissance de gestes requiert quant à elle une décomposition du mouvement en séquences « tenue-mouvement-tenue » (« *hold-movement-hold* »). Pour chaque séquence, la trajectoire du centroïde de la main définie une série angulaire que les auteurs utilisent en entrée de HMM [**98**]. La méthode est testée sur un corpus de seize gestes effectués par dix utilisateurs différents. Pas moins de cinq de ces gestes ont des taux de reconnaissance supérieurs à 90%, tandis que pour trois d'entre eux, ils restent inférieurs à 50%.

Comme nous venons de le voir, de nombreuses approches dédiées à la détection de saillance spatio-temporelle dans les vidéos 2D se montrent aptes à saisir des informations cruciales présentes dans les contenus gestuels. Plusieurs méthodes suggèrent même qu'il est possible d'élargir la notion de « pose », initialement centrée autour de déterminations strictement spatiales et globales (*cf.* section III.1.1), aux aspects locaux du mouvement, de telle sorte qu'on puisse le réduire à ses indices spatio-temporels clés.

En outre, nous estimons que l'apparition récente de technologies de capture capables de rendre à chaque instant la pose 3D du corps peut nous aider dans ce sens. En effet, la précision et la robustesse qu'offrent ces technologies tendent de plus en plus à réduire les difficultés de suivi corporel que posent de façon accrue les vidéos 2D, et permettent d'envisager une description plus granuleuse des aspects dynamiques et structurels du geste.

Au contraire de ce qui semble transparaître dans certaines approches récentes, à savoir que le simple suivi de trajectoires 3D d'articulations clés du corps se suffirait quasiment à lui-même en vue de la reconnaissance d'actions [**129**] [**130**] [**131**] [**132**] [**133**], nous avons estimé que l'apparition de la Kinect nous invitait à nous tourner vers une compréhension plus approfondie du geste, capable de prendre en compte ses aspects structurels voire linguistiques, ce qui étaient hautement difficile à envisager dans le contexte traditionnel de flux vidéo 2D. Analysons, dans la section suivante, les approches présentant un intérêt à cet égard.

III.1.4. Représentations structurelles 3D et morphologie du mouvement

Pour l'ensemble des approches présentées dans cette section, l'acquisition des séquences de geste est effectué à l'aide d'une caméra de type Kinect. Les informations disponibles en sortie de ce type de caméra sont donc :

- la position 3D des articulations correspondant à un squelette humanoïde (*les détails relatifs à ces articulations sont donnés au début de la* section IV) ;
- la carte de profondeur ;
- les images de texture RGB des séquences acquises.

Les méthodes présentées dans la suite exploitent exclusivement les positions 3D des articulations ou des représentations du corps similaires ou semblables à celles que fournissent les capteurs Kinect (*i.e.*, des trajectoires d'articulations de référence du corps).

Xia *et al.* [**37**] proposent une représentation des poses 3D fournies par une caméra Kinect, fondée sur une représentation de la trajectoire des articulations du squelette en coordonnées sphériques dans un repère centré au niveau du pelvis (Figure III.6). Les échantillonnages respectifs de l'espace de l'angle d'élévation en 7 valeurs et de celui de l'angle d'azimut en 12 valeurs produisent une partition en $n = 84$ bins. Ainsi à chaque instant t, chacune des articulations considérées se voit distribuée sur ces n valeurs. Pour chacune, des poids de votes sont donnés aux bins géométriquement voisins de sa position selon une

distribution gaussienne (les votes pour l'angle d'inclination et pour l'angle d'azimut sont calculés séparément, car les deux dimensions sont indépendantes). L'histogramme de *n* valeurs qui cumule les votes pour toutes les articulations décrit la posture à chaque instant.

Un dictionnaire de poses-clés de mouvement est constitué à l'aide de l'algorithme de clustering k-means [**110**], appliqué à un ensemble de séquences d'apprentissage. Chaque pose peut alors être référée au mot visuel le plus proche. Les séquences de mots visuels qui en résultent sont utilisées comme observations dans une méthode de reconnaissance d'actions par HMM [**98**] dont les états cachés sont lesdites actions.

L'approche est testée sur le corpus MSR Action 3D [**38**] (*cf.* section V.1) avec des taux de reconnaissance globalement supérieurs à 93%, sauf dans le cas de l'approche inter-sujet où les gestes de la moitié des participants ont été utilisés pour l'entrainement et les gestes de l'autre moitié pour l'étape de test.

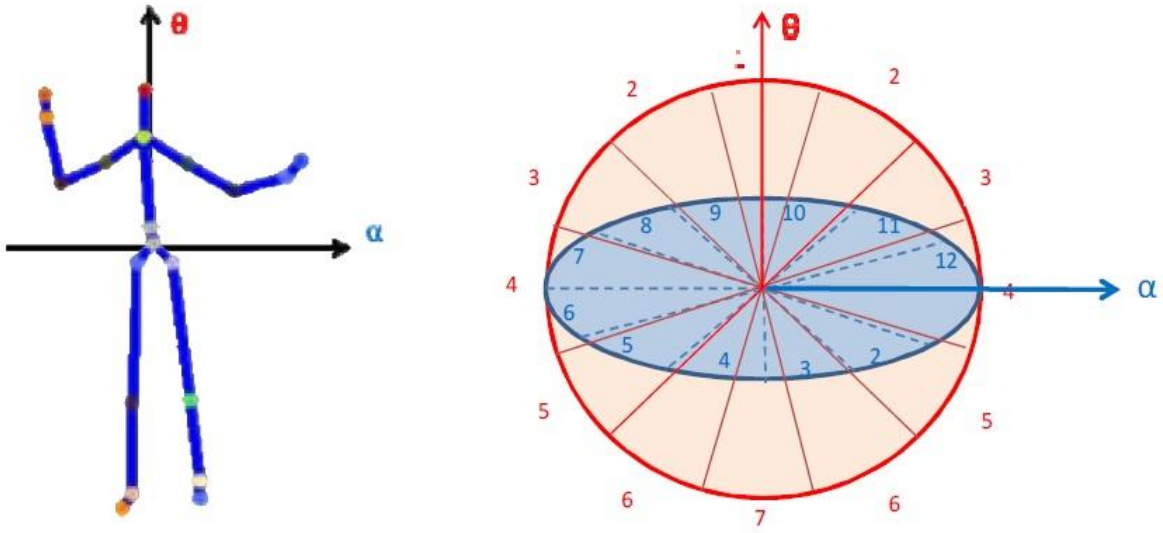


Figure III.6 Changement de repère et coordonnées sphériques proposées par Xia *et al.* (source : [**37**]).

Dans [**134**], Hussein *et al.* proposent une méthode pour la reconnaissance d'actions qui exploite la matrice de covariance des positions des articulations du squelette fournies par une caméra Kinect au cours du temps. Les descripteurs de mouvement globaux utilisés cumulent un ensemble de matrices de covariance calculées à différentes échelles, sur des sous-séquences hiérarchiques inspirées par les méthodes d'échelonnage spatial pyramidal. Le nombre de niveaux hiérarchiques de sous-séquences peut être paramétré.

L'approche est testée sur trois bases de gestes : MSR Action 3D [**38**], MSRC-12 [**36**] (*cf.* section V.1) et HDM05-MoCap [**90**] (*cf.* section III.1.1), avec comme taux de reconnaissance de 90.5%, 98.7% (meilleur résultat parmi ceux obtenus pour différentes stratégies) et 95.4% (idem), respectivement.

Dans [**135**], Jiang *et al.* proposent un modèle de mouvement fondé sur une représentation hiérarchique des positions 3D des articulations rendues par une caméra Kinect. Dans un premier temps, ils assignent chaque geste à un groupe, en fonction des parties du corps en mouvement au cours de l'action. Quatre parties sont considérées, qui sont le bras gauche, le bras droit, le bas du corps, et la zone « restante » englobant la tête, le torse et les hanches – pour un nombre maximal de groupes possibles égal à seize, si l'on envisage toutes les configurations possibles selon lesquelles chacune de ces quatre zones contribue

ou non à la réalisation du geste ($2^{nombre\ de\ parties} = 2^4 = 16$ configurations possibles). Ensuite, pour chacun de ces groupes spécifiques à un type de mouvement, un classifieur *k-NN* est entrainé en vue de l'étape future de reconnaissance d'actions. Ce classifieur prend comme entrée les mouvements des articulations et leurs positions relatives par rapport à une articulation stable pour une trame donnée. Des sacs de mots sont utilisés au sein de chaque groupe pour réduire la dimensionnalité du problème.

Lors de l'étape de test, le geste est en premier lieu affecté à un groupe, et le classifieur *k-NN* dudit groupe est utilisé pour donner au geste son label. Le label final est obtenu en fusionnant la classification obtenue pour chaque trame, avec néanmoins un système de pondérations qui permet de mettre de l'emphase sur les trames représentatives, en gérant notamment les erreurs de suivi du squelette dues aux occlusions.

L'approche est testée sur la base UTKinect-HumanDetection [**37**] (*cf.* section V.1), avec des résultats en termes de taux de reconnaissance proches de 97%. Pour le corpus MSRC-12 [**36**] (*cf.* section V.1), les taux de reconnaissance avoisinent les 100% pour certains gestes, bien que certaines confusions demeurent (les taux les plus faibles étant aux alentours de 85%).

Dans [**136**], Etemad et Arya introduisent une nouvelle méthode de déformation temporelle pour rendre plus aisée la comparaison entre les gestes. Chaque séquence de mouvement est décrite par des séries angulaires consacrées aux articulations 3D. Le principe consiste à mettre en œuvre une procédure d'appariement de deux séries à comparer, fondé sur une décomposition de chaque séquence en segments. Chaque segment source est alors déformé pour atteindre un segment cible. La liste des degrés de déformation propres aux segments est calculée en maximisant une fonction objectif intégrant le coefficient de corrélation linéaire de Pearson (*Pearson's linear correlation coefficients* : *PCC*) entre les trajectoires des articulations initiales et les trajectoires objectifs, ainsi que les différents poids associés auxdites articulations. La distance normalisée et les résultats de corrélation obtenus pour la déformation de différentes actions issues du corpus Carnegie Mellon University [**137**] se montrent supérieurs à ceux obtenus avec d'autres techniques de déformation temporelle (déformation temporelle uniforme, déformation temporelle dynamique….).

Dans [**138**], Zhao *et al.* proposent une nouvelle approche qu'ils désignent sous le nom de « *Structured Streaming Skeletons* » (*SSS*), qui consiste à représenter un geste par une combinaison de mouvements de parties du corps. L'approche exploite l'ensemble des séries temporelles scalaires qui représentent les distances entre deux articulations données au cours du temps. Ces distances sont normalisées par rapport à la longueur du chemin qui sépare les deux articulations dans le squelette. Un tel modèle présente l'avantage d'être indépendant de la position du corps dans l'espace, de sa taille et de son orientation. Le mouvement se retrouve ainsi décrit dans un espace à $\frac{20\times19}{2} = 190$ dimensions (dans la mesure où une caméra Kinect donne accès à 20 articulations de référence).

Après avoir manuellement segmenté les gestes de leur propre corpus, Zhao *et al.* constituent pour chacune des 190 dimensions du geste (*e.g.*, pour chacun des modèles) un dictionnaire de clés du mouvement à l'aide d'un algorithme de clustering spectral [**139**]. Les gestes du corpus peuvent ainsi être représentés sur des espaces réduits pour entrainer des modèles d'actions et classifier des mouvements corporels selon ces modèles.

Yang *et al.* [**140**] proposent une représentation du mouvement qui reprend et étend des éléments de celle de Zhao *et al.* [**138**]. A chaque trame temporelle *t*, le geste est également décrit par une série de valeurs qui correspondent chacune à une distance entre deux articulations différentes. A ces

caractéristiques de posture, ils ajoutent une description du mouvement, constituée de toutes les distances possibles entre la position d'une articulation à l'instant *t* et la position d'une articulation à l'instant *t-1* (ce qui ajoute $20^2 = 400$ valeurs à la représentation). Enfin, ils intègrent l'expression du décalage par rapport à la position initiale (considérée comme une position neutre) par les distances possibles entre une articulation à l'instant *t* et une articulation à l'instant initial (ce qui ajoute $20^2 = 400$ valeurs). Les auteurs normalisent les descripteurs et appliquent une ACP pour réduire la taille et la redondance de leurs données. Cela conduit à une représentation dite « *compact Eigen Joints representation* ». Les auteurs utilisent un classifieur naïf de Bayes qui s'appuie sur la méthode des plus proches voisins (*Naïve-Bayes-Nearest-Neighbors classifier*) et qu'ils entrainent en vue de la reconnaissance d'actions. Les taux de reconnaissance obtenus sont supérieurs à 95% sur le corpus MSR Action 3D [**38**] (*cf*. section V.1). La même représentation de mouvement est utilisée dans l'approche présentée dans [**141**], où elle est fusionnée avec des caractéristiques spatio-temporelles semblables à celles qui sont présentées par Wang *et al.* dans [**108**].

Dans leur étude consacrée à la segmentation et à l'analyse des gestes de clarinettistes, Caramiaux *et al.* [**142**] s'appuient sur l'hypothèse sous-jacente qu'un geste musical correspond à une séquence d'actions primitives de bases et que sa structure est liée à celle de ladite musique. L'espace des actions possibles est alors doté d'une structure linguistique à plusieurs niveaux. Caramiaux *et al.* tentent de démontrer qu'il est possible d'utiliser un dictionnaire de modèles pour caractériser la structure temporelle de la gestuelle du clarinettiste. Quatre HMM [**98**] sont entrainés, chacun exploitant un dictionnaire de taille différente, composé d'un ensemble de primitives du mouvement. Les mouvements de la clarinette sont décrits par l'angle azimutal et l'angle d'élévation dans un système de coordonnées sphériques solidaire au corps de la clarinette. Les états cachés des modèles correspondent aux primitives (classes) de mouvement et à la durée de ces segments-primitives. Tout ajout d'une nouvelle primitive dans un dictionnaire nécessite de s'assurer au préalable que la log-vraisemblance de l'observation relativement au nouveau modèle que constitue l'ajout de la primitive est en augmentation. Dans le cas contraire, le symbole est jugé inutile et n'est pas intégré. La comparaison entre les prédictions (classes) des différents modèles de Markov et les signaux originaux (mouvements représentés dans le système sphérique) montrent :

- qu'un nombre élevé de symboles dans le dictionnaire ne diminue pas nécessairement la distance entre le modèle et les signaux,
- que la prédiction ne tend à s'améliorer que si les primitives du mouvement qui s'ajoutent dans le dictionnaire sont pertinentes (c'est-à-dire représentatives de l'information).

Par ailleurs, les auteurs constatent qu'une forte concentration de symboles de courte durée, qui correspondent à des articulations dans le mouvement, est détectée entre des séquences récurrentes de symboles longs, ce qui est dépeint parfaitement le profil d'une phrase musicale.

Une telle représentation symbolique du geste fournit un puissant outil d'indexation de motifs dans les signaux gestuels continus.

Une approche similaire, qui consiste à segmenter le geste en primitives d'actions est également proposée par Guerra-Filho and al. [**143**]. L'analogie avec la linguistique est signalée par l'appariement entre les concepts de *phonèmes* et de *kinétèmes*. Il s'agit donc de s'intéresser à la phonologie du mouvement. Pour chaque actionneur d'intérêt (exemple : un doigt), trois signaux correspondant chacun à un angle de rotation peuvent être suivis au cours du temps. Pour chaque signal, le signe (positif ou négatif) de la vitesse et de l'accélération définissent un système à quatre états (vitesse positive/accélération positive, vitesse positive/accélération négative, vitesse négative/accélération

positive, vitesse négative/accélération négative) parmi lesquels le signal prend une valeur à chaque instant (et pour un certain temps). Les séries angulaires de chaque actionneur donnent donc lieu à un « *actiongram* » qui relate l'aspect qualitatif (état) et l'aspect quantitatif (durée de persistance dans l'état, et déplacement angulaire au cours de la persistance dans l'état) du mouvement. Les auteurs discutent la pertinence de leur modèle au regard de cinq propriétés que sont :

- la *compacité*, qui consiste à décrire le mouvement avec un minimum d'information,
- l'*invariance* de la description selon la vue, qui témoigne donc du caractère absolu de la représentation du geste,
- la *reproductibilité* intra-personnelle et extra-personnelle, qui mesure la capacité d'un système à fournir la même représentation d'un geste instancié plusieurs fois par la même personne, et par différentes personnes,
- la *sélectivité*, où la capacité de la représentation à fournir une distinction entre des actions différentes,
- la *re-constructivité*, où la possibilité offerte de reconstruire au moins partiellement le signal original à partir de la représentation.

Ils proposent ensuite une méthode d'apprentissage des morphologies des différentes actions à partir de leurs représentants, fondée sur des grammaires parallèles synchronisées (*Parallel Synchronous Grammar System* : *PSGS*).

Dans [**144**], Liutkus *et al.* se proposent de décomposer le mouvement dansé en primitives de mouvement. A partir de dix-sept articulations acquises avec une caméra Kinect, les auteurs cherchent à décomposer la trajectoire dansante du squelette en quatre composantes latentes : les deux premières dénotent le caractère répétitif du mouvement, les deux autres incarnent respectivement ses variations lentes et son caractère imprévisible.

Chaque composante est alors modélisée par une distribution gaussienne de moyenne nulle et dont covariance est supposée séparable en deux composantes, l'une spatiale et l'autre temporelle. Liutkus *et al.* estiment qu'une telle décomposition du mouvement rend possible la reconnaissance de gestes, en fournissant des approximations satisfaisantes. Néanmoins, les auteurs ne proposent pas une vraie étude de ces aspects avec résultats de reconnaissance objectifs.

L'analyse des approches présentées prouve que la connaissance précise des lieux et trajectoires du mouvement offre des possibilités foisonnantes en termes de caractérisations structurale et processuelle du geste. La fiabilité des squelettes conduit à l'élaboration de modèles intermédiaires du mouvement corporel qui pouvaient difficilement voir le jour dans le cas de vidéos 2D. Par ailleurs, cette fiabilité visuelle rend la prise en compte des aspects locaux et dynamiques du geste toujours plus congruente avec la réduction de la représentation à des poses ou postures clés.

Néanmoins, les descripteurs utilisés pour définir ou extraire ces postes ou atomes sont généralement dédiés à des indices purement visuels et structurels du mouvement. Ce faisant, ils passent à côté de plus hauts niveaux de compréhension du geste, et à savoir ses aspects communicationnel, intentionnel, ou encore expressif.

D'autres champs de recherches tentent notamment de prendre en compte la caractérisation sémantique des mouvements corporels. L'hypothèse sous-jacente est alors que le contenu gestuel ne peut pas se référer à des actions spécifiques ou à des mouvements précis dans l'espace, mais aussi à son caractère intersubjectif. Ces aspects sont présentés dans le paragraphe suivant.

III.2 Activité de groupe et créativité

Des études récentes se consacrent aux activités de groupes. Les mouvements corporels des protagonistes y aident à détecter l'établissement de la hiérarchie lors d'interactions entre individus, et plus généralement à étudier la façon dont ces individus coordonnent leurs actions. Les notions de *hiérarchie* et de *leadership* sont à mettre en relation avec la dimension de *domination* du modèle PAD (*Pleasure-Arousal-Dominance*) d'émotions de Mehrabian [**54**], ou avec les espaces de *contrôle*/*soumission* et *propension*/*obstruction* proposés par Scherer [**46**]. Si l'on se réfère aux travaux de Sander *et al.* [**60**], l'établissement de hiérarchies a trait à la *puissance* et au *contrôle*, qui sont des éléments constitutifs de la formation de l'émotion.

Pour Glowinski *et al.* [**85**], la musique classique est un champ dans lequel l'analyse du leadership présente un intérêt spécifique, dans la mesure où la « vérité terrain » de ce leadership est posée dès le départ par la partition. Glowinski *et al.* travaillent sur l'établissement de ce leadership au sein d'un quatuor à cordes. Ils étudient les mouvements de tête (Figure III.7) – décisif lors de performances musicales pour ce qui est d'indiquer le début d'une phrase ou pour diriger un changement de vitesse – des quatre musiciens à travers différentes situations :

- répétition ordinaire ;
- répétition avec des changements de rôles entre les premier et deuxième violons ;
- répétition avec battue au métronome ;
- répétition avec demande de « sur-expressivité » – *e.g.*, d'exagération expressive volontaire ;
- situation de concert.

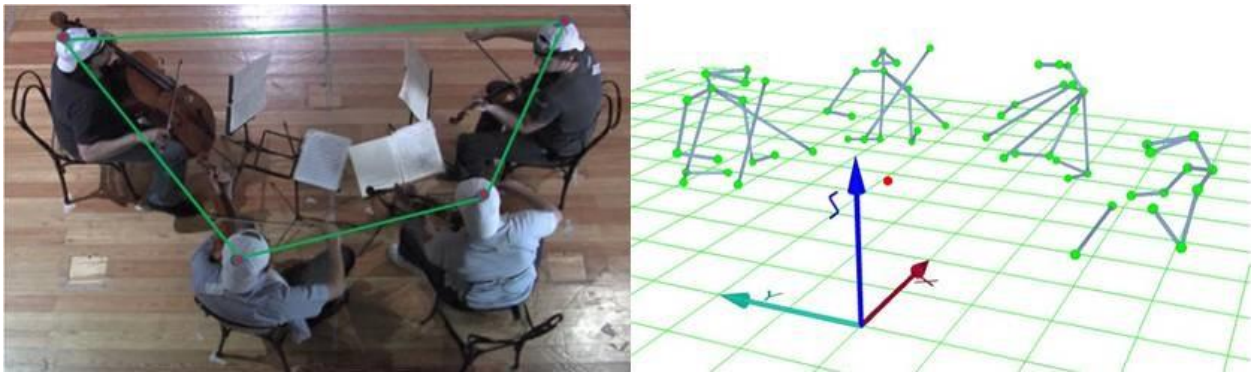


Figure III.7 Etude du leadership en musique : illustration du polygone formé par les têtes des quatre musiciens dont il est question dans [**85**] (a) ; squelettes corporels capturés lors de l'expérience présentée dans [**86**] et dont on cherche à étudier le mouvement par rapport au point de référence représenté en rouge (b).

Les auteurs analysent la complexité de cinq séries temporelles d'intérêt, que sont les séries des vitesses du mouvement de la tête pour chacun des quatre musiciens, ainsi que la série constituée à chaque instant de l'aire d'un polygone ayant pour sommet les quatre têtes (Figure III.7.a). Cette complexité est mesurée à l'aide de l'indice de « *Sample Entropy* » (*SampEn*) introduite par Richman et Moorman [**145**]. Cet indice définit la probabilité, pour deux séquences parallèles similaires sur un certain nombre de points, de le rester lors de l'ajout ou de la prise en compte d'un nouvel élément. Les auteurs démontrent que la complexité de la série temporelle du premier violon (qui assume la conduction) est globalement plus corrélée à la complexité de la série des aires du polygone, qui à son tour n'est pas corrélée à la complexité des séries des autres musiciens.

Dans [**86**], les mêmes auteurs proposent une autre « caractérisation sociale » de la musique d'ensemble. Ils s'intéressent aux distances de chacune des têtes des musiciens à une position de référence surélevée au milieu du quatuor, qu'ils appellent « centre subjectif ». Ce centre de référence est censé aider les participants à se coordonner pour atteindre une cohérence sonore (Figure III.7.b). Ils proposent notamment d'étudier comment la distance entre un musicien et le point de référence est peut être conditionnée par la distance des autres musiciens à ce point.

D'Ausilio *et al* [**87**] se focalisent pour leur part sur l'émergence de leadership dans un orchestre. Ils analysent la cinématique d'un ensemble de huit instrumentistes à cordes et de deux chefs d'orchestre et établissent des relations causales de chef à musicien et de musicien à musicien. Une étude similaire est également proposée dans [**84**].

Outre les dimensions de contrôle et de domination évoquées plus haut, ces travaux ne sont pas sans évoquer la qualité de *Relation* du modèle d'analyse de mouvement de Laban (LMA – *Laban Mouvement Analysis*), qui a été introduite et discutée à la section II.3. Ici, la caractérisation de l'intersubjectivité se limite à l'étude de corrélations entre différents mouvements.

Intéressons-nous à présent aux aspects de plus haut niveau, qui concernent notamment la dimension émotionnelle des gestes et des mouvements correspondants. L'objectif de ces travaux, présentés dans la section suivante, est notamment d'établir dans quelle mesure il est possible de détecter et reconnaitre des émotions humaines à partir de séquences de gestes.

III.3. Analyse émotionnelle du mouvement

Dans [**146**], Cimen *et al.* proposent 3 types de descripteurs de mouvement exploitant la position 3D de cinq articulations (fournies par une caméra Kinect), que sont les deux poignets, les deux chevilles et la tête, pour caractériser la posture, la dynamique du mouvement, ainsi que sa fréquence. La posture est décrite en fonction de la position des cinq articulations, de leur orientation, du volume du parallélépipède englobant du corps, et de son centre de masse. Les caractéristiques dynamiques du geste sont la vitesse, l'accélération et l'à-coup (*e.g.*, dérivée troisième du mouvement) des trajectoires des articulations d'intérêt. Enfin, les auteurs caractérisent la fréquence du mouvement à l'aide de coefficients issus de l'application de la transformée de Fourier aux trajectoires des cinq articulations.

Ces descripteurs sont utilisés dans un contexte d'analyse émotionnelle mettant en jeu quatre émotions de base : la *tristesse*, la *joie*, la *colère*, et le *calme*. Pour cela, Cimen *et al.* présentent un corpus de 96 gestes actés par 12 acteurs, chacun ayant été requis de simuler deux fois chacune des quatre émotions en marchant en ligne droite. La classification est effectuée grâce à des SVM [**104**] pour chacun des 7 « lots » de descripteurs qu'il est possible de constituer à partir des 3 types initialement présentés (*e.g.*, descripteurs de posture seuls, descripteurs de dynamique seuls, descripteurs de fréquence seuls, descripteurs de posture et de dynamique, etc.). Les taux de reconnaissance sont remarquables, s'élevant à 91%.

Dans [**147**], Bernhardt et Robinson proposent une approche pour la reconnaissance de contenus émotionnels fondée sur une représentation simplifiée des gestes en primitives du mouvement. Les séquences de squelettes sont acquises en utilisant des techniques basées sur Character Studio (plug-ins pour 3D Studio MAX, AutoDesk, Inc.) et MATLAB (The MathWorks, Inc.) [**148**]. Ces squelettes sont légèrement différents de ceux qui sont fournis par la Kinect, et se composent de 15 articulations. Les séquences utilisées dans cette étude sont au nombre de 1200, et correspondent aux mouvements *knocking*,

throwing, *lifting* et *walking*, effectués par une trentaine d'individus. Une segmentation des séquences au sens du mouvement est effectuée, à l'aide d'un échantillonnage d'une mesure d'énergie globale du corps à chaque instant, exprimée en fonction des positions et rotations des différentes articulations. Cela permet de décomposer le mouvement en un ensemble de classes d'énergie. Les primitives du mouvement sont calculées à l'aide de l'algorithme k-means [**110**] qui prend pour groupements initiaux les classes d'énergie résultant de la segmentation. Les descripteurs utilisés pour l'analyse de l'affect sont globaux et caractérisent la dynamique de la main ainsi que la distance entre la main et le corps. Ils sont utilisés pour nourrir des SVM à noyaux polynomiaux [**104**], afin de classifier les gestes selon quatre catégories émotionnelles : *émotion neutre*, *bonheur*, *colère* et *tristesse*.

Certains travaux se penchent plus spécifiquement sur les parties supérieures du corps. Dans [**149**], Balomenos *et al.* proposent une méthodologie d'analyse émotionnelle de vidéos 2D qui combine le traitement de caractéristiques faciales et celui de la position des mains, position qu'ils obtiennent par des procédures plus ou moins classiques de segmentation. Ces deux traitements ont lieu séparément. Un module dédié à la reconnaissance d'émotions par la seule analyse du visage est complété par un outil d'analyse du geste à base de HMM [**98**] et dédié à la classification en classes de gestes. La correspondance entre le type de geste détecté et l'émotion que l'individu est supposé exprimer par ce geste est établie par des règles de décision.

La méthode est validée à l'aide d'un corpus vidéo mettant en jeu 6 catégories d'émotions : la *joie*, la *tristesse*, la *colère*, la *peur*, le *dégoût* et la *surprise*. Outre son annotation en termes d'émotions, ce corpus propose les sept classes de gestes suivantes : *hand clapping – high frequency*, *hand clapping – low frequency*, *lift of the hand – low speed*, *lift of the hand – high speed*, *hands over the head – gesture*, *hands over the head – posture*, et *italianate gestures*. Ces gestes ont été réalisés par trois individus, à raison d'une quinzaine de séquences par classe.

L'entrainement des HMM conduit à un taux de reconnaissance des classes de geste globalement égal à 94.3%. L'utilisation du module de reconnaissance de geste conjointement au dispositif d'analyse du visage permet de classifier les vidéos en termes d'émotions avec un taux de reconnaissance global de 85%.

Dans [**150**], H. Gunes et M. Piccardi extraient d'images de vidéos 2D des régions incluant les deux mains et la tête. Les variations de ces régions en termes de position du centroïde, de rotation, de hauteur et de largeur, par rapport à des positions de référence, sont utilisées comme descripteurs en entrée de classificateurs bayésiens pour inférer 6 classes d'émotion que sont l'*anxiété*, la *colère*, le *dégoût*, la *peur*, le *bonheur* et l'*incertitude*. Les auteurs testent leur approche sur un corpus de 385 vidéos annotées à partir de ce lexique.

Plusieurs options sont étudiées pour ce qui est de la gestion des deux canaux (canal du corps – *e.g.*, mains – et canal du visage) :

- Une première option, monomodale, considère chaque canal indépendamment de l'autre, et donne des taux de reconnaissance d'émotion égaux à 76.4% et 89.9% respectivement pour le visage et les mains. Dans le premier cas, la *colère* tend à être confondue avec le *dégoût*. Dans l'autre, elle est confondue avec l'*anxiété*.
- Une seconde option, bimodale cette fois, consiste à concaténer les descripteurs faciaux et corporels dans un même vecteur selon une démarche dite de « *feature-level fusion* ». Le taux de reconnaissance s'élève à 94.02% et la confusion entre *colère* et *anxiété* tend à s'atténuer.

- Une troisième option, également bimodale, propose une méthode de fusion dite « *decision-level* », où les résultats obtenus sur chaque canal indépendamment sont fusionnés par suite selon une règle de décision au choix parmi diverses possibles. Les règles de somme, de produit, et de somme pondérée sont testées, et donnent des taux de reconnaissance globaux respectivement égaux à 91.1%, 87.3% et 79.7%.

Dans [**151**], Nicolaou *et al.* se proposent de reconnaître le profil émotionnel de vidéos à partir de signaux provenant de trois canaux : le visage, les épaules, et le signal audio. Ils utilisent le corpus *Sensitive Artificial Listener* (*SAL* [**152**]), qui centralise des réactions de diverses personnes aux sollicitations d'avatars, et met en jeu quatre « attitudes affectives » : *happy*, *gloomy*, *angry* et *pragmatic*. Ce corpus vidéo est annoté manuellement selon l'espace valence-éveil [**52**] (*cf.* section II.2.2). La vérité terrain ainsi obtenue, qui définit notamment les trames de début et de fin de chacune des occurrences d'un type d'émotion particulier, permet de segmenter les données.

L'acquisition des signaux visuels – une vingtaine de points pour le visage, cinq pour les épaules – requière l'usage de diverses méthodes, incluant notamment des classifieurs de Haar [**153**], une segmentation en régions d'intérêt et une extraction de caractéristiques par filtres de Gabor [**154**]. Le descripteur acoustique comporte pour sa part une quinzaine de valeurs.

Différentes stratégies de traitement de la multi-modalité sont proposées pour la classification.

- Dans une première approche, chaque modalité du comportement (*e.g.*, chacun des trois canaux) est considérée indépendamment des autres. Un classifieur est donc construit pour chaque type de signal. Les auteurs comparent les performances de SVM [**104**] à celle de *Bidirectional Long Short-Term Memory Neural Networks* (BLSTM-NN), régulièrement utilisés pour la prédiction d'affect à partir de signaux audio [**155**], et combinant les spécificités de *Bidirectional Reccurent Neural Networks* (BRNN [**156**]) et de *Long Short-Term Memory Neural Networks* (LSTM-NN [**157**]). Les premiers (BRNN) présentent l'avantage de prendre en compte les états passés et futurs au regard de la trame courante lors de la procédure d'apprentissage, en présentant chaque séquence d'entrainement en ordres ascendant et descendant. Les seconds (LSTM-NN) permettent de parer aux difficultés rencontrées avec des réseaux de neurones classiques pour ce qui est de l'apprentissage de dépendances temporelles sur le long terme, en procédant notamment par blocs de mémoire interconnectés (et non simplement par nœuds).
- La seconde méthode utilise également l'approche BLSTM-NN et prédit les émotions en suivant une stratégie de « *feature-level fusion* » ; un unique vecteur réunit l'intégralité des caractéristiques audio-visuelles.
- La troisième stratégie procède par « *model-level fusion* »; les canaux audio et visuels n'étant pas ici considérés comme indépendants.
- La dernière proposition consiste à effectuer une prédiction émotionnelle par « *output-associative fusion* ». Non seulement les canaux audio et visuels ne sont pas considérés comme indépendants, mais c'est également le cas pour les « sorties », c'est-à-dire les prédictions sur les dimensions valence et éveil.

Les auteurs proposent plusieurs conclusions à leur étude.

- Globalement, les taux de reconnaissance sont meilleurs avec la prise en compte de la multi-modalité.
- L'éveil est mieux prédit à l'aide des signaux audio qu'à partir du visage ou du mouvement des épaules. Le contraire peut être constaté dans le cas de la dimension de valence.
- Enfin, la multi-modalité semble plus adaptée à la prédiction de la valence que de l'éveil.

Dans leur revue des approches dédiées au traitement automatique des émotions selon leurs représentations dimensionnelles (*cf.* section II.2.2), Gunes *et al.* présentent des conclusions similaires en mettant davantage l'accent sur les signaux vocaux. La hauteur de la voix (le *pitch*), l'intensité, l'énergie des hautes fréquences et le timbre renseignent énormément sur l'éveil. Une forte « puissance » (*cf.* dimension *contrôle*/*soumission* section II.2.2) semble corrélée aux faibles fréquences et aux longues voyelles. Des émotions positives semblent liées à un flux de parole élevé, à des voyelles longues et à des bandes de fréquences larges.

D'autres travaux sont consacrés à la génération et à synthèse d'états émotionnels sur des avatars ou sur des agents dits « rationnels » ou « conversationnels ». La plupart de ces agents s'appuient sur le modèle émotionnel *OCC*. [**58**] présenté à la section II.2.3.1. Ci-dessous, nous présentons les différents travaux dédiés à la génération d'émotions ou d'attitudes émotionnelles sur des systèmes conversationnels en lien avec leur représentation cognitive.

Tableau III.1 Configuration du critère de désirabilité et attitudes mentales associées (Sadek *et al*. [**61**]).

Valeurs possibles pour le critère de *désirabilité*	**Configuration d'attitudes mentales primitives**	**Emotion associée**
désirabilité présente de l'évènement *E* pour le choix *C* effectué par *A*	*A* ***désire*** que *C* soit vrai. *A* ***croit*** que : *E* a eu lieu avant quoi il ***estimait*** avec une certaine probabilité que *C* était vrai. *A* ***estime*** avec une certaine probabilité que *C* est vrai. *A* ***croit*** que *C* a plus de probabilité d'être vrai maintenant qu'avant *E*.	joie
indésirabilité présente de l'évènement *E* pour le choix *C* effectué par *A*	*A* ***désire*** que *C* soit vrai. *A* ***croit*** que : *E* a eu lieu avant quoi il ***estimait*** avec une certaine probabilité que *C* était vrai. *A* ***estime*** avec une certaine probabilité que *C* est vrai. *A* ***croit*** que *C* avait plus de probabilité d'être vrai avant *E* que maintenant.	mécontentement
désirabilité future de l'évènement *E* pour le choix *C* effectué par *A*	*A* ***désire*** que *C* soit vrai. *A* ***estime*** avec une certaine probabilité que *C* est vrai. *A* ***estime*** avec une certaine probabilité que : *E* peut avoir lieu, après quoi il ***estime*** avec une certaine probabilité que *C* sera vrai. *A* ***croit*** que *C* aura plus de probabilité d'être vrai après *E* que maintenant.	espoir
indésirabilité future de l'évènement *E* pour le choix *C* effectué par *A*	*A* ***désire*** que *C* soit vrai. *A* ***estime*** avec une certaine probabilité que *C* est vrai. *A* ***estime*** avec une certaine probabilité que : *E* peut avoir lieu, après quoi il ***estime*** avec une certaine probabilité que *C* sera vrai. *A* ***croit*** que *C* a plus de probabilité d'être vrai maintenant qu'après *E*.	peur

La technologie *ARTIMIS.* [**61**] développée par France Télécom utilise comme cadre conceptuel la *Théorie de l'Interaction Rationnelle* (« *Theory of Rational Interaction* » : *TRI.*), qui représente une extension de l'approche dite « *Croyances, Désirs et Intentions* » (« *Beliefs, Desires and Intentions* » : *BDI*) précédemment introduite dans [**158**]. Dans ce cadre, une perception potentiellement inductrice d'émotions est représentée par des « *attitudes mentales* ». La configuration de telles attitudes mentales donne lieu à un *état mental*, auquel est associée une émotion.

Le cadre de la *TRI* définit trois types d'attitudes mentales primitives pour l'agent rationnel au regard d'une situation particulière ou d'un élément particulier – *évènement*, *action*, ou *objet* si l'on se réfère au modèle *OCC* [**58**]) – : les *croyances*, les *incertitudes*, et les *choix* :

- une *croyance* est une proposition considérée comme vraie par l'agent ;
- une *incertitude* est une proposition dont l'agent n'est pas tout à fait certain qu'elle est vraie ; il fait l'*estimation* qu'elle est vraie avec un certain degré de probabilité ;
- un *choix* est une proposition que l'agent souhaiterait voire satisfaite par le monde.

Pour la technologie *ARTIMIS*, Ochs *et al.* utilisent le modèle *OCC* uniquement pour la classe des émotions liées à des *évènements*, et plus particulièrement pour la *joie*, le *mécontentement*, l'*espoir* et la *peur*. Le critère d'évaluation est la *désirabilité*. Le Tableau III.1 présente, sous forme d'énumération de propositions logiquement jointes, les configurations d'attitudes mentales primitives pour un agent *A*, relativement à la désirabilité d'un évènement *E* présent, passé ou futur, au regard d'un choix *C* effectué. Y sont également indiquées les émotions correspondant aux valeurs possibles de désirabilité.

Dans le cas du modèle de Prendinger *et al.* [**63**], un module évalue la signification émotionnelle de l'évènement en termes de *but*, de *principes* et de *préférences*. L'incertitude n'est pas prise en compte.

Avec le modèle *EMA* (*Emotion and Adaptation*) [**62**], les émotions peuvent être générées sur le faciès des agents en fonction d'un certain nombre de variables d'évaluation qui ont trait à :

- la *pertinence* ou l'*utilité* d'un évènement pour l'agent ;
- la *désirabilité* d'un évènement au regard des préférences de l'agent ;
- la *vraisemblance* ou le *degré de certitude* d'un évènement pour l'agent ;
- l'*attribution causale*, qui définit si l'agent qui a causé un évènement mérite de la reconnaissance ou si on doit lui imputer une faute ;
- le potentiel de *prise en main* et de *contrôle* de la situation par l'agent ;
- l'évaluation du *caractère changeable* d'un évènement sans intervention direct de la part d'un agent extérieur.

La *désirabilité* de l'évènement et son *degré de certitude* sont conformes au modèle de simulation d'émotions PETEEI (« *PET with Evolving Emotional Intelligence* » [**159**]). Le modèle *EMA* prend donc également en compte des évaluations relatives au *type d'agent* responsable de l'évènement, au degré de *contrôle* et à la capacité de l'agent à faire face à la situation rencontrée. Il s'appuie sur une représentation *causale* des évènements et sur les états de l'agent qui en résultent.

Dans leurs travaux sur l'expressivité des agents conversationnels [**65**], J. C. Martin *et al.* définissent un modèle d'expressivité du geste avec six caractéristiques : l'extension spatiale, l'extension temporelle, la puissance, la fluidité, le caractère répétitif et l'activité globale. Un premier test leur permet de vérifier que ces dimensions sont aisément perçues par des observateurs humains (auquel cas une variation dans une seule des six dimensions sera détectée et correctement attribuée). Il s'agit de vérifier si les utilisateurs

sont capables ou non de *détecter* les variations d'un paramètre. Les extensions spatiale et temporelle sont les mieux reconnues. Dans un second test, les auteurs se demandent si la combinaison de paramètres en vue de refléter au mieux une communication intentionnelle donnée résultera en une meilleure crédibilité de l'agent conversationnel. Il s'agit alors de vérifier si les utilisateurs sont capables ou non d'*interpréter* les paramètres dimensionnels. Trois comportements – ou qualités – sont définis : *abrupte*, *paresseux* et *vigoureux*, qui sont chacun déterminés par une configuration particulière des six paramètres de geste énumérés ci-dessus. Il est demandé aux observateurs de classifier quatre instanciations de chacune de ces trois qualités, de la moins appropriée à la plus appropriée, au regard du contenu expressif visible. Les qualités *abrupte* et *vigoureux* sont correctement perçues.

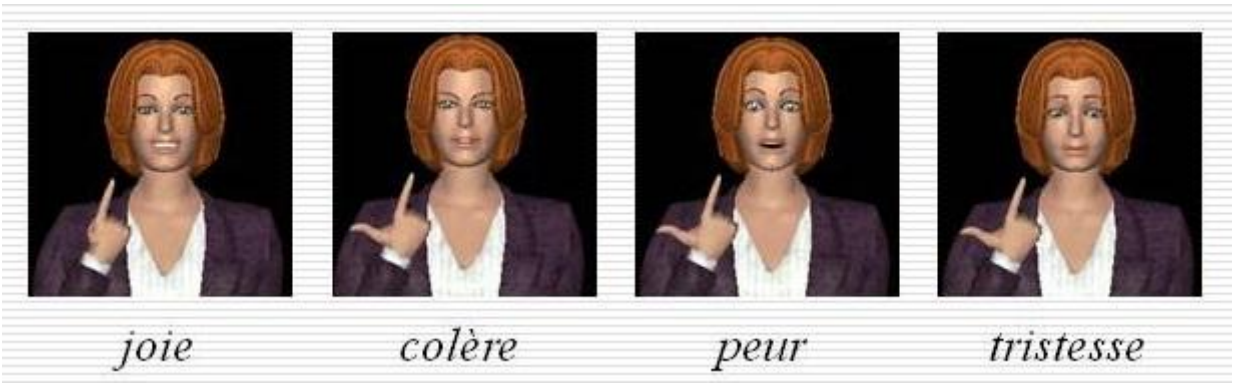


Figure III.8 Exemples d'expressions faciales d'émotions sur l'agent conversationnel Greta (Pasquariello et Pelachaud [**160**]).

Dans [**66**], Ochs *et al.* utilisent le prototype d'agent conversationnel Greta [**160**] (Figure III.8) et suggèrent des règles de génération pour deux types de mélanges émotionnels sur le visage : la *superposition* de deux émotions déclenchées et le *masquage* d'une émotion déclenchée par une émotion exprimée. Ils donnent des exemples d'émotions faciales résultant de ces complexes émotionnels.

- La superposition de deux émotions déclenchées se fait par l'attribution à la partie supérieure du visage de l'émotion « la plus négative », et l'attribution à la partie inférieure du visage de l'émotion « la plus positive », selon des règles fixées par Ekman et Friesen dans [**161**].
- Le masquage d'une émotion ressentie par une émotion exprimée est permis par la connaissance des régions du visage qui sont le siège de l'expression caractéristique de l'émotion ressentie. Les caractéristiques faciales liées à cette dernière sont difficiles à inhiber, et justement absentes lorsque l'émotion est simulée.

De Rosis *et al.* [**64**] utilisent également l'agent Greta avec un réseau dynamique de croyances : les émotions sont générées quand l'agent subit un *changement de croyance* concernant la réalisation de ses buts ou la menace qui pèse sur l'un d'eux.

Dans [**67**], Grizard *et al.* travaillent à partir du modèle évaluatif de Scherer [**60**] pour animer un robot développé par Philips (la plateforme « *iCat* ») et un avatar graphique créé avec l'outil commercial « *Haptek* ». Les expressions faciales exploitent les unités d'actions faciales (« *Facial Action Units* » : *FACS* [**49**]) de Ekman et Friesen. Pour les deux agents, ils évaluent la crédibilité des cinq émotions que sont la *joie*, le *dégoût*, la *tristesse*, la *colère* et la *peur*.

Dans cette section, nous avons présenté différents travaux dédiés à l'analyse et à la synthèse d'émotions. Comme nous l'avons expliqué à la fin de la section II.2, l'analyse d'émotions exprimées par un sujet nécessite de se référer au mouvement corporel qui porte ces émotions. Inversement, le geste, de par son caractère intentionnel et communicationnel exprime l'état émotionnel de celui qui l'effectue. Dans la section suivante, nous nous concentrons sur des modèles descriptifs du geste qui incorporent sa dimension expressive.

III.4. Analyses à base de l'expressivité

Nous nous penchons enfin sur des travaux dédiés à l'analyse de l'expressivité que nous regroupons en trois familles d'approches:

- Les premières consistent à définir des descripteurs mi-niveau du geste dans ses dimensions communicationnelles et expressives : elles ont pour objectif de permettre la qualification du geste selon des descriptions sémantiques de plus haut-niveau effectuées par annotateurs, et sont donc utilisées en entrée des procédures d'apprentissage supervisé.
- Les secondes sont proprement dédiées à l'étude du geste selon les concepts du modèle de mouvement LMA de Laban. Il s'agit de décrire les gestes en termes de qualités ou sous-qualités de Laban, à l'aide de systèmes d'apprentissage fondés sur des descripteurs visuels ou kinesthésiques mi-niveau.
- Les troisièmes méthodes consistent à quantifier les concepts de Laban pour en faire le substrat d'une description sémantique mi-niveau du geste, dédiée à la reconnaissance de contenus haut-niveau selon lesquels le geste a été indexé ou annoté.

III.4.1. Descripteurs expressifs mi-niveau

Une première famille d'approches consiste donc à prendre appui sur l'expressivité gestuelle et sur les caractéristiques communicatives du geste pour bâtir des descripteurs mi-niveau du mouvement corporel. Ce modèle intermédiaire est généralement couplé à l'utilisation de procédures d'apprentissage supervisé pour déterminer des contenus haut-niveau comme des émotions ou des états affectifs.

Citons tout d'abord l'approche de Glowinski *et al.* [**162**], qui proposent une représentation minimale de l'expressivité du geste, basée sur les trajectoires des deux mains et de la tête extraites de portraits vidéo. Un premier module est dédié à l'extraction de caractéristiques « bas-niveau » que sont les positions et vitesses des mains et de la tête. Un second module a pour fonction de calculer vingt-cinq indices « mi-niveau » ayant trait à l'expressivité : l'énergie, l'extension spatiale, la régularité et la souplesse du mouvement, la symétrie, le mouvement global de la tête. Enfin, un troisième module opère une réduction de l'espace à l'aide d'une Analyse en Composantes Principales. Les auteurs disposent d'un corpus 120 vidéos annotées à l'aide de douze catégories émotionnelles et également selon l'espace valence-éveil [**52**] (*cf.* section II.2.2). Après application de leurs descripteurs à ce corpus et réduction de la dimensionnalité, les auteurs parviennent à constituer quatre clusters de gestes qui recoupent les quatre cadrans de l'espace valence-éveil (*e.g.*, *plaisant-éveillé*, *plaisant-calme*, *déplaisant-éveillé*, *déplaisant-calme*). La partition qui en résulte contient des lots de catégories émotionnelles bien spécifiques (par exemple, le groupe « *plaisant-éveillé* » contient de fait les portraits représentatifs de l'*allégresse*, de l'*amusement* et de la *fierté*).

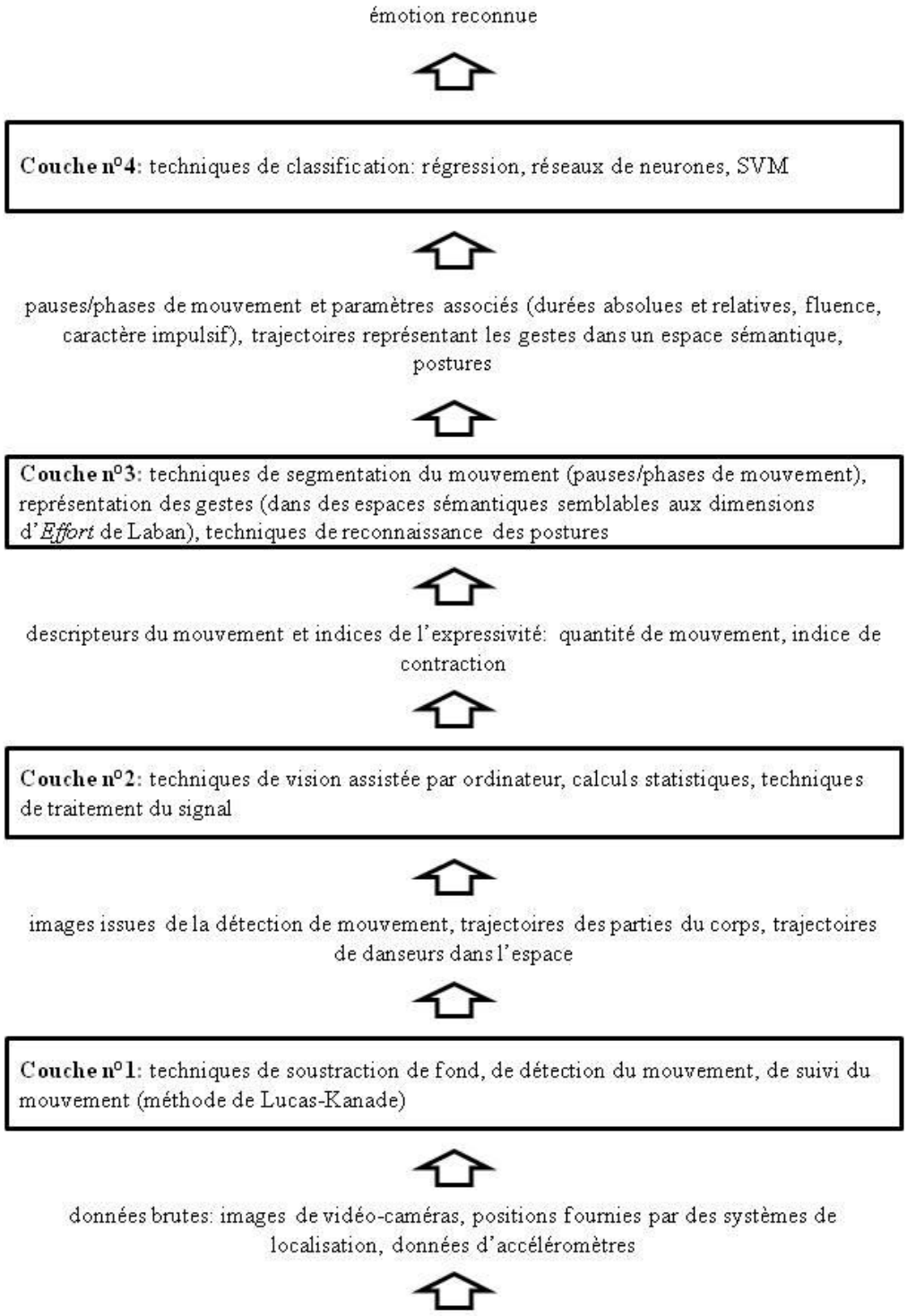


Figure III.9 Reproduction du schéma proposé par Camurri *et al.* dans [**83**], présentant l'approche par couche dédiée à l'extraction de contenus émotionnels dans le geste dansé.

Dans [**163**], Camurri *et al*. privilégient une approche par couche dont les principes sont résumés sur la Figure III.9. A partir de signaux 2D bas-niveau, le principe consiste à décrire les mouvements de musiciens et de danseurs en fonction de leur énergie, de leur fluidité, de la quantité de mouvement, de leur degré de contraction, de manière à extraire des contenus sémantiques et expressifs, notamment en

s'inspirant des notions développées par Rudolf Laban (*cf.* section II.3). Ensuite, ces contenus expressifs sont utilisés pour qualifier d'autres contenus haut-niveau comme des émotions ou des affects à l'aide de méthodes d'apprentissage. Dans ce cas, il s'agit d'utiliser des arbres de décision de façon à attribuer le geste à une catégorie émotionnelle parmi un ensemble prédéfini discret : *joie*, *peur*, *souffrance*, *colère*. Les auteurs comparent les résultats en termes de classification avec ceux qui sont reportés dans [**83**] et qui dénotent l'évaluation par des spectateurs du contenu émotionnel des gestes artistiques.

Intéressons-nous à présent aux approches dédiées à la caractérisation des qualités/sous-qualités du modèle d'analyse de mouvement de Laban.

III.4.2. Analyse des qualités de Laban

Dans l'approche introduite dans [**79**], Swaminathan *et al.* travaillent sur des gestes issus d'improvisations dansées et les caractérisent selon que le corps avance/recule, s'élève/s'abaisse ou s'élargit/se rétrécit. Cela revient à décrire la sous-composante *Mise en form*e de la qualité *Forme*. Ils utilisent un réseau bayésien dynamique qui prend pour entrée des positions 3D de 34 marqueurs repérés sur le corps et calcule des « sous-indicateurs » du geste.

Zhao et Badler [**80**] proposent une analyse des corrélations existant entre les sous-espaces d'*Effort* de Laban. Les annotations LMA de gestes effectuées par des experts sont comparées ici avec les classifications de ces gestes opérées par des classifieurs à base de réseaux de neurones, chacun étant consacré à une sous-qualité d'*Effort* de Laban et selon deux caractéristiques globales : (1) le performeur se livre pleinement dans le sens de la qualité en question (*indulging in the quality*) ; (2) il lutte contre la qualité en question (*fighting against the quality*). Dans cette étude, les descripteurs qui servent d'observation aux réseaux de neurones sont globaux et concernent la courbure et la torsion de la trajectoire du corps, différents angles de référence au niveau du poignet et du coude, ainsi qu'une appréciation de la taille du déploiement corporel.

Dans [**81**], Bouchard et Badler utilisent le modèle LMA pour des problématiques de segmentation des gestes. Les auteurs montrent que les méthodes automatiques de segmentation par apprentissage sont aptes à diviser sémantiquement le mouvement, mais ont tendance à ne fonctionner que sur des classes de mouvements de natures bien spécifiques et restreintes (*e.g.*, gestes dansés, postures, langue des signes). Pour segmenter des mouvements plus génériques, il est nécessaire de disposer d'une caractérisation par descripteurs dits « ouverts ». Les caractéristiques LMA sont considérés par les auteurs comme étant plus pertinentes en vue d'une segmentation des gestes que la cinématique du mouvement.

Pour chaque dimension d'*Effort* de Laban, un réseau de neurones est entrainé pour catégoriser le mouvement entre trois états : *indulging in the quality*, *neutral* et *fighting against the quality*. Le signal en entrée des réseaux de neurones est un descripteur de mouvement qui s'appuie sur des marqueurs de référence pour caractériser la position de certaines articulations, dans le cadre d'une approche cinématique de description du geste.

L'entrainement des réseaux de neurones est effectué à l'aide d'une base de gestes spécialement conçue, qui met en jeu douze mouvements exécutés de douze façons différentes. Par ailleurs, chaque geste est segmenté en fonction de la cinématique du mouvement, en suivant un protocole hybride mêlant segmentations manuelle et automatique. L'expérience de Bouchard et Badler consiste à comparer la segmentation hybride avec une segmentation automatique qui s'appuie sur une caractérisation de la

saillance du mouvement. Cette saillance est associée à des changements dans les poids des couches cachées des réseaux de neurones et est représentée par un histogramme qui globalise l'information tout au long du mouvement. La comparaison entre les deux types de segmentation prouve la pertinence de l'approche proposée.

Dans [**77**], Hachimura *et al*. quantifient les sous-qualités de *Poids*, d'*Espace* et de *Temps* de la qualité d'*Effort*, ainsi que les sous-qualités de *Mise en Forme* et de *Flux de Forme*. Ces quantifications sont locales. Pour chaque sous-qualité, une ou plusieurs caractéristiques décrivent le geste à l'instant *t*. La comparaison entre l'analyse de Laban dérivée de ces indices et l'annotation correspondante fournie par des spécialistes montrent une correspondance seulement dans le cas de certaines qualités : il s'agit des périodes de mouvements effectivement lourds et légers (dimension *Poids*), ainsi que des périodes de mouvements effectivement soudains (dimension *Temps*), qui sont bien reconnues par le système.

Dans [**25**], les deux études présentées traitent de la caractérisation du mouvement dansé proposée par Rudolf Laban. L'analyse de Laban est mise à profit pour développer des descripteurs de geste à des fins de reconnaissance automatique. Les gestes expressifs pour l'apprentissage et la reconnaissance exploitent quatre composantes dynamiques de qualités de mouvement, élaborées avec des danseurs et des chorégraphes, que sont l'inspiration (*Breathing*), le saut (*Jumping*), l'extension (*Expanding*), et la réduction (*Reducing*). En outre, chacun de ces éléments est échelonné selon de multiples nuances.

Dans la première approche, les dynamiques du mouvement sont modélisées par des systèmes de « masses-ressorts » gouvernés par une équation de mouvement dont les paramètres sont utilisés pour entrainer des classificateurs dédiés. La vérité terrain correspond aux quatre composantes dynamiques des qualités de mouvement susmentionnées. La classification est effectuée avec deux méthodes distinctes, l'une par moindres carrés et l'autre par filtrage de particules.

Dans la seconde étude, il s'agit de calculer des descripteurs du mouvement dansé pour entrainer des modèles HMM [**98**] dédiés à la reconnaissance des quatre composantes dynamiques des qualités de mouvement. Un ensemble de descripteurs, à la fois spatiaux et temporels, sont proposés. Parmi les descripteurs spatiaux, citons :

- la *verticalité* : rapport entre la hauteur et la largeur de la silhouette,
- l'*angle des aisselles* : angle séparant l'axe défini par le bras et l'axe vertical défini par le buste ;
- l'*extension* : distance maximale entre le centre de masse et l'ensemble des extrémités du corps (mains, pieds, épaules, tête...),
- l'*ouverture des jambes* : distance entre les deux pieds,
- le *transfert de poids* : distance entre l'abscisse du centre de masse et le centre du segment défini par les deux pieds.

En ce qui concerne les descripteurs temporels, trois éléments sont élaborés :

- la *périodicité*, définie comme la moyenne du coefficient d'auto-corrélation des quatre extrémités délimitant la silhouette,
- le rapport extension/contraction (*increase/decrease*), qui décrit l'évolution temporelle (taux de variation) d'un paramètre spatial donné (*i.e.*, parmi les descripteurs spatiaux précédemment présentés),
- la *quantité de mouvement* : variation entre deux trames successives en termes de pixels de la silhouette captée du participant.

Dans [**82**], D. Chi *et al.* présentent le système *EMOTE* (« *Expressive MOTion Engine* »), dédié à l'animation de personnages 3D, qui met en pratique les composantes d'*Effort* et de *Forme* du modèle

LMA pour caractériser et en même temps éditer l'expressivité de mouvements-types. La Figure III.10 illustre l'interface que propose le système pour sculpter le geste en termes de sous-qualités d'*Effort* en le « repérant » dans le graphe d'Effort de Laban (*cf*. section II.3.1).

Pour des raisons d'équilibre et de cohérence du mouvement d'ensemble, les membres inférieurs ne sont pas affectés. Seuls le torse et les bras sont pris en compte. Les mouvements-types du corps sont donc spécifiés pour ces parties spécifiques, en tant que paramétrage temporel des poses, et peuvent provenir d'une librairie de mouvement ou directement d'une captation gestuelle. Les variations des indices de *Forme* modifient l'attitude des bras et du torse, et donc la pose globale du corps, tandis que seuls les bras sont concernés par les dimensions d'*Effort*.

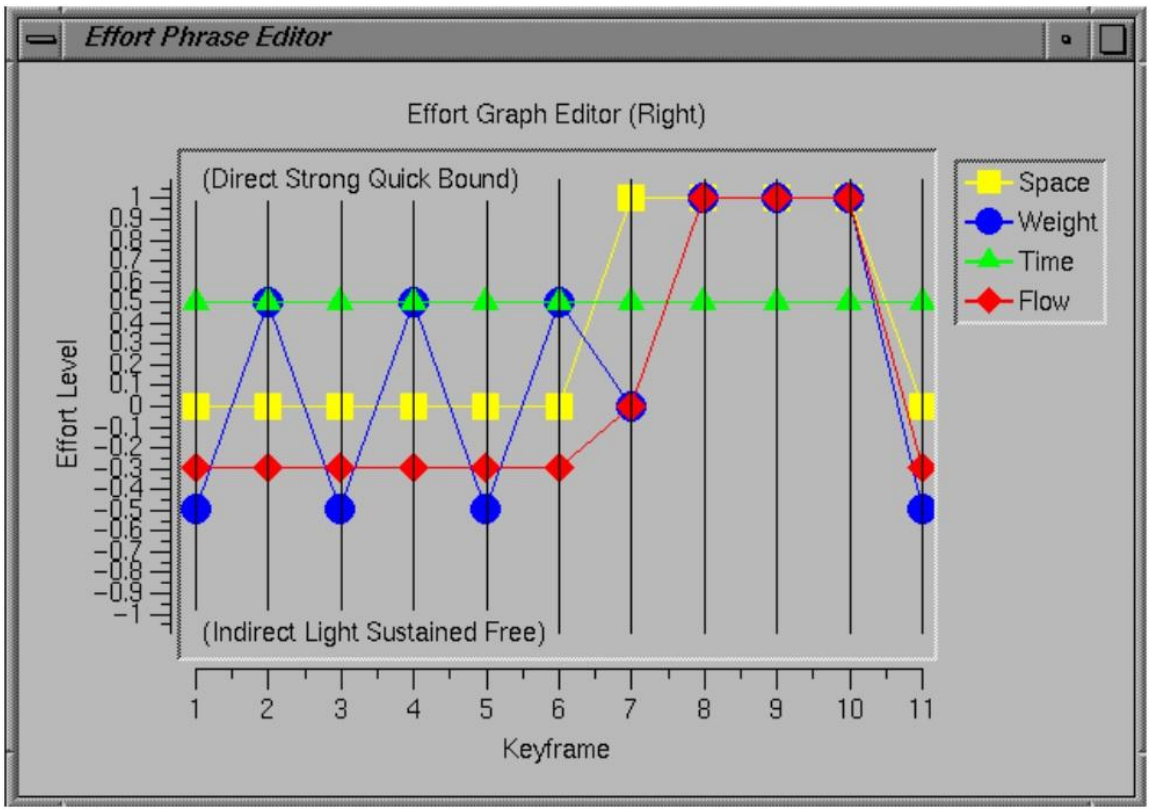


Figure III.10. Editeur du système EMOTE permettant de définir le phrasé d'*Effort* des bras (source : [**82**]).

Les approches présentées dans cette partie consistent essentiellement, à partir de descripteurs visuels ou kinesthésiques mi-niveau, à classifier les gestes selon les sous-qualités d'*Effort* et de *Forme* de Laban, ou à les exprimer à l'aide de ces concepts. Au demeurant, ces sous-qualités d'*Effort* et de *Forme* sont les seuls concepts de la LMA qui se prêtent immédiatement à la quantification, de par leur dimensionnalité (*cf.* continuum *indulging in the quality*/*fighting against the quality* dans le cas des sous-dimensions d'*Effort*, continuum de *Mise en Forme*, etc.). Toujours est-il que les approches décrites montrent dans leur ensemble que ces éléments de la LMA sont parfaitement identifiables et qu'une représentation du geste qui prenne appui sur eux est tout à fait possible.

Il s'agit désormais de trouver une mesure effective pour l'intégralité des qualités de Laban. Intéressons-nous donc aux méthodes qui visent établir une mesure/quantification objective des différents concepts de la LMA, qui nous approchent de la spécification d'un modèle du geste mi-niveau.

III.4.3. Quantification des qualités de Laban

Dans [**74**], Kapadia *et al.* utilisent le mdoèle LMA pour des objectifs d'indexation de contenus gestuels. Les auteurs définissent un ensemble de descripteurs locaux de mouvement, associés à chaque trame de la séquence du geste. Ces différents descripteurs visent à quantifier les concepts de Laban suivants :

- qualité de *Corps*,
- sous-qualités d'*Effort*,
- sous-qualité de *Flux de Forme*.

En outre, ils peuvent être appliqués à de multiples niveaux de détail (articulations individuelles, zones du corps spécifiques, ou corps dans son entièreté).

L'approche proposée permet de mettre en œuvre un système de requêtes dans de larges bases de segments gestuels et rend possible l'utilisation à la fois de valeurs clés et de contraintes temporelles.

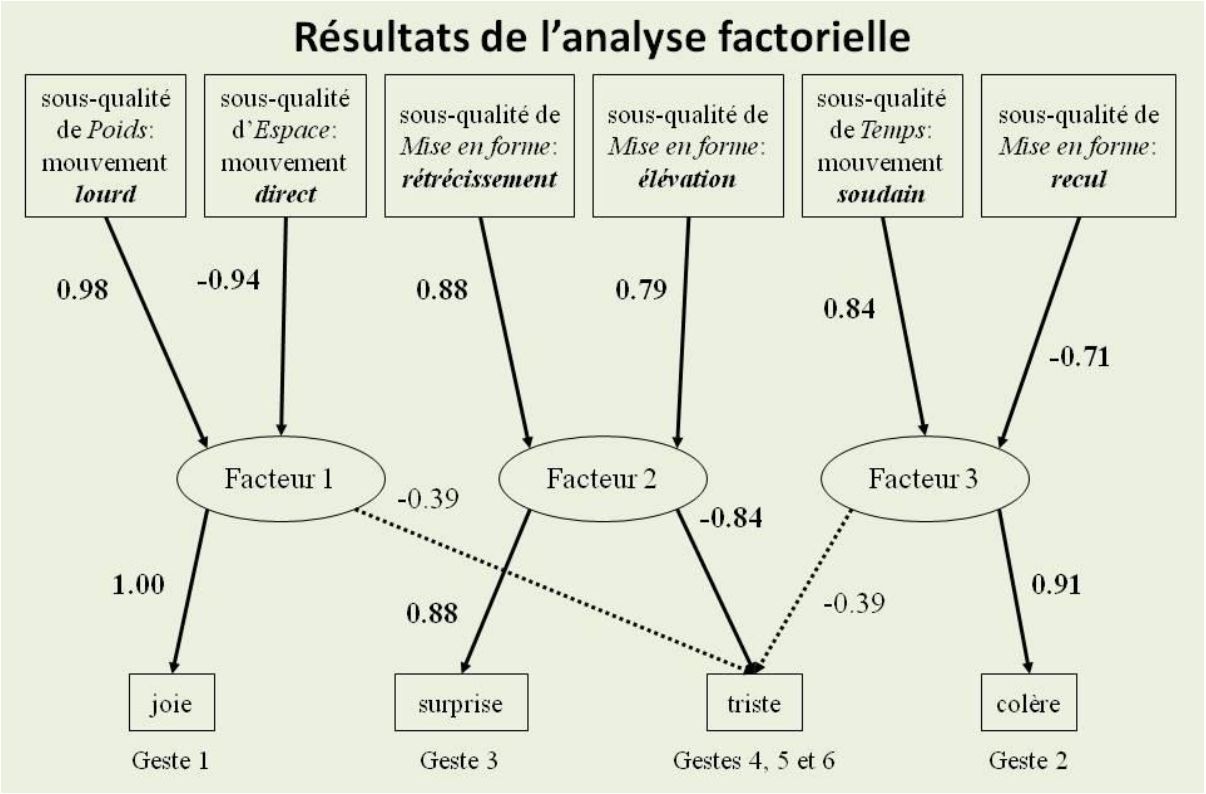


Figure III.11 Diagramme de présentation des résultats de l'analyse factorielle proposée par Nakata *et al.* (source : [**75**]).

Nakata *et al.* [**75**] proposent une série de descripteurs globaux du mouvement, chacun étant dédié à une qualité dérivée du modèle LMA. Plus précisément, les qualités retenues sont les suivantes :

- sous-qualités de *Poids* (lourd/léger), d'*Espace* (direct/indirect) et de *Temps* (soudain/soutenu) de la qualité d'*Effort*,
- sous-qualité *Mise en Forme* de la qualité de *Forme* (élévation/abaissement, avancement/recul, élargissement/rétrécissement).

Les descripteurs gestuels développés sont utilisés pour décrire cinq gestes de robots dansants annotés selon quatre catégories émotionnelles que sont la *joie*, la *surprise*, la *tristesse* et la *colère*. Chaque catégorie émotionnelle est décrite avec niveau d'intensité sur une échelle de 0 à 3. Une analyse factorielle

est utilisée pour établir des relations de causalité entre les qualités de Laban et les émotions. La structure de causalité qui en résulte est présentée sous forme d'un diagramme, illustré Figure III.11. Le sens des flèches dénote la dépendance causale et les nombres désignent les coefficients de saturation : un coefficient supérieur en valeur absolue à 0.7 correspond à une causalité forte.

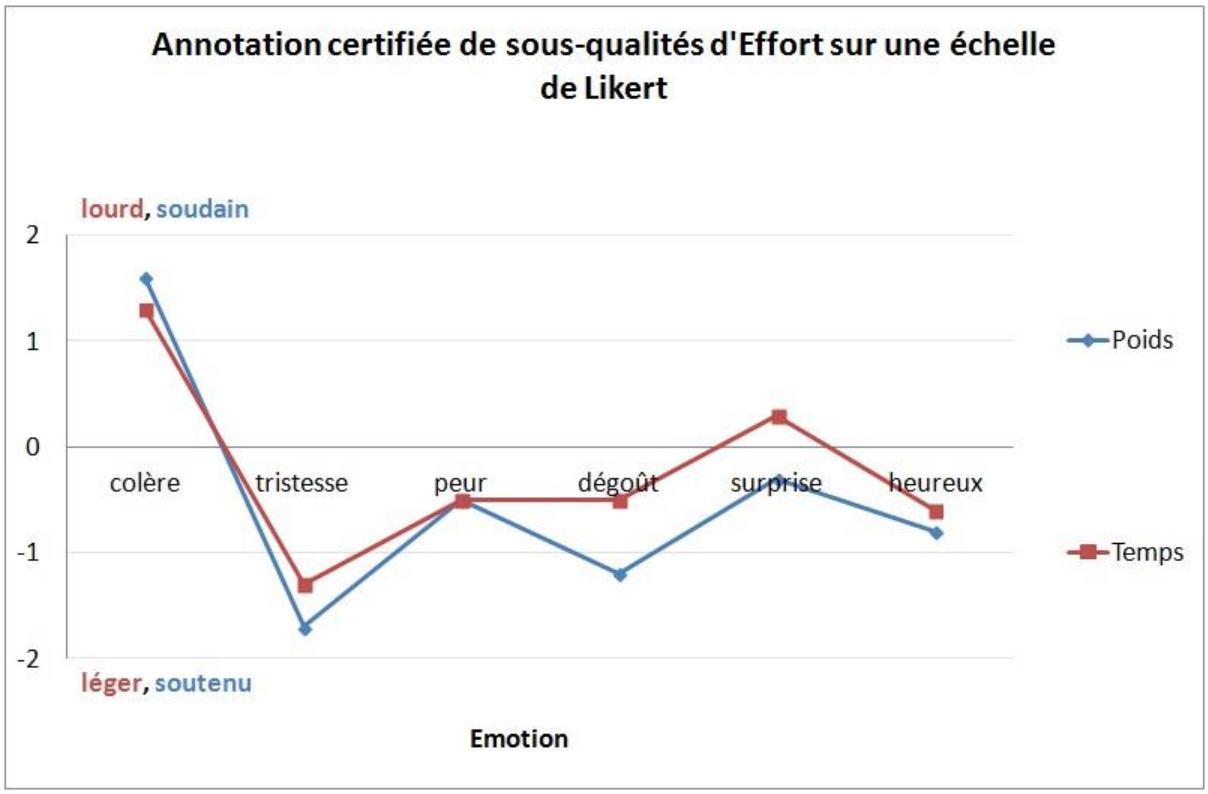


Figure III.12 Quantifications moyennes du *Poids* et du *Temps* sur une échelle de Likert à 5 niveaux dans l'approche de Samadani *et al.* [**76**].

Dans [**76**], Samadani *et al.* s'inspirent des travaux introduits dans [**75**] et [**77**] pour proposer des descripteurs globaux (le détail de ces descripteurs sera évoqué dans la partie IV) de geste dédiés à la quantification des sous-qualités d'*Effort* de Laban, ainsi qu'à la quantification du *Mouvement directionnel*, sous-composante de la qualité de *Forme*. Les auteurs définissent une quantification globale de ces différentes sous-qualités, s'appuyant soit sur une caractérisation intégrale du corps, soit sur des sous-ensembles d'articulations d'intérêt.

Les descripteurs développés sont appliqués à des gestes prédéfinis mettant en jeu les mains et la tête. Ces gestes sont annotés selon les sous-qualités d'*Effort* de Laban. Pour chaque sous-qualité, la caractérisation du geste sur le continuum *indulging/fighting in/against the quality*, est ramenée à un positionnement sur une échelle de Likert de 5 valeurs : {-2, -1, 0, 1, 2}. En termes d'émotions, 6 catégories sont proposées, incluant *colère*, *tristesse*, *peur*, *dégoût*, *surprise* et *bonheur*.

Les auteurs étudient les phénomènes de corrélation entre l'annotation certifiée de l'*Effort* de Laban sur les échelles de Likert et les quantifications des sous-qualités d'*Effort* : les dimensions de *Poids* et de *Temps* montrent un taux de corrélation assez fort entre l'annotation certifiée et les quantifications ayant trait à l'intégralité du corps.

Par ailleurs, les auteurs s'intéressent à la corrélation entre les annotations de l'*Effort* et les qualifications en émotions (Figure III.12). Ainsi pour les annotateurs, la *colère* apparaît comme *lourde*

(dimension de *Poids*) et *soudaine* (dimension de *Temps*), tandis que la *tristesse* est conjointement *légère* (*Poids*) et *soutenue* (*Temps*).

Dans [**78**], Aristidou et Chrysanthou quantifient à chaque instant les qualités de *Corps*, d'*Espace*, ainsi que les sous-qualités d'*Effort* et de *Forme*, et se proposent d'en étudier l'impact dans la classification de mouvements dansés en termes d'émotions. Six catégories d'états affectifs sont proposées : le *bonheur*, la *tristesse*, la *nervosité*, la *curiosité*, l'*activité* et la *peur*.

Au regard de la qualité de *Corps*, le *bonheur* et la *peur* présentent un coefficient de corrélation linéaire élevé, ainsi que la *peur* et l'*activité*. Les sous-qualités d'*Effort* permettent en revanche de séparer aisément les émotions. Par ailleurs, l'utilisation conjointe des composantes de *Forme* et d'*Espace* engendrent des confusions entre les catégories de *curiosité* et de *nervosité*, ainsi qu'entre *bonheur* et *tristesse*. L'utilisation de toutes les qualités de Laban retenues permet en revanche une compréhension quasi parfaite des émotions et intentions des acteurs. Cela prouve une fois de plus que l'état affectif des participants influence les qualités du mouvement.

Soulignons tout d'abord que la plupart des approches des sections III.4.2 et III.4.3 requièrent l'intervention d'experts LMA pour pouvoir annoter un corpus de gestes, et ainsi définir une vérité terrain correspondant aux concepts de Laban que les descripteurs mi-niveau ont pour objectif d'inférer.

Par ailleurs, remarquons que la construction de tels descripteurs mi-niveau dans l'optique d'une caractérisation du geste en termes de qualités ou de sous-qualités de Laban, tout autant que la quantification des concepts LMA ou de notions inspirées de la LMA en vue de la reconnaissance d'indices haut-niveau portés par le geste (*e.g.*, émotions, actions), n'ont jusqu'à présent été testées que sur des mouvements corporels *actés*, et spécifiquement dédiés à l'analyse de l'expressivité ou à une étude poussée de la LMA :

- portraits actés, ne rendant compte que du haut du corps [**162**] ;
- gestes dansés de robots prédéfinis et annotés en émotions [**75**] ;
- prototypes d'utilisation des mains et des bras conçus par des professionnels du mouvement, et utilisés par des acteurs pour exprimer des émotions [**76**] ;
- mouvements humains dansés aux profils d'expressivités précis, présentés comme une extension de la théorie de l'Effort de Laban [**25**] ;
- mouvements humains dansés avec pour but d'analyser la *Mise en forme* [**79**] ;
- mouvements dansés explicitement pour être analysés en termes d'émotions, et dont l'intérêt expressif et émotionnel, en lien avec les représentations classiques de l'affect en termes de catégories (*cf.* section II.2.1), transparaissait nécessairement aux yeux des exécutants au moment de leur performance [**163**] [**78**];
- « plan chorégraphique » bien délimité en termes de direction spatiale (exemple : vers l'avant, vers l'arrière, ascendant, descendant, diagonal, horizontal, sagittal), d'exploration de l'espace (exemple : mi-portée, proche, mi-gauche, mi-droite, lointain), de forme globale (discursif, arqué, circulaire), de forme de la main (ouverte, poing fermé, griffe) [**164**] [**81**].

Les divers travaux étudiés dans cette section et se proposant d'analyser le geste en termes de qualités de Laban présentent diverses limitations auxquelles nous tenterons de répondre dans les chapitres suivants.

Les approches de Zhao et Badler [**80**] et de S. F. Alaoui [**25**] proposent une qualification du mouvement en termes de qualités de Laban ou de concepts hérités, à partir de descripteurs mi-niveau dédiés à l'intégralité de la période gestuelle. La reconnaissance de qualités de Laban y est donc un objectif en soi, qui requiert la définition intuitive de descripteurs mi-niveau capables de saisir des aspects décisifs du geste et nécessaires à son interprétation. Par ailleurs, le calcul de tels descripteurs « globaux» nécessite la réalisation complète du geste à analyser et ne se prête pas à une reconnaissance de contenu qui soit dynamique (*e.g.*, en temps réel).

L'approche de Swaminathan *et al.* [**79**]) répond à cette problématique, en ce qu'elle consiste à quantifier localement (par trame) le mouvement en vue d'une analyse LMA tout au long des gestes selon la sous-dimension de *Mise en Forme*. Mais ici encore, la caractérisation du geste selon certaines dimensions de la LMA nécessite l'élaboration de descripteurs intuitifs du mouvement.

D'autres contributions, plus proches de nos objectifs, fournissent des descripteurs du geste spécifiques à certaines parties du corps ou pour le corps entier, et qui plus est dédiés à la quantification des concepts LMA en vue de la reconnaissance d'émotions (Nakata *et al.* [**75**], Samadani *et al.* [**76**]). Les caractéristiques qui servent à décrire le geste s'appuient donc sur des concepts liés au modèle de l'expressivité corporelle que constitue la LMA, et incorporent ce faisant un certain degré de sémantique. Reste que dans ces approches, les descripteurs sont une fois de plus globaux, c'est-à-dire dédiés à la durée de vie entière du geste.

Enfin, certaines approches effectuent des quantifications locales (*e.g.*, par trame ou par segment temporel de courte durée) de certaines qualités de Laban. Les indices résultants sont comparés aux annotations par trame basées sur le LMA et effectuées par des experts (Hachimura *et al.* [**77**]) ou utilisés pour indexer les gestes par clés de mouvement dans une base de données (Kapadia *et al.* [**74**]). Ces dernières approches fournissent des quantifications objectives de concepts clés de la LMA.

Notre objectif est de conceptualiser un nouveau modèle expressif du geste ayant pour but de quantifier des concepts abstraits de l'analyse de Laban, incorporant les éléments expressifs, intentionnels, et locaux du geste. Ces quantifications seront directement exploitables pour caractériser des contenus haut niveau comme des actions ou des émotions dans un cadre d'apprentissage supervisé, sans chercher explicitement à élaborer la description intermédiaire du mouvement dans les espaces abstraits de la LMA. Nous aurons à cœur de montrer, aux sections VI et VII, que notre modèle descriptif du geste est capable de répondre à des enjeux de classification d'actions et d'analyse de contenus émotionnels dans un contexte plus générique, à partir de gestes non-spécifiquement dédiés au traitement de l'expressivité, et donc non-actés à cet effet.

IV. Descripteurs LMA

Dans ce chapitre, nous présentons notre modèle descriptif de l'expressivité du geste inspiré par la Laban Movement Analysis (LMA – cf. section II.3). Pour chaque qualité ou sous-qualité d'intérêt issue de la LMA, nous analysons tout d'abord les approches de l'état de l'art précédemment dédiées à sa caractérisation par descripteurs numériques. Ensuite, nous proposons notre propre modèle de description. Ces éléments de description du geste seront à la base de nos modèles descriptifs globaux et locaux mis en œuvre dans les sections VI et VII, et s'appuient sur une représentation de la pose corporelle dans un format identique à celui délivré par la Kinect à chaque trame du mouvement (e.g., 20 articulations de référence). Nous illustrons le caractère discriminant de certains de ces descripteurs en les appliquant à des gestes variés, pour constater les différences de profils qu'ils peuvent engendrer.

Pour l'élaboration de nos descripteurs, nous avons donc utilisé les concepts du modèle LMA, qui recouvre parfaitement l'expressivité, le dynamisme et l'aspect communicationnel et intentionnel inhérents à tout mouvement corporel.

Les descripteurs proposés sont associés aux trajectoires 3D des articulations d'un squelette corporel tel qu'il est détecté et enregistré à l'aide d'une caméra Kinect à la fréquence de 30 trames par seconde. Cela correspond à un nombre maximum de 20 articulations, associées aux parties du corps suivantes : *centre des hanches*, *bas de la colonne vertébrale*, *centre des épaules*, *tête*, *épaule gauche*, *coude gauche*, *poignet gauche*, *main gauche*, *épaule droite*, *coude droit*, *poignet droit*, *main droite*, *hanche gauche*, *genou gauche*, *cheville gauche*, *pied gauche*, *hanche droite*, *genou droit*, *cheville droite*, *pied droit* (Figure IV.1).

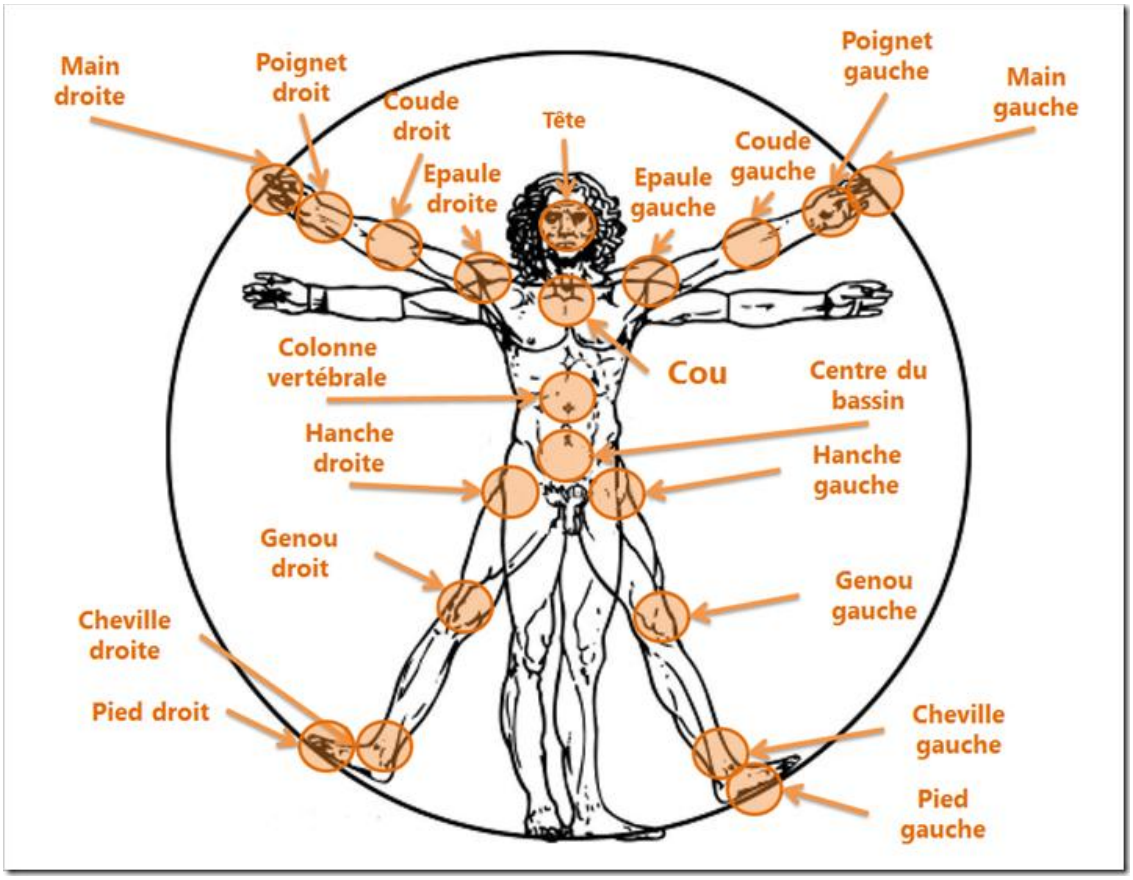


Figure IV.1 Articulations fournies par une caméra Kinect.

Pour chaque articulation *i*, une trajectoire est obtenue, représentée par une séquence de coordonnées 3D, notée $\{P_i = (x_{i,t}, y_{i,t}, z_{i,t})\}_{t=1}^{T}$ dans le système cartésien *(Oxyz)*, où *T* correspond au nombre de trames de la séquence gestuelle. Avant d'extraire les descripteurs, une procédure de normalisation est appliquée, réalisée pour chaque trame *t* à l'aide de plusieurs transformations géométriques globales. L'objectif est d'aligner le squelette selon une nouvelle configuration spatiale $P_{i,t}^{trans} = (x_{i,t}^{trans}, y_{i,t}^{trans}, z_{i,t}^{trans})$ qui rend les plans *(xOy)*, *(yOz)* and *(zOx)* du repère cartésien considéré parallèles aux plans sagittal, vertical et horizontal, respectivement (plans corporels représentés sur la Figure IV.2.b et introduits à la section II.3.1 sur la Figure II.9).

Plus précisément, les transformations suivantes sont appliquées :

1. une translation du corps qui positionne le centre des hanches au centre du repère ;

2. une rotation autour de l'axe *(Oy)* de telle sorte que les épaules gauche et droite se retrouvent dans un plan parallèle au plan *(yOz)* ;
3. une rotation autour de l'axe *(Oz)* qui vise à positionner le segment reliant le centre des épaules et le centre des hanches dans un plan parallèle au plan *(yOz)* ;
4. une rotation autour de l'axe *(Ox)* alignant les épaules gauche et droite dans un plan parallèle au plan *(zOx)* ;
5. enfin, une translation finale pour restaurer le centre des hanches à sa position initiale.

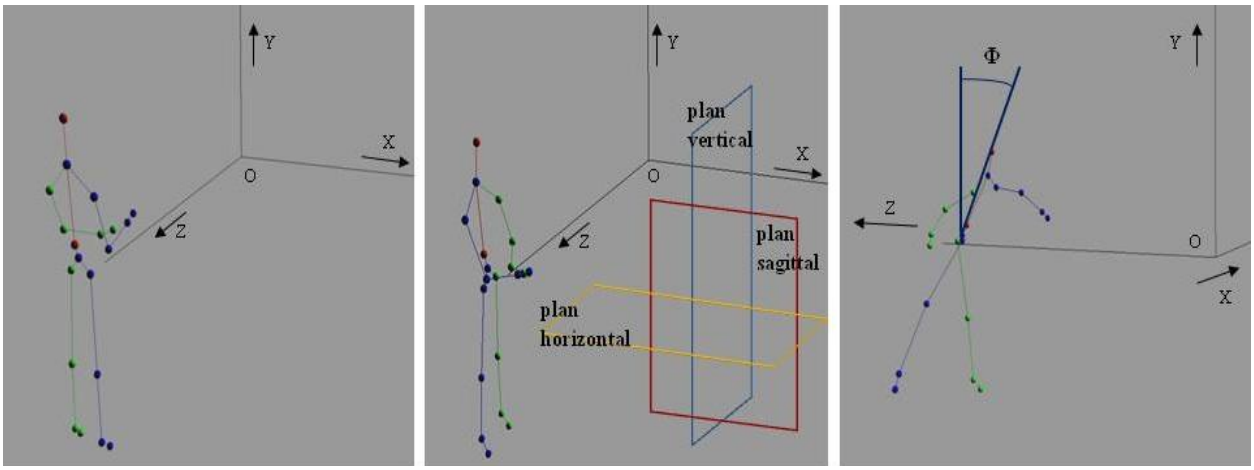


Figure IV.2 Squelette à un instant donné (a) ; illustration des plans corporels et de leurs correspondances avec les plans du repère cartésien après les transformations géométriques de normalisation (b) ; définition de l'angle de penchement (c).

Le résultat de cette séquences de transformations à un instant *t* est illustré Figure IV.2.b. Explicitons à présent les caractéristiques proposées pour notre modèle du geste expressif. L'objectif est de dériver une description pertinente et discriminante pour chaque composante (ou sous-composante) du modèle LMA retenu. Le Tableau IV.1 rappelle les différents éléments de Laban sous l'étude.

Tableau IV.1 Rappel des concepts de Laban d'intérêt pour notre étude du geste.

Qualité de Laban	**Sous-qualité de Laban**
Corps	
Espace	
Forme	*Flux de forme*
	Mouvement directionnel
	Mise en forme
Effort	*Espace*
	Temps
	Flux
	Poids

Notons que certaines de ces caractéristiques sont locales, visant à décrire le geste à un instant donné, tandis que d'autres en décrivent l'expressivité en relation avec sa réalisation globale. Ces derniers indices ne seront donc utilisés qu'en vue d'une reconnaissance globale – ou *a posteriori* – du geste (*cf.* chapitre VI).

Par ailleurs, les figures proposées dans ce chapitre pour illustrer les gestes obtenus à l'aide d'une caméra Kinect se composent d'images acquises face aux exécutants, et qui ont été enregistrées en miroir de la réalité. Contrairement à celles de autres chapitres, ces images de trames gestuelles supposent donc qu'on les interprète une fois leur retournement horizontal (*e.g.*, interversion des côtés gauche et droit) achevé.

IV.1. Qualité de *Corps*

La qualité de ***Corps*** fait référence aux parties du corps qui sont utilisées, à l'initialisation et au séquençage du mouvement.

Dans [**74**], Kapadia et *al.* spécifient un ensemble d'indicateurs de cette dimension corporelle pour un instant donné, caractérisant :

- le déplacement ou orientation de la main par rapport à l'épaule du même membre ;
- la distance de chaque main au plus proche segment corporel ;
- l'équilibre du corps, exprimé par une valeur booléenne indiquant la position du centre de masse par rapport au polygone support du squelette corporel ;
- l'identification du support du geste, c'est-à-dire du segment corporel utilisé pour supporter le poids du corps et en contact avec le sol.

Aristidou et Chrysanthou [**78**] représentent la composante de *Corps* à chaque instant par la distance entre la main et l'épaule pour chaque bras, ainsi que la distance entre la main gauche et la main droite. Ils se rapportent également à la hauteur des hanches par rapport au sol.

Dans le cadre de nos rencontres avec des chefs d'orchestre en vue de la constitution de notre base de données de gestes de direction d'ensembles (*cf.* section V.2.1), l'ouverture/fermeture des bras par rapport à l'axe longitudinal nous a été décrite comme un élément particulier de la direction orchestrale, constitutif de l'engagement dans la musique.

Par ailleurs, dans son étude du mouvement dansé, Sarah Fdili Alaoui [**25**] inclut parmi ses indices de l'expressivité la mesure de l'angle des aisselles, c'est-à-dire l'angle formé par l'axe défini par le bras et l'axe vertical défini par le buste.

Dans nos travaux, nous avons choisi de quantifier la qualité de *Corps* pour chaque trame à l'aide de 3 valeurs. Comme dans [**78**], nous retenons deux distances, l'une entre la main et l'épaule gauches, et l'autre entre la main et l'épaule droite :

$$D^g(t) = \left\| P_{épaule\ gauche,t} - P_{main\ gauche,t} \right\|, \tag{IV.1}$$

$$D^d(t) = \left\| P_{épaule\ droite,t} - P_{main\ droite,t} \right\|, \tag{IV.2}$$

où $\|\ \ \|$ désigne la distance Euclidienne dans l'espace 3D.

Ces deux distances décrivent l'utilisation des membres supérieurs et doivent permettre de discriminer des gestes aussi différents qu'un lancer d'objet (Figure IV.3) et un mouvement périodique de chef d'orchestre battant la mesure (Figure IV.4). Pour ces deux gestes, les profils de ces distances sont représentés Figure IV.5 et Figure IV.6. Les séries temporelles obtenues soulignent bien leur caractère discriminant : le mouvement de battue (courbes bleues) se traduit par des oscillations, et le geste de lancer

(courbes rouges) est parfaitement identifiable sur la Figure IV.6 : la distance entre la main et l'épaule droites se stabilise après le lancer de l'objet, à mesure que le bras revient le long du corps.

Figure IV.3 Participant réalisant le geste *lancer un objet devant soi*, pour la constitution du corpus HTI 2014-2015 (*cf.* section V.2.2).

Figure IV.4 Chef d'orchestre battant la mesure par de petits mouvements dynamiques, symétriques et périodiques, de la base de direction orchestrale (*cf.* section V.2.1).

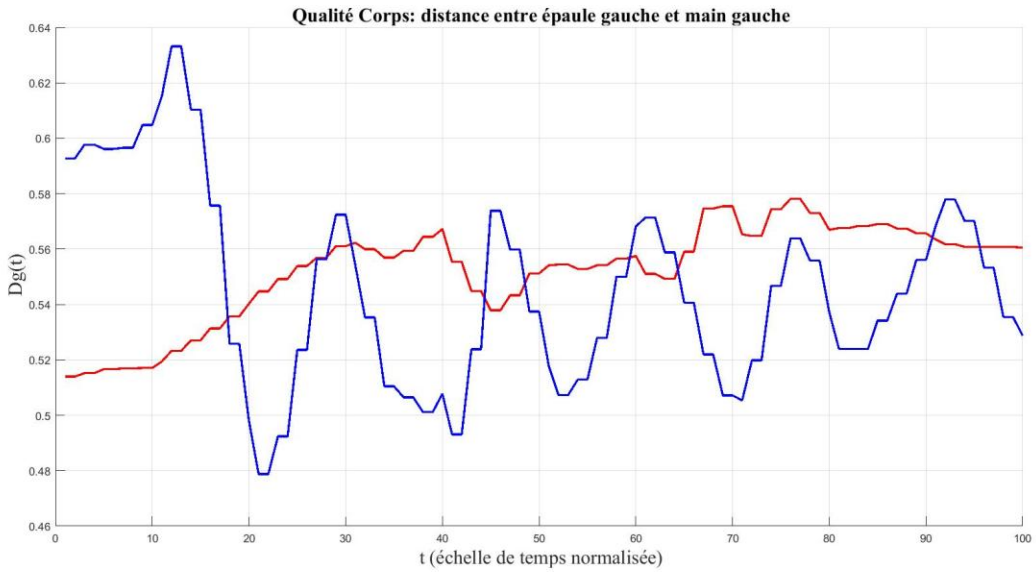


Figure IV.5 Séries des distances $D^g(t)$ entre l'épaule gauche et la main gauche pour le geste *lancer un objet devant soi* (courbe rouge – *cf.* Figure IV.3) et le geste dynamique de battue de la mesure (courbe bleue – *cf.* Figure IV.4).

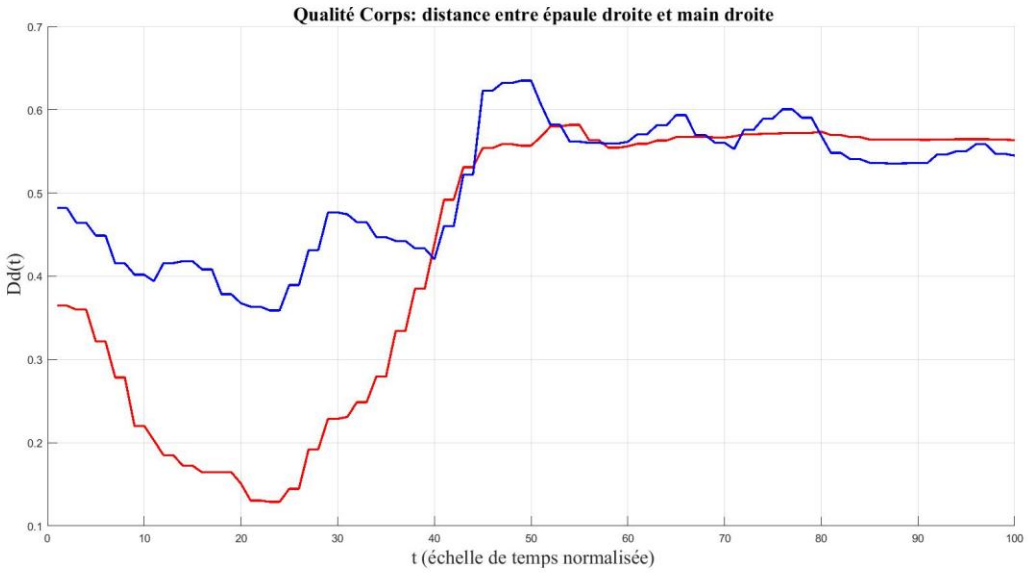


Figure IV.6 Séries des distances $D^d(t)$ entre l'épaule droite et la main droite pour le geste *lancer un objet devant soi* (courbe rouge – *cf.* Figure IV.3) et le geste dynamique de battue de la mesure (courbe bleue – *cf.* Figure IV.4).

Aussi, nous utilisons un troisième indice, inspiré des travaux de Glowinski *et al.* [**162**] qui vise à caractériser la dissymétrie spatiale du corps. Cet indice de dissymétrie est exprimé par l'équation suivante :

$$Dis(t) = \frac{d_{gauche,\,centre}(t)}{d_{gauche,\,centre}(t)+d_{droite,\,centre}(t)}, \qquad \text{(IV.3)}$$

où $d_{gauche/droite,\,centre}(t)$ désigne la distance entre la main gauche/droite et sa projection sur le tronc (*e.g.*, axe centre des hanches - centre des épaules). $Dis(t)$ prend ses valeurs dans l'intervalle [0, 1]. Sa valeur est égale à 0.5 dans le cas d'une parfaite symétrie corporelle.

Cette symétrie, considérée comme de première importance par les chefs d'orchestre pour leur gestuelle de direction, permet par exemple de différencier de simples gestes de battue, avec toute la variabilité qu'ils peuvent recouvrir (Figure IV.4), des gestes plus intentionnels et marqués (comme celui illustré Figure IV.7). La capacité de discrimination de la série *Dis*(*t*) est illustrée Figure IV.8. Ici, nous pouvons distinguer nettement le profil de battue (courbe rouge), légèrement oscillant, de la forme de courbe correspondant à la demande de tenue du son exprimée par le chef d'orchestre (courbe bleue), pour laquelle l'immobilisation relative d'un bras se traduit par une symétrie tendant à se stabiliser. Au-delà de l'univers orchestral, cette mesure nous paraît tout à fait pertinente pour la discrimination entre d'autres types d'actions.

Figure IV.7 Chef d'orchestre stabilisant le bras gauche pour demander une tenue où se conjugent force et tension (base de direction orchestrale – *cf.* section V.2.1).

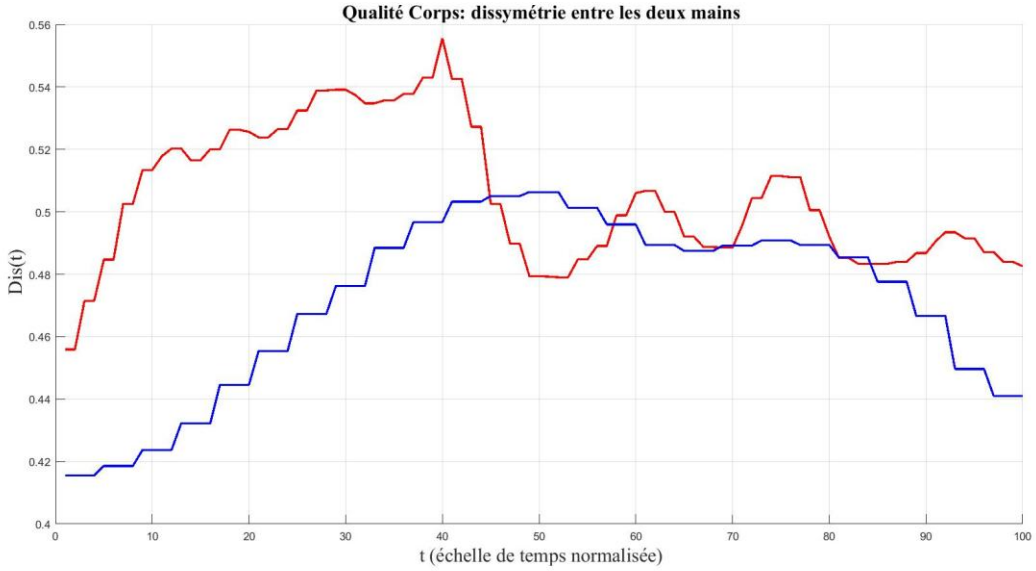


Figure IV.8 Série des dissymétries spatiales $Dis(t)$ pour un profil de simple battue rythmique (courbe rouge – *cf.* Figure IV.4) et une demande de tenue du son exprimée par le chef d'orchestre (courbe bleue – *cf.* Figure IV.7).

La deuxième qualité LMA que nous nous attachons à décrire est celle de l'*Espace*.

IV.2. Qualité d'*Espace*

La qualité d'***Espace*** se réfère à l'exploration des environs et au chemin du mouvement. Dans le cas de la danse, et même plus généralement de gestes artistiques, ce caractère exploratoire est particulièrement insistant. La Figure IV.9 présente un exemple de mouvement durant lequel un chef d'orchestre va chercher le son vers un lieu particulier de l'orchestre, avant de revenir à sa position initiale.

Dans [**65**], J. C. Martin *et al.* travaillent sur la génération de contenus expressifs sur des agents conversationnels et utilisent la notion d'extension spatiale, qui est présentée comme une des six caractéristiques de base d'un modèle de l'affect.

La composante d'*Espace* est également considérée dans [**78**], où les auteurs proposent de la caractériser par deux indices : l'aire totale au sol parcourue par le corps rapportée à la durée entière du geste, et la distance parcourue sur une période, dont la durée peut varier de 5 à 30 secondes.

Pour décrire la qualité d'espace, nous avons tout d'abord retenu une caractéristique globale du geste, notée l_{CH} et définie comme la longueur du chemin parcouru par le centre des hanches. Soulignons également la possibilité d'utiliser dans ce cadre le volume englobant total parcouru comme caractérisation de ce chemin parcouru, au même titre que la projection de ce volume sur le sol (c'est-à-dire sur le plan *(zOx)* – Figure IV.2.b) qui correspond à l'aire totale parcourue.

Nous quantifions également le caractère exploratoire du geste par un indice local, relatant le mouvement vers l'avant/l'arrière, mesurable à chaque instant du mouvement. Cet indice est la composante selon l'axe *x* de la tête après les transformations élémentaires subies par le squelette (Figure IV.2.b), c'est-à-dire dans la direction parallèle au plan vertical : $x_{tête,t}^{trans}$.

Ces deux premières caractérisations du mouvement différencieront les gestes pour lesquels le corps se déplace globalement (comme le geste orchestral illustré Figure IV.9) des gestes mettant surtout en jeu la

corrélation entre les mouvements de différentes parties du corps (comme le geste *faire ses lacets* présenté Figure IV.10).

Figure IV.9 Mouvement de direction musicale brusque vers l'avant, ample, suivi d'un battue légère avec mouvements de la tête saccadés, puis retour progressif vers l'arrière (base de direction orchestrale – *cf.* section V.2.1).

Figure IV.10 Réalisation du geste *faire ses lacets* pour la base HTI 2014-2015 (*cf.* section V.2.2).

Pour chaque trame, nous utilisons enfin l'« angle de penchement vers l'avant » $\Phi(t)$, qui dénote une forme d'insistance, et que nous définissons comme l'angle entre la direction vertical *(Oy)* et l'axe reliant le centre des hanches et la tête, exprimé en radians (Figure IV.2.c). Cet indice supplémentaire nous intéresse particulièrement, puisque le caractère plus ou moins courbé du mouvement peut nous permettre de différencier des gestes structurellement aussi proches que *faire ses lacets* (Figure IV.10) et *se mettre à genoux* ; leur seule différence concerne le buste de l'individu, qui est censé rester vertical une fois l'agenouillement conclu dans le cas du geste *se mettre à genoux* (Figure IV.11). Cette capacité de discrimination est illustré Figure IV.12. Ici, une telle distinction apparaît clairement : la courbe dédiée au geste *faire ses lacets* (en rouge) atteint un palier qui correspond au penchement de l'individu vers son pied, là où le geste *se mettre à genoux* (courbe bleue) conserve une valeur d'angle de penchement proche de zéro du fait que l'individu maintient son buste droit tout au long du geste.

Figure IV.11 Réalisation du geste *se mettre à genoux* pour la base HTI 2014-2015 (*cf.* section V.2.2).

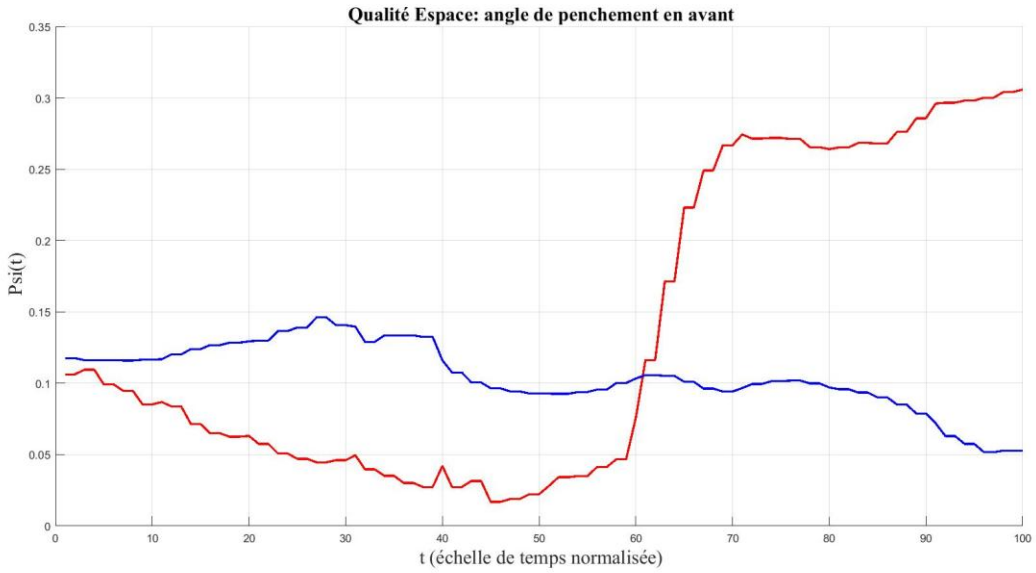


Figure IV.12 Série des angles de penchement vers l'avant $\Phi(t)$ pour le geste *faire ses lacets* (courbe rouge – *cf.* Figure IV.10) et pour le geste *se mettre à genoux* (courbe bleue – *cf.* Figure IV.11).

Intéressons-nous à présente à la caractérisation de la qualité de *Forme* et de ses sous-composantes.

IV.3. Qualité de *Forme*

La troisième qualité de mouvement issue de la LMA à laquelle nous nous intéressons est dédiée à la ***Forme*** que prend le corps au cours d'un geste. Elle est décrite selon trois types d'information : le *Flux de forme*, le *Mouvement directionnel* et la *Mise en forme*.

IV.3.1. Sous-qualité de *Flux de forme*

Nous nous sommes intéressés en premier lieu à la sous-qualité ***Flux de forme***, qui traite des relations changeantes entre les différentes parties du corps.

Pour Hachimura *et al.* [**77**] et Aristidou et Chrysanthou [**78**], le *Flux de forme* peut-être caractérisé à chaque instant par le volume d'un parallélépipède rectangle englobant le corps, par le volume d'une coque

convexe basée sur des articulations de référence du corps (*e.g.*, tête, extrémités des bras, épaules, hanches, pieds).

Kapadia *et al.* [**74**] étendent la possibilité d'une telle description du *Flux de forme* à des sous-parties du corps (par exemple, aux membres).

Pour décrire l'expressivité du geste dansé, Alaoui [**25**] estime qu'il est nécessaire de tenir compte de la distance maximale entre le centre de masse et chacune des extrémités du corps (mains, pieds, épaules, tête...).

Pour étudier le geste musical, les chefs d'orchestre que nous avons interviewés nous ont conseillé de particulièrement prêter attention à l'agrandissement ou à la réduction de la forme du corps (notamment du torse).

Dans leur représentation minimale de l'expressivité corporelle, Glowinski *et al.* [**162**] comptent parmi leurs descripteurs l'aire d'un triangle dont les trois sommets sont la tête, la main gauche et la main droite. Un tel indice se rapproche également de notre notion de *Flux de Forme*.

Dans notre cas, nous proposons de quantifier le *Flux de forme* par un indice relatant la contraction du corps, défini comme décrit par l'équation suivante :

$$C(t) = \frac{(\|P_{centre\ des\ hanches,t} - P_{main\ gauche,t}\| + \|P_{centre\ des\ hanches,t} - P_{main\ droite,t}\|)}{2} \qquad \text{(IV.4)}$$

Ce descripteur caractérise, pour chaque trame de la séquence de geste, l'extension des membres par rapport au centre du corps.

Cet indice permet de séparer des gestes impliquant un certain repliement du corps (*faire ses lacets* – Figure IV.10, *se mettre à genoux* – Figure IV.11) d'autres mouvements naturellement plus amples (Figure IV.13). La Figure IV.14 montre le profil en cloches que peut dessiner un mouvement ample et répétitif de chef d'orchestre (courbe bleue), par contraste avec des mouvements nécessitant plus de repli sur soi-même, comme le fait de lacer ses chaussures (courbe rouge).

Figure IV.13 Mouvement large, ample, calme, incarnant la profondeur de la musique (base de direction orchestrale – *cf.* section V.2.1).

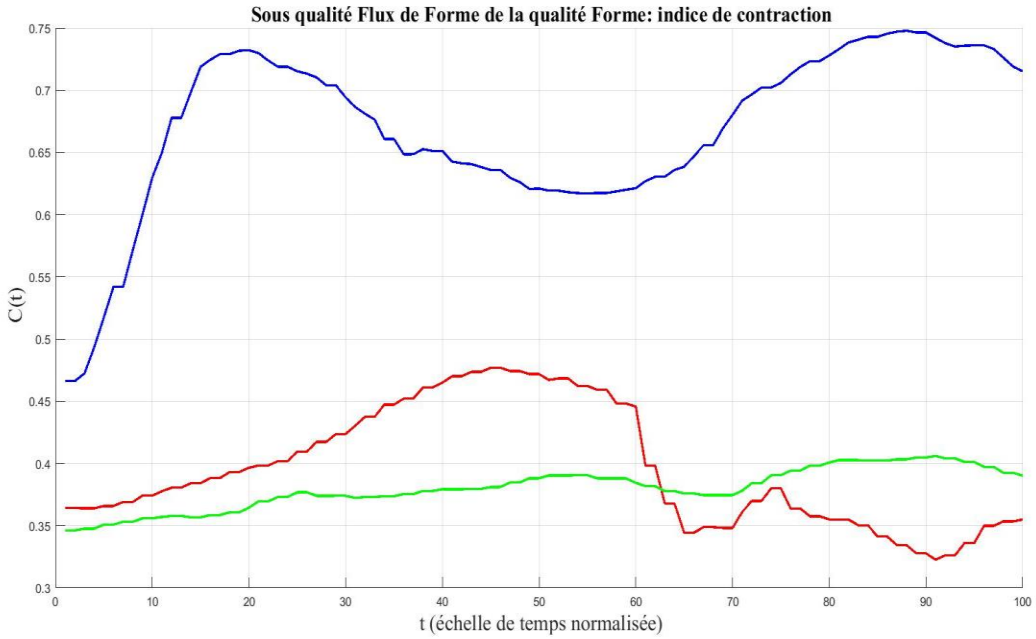


Figure IV.14 Série des indices de contraction $C(t)$ pour le geste *faire ses lacets* (courbe rouge – *cf.* Figure IV.10), pour le geste *se mettre à genoux* (courbe verte – *cf.* Figure IV.11) et pour un geste de chef d'orchestre large et ample (courbe bleue – *cf.* Figure IV.13).

La deuxième sous-qualité de la dimension *Forme* est le *Mouvement directionnel*, dont la caractérisation est présentée dans la section suivante.

IV.3.2. Sous-qualité de *Mouvement directionnel*

La sous-qualité de ***Mouvement directionnel*** décrit la direction éventuelle du mouvement vers un point particulier de l'espace et traduit le but.

Pour Samadani *et al.* [**76**], elle est le mieux décrite par la courbure moyenne du mouvement dans le plan 2D au sein duquel le déplacement le plus large apparaît. Il s'agit donc d'appliquer une méthode d'analyse en composantes principales (A.C.P.) aux trajectoires 3D des articulations du haut du corps, et d'en extraire les deux composantes principales qui forment l'espace 2D recherché.

Dans [**80**], Zhao et Badler rappellent que la courbure est proéminente quand : 1) le mouvement débute (ou après une période de repos), 2) le mouvement prend fin, ou 3) un changement de direction a lieu, et décrit donc parfaitement l'intention subite de relier l'action à un point singulier de l'environnement.

Glowinski *et al.* [**162**] considèrent également la courbure dans les deux directions principales du mouvement comme un indice pertinent de l'expressivité. En effet, ils en retiennent sa valeur moyenne, sur la globalité de la séquence de geste.

Les gestes qui nous ont servi de base de travail ne nécessitant généralement que peu de déplacement global dans l'espace (gestes de chefs d'orchestre devant un pupitre, interactions avec une interface homme machine) –, nous n'avons pas retenu cette sous-qualité de *Forme* dans notre modèle de geste expressif.

IV.3.3. Sous-qualité de *Mise en forme*

La sous-qualité ***Mise en forme*** caractérise les changements de la forme du corps à l'épreuve de son interaction avec l'extérieur.

Dans [**75**], Nakata *et al.* quantifient la sous-qualité de *Mise en Forme* sur le plan horizontal (*cf.* élargissement/rétrécissement), à partir de la projection sur celui-ci des deux mains et d'un point situé à quelques centimètre au-delà de la ligne du regard. Pour le plan vertical (*cf.* élévation/abaissement), ils utilisent la somme des sinus des angles d'élévation des deux bras et de la tête, pondérés par les poids des parties respectives. Enfin, la quantification sur le plan sagittal (*cf.* avancement/recul) prend pour éléments de base la somme de la vitesse d'avancement/retrait du corps ainsi que les vitesses longitudinales aux extrémités des deux bras, pondérées par leurs masses respectives.

Pour Hachimura *et al.* [**77**], la *Mise en Forme* est quantifiée à chaque instant sur les trois plans. L'aire de la projection sur le sol de l'enveloppe convexe du corps quantifie la *Mise en Forme* sur le plan horizontal. Le déplacement vertical du centre des hanches quantifie la *Mise en Forme* sur le plan de vertical (si le déplacement est positif, il y a élévation, sinon il y a abaissement). Enfin, la mesure de l'avancement du centre des hanches vers l'avant quantifie la *Mise en Forme* sur le plan sagittal (avancement/recul).

La sous-composante de *Mise en Forme* est celle qui caractérise le plus directement l'adaptation d'un sujet au milieu que lui imposent ses contraintes physiques. Nous la quantifions pour notre part à chaque trame à l'aide de trois valeurs, qui correspondent aux amplitudes corporelles selon les directions perpendiculaires aux plans vertical, horizontal et sagittal (Figure IV.2.b). Elles sont notées respectivement par $A^x(t)$, $A^y(t)$, et $A^z(t)$ et définies comme décrit dans les équations suivantes :

$$A^x(t) = (max_i(\{x_{i,t}^{trans}\}) - min_i(\{x_{i,t}^{trans}\})), \quad \text{(IV.5)}$$

$$A^y(t) = (max_i(\{y_{i,t}^{trans}\}) - min_i(\{y_{i,t}^{trans}\})), \quad \text{(IV.6)}$$

$$A^z(t) = (max_i(\{z_{i,t}^{trans}\}) - min_i(\{z_{i,t}^{trans}\})), \quad \text{(IV.7)}$$

où *i* indexe les articulations du squelette.

Figure IV.15 Chef d'orchestre au mouvement ample et lourd, dénotant force et conviction (base de direction orchestrale – *cf.* section V.2.1).

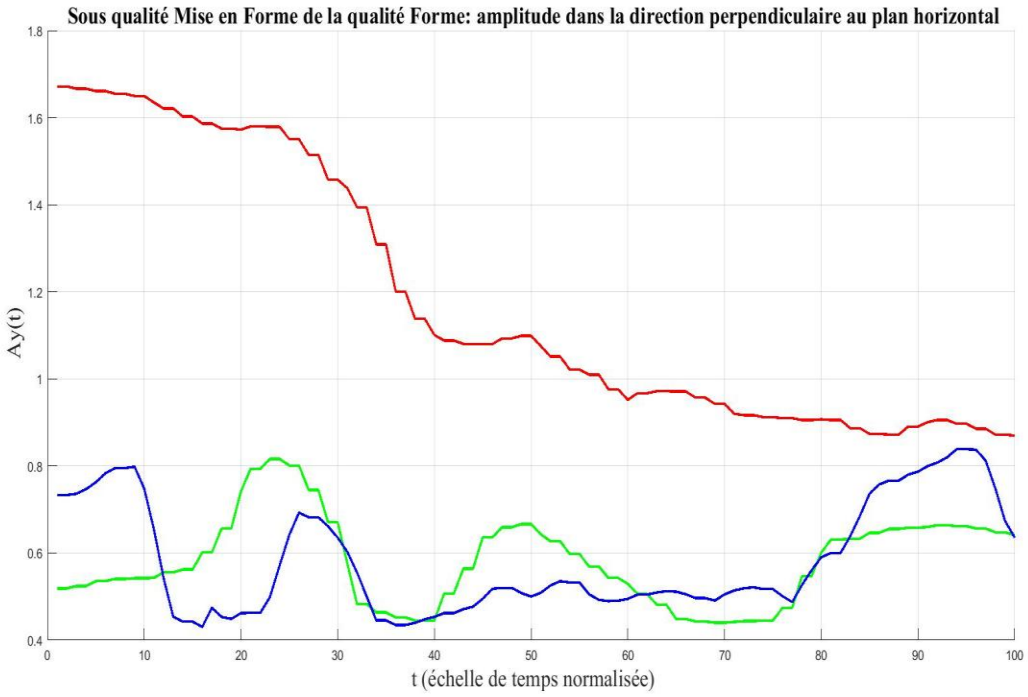


Figure IV.16 Série des amplitudes $A^y(t)$ dans la direction perpendiculaire au plan horizontal pour le geste *faire ses lacets* (courbe rouge – *cf.* Figure IV.10), pour un geste ample de chef d'orchestre (courbe verte – *cf.* Figure IV.15), et pour un geste dénotant successivement avancement et retrait (courbe bleue – *cf.* Figure IV.9).

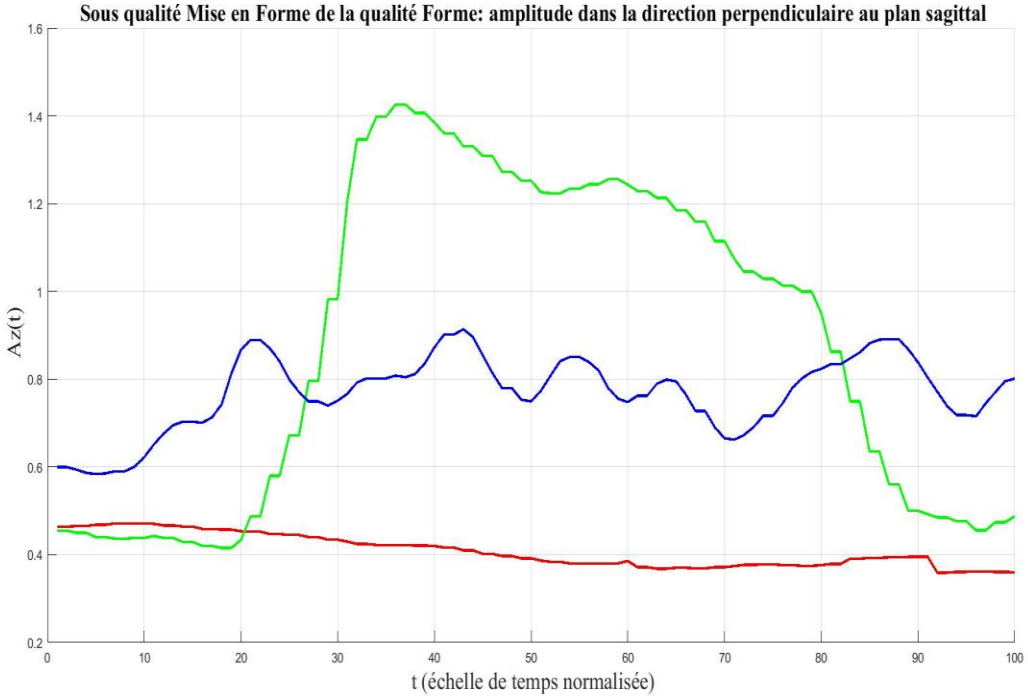


Figure IV.17 Série des amplitudes $A^z(t)$ dans la direction perpendiculaire au plan sagittal pour le geste *faire ses lacets* (courbe rouge – *cf.* Figure IV.10), pour un geste ample de chef d'orchestre (courbe verte –

cf. Figure IV.15), et pour un geste dénotant successivement avancement et retrait (courbe bleue – *cf.* Figure IV.9).

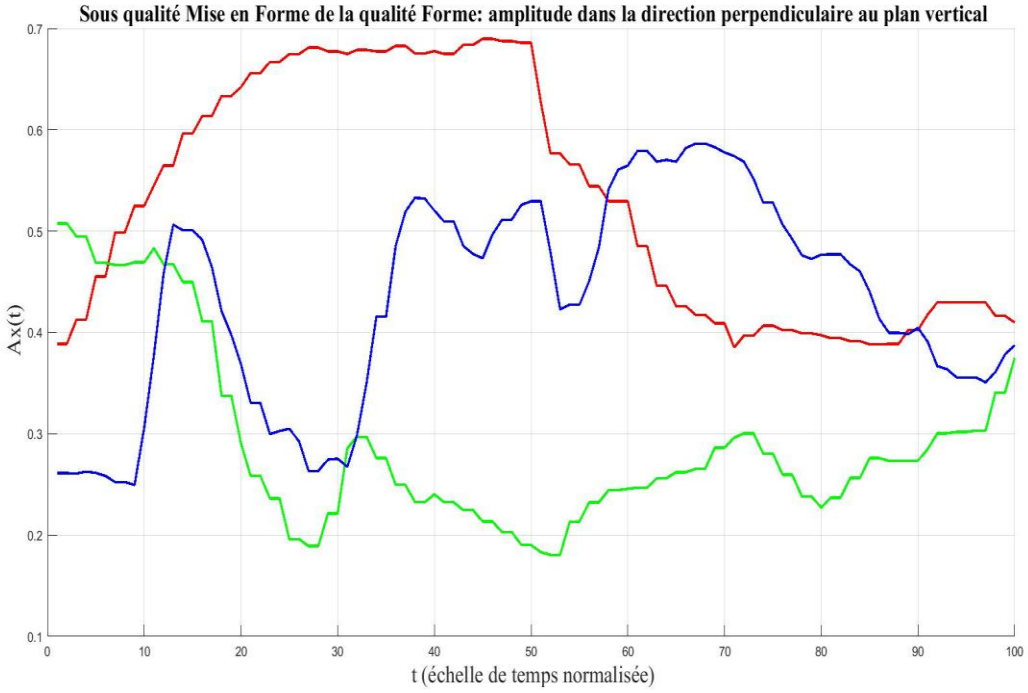


Figure IV.18 Série des amplitudes $A^x(t)$ dans la direction perpendiculaire au plan vertical pour le geste *faire ses lacets* (courbe rouge – *cf.* Figure IV.10), pour un geste ample de chef d'orchestre restant relativement sur place (courbe verte – *cf.* Figure IV.15), et pour un geste dénotant avancements et retraits successifs (courbe bleue – *cf.* Figure IV.9).

L'affichage des profils de ces trois types d'amplitudes $A^x(t)$, $A^y(t)$ et $A^z(t)$ pour trois gestes différents permet de saisir leur caractère discriminant. Ainsi, la Figure IV.16 dédiée à la série des amplitudes $A^y(t)$ selon la direction perpendiculaire au plan horizontal met en valeur une nette diminution des valeurs pour le geste *faire ses lacets* (Figure IV.10) qui nécessite un abaissement (courbe rouge). La Figure IV.17 montre les séquences d'amplitudes $A^z(t)$ sur la direction perpendiculaire au plan sagittal et donne à voir un profil en cloche spécifique (courbe verte) pour le geste ample du chef d'orchestre représenté à la Figure IV.15. Enfin, la Figure IV.18, illustrant l'amplitude $A^x(t)$ sur la direction perpendiculaire au plan vertical, permet d'isoler le mouvement du chef d'orchestre illustré à la Figure IV.9, mêlant penchements successifs du corps vers l'avant et donc extensions/réductions dans la direction concernée (courbe bleue), ainsi que le geste *faire ses lacets* (*cf.* Figure IV.10) pour lequel le corps doit s'étendre vers l'avant au moment où il se plie (courbe rouge).

Introduisons à présent les descripteurs considérés pour représenter la qualité d'*Effort*.

IV.4. Qualité d'*Effort*

Une première composante de la qualité d'***Effort*** concerne la sous-qualité d'*Espace*.

IV.4.1. Sous-qualité d'*Espace*

La sous-qualité d'***Espace*** vise à discerner entre mouvement *direct* ou *rectiligne* d'un mouvement *indirect* ou *flexible*

Pour Nakata *et al.* [**75**], elle est quantifiée par la somme totale des différences entre les angles d'élévation des deux bras et l'angle de hochement de la tête. Ces différences sont pondérées par des facteurs déterminés expérimentalement.

Pour Hachimura *et al.* [**77**], elle est caractérisée à chaque instant par le produit scalaire entre le vecteur symbolisant la direction normale au visage et le vecteur de la direction tangente à la trajectoire du centre des hanches.

Avec Samadani *et al.* [**76**], un autre produit scalaire est considéré : celui entre la tangente à la trajectoire du torse et la tangente à la trajectoire du poignet. Les auteurs réfèrent également à la possibilité de calculer le rapport entre la distance parcourue sur toute la série par une articulation d'intérêt (exemple : une main) et la distance qui sépare son premier lieu de son dernier. Kapadia *et al.* [**74**] reprennent d'ailleurs cette idée.

Aristidou et Chrysanthou [**78**] s'intéressent à l'angle qui relate la direction du corps par rapport à celle du mouvement.

Cette sous-composante d'*Espace*, qui traite de l'attention que le sujet peut porter sur les environs, est intimement liée à la sous-qualité de *Mouvement directionnel*, qui traite de la direction éventuelle du mouvement vers un point particulier (dont les quantifications possibles ont été présentées à la section IV.3.3). Pour les mêmes raisons que pour cette dernière, nous n'avons pas cherché à nous pencher sur cette sous-composante, du fait du peu de mouvement global présent dans les gestes considérés.

IV.4.2. Sous-qualité de *Temps*

La sous-dimension de *Temps* est souvent ramenée à une étude de l'accélération, dans la mesure où elle consiste à différencier les mouvements *soudains* des mouvements plus *soutenus*, *continus*.

Ainsi, elle est caractérisée par Hachimura *et al.* [**77**], pour chaque trame par la somme des accélérations du centre des hanches, des mains et des pieds.

Samadani *et al.* [**76**] se réfèrent également aux accélérations de telles parties du corps pour calculer des descripteurs de *Temps* du geste global.

Kapadia *et al.* [**74**] s'intéressent aussi à l'accélération pour des parties du corps spécifiques et pour des segments de geste correspondant à des intervalles temporels prédéfinis. Ainsi, sur la période d'intérêt, ils proposent une mesure d'accélération cumulée. Un mouvement *soudain* se voit décrit par des pics d'accélération, tandis qu'un mouvement *soutenu* présente une accélération globale faible.

Pour Alaoui [**25**], il s'agit plutôt d'étudier des caractéristiques globales du geste à partir de séries temporelles représentant les caractéristiques spatiales du mouvement :

- la *périodicité* et son inverse, la *fréquence*, sont calculées comme la moyenne du coefficient d'auto-corrélation des quatre extrémités délimitant la silhouette ;
- le paramètre *increase/decrease* décrit l'évolution temporelle (taux de variation) de différents paramètres spatiaux, chacun étant calculable pour une trame gestuelle donnée (la liste de ces paramètres a déjà été présentée à la section III.4.2) ;
- la *quantité de mouvement* exprime la variation entre deux trames successives des aires (exprimées en en termes de nombre pixels les composant) de la silhouette 2D du corps.

Dans leurs travaux sur l'expressivité des agents conversationnels [**65**], J. C. Martin *et al.* proposent le même type d'approche et cherche à analyser l'activité globale (généralement liée à énergie cinétique), ou encore le caractère répétitif ou non du mouvement.

Nous avons privilégié le second type d'approche pour caractériser le phénomène temporel, qui offre l'avantage d'ajouter à la représentation une certaine dimension rythmique. Ainsi, nous considérons en premier lieu une segmentation du geste en périodes dites de faible activité (ou pauses) et périodes de moyenne ou haute activité. Cette division est effectuée à partir de la séquence d'énergie cinétique, définie pour un instant donné *t* comme décrit par l'équation suivante:

$$E_c(t) = \sum_i \left(m_i . v_i^2(t)\right), \; t \in \{1; T\}, \tag{IV.8}$$

où i indexe les articulations correspondant à la main gauche, à la main droite et à la tête, m_i la masse associée et v_i la vitesse correspondante. Les masses associées à la tête et aux mains correspondent à des valeurs moyennes calculées sur un échantillon d'individus comme il est proposé dans [**165**].

Chaque instant du geste est attribué à une période de faible ou forte activité si l'énergie $E_c(t)$ est supérieure à 1/10 de l'énergie maximale observée au cours du geste. Dans le cas contraire, il est considéré comme un instant de pause. Nous calculons premièrement la durée totale des périodes de pause relativement à la durée totale du geste. Ensuite, nous considérons deux séries numériques qui respectivement aux durées des pauses et aux durées des périodes d'activité ; pour chacune de ces deux séries, nous calculons trois paramètres qui sont : la moyenne, l'écart-type et la valeur maximale. Enfin, nous retenons également la durée totale du geste. Au final donc, la sous-composante *Temps* est donc décrite avec 8 valeurs.

Relater de la sorte la rythmique du mouvement permet, sur une période de mouvement suffisamment longue, de discriminer les gestes temporellement amples, voire hors tempo (comme l'exemple illustré Figure IV.13), des gestes extrêmement répétitifs et rythmés comme celui illustré Figure IV.19.

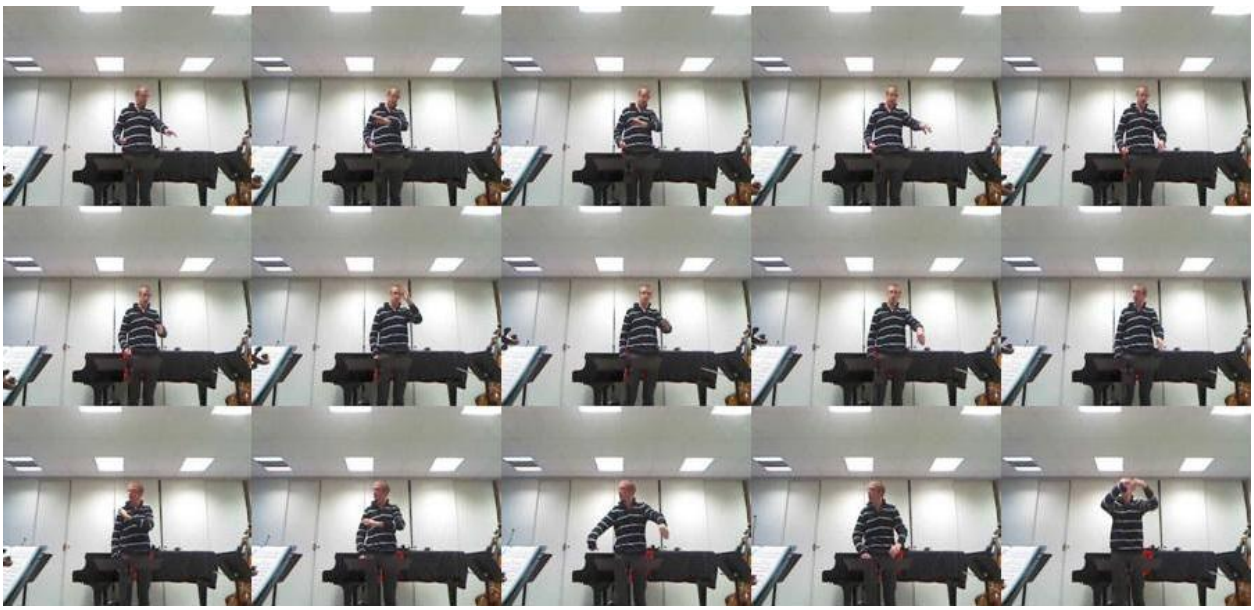


Figure IV.19 Chef d'orchestre au mouvement brusque, à la battue presque militaire, très rapide (base de direction orchestrale – *cf.* section V.2.1).

IV.4.3. Sous-qualité de *Flux*

La sous-qualité de ***Flux*** renseigne sur le caractère tendu ou non du geste. Elle traduit l'attitude de *contrôle*, et distingue un mouvement *libre* d'un mouvement *lié*, *contraint* ou *contrôlé*. Dans leurs travaux sur l'expressivité des agents conversationnels [**65**], J. C. Martin *et al.* invoquent la notion de *fluidité* pour l'intégrer dans un modèle d'agent expressif.

Le *Flux* est généralement décrit à l'aide de la troisième dérivée temporelle de la trajectoire du mouvement, également appelée « à-coup » (*jerk* dans la terminologie anglophone). Pour chaque articulation, Samadani *et al.* [**76**] quantifient cette sous-composante d'*Effort* par la somme sur toute la durée du geste des modules des vecteurs d'à-coup. Kapadia *et al.* [**74**] raisonnent d'une façon similaire pour des parties du corps spécifiques et sur des segments de mouvement d'empans temporels variables. Un mouvement *lié* présente alors des à-coups très forts, tandis qu'un mouvement *libre* présente un à-coup faible du fait de l'uniformité de l'accélération. Aristidou et Chrysanthou [**78**] utilisent également l'à-coup pour caractériser le *Flux*, dans la mesure où ils calculent pour chaque instant la dérivée de l'accélération du centre des hanches.

Pour caractériser le *Flux*, nous retenons la valeur absolue de l'à-coup, que nous calculons pour les trajectoires des deux mains, comme décrit dans les équations suivantes :

$$Acoup^{x,y,z}_{main\ gauche}(t) = \frac{d^3 OPmain\ gauche(t)}{dt^3} \quad \text{(IV.9)}$$

$$Acoup^{x,y,z}_{main\ droite}(t) = \frac{d^3 OPmain\ droite(t)}{dt^3} \quad \text{(IV.10)}$$

Figure IV.20 Réalisation du geste *intercepter un objet* pour la base HTI 2014-2015 (*cf.* section V.2.2).

Figure IV.21 Réalisation du geste *dire merci en langue des signes* pour la base HTI 2014-2015 (*cf.* section V.2.2).

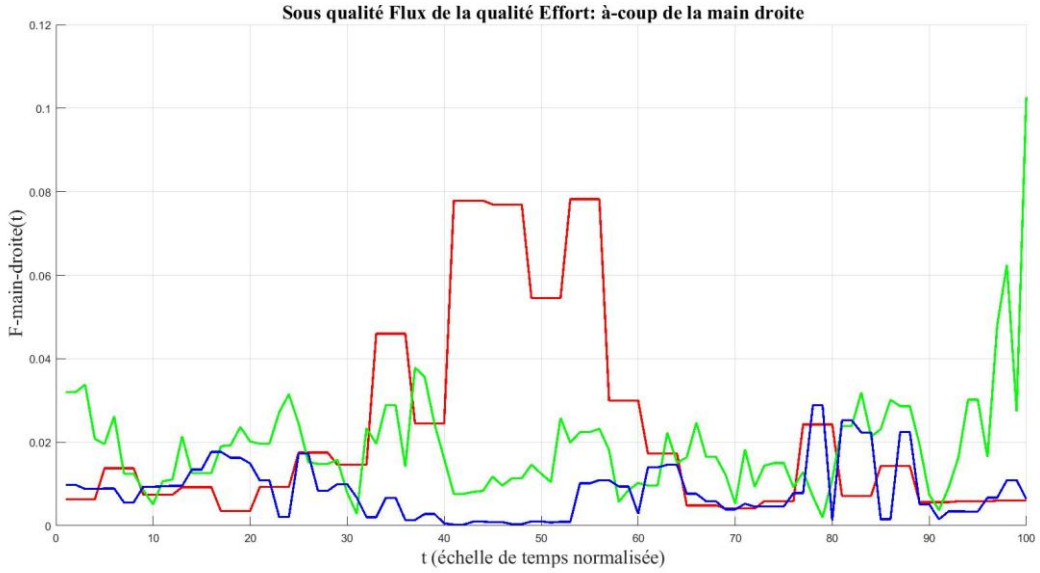


Figure IV.22 Série des indices de flux du mouvement de la main droite $Acoup^{x,y,z}_{main\ droite}(t)$ pour le geste *intercepter un objet* (courbe rouge – *cf.* Figure IV.20), pour un geste de battue sobre et déterminée (courbe verte – *cf.* Figure IV.19) et pour le geste *dire merci en langue des signes* (courbe bleue – *cf.* Figure IV.21).

Les gestes extrêmement secs, quasi inexpressifs (*intercepter un objet* – Figure IV.20) ou saccadés (Figure IV.19), ont selon cette sous-qualité de mouvement un profil qui diffère significativement de celui de gestes plus lisses ou dénotant une forme de facilité d'exécution (exemple : *dire merci en langue des signes* – Figure IV.21). L'illustration du caractère pertinent de notre indice de flux du mouvement est présentée Figure IV.22. Y est mis en évidence un pic d'à-coup pour le geste *intercepter un objet* (courbe rouge) ainsi qu'une série de valeurs d'à-coup très faibles pour le geste *merci en langue des signes* (courbe bleue).
Ces deux gestes, bien que proches du point de vue de leur structuration, ont pu être différenciés convenablement par notre indice de *Flux*.

IV.4.4. Sous-qualité de *Poids*

La sous-qualité ***Poids*** caractérise l'impact du mouvement et différencie un mouvement *lourd* ou *ferme* d'un mouvement *léger*. Elle recoupe la notion de *puissance* que développent pour leurs agents émotionnels Martin *et al.* [**65**], et peut s'avérer utile à la différenciation entre, d'une part, des mouvements hésitants ou dénotant peu d'*Effort* (Figure IV.13), et d'autre part des mouvements dénotant une forme relativement forte d'affirmation (Figure IV.15).

Dans [**163**], Camurri *et al.* réfèrent *Poids* comme à la composante verticale de l'accélération du mouvement.

Dans la plupart des travaux de l'état de l'art, le *Poids* est décrit à un moment donné du geste à partir de l'énergie cinétique. Ainsi pour Nakata *et al.* [**75**], le *Poids* est quantifié par la somme totale des énergies cinétiques à chaque articulation en mouvement, pondérées par le poids de ladite articulation. Il en est de même pour Hachimura *et al.* [**77**]. Samadani *et al.* [**76**], qui utilisent l'énergie cinétique pour les avant-bras, les bras et les doigts.

Dans une approche différente, Samadani *et al.* caractérisent le geste à partir du maximum de la série des décélérations des articulations considérées. Kapadia *et al.* [**74**] et Aristidou et Chrysanthou [**78**]

adoptent un principe similaire. Un mouvement *lourd* est caractérisé par une grande décélération du mouvement des extrémités des membres, là où un mouvement *léger* est représentative d'une décélération faible ou nulle.

Dans notre cas, nous avons adopté une approche par vitesses et accélérations verticales – la direction verticale étant celle d'application du poids – ($y'_{.,t}$ et $y''_{.,t}$) des deux mains et du centre des hanches. Cela conduit à un ensemble de 6 séries temporelles pour décrire la sous-composante de poids :

$$Poids(t) = (v^y_{main\ gauche}(t), v^y_{main\ droite}(t), v^y_{centre\ des\ hanches}(t), a^y_{main\ gauche}(t), a^y_{main\ droite}(t), a^y_{centre\ des\ hanches}(t)) \quad \text{(IV.11)}$$

où :

$$v^y_{main\ gauche}(t) = \frac{dy_{main\ gauche,t}}{dt} \quad \text{(IV.12)}$$

$$v^y_{main\ droite}(t) = \frac{dy_{main\ droite,t}}{dt} \quad \text{(IV.13)}$$

$$v^y_{centre\ des\ hanches}(t) = \frac{dy_{centre\ des\ hanches,t}}{dt} \quad \text{(IV.14)}$$

$$a^y_{main\ gauche}(t) = \frac{d^2y_{main\ gauche,t}}{dt^2} \quad \text{(IV.15)}$$

$$a^y_{main\ droite}(t) = \frac{d^2y_{main\ droite,t}}{dt^2} \quad \text{(IV.16)}$$

$$a^y_{centre\ des\ hanches}(t) = \frac{d^2y_{centre\ des\ hanches,t}}{dt^2} \quad \text{(IV.17)}$$

Cela conclut notre représentation de la sous-qualité de *Poids* de la qualité d'*Effort*.

IV.4. Conclusion

Dans ce chapitre, nous avons développé un ensemble de descripteurs, à la fois globaux (décrivant le geste dans sa globalité, sur sa durée de vie) et locaux (pour la caractérisation des trames individuelles), dédiés aux différentes qualités et sous-qualité du modèle LMA, qui visent à offrir une représentation objective et mesurable des celles-ci.

Les différents descripteurs proposés sont résumés dans le Tableau IV.2, en précisant si l'indice choisi est global ou non. Nous pouvons observer que la majorité des indices sont locaux (*e.g.*, dédiés à une trame temporelle donnée), et seront par conséquent utilisables tout à la fois dans le cas d'une expérience de reconnaissance dynamique du geste (*cf.* chapitre VII), et dans un cadre de reconnaissance globale (*e.g.*, à partir de la description de l'intégralité de l'empan temporel durant lequel un geste est effectué – *cf.* chapitre VI). Dans ce dernier cas, tout indice local sera étudié en tant que série temporelle pour en extraire une caractérisation globale et statistique.

Dans les sections suivantes, nous étudions dans quelle mesure nos descripteurs de Laban, bien que spécifiquement dédiés à l'expressivité et en partie inspirés par le geste artistique et la direction orchestrale, peuvent se montrer aptes à caractériser des gestes quelconques.

Tableau IV.2 Résumé des quantifications de qualités et sous-qualités de Laban.

Qualité de Laban	**Sous-qualité**	**Descripteurs**
Corps		• Distance entre la main gauche et l'épaule gauche : $D^g(t)$ • Distance entre la main droite et l'épaule droite : $D^d(t)$ • Indice de dissymétrie : $Dis(t)$
Espace		• Longueur du chemin parcouru par le centre des hanches : l_{CH} (indice **global**) • Position de la tête vers l'avant : $x_{tête,t}^{trans}$ • Angle de penchement vers l'avant : $\Phi(t)$
Forme	*Flux de forme*	• Indice de contraction : $C(t)$
	Mise en forme	• Amplitude corporelle selon la direction perpendiculaire au plan vertical : $A^x(t)$ • Amplitude corporelle selon la direction perpendiculaire au plan horizontal : $A^y(t)$ • Amplitude corporelle selon la direction perpendiculaire au plan sagittal : $A^z(t)$
Effort	*Temps*	• Caractéristiques **globales** et statistiques, extraites de la segmentation du mouvement en périodes de forte et faible activité, basée sur l'énergie cinétique $E_c(t)$
	Flux	• A-coup pour la trajectoire de la main gauche : $Acoup_{main\ gauche}^{x,y,z}(t)$ • A-coup pour la trajectoire de la main droite : $Acoup_{main\ droite}^{x,y,z}(t)$
	Poids	• Composante verticale de la vitesse pour la trajectoire de la main gauche : $v_{main\ gauche}^{y}(t)$ • Composante verticale de la vitesse pour la trajectoire de la main droite : $v_{main\ droite}^{y}(t)$ • Composante verticale de la vitesse pour la trajectoire du centre des hanches : $v_{centre\ des\ hanches}^{y}(t)$ • Composante verticale de l'accélération pour la trajectoire de la main gauche : $a_{main\ gauche}^{y}(t)$ • Composante verticale de l'accélération pour la trajectoire de la main droite : $a_{main\ droite}^{y}(t)$ • Composante verticale de l'accélération pour la trajectoire du centre des hanches : $a_{centre\ des\ hanches}^{y}(t)$

V. Corpus de gestes 3D

L'objectif de ce chapitre est de présenter le matériel disponible en vue de l'application de nos descripteurs à la reconnaissance d'actions ou d'émotions. A ce jour, peu de bases de gestes existent qui, outre le fait de proposer des actions pré-segmentées, fournissent également un suivi des articulations de référence du corps tout au long du mouvement. Après avoir présenté les corpus disponibles publiquement, et que nous utiliserons dans nos expériences aux sections VI et VII, nous introduisons les deux bases de gestes que nous avons construites. La première, ORCHESTRE-3D, se compose de gestes de chefs d'orchestre pré-segmentés et est annotée à l'aide d'un lexique d'émotions musicales. La deuxième, HTI 2014-2015, propose onze types d'action différents, répartis sur des séquences mettant chacune en jeu plusieurs d'entre eux.

Nous nous sommes intéressés à l'analyse de gestes acquis avec des caméras 3D qui rendent disponibles les trajectoires corporelles (*i.e.*, positions des articulations) en 3D. Avec l'émergence relativement récente des caméras de types Kinect, plusieurs bases de données, dédiées à la reconnaissance des gestes ont été constituées. De telles bases de données présentent l'avantage de fournir des positions de références du corps (*e.g.*, 20 articulations de référence) à un taux d'acquisition raisonnable (*e.g.*, 15-30 trames par seconde) et avec une incertitude n'excédant pas deux centimètres par position d'articulation. La fiabilité d'un tel matériau de base permet d'envisager une analyse des contenus haut-niveau du geste de façon plus aisée que dans le cas de vidéos 2D, où l'extraction d'indices corporels constitue déjà une difficulté en soi et rend alors difficile l'élaboration de modèles sémantiques de la gestualité.

Présentons donc les différentes bases de test de gestes 3D disponibles.

V.1 Bases de données de gestes 3D

Le corpus ***MSR Action 3D*** [**38**] propose 20 types d'actions, choisies dans le contexte d'une interaction avec des consoles de jeu. Ces actions recouvrent des mouvements variés des bras, des jambes, du torse, ainsi que leurs combinaisons. Il a par ailleurs été spécifié aux sujets exécutants d'utiliser leur bras ou leur jambe droit(e) dans le cas d'actions n'impliquant qu'un seul membre.

Les actions sont réparties en trois sous-ensembles, résumés dans le Tableau V.1.

Les ensembles A1 et A2 regroupent des actions proches voire similaires, tandis que l'ensemble A3 met en jeu des actions plus complexes. Les Figure V.1, Figure V.2 et Figure V.3 illustrent des classes de gestes extraites de chacun de ces trois sous-ensembles en proposant des successions de squelettes correspondant à la réalisation des gestes.

Tableau V.1 Classes de gestes du corpus MSR Action 3D [**38**] regroupées par ensemble, avec pour chacune le nombre N_{occ} d'occurrences dans le corpus, ainsi que les durées minimale L_{min} et maximale L_{max} exprimées en secondes.

Ensemble A1				**Ensemble A2**				**Ensemble A3**			
Classe	N_{occ}	L_{min}	L_{max}	**Classe**	N_{occ}	L_{min}	L_{max}	**Classe**	N_{occ}	L_{min}	L_{max}
***se pencher* (*bend*)**	20	1.4	3.7	***dessiner un cercle* (*draw circle*)**	30	1.5	3.8	***coup de pied vers l'avant* (*forward kick*)**	29	1.1	3.8
***coup de poing* (*forward punch*)**	26	1.3	5.0	***cocher une case* (*draw tick*)**	30	1.7	2.9	***swing de golf* (*golf swing*)**	30	2.0	4.7
***coup de marteau* (*hammer*)**	27	1.9	4.5	***dessiner une croix* (*draw x*)**	27	1.4	4.4	***lancer au loin* (*high throw*)**	26	1.5	3.7
***taper des mains* (*hand clap*)**	30	1.1	3.2	***coup de pied vers l'avant* (*forward kick*)**	29	1.1	3.8	***jogging* (*jogging*)**	30	1.8	3.8
***lancer au loin* (*high throw*)**	26	1.5	3.7	***attraper d'une main* (*hand catch*)**	26	1.3	6.6	***ramasser et jeter* (*pick up & throw*)**	22	2.2	3.6
***signe horizontal de la main* (*horizontal arm wave*)**	27	1.9	4.3	***signe de la main* (*high arm wave*)**	27	2.0	4.4	***coup de pied sur le côté* (*side kick*)**	20	1.2	3.7
***ramasser et jeter* (*pick up & throw*)**	22	2.2	3.6	***boxer sur le côté* (*side boxing*)**	30	1.0	4.0	***service cuillère* (*tennis serve*)**	30	0.8	4.5
service cuillère* (*tennis serve)	30	0.8	4.5	***signe avec les deux mains* (*two hand wave*)**	30	0.9	4.7	***service volée* (*tennis swing*)**	30	1.7	3.7

Les actions ont été enregistrées à l'aide d'une caméra Kinect à une cadence d'acquisition de 15 trames par seconde. Le corpus ne propose pas de séquence de succession d'actions, chaque séquence de mouvement correspondant à la réalisation d'une seule action. Par conséquent, un tel corpus ne peut se révéler utile pour des expériences de décodages d'actions successives au sein de séquences gestuelles complexes. Pour la composition du corpus, chaque action a été exécutée deux ou trois fois par une dizaine de sujets, pour un total de 547 séquences gestuelles.

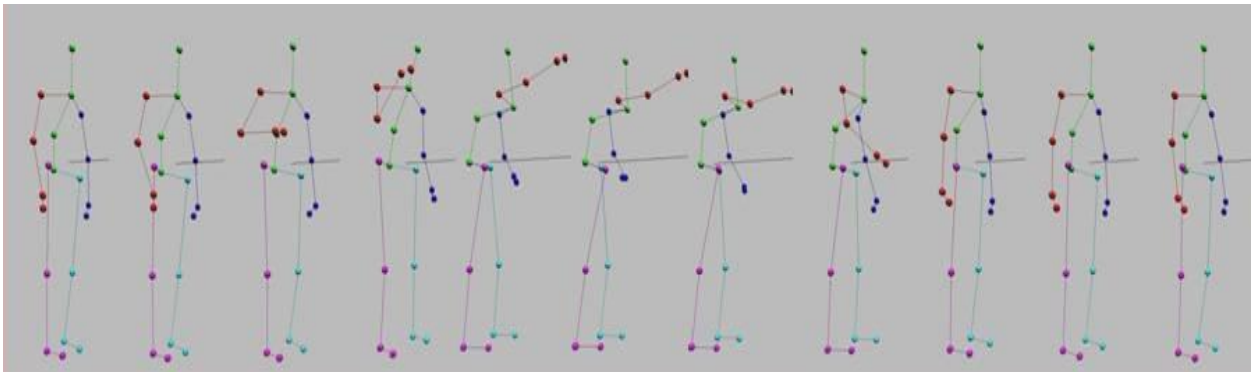


Figure V.1 Train de squelettes correspondant à la réalisation du geste *coup de poing* présent dans le sous-ensemble A1 du corpus MSR Action 3D [**38**].

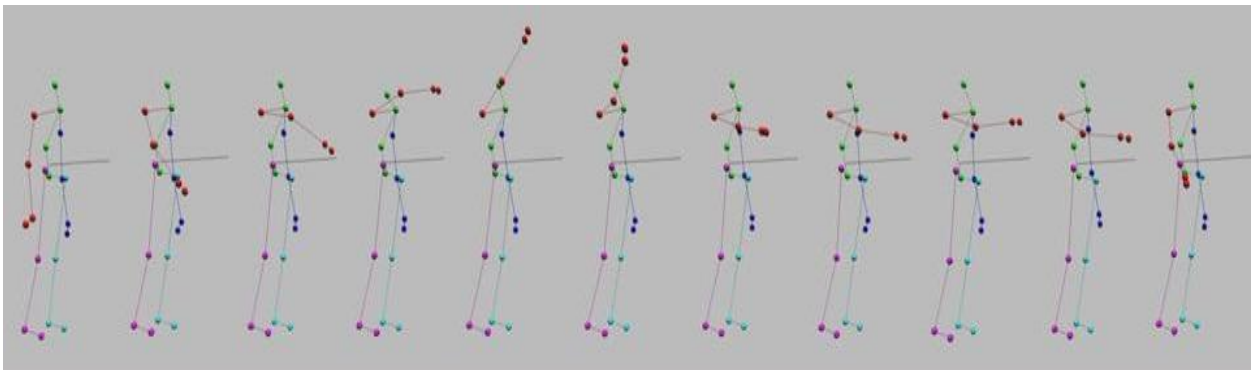


Figure V.2 Train de squelettes correspondant à la réalisation du geste *dessiner un cercle* présent dans le sous-ensemble A2 du corpus MSR Action 3D [**38**].

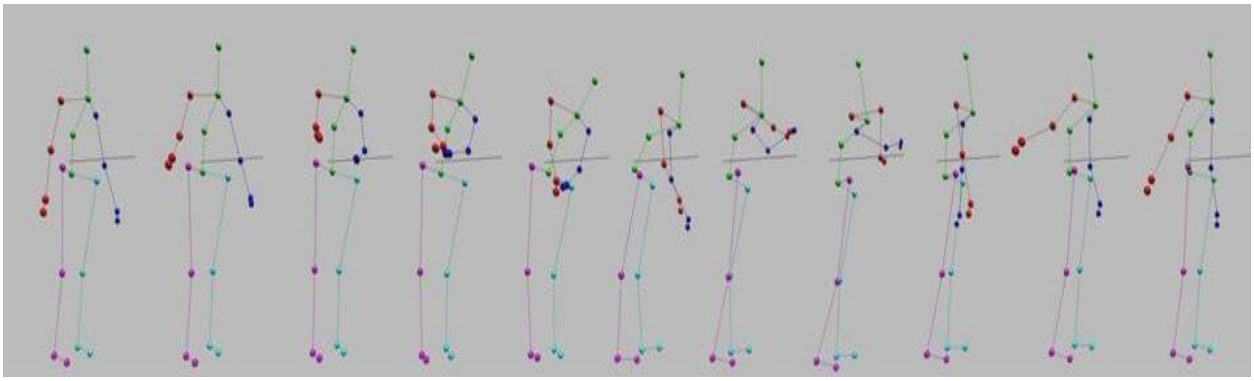


Figure V.3 Train de squelettes correspondant à la réalisation du geste *swing de golf* présent dans le sous-ensemble A3 du corpus MSR Action 3D [**38**].

Une deuxième base de test 3D est le corpus ***UTKinect-HumanDetection*** [**37**], qui propose des séquences de gestes mettant en jeu 10 catégories d'action différentes, résumées dans le Tableau V.2.

Les squelettes 3D des actions ont été acquis à l'aide d'une caméra Kinect. Les auteurs précisent que la présence de trames RGB pour lesquelles aucun squelette n'a pas pu être rendu tend à faire baisser le taux d'acquisition des squelettes à 15 trames par seconde.

Par ailleurs, les séquences gestuelles proposées consistent en des successions d'actions identiques, c'est-à-dire systématiquement effectuées dans le même ordre.

Tableau V.2 Classes de gestes du corpus UTKinect-HumanDetection [**37**], avec pour chacune le nombre N_{occ} d'occurrences dans le corpus, ainsi que les durées minimale L_{min} et maximale L_{max} exprimées en secondes.

Classe	N_{occ}	L_{min}	L_{max}
porter (***carry***)	19	3.8	15.1
taper des mains (***clap hands***)	20	2.0	7.1
ramasser (***pick up***)	20	2.5	7.5
tirer (***pull***)	20	1.1	5.7
pousser (***push***)	20	0.8	2.2
s'asseoir (***sit down***)	20	2.5	5.6
se lever (***stand up***)	20	1.6	4.4
jeter (***throw***)	20	0.7	3.0
marcher (***walk***)	20	1.2	6.0
faire signe avec les mains (***wave hands***)	20	3.6	13.5

Les Figure V.4 et Figure V.5 illustrent respectivement les classes de gestes *marcher* et *taper des mains*, à l'aide de successions de squelettes correspondantes.

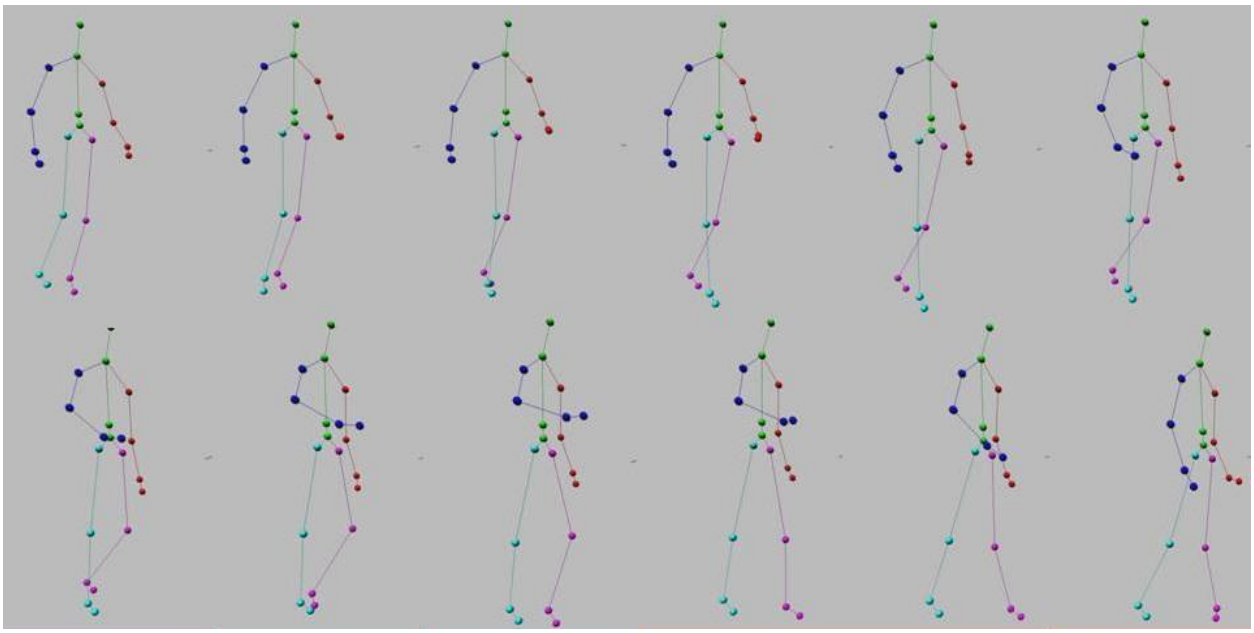


Figure V.4 Train de squelettes correspondant à la réalisation du geste *marcher* présent dans le corpus UTKinect-HumanDetection [**37**].

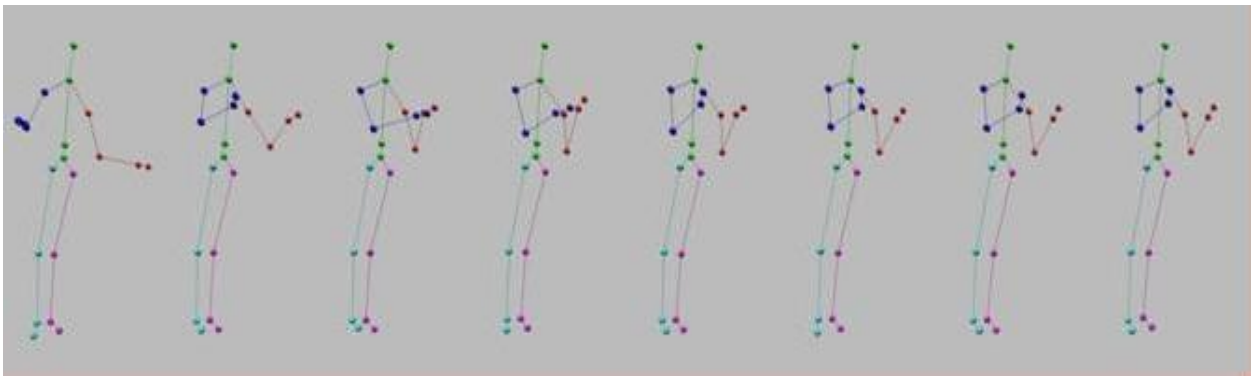


Figure V.5 Train de squelettes correspondant à la réalisation du geste *taper des mains* présent dans le corpus UTKinect-HumanDetection [**37**].

Le corpus ***Microsoft Research Cambridge-12*** [**36**] (MSRC-12 – Figure V.6) propose également des poses de squelette enregistrées à l'aide d'une Kinect à une fréquence de 30Hz. Le corpus MSRC-12 rend disponible 12 actions correspondant à deux catégories de référence de la taxonomie de David McNeill [**41**] (*cf.* section II.1.1) : il s'agit :

- de *gestes iconiques*, qui représentent un objet, par des évocations de forme, d'extension spatiale, de taille, d'orientation, ou une action concrète
- et de gestes *métaphoriques* qui tentent de représenter des idées plus abstraites.

Figure V.6 Exemples de poses enregistrées en vue de la constitution du MSRC-12 dataset [**36**].

Tableau V.3 Classes de gestes du corpus Microsoft Research Cambridge-12 [**36**], avec pour chacune le nombre N_{occ} d'occurrences dans le corpus, ainsi que les durées minimale L_{min} et maximale L_{max} exprimées en secondes.

Gestes iconiques				Gestes métaphoriques			
Classe	N_{occ}	L_{min}	L_{max}	**Classe**	N_{occ}	L_{min}	L_{max}
***s'accroupir ou se cacher* (*Crouch or hide*)**	450	1.2	44.0	***lancer la musique/monter le volume* (*Start music/Raise volume*)**	458	1.3	23.1
***tirer au pistolet* (*Shoot a pistol*)**	462	1.5	11.2	***naviguer vers le menu suivant* (*Navigate to next menu*)**	471	1.6	11.3
***jeter un objet* (*Throw an object*)**	465	1.5	10.2	***stopper la musique* (*Wind up the music*)**	601	0.4	32.8
***changer d'arme* (*Change weapon*)**	450	1.7	8.0	***s'incliner pour clore la session musicale* (*Take a bow to end music session*)**	458	1.7	10.2
***donner un coup de pied* (*Kick*)**	454	1.5	7.6	***protester contre la musique* (*Protest the music*)**	458	1.9	14.3
***mettre des lunettes de protection* (*Put on night vision goggles*)**	458	1.9	8.8	***accélérer le tempo* (*Move up the tempo of the song*)**	468	1.2	30.1

Les différentes catégories proposées ainsi que leurs caractéristiques en termes de nombre d'occurrences et longueurs minimale et maximale sont résumées dans le Tableau V.3.

La base de données correspond à un total de 6 heures et 40 minutes de mouvements et de plus de 5500 gestes répartis sur environ 600 séquences gestuelles réalisées par trente individus. Chacune des séquences de mouvement correspond à plusieurs réalisations successives d'une même action. Par conséquent, un tel corpus ne peut pas se révéler utile en vue de l'analyse de séquences gestuelles correspondant à des successions d'actions.

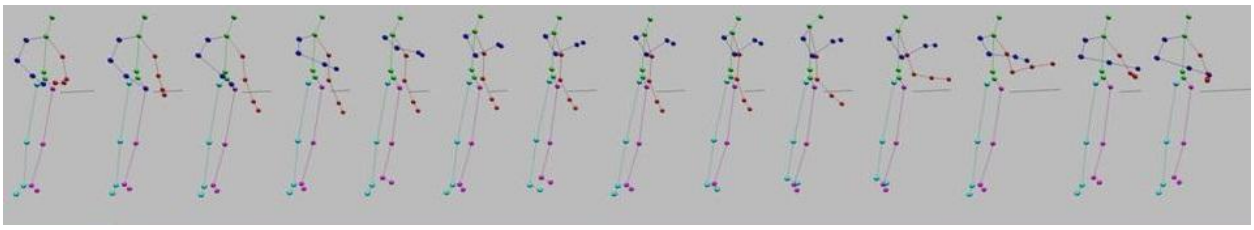


Figure V.7 Train de squelettes correspondant à la réalisation du geste iconique *changer d'arme* présent dans le corpus MSRC-12 [**36**].

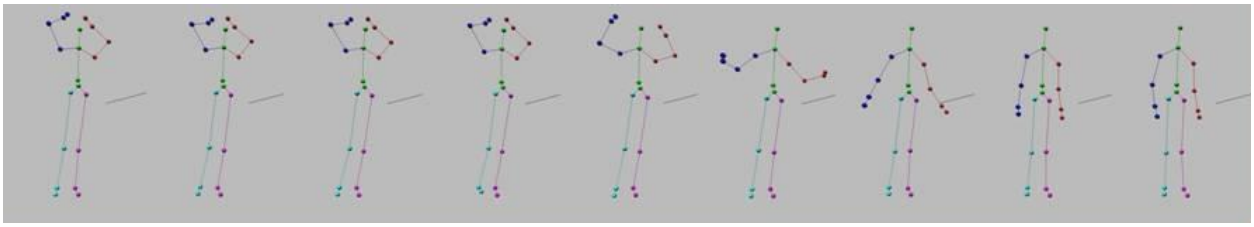


Figure V.8 Train de squelettes correspondant à la réalisation du geste métaphorique *protester contre la musique* présent dans le corpus MSRC-12 [**36**].

Pour illustrer le type de contenu gestuel présent dans le corpus MSRC-12, nous représentons aux Figure V.7 et Figure V.8 des séquences de squelettes correspondant chacune à un type de geste présent dans la base : *changer d'arme* (geste iconique) et *protester contre la musique* (geste métaphorique).

Par comparaison avec les autres bases de gestes 3D, le corpus MSRC-12 présente un certain nombre d'avantages : taux d'acquisition élevé d'environ 30 trames par seconde, nombre de gestes élevé, pertinence des diverses actions proposées. Nous avons également constaté à la visualisation des séries de squelettes que les poses 3D sont fiables et de bonne qualité. Par ailleurs, dans le cas des séquences où une même action est réalisée plusieurs fois d'affilée, une pré-segmentation de ces séquences en réalisations unitaires de l'action considérée est également disponible.

Pour toutes ces raisons, nous avons privilégié le corpus MSRC-12 pour nos expérimentations concernant l'analyse globale de gestes (*cf.* section VI).

En ce qui concerne la reconnaissance dynamique du geste (*cf.* section VII), nous avons retenu l'ensemble des corpus MSR Action 3D, MSRC-12 et UTKinect-HumanDetection, ce qui nous permet notamment de nous comparer à différentes approches de la littérature.

Notons cependant que les bases de données 3D disponibles sont très orientées action, sans prendre en compte les caractéristiques de plus haut niveau et notamment les aspects émotionnelles des gestes. Afin de s'affranchir de cette limitation, nous avons constitué par nos propres moyens deux nouvelles bases de données. Une première, appelée ORCHESTRE-3D, est spécifiquement dédiée à l'analyse émotionnelle de la direction orchestrale. La seconde, appelée HTI 2014-2015, est quant à elle dédiée à la reconnaissance

dynamique des gestes et inclut des séquences correspondant à différentes succession d'actions (4 à 6). Présentons donc maintenant ces deux nouvelles bases, avec protocoles de constitution respectifs et caractéristiques associées.

V.2. Corpus constitués et enjeux

V.2.1. ORCHESTRE-3D : base de données dédiée à l'analyse émotionnelle de la direction orchestrale

Lors de l'introduction à nos travaux, nous avons souligné que l'art, et plus spécifiquement la musique étaient des lieux privilégiés de l'expressivité gestuelle (*cf.* section I.3). Nous avons présenté cette expressivité dans une perspective communicatrice et intersubjective, qui intègre l'exploration corporelle et la respiration [**30**] [**72**], l'accompagnement gestuel de la production du son [**32**] [**31**], les jeux d'anticipation [**8**], les dimensions symbolique et conventionnelle [**43**], ou encore l'expression d'émotions ou de sentiments [**35**] [**17**]. Par ailleurs, nous avons vu [**28**] que l'expressivité musicale tend à s'élaborer à partir de messages sensoriels du corps soulignés le plus souvent par la vue, et désignés comme étant des « abstraits sentimentaux », de sorte que des correspondances en viennent à se dessiner entre des figures musicales et des profils spatio-temporels naturels.

Le chef d'orchestre occupe au sein de la communauté artistique une position très singulière. Son rôle est d'incarner la *musique à immédiatement venir* à l'aide de son corps, en exprimant métaphoriquement ce qu'il attend de la part de ses musiciens. Outre des éléments gestuels relatifs aux nuances, aux tempi, à la battue, aux attaques, ou encore à la qualité du son, il semble bien qu'une telle *incarnation* de la musique repose sur des représentations communes aux musiciens et au chef d'orchestre – représentations qu'en général, ce dernier ne cherche pas particulièrement à verbaliser.

Nous avons alors supposé que ce *quelque chose de commun* repose sur des représentations émotionnelles. Selon cette hypothèse, la communication gestuelle du chef d'orchestre implique des émotions sous-jacentes. De telles émotions sont un *quelque chose* qui se joue à la périphérie de la conscience du musicien, et qui dans un premier temps échappe à cette conscience, tout en étant « vécu » par elle. Ces « émotions vécues » sont ce que le philosophe Merleau-Ponty appelle des « configurations » (*e.g.*, *gestalt* dans la terminologie allemande) dont il échoit à la conscience d'enfermer l'existence dans le concept même d' « émotion ».

Plusieurs chefs d'orchestre et musiciens, dont certains professionnels et intéressés par la transdisciplinarité entre la musique et les sciences sociales, ont fait leur l'idée que la caractérisation de gestes musicaux à l'aide de catégories d'émotions de base, ou même musicales, constituerait une avancée dans l'appréhension d'un langage émotionnel sous-jacent à la pratique de la musique, et qui plus est permettrait d'établir des ponts entre émotions musicales et indices dynamiques du mouvement. Six chefs d'orchestre nous ont listé chacun un certain nombre de caractéristiques visuelles décisives qu'ils ont l'habitude de mettre en jeu dans la production de leurs gestes directeurs.

Les caractéristiques communes qui sont ressorties de ce sondage sont les suivantes :

- symétrie ou dissymétrie générale du corps,
- extension spatiale,
- poids du mouvement,
- caractère plus ou moins franc avec lequel le mouvement est effectué,
- degré de penchement vers avant,
- quantité de mouvement vers l'avant,

- dynamique de la main non-dominante (relatant le type de contenu sonore demandé).

De tels indices s'intègrent parfaitement dans une description du geste du type de celle que nous avons construite en nous inspirant de l'analyse LMA, et semblent tout indiqués pour l'analyse du contenu émotionnel des gestes orchestraux (*cf.* section VI.3).

Figure V.9 Exemples de chefs d'orchestre enregistrés en répétition pour constituer notre base de gestes.

La constitution de notre corpus de direction orchestrale s'est faite comme suit : nous avons enregistré 8 répétitions d'orchestres avec une caméra Kinect, et stocké les articulations de référence des chefs d'orchestre nécessaires à l'application de notre modèle (Figure V.9). Ces répétitions touchaient à des musiques de styles variés (classique, musique de film, jazz…).

Chaque session d'enregistrement a donné lieu à des extraits musicaux que nous avons segmentés scrupuleusement à la main, pour adresser et résoudre divers problèmes.

- La tourne des pages de la partition provoque généralement un avancement du chef vers le pupitre et une perte de la captation. Les segments vidéo correspondants ont donc été supprimés.
- Divers gestes fortuits, insignifiants du point de vue musical, n'ont pu être conservés.
- Les occlusions dues à la présence d'un musicien jouant debout à proximité du chef d'orchestre, et malheureusement capté par la Kinect à sa place, nous ont imposé des suppressions de pans entiers d'enregistrements.
- Le stoppage de la musique par le chef pour recommandations à l'orchestre nous a obligés à interrompre régulièrement la captation.
- Les effets d'exagération voire d'« insistance » dans la gestualité, causés par la répétition du même passage par l'orchestre jusqu'à obtention de la qualité de son voulue, a engendré une certaine redondance. Dans un tel cas, nous avons conservé pour notre corpus seulement la première occurrence des passages musicaux en question.

A cause de ces conditions de répétition (et même de répétition générale d'avant concert pour l'une d'entre elles), nous n'avons donc pas pu nous contenter d'enregistrer des séquences correspondant à des pièces ou des mouvements entiers. Nous avons dû examiner scrupuleusement toutes les séquences enregistrées de façon à fournir aux futurs annotateurs des échantillons gestuels cohérents, pouvant chacun être considérés comme un tout interprétable sans ambiguïté et avec un squelette corporel de qualité.

Nous avons ainsi recueilli 892 gestes pré-segmentés et de durées variables (de 2 à 20 secondes) que nous avons présentés à des musiciens afin qu'ils les annotent à l'aide de catégories émotionnelles.

Pour cela, nous avons tout d'abord bâti avec de jeunes musicologues un lexique (quinze musiciens ont participé bénévolement à ce travail), inspiré des modèles classiques de catégories d'émotions (*cf.* section II.2.1), mais amélioré de façon à proposer des catégories relatant des émotions musicales, similairement au travail effectué dans [**35**] où il s'agissait de qualifier avec des émotions des morceaux de musique classique proposés à l'écoute. Au total, notre lexique contenait les 17 émotions suivantes : *calme*, *agité*, *éveillé*, *tendu*, *serein*, *magique*, *mystérieux*, *heureux*, *triste*, *facile*, *surprenant*, *troublé*, *mélancolique*, *inquiétant*, *colérique*, *tragique*, et *inexpressif*.

Figure V.10 Visualisation de notre espace d'annotation des gestes de direction orchestrale. Les éléments rouges désignent les classes d'émotion qui ont été choisies par le participant lors d'un passage précédent par l'interface d'annotation. Un tel rappel est fait pour que ce dernier se remémore ses choix précédents dans le cas où il en aurait besoin en vue de leur modification, auxquel cas ses nouveaux choix correspondent au cochage des cases.

Ensuite, nous avons demandé à 15 musiciens à la fois professionnels et amateurs d'annoter le corpus à l'aide de ces catégories. Pour cela, nous avons développé une interface web permettant la mise en ligne des échantillons vidéo à annoter en termes émotionnels. Cette interface est illustrée Figure V.10. Pour

chaque segment gestuel, l'annotateur doit choisir entre 1 et 3 catégories émotionnelles pour caractériser l'intentionnalité expressive du chef exécutant le mouvement. Pour qualifier un extrait en termes d'émotions, l'annotateur musicien fait donc l'expérience d'une « configuration » émotionnelle spécifique à l'œuvre dans la gestuelle, avant de la ramener à un concept – qu'il doit donc définir ici à l'aide de 1 à 3 mots.

Précisons ici que les participants à l'annotation n'ont pas eu accès aux échantillons sonores correspondants aux vidéos. En effet, nous avons estimé que l'accès au contenu sonore aurait occasionné une interprétation naturelle de la musique, et aurait dissuadé les individus de se concentrer pleinement sur la gestualité.

Pour chaque échantillon gestuel du corpus, un histogramme de choix d'émotions par les participants a ainsi pu être construit, dont seules les 3 émotions les plus représentées ont été gardées comme annotation finale (*e.g.*, vérité terrain).

La Figure V.11 illustre la distribution des catégories émotionnelles qui a résulté de cette étape d'annotation. On y observe que parmi les émotions considérées initialement, un certain nombre n'a été que peu voir pas du tout considéré par les annotateurs. Comme notre objectif était d'exploiter cette base en tant que vérité terrain dans le cadre d'une procédure d'apprentissage supervisé, il était indispensable de ne considérer que les émotions suffisamment représentées et pouvant faire l'objet d'études statistiques. Nous avons donc rejeté les classes dont la probabilité d'apparition au sein du corpus était inférieure à 20%. Parmi celles-ci, nous avons cependant retenu les classes *mystérieux* et *colérique*, parce que la variabilité de leurs expressions potentielles nous intéressait. Nous avons également éliminé la classe *heureux*, dans la mesure où il ne nous paraissait pas légitime de la conserver en l'absence de la classe *triste*. Cette étape de troncature nous a conduits à un lexique consolidé regroupant les 9 émotions suivantes : *calme*, *agité*, *éveillé*, *tendu*, *serein*, *magique*, *mystérieux*, *facile* et *colérique*. Notre corpus final se compose de 882 séquences de gestes annotées selon les 9 catégories émotionnelles obtenues et disponibles pour une expérience de reconnaissance d'émotions (*cf.* section VI.3).

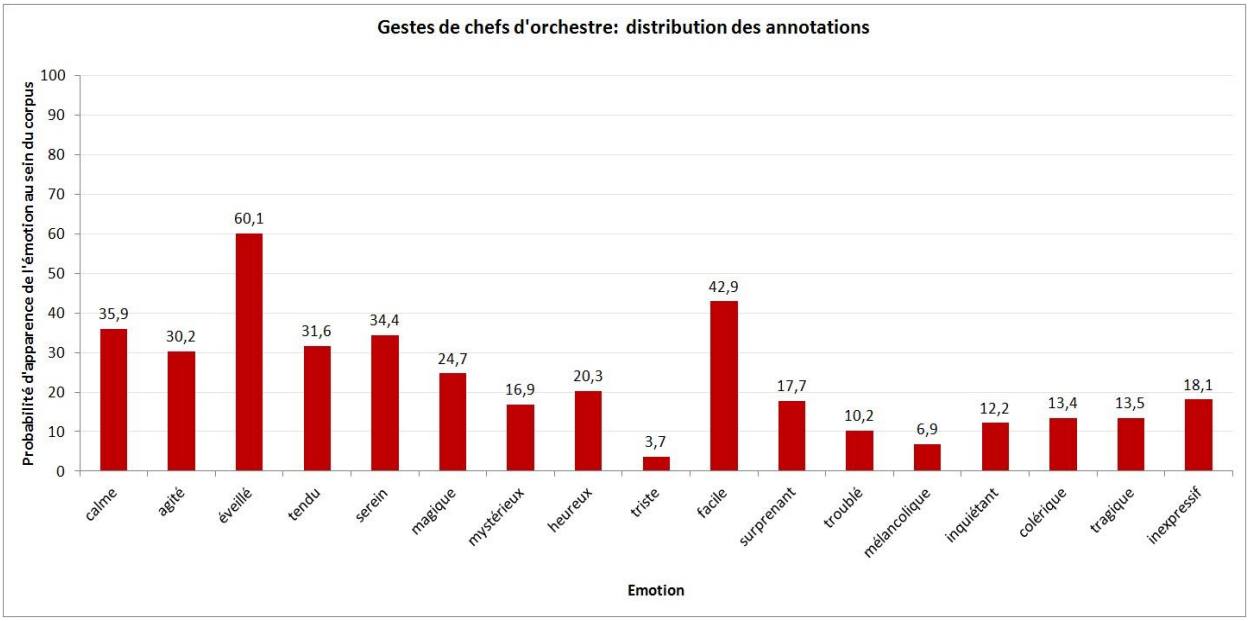


Figure V.11 Distribution des choix de catégories émotionnelles à la fin de l'annotation du corpus des gestes de direction orchestrale. Pour chaque classe, la valeur indiquée correspond à la probabilité pour ladite classe d'être parmi les 3 émotions les plus choisies par les participants lors de l'annotation d'un échantillon quelconque.

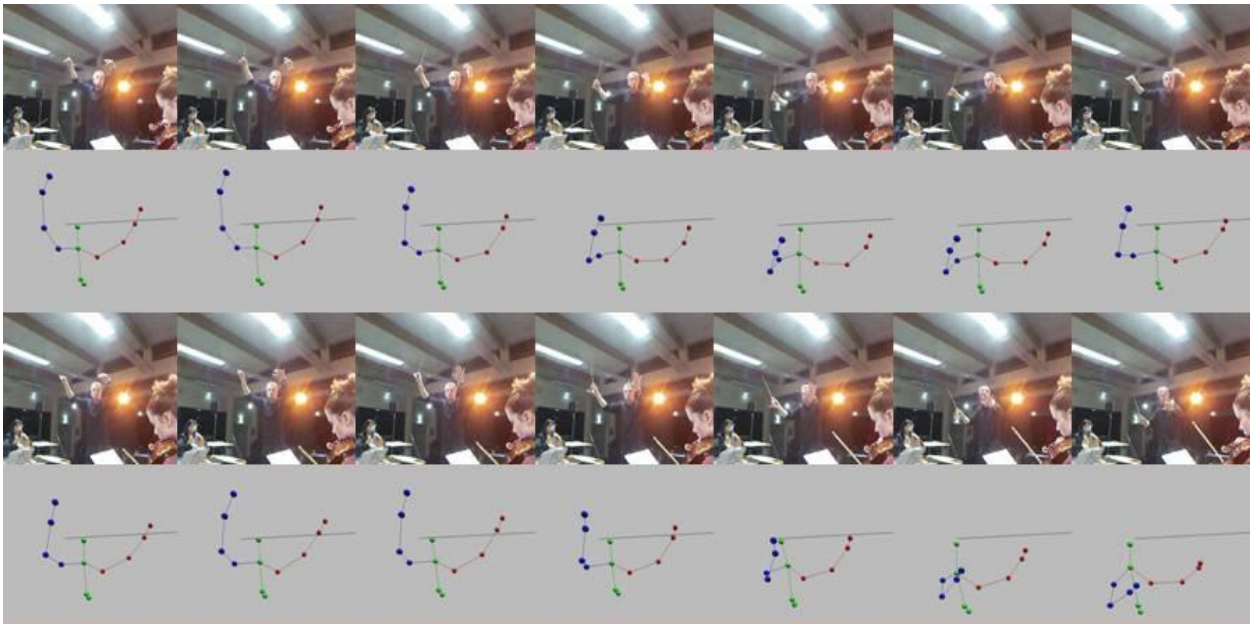


Figure V.12 Séquence d'images, avec les squelettes corporels correspondants, enregistrée lors de la répétition générale d'un concert à l'Etang-la-Ville (les Yvelines, France) en février 2013, pour la constitution du corpus ORCHESTRE-3D. Par des mouvements de bras lents et très communicatifs, le chef d'orchestre exprime calme et sérénité.

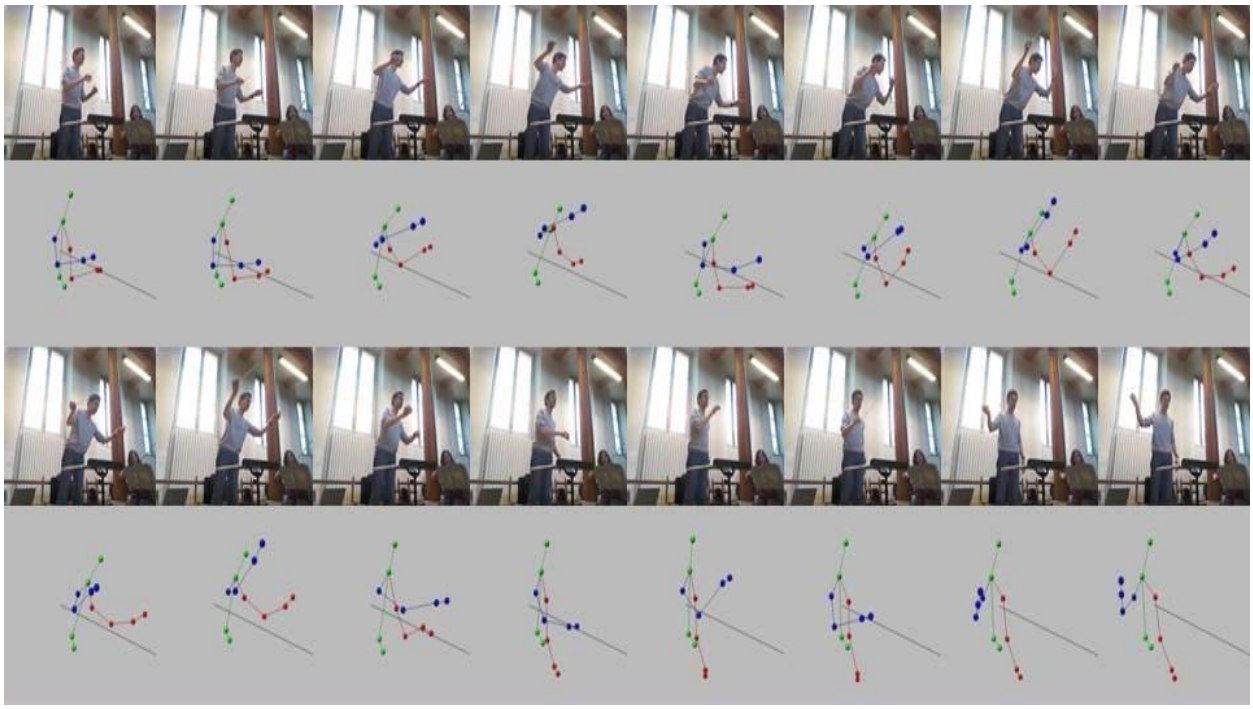


Figure V.13 Séquences d'images et de squelettes corporels enregistrées au lycée Louis-le-Grand (ville de Paris, France) en avril 2013 lors d'une répétition de l'orchestre de l'établissement, pour la constitution du corpus ORCHESTRE-3D. Mouvement heureux, agréable et engagé dans la musique.

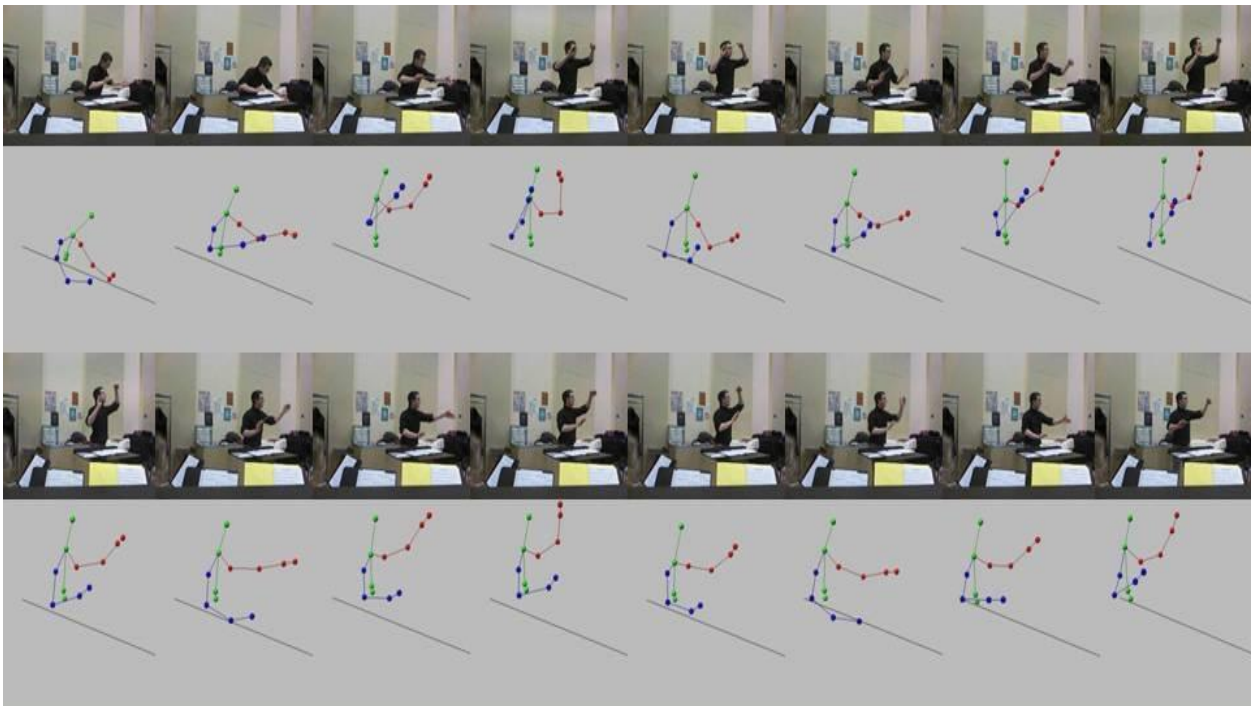


Figure V.14 Séquences d'images et de squelettes enregistrées lors d'une répétition d'un orchestre universitaire à la Maison de la Musique de la ville de Nanterre (Hauts-de-Seine, France) en avril 2013, pour la constitution du corpus ORCHESTRE-3D. Le geste est allant, entrainant, et dénote même une forme d'insistance.

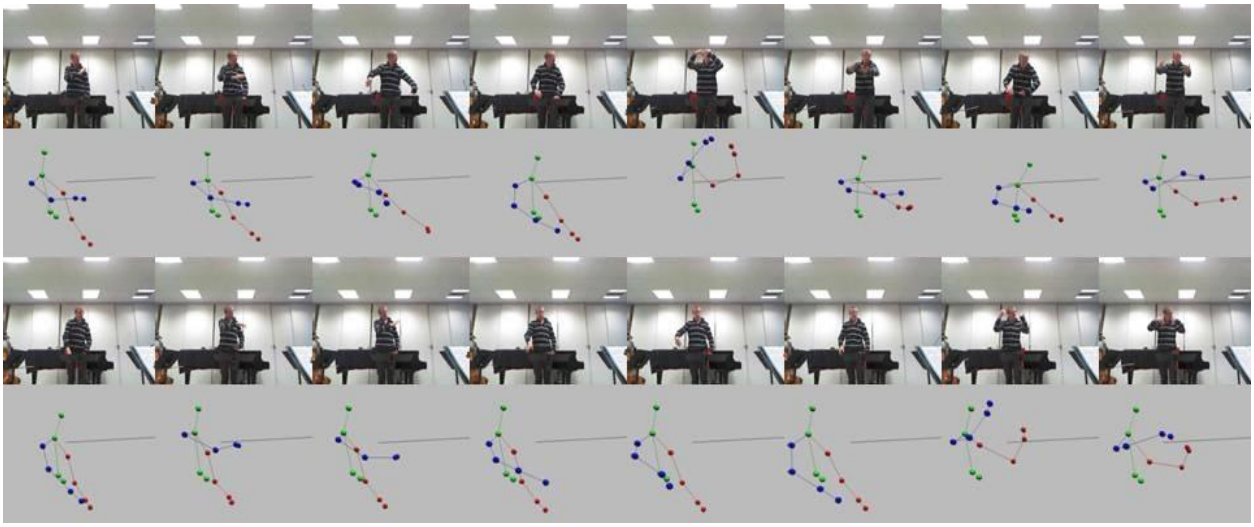


Figure V.15 Séquence d'images, avec les squelettes corporels correspondants, enregistrée lors d'une répétition d'orchestre à l'Ecole de Musique du Centre de Lille (Nord, France) en mai 2013, pour la constitution du corpus ORCHESTRE-3D. Gestuelle colérique, tendue, voire presque militaire.

Les Figure V.12, Figure V.13, Figure V.14 et Figure V.15 proposent des gestes issus de différentes sessions d'enregistrement. Sur chaque figure, la séquence d'images correspond à un geste de direction orchestral pré-segmenté. Pour chaque train d'images représenté en ligne, la série des squelettes corporels enregistrés par la Kinect est également illustrée à la ligne suivante. Seul le haut du squelette est représenté, du fait du masquage quasi constant du corps par les pupitres du chef d'orchestre ou des musiciens.

Avant de poursuivre, nous souhaitons apporter quelques compléments de précisions nécessaires à propos de l'annotation de notre corpus de gestes musicaux.

- Nous n'avons donné aucune indication précise aux annotateurs quant à l'utilisation du lexique d'émotions, en dehors de l'unique consigne de réduire le nombre de catégories choisies au maximum de 3. Si certains individus ont sans doute essayé pour chaque échantillon de trouver la ou les catégories qui semblaient représenter le mieux leur impression, il est possible que d'autres, par exemple, aient préféré de polariser leur choix autour de certaines catégories spécifiques et en apparence génériques (*éveillé*, *calme*…), pour ensuite ajuster ou approfondir leur description émotionnelle à travers des catégories plus spécifiques (*tendu*, *colérique, mélancolique)*. Il est également plausible que des individus se soient sentis poussés, par la variabilité de la palette d'émotions qui se proposait à eux, à choisir systématiquement 3 classes pour qualifier chaque geste, quitte à forcer l'usage du dictionnaire pour certains gestes qu'en réalité ils auraient pu se contenter de ne qualifier qu'avec une seule catégorie. Nous avons en quelque sorte laissé nos participants élaborer leur propre stratégie de « composition » en matière de description de la communication émotionnelle.
- En outre, et en vertu de l'absence d'étape de préparation à la manipulation du lexique, nous n'avons pu disposer d'aucune information qui puisse relater la régularité ou l'évolution de la stratégie d'annotation de chacun des annotateurs. Pour chacun d'entre eux, nous n'avons eu aucune possibilité de savoir si la contribution affectée à chaque mot du lexique dans la modélisation de l'émotion était restée stable ou avait varié avec la perception progressive du corpus par l'annotateur, sans cesse renouvelée à mesure que ce dernier cheminait en son sein.
- Au demeurant, nous avons laissé aux participants la possibilité de modifier leurs choix jusqu'à la clôture définitive de l'annotation du corpus, au cas où ils auraient estimé devoir vérifier la consistance de leur propre usage du lexique d'émotions (Figure V.10).

L'impossibilité de nous référer à une quelconque prise en main du système d'annotation par les participants justifie que dans la section VI.3 dédiée à l'analyse émotionnel des gestes de ce corpus, nous ayons considéré chaque classe comme étant indépendante des autres (*e.g.*, un classifieur par catégorie).

A la différence de la majorité des bases de données gestuelles, qui sont constituées de mouvements *actés*, les gestes musicaux traités ici ont un caractère *spontané* et ont été enregistrés en conditions réelles, mettant alors en jeu des expressions corporelles authentiquement communicationnelles. Toutefois, les chefs d'orchestre impliqués ont témoigné du fait qu'être filmé et se savoir être l'objet d'une étude scientifique avait forcément dû modifier leur comportement et la théâtralité de leur gestuelle, et cela d'autant plus qu'il s'agissait bien ici de gestes expressifs censés *représenter des émotions*, et non d'émotions prises sur le vif. Nous aurions souhaité pouvoir comparer ces gestes orchestraux enregistrés en répétition avec les mouvements des mêmes individus en situation de concert, mais ni les conditions ni le temps ne nous l'ont malheureusement permis.

Dans une optique bien différente, la deuxième base de gestes 3D, appelée HTI 2014-2015, est dédié à la reconnaissance dynamique, à la volée, d'actions.

V.2.2. Base de données d'actions : HTI 2014-2015

Les bases de gestes 3D introduites dans le paragraphe V.1.2 incluent des gestes individuels (MSR Action 3D dataset), des séquences de répétition d'un même geste (MSRC-12 dataset), ou des séquences

où les gestes se succèdent systématiquement dans le même ordre (UTKinect-HumanDetection dataset). Par ailleurs, hormis le MSRC-12 dataset, les taux d'acquisition des squelettes sont relativement faibles (15 trames/s). Pour des objectifs de caractérisation dynamique du geste (*cf.* chapitre VII), nous avons décidé d'élaborer notre propre corpus d'actions.

La base de données fournit des séquences de 4 à 6 actions successives au format Kinect (*e.g.*, séries de squelettes corporels enregistrés au taux de 30 trames par seconde).

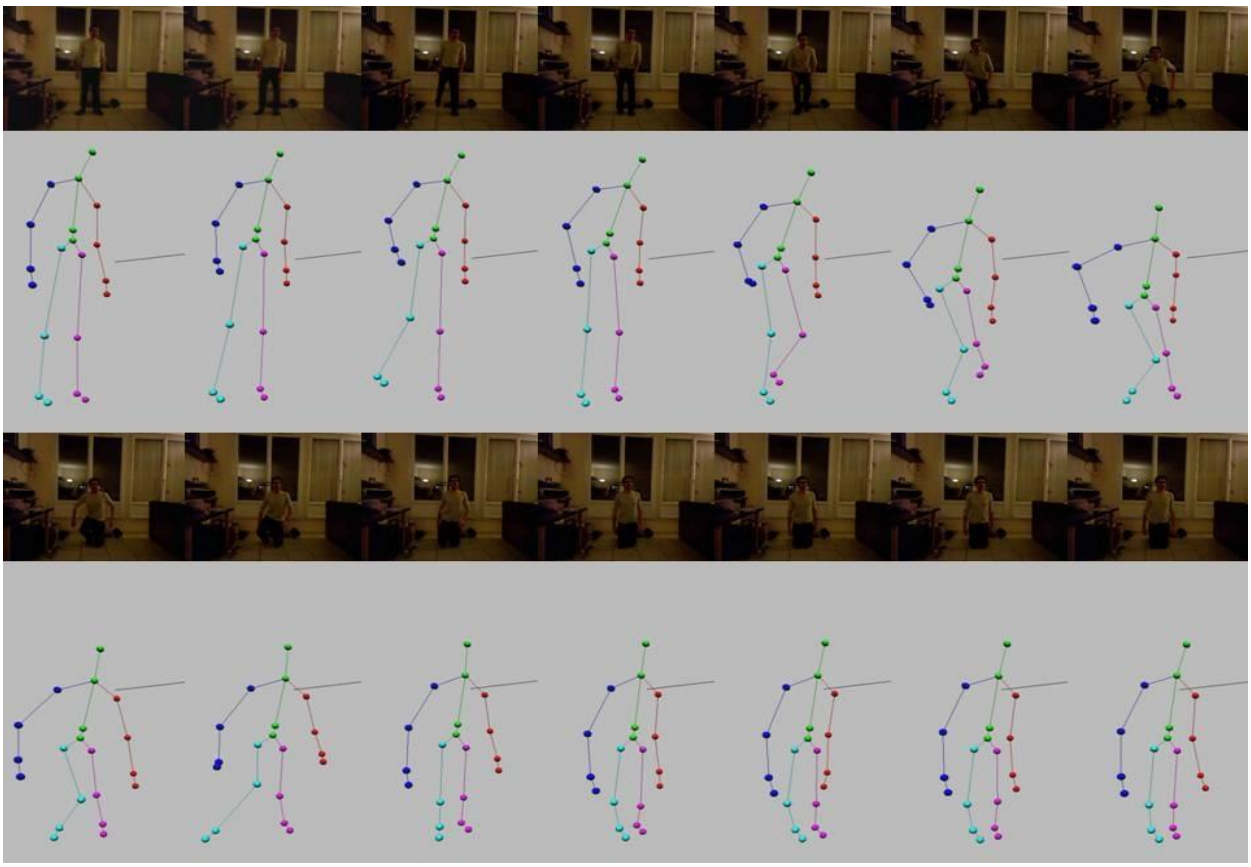


Figure V.16 Visualisation d'images et de squelettes coprorels lors de l'exécution du geste *se mettre à genoux* par un étudiant de la classe HTI 2014-2015 lors de la constitution du corpus de gestes du même nom.

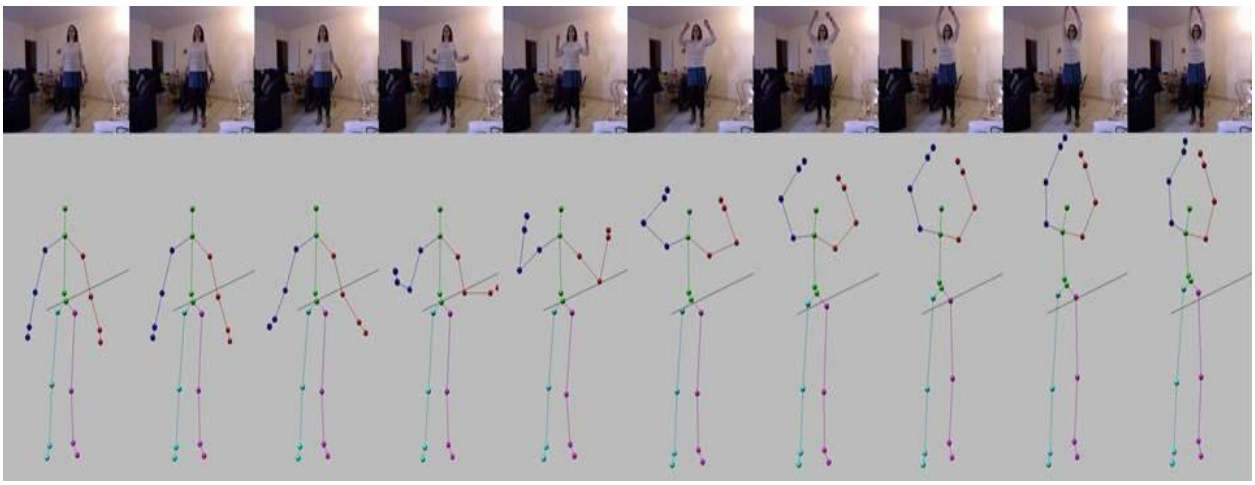


Figure V.17 Visualisation d'images et de squelettes coprorels lors de l'exécution du geste *s'étirer* par une étudiante de la classe HTI 2014-2015 lors de la constitution du corpus de gestes du même nom.

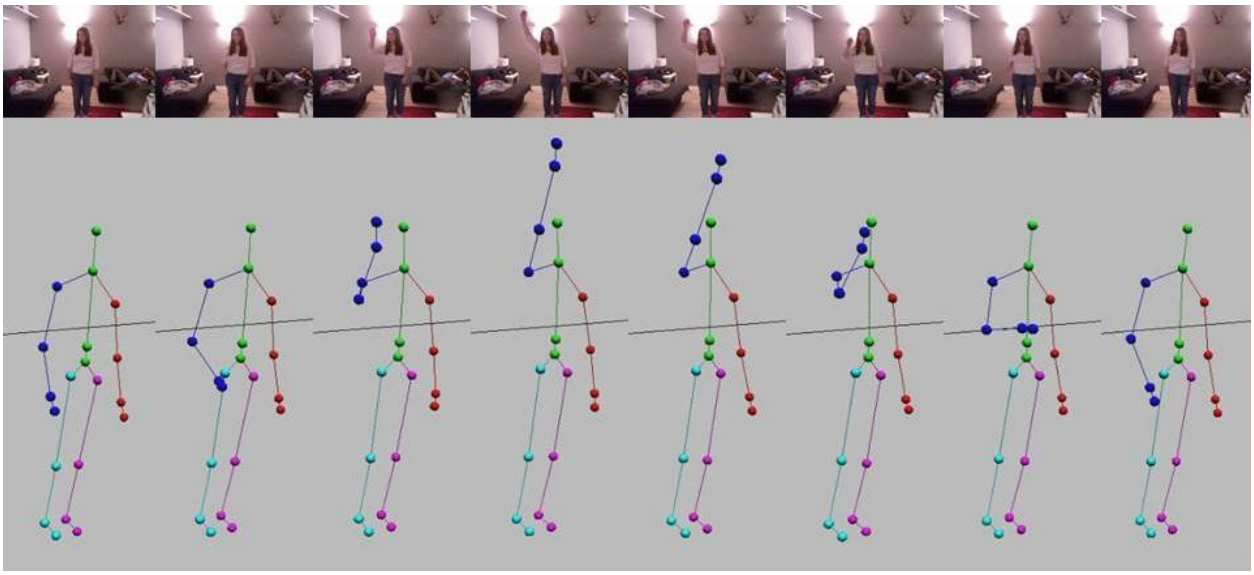


Figure V.18 Visualisation d'images et de squelettes coprorels lors de l'exécution du geste *se intercepter un objet* par une étudiante de la classe HTI 2014-2015 lors de la constitution du corpus de gestes du même nom.

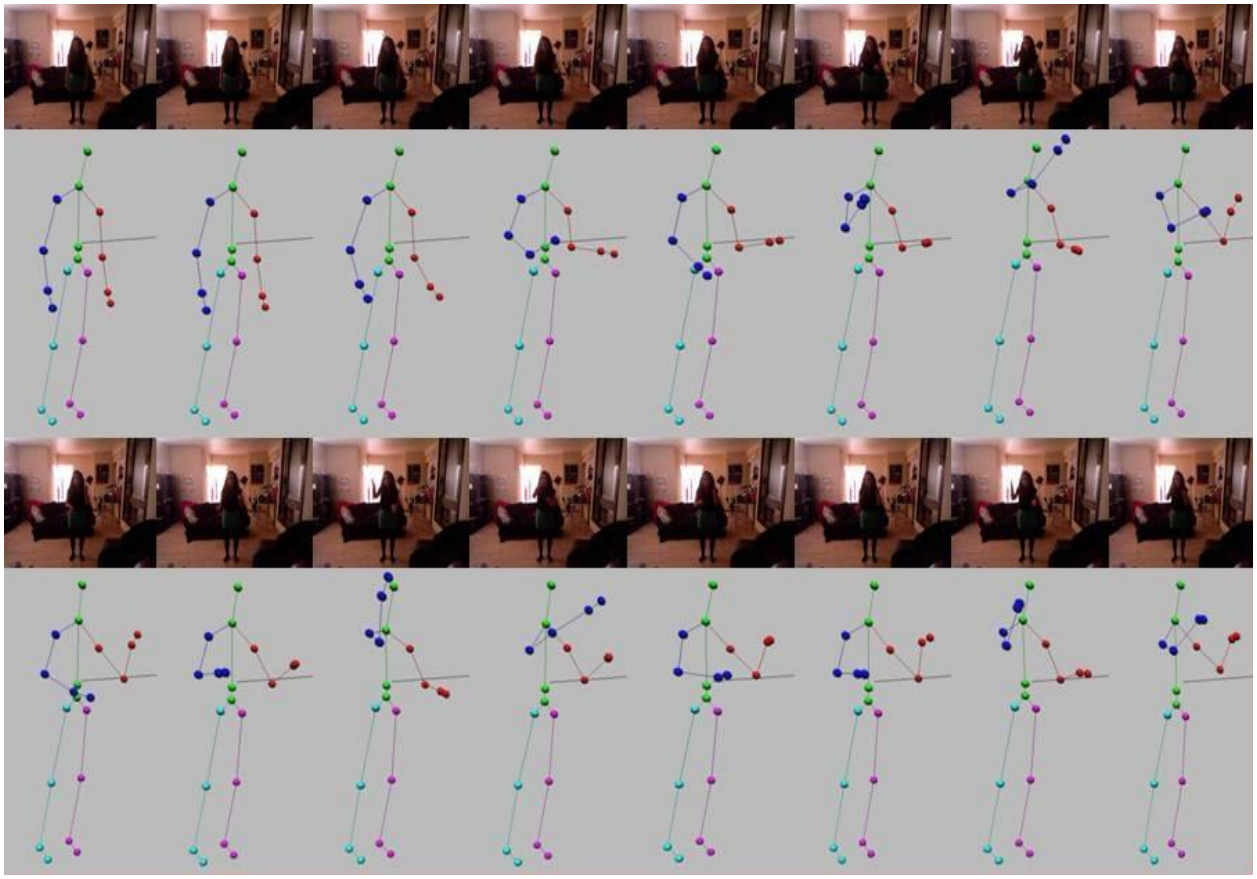


Figure V.19 Visualisation d'images et de squelettes coprorels lors de l'exécution du geste *jongler* par une étudiante de la classe HTI 2014-2015 lors de la constitution du corpus de gestes du même nom.

Nous avons demandé à deux étudiants de la majeure *High-Tech Imaging* à TELECOM SudParis pour l'année 2014-2015 (*HTI 2014-2015*) de conceptualiser un lexique d'actions courantes, en s'inspirant des instructions données dans [**36**] relativement à l'élaboration du corpus MSRC-12. Les 11 catégories de gestes résultant de cette conceptualisation sont énumérées dans le Tableau V.4. Ces actions sont

relativement variées, bien que certaines aient été volontairement choisies comme étant proches les unes des autres (exemple : *faire ses lacets* et *se mettre à genoux*, ou encore *se boucher les oreilles* et *se frotter les yeux*). Les deux étudiants ont ensuite pris en charge et supervisé la constitution d'un corpus de séquences de gestes mettant en jeu ces 11 actions. Les 11 étudiants[1] de la filière ont contribué à la constitution d'un tel corpus de séquences d'actions. Il a été demandé à chacun d'exécuter une dizaine de séquences prédéfinies de 6 gestes différents issus du lexique. Une telle prédéfinition fournit une segmentation implicite des séquences en ses différentes actions. Ce sont ces découpages en échantillons dédiés chacun à une action que nous utilisons dans l'expérience du chapitre VII comme gestes.

Le nombre d'instanciations que propose le corpus HTI 2014-2015 pour chaque classe est indiqué dans le Tableau V.4. Notons qu'au sein d'une séquence, deux actions successives sont séparées par le retour de l'individu qui exécute le mouvement à la position de repos pour un temps indéfini. Les Figure V.16, Figure V.17, Figure V.18 et Figure V.19 illustrent des images et squelettes corporels correspondant à la réalisation de quatre actions différentes parmi celles qu'énumère le Tableau V.4.

Tableau V.4 Catégories gestuelles du corpus HTI 2014-2015 avec leurs nombres d'apparitions N_{occ} dans le corpus, ainsi que les durées minimale L_{min} et maximale L_{max} exprimées en secondes.

Classe	N_{occ}	L_{min}	L_{max}
***dire merci en langage des signes* (*say "thank you" in ASL*)**	53	1.9	6.2
***faire ses lacets* (*tie shoelaces*)**	57	3.5	12.2
***faire un cercle avec le bras droit* (*draw a circle with the right arm*)**	54	1.2	6.1
***faire un tour sur soi-même* (*rotate on oneself*)**	49	2.0	5.4
***intercepter un objet* (*catch an object*)**	48	1.4	8.0
***jongler* (*juggle*)**	51	2.8	8.2
***lancer un objet devant soi* (*throw an object in front*)**	49	1.5	6.1
***se boucher les oreilles* (*cover one's ears*)**	53	2.2	8.5
***se frotter les yeux* (*rub one's eyes*)**	44	2.3	7.8
***se mettre à genoux* (*kneel*)**	54	3.9	15.4
***s'étirer* (*stretch out*)**	53	2.4	9.4

Un examen minutieux des séquences acquises nous a conduits à exclure du corpus quelques séquences aux squelettes inexploitables. Nous avons finalement obtenu 107 séquences gestuelles, composées chacune de 4 à 6 actions réalisées successivement, pour un total de 565 gestes. La durée totale du corpus est de 48 minutes et 54 secondes. Les séquences ont des durées variables allant de 16 secondes à 1 minute et 7 secondes. La durée des gestes individuels varie de 1 seconde à 15 secondes.

[1] Mentionnons que l'égalité entre le nombre de catégories d'actions retenues et le nombre d'étudiants participant à l'expérimentation, 11 dans les deux cas, est purement fortuite.

VI. Reconnaissance globale de gestes

Dans cette première expérience dite de « reconnaissance globale », le geste est décrit à l'aide d'une caractérisation mi-niveau sur l'intégralité de sa période de réalisation. Cette caractérisation prend la forme d'un descripteur global qu'il s'agit d'utiliser pour entrainer des dispositifs d'apprentissage supervisé (e.g., SVM, Random Forests), afin d'analyser des contenus haut-niveau. Après avoir introduit notre descripteur global dans un premier paragraphe, nous en évaluons les performances sur deux applications différentes. Une première concerne la reconnaissance d'actions et prend appui sur le corpus Microsoft Research Cambridge-12 (MSRC-12 [**36**]*). La deuxième vise à reconnaître les contenus émotionnels présents dans les échantillons de notre base de gestes de direction orchestrale (ORCHESTRE-3D – cf. section V.2.1).*

VI.1. Descripteurs globaux

Dans le cadre de la reconnaissance globale de gestes, nous supposons que la séquence gestuelle est disponible dans sa totalité au moment de l'analyse. L'hypothèse sous-jacente à une telle approche est le fait que chaque séquence représente un et unique geste, que l'on souhaite reconnaître à partir d'un lexique gestuel donné. Le principe consiste alors à caractériser chaque séquence à l'aide d'un ensemble de paramètres globaux, à la fois statistiques [**80**], [**162**], [**76**] et déterministes, associés dans notre cas aux séries temporelles correspondant aux différents descripteurs de Laban introduits au chapitre IV. A cela s'ajoutent quelques autres éléments descriptifs du geste, non appuyés sur des séries.

La quantification globale de la qualité de *Corps* prend appui sur la série constituée par la dissymétrie spatiale du corps à chaque instant *t* :

$$Dis(t) = \frac{d_{gauche,\,centre}(t)}{d_{gauche,\,centre}(t)+d_{droite,\,centre}(t)} \tag{VI.1}$$

La représentation globale associée consiste en 6 paramètres relatifs à cette série numérique : la moyenne (μ_{Dis}), l'écart-type (σ_{Dis}), le rapport entre maximum et moyenne ($\rho_{Dis}^{max,\mu}$), le rapport entre minimum et moyenne ($\rho_{Dis}^{min,\mu}$), le nombre d'extrema locaux ($n_{Dis}^{extrema}$) et l'instant relatif d'atteinte du maximum global (t_{Dis}^{max}). Ces valeurs sont résumées dans le Tableau VI.1.

Tableau VI.1 Paramètres globaux associés à la qualité *Corps* à partir de la série des valeurs de dissymétrie spatiale du corps $Dis(t)$.

Paramètre et formule	
$\mu_{Dis} = E\{Dis(t)\} = \frac{1}{T}\sum_{t=0}^{T-1} Dis(t)$	(VI.2)
$\sigma_{Dis} = E\{(Dis(t) - \mu_{Dis})^2\} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(Dis(t) - \mu_{Dis})^2}$	(VI.3)
$\rho_{Dis}^{min,\mu} = \frac{min_{t\in\{0,\cdots,T-1\}}\{Dis(t)\}}{\mu_{Dis}}$	(VI.4)
$\rho_{Dis}^{max,\mu} = \frac{max_{t\in\{0,\cdots,T-1\}}\{Dis(t)\}}{\mu_{Dis}}$	(VI.5)
$n_{Dis}^{extrema} = Card\{\mathcal{E}_{Dis}\}$, où $\mathcal{E}_{Dis}$ désigne l'ensemble des points d'extrema locaux de la série *Dis*(*t*)	(VI.6)
$t_{Dis}^{max} = \frac{arg\,max_{t\in\{0,\cdots,T-1\}}\{Dis(t)\}}{T}$	(VI.7)

La qualité de mouvement *Espace* est décrite à l'aide de 9 valeurs.

Nous considérons tout d'abord la longueur totale de la trajectoire du centre des hanches, noté par l_{CH} et définie comme décrit dans l'équation suivante :

$$l_{CH} = \sum_{t=1}^{T-1} \left\| P_{centre\ des\ hanches,t-1} - P_{centre\ des\ hanches,t} \right\| \quad \text{(VI.8)}$$

Ensuite, nous nous référons à la trajectoire de la tête et calculons le nombre de passages par zéro (*number of zero crossings* : $n_{avant/arrière}^{ZC}$) de la première dérivée dans la direction parallèle au plan vertical :

$$n_{avant/arrière}^{ZC} = n^{ZC}\left\{x'^{trans}_{tête,t}\right\}_{t=0}^{T-1}, \quad \text{(VI.9)}$$

afin de caractériser le mouvement vers l'avant/arrière.

Nous considérons également l'amplitude du mouvement de la tête $A_{tête}$ dans cette même direction :

$$A_{tête} = max_{t\in\{0,T-1\}}\left\{x_{tête,t}^{trans}\right\} - min_{t\in\{0,T-1\}}\left\{x_{tête,t}^{trans}\right\}, \quad \text{(VI.10)}$$

ainsi que le moment d'atteinte de son maximum relativement au geste entier :

$$t_{tête,x}^{max} = \frac{1}{T} arg\ \max_{t\in\{0,T-1\}}\left\{x_{tête,t}^{trans}\right\} \quad \text{(VI.11)}$$

Tableau VI.2 Paramètres globaux associés à la qualité de mouvement *Espace* à partir de la série des valeurs de l'angle de penchement $\Phi(t)$.

Paramètre et formule	
$\mu_\Phi = E\{\Phi(t)\} = \frac{1}{T}\sum_{t=0}^{T-1}\Phi(t)$	(VI.12)
$\sigma_\Phi = E\left\{(\Phi(t) - \mu_{\mu_\Phi})^2\right\} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(\Phi(t) - \mu_\Phi)^2}$	(VI.13)
$\rho_\Phi^{max,\mu} = \frac{max_{t\in\{0,\cdots,T-1\}}\{\Phi(t)\}}{\mu_\Phi}$	(VI.14)
$n_\Phi^{maxima} = Card\{\mathcal{M}_\Phi\}$, où $\mathcal{M}_\Phi$ désigne l'ensemble des points de maxima locaux de la série Φ (t)	(VI.15)
$t_\Phi^{max} = \frac{arg\ max_{t\in\{0,\cdots,T-1\}}\{\Phi(t)\}}{T}$	(VI.16)

Enfin, nous utilisons la séquence des angles de penchement vers l'avant $\Phi(t)$, dont nous calculons les cinq caractéristiques suivantes : moyenne (μ_Φ), écart-type (σ_Φ), rapport entre le maximum et la moyenne

($\rho_{\Phi}^{max,\mu}$), nombre de maxima locaux (n_{Φ}^{maxima}) et l'instant relatif d'atteinte du maximum global (t_{Φ}^{max}). Ces différents paramètres sont résumés dans le Tableau VI.2.

La sous-qualité de *Mise en forme* du concept de *Forme* est décrite par les séquences des amplitudes du mouvement dans les directions respectivement perpendiculaires aux plans vertical, horizontal et sagittal :

$$A^x(t) = \left(max_i\left(\left\{x_{i,t}^{trans}\right\}\right) - min_i\left(\left\{x_{i,t}^{trans}\right\}\right)\right), \quad \text{(VI.17)}$$

$$A^y(t) = \left(max_i\left(\left\{y_{i,t}^{trans}\right\}\right) - min_i\left(\left\{y_{i,t}^{trans}\right\}\right)\right), \quad \text{(VI.18)}$$

$$A^z(t) = \left(max_i\left(\left\{z_{i,t}^{trans}\right\}\right) - min_i\left(\left\{z_{i,t}^{trans}\right\}\right)\right), \quad \text{(VI.19)}$$

où *i* indexe les articulations du corps.

D'une manière similaire que pour la série $Dis(t)$, pour chacune de ces trois séquences numériques nous avons retenu les six paramètres suivants : moyenne ($\mu_{A^x}, \mu_{A^y}, \mu_{A^z}$), écart-type ($\sigma_{A^x}, \sigma_{A^y}, \sigma_{A^z}$), rapport entre maximum et moyenne ($\rho_{A^x}^{max,\mu}, \rho_{A^y}^{max,\mu}, \rho_{A^z}^{max,\mu}$), rapport en minimum et moyenne ($\rho_{A^x}^{min,\mu}, \rho_{A^y}^{min,\mu}, \rho_{A^z}^{min,\mu}$), nombre d'extrema locaux ($n_{A^x}^{extrema}, n_{A^y}^{extrema}, n_{A^z}^{extrema}$) et instant relatif d'atteinte du maximum global ($t_{A^x}^{max}, t_{A^y}^{max}, t_{A^z}^{max}$). Cela conduit à un nombre total de 18 composantes de *Mise en forme*.

Nous caractérisons ensuite la sous-composante *Temps* de la qualité *Effort*.

En premier lieu, nous retenons la durée totale du geste (T), exprimée en nombre de trames.

Ensuite, nous calculons sept valeurs qui sont toutes définies à partir de la segmentation du geste en périodes dites de faible activité (ou pauses) et périodes de moyenne ou haute activité, effectuée à partir de la séquence d'énergie cinétique $\{E_c(t)\}_{t=1}^{T}$. (*cf.* section IV.4.2) A partir de cette segmentation, plusieurs valeurs sont calculées.

Nous considérons pour cela deux séries numériques : une première série constituée de la durée des pauses (qui sont au nombre de N_{pauses}), une deuxième constituée de la durée des périodes d'activité, également appelées « cloches de mouvement » (qui sont au nombre de N_{act}) :

$$\left\{\left(T_{pause}\right)_p = durée\left(pause_p\right)\right\}_{p\in\{1,\cdots,N_{pauses}\}} \quad \text{(VI.20)}$$

$$\{(T_{act})_a = durée(période\ d'activité_a)\}_{a\in\{1,\cdots,N_{act}\}} \quad \text{(VI.21)}$$

Nous calculons premièrement la durée totale des périodes de pause relativement à la durée totale du geste :

$$\rho_{T_{pause}}^{somme,T} = \frac{somme_{p\in\{1,\cdots,N_{pauses}\}}\left\{\left(T_{pause}\right)_p\right\}}{T} \quad \text{(VI.22)}$$

Pour chacune des deux séries des durées des pauses et des périodes de mouvement (*cf.* équation VI.20, équation VI.21), nous calculons trois paramètres qui sont : la moyenne ($\mu_{T_{pause}}, \mu_{T_{act}}$), l'écart-type ($\sigma_{T_{pause}}, \sigma_{T_{act}}$) et le maximum ($max_{T_{pause}}, max_{T_{act}}$). Ces paramètres sont résumés aux Tableau VI.3 et Tableau VI.4.

Au final donc, la sous-composante de *Temps* est donc décrite avec 8 valeurs.

Tableau VI.3 Paramètres globaux associés à la sous-composante de *Temps* de la qualité d'*Effort* à partir de la série des durées des pauses.

Paramètre et formule	
$\mu_{T_{pause}} = E\left\{\left(T_{pause}\right)_p\right\} = \frac{1}{N_{pauses}} \sum_{p=1}^{N_{pauses}} \left(T_{pause}\right)_p$	(VI.23)
$\sigma_{T_{pause}} = E\left\{\left(\left(T_{pause}\right)_p - \mu_{T_{pause}}\right)^2\right\} = \sqrt{\frac{1}{N_{pauses}} \sum_{p=1}^{N_{pauses}} \left(\left(T_{pause}\right)_p - \mu_{T_{pause}}\right)^2}$	(VI.24)
$max_{T_{pause}} = max_{p\in\{1,\cdots,N_{pauses}\}}\left\{\left(T_{pause}\right)_p\right\}$	(VI.25)

Tableau VI.4 Paramètres globaux associés à la sous-composante de *Temps* de la qualité d'*Effort* à partir de la série des durées des périodes d'activité.

Paramètre et formule	
$\mu_{T_{act}} = E\{(T_{act})_a\} = \frac{1}{N_{act}} \sum_{a=1}^{N_{act}} (T_{act})_a$	(VI.26)
$\sigma_{T_{act}} = E\left\{\left((T_{act})_a - \mu_{T_{act}}\right)^2\right\} = \sqrt{\frac{1}{N_{act}} \sum_{a=1}^{N_{act}} \left((T_{act})_a - \mu_{T_{act}}\right)^2}$	(VI.27)
$max_{T_{act}} = max_{a\in\{1,\cdots,N_{act}\}}\{(T_{act})_a\}$	(VI.28)

Pour la sous-qualité de *Flux* de la qualité d'*Effort*, nous nous intéressons aux séries des modules de l'à-coup (dérivée troisième) associées aux trajectoires 3D des deux mains :

$$Acoup^{x,y,z}_{main\ gauche}(t) = \frac{d^3 OPmain\ gauche(t)}{dt^3} \quad \text{(VI.29)}$$

$$Acoup^{x,y,z}_{main\ droite}(t) = \frac{d^3 OPmain\ droite(t)}{dt^3} \quad \text{(VI.30)}$$

Cinq éléments sont retenus pour caractériser chaque série : moyenne ($\mu_{Acoup_{main\ gauche}}$, $\mu_{Acoup_{main\ droite}}$), écart-type ($\sigma_{Acoup_{main\ gauche}}$, $\sigma_{Acoup_{main\ droite}}$), rapport entre maximum et moyenne ($\rho_{Acoup_{main\ gauche}}$, $\rho_{Acoup_{main\ droite}}$), nombre de maxima locaux ($n^{maxima}_{Acoup_{main\ gauche}}$, $n^{maxima}_{Acoup_{main\ droite}}$) et instant relatif d'atteinte du maximum global ($t^{max}_{Acoup_{main\ gauche}}$, $t^{max}_{Acoup_{main\ droite}}$), ce qui conduit à un total de 10 valeurs pour la sous-qualité de *Flux*.

Enfin, la sous-qualité de *Poids* de la qualité d'*Effort* est quantifiée à partir des séries des vitesses et accélérations verticales des deux mains et du centre des hanches :

$$v^{y}_{main\ gauche}(t) = \frac{dy_{main\ gauche,t}}{dt} \tag{VI.31}$$

$$v^{y}_{main\ droite}(t) = \frac{dy_{main\ droite,t}}{dt} \tag{VI.32}$$

$$v^{y}_{centre\ des\ hanches}(t) = \frac{dy_{centre\ des\ hanches,t}}{dt} \tag{VI.33}$$

$$a^{y}_{main\ gauche}(t) = \frac{d^2y_{main\ gauche,t}}{dt^2} \tag{VI.34}$$

$$a^{y}_{main\ droite}(t) = \frac{d^2y_{main\ droite,t}}{dt^2} \tag{VI.35}$$

$$a^{y}_{centre\ des\ hanches}(t) = \frac{d^2y_{centre\ des\ hanches,t}}{dt^2} \tag{VI.36}$$

Chacune de ces six séries est représentée globalement par caractéristiques suivantes : moyenne ($\mu_{v^y_{main\ gauche}}, \mu_{v^y_{main\ droite}}, \mu_{v^y_{centre\ des\ hanches}}, \mu_{a^y_{main\ gauche}}, \mu_{a^y_{main\ droite}}, \mu_{a^y_{centre\ des\ hanches}}$), écart-type ($\sigma_{v^y_{main\ gauche}}, \sigma_{v^y_{main\ droite}}, \sigma_{v^y_{centre\ des\ hanches}}, \sigma_{a^y_{main\ gauche}}, \sigma_{a^y_{main\ droite}}, \sigma\mu_{a^y_{centre\ des\ hanches}}$), amplitude – *e.g.*, différence entre le maximum et le minimum sur toute la durée du geste – ($A_{v^y_{main\ gauche}}, A_{v^y_{main\ droite}}, A_{v^y_{centre\ des\ hanches}}, A_{a^y_{main\ gauche}}, A_{a^y_{main\ droite}}, A_{a^y_{centre\ des\ hanches}}$), nombre de minimaux locaux ($n^{minima}_{v^y_{main\ gauche}}, n^{minima}_{v^y_{main\ droite}}, n^{minima}_{v^y_{centre\ des\ hanches}}, n^{minima}_{a^y_{main\ gauche}}, n^{minima}_{a^y_{main\ droite}}, n^{minima}_{a^y_{centre\ des\ hanches}}$) et instant relatif d'atteinte du minimum ($t^{min}_{v^y_{main\ gauche}}, t^{min}_{v^y_{main\ droite}}, t^{min}_{v^y_{centre\ des\ hanches}}, t^{min}_{a^y_{main\ gauche}}, t^{min}_{a^y_{main\ droite}}, t^{min}_{a^y_{centre\ des\ hanches}}$). Cela conduit à une représentation de la qualité de *Poids* en 30 valeurs.

Le vecteur descripteur d qui résulte de cette procédure se compose de D = 81 valeurs.

$$d = (d_1, d_2, d_3, \dots, d_D) \tag{VI.37}$$

Ces différentes composantes sont résumées dans le Tableau VI.5.

Dans le paragraphe suivant, nous détaillons notre usage d'un tel descripteur global en vue de l'analyse de geste.

Tableau VI.5 Résumé des quantifications de qualités et sous-qualités de Laban donnant lieu au vecteur descripteur global d (*cf.* équation VI.37). Pour chaque qualité ou sous-qualité sont préciseés les composantes qui la quantifient, et la taille du sous-descripteur ainsi engendré.

Qualité de Laban	**Sous-qualité**	**Nombre de composantes**	**Composantes**
Corps		6	• Indice de dissymétrie $Dis(t)$: μ_{Dis}, σ_{Dis}, $\rho_{Dis}^{min,\mu}$, $\rho_{Dis}^{max,\mu}$, $n_{Dis}^{extrema}$, t_{Dis}^{max}
Espace		9	• l_{CH} • $n_{avant/arrière}^{ZC}$ • $A_{tête}$ • $t_{tête,x}^{max}$ • Angle de penchement vers l'avant $\Phi(t)$: μ_{Φ}, σ_{Φ}, $\rho_{\Phi}^{max,\mu}$, n_{Φ}^{maxima}, t_{Φ}^{max}
Forme	*Mise en forme*	18	• Amplitude corporelle selon la direction perpendiculaire au plan vertical $A^x(t)$: μ_{A^x}, σ_{A^x}, $\rho_{A^x}^{max,\mu}$, $\rho_{A^x}^{min,\mu}$, $n_{A^x}^{extrema}$, $t_{A^x}^{max}$ • Amplitude corporelle selon la direction perpendiculaire au plan horizontal $A^y(t)$: μ_{A^y}, σ_{A^y}, $\rho_{A^y}^{max,\mu}$, $\rho_{A^y}^{min,\mu}$, $n_{A^y}^{extrema}$, $t_{A^y}^{max}$ • Amplitude corporelle selon la direction perpendiculaire au plan sagittal $A^z(t)$: μ_{A^z}, σ_{A^z}, $\rho_{A^z}^{max,\mu}$, $\rho_{A^z}^{min,\mu}$, $n_{A^z}^{extrema}$, $t_{A^z}^{max}$
Effort	*Temps*	8	• Durée du geste T • Pourcentage de faible activité relativement au à la durée du geste $\rho_{T_{pause}}^{somme,T} = \frac{somme_{p\in\{1,\cdots,N_{pauses}\}}\{(T_{pause})_p\}}{T}$ • Durée des pauses : $\mu_{T_{pause}}$, $\sigma_{T_{pause}}$, $max_{T_{pause}}$ • Durée des périodes d'activité : $\mu_{T_{act}}$, $\sigma_{T_{act}}$, $max_{T_{act}}$
	Flux	10	• A-coup pour la trajectoire de la main gauche $Acoup_{main\ gauche}^{x,y,z}(t)$: $\mu_{Acoup_{main\ gauche}}$, $\sigma_{Acoup_{main\ gauche}}$, $\rho_{Acoup_{main\ gauche}}$, $n_{Acoup_{main\ gauche}}^{maxima}$, $t_{Acoup_{main\ gauche}}^{max}$ • A-coup pour la trajectoire de la main droite $Acoup_{main\ droite}^{x,y,z}(t)$: $\mu_{Acoup_{main\ droite}}$, $\sigma_{Acoup_{main\ droite}}$, $\rho_{Acoup_{main\ droite}}$, $n_{Acoup_{main\ droite}}^{maxima}$, $t_{Acoup_{main\ droite}}^{max}$
	Poids	30	• Composante verticale de la vitesse pour la trajectoire de la main gauche $v_{main\ gauche}^{y}(t)$: $\mu_{v_{main\ gauche}^{y}}$, $\sigma_{v_{main\ gauche}^{y}}$, $A_{v_{main\ gauche}^{y}}$, $n_{v_{main\ gauche}^{y}}^{minima}$, $t_{v_{main\ gauche}^{y}}^{min}$ • Composante verticale de la vitesse pour la trajectoire de la main droite $v_{main\ droite}^{y}(t)$: $\mu_{v_{main\ droite}^{y}}$, $\sigma_{v_{main\ droite}^{y}}$, $A_{v_{main\ droite}^{y}}$, $n_{v_{main\ droite}^{y}}^{minima}$, $t_{v_{main\ droite}^{y}}^{min}$ • Composante verticale de la vitesse pour la trajectoire du centre des hanches $v_{centre\ des\ hanches}^{y}(t)$: $\mu_{v_{centre\ des\ hanches}^{y}}$, $\sigma_{v_{centre\ des\ hanches}^{y}}$, $A_{v_{centre\ des\ hanches}^{y}}$, $n_{v_{centre\ des\ hanches}^{y}}^{minima}$, $t_{v_{centre\ des\ hanches}^{y}}^{min}$ • Composante verticale de l'accélération pour la trajectoire de la main gauche $a_{main\ gauche}^{y}(t)$: $\mu_{a_{main\ gauche}^{y}}$, $\sigma_{a_{main\ gauche}^{y}}$, $A_{a_{main\ gauche}^{y}}$, $n_{a_{main\ gauche}^{y}}^{minima}$, $t_{a_{main\ gauche}^{y}}^{min}$ • Composante verticale de l'accélération pour la trajectoire de la main droite $a_{main\ droite}^{y}(t)$: $\mu_{a_{main\ droite}^{y}}$, $\sigma_{a_{main\ droite}^{y}}$, $A_{a_{main\ droite}^{y}}$, $n_{a_{main\ droite}^{y}}^{minima}$, $t_{a_{main\ droite}^{y}}^{min}$ • Composante verticale de l'accélération pour la trajectoire du centre des hanches $a_{centre\ des\ hanches}^{y}(t)$: $\mu_{a_{centre\ des\ hanches}^{y}}$, $\sigma_{a_{centre\ des\ hanches}^{y}}$, $A_{a_{centre\ des\ hanches}^{y}}$, $n_{a_{centre\ des\ hanches}^{y}}^{minima}$, $t_{a_{centre\ des\ hanches}^{y}}^{min}$

VI.2. Reconnaissance d'actions

VI.2.1. Utilisation du corpus Microsoft Research Cambridge-12

Notre première expérimentation d'analyse globale du geste est dédiée à la reconnaissance d'actions spécifiques. Elle prend appui sur le corpus Microsoft Research Cambridge-12 (MSRC-12) (*cf.* section V.1), qui est tout à fait adapté pour ce type d'expérimentation de la reconnaissance globale.

Afin d'illustrer le caractère discriminant des descripteurs retenus, le Tableau VI.6 illustre les valeurs (moyennes et écarts-types sur l'ensemble des actions considérées) que prennent 4 composantes de notre descripteur global du geste, pour 3 actions de MSRC-12 dataset : *lancer la musique/monter le volume* (A1), *s'incliner pour clore la session musicale* (A2) et *changer d'arme* (A3). Les 4 caractéristiques retenues sont les suivantes :

- μ_{Dis} : moyenne de la série des dissymétries spatiales, quantifiant la qualité de *Corps* (*cf.* équation VI.2) ;
- μ_{A^y} : moyenne de la série des amplitudes corporelles dans la direction perpendiculaire au plan horizontal (caractérisant l'élévation/abaissement), quantifiant la sous-composante de *Mise en Forme* de la qualité de *Forme* (*cf.* équation VI.18) ;
- $A_{tête}$: amplitude du mouvement de la tête dans la direction perpendiculaire au plan vertical, (caractérisant le mouvement vers l'avant/arrière), quantifiant la qualité d'*Espace* (*cf.* équation VI.10) ;
- $\mu_{T_{act}}$: moyenne de la durée des périodes de moyenne ou forte d'activité, quantifiant la sous-composante de *Temps* de la qualité d'*Effort* (*cf.* équation VI.26).

Tableau VI.6 Moyennes et écarts-types basés sur les valeurs prises par 4 composantes de notre descripteur global, pour 3 actions différentes du MSRC-12 dataset.

Action	***lancer la musique/monter le volume* (A1)**	***s'incliner pour clore la session musicale* (A2)**	***changer d'arme* (A3)**
μ_{Dis} : moyenne de la série des dissymétries spatiales			
Moyenne	0.498	0.515	0.555
écart-type	0.008	0.045	0.039
μ_{A^y} : moyenne de la série des amplitudes corporelles dans la direction perpendiculaire au plan horizontal			
Moyenne	0.953	0.675	0.976
écart-type	0.027	0.101	0.027
$A_{tête}$: amplitude du mouvement de la tête dans la direction perpendiculaire au plan vertical			
Moyenne	0.048	0.157	0.089
écart-type	0.067	0.084	0,050
$\mu_{T_{act}}$: moyenne de la durée des périodes de moyenne ou forte d'activité			
Moyenne	0.599	0.448	0.323
écart-type	0.190	0.249	0.108

Dans le cas de μ_{Dis}, le calcul de la moyenne et de l'écart-type pour les 3 actions montre que cette caractéristique permet de distinguer aisément les gestes A2 et A3, ainsi que de différencier les gestes A1 et A3. En revanche, le seul paramètre μ_{Dis} ne suffit pas à distinguer les actions A1 et A2, car seul un

faible écart de 0.017 mètres sépare les moyennes des μ_{Dis} pour ces deux classes d'actions. Pour ce qui est de la composante $A_{tête}$, des considérations similaires font apparaître un unique problème de distinction entre les actions A1 et A3. La discrimination basée sur la composante μ_{Ay} rend difficile la distinction entre A1 et A3, car les moyennes et écarts-types sont très proches dans le cas de ces deux actions. Au contraire, la grande distance avec la valeur moyenne pour l'action A2 semble faciliter la reconnaissance de cette dernière parmi les occurrences des actions A1 et A3. Enfin, l'étude de la moyenne et de la dispersion des valeurs prises par la composante $\mu_{T_{act}}$ montre que seule une distinction entre les actions A1 et A3 peut être établie à l'aide de cette dernière.

Ces exemples montrent que l'usage de combinaisons d'indices de l'expressivité peut nous aider à reconnaître différentes actions présentes dans notre corpus gestuel.

VI.2.2. Méthodes de classification

Nous avons comparé les performances de deux différentes méthodes de classification. Dans les deux cas, nous avons utilisé l'implantation disponible dans la boîte à outils python « *scikit-learn* » [**166**].

La première utilise les ***Support Vector Classifiers*** (*SVC* [**104**]) avec la stratégie *one versus one* [**167**]. Cette stratégie d'apprentissage multi-classes entraine un classifieur SVM pour chaque paire de classes. Lors de l'étape de test, la classe qui collecte le score le plus haut, *i.e.*, le plus grand nombre de votes de classifieur, est retenue.

La librairie *scikit-learn* propose différents paramètres d'optimisation pour les SVC. Ci-dessous, nous les énumérons, ainsi que les valeurs que nous avons testées pour chacun d'eux lors des expériences :

- paramètre de pénalité *C* du terme d'erreur : 10^{-2}, 10^{0}, 10^{2}, 10^{3} ;
- type de noyau : polynomial, gaussien (« *rbf* »), linéaire, sigmoïdal ;
- degré du noyau : 10^{1}, 10^{2}, 10^{3}, 10^{4} ;
- coefficient γ du noyau (quand nécessaire) : 10^{-5}, 10^{-2}, 10^{1}.

Le Tableau VI.8 précise les paramètres avec lesquels nous avons obtenu les meilleurs résultats.

Le second classifieur utilise un type de forêts aléatoires nommé ***Extremely Randomized Trees*** (*Extra Trees*) [**102**] [**168**] qui consiste en une construction de plusieurs arbres aléatoires entrainés indépendamment, dont le principe est rappelé ci-dessous. Conformément au principe général de construction d'arbres de décision, l'arborescence des Extra Trees est construite récursivement par divisions successives de l'ensemble des échantillons d'entrainement. Ainsi, à partir de la racine (à laquelle « appartient » l'intégralité des échantillons d'entrainement), les « nœuds » se subdivisent progressivement en nœuds enfants, jusqu'à ce que l'ensemble des nœuds soient déclarés « terminaux » (ou « feuilles »). L'arbre qui en résulte est binaire, chaque nœud pouvant être décomposé en deux nœuds enfants, appelés par convention *gauche* et *droit*.

A chaque nœud – comprenant un certain nombre d'échantillons d'apprentissage –, la procédure suivante est appliquée.

- Si la profondeur maximale est atteinte, alors le nœud est déclaré comme étant une feuille.
- Sinon, un ensemble de tests est effectué. Chacun de ces tests identiques a lieu comme suit :
    - Un indice de composante i_{rand} du vecteur caractéristique qui sert d'entrée (dans notre cas, notre descripteur global du geste, *cf.* équation VI.37, section VI.1) est choisi aléatoirement. Un seuil $v_{i_{rand}}$ associé à la i_{rand}-ième composante est également

déterminé aléatoirement sur une plage de variation $[d_{i_{rand}}^{min}, d_{i_{rand}}^{max}]$ où $d_{i_{rand}}^{min}$ et $d_{i_{rand}}^{max}$ correspondent respectivement aux valeurs minimale et maximale prises par la composante $d_{i_{rand}}$ sur l'ensemble du corpus.

- Les échantillons pour lesquels la valeur de la composante est supérieure au seuil sont attribués au nœud enfant gauche, tandis que ceux pour lesquels cette valeur lui est inférieure sont attribués au nœud enfant droit. De cette manière, la distribution des échantillons pour le nœud considéré est donc séparée en deux.
- Un score d'appréciation de la qualité de la division est estimé. Son rôle est de mesurer l'homogénéité des deux distributions résultantes (les critères possibles sont introduits plus loin) pour le test.

- La séparation retenue est celle qui correspond au meilleur score d'appréciation. Les deux sous-ensembles d'échantillons ainsi définis sont attribués à deux nœuds fils : gauche et droite. Pour chacun de ces nœuds, la procédure de séparation avec test d'hétérogénéité est appelée récursivement à deux conditions :
  - le sous-ensemble d'échantillons attribué doit être suffisamment hétérogène du point de vue de la distribution des classes (un seuil minimal d'hétérogénéité est ici utilisé),
  - le sous-ensemble doit contenir suffisamment d'échantillons (le nombre minimal toléré d'échantillons étant un paramètre à définir).

  Autrement le nœud fils est déclaré feuille.

Ainsi, lorsqu'un nœud est déclaré feuille, l'ensemble des échantillons qui lui ont été attribués définissent une distribution en termes de classes. Cette distribution peut prendre la forme d'un histogramme où chaque classe est représentée par une barre dénotant le nombre d'échantillons qui lui appartiennent.

Le nombre d'arbres aléatoires de la forêt ayant été choisi au préalable, chaque arbre est construit (*e.g.*, entrainé) selon le processus explicité ci-dessus. Une fois la forêt construite, l'étape de reconnaissance peut avoir lieu.

Lors de cette étape de reconnaissance (*i.e.*, phase de test), chaque échantillon à classifier parcourt chacun des arbres de la forêt, de nœud en nœud à partir de la racine. A chaque nœud parcouru, le test de séparation qui a été défini pour celui-ci sert à établir si l'échantillon sera attribué au nœud fils de gauche ou au nœud fils de droite. Lorsque l'échantillon arrive sur une feuille, les labels qui lui sont attribués correspondent à la distribution de classes constituée par les échantillons d'entrainement affectés à ladite feuille lors de la construction de l'arbre.

Le vecteur descripteur est traité par tous les arbres de la forêt, de telle sorte qu'il se voit affecter une distribution par arbre, sous forme d'un histogramme de classes. La somme de ces histogrammes fournit un histogramme de classes global, dont on retient comme classe effectivement reconnue celle qui collecte le plus grand nombre de votes. La Figure VI.1 permet de visualiser le traitement du descripteur d (*cf.* équation VI.37) de l'échantillon de test lorsque celui–ci est placé en entrée de la forêt pour que sa classe d'appartenance soit déterminée.

Comme pour les SVC, les valeurs de différents paramètres doivent être choisies :

- le nombre d'arbres de la forêt : 10^1, 10^2, 10^3 ;
- le critère d'appréciation de la qualité de la division pour chacun des deux ensembles résultant du test de nœud :
  - entropie ;

- indice de diversité de Gini ;

- le nombre de caractéristiques parmi les N valeurs de descripteur (*cf.* équation VI.37) à considérer lors du test consistant à trouver la meilleure séparation à un nœud :
  - N ;
  - $\sqrt{N}$;
  - $\log_2(N)$.

Comme dans le cas des SVC, le Tableau VI.9 précise plus bas les paramètres de forêt d'arbres aléatoires avec lesquels nous avons obtenu les meilleurs résultats.

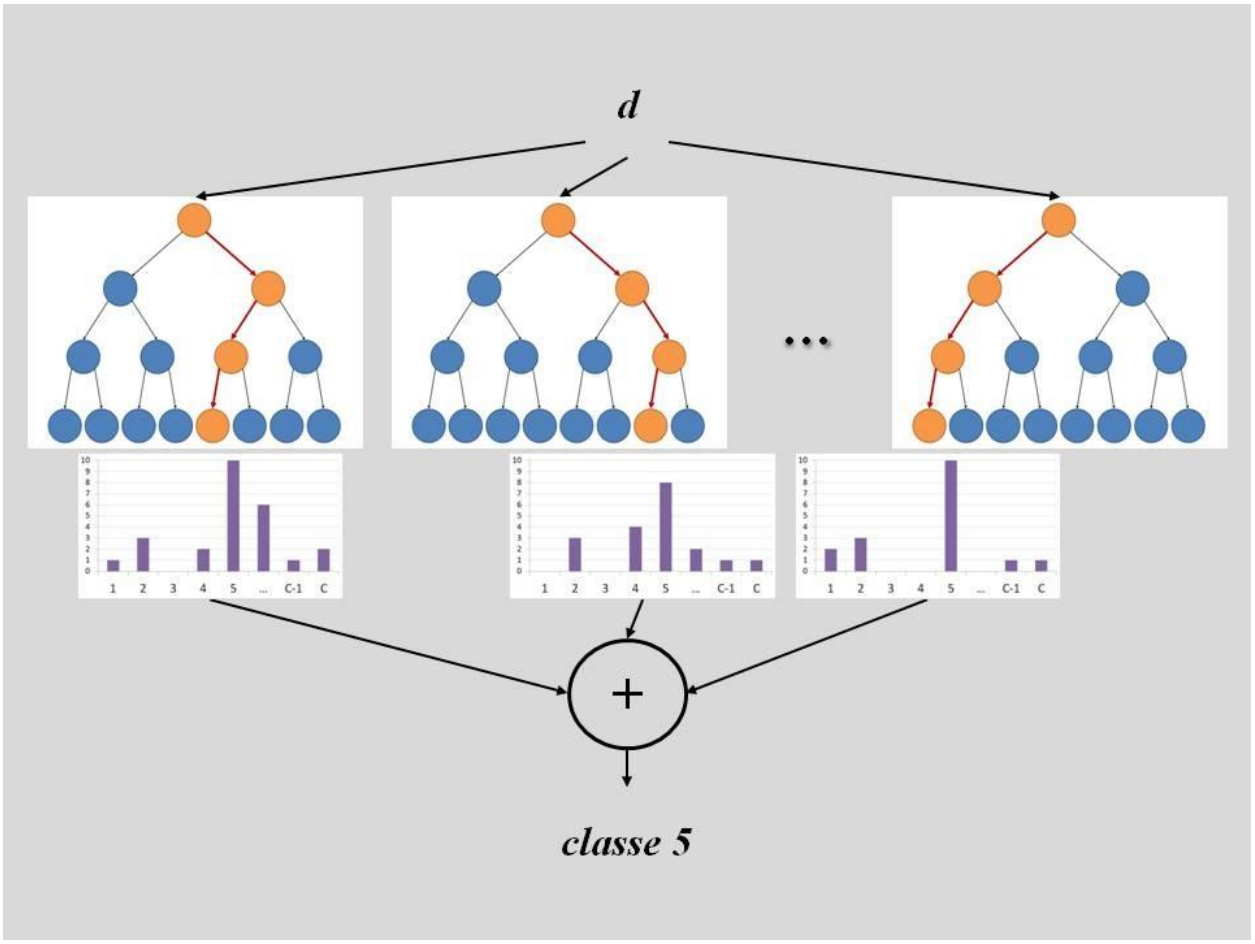


Figure VI.1 Illustration de la procédure de reconnaissance de classe dans le cas d'une forêt d'arbres aléatoires. Dans chacun des arbres contruits durant l'étape d'entrainement, le descripteur d (*cf.* équation VI.37) de l'échantillon à classer effectue un parcours (en orange) jusqu'à une feuille où lui est assigné un histogramme de labels. Sur l'histogramme résultant (*e.g.*, somme des histogrammes renvoyés par chacun des arbres), la classe qui obtient le score maximal est décrétée classe de l'échantillon.

VI.2.3. Protocole d'évaluation

Nous avons évalué la méthode de reconnaissance proposée pour les trois ensembles de gestes envisageables sur le corpus MSRC-12 :

1. l'ensemble des gestes iconiques,
2. l'ensemble des gestes métaphoriques,
3. l'intégralité du corpus MSRC-12.

Dans chacun de ces trois cas, nous avons utilisé un schéma classique de « validation croisée » (*cross validation* dans la littérature anglophone [**169**]) en cinq étapes, avec un rapport quantitatif entre données d'entrainement et données de test de 80%/20%. Cette procédure de validation croisée a été appliquée à partir d'une division des données gestuelles concernées (*i.e.*, gestes iconiques seuls, gestes métaphoriques seuls, entièreté du corpus) en cinq blocs préservant au mieux la distribution initiale des classes sur l'ensemble du corpus. De cette manière, les classes sont représentées de la même manière dans chacun des blocs.

Nous utilisons le F-score [**170**], qui hybride les critères habituels de précision et de rappel en un score unique, comme mesure de la performance de notre approche :

$$F - score = 2\frac{precision\ .rappel}{precision+rappel}, \tag{VI.38}$$

où $precision$ désigne le pourcentage d'échantillons vrais positifs en rapport avec la totalité des échantillons classés comme positifs, et $rappel$ le pourcentage d'échantillons justement classés comme positifs parmi les échantillons qui le sont effectivement.

$$precision = \frac{nombre\ de\ vrais\ positifs}{nombre\ de\ vrais\ positifs+nombre\ de\ faux\ positifs} \tag{VI.39}$$

$$rappel = \frac{nombre\ de\ vrais\ positifs}{nombre\ de\ vrais\ positifs+nombre\ de\ faux\ négatifs} \tag{VI.40}$$

VI.2.4. Résultats expérimentaux

Les résultats pour les deux sous-ensembles de gestes iconiques et métaphoriques, et pour le corpus entier sont reportés respectivement dans la Figure VI.2, la Figure VI.3 et la Figure VI.4. Les F-scores moyens par sous-ensemble sont résumés dans le Tableau VI.7.

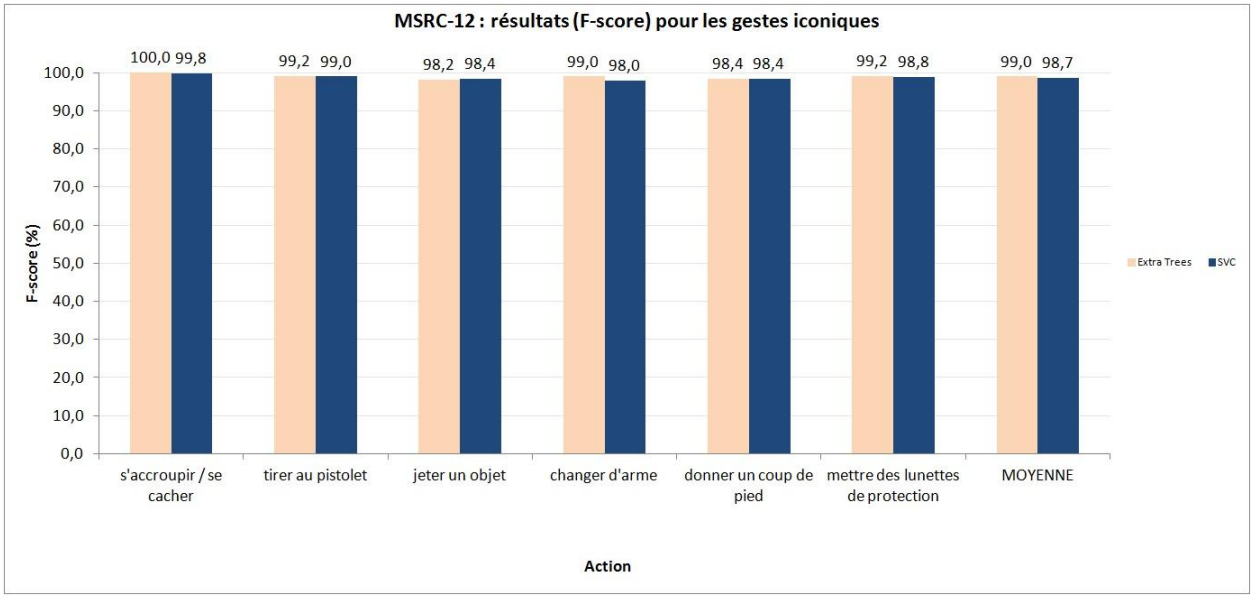


Figure VI.2 F-scores (en %) obtenus par classe pour les différentes méthodes de classification retenues, pour le sous-ensemble des gestes iconiques du corpus MSRC-12.

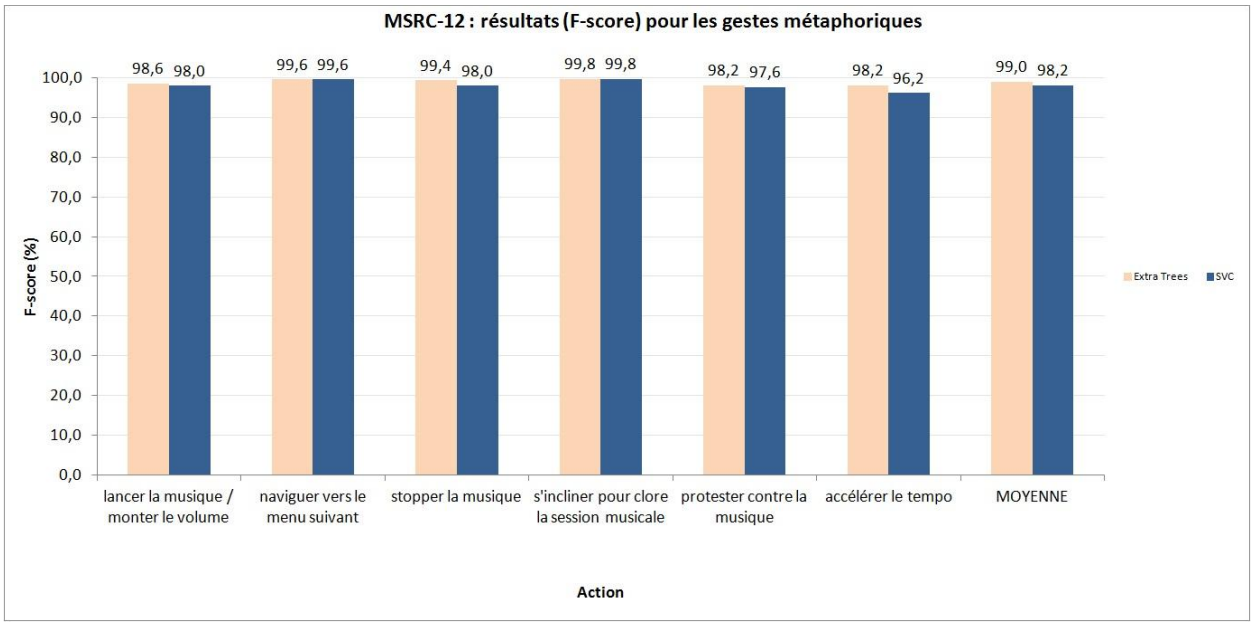


Figure VI.3 F-scores (en %) obtenus par classe pour les différentes méthodes de classification retenues pour le sous-ensemble des gestes métaphoriques du corpus MSRC-12.

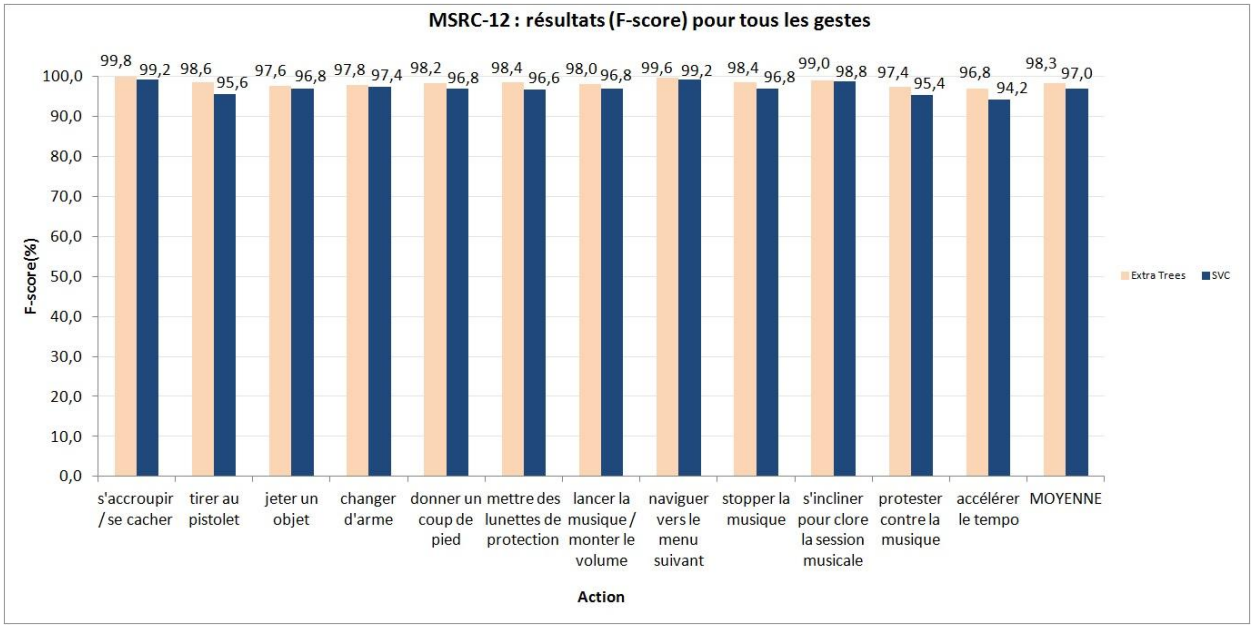


Figure VI.4 F-scores (en %) obtenus par classe pour les différentes méthodes de classification retenues, pour l'intégralité du corpus MSRC-12.

Tableau VI.7 F-scores moyens (en %) pour les ensembles de gestes concernés du corpus MSRC-12.

Gestes concernés	**Gestes iconiques**	**Gestes métaphoriques**	**Tous les gestes**
Résultats Extra Trees	99.0	99.0	98.3
Résultats SVC	98.7	98.2	97

Ces résultats correspondent aux meilleurs scores obtenus à travers différentes combinaisons de paramètres possibles propres à chaque stratégie de classification (*e.g.*, SVC ou Extra Trees). Les

paramètres es utilisés sont résumés dans les Tableau VI.8 et Tableau VI.9, pour les SVC et les Extra Trees, respectivement.

Le F-score moyen obtenu est dans tous les cas supérieur à 97%, quelle que soit la stratégie de classification mise en œuvre.

Les Extra Trees tendent à avoir un niveau de performance légèrement supérieur à celui des SVC, avec un gain de F-score d'environ 1%.

Les taux de reconnaissance sont légèrement supérieurs pour les gestes iconiques que pour les gestes métaphoriques dans le cas de l'utilisation des SVC (ce qui paraît être cohérent avec le regain prévisible de difficultés à caractériser des gestes à mesure qu'ils sont abstraits, et donc réalisés de manière plus variée que des gestes iconiques).

Lorsque l'intégralité des gestes (iconiques et métaphoriques) est utilisée, les performances se dégradent très légèrement (avec des diminutions de F-score respectives de 1.4% et 0.7% pour les SVC et les Extra Trees).

Dans tous les cas, les scores dépassent très largement ceux qui étaient obtenus dans [**130**], où les meilleurs F-scores atteints séparément pour les sous-ensembles de gestes iconiques et de gestes métaphoriques sont respectivement de 95% et 81%. Ils dépassent également l'approche de [**131**] où le niveau moyen de précision atteint sur l'ensemble du corpus ne dépasse pas 93%, ou encore l'approche de [**135**] où cette précision moyenne est de 94.6%. Enfin, ils montrent une supériorité en comparaison avec l'approche introduite dans [**138**], où les résultats en termes de F-score moyen n'excèdent pas 73%, et sont présentés par sous-ensemble de gestes, en fonction du type de consigne ou d'amorce utilisé pour indiquer aux participants les actions à réaliser lors de la constitution de la base de données. Cela démontre la pertinence de l'approche proposée et la capacité des descripteurs LMA à capturer efficacement les caractéristiques déterminantes du mouvement corporel, qu'il soit abstrait ou métaphorique.

Tableau VI.8 Combinaison de paramètres pour les SVC, en fonction de l'ensemble de gestes d'intérêt dans l'expérience de reconnaissance d'action du MSRC-12 dataset.

Gestes concernés	**Gestes iconiques**	**Gestes métaphoriques**	**Toutes les gestes**
C	0.01	0.01	1000
Type de noyau	Polynomial	polynomial	Polynomial
Degré du noyau	4	3	3
Coefficient γ	10.0	10.0	0.01

Tableau VI.9 Combinaison optimale de paramètres pour les Extra Trees, en fonction de l'ensemble de gestes d'intérêt dans l'expérience de reconnaissance d'action du MSRC-12 dataset.

Gestes concernés	**Gestes iconiques**	**Gestes métaphoriques**	**Toutes les gestes**
Nombre d'arbres	1000	1000	1000
Critère d'appréciation de la qualité de la division	Entropie	entropie	indice de diversité de Gini
Nombre de caractéristiques parmi les N valeurs de descripteur à considérer lors du test	$\sqrt{N}$	$\sqrt{N}$	N

Ces différents résultats démontrent que l'approche proposée est tout à fait adaptée pour des objectifs de reconnaissance d'actions. Examinons à présent dans quelle mesure notre caractérisation globale et expressive du geste peut être utilisée en vue d'une analyse de plus haut niveau, notamment émotionnelle, dans l'optique des travaux présentés dans [**146**] [**147**] [**149**] [**163**] [**162**].

VI.3. Analyse émotionnelle

Dans ce cadre, nous avons utilisé notre corpus de gestes pré-segmentés de chefs d'orchestre, annoté en termes d'émotions musicales (*cf.* section V.2.1). Nous avons donc couplé nos descripteurs expressifs à des méthodes d'apprentissage supervisé pour tenter d'inférer l'espace multi-émotionnel composé de nos catégories d'émotions musicales à partir de la description expressive du geste.

VI.3.1. Méthodes de classification et protocole d'évaluation

Trois différentes méthodes d'apprentissage supervisé ont été retenues pour tester la capacité de nos descripteurs à reconnaître les émotions exprimées par les gestes de chefs d'orchestre :

- une méthode fondée sur les ***Extremely Randomized Trees*** (*Extra Trees*) [**168**] ;
- une implémentation des ***Support Vector Classifiers*** (*SVC*) [**104**] avec la stratégie *one versus one* ;
- une dernière technique à base de ***One Class SVM*** [**171**].

Cette dernière technique consiste, pour une classe donnée, à créer un modèle des échantillons d'entrainement qui la représentent, c'est-à-dire à définir une « frontière douce » les délimitant, de sorte à être capable de classifier les nouveaux points trop éloignés de cette frontière comme non-représentants de la classe. Les données d'entrainement sont séparées de l'origine (dans l'espace caractéristique) et la distance entre l'hyperplan et l'origine est maximisée, de telle sorte que la frontière capture les régions dans l'espace de départ (*e.g.*, l'espace des descripteurs) où résident les données.

Pour chacune de ces méthodes, nous avons également utilisé des éléments disponibles de la boîte à outils « *scikit-learn* » [**166**].

Une stratégie d'optimisation des paramètres est suivie sur le même modèle que pour l'expérience précédente. Nous avons déjà introduit les paramètres d'optimisation pour les Extra Trees et les SVC (*cf.* section VI.2.2). Pour la technique *One Class SVM*, les paramètres utilisés sont les mêmes que pour les SVC. A ces valeurs s'ajoute ici paramètre ν qui correspond à une borne supérieure sur la fraction des erreurs d'entrainement (*e.g.*, les données d'entrainement qui sont considérées comme n'appartenant pas à la classe) et à une borne inférieure sur la fraction des données d'entrainement qui constituent le vecteur support, et qui prend les valeurs suivantes : 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}. Les Tableau VI.10, Tableau VI.11 et Tableau VI.12 fournissent les jeux de paramètres optimaux (*i.e.*, pour lesquels les meilleurs scores ont été atteints) respectivement pour les Extra Trees, les SVC et les One Class SVM.

Comme pour l'expérience précédente dédiée à la reconnaissance d'actions, nous avons utilisé un schéma classique de validation croisée en cinq étapes, avec un rapport de données d'entrainement/test égal à 80%/20%. Pour cela nous avons effectué une division du corpus en cinq blocs de manière à préserver au maximum la distribution initiale des classes. Comme mesure de performance, ici encore nous avons retenu le F-score (*cf.* équation VI.38).

Notons qu'au contraire de l'expérience précédente, nous avons décidé de construire un classificateur par catégorie (*e.g.*, pour chacune des 9 émotions), quelle que soit la méthode de classification retenue.

Chaque classe est alors considérée indépendamment des autres, de sorte à traiter convenablement le problème de l'étiquetage multiple sur un corpus aussi réduit que le nôtre (882 gestes annotés). En effet, la petite taille de notre base de données aurait rendu totalement illégitime toute conclusion relative à un « mélange » d'émotions. Chaque classifieur renvoie donc comme résultat l'appartenance ou non de l'échantillon de test à la classe à laquelle il est dédié. L'appartenance de l'échantillon aux différentes classes nous est ainsi donnée par l'ensemble des résultats binaires fournis par tous ces classifieurs pris séparément.

VI.3.2. Résultats expérimentaux

Les résultats obtenus sont présentés Figure VI.5. Pour chaque émotion et chaque méthode, le résultat affiché correspond au meilleur score, obtenu pour une combinaison de paramètres propre à chacune des trois méthodes d'apprentissage. Les paramètres d'optimisation sont présentés dans le Tableau VI.10, le Tableau VI.11 et le Tableau VI.12, respectivement, pour les Extra Trees, les SVC et les One Class SVM.

L'approche SVC donne des résultats supérieurs à ceux des autres méthodes, avec un F-score moyen égal à 57.0%, contre 52.8% pour les Extra Trees et seulement 49.9% pour les One Class SVM. La méthode SVC conduit également aux meilleurs résultats par classe, sauf dans le cas des catégories *éveillé*, *magique*, *facile*.

Globalement, les résultats des SVC excèdent ceux des Extra Trees de 4.1%, avec des écarts de scores pour les émotions *tendu*, *serein* et *colérique* supérieurs à 5% et conduisant à un gain de jusqu'à 12% pour la catégorie *magique*.

Les SVC surpassent les One Class SVM, avec une différence de F-score moyen d'environ 7%. Cette différence est encore plus prononcée (plus de 11%) pour les émotions *calme*, *agité* et *serein*, et atteint même 20.4% pour l'émotion *colérique*.

Les différences de performance sont plus relatives lorsqu'on compare les performances des approches *Extra Trees One Class SVM*, avec un gain global de 3.0% pour les *Extra Trees*. Les différences les plus fortes sont obtenues pour les classes *calme*, *agité* et *colérique*, avec des écarts respectifs de 11.6%, 10.5% et 11.9%. Néanmoins, les résultats des *One Class SVM* sont supérieurs de plus de 4% à ceux des *Extra Trees* pour les émotions *tendu* et *facile*. Le meilleur gain en termes de F-score pour l'approche *One Class SVM* est obtenu pour la catégorie *magique* (14.1%).

Les résultats obtenus à l'aide des *One Class SVM* sont globalement les moins bons, mis à part pour les catégories *tendu*, *magique* et *facile*. On notera que ce sont ces *One Class SVM* qui donnent le meilleur F-score pour la catégorie *magique*, qui est néanmoins la classe la moins représentée (*e.g.*, la moins choisie lors de l'annotation).

L'approche *Extra Trees* donne le meilleur résultat pour la seule classe *éveillé*. De manière générale, cela montre la meilleure efficacité des méthodes à base de SVM sur des problèmes de classification où le nombre d'échantillons est relativement faible.

Les résultats obtenus dans cette expérience dédiée à l'analyse de gestes expressifs haut-niveau sont, sans surprise, significativement inférieurs à ceux obtenus dans le cas des gestes représentant des actions spécifiques (*cf.* section VI.2). Cela s'explique sans doute par la multiplicité des expressions possibles de chacune des émotions, multiplicité qui contraste avec la plus forte homogénéité unissant les différentes instanciations de catégories d'actions prédéfinies.

L'élaboration de notre modèle du geste expressif LMA tente d'apporter notamment une réponse à la problématique de variabilité intra-classe posée par l'expression d'émotions. Les résultats obtenus sont particulièrement pertinents pour les émotions suivantes : *calme*, *agité*, *éveillé*, *serein* et *facile* (avec F-scores de 76.2, 59.1, 79.1, 66.7 et 60.8%, respectivement). Soulignons également que, pour presque la moitié des émotions, au moins une des méthodes considérée conduit à un F-score supérieur à 60%, avec une performance atteignant presque les 80% pour l'émotion *éveillé*.

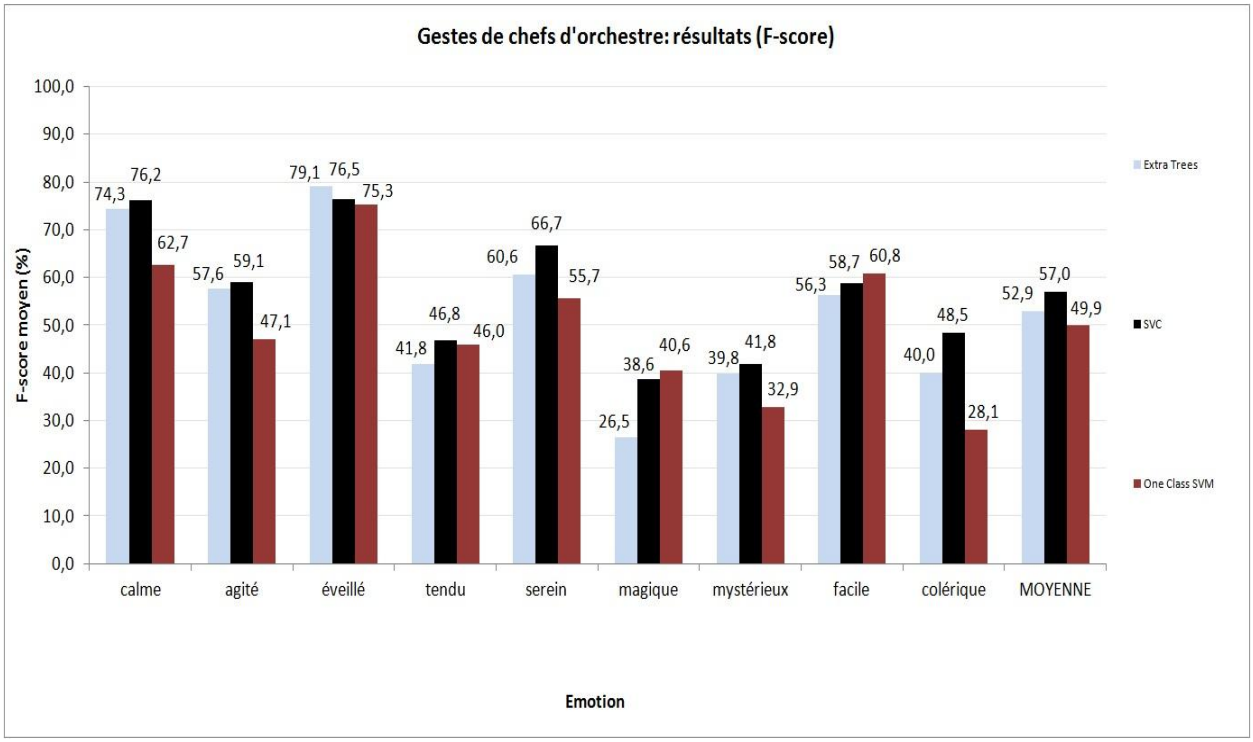


Figure VI.5 F-scores (en %) obtenus par classe pour les différentes méthodes de classification retenues, et pour le corpus des gestes de direction orchestrale.

Tableau VI.10 Combinaison optimale de paramètres pour les Extra Trees dans l'expérience de reconnaissance d'émotions basée sur le corpus des gestes de direction orchestrale.

Paramètre	**Nombre d'arbres**	**Critère d'appréciation de la qualité de la division**	**Nombre de caractéristiques parmi les N valeurs de descripteur à considérer lors du test**
calme	1000	entropie	N
agité	100	indice de diversité de Gini	N
éveillé	100	indice de diversité de Gini	$\sqrt{N}$
tendu	1000	entropie	N
serein	1000	entropie	N
magique	10	entropie	N
mystérieux	1000	indice de diversité de Gini	N
facile	100	indice de diversité de Gini	$\sqrt{N}$
colérique	1000	entropie	N

Tableau VI.11 Combinaison optimale de paramètres pour les SVC dans l'expérience de reconnaissance d'émotions basée sur le corpus des gestes de direction orchestrale.

Paramètre	**C**	**Type de noyau**	**Degré du noyau**	**Coefficient γ**
calme	1	linéaire	3	0
agité	100	polynomial	1	10
éveillé	100	polynomial	4	0.01
tendu	100	gaussien	3	0.01
serein	1000	polynomial	1	0.01
magique	100	gaussien	3	0.01
mystérieux	100	sigmoïdal	3	0.01
facile	100	polynomial	2	0.01
colérique	1	polynomial	1	10

Tableau VI.12 Combinaison optimale de paramètres pour les One Class SVM dans l'expérience de reconnaissance d'émotions basée sur le corpus des gestes de direction orchestrale.

Paramètre	**Type de noyau**	**Coefficient γ**	**Coefficient ν**
calme	sigmoïdal	10	0.1
agité	gaussien	0.00001	0.1
éveillé	gaussien	0.01	0.0001
tendu	Gaussien	0.01	0.01
serein	Sigmoïdal	10	0.1
magique	Gaussien	0.01	0.01
mystérieux	Sigmoïdal	10	0.1
facile	Gaussien	0.01	0.01
colérique	Gaussien	0.01	0.0001

Le Tableau VI.13 présente les résultats de quelques autres méthodes d'analyse des émotions basées sur l'étude du mouvement corporel. Si la plupart de ces approches ont abouti à des résultats bien supérieurs aux nôtres, il nous paraît nécessaire de rappeler certains éléments cruciaux.

- La plupart des cas présentés dans le tableau mettent en jeu 4 à 6 émotions à reconnaître, c'est-à-dire un nombre inférieur aux 9 émotions musicales dont nous nous sommes servis.
- Par ailleurs, il faut noter qu'au contraire des corpus qui ont pu servir aux expériences décrites dans le tableau, et plus généralement à la différence de la majorité des bases de données gestuelles de l'état de l'art présentées aux sections III.1.1 (corpus 2D) et V.1 (corpus 3D), nos gestes de direction orchestrale avaient un caractère *spontané*, dans la mesure où ils avaient été enregistrés en conditions réelles. La variabilité des profils gestuels au sein de chacune de nos classes d'émotions ne pouvait que s'en retrouver décuplée, et ainsi rendre d'autant plus difficile la classification.
- Nous aurions sans doute pu obtenir de meilleurs résultats à condition de disposer d'un corpus beaucoup plus grand, offrant de nombreuses instanciations de chacune de nos émotions musicales – on rappellera qu'une base de 882 gestes annotée à l'aide de 9 émotions est en fait relativement petite à l'égard de l'étude que nous avons souhaité mener.
- Enfin, nous rappellerons le caractère tout à fait singulier de notre hypothèse de départ, à savoir que les gestes de direction orchestrale expriment des émotions musicales, la seule validation de cette hypothèse résidant dans les accords minimaux qui avaient pu être trouvés par les annotateurs au moment de l'étiquetage des échantillons du corpus ORCHESTRE-3D. Cette spécificité contraste

avec des expériences menées sur des bases de contenus émotionnels actées à la demande, et qui plus est se référent à des catégorisations courantes des émotions (*cf.* section II.2.1).

Pour toutes les raisons qui viennent d'être évoquées, la seule expérience réellement comparable à la nôtre est celle de Camurri *et al.* [**163**] (*cf.* dernière ligne du Tableau VI.13), pour qui il s'agissait de reconnaître des émotions dans des gestes dansés. La comparaison avec leurs résultats montre que nos méthode et performances ne sont pas en reste.

Tableau VI.13 Présentation de résultats obtenus dans des approches de la littérature consacrées à la reconnaissance de catégories émotionnelles à partir du geste.

Auteurs	**Emotions**	**Descripteurs, stratégie**	**Taux de reconnaissance**
Cimen *et al.* [**146**]	*tristesse*, *joie*, *colère*, et *calme* (4)	posture, dynamique et fréquence **(3D)**	91%
Bernhardt et Robinson [**147**]	*émotion neutre*, *bonheur*, *colère*, et *tristesse* (4)	primitives du mouvement **(3D)**	81,1%
Balomenos *et al.* [**149**]	*joie*, *tristesse*, *colère*, *peur*, *dégoût* et *surprise* (6)	reconnaissance de geste et analyse du visage **(2D)**	85%
Gunes et Piccardi [**150**]	*anxiété*, *colère*, *dégoût*, *peur*, *bonheur* et *incertitude* (6)	2 canaux : visage et mains **(2D)**	Visage : 76,4% Mains : 89,9% Concaténation : 94,02% Somme : 91,1% Produit : 87,3% Somme pondérée : 79,7%
Camurri *et al*. [**163**]	4 émotions dans des mouvements dansés : *colère*, *peur*, *souffrance*, *joie*	Quantité de mouvement, contraction **(2D)**	30,8-71,9%,

VII. Reconnaissance dynamique de gestes

Dans cette deuxième expérience dite de « reconnaissance locale », l'objectif est de reconnaître dynamiquement, à la volée, l'action au fur et à mesure de sa réalisation. Le geste est désormais caractérisé à chaque instant (e.g., trame) par un vecteur descripteur, que nous présentons dans un premier temps. Les gestes sont alors représentés par des séquences de vecteurs d'observation. De telles séquences vectorielles sont utilisées dans une approche couplant des Modèles de Markov Cachés (Hidden Markov Models) avec une procédure de soft assignment dont l'objectif est de localiser le vecteur d'observation parmi une liste de prototypes préalablement constituée pour établir une représentation simplifiée des gestes. La méthode est testée sur notre propre base d'action (HTI 2014-2015 – cf. section V.2.2), ainsi que sur les corpus de gestes 3D disponibles dans la littérature (cf. section V).

La caractérisation du mouvement à chaque instant *t* satisfait les pré-requis pour une classification du geste en temps réel. Les données peuvent alors être traitées dynamiquement, à la volée avant même la fin de la réalisation du geste et sans pré-segmentation préalable.

VII.1. Descripteurs par trame temporelle

Pour cette approche donc, chaque trame gestuelle *t* est décrite à l'aide d'un vecteur descripteur :

$$v(t) = \big(v_1(t), v_2(t), v_3(t), \dots, v_P(t)\big), \qquad \text{(VII.1)}$$

dont chaque composante est dédiée à une qualité ou sous-qualité issue de la modélisation LMA. Ces différentes composantes sont directement issues du modèle que nous avons élaboré et présenté à la section IV. Elles correspondent précisément aux indices qui avaient été définis par trame et sont présentées dans le Tableau IV.2. Ces quantifications nous conduisent à un total de $P = 17$ indices descriptifs de l'expressivité du geste à chaque instant *t*.

Dans le paragraphe suivant, nous détaillons l'usage de ces descripteurs de trame en vue de l'analyse dynamique geste.

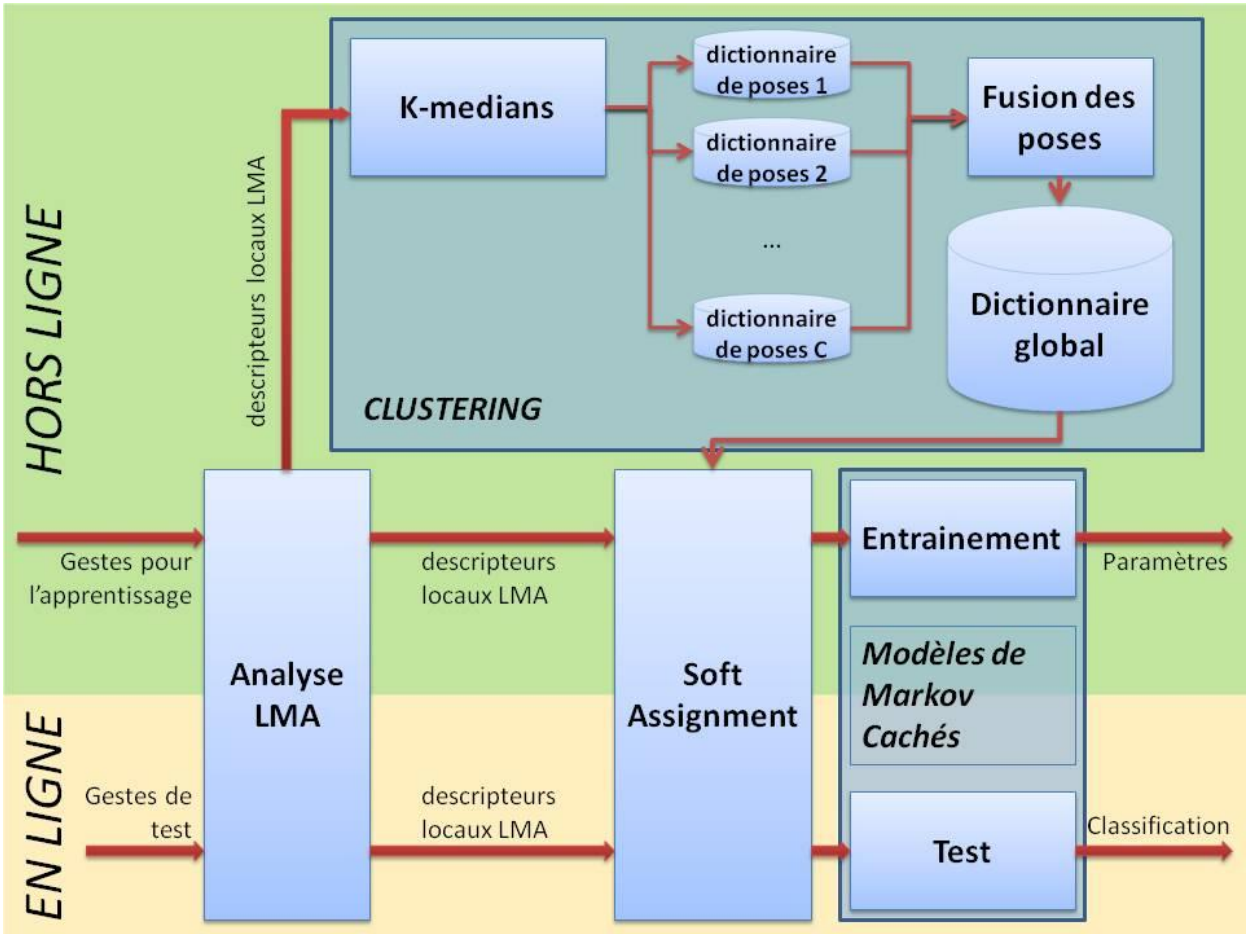


Figure VII.1 Schéma synoptique du cadre d'analyse dynamique du geste.

VII.2. Protocole d'analyse

La méthode d'analyse dynamique du geste est présentée Figure VII.1, avec à la fois l'étape d'entrainement (hors ligne) et l'étape de test/classification (en ligne).

Dans l'étape hors ligne, l'élément central correspond à la construction d'un dictionnaire de poses-clés, dont le rôle est de quantifier de manière pertinente l'espace des poses 3D des gestes. Des sous-dictionnaires individuels sont tout d'abord construits pour chaque catégorie considérée. Ensuite, un processus de fusion de poses est appliqué, afin de déterminer un dictionnaire de poses global, pouvant être appliqué à la totalité des catégories gestuelles. L'objectif ici est d'obtenir un ensemble ref de poses de références à la fois représentatives et non-redondantes pour les poses 3D des gestes considérées, pouvant servir de « mots visuels 3D ».

Ce dictionnaire de poses est ensuite utilisé dans la phase en-ligne de classification. Pour chaque trame d'une séquence de gestes, une procédure d'affectation douce (*soft assignment*) de la pose courante par rapport au dictionnaire de poses ref est considérée. Cela permet notamment d'éviter les problématiques liées à une quantification trop brutale de l'espace des gestes que rencontrent en général les classifications en dur. Dans ce cadre, au lieu de quantifier chaque pose courante par rapport à une pose-clé du dictionnaire, un vecteur de probabilités d'affectation par rapport à l'ensemble des poses du dictionnaire est construit. Les séries temporelles des vecteurs d'affectation qui en résultent sont ensuite exploitées dans le cadre d'un modèle HMM afin d'obtenir une classification en temps réel, à la volée des séquences considérées. Soulignons, que le même processus d'affectation douce est également utilisé dans la phase hors-ligne pour l'entrainement du modèle HMM.

Détaillons, en premier lieu, le processus de construction du dictionnaire des poses de référence.

VII.2.1. Extraction de poses-clés

L'objectif de cette étape hors ligne est de déterminer un lexique de poses de référence distinctes qui peuvent conduire à une représentation simplifiée des gestes. Pour cela, nous supposons disponible une base de gestes d'entrainement, catégorisés selon un lexique de catégories d'actions (pour cette expérience, nous avons notamment considéré la base données HTI 2014-2015 et ses onze classes – *cf.* section V.2.2 pour une description du corpus).

Pour chaque catégorie gestuelle (*i.e.*, classe d'action), nous déterminons en premier lieu une liste de poses de référence, de manière indépendante des autres catégories.

Si l'on considère la liste de l'ensemble des séquences gestuelles représentant les instanciations d'une catégorie de gestes donnée *G*, de taille $|G|$:

$$(S_1^G, S_2^G, \dots, S_{|G|}^G), \tag{VII.2}$$

chaque instanciation *i* est représentée par une série de descripteurs de trame :

$$S_i^G = (v^{G,i}(1), v^{G,i}(2), \dots, v^{G,i}(T^{G,i}), \tag{VII.3}$$

où $T^{G,i}$ désigne le nombre de trames de la *i*-ième séquence gestuelle de la catégorie *G*.

Les poses-clés sont calculées en prenant pour base l'intégralité des séquences $(S_1^G, S_2^G, \dots, S_{|G|}^G)$ à l'aide d'un algorithme de *k-medians clustering* [**110**]. Les centroïdes sont initialisé de manière aléatoire – *i.e.*, les *k* centroïdes initiaux sont tirés au hasard parmi l'intégralité des poses de la série des $|\,G\,|$ séquences. L'algorithme de *k-medians* assure que les centroïdes retenus, qui sont donc exprimés dans l'espace des descripteurs, correspondent bien à des poses existantes extraites du corpus d'entrainement.

A chaque itération de l'algorithme de *k-medians*, la pertinence des groupements fournis est évaluée à l'aide d'une mesure de validité, définie comme le rapport en les similarités intra- et inter-classes :

$$validité = \frac{D_{intra}}{D_{inter}}, \tag{VII.4}$$

où D_{intra} est une mesure de compacité intra-classe définie comme:

$$D_{intra} = \frac{1}{N_G}\sum_{k\in\{1,K\}} \left(\sum_{x \quad c_k} d(x,\mu_k)^2\right) \tag{VII.5}$$

et D_{inter} correspond à la distance inter-groupements :

$$D_{inter} = min_{k\in\{1,K-1\},\ l\in\{k+1,K\}}(d(\mu_k,\mu_l)^2) \tag{VII.6}$$

Dans les équations introduites ci-dessus, les notations suivantes sont utilisées :

- N_G est le nombre total de trames pour l'intégralité des gestes concernés (c'est-à-dire la somme des nombres de trames sur chaque séquence) pour la catégorie G,
- *K* est le nombre de clusters,
- C_k désigne le *k*-ième cluster,
- μ_k désigne le centroïde (moyenne) du *k*-ième cluster,
- $d(x_1, x_2)$ dénote la distance euclidienne normalisée [**172**] entre deux vecteurs x_1 and x_2 dans notre espace de descripteurs LMA.

Soulignons que le nombre de clusters *K* doit être suffisamment important pour pouvoir capter la variabilité des poses au coures de chaque catégorie gestuelle considérée sans pour autant péricliter la pertinence de la représentation en introduisant trop de redondances. Dans nos travaux, nous avons retenu pour le paramètre *K* la valeur 10, en considérant qu'un geste donné peut être intuitivement représenté par un tel nombre de poses de références, sauf dans le cas du MSRC-12 dataset (*cf.* section V.1) où un nombre de 5 poses par geste semblait amplement suffisant, en raison d'une certaine homogénéité de mouvement observée à l'intérieur des classes de gestes.

L'algorithme de *k-medians* tente de minimiser itérativement la mesure de $validité$. A chaque itération, les séquences de l'ensemble d'apprentissage sont affectées au plus proche prototype du dictionnaire courant. Une fois l'affectation réalisée pour l'ensemble des séquences, les prototypes sont ré-calculés en tant que médians de la classe correspondante. L'algorithme de *clustering* s'arrête lorsque la variation de la mesure validité entre deux itérations successives se retrouve en dessous d'un certain seuil ou lorsqu'un nombre maximum d'itérations possibles est atteint, ce dernier étant un paramètre à définir.

A la fin du processus de *clustering*, nous obtenons un vecteur de poses de référence, noté par :

$$(P_1^G, P_2^G, \dots, P_K^G) \quad \text{(VII.7)}$$

qui va caractériser la catégorie gestuelle donnée G. Chaque pose P_j^G correspond donc à un paramétrage de la pose du squelette corporel et à son vecteur caractéristique associé dans l'espace de nos descripteurs de Laban.

Cette stratégie de calcul des poses de référence par catégorie gestuelle offre l'avantage d'une représentation de chaque geste par un nombre réduit de poses-clés bien distinctes. Néanmoins, il est probable que l'on obtienne des poses similaires dans des dictionnaires associés à des catégories différentes, ce qui conduirait à une représentation redondante. Pour éviter une telle redondance, nous appliquons un procédé inter-catégorie de fusion des poses. Cela permet de s'assurer de l'obtention d'un dictionnaire unique, pouvant couvrir l'ensemble des catégories tout en s'assurant de regrouper des poses distinctes les unes des autres.

Pour cela, le principe consiste à fusionner les *clusters* dont les deux centroïdes (*i.e.*, les deux poses-clés) sont séparés par une distance inférieure à un seuil prédéfini ϱ. La liste de poses qui résulte de ce procédé constitue un dictionnaire global de *M* poses de référence, représentatives pour tout le corpus d'entrainement :

$$(P_1^{ref}, P_2^{ref}, \dots, P_M^{ref}) \quad \text{(VII.8)}$$

La disponibilité de telles poses permet alors d'obtenir une représentation de chaque séquence gestuelle, capable d'intégrer la variabilité des gestes réalisés par différents individus. Pour cela, il suffit d'assigner chaque trame de ladite séquence à son prototype le plus proche dans le dictionnaire.

Toutefois, une telle méthode d'affectation dure, dite de « *hard assignment* », se heurte généralement à des erreurs de quantification. Nous avions déjà souligné ce problème à la section III.1.3 à propos de l'approche introduite par Li *et al.* dans [**121**], où la reconnaissance d'actions était également fondée sur des mots visuels. Les auteurs se penchaient sur la prise en compte des capacités à reconstruire une caractéristique spatio-temporelle d'intérêt non seulement du mot le plus proche, mais également d'autres mots voisins (au sens des descripteurs associés).

Pour cette raison, et en lien avec l'approche de Li *et al.* [**121**], nous avons considéré une représentation plus graduelle de chaque trame gestuelle, dite d'affectation douce (*soft assignment*) et présentée dans la section suivante.

VII.2.2. Représentation par affectation douce

La méthode d'affectation douce est utilisée pour localiser un vecteur descripteur parmi une liste de vecteurs prototypes.

Pour chaque descripteur de trame $v(t) = (v_1(t), \dots, v_P(t))$, nous calculons la distance $d_j(t)$ de $v(t)$ à chaque pose-clé P_j^{ref} du dictionnaire global :

$$d_j(t) = d\left(v(t), P_j^{ref}\right), j \in \{1, M\} \quad \text{(VII.9)}$$

Le vecteur d'affectation douce $o(t)$ est défini par une série de distances normalisées :

$$o(t) = (d'_1(t), d'_2(t), \dots, d'_M(t)), \quad \text{(VII.10)}$$

où chaque composante $d'_j(t)$ représente une distance normalisée, définie comme décrit dans l'équation suivante :

$$d'_j(t) = \frac{d_j(t)}{\sum_{i\in\{1,M\}} (d_i(t))}, j \in \{1, M\} \quad \text{(VII.11)}$$

Le vecteur $o(t)$ décrit la position relative du vecteur $v(t)$ à l'instant *t* dans l'espace que dressent les poses de référence.

Dans la section suivante, nous décrivons de quelle manière une séquence de vecteurs de *soft assignment* $(o(1), o(2), \dots, o(T))$ peut être utilisée dans une approche reconnaissance de gestes par modèles HMM.

VII.2.3. Reconnaissance de gestes par modèles HMM

VII.2.3.1. Les modèles HMM : présentation générale

Les modèles – ou chaînes – de Markov caché(e)s (*Hidden Markov Models*: *HMM*) [**98**] sont un cadre probabiliste idéal pour la modélisation des actions humaines, car ils permettent de prendre en compte le caractère séquentiel des données gestuelles, en prenant en compte les probabilités de transition entre poses successives.

Les phénomènes temporels y sont modélisés par une variable temporelle d'observation $o(t)$, sous-tendue par un nombre N_E de valeurs parcourues par une variable d'état caché $e(t)$ à chaque instant *t* :

$$\{E_i\}_{i=1}^{N_E} \quad \text{(VII.12)}$$

De façon générale, les HMM sont décrits à l'aide :

- des probabilités d'état initial :

$$\Pi = \{\pi_i = P(e(0) = E_i)\}_{i=1}^{N_E} \quad \text{(VII.13)}$$

- des probabilités de transitions d'un état caché à un autre :

$$A = \{a_{ij} = P\big(e(t+1) = E_j \,\big|\, e(t) = E_i\big)\}_{i\in\{1,N_E\}, j\in\{1,N_E\}} \quad \text{(VII.14)}$$

- des probabilités d'émission d'une observation o_t à l'instant *t* pour chaque état caché, aussi appelées probabilités de sortie (*output probabilities*) :

$$B = \{b_j(o_t) = P(o_t \mid e(t) = E_j)\}_{j=1}^{N_E} \quad \text{(VII.15)}$$

qui peuvent être modélisées par des distributions probabilistes discrètes, dans le cas d'observations prenant leurs valeurs dans une liste de labels, ou par des fonctions de densité continues.

Ces divers paramètres de modèle sont stockés dans un vecteur global, noté :

$$= (\Pi, A, B) \quad \text{(VII.16)}$$

Les modèles HMM sont construits selon les trois hypothèses suivantes :

- l'état caché $e(t)$ à l'instant *t* ne dépend que de l'état caché $e(t-1)$ à l'instant *t-1* :

$$P(e(t) = e_t \mid \cap_{t'=0}^{t'=t-1} e(t') = e_{t'}) = P(e(t) = e_t \mid e(t-1) = e_{t-1}) \quad \text{(VII.17)}$$

- les probabilités d'émission des états cachés sont indépendantes les unes des autres :

$$P\left(\cap_{t'=1}^{t'=t} o(t') = o_{t'} \,\middle|\, \cap_{t'=0}^{t'=t} e(t') = e_{t'}\right) = \prod_{t'=1}^{t'=t} P\left(o(t') = o_{t'} \,\middle|\, \cap_{t''=0}^{t''=t} e(t'') = e_{t''}\right) \quad \text{(VII.18)}$$

- l'observation $o(t)$ à l'instant *t* ne dépend que de l'état caché $e(t)$ à l'instant *t* :

$$P(o(t) = o_t \mid \cap_{t'=0}^{t'=t} e(t') = e_{t'}) = P(o(t) = o_t \mid e(t) = e_t) \quad \text{(VII.19)}$$

De telles restrictions rendent plus aisée l'expression de la probabilité jointe d'une séquence d'observations $\mathcal{O} = (o_1, o_2, \dots, o_{\mathcal{T}})$ et de la séquence d'états cachés correspondante $\mathcal{E} = (e_0, e_1, \dots, e_{\mathcal{T}})$ où $\mathcal{T}$ est le nombre de trames de la séquence d'observations (on considère que le premier état caché e_0 est non-émetteur d'observation). Cette probabilité peut alors s'exprimer comme :

$$\mathrm{P}(\mathcal{O}, \mathcal{E} \mid {}^{\mathrm{opt}}) = [\prod_{t=1}^{t=\mathcal{T}} \mathrm{P}(\mathrm{o(t)} = \mathrm{o_t} \mid \mathrm{e(t)} = \mathrm{e_t})].[\prod_{t=1}^{t=\mathcal{T}} \mathrm{P}(\mathrm{e(t)} = \mathrm{e_t} \mid \mathrm{e(t-1)} = \mathrm{e_{t-1}})].\mathrm{P}(\mathrm{e(0)} = \mathrm{e_0}) \quad \text{(VII.20)}$$

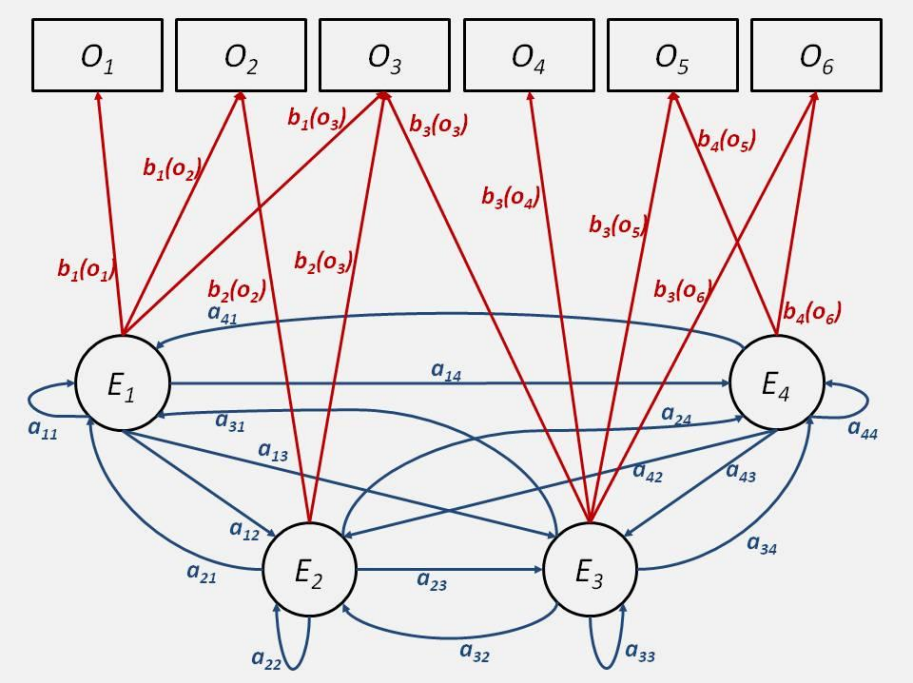


Figure VII.2 Exemple de configuration de HMM pour des observations discrètes et 4 valeurs d'état caché. Les probabilités de transition entre les états sont représentées en bleu, et les distributions d'émissions par les états en rouge.

Un exemple de modèle HMM est illustré Figure VII.2, Ici, nous représentons ce que peut être une configuration de HMM pour quatre valeurs possibles d'états cachés et un espace d'observations discret composé de six labels : $\{O_n\}_{n=1}^{6}$.

Notons que pour cet exemple (et cela vaudra plus généralement pour la suite de ce paragraphe), le modèle HMM est ergodique. La conséquence de cette hypothèse est que tout état peut être atteint depuis n'importe quel état en un nombre fini de pas. Par ailleurs, on constate que les états cachés, selon leur valeur, n'émettent pas nécessairement la totalité des symboles du lexique d'observations (par exemple, l'état de valeur E_3 n'émet que les symboles O_3, O_4, O_5 et O_6 avec les probabilités respectives $b_3(O_3), b_3(O_4)$, $b_3(O_5)$ et $b_3(O_6)$).

VII.2.3.2. Mise en œuvre des modèles HMM

Nous avons décidé d'élaborer notre propre bibliothèque de modèles HMM, de façon à rendre possibles plusieurs choix relatifs aux modèles de probabilités d'émissions et aux procédures d'entrainement et de classification.

En effet, lorsque nous avons souhaité étendre notre modèle de geste expressif à la caractérisation dynamique du mouvement, nous comptions au départ tester deux approches : l'une consistant en une utilisation immédiate de nos descripteurs locaux de Laban en guise d'observation $o(t)$; l'autre étant basée sur du *hard assignment* à partir de notre dictionnaire de poses (*cf.* équation VII.8). Dans ce cas l'usage des lois de probabilité de sortie discrètes auraient été de cours, dans la mesure où chaque observation $o(t)$ aurait été réduite à une pose-clé. Il nous paraissait donc de premier ordre de rendre modulable le type de modèle de probabilités de sortie.

VII.2.3.2.1. Distributions probabilistes pour l'émission des observations

Nous nous sommes donc donné les moyens de choisir entre plusieurs modèles B d'émission d'observations par les états cachés (*cf.* équation VII.15). Nous en avons considéré et développé trois types « classiques ».

1. Dans un premier cas, les émissions sont discrètes et mettent en jeu une liste de labels observables :

$$\left\{\left\{o_{j_m}\right\}_{j_m=1}^{M_j}\right\}_{j=1}^{N_E}, \tag{VII.21}$$

où chaque M_j désigne le nombre d'observations distinctes que l'état E_j peut émettre. Les probabilités d'émissions prennent alors la forme suivante :

$$B = \left\{\left\{P\big(o(t) = o_{j_m} \,\middle|\, e(t) = E_j\big)\right\}_{j_m=1}^{M_j}\right\}_{j=1}^{N_E} \tag{VII.22}$$

2. Dans une seconde configuration, les émissions sont régies par des fonctions de densité gaussiennes définies chacune par une observation moyenne et une matrice de covariance :

$$B = \left\{b_j(o_t) \sim \mathcal{N}(o_t; \mu_j, \Sigma_j)\right\}_{j=1}^{N_E} \tag{VII.23}$$

3. Dans la troisième situation, les émissions d'observations sont modélisées par des mélanges de distributions gaussiennes :

$$B = \left\{ b_j(o_t) \sim \sum_{j_m=1}^{M_j} c_{j_m} \mathcal{N}(o_t; \mu_{j_m}, \Sigma_{j_m}) \right\}_{j=1}^{N_E}, \quad \text{(VII.24)}$$

où pour chaque état E_j, M_j désigne le nombre de distributions gaussiennes de la mixture modélisant l'émission d'observations par l'état E_j, μ_{j_m} correspond à la moyenne de la j_m-ième composante gaussienne pour l'état E_j, Σ_{j_m} à sa matrice de covariance, et c_{j_m} à son poids dans le mélange, de sorte que pour tout état E_j : $\sum_{j_m=1}^{M_j} c_{j_m} = 1$.

VII.2.3.2.2. Procédures d'entrainement/décodage

Nous avons implémenté deux méthodes d'entrainement itératif. Dans la suite, on considère un nombre N_{Seq} de séquences gestuelles. Désignons respectivement par $(\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_{N_{Seq}})$, $(\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_{N_{Seq}})$ et $(\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{N_{Seq}})$ les séquences d'observations correspondant à ces séquences gestuelles, leurs séries d'états cachés correspondants, ainsi que leurs tailles (*e.g.*, nombre de trames de chacune d'elles).

VII.2.3.2.2.1. Entrainement de Baum-Welch

La première méthode d'entrainement implémente l'***algorithme de Baum-Welch*** [**98**], dont l'objectif consiste à estimer itérativement les paramètres de sorte à maximiser l'espérance globale des observations :

$$^{opt} = arg\ max\ \left(\sum_{s=1}^{N_{Seq}} P(\mathcal{O}_s|\ )\right) \quad \text{(VII.25)}$$

Pour une itération donnée, notée $iter$, l'algorithme *forward-backward* [**98**] est utilisé. Il consiste à calculer récursivement les probabilités α (dites « *forward* ») et β (dites « *backward* ») définies de la façon suivante pour chaque valeur d'état caché E_j, à chaque instant t de chaque séquence s :

$$\alpha_j^s(t) \triangleq P\left(\mathcal{O}_s(1), \mathcal{O}_s(2), \dots, \mathcal{O}_s(t), \mathcal{E}_s(t) = E_j \mid \ ^{iter}\right) = \begin{cases} \pi_i \ si\ t = 0 \\ \left(\sum_{i=1}^{N_E} \alpha_i^s(t-1).a_{ij}\right).b_j\left(o(t)\right)\ sinon \end{cases} \quad \text{(VII.26)}$$

$$\beta_i^s(t) \triangleq P\left(\mathcal{O}_s(t+1), \mathcal{O}_s(t+2), \dots, \mathcal{O}_s(\mathcal{T}_s) \mid \mathcal{E}_s(t) = E_i, \ ^{iter}\right) \begin{cases} 1\ si\ t = \mathcal{T}_s \\ \sum_{j=1}^{N_E} a_{ij}.b_j\left(o(t+1)\right).\beta_j^s(t+1)\ \ sinon \end{cases} \quad \text{(VII.27)}$$

Grâce à ces éléments, il est possible de calculer les probabilités d'intérêt suivantes en vue la mise à jour des paramètres pour l'itération suivante de l'algorithme d'entrainement :

$$\text{Probabilité d'occupation : } P(\mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \ ^{iter}) = \frac{\alpha_i^s(t).\beta_i^s(t)}{\sum_{j=1}^{N_E} \alpha_j^s(t).\beta_j^s(t)} \quad \text{(VII.28)}$$

Probabilité de transition : $P(\mathcal{E}_s(t+1) = E_j, \mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad ^{iter}) = \frac{\alpha_i^s(t).a_{ij}.b_j(o(t+1)).\beta_j^s(t+1)}{\sum_{l=1}^{N_E} \alpha_l^s(t).\beta_l^s(t)}$ (VII.29)

Les paramètres du modèle HMM pour l'itération suivante $^{iter+1}$ peuvent alors être mis à jour. Plus précisément, les probabilités initiales (équation VII.13) et transitives (équation VII.14) sont mises à jour comme suit :

$$\Pi^{iter+1} = \left\{ \pi_i^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} P\left(\mathcal{E}_s(0) = E_i \mid \mathcal{O}_s, \quad ^{iter}\right)}{N_{Seq}} \right\}_{i=1}^{N_E} \quad \text{(VII.30)}$$

$$A^{iter+1} = \left\{ a_{ij}^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=0}^{\mathcal{T}s-1} P\left(\mathcal{E}_s(t+1) = E_j, \mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad ^{iter}\right)}{\sum_{s=1}^{N_{Seq}} \sum_{t=0}^{\mathcal{T}s-1} P\left(\mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad ^{iter}\right)} \right\}_{i\in\{1,N_E\}, j\in\{1,N_E\}} \quad \text{(VII.31)}$$

Les probabilités émettrices (équation VII.15) sont mises à jour différemment, en fonction de leur type.

1. Dans le cas de probabilités discrètes (*cf.* équation VII.22) :

$$B^{iter+1} = \left\{ \left\{ P(o(t) = o_{j_m} \mid e(t) = E_j) = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} P\left(\mathcal{E}_s(t) = E_j, \mathcal{O}_s(t) = o_{j_m} \mid \mathcal{O}_s, \quad ^{iter}\right)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad ^{iter}\right)} \right\}_{j_m=1}^{M_j} \right\}_{j=1}^{N_E} \quad \text{(VII.32)}$$

2. Dans le cas de distributions gaussiennes (*cf.* équation VII.23), pour chaque valeur E_j :

$$\mu_j^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} \mathcal{O}_s(t).P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad ^{iter}\right)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad ^{iter}\right)} \quad \text{(VII.33)}$$

$$\Sigma_j^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} (\mathcal{O}_s(t) - \mu_j{}^{iter}) * (\mathcal{O}_s(t) - \mu_j{}^{iter})^T .P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad ^{iter}\right)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad ^{iter}\right)} \quad \text{(VII.34)}$$

3. Dans le cas de mixture de distributions gaussiennes (*cf.* équation VII.24), pour chaque valeur E_j, les paramètres de chaque composante j_m sont calculés selon la formule des mixtures de gaussiennes (*Gaussian Mixture Models* : *GMM* [**173**]) :

$$\mu_{j_m}^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} \mathcal{O}_s(t).\gamma_{j_m}{}^{iter}(t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} \gamma_{j_m}{}^{iter}(t)} \quad \text{(VII.35)}$$

$$\Sigma_{j_m}^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} (\mathcal{O}_s(t) - \mu_{j_m}{}^{iter}) * (\mathcal{O}_s(t) - \mu_{j_m}{}^{iter})^T .\gamma_{j_m}{}^{iter}(t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} \gamma_{j_m}{}^{iter}(t)} \quad \text{(VII.36)}$$

$$c_{j_m}^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{T_S} \gamma_{j_m}{}^{iter}(t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{T_S} \sum_{j'_m=1}^{M_j} \gamma_{j'_m}{}^{iter}(t)} \quad \text{(VII.37)}$$

où pour chaque instant *t* :

$$\gamma_{j_m}{}^{iter}(t) = P\left(\mathcal{E}_s(t) = E_j \mid \mathcal{O}_s, \quad {}^{iter}\right) . \frac{c_{j_m}{}^{iter} \mathcal{N}(o(t); \mu_{j_m}{}^{iter}, \Sigma_{j_m}{}^{iter})}{\sum_{j'_m=1}^{M_j} c_{j'_m}{}^{iter} \mathcal{N}(o(t); \mu_{j'_m}{}^{iter}, \Sigma_{j'_m}{}^{iter})}, \quad \text{(VII.38)}$$

L'entrainement itératif s'arête lorsque la différence entre deux espérances globales successives, calculées à des pas d'itération $iter$ et $iter + 1$ à partir de la mise à jour des paramètres de modèle , sont inférieures à un certain seuil $\varepsilon = 0.001$. Le dernier jeu de paramètres calculé est alors conservé.

VII.2.3.2.2.2. Entrainement de Viterbi

La deuxième méthode d'entrainement utilise l'***algorithme de Viterbi*** qui vise à trouver les séquences d'états cachés, émettrices des séquences d'observations, qui maximisent les probabilités *a posteriori* :

$${}^{opt} = arg\ max\ \left(\sum_{s=1}^{N_{Seq}} P(\mathcal{E}_s | \mathcal{O}_s, \quad)\right) \quad \text{(VII.39)}$$

Cette méthode d'apprentissage est plus rapide que la précédente, dans la mesure où elle ne nécessite pas l'emploi de la procédure forward-backward, et ne cherche à calculer ni les probabilités d'occupation des états cachés ($P(\mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad {}^{iter})$: équation VII.28) ni les probabilités de changement d'état ($P\left(\mathcal{E}_s(t+1) = E_j, \mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad {}^{iter}\right)$: équation VII.29). En effet, elle vise à tracer de façon dynamique le chemin de plus forte probabilité au sein des états cachés.

Ainsi, à un instant *t* d'une séquence s, une valeur $\delta_j^s(t)$ est calculé, qui dénote la plus grande probabilité d'un chemin de taille *t* se terminant en E_j et tient compte des observations :

$$\delta_j^s(t) \triangleq \max_{(\mathcal{E}_s)_{t'=0}^{t'=t}} P\left(\{\mathcal{O}_s\}_{t'=1}^{t'=t}, \{\mathcal{E}_s\}_{t'=0}^{t'=t-1}, \mathcal{E}_s(t) = E_j \mid \quad {}^{iter}\right) = \begin{cases} \pi_i \ si\ t = 0 \\ \max_{i \in \{1, N_E\}} \delta_i^s(t-1).a_{ij}.b_j\left(o(t)\right) \ sinon \end{cases} \quad \text{(VII.40)}$$

Les valeurs de $\boldsymbol{\delta}$ peuvent ainsi être calculées itérativement. En outre, à chaque instant *t* et chaque état E_j, le calcul de $\delta_j^s(t)$ fourni le « meilleur état précédent » :

$$\mathcal{E}_s^{Best}(t-1) = \underset{i \in \{1, N_E\}}{arg\ max}\ \delta_i^s(t-1).a_{ij}.b_j\left(o(t)\right) \quad \text{(VII.41)}$$

Par un parcours en marche arrière à partir de la dernière trame *t*, il est donc possible de reconstituer le chemin $(\mathcal{E}_s)_{t'=0}^{t'=t}$ qui maximise $\delta_j^s(t)$. En appliquant un tel procédé de reconstruction du « meilleur » chemin pour toutes les séquences, il est donc possible de définir des fonctions indicatrices relativement aux états occupés et aux transitions d'état à état, qui remplissent respectivement les mêmes rôles que les probabilités d'occupation des états cachés ($P(\mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad {}^{iter})$: équation VII.28) et de changement d'état ($P\left(\mathcal{E}_s(t+1) = E_j, \mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \quad {}^{iter}\right)$: équation VII.29) pour l'algorithme Baum-Welch :

$$I^s(i,t) = \begin{cases} 1 \; si \; \mathcal{E}_s(t) = E_i \\ 0 \; sinon \end{cases} \quad \text{(VII.42)}$$

$$I^s(i,j,t) = \begin{cases} 1 \; si \; \mathcal{E}_s(t+1) = E_j \; et \; \mathcal{E}_s(t) = E_i \\ 0 \; sinon \end{cases} \quad \text{(VII.43)}$$

$$I^s(j,o_t,t) = \begin{cases} 1 \; si \; \mathcal{E}_s(t) = E_j \; et \mathcal{O}_s(t) = o_t \\ 0 \; sinon \end{cases} \quad \text{(VII.44)}$$

Les probabilités initiales (équation VII.13) et de transition (équation VII.14) peuvent alors être mises à jour comme suit :

$$\Pi^{iter+1} = \left\{ \pi_i^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} I^s(i,t)}{N_{Seq}} \right\}_{i=1}^{N_E} \quad \text{(VII.45)}$$

$$A^{iter+1} = \left\{ a_{ij}^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=0}^{\mathcal{T}s-1} I^s(i,j,t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=0}^{\mathcal{T}s-1} I^s(i,t)} \right\}_{i \in \{1,N_E\}, j \in \{1,N_E\}} \quad \text{(VII.46)}$$

Comme pour l'algorithme précédent, les probabilités émettrices (équation VII.15) sont mises à jour différemment selon leur type :

1. Dans le cas de probabilités discrètes (équation VII.22) :

$$B^{iter+1} = \left\{ \left\{ P\left(o(t) = o_{j_m} \mid e(t) = E_j\right) = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} I^s(j,o_{j_m},t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} I^s(j,t)} \right\}_{j_m=1}^{M_j} \right\}_{j=1}^{N_E} \quad \text{(VII.47)}$$

2. Dans le cas de distributions gaussiennes (équation VII.23), pour chaque valeur d'état caché E_j :

$$\mu_j^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} \mathcal{O}_s(t).I^s(j,t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} I^s(j,t)} \quad \text{(VII.48)}$$

$$\Sigma_j^{iter+1} = \frac{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} (\mathcal{O}_s(t) - {\mu_j}^{iter}) * (\mathcal{O}_s(t) - {\mu_j}^{iter})^T . I^s(j,t)}{\sum_{s=1}^{N_{Seq}} \sum_{t=1}^{\mathcal{T}s} I^s(j,t)} \quad \text{(VII.49)}$$

3. Dans le cas de mixture de distributions gaussiennes (équation VII.24), les mises à jour suivent une règle similaire que dans le cas de l'algorithme Baum-Welch (équation VII.25), à une différence près. Ici, pour chaque instant *t*, l'indice ${\gamma_{j_m}}^{iter}(t)$ est remplacé par le suivant :

$$\gamma_{j_m}^{iter,bis}(t) = I^s(j,t).\frac{{c_{j_m}}^{iter} \mathcal{N}(o(t); {\mu_{j_m}}^{iter}, {\Sigma_{j_m}}^{iter})}{\sum_{j'_m \in [1,M_j]} {c_{j'_m}}^{iter} \mathcal{N}(o(t); {\mu_{j'_m}}^{iter}, {\Sigma_{j'_m}}^{iter})} \quad \text{(VII.50)}$$

Comme pour l'algorithme Baum-Welch, l'entrainement itératif se termine lorsque la différence entre deux sommes des probabilités *a posterori* successives, sont inférieures à un seuil $\varepsilon = 0.001$.

VII.2.3.2.2.3. Procédure de décodage de Viterbi

Pour chaque séquence de test $\mathcal{O}_{test}$, l'étape de décodage, c'est-à-dire la détermination de la séquence $\mathcal{E}^{opt}$ d'états cachés la plus probable au vu de la succession des observations, utilise également l'algorithme de Viterbi [**98**], qui maximise les probabilités « a posterori » :

$$\mathcal{E}^{opt} = arg\ max_{\mathcal{E}}(P(\mathcal{E}|\mathcal{O}_{test},\ \ ^{opt})) \tag{VII.51}$$

La reconstitution de ce chemin $\mathcal{E}^{opt}$ est permise par l'usage des équations VII.40 et VII.41.

Quel que soit l'algorithme utilisé, pour des séquences relativement longues (*i.e.*, plusieurs milliers de trames), les valeurs des probabilités mises en œuvre tendent à devenir extrêmement petites. Cela pose un réel problème, puisque la précision finie des nombre flottants représentés sur un ordinateur empêche de pouvoir appliquer directement ces algorithmes. Afin de s'affranchir de cette limitation, nous adopté la méthode décrite dans la section suivante.

VII.2.3.3. Gestion de la problématique de précision finie

Pour remédier à ce problème de précision, nous avons suivi les recommandations proposées dans [**174**]. Le principe consiste à utiliser de logarithmes de probabilités. Nous avons pour cela défini un seuil de lissage du logarithme népérien pour des valeurs proches de zéro. Cette valeur, notée par *_SMOOTHED_ZERO_LOG*, correspond à la valeur « *Not a Number* » fournie par l'arithmétique des nombres flottants dans la plupart des implantations.

Conformément aux recommandations données dans [**174**], nous avons ainsi redéfini plusieurs opérateurs permettant de gérer l'usage des fonctions exponentielle et logarithmique pour la valeur *_SMOOTHED_ZERO_LOG*, rendant alors possible le calcul et le maintien de probabilités réelles aussi faibles que possible :

$$exp^{stable}(x) = \begin{cases} 0\ si\ x = \ \_SMOOTHED\_ZERO\_LOG \\ exp(x)\ sinon \end{cases} \tag{VII.52}$$

$$ln^{stable}(x) = \begin{cases} \_SMOOTHED\_ZERO\_LOG\ si\ x = \ 0 \\ ln(x)\ sinon\ (\text{à}\ condition\ que\ x > 0) \end{cases} \tag{VII.53}$$

$$somme\_logarithmique^{stable}\left(ln^{stable}(x), ln^{stable}(y)\right) = \begin{cases} ln^{stable}(x+y) si\ x > 0\ et\ y > 0 \\ ln^{stable}(x)\ si\ y = 0 \\ ln^{stable}(y)\ si\ x = 0 \end{cases} \tag{VII.54}$$

$$produit\_logarithmique^{stable}\left(ln^{stable}(x), ln^{stable}(y)\right) = \begin{cases} ln^{stable}(x) + ln^{stable}(x) si\ x > 0\ et\ y > 0 \\ \_SMOOTHED\_ZERO\_LOG\ si\ x = \ 0\ ou\ y = \ 0 \end{cases} \tag{VII.55}$$

VII.2.3.4. Stockage des probabilités et gain de temps

Le caractère itératif des algorithmes d'apprentissage des HMM (*cf.* section VII.2.3.2.2) requiert de garder en mémoire les paramètres de modèle, ainsi que les probabilités de référence nécessaires à la mise

à jour dudit modèle , calculées pour chaque état caché possible E_j à chaque trame *t* de chaque séquence gestuelle (*e.g.*, $b_j(o(t))$, $\alpha_j^s(t)$, $\beta_j^s(t)$, $\delta_j^s(t)$).

Ces probabilités de référence nécessitent notamment l'accès aux valeurs des probabilités d'émissions d'observations B (équation VII.15) pour chaque état j et chaque trame *t* de toute séquence $\mathcal{E}_s$. Dans le cas de probabilités d'émission d'observations modélisées spécifiquement par des gaussiennes (équation VII.23) ou des mixtures de gaussiennes (équation VII.24), les calculs peuvent s'avérer d'autant plus longs et coûteux en termes de mémoire que la taille des observations est grande. D'autre part, les appels aux fonctions de densité concernées par ces modèles probabilistes utilisent des déterminants des matrices de covariance des distributions gaussiennes en jeu, ainsi que les inverses de ces matrices :

- $\{(|\Sigma_j|, \Sigma_j^{-1})\}_{j=1}^{N_E}$ dans le cas de distributions gaussiennes ;
- $\left\{\{|\Sigma_{j_m}|, \Sigma_{j_m}^{-1}\}_{j_m=1}^{M_j}\right\}_{j=1}^{N_E}$ dans le cas de mixtures de gaussiennes,

Nous avons en conséquence veillé, pour chaque nouvelle itération de calcul du modèle , à stocker ces valeurs et matrices, pour les rendre immédiatement disponibles à chaque appel d'une probabilité de sortie de type $b_j(o_t)$, et ainsi éviter des calculs trop laborieux.

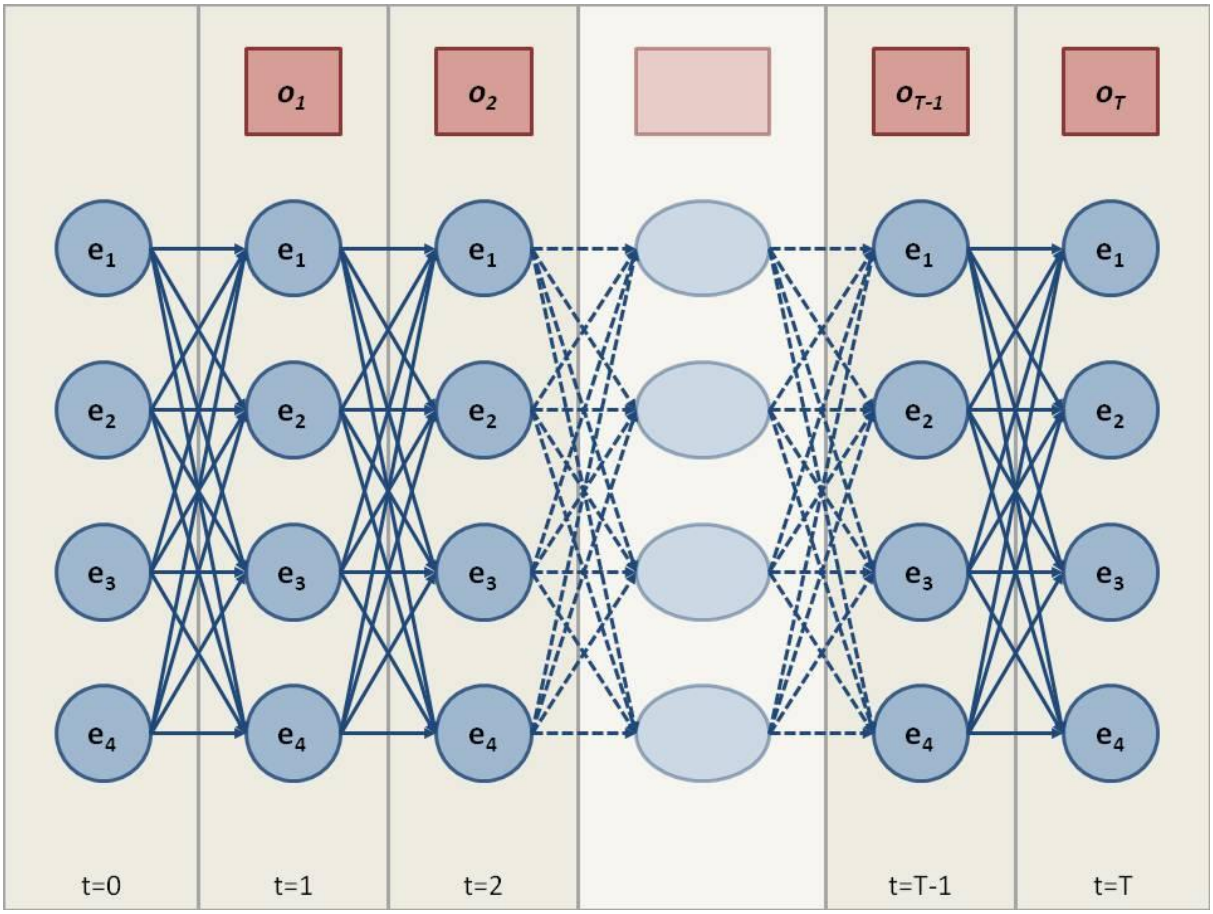


Figure VII.3 Illustration d'un maillage de HMM pour 4 valeurs possibles d'état caché à chaque trame *t*. Le premier état caché est non-émetteur d'observation. Pour le reste, chaque trame *t* (*Timeslot*, correspondant à une colonne grise) est représentée par l'observation o_t émise par l'état $e(t)$ ainsi que par les valeurs candidates (en bleu) pour ce dernier. Dans notre formalisme, ces valeurs candidates correspondent à des nœuds (*Nodes*). Les transitions entre l'état à un instant *t* et l'état à l'instant suivant sont modélisées par des flèches.

Par ailleurs, nous avons opté pour un formalisme de programmation objet consistant à représenter par un nœud (*Node*) chaque valeur possible pour un état caché $e(t)$ à tout instant *t* d'une séquence. La Figure VII.3 montre ainsi le « treillis » que constituent ces nœuds dans notre implantation, et les transitions qui sont possibles entre eux, selon les probabilités de transition *A* (équation VII.14).

Pour chaque instant *t* d'une séquence $\mathcal{E}_s$, chaque nœud (représenté par un cercle bleu) correspond à une valeur possible E_j pour l'état caché $\mathcal{E}_s(t)$, émettrice de l'observation. Dans notre architecture, un tel nœud se voit doté de propriétés suivantes :

$$\_logBeO = ln^{stable}\left(b_j(o(t))\right), \tag{VII.56}$$

$$\_logAlpha = ln^{stable}\left(\alpha_j^s(t)\right), \tag{VII.57}$$

$$\_logBeta = ln^{stable}\left(\beta_j^s(t)\right), \tag{VII.58}$$

$$\_logDelta = ln^{stable}\left(\delta_j^s(t)\right), \tag{VII.59}$$

Ces probabilités ne sont calculées qu'une seule fois, et sont donc stockées en tant qu'attributs du nœud jusqu'à leur mise à jour éventuelle (*e.g.*, dans le cas d'entrainement des HMM, si une itération d'estimations supplémentaire est requise). Elles peuvent être directement réutilisées telles quelles sans avoir à être systématiquement recalculées. Cela représente un gain de temps non-négligeable, notamment pour ce qui est des probabilités de sortie $b_j(o(t))$, fortement sollicitées :

1. dans le cas de l'entrainement effectué avec l'algorithme Baum-Welch, elles sont calculées leur unique fois au moment de l'emploi de l'algorithme *forward* (équation VII.26) ;
2. dans le cas de l'entrainement ou du décodage effectués avec l'algorithme de Viterbi, elles sont calculées pour l'instant *t* concerné lors de l'évaluation $\delta_j^s(t)$ du chemin garantissant la probabilité maximale a posteriori (équation VII.40).

On constate par ailleurs que lors de l'entrainement de Baum-Welch, les probabilités dites d'« occupation » $P(\mathcal{E}_s(t) = E_i \mid \mathcal{O}_s, \ ^{iter})$ (équation VII.28) sont très fortement utilisées lors de l'étape de mise à jour des paramètres de modèle . L'estimation d'une telle probabilité fait appel au calcul de la probabilité $\sum_{j=1}^{N_E} \alpha_j^s(t).\beta_j^s(t)$ dont il est facile de montrer qu'elle est égale à $P\left(\mathcal{O}_s \mid \ ^{iter}\right)$:

$$P\left(\mathcal{O}_s \mid \ ^{iter}\right) = \sum_{j=1}^{N_E} \alpha_j^s(t).\beta_j^s(t), \ \forall\, t \in \{0, \mathcal{T}s\}, \tag{VII.60}$$

Or, cette probabilité n'a de sens que pour la trame *t* de la séquence $\mathcal{E}_s$, et non pour un nœud particulier. Pour continuer à gagner du temps et de la mémoire, nous avons donc décidé de ne calculer cette valeur qu'une seule fois au moment de l'appel de la procédure *forward* qui consacre le calcul des α nécessaires (équation VII.26) et de la stocker dans un objet prenant le nom générique de *TimeSlot*, prenant comme attribut la liste des nœuds représentant les valeurs d'état possibles. Un tel objet correspond à une colonne grise sur la Figure VII.3.

Un *TimeSlot* incorpore donc une observation $o(t) = o_t$ et une série de nœuds correspondants aux valeurs d'état susceptibles de la générer.

Ces diverses stratégies de stockage de valeurs ou de matrices aux calculs lourds ont permis un gain de temps de calcul considérable, malgré l'augmentation de ce temps de calcul avec la taille des observations.

VII.2.4. Choix des méthodes et protocole d'évaluation

Dans notre approche, les états cachés correspondent à nos catégories gestuelles. Le vecteur d'observation $o(t)$ pour un instant *t* est celui qui résulte de l'étape de *soft assignment* (*cf.* section VII.2.2). Les fonctions de densité B sont données par des distributions gaussiennes (équation VII.23).

Les procédures de paramétrage des HMM avec l'algorithme Baum-Welch et avec l'entrainement de Viterbi n'ayant pas donné de différence significative du point de vue des résultats de classification, nous avons décidé de nous focaliser sur les résultats obtenus à l'aide de l'algorithme Baum-Welch.

Dans la section suivante, nous présentons et analysons les résultats obtenus sur différents corpus de gestes.

Pour chacun de ces corpus, le protocole d'évaluation retenue est le suivant.

- Nous avons utilisé un schéma classique de validation croisée en cinq étapes, avec un rapport quantitatif entre données d'entrainement et données de test de 80%/20%. Cette validation croisée s'appuie sur une division du corpus ou sous-corpus de gestes concerné en cinq blocs préservant au mieux la distribution initiale des classes.
- Lors de l'étape de test (*i.e.*., étape de reconnaissance effective du geste), chaque séquence d'observations $\mathcal{O}_{test}$ est décodée suivant la procédure de décodage de Viterbi (*cf.* section VII.2.3.2.2.3). Il en résulte alors la séquence d'états cachés optimale $\mathcal{E}_{\mathcal{O}_{test}}^{opt}$, composée d'une suite de symboles correspondant à des catégories gestuelles :

$$\mathcal{E}_{\mathcal{O}_{test}}^{opt} = \left(\mathcal{E}_{\mathcal{O}_{test}}^{opt}(0),\ \mathcal{E}_{\mathcal{O}_{test}}^{opt}(1),\ \mathcal{E}_{\mathcal{O}_{test}}^{opt}(2), \ldots,\ \mathcal{E}_{\mathcal{O}_{test}}^{opt}(\mathcal{T}) \right), \tag{VII.61}$$

et dont il s'agit de déduire le geste effectivement reconnu.

Pour évaluer cette opération de reconnaissance, nous identifions dans la séquence de décodage $\mathcal{E}_{\mathcal{O}_{test}}^{opt}$ les trois classes les plus représentées, ordonnées selon leur fréquence d'apparition relative au regard de la taille $\mathcal{T}$ de la séquence. La méthode retenue est illustrée Figure VII.4, où pour faciliter la lecture, le lexique des classes est composé de lettres capitales. Les symboles de la séquence $\mathcal{E}_{\mathcal{O}_{test}}^{opt}$ permettent de définir un histogramme de décodage présentant pour chacune des classes le nombre de fois où son symbole a été décodé. Alors, la classe $\mathcal{E}_{\mathcal{O}_{test}}(1)$ dont l'effectif est le plus grand est attribuée au geste. Nous conservons également les noms $\mathcal{E}_{\mathcal{O}_{test}}(2)$ et $\mathcal{E}_{\mathcal{O}_{test}}(3)$ des classes reconnues respectivement en seconde et troisième positions. Notons que plusieurs classes peuvent occuper un même rang, auquel cas $\mathcal{E}_{\mathcal{O}_{test}}(2)$ et $\mathcal{E}_{\mathcal{O}_{test}}(3)$ peuvent parfois ne se voir attribuer aucune valeur. Ainsi, dans le cas de l'exemple donné à la Figure VII.4,

  - $\mathcal{E}_{\mathcal{O}_{test}}(1)$ prend une unique valeur *A* ;

- $\mathcal{E}_{\mathcal{O}_{test}}(2)$ prend trois valeurs, car la seconde position est partagée par trois classes : *C*, *D* et *G* ;
- $\mathcal{E}_{\mathcal{O}_{test}}(3)$ n'a aucun sens, dans la mesure où plusieurs classes se partagent le second rang.

Un tel calcul pour chaque séquence de test permet de définir pour chaque classe gestuelle G trois taux de reconnaissance $TR^G(1)$, $TR^G(2)$ et $TR^G(3)$ définis comme les pourcentages de gestes où la catégorie est correctement identifiée respectivement en première, seconde et troisième positions.

$$TR^G(i) = \frac{\sum_{s=1}^{N_{test}} I_i^s(G,G)}{\sum_{s=1}^{N_{test}} I^s(G)}, \forall\, i \in \{1,2,3\}, \tag{VII.62}$$

où :

$$I_i^s(G,G') = \begin{cases} 1\ si\ \mathcal{E}_{\mathcal{O}_s}(i) = G\ et\ Vérité\ Terrain_{\mathcal{O}_s} = G' \\ 0\ sinon \end{cases}, \tag{VII.63}$$

$$I^s(G) = \begin{cases} 1\ si\ Vérité\ Terrain_{\mathcal{O}_s} = G \\ 0\ sinon \end{cases}, \tag{VII.64}$$

et N_{test} désigne le nombre de séquences de tests et $Vérité\ Terrain_{\mathcal{O}_s}$ la classe à laquelle le geste s appartient effectivement.

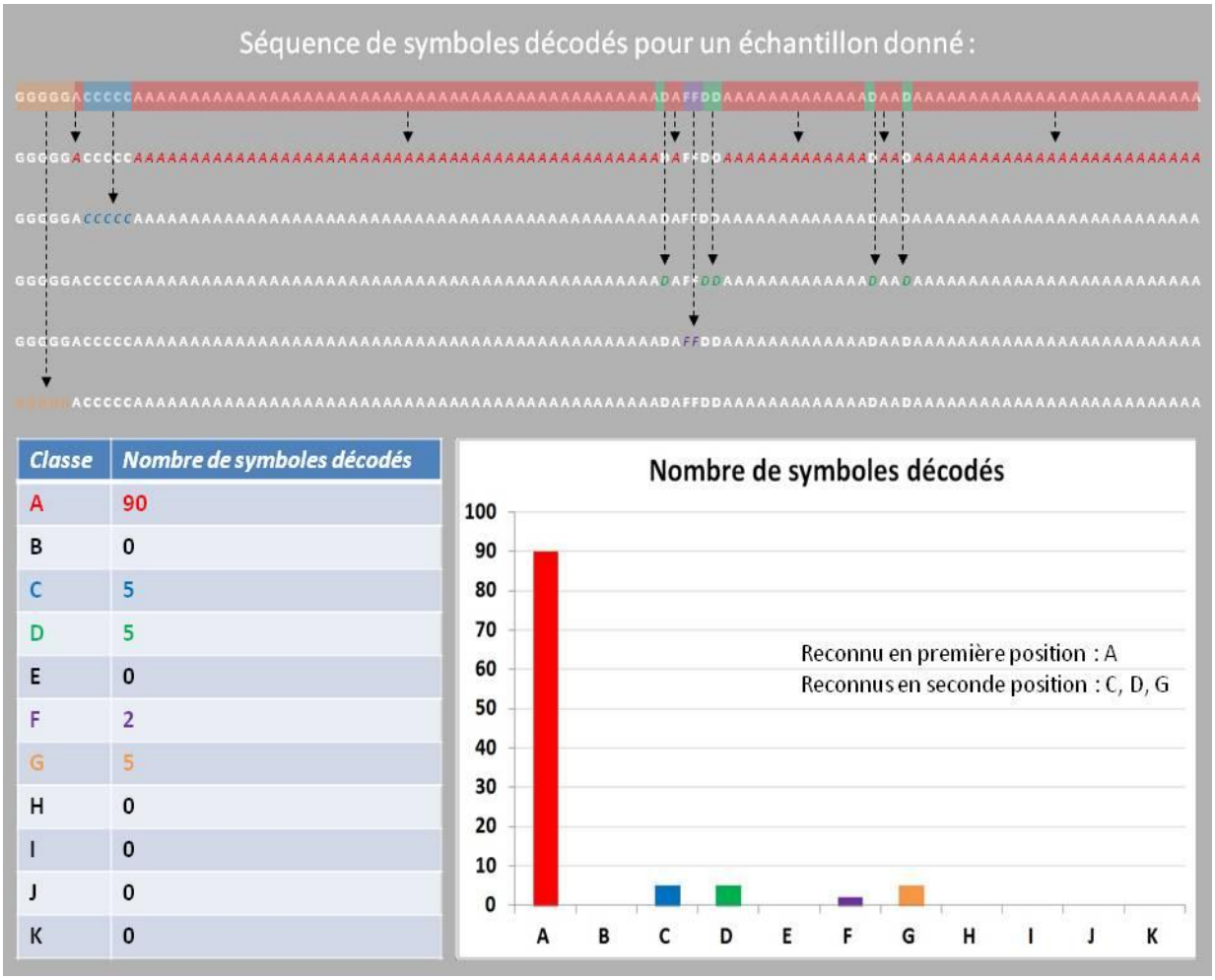
Nous avons également considéré les taux de reconnaissances cumulés $TR_{cum}^G(1)$, $TR_{cum}^G(2)$ et $TR_{cum}^G(3)$ dénotant respectivement les taux de reconnaissance en première position, parmi les deux premières positions, et enfin parmi les trois premières positions :

$$TR_{cum}^G(i) = \sum_{k=1}^{i} TR(i)\,, \forall\, i \in \{1,2,3\} \tag{VII.65}$$

Pour compléter l'étude de la performance de notre système d'analyse dynamique du geste, nous proposons également d'analyser les taux de reconnaissance par trame. Ces taux sont présentés dans une matrice de confusion C, où chaque élément $C(i,j)$ dénote le pourcentage de fois où une trame consacrant la réalisation du geste *i* a été classifiée dans la catégorie *j*, de telle sorte que :

$$\sum_{j=1}^{G} C(i,j) = 1, \forall\, i \in \{1,G\} \tag{VII.66}$$

Les éléments de la diagonale correspondent donc aux taux de reconnaissance des gestes, c'est-à-dire aux pourcentages de classification correcte des trames gestuelles. Pour chaque ligne d'une telle matrice, le meilleur score sera représenté en gras. Les éléments diagonaux qui ne correspondent pas à la valeur maximale de leur ligne seront représentés en rouge, auquel cas on déduira que les trames d'exécution de la catégorie gestuelle correspondant à la ligne en question n'ont pas été correctement classifiées, dans la mesure où une autre catégorie aura obtenu un pourcentage de décodage supérieur. Les taux de reconnaissance par trame ont été calculés pour l'intégralité des trames gestuelles du corpus ou sous-corpus d'intérêt.



Classe	Nombre de symboles décodés
A	90
B	0
C	5
D	5
E	0
F	2
G	5
H	0
I	0
J	0
K	0

Figure VII.4 Illustration de l'étiquetage d'un échantillon lors de l'étape de reconnaissance. Pour faciliter la lecture, nous avons pris l'exemple d'un lexique de onze classes, représentées par des lettres en capitales. La séquence de décodage est reproduite sur plusieurs lignes pour mettre en valeur les symboles reconnus. C'est la classe A qui est reconnue en première position. Trois classes arrivent au second rang. Du fait qu'elles sont au nombre de trois, aucune classe n'est reconnue en troisième position.

VII.3. Evaluation expérimentale

Nous avons appliqué les descripteurs locaux présentés à la section VII.1 à plusieurs corpus de gestes et pour chacun d'eux, nous avons suivi l'approche présentée à la section VII.2 : nous avons constitué un lexique de poses pour chaque catégorie gestuelle, et réuni les clés dans un dictionnaire global. Dans les sections suivantes, nous présentons les résultats obtenus sur quatre bases différentes, qui sont les corpus MSR Action 3D, MSRC-12, UTKinect-Human Detection et HTI 2014-2015 (*cf.* section V.2.2).

VII.3.1. Résultats obtenus sur le corpus MSR Action 3D

La première base de test pour notre approche dynamique est le corpus MSR Action 3D (*cf.* section V.1), et plus précisément les sous-ensembles qui sont définis par ses auteurs [**38**] :

Les actions sont réparties en trois sous-ensembles :

- A1 : *se pencher*, *coup de poing*, *coup de marteau*, *taper des mains*, *lancer au loin*, *signe horizontal de la main*, *ramasser et jeter*, et *service cuillère* ;

- A2 : *dessiner un cercle*, *cocher une case*, *dessiner une croix*, *coup de pied vers l'avant*, *attraper d'une main*, *signe de la main*, *boxer sur le côté* et *signe avec les deux mains* ;
- A3 : *coup de pied vers l'avant*, *swing de golf*, *lancer au loin*, *jogging*, *ramasser et jeter*, *coup de pied sur le côté*, *service cuillère* et *service volée*.

Pour chacun de ces sous-ensembles donc, nous suivons le protocole présenté à la section VII.2, à partir de la formation de dictionnaires de classes de gestes de $K = 10$ poses. Pour chacun de ces sous-ensembles, nous présentons les résultats obtenus avec pour choix de seuil de fusion des dictionnaires de classe ϱ=3.0. Pour chaque sous-ensemble, c'est presque systématiquement avec cette valeur de seuil que les meilleurs résultats ont été obtenus, à quelques 4 ou 5% près.

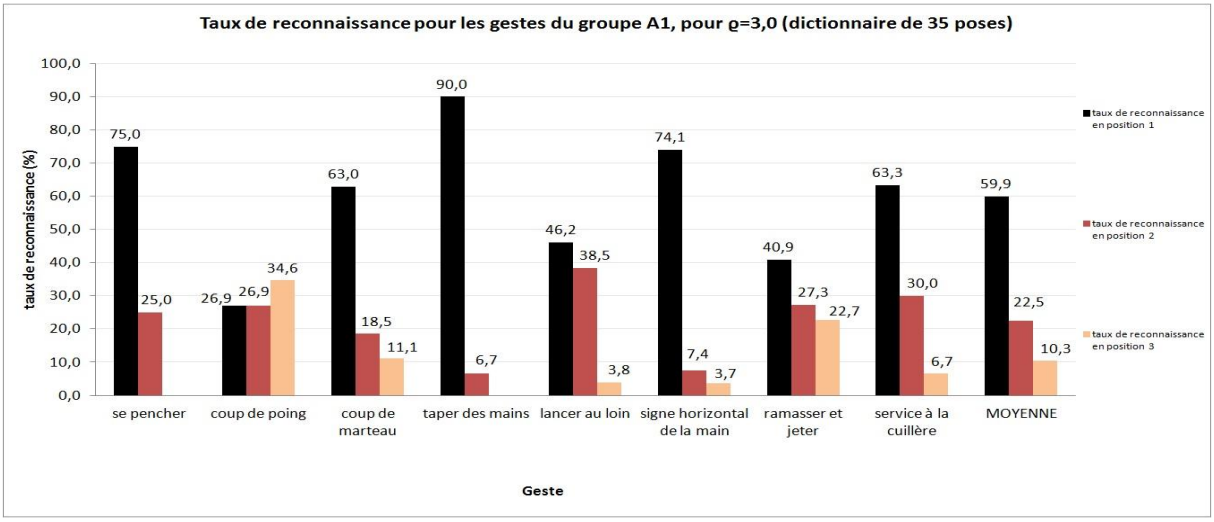


Figure VII.5 Taux de reconnaissance par classe pour les gestes du groupe A1 du corpus MSR Action 3D pour un dictionnaire de 35 poses.

Tableau VII.1 Matrice de confusion par trame pour les gestes du groupe A1 du corpus MSR Action 3D pour ϱ=3,0 (dictionnaire de 35 poses).

Matrice de confusion	se pencher	coup de poing	coup de marteau	taper des mains	lancer au loin	signe horizontal de la main	ramasser et jeter	service cuillère
se pencher	**46,8**	0,4	2,3	4,1	0,0	31,7	14,0	0,8
coup de poing	0,3	26,5	26,1	0,9	4,2	**35,7**	2,2	4,2
coup de marteau	1,6	4,6	**43,1**	1,6	2,9	37,9	3,3	5,0
taper des mains	0,1	0,9	0,9	**64,6**	0,0	27,0	0,3	6,2
lancer au loin	0,0	10,6	24,6	1,5	**30,3**	26,7	1,7	4,6
signe horizontal de la main	1,8	13,6	10,0	11,6	0,4	**57,8**	0,6	4,3
ramasser et jeter	22,0	3,2	10,9	1,6	15,6	13,5	**25,8**	7,4
service cuillère	4,3	2,6	6,3	7,2	2,9	28,3	8,1	**40,5**

Dans le cas du sous-ensemble A1, seules les trois classes *taper des mains*, *signe horizontal de la main* et *se pencher* obtiennent des taux de reconnaissance supérieurs à 74%, et trois classes ont des taux inférieurs à 50% : *lancer au loin*, *coup de poing*, et *ramasser et jeter*, pour un taux de reconnaissance global moyen de 59.9% (Figure VII.5). Les cumuls des taux de reconnaissance aux deux premières positions $TR_{cum}^{moyen}(2)$ et aux trois premières positions $TR_{cum}^{moyen}(3)$ sont en moyenne respectivement égaux à 82.5 et 92.8%.

Pour chaque classe, les taux reconnaissance par trame gestuelle (Tableau VII.1) des confusions peuvent exister entre la classe *coup de poing* et les deux classes *coup de marteau* et *signe horizontal de la main*, entre la classe *lancer au loin* et ces deux dernières, ou encore entre *ramasser et jeter* et *se pencher*, confusions que l'on peut attribuer à des ressemblances en termes d'exécution des gestes concernés – on rappellera notamment que pour ce corpus MSR Action 3D, il avait été spécifié aux sujets exécutants d'utiliser leur « côté droit » pour les gestes n'impliquant qu'un seul membre. (*cf.* section V.1).

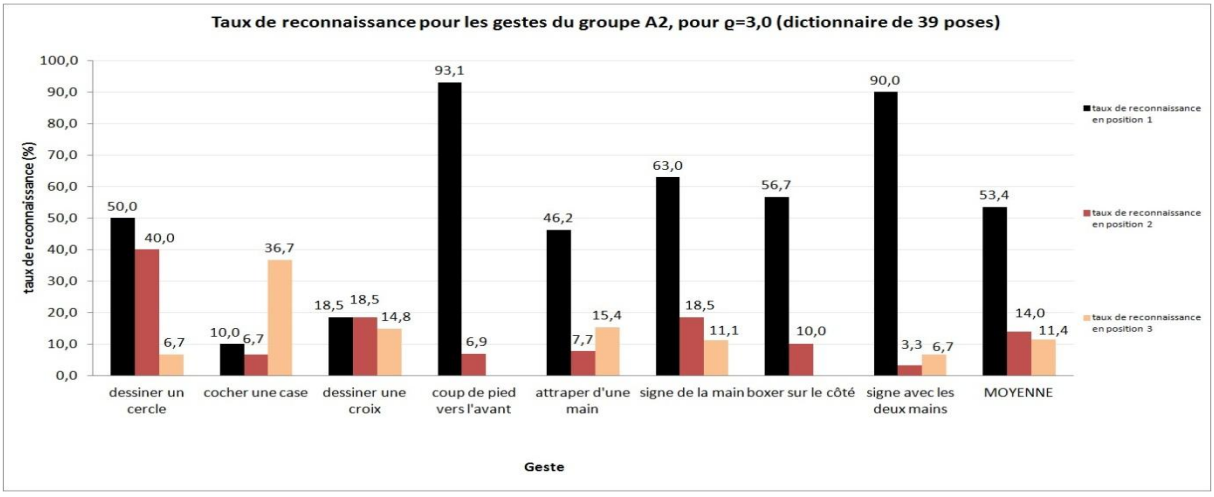


Figure VII.6 Taux de reconnaissance par classe pour les gestes du groupe A2 du corpus MSR Action 3D pour un dictionnaire de 39 poses.

Tableau VII.2 Matrice de confusion par trame pour les gestes du groupe A2 du corpus MSR Action 3D pour ϱ=3,0 (dictionnaire de 39 poses).

Matrice de confusion	dessiner un cercle	cocher une case	dessiner une croix	coup de pied vers l'avant	attraper d'une main	signe de la main	boxer sur le côté	signe avec les deux mains
dessiner un cercle	**40,4**	10,7	11,1	21,3	4,2	11,4	1,0	0,0
cocher une case	**31,4**	14,4	12,3	22,0	5,6	13,1	1,2	0,0
dessiner une croix	**28,9**	8,6	19,9	24,4	5,5	10,7	2,0	0,0
coup de pied vers l'avant	1,7	0,0	0,2	**79,8**	0,2	0,0	18,1	0,0
attraper d'une main	16,3	4,8	5,8	23,3	**30,4**	15,5	3,7	0,2
signe de la main	16,4	9,6	2,9	21,7	2,9	**37,0**	9,4	0,1
boxer sur le côté	4,0	1,3	3,8	24,1	13,3	6,0	**47,3**	0,1
signe avec les deux mains	4,6	0,0	5,2	20,5	0,6	0,0	4,7	**64,4**

Avec le sous-ensemble A2, deux classes obtiennent des scores supérieurs à 90% : *coup de pied vers l'avant* et *signe avec les deux mains*. Par ailleurs, les catégories *dessiner un cercle*, *signe de la main* et *boxer sur le côté* voient leurs scores au-dessus de 50%, pour un taux de reconnaissance moyen de 53.4% (Figure VII.6).

Pour ce qui est des taux reconnaissance par trame gestuelle (Tableau VII.2), les confusions les plus fortes ont une fois de plus lieu entre des gestes très proches du point de vue de leur réalisation effective, à savoir *dessiner un cercle*, *cocher une case* et *dessiner une croix*. Ces confusions à un niveau local (*e.g.*, par trame) se répercutent nécessairement lors du calcul des taux de reconnaissance globaux $TR^G(1)$, $TR^G(2)$ et $TR^G(3)$, et peuvent expliquer les mauvais scores et fluctuations obtenus notamment pour les classes *cocher une case* et *dessiner une croix*, pour lesquelles les taux de reconnaissance obtenus sont respectivement égaux à 10.0 et 18.5%.

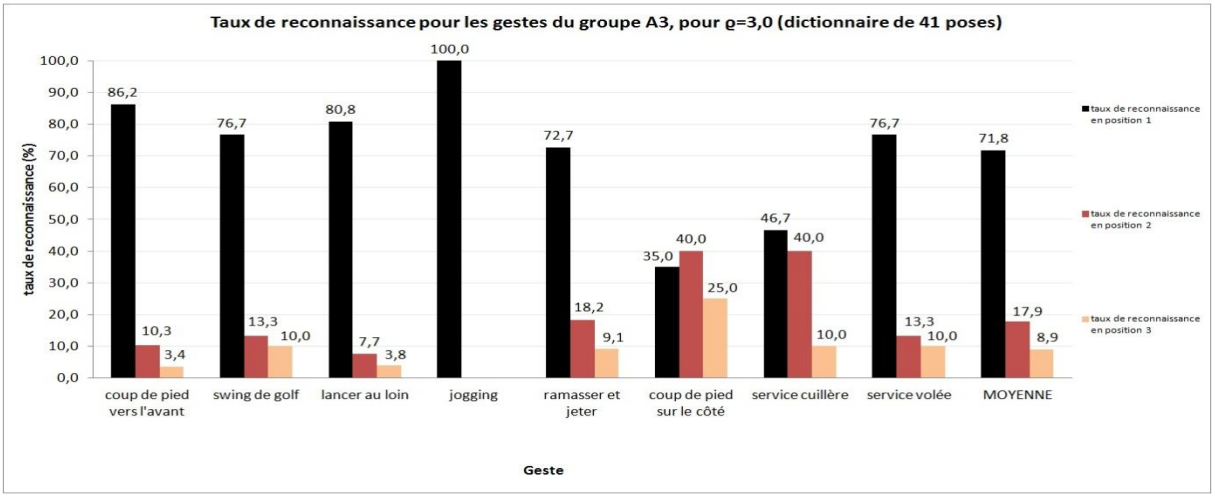


Figure VII.7 Taux de reconnaissance par classe pour les gestes du groupe A3 du corpus MSR Action 3D pour un dictionnaire de 41 poses.

Tableau VII.3 Matrice de confusion par trame pour les gestes du groupe A3 du corpus MSR Action 3D pour ϱ=3,0 (dictionnaire de 41 poses).

Matrice de confusion	coup de pied vers l'avant	swing de golf	lancer au loin	jogging	ramasser et jeter	coup de pied sur le côté	service cuillère	service volée
coup de pied vers l'avant	**73,9**	0,6	0,0	0,2	4,2	14,3	0,0	6,8
swing de golf	16,6	**47,9**	0,1	9,1	7,1	9,6	4,9	4,7
lancer au loin	22,7	2,6	**46,1**	1,0	0,1	5,3	4,2	18,0
jogging	22,1	1,4	0,1	**62,4**	0,4	11,3	0,2	2,1
ramasser et jeter	15,3	5,9	21,0	2,8	**34,5**	2,6	8,3	9,4
coup de pied sur le côté	**43,5**	0,0	0,0	0,8	0,4	41,2	0,0	14,0
service cuillère	19,8	6,4	4,1	6,4	4,8	14,2	**34,9**	9,4
service volée	19,8	5,7	13,1	0,6	1,0	14,2	3,2	**42,4**

Pour finir, le taux de reconnaissance moyen obtenu sur le sous-ensemble de gestes A3 est de 71.8% (Figure VII.7). Six des huit classes du sous-ensemble ont des scores supérieurs à 72%, dont trois supérieurs à 80% qui correspondent aux classes *coup de pied vers l'avant*, *lancer au loin* et *jogging* (qui atteint 100%). Le cumul des taux de reconnaissance aux deux premières positions $TR_{cum}^{G}(2)$ est pour chaque geste supérieur à 75% (pour une moyenne de 89.7%). Un même cumul sur les trois premières positions $TR_{cum}^{G}(3)$ fournit des taux tous supérieurs à 92.3% (pour une moyenne 98.6% et six catégories sur huit à 100%).

Les taux reconnaissance par trame gestuelle (Tableau VII.3) indiquent une confusion entre les classes *coup de pied vers l'avant* et *coup de pied sur le côté*, confusion qui s'explique une fois de plus par une proximité en termes de structuration du mouvement.

De façon générale, les taux de reconnaissance obtenus pour le sous-ensemble A3 (moyenne : 71.8%) sont nettement supérieurs à ceux obtenus sur les sous-ensembles A1 et A2 (moyennes respectives : 59.9 et 53.4%). Nous avons expliqué plus haut que les confusions entre les classes des sous-ensembles A1 et A2 (exprimées dans les Tableau VII.1 et Tableau VII.2) s'expliquaient par les fortes proximités structurales entre les gestes. Rappelons que par construction, les sous-ensembles A1 et A2 (*cf.* section V.1) étaient justement dédiés au regroupement de mouvements similaires (dans le cas de A1 : *coup de poing*, *coup de marteau*, *signe horizontal de la main* ; dans le cas de A2 : *dessiner un cercle*, *cocher une case*, *dessiner une croix*). Le sous-corpus A3 était quant à lui réputé se composer d'actions plus complexes (exemple : *swing de golf*, *ramasser et jeter*). Il semble que notre approche, associant des descripteurs de trame avec des Chaînes de Markov Cachées, se soit montrée plus apte à saisir la structuration de tels mouvements complexe.

En dépit des bons scores obtenus pour certaines catégories, les performances de reconnaissance restent légèrement inférieures à ceux qui ont été obtenus pour les sous-ensembles concernés par Li *et al.* [**38**], Xia *et al.* [**37**], Yang *et al.* [**140**] ou encore Zhu *et al.* [**141**], où les auteurs profitent de la précision avec laquelle la technologie fournit la trajectoire des articulations pour bâtir des descripteurs exclusivement dédiés à la structuration précise du mouvement dans l'espace, ou aux relations qu'entretiennent les parties du corps entre elles au cours du geste. On notera qu'hormis la méthode de Xia *et al.* [**37**] qui fait usage de HMM, les autres approches citées utilisent des méthodes globales (Naïve-Bayes-Nearest-Neighbor classifier, Random Forests, …) pour reconnaître les actions du corpus en fonction de leur caractérisation sur l'entièreté de leur durée de réalisation.

VII.3.2. Microsoft Gesture dataset : MSRC-12

Le MSRC-12 dataset (*cf.* section V.1) est notre deuxième corpus de test retenu pour l'évaluation de notre méthode de reconnaissance dynamique du geste. Au contraire de la base de mouvement précédemment utilisée, nous avons constaté que le corpus MRSC-12 présentait une certaine homogénéité visuelle intra-classe, c'est-à-dire une faible variabilité dans la réalisation des gestes d'une même catégorie d'action. Dans ce cas, nous avons utilisé un paramètre $K = 5$, pour déterminer les poses de références des catégories individuelles.

Nous présentons les résultats obtenus pour

1. l'ensemble des gestes iconiques ;
2. l'ensemble des gestes métaphoriques.

Pour ces sous-ensembles de gestes iconiques et métaphoriques pris séparément, les meilleurs résultats sont obtenus pour la taille maximale de dictionnaire atteinte (*e.g.*, pour ϱ = 0.5), qui dans les deux cas correspond à l'entièreté des poses issues des dictionnaires de classe. En effet, pour l'un ou l'autre des sous-ensembles de gestes (*e.g.*, gestes iconiques ou gestes métaphoriques), à raison de 5 poses pour chacune des 6 catégorie impliquées, la totalité des poses de chacun des dictionnaires de classe résulte en un ensemble global de 30 poses. L'utilisation d'une valeur faible (*i.e.*, en dessous de 0.5) du paramètre de fusion ϱ ne permet dans ce cas aucune fusion.

Les résultats obtenus sont présentés Figure VII.8, Figure VII.9, Figure VII.10 et Figure VII.11 (respectivement sur les Figure VII.12, Figure VII.13, Figure VII.14 et Figure VII.15) dans le cas des gestes iconiques (respectivement métaphoriques).

Pour un dictionnaire de 30 poses, on constate que les gestes iconiques ont des taux de reconnaissance supérieurs à 80%, pour un taux moyen de 88.6% (Figure VII.11), ce qui positionne nos résultats au même rang que ceux obtenus dans [**130**] [**131**] [**134**] [**138**] sur le même corpus de gestes. Néanmoins au contraire de notre système de reconnaissance en temps réel, les approches citées utilisent des algorithmes d'apprentissage supervisés non-dynamiques (SVM, Hidden Conditional Random Fields, Random Decision Forests, méthodes de régression basées sur des descripteurs de corrélations entre les mouvements des articulations…).

Tableau VII.4 Taille *M* du dictionnaire global en fonction du seuil de fusion ϱ pour les gestes iconiques du corpus MSRC-12, à raison de 5 poses par classe.

seuil ϱ	**Taille *M* du dictionnaire global**
2.0	24
1.5	25
1.0	26
0.5	30

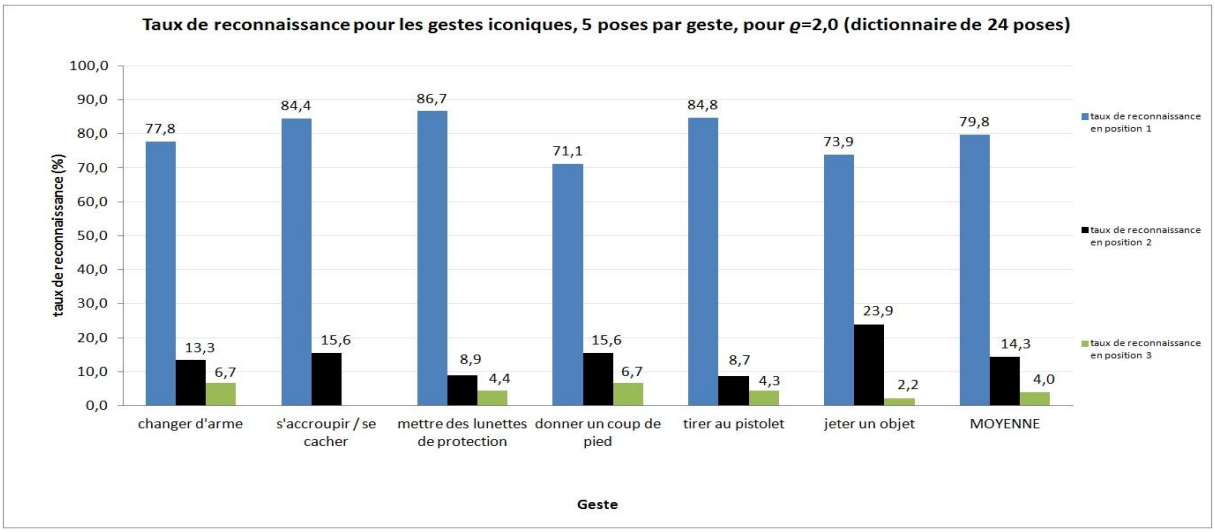


Figure VII.8 Taux de reconnaissance par classe pour les gestes iconiques du corpus MSRC-12 pour un dictionnaire de 24 poses.

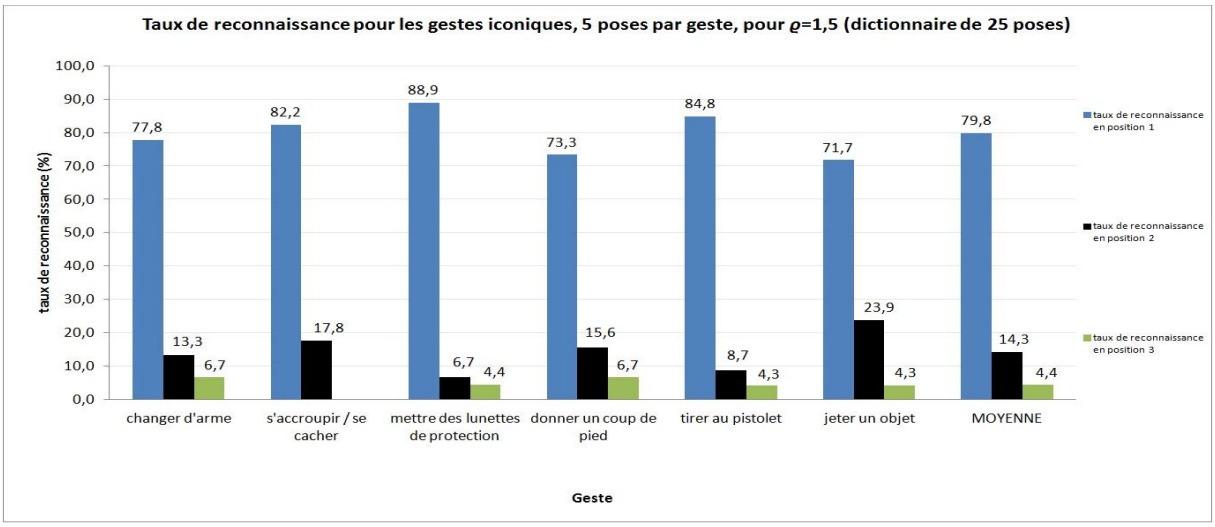


Figure VII.9 Taux de reconnaissance par classe pour les gestes iconiques du corpus MSRC-12 pour un dictionnaire de 25 poses.

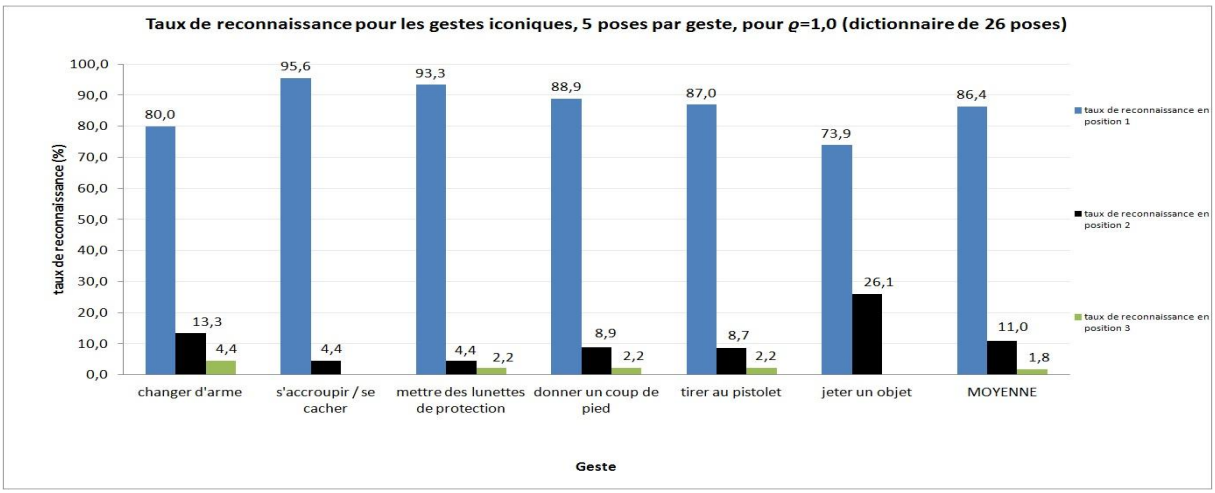


Figure VII.10 Taux de reconnaissance par classe pour les gestes iconiques du corpus MSRC-12 pour un dictionnaire de 26 poses.

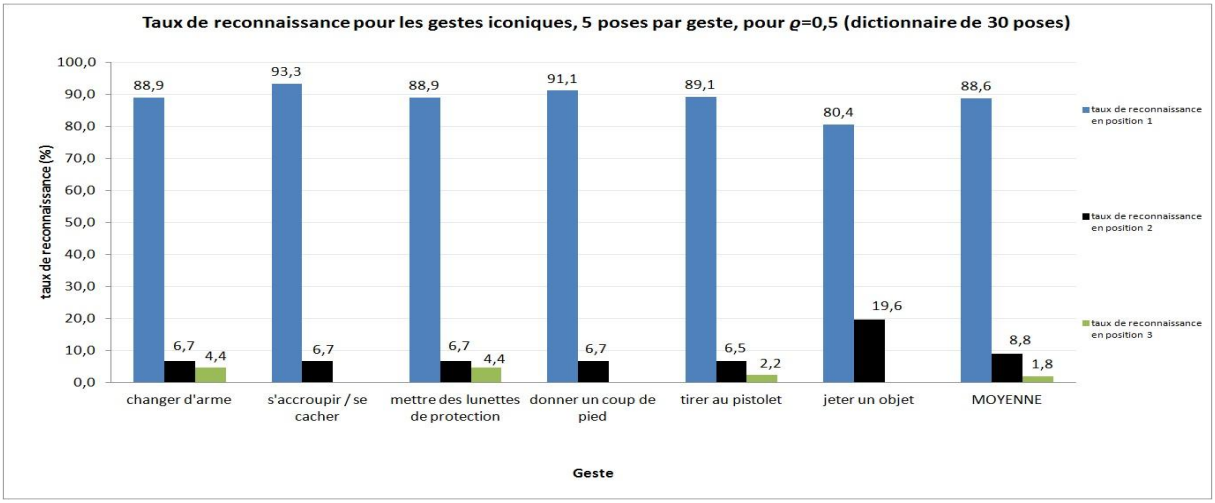


Figure VII.11 Taux de reconnaissance par classe pour les gestes iconiques du corpus MSRC-12 pour un dictionnaire de 30 poses.

C'est pour cette taille de dictionnaire que les catégories de gestes iconiques voient leur score maximisé, à l'exception des catégories *s'accroupir/se cacher* et *mettre des lunettes de protection*. Toujours est-il que pour ces deux dernières classes, le score reste relativement élevé (93.3 et 88.9% respectivement). Le taux de reconnaissance pour la classe *donner un coup de pied* connaît un saut de 20% entre ce qu'il est pour le dictionnaire le plus petit (24 poses, Figure VII.8) et ce qu'il devient pour le dictionnaire le plus grand (30 poses, Figure VII.11). Les taux des catégories *changer d'arme* et *s'accroupir/se cacher* connaissent quant à eux des accroissements respectifs de 11,1 et 8.9%.

Tableau VII.5 Matrice de confusion par trame pour les gestes iconiques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=2,0 (dictionnaire de 24 poses).

Matrice de confusion	changer d'arme	s'accroupir / se cacher	mettre des lunettes de protection	donner un coup de pied	tirer au pistolet	jeter un objet
changer d'arme	**61,7**	0,0	6,2	19,6	5,8	6,7
s'accroupir / se cacher	3,0	**56,0**	0,2	37,4	0,0	3,4
mettre des lunettes de protection	3,1	0,0	**57,1**	31,0	7,9	0,8
donner un coup de pied	13,4	4,1	1,4	**49,7**	1,1	30,2
tirer au pistolet	5,7	0,1	11,4	27,6	**49,4**	5,8
jeter un objet	16,7	0,4	0,1	28,0	0,5	**54,3**

Tableau VII.6 Matrice de confusion par trame pour les gestes iconiques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=1,5 (dictionnaire de 25 poses).

Matrice de confusion	changer d'arme	s'accroupir / se cacher	mettre des lunettes de protection	donner un coup de pied	tirer au pistolet	jeter un objet
changer d'arme	**61,7**	0,1	6,3	19,8	5,3	6,8
s'accroupir / se cacher	2,7	**56,0**	0,2	37,9	0,1	3,0
mettre des lunettes de protection	3,1	0,0	**62,2**	31,1	2,7	1,0
donner un coup de pied	13,1	3,6	1,8	**50,9**	0,8	29,8
tirer au pistolet	5,6	0,1	10,6	27,7	**49,7**	6,2
jeter un objet	16,8	0,3	0,2	28,8	0,4	**53,5**

Tableau VII.7 Matrice de confusion par trame pour les gestes iconiques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=1,0 (dictionnaire de 26 poses).

Matrice de confusion	changer d'arme	s'accroupir / se cacher	mettre des lunettes de protection	donner un coup de pied	tirer au pistolet	jeter un objet
changer d'arme	**62,7**	5,0	7,3	14,8	5,3	4,9
s'accroupir / se cacher	2,5	**69,8**	2,4	23,1	0,0	2,2
mettre des lunettes de protection	3,0	10,9	**68,9**	13,3	3,5	0,4
donner un coup de pied	7,6	5,8	4,6	**67,9**	1,4	12,7
tirer au pistolet	5,9	6,3	10,9	19,8	**52,6**	4,6
jeter un objet	10,5	3,5	0,5	28,3	0,8	**56,3**

Tableau VII.8 Matrice de confusion par trame pour les gestes iconiques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=0,5 (dictionnaire de 30 poses).

Matrice de confusion	changer d'arme	s'accroupir / se cacher	mettre des lunettes de protection	donner un coup de pied	tirer au pistolet	jeter un objet
changer d'arme	**66,1**	8,5	5,1	10,3	7,0	3,0
s'accroupir / se cacher	2,4	**77,2**	2,3	15,5	0,2	2,4
mettre des lunettes de protection	3,3	15,8	**64,1**	9,8	6,6	0,4
donner un coup de pied	8,7	11,3	2,5	**65,2**	2,7	9,6
tirer au pistolet	5,8	10,5	10,0	14,9	**54,5**	4,3
jeter un objet	8,2	10,8	0,0	20,4	1,2	**59,4**

Les tableaux présentant les matrices de confusion par trame (Tableau VII.5, Tableau VII.6, Tableau VII.7 et Tableau VII.8) en fonction du seuil de fusion ϱ confirment la tendance à l'amélioration des performances de reconnaissance à mesure que la taille du dictionnaire global grandit, avec un bond moyen en termes de taux reconnaissance par trame de 9.7%, et même un saut supérieur à 21% dans le cas de la catégorie *s'accroupir/se cacher*.

Pour ce qui est de l'étude réservée aux gestes métaphoriques du corpus (Figure VII.12, Figure VII.13, Figure VII.14 et Figure VII.15), la performance est également fonction croissante de la taille du dictionnaire. Si dans le meilleur des cas (*e.g.*, dictionnaire de 30 poses, Figure VII.15), les taux sont globalement inférieurs à ce qu'ils sont pour les gestes iconiques (taux de reconnaissance moyen de 75.2%), 3 catégories dépassent tout de même les 86.7% : *s'incliner pour clore la session musicale*, *protester contre la musique* et *naviguer vers le menu suivant* obtiennent des scores respectifs égaux à 95.6, 86.7 et 95.7% (Figure VII.15). Le dictionnaire « maximal » (*i.e.*, 30 poses) maximise les taux de reconnaissance de tous les gestes métaphoriques, avec des gains atteignant 24.5% et même 57.4% pour les classes *s'incliner pour clore la session musicale* et *naviguer vers le menu suivant*. En revanche, le score reste faible pour le geste *accélérer le tempo*, avec un taux restant inférieur à 30%. On notera enfin que les taux de reconnaissances cumulés moyens aux deux premières positions $TR_{cum}^{moyen}(2)$ et trois premières positions $TR_{cum}^{moyen}(3)$ sont respectivement de 92.5 et 96.6%.

Les tableaux présentant les matrice de confusion par trame (Tableau VII.10, Tableau VII.11, Tableau VII.12 et Tableau VII.13) confirment la tendance à l'amélioration des performances de reconnaissance à mesure que la taille du dictionnaire global grandit, avec un gain moyen en termes de taux reconnaissance par trame de 9.7%, et même un saut supérieur à 21% dans le cas de la catégorie *s'accroupir/se cacher*. L'accroissement de la taille de dictionnaire tend à faire disparaître les confusions pouvant exister entre certaines catégories. Les taux de reconnaissance par trame pour les gestes *s'incliner pour clore la session musicale* et *naviguer vers le menu suivant* augmentent même 23.6 et 35.6%.

Tableau VII.9 Taille M du dictionnaire global en fonction du seuil de fusion ϱ pour les gestes métaphoriques du corpus MSRC-12, à raison de 5 poses par classe.

seuil ϱ	**Taille M du dictionnaire global**
2.0	19
1.5	21
1.0	24
0.5	30

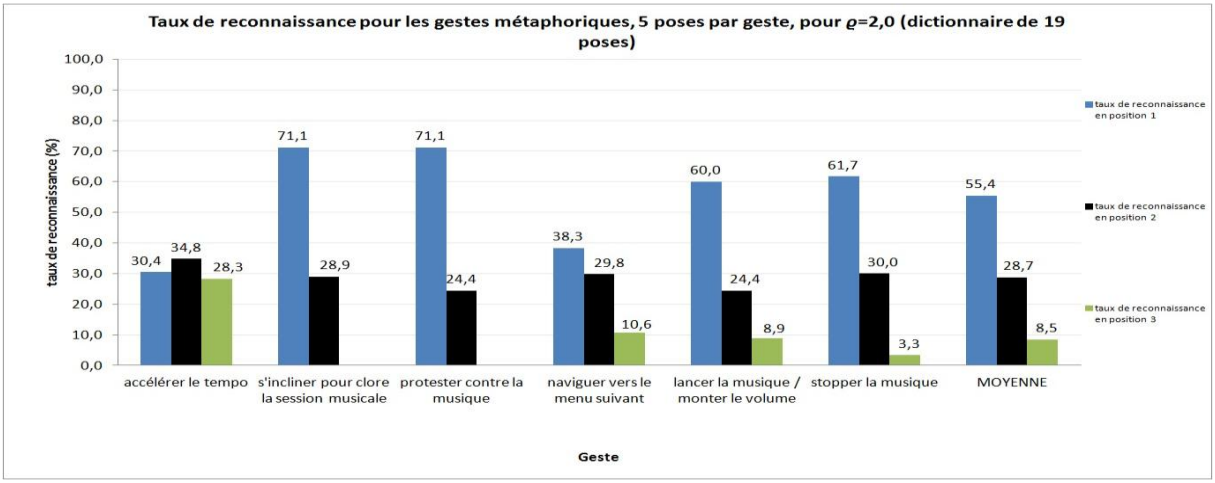


Figure VII.12 Taux de reconnaissance par classe pour les gestes métaphoriques du corpus MSRC-12 pour un dictionnaire de 19 poses.

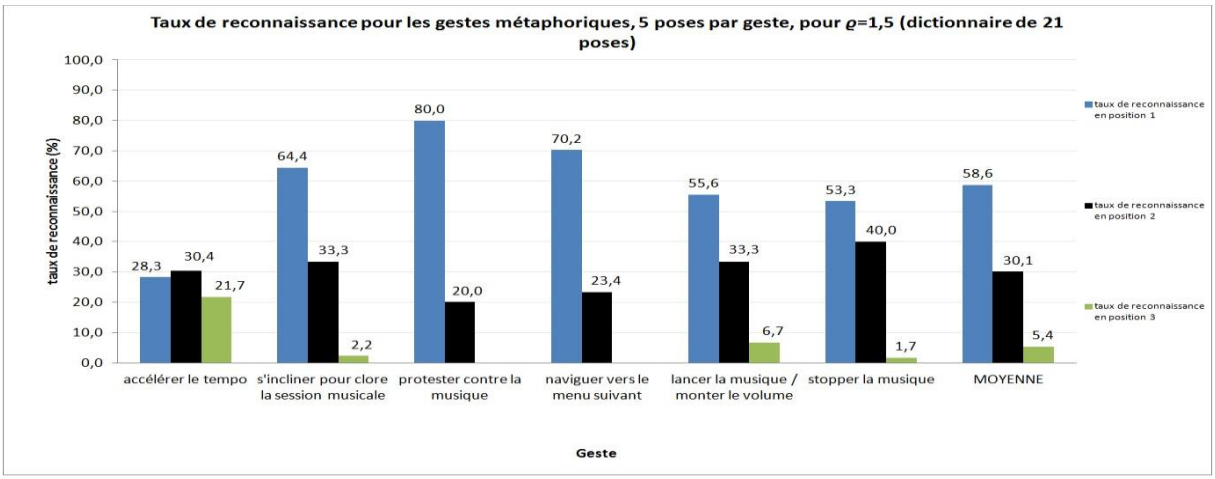


Figure VII.13 Taux de reconnaissance par classe pour les gestes métaphoriques du corpus MSRC-12 pour un dictionnaire de 21 poses.

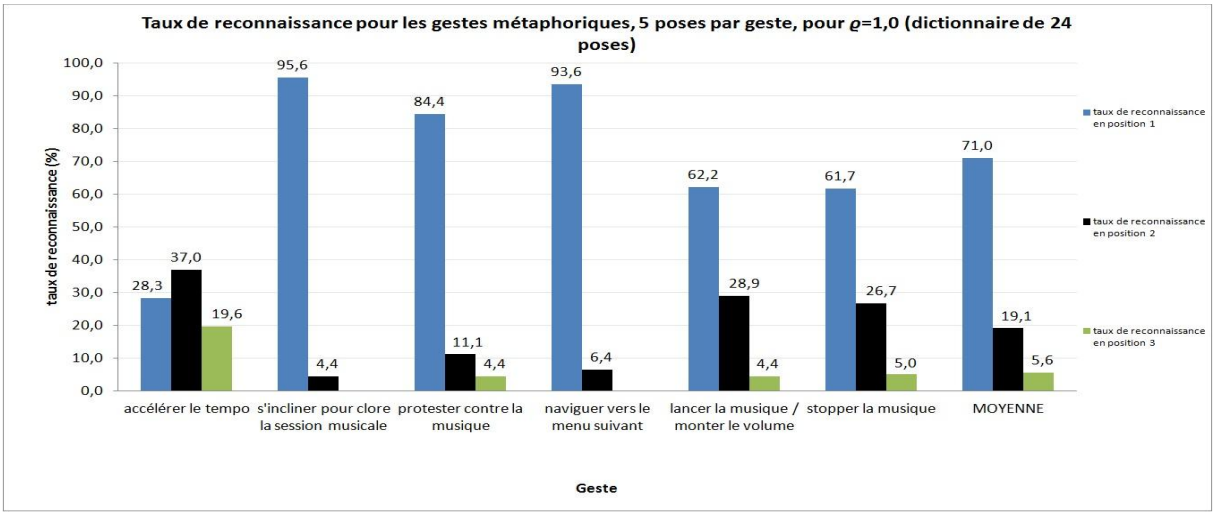


Figure VII.14 Taux de reconnaissance par classe pour les gestes métaphoriques du corpus MSRC-12 pour un dictionnaire de 24 poses.

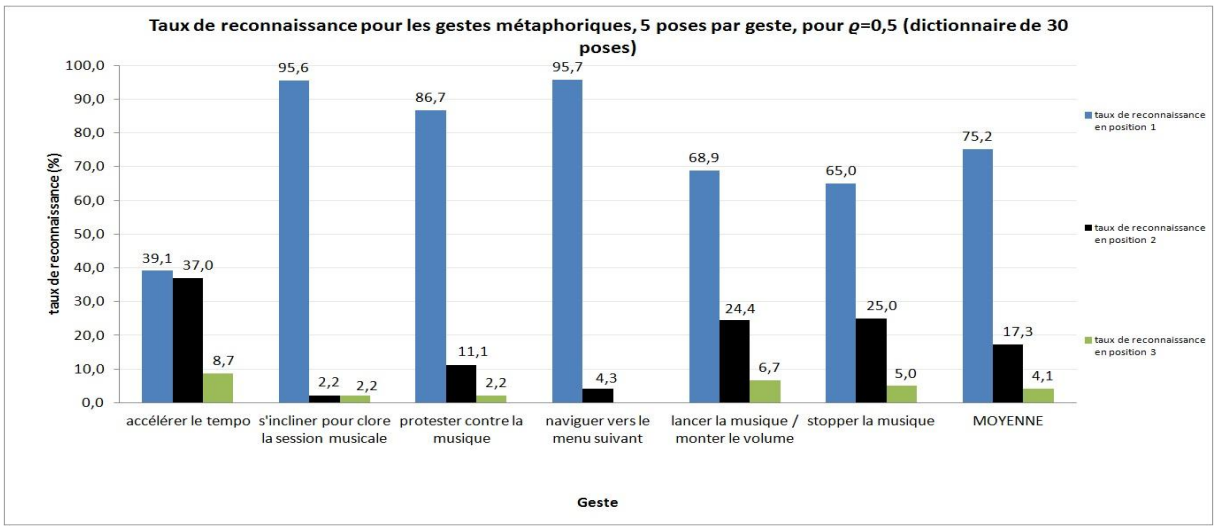


Figure VII.15 Taux de reconnaissance par classe pour les gestes métaphoriques du corpus MSRC-12 pour un dictionnaire de 30 poses.

Tableau VII.10 Matrice de confusion par trame pour les gestes métaphoriques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=2,0 (dictionnaire de 19 poses).

Matrice de confusion	accélérer le tempo	s'incliner pour clore la session musicale	protester contre la musique	naviguer vers le menu suivant	lancer la musique / monter le volume	stopper la musique
accélérer le tempo	**27,9**	1,5	11,8	22,7	19,7	16,4
s'incliner pour clore la session musicale	0,1	**53,8**	0,0	39,3	0,6	6,2
protester contre la musique	14,8	1,5	**35,6**	29,1	6,1	12,9
naviguer vers le menu suivant	0,2	14,9	0,0	32,7	13,1	**39,1**
lancer la musique / monter le volume	16,4	0,6	0,3	**35,5**	33,5	13,9
stopper la musique	15,2	1,8	1,7	28,9	6,9	**45,5**

Tableau VII.11 Matrice de confusion par trame pour les gestes métaphoriques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=1,5 (dictionnaire de 21 poses).

Matrice de confusion	accélérer le tempo	s'incliner pour clore la session musicale	protester contre la musique	naviguer vers le menu suivant	lancer la musique / monter le volume	stopper la musique
accélérer le tempo	23,8	1,1	15,4	**24,2**	19,9	15,6
s'incliner pour clore la session musicale	0,0	**51,7**	0,0	42,9	1,1	4,2
protester contre la musique	9,4	1,0	**40,1**	30,4	6,8	12,2
naviguer vers le menu suivant	0,2	9,8	0,0	**52,1**	8,8	29,1
lancer la musique / monter le volume	14,6	0,3	1,5	**37,4**	34,6	11,6
stopper la musique	19,2	1,5	2,2	32,7	5,6	**38,8**

Tableau VII.12 Matrice de confusion par trame pour les gestes métaphoriques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=1,0 (dictionnaire de 24 poses).

Matrice de confusion	accélérer le tempo	s'incliner pour clore la session musicale	protester contre la musique	naviguer vers le menu suivant	lancer la musique / monter le volume	stopper la musique
accélérer le tempo	**25,8**	17,3	11,2	9,8	20,9	15,0
s'incliner pour clore la session musicale	1,1	**76,4**	0,8	16,8	2,2	2,7
protester contre la musique	10,0	24,4	**38,5**	10,1	6,6	10,4
naviguer vers le menu suivant	0,4	19,6	0,0	**71,4**	1,4	7,2
lancer la musique / monter le volume	10,7	29,6	0,5	9,8	**36,6**	12,8
stopper la musique	14,6	16,5	2,3	21,9	5,7	**39,0**

Tableau VII.13 Matrice de confusion par trame pour les gestes métaphoriques du corpus MSRC-12 à raison de 5 poses par geste, pour ϱ=0,5 (dictionnaire de 30 poses).

Matrice de confusion	accélérer le tempo	s'incliner pour clore la session musicale	protester contre la musique	naviguer vers le menu suivant	lancer la musique / monter le volume	stopper la musique
accélérer le tempo	**27,2**	19,9	11,0	7,1	21,4	13,3
s'incliner pour clore la session musicale	1,1	**77,4**	1,1	14,8	1,9	3,7
protester contre la musique	10,3	26,5	**41,1**	8,3	6,4	7,4
naviguer vers le menu suivant	0,0	22,2	0,0	**68,3**	0,5	9,0
lancer la musique / monter le volume	8,6	31,7	0,9	8,0	**40,0**	10,8
stopper la musique	13,1	22,2	2,3	15,9	6,3	**40,2**

VII.3.3. Résultats sur le corpus UTKinect-Human Detection

Nous avons également testé notre approche de reconnaissance dynamique sur le corpus UTKinect-Human Detection (*cf.* section V.1), avec K = 10 poses représentatives par classe. Le Tableau VII.14 fournit la taille M du dictionnaire global en fonction du seuil de fusion ϱ, et les Figure VII.16, Figure VII.17, Figure VII.18 et Figure VII.19 présentent les taux de reconnaissance $\{TR^G(i)\}_{\forall\, i\in\{1,2,3\}}$ (équation VII.62) obtenus en fonction des valeurs prises par ce seuil.

Ici encore, on note une amélioration des résultats à mesure que l'on augmente la taille du dictionnaire, sauf dans l'unique cas de la catégorie *s'asseoir* qui obtient son meilleur taux de reconnaissance pour un nombre de poses minimal (ϱ = 4.5, M = 19). Globalement, les meilleurs résultats sont obtenus pour ϱ = 3.0 et un total de 39 poses de référence, avec un taux de reconnaissance moyen en première position $TR^{moyen}(1)$ égal à 86.5% (Figure VII.19). Les taux de reconnaissance cumulatifs moyens aux deuxième et troisième ordres sont respectivement égaux à 97.5% et 99.5%. Hormis les classes d'actions *pousser* et *s'asseoir*, les catégories gestuelles obtiennent des taux de reconnaissance supérieurs à 80%, voire supérieurs à 95% pour une petite moitié d'entre elles. De tels résultats sont comparables à ceux de Xia *et al.* [**37**] pour qui le taux de reconnaissance moyen atteint 90.9% et cinq catégories dépassent les 96,5%, ou encore les performances de Zhu *et al.* [**141**] où les différentes approches testées donnent des taux de reconnaissance moyens variant entre 80.8 et 91.9%. Une fois de plus, on soulignera qu'au contraire de cette dernière approche, notre démarche permet de discriminer les actions en temps réel.

Tableau VII.14 Taille M du dictionnaire global en fonction du seuil de fusion ϱ pour le corpus UTKinect-Human Detection dataset, à raison de 10 poses par classe.

seuil ϱ	**Taille M du dictionnaire global**
4.5	19
4.0	24
3.5	32
3.0	39

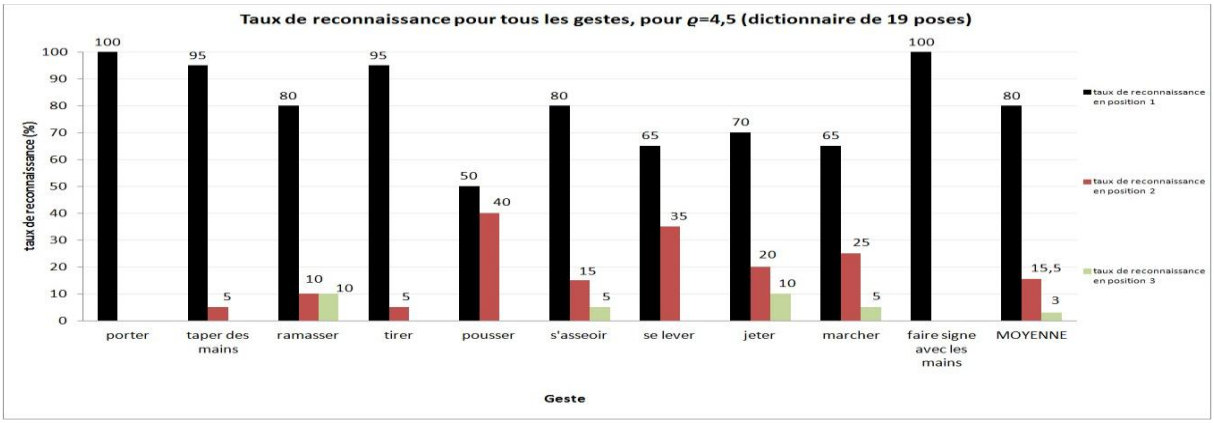


Figure VII.16 Taux de reconnaissance par classe pour le corpus UTKinect-HumanDetection pour un dictionnaire de 19 poses.

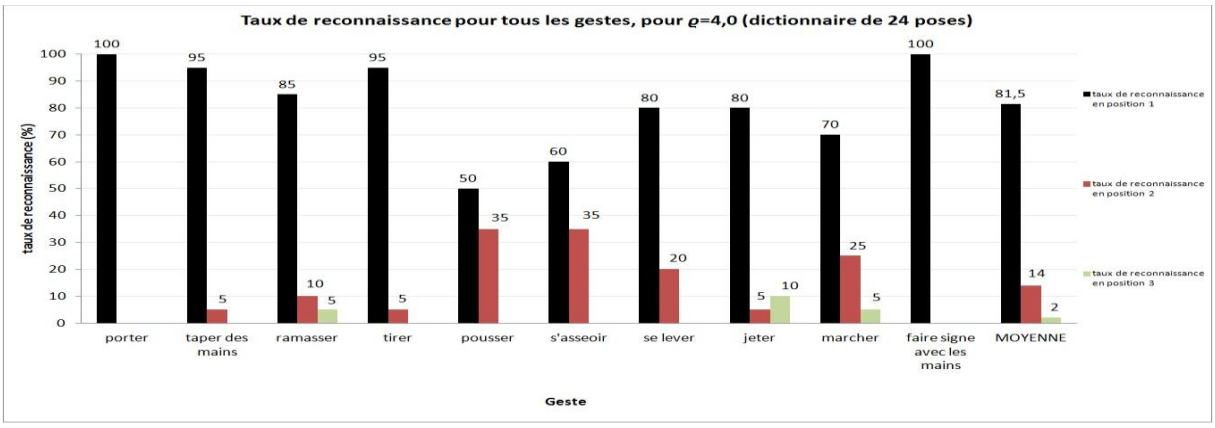


Figure VII.17 Taux de reconnaissance par classe pour le corpus UTKinect-HumanDetection pour un dictionnaire de 24 poses.

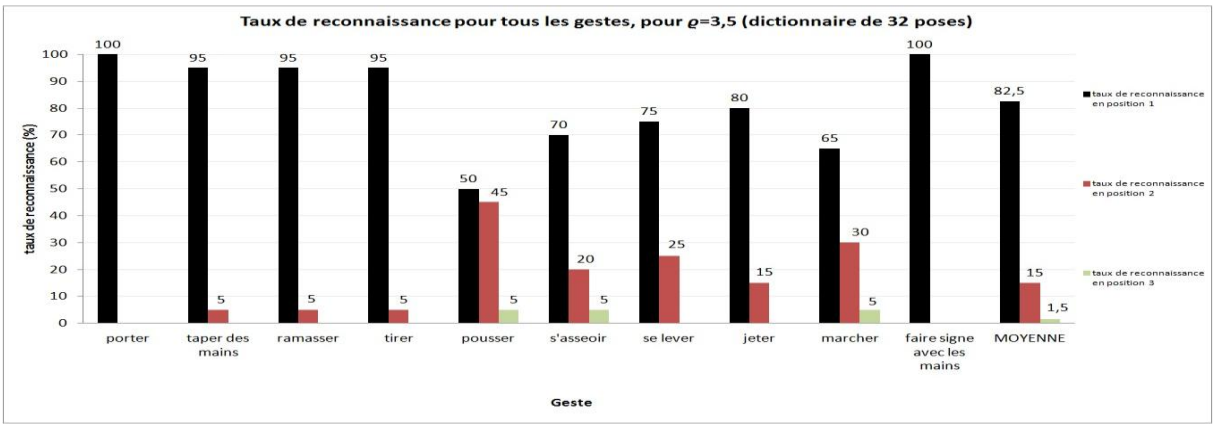


Figure VII.18 Taux de reconnaissance par classe pour le corpus UTKinect-HumanDetection pour un dictionnaire de 32 poses.

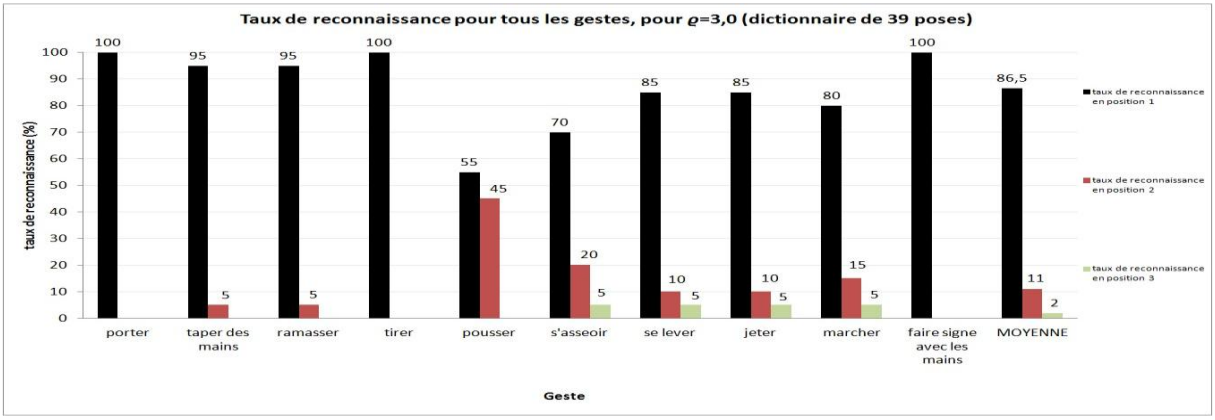


Figure VII.19 Taux de reconnaissance par classe pour le corpus UTKinect-HumanDetection pour un dictionnaire de 39 poses.

Les résultats qu'expriment les matrices de confusion dédiées aux taux de reconnaissance par trame gestuelle confirment cette tendance, et sont présentés dans les Tableau VII.15, Tableau VII.16, Tableau VII.17 et Tableau VII.18 pour les différentes valeurs de ϱ. Le meilleur taux de reconnaissance par trame est obtenu pour le plus grand dictionnaire testé (ϱ = 3.0, M = 39) pour toutes les classes, hormis le geste

s'asseoir (51.0% pour ϱ = 4.5). Ainsi, les catégories du UTKinect-HumanDetection dataset sont toutes convenablement reconnues.

Tableau VII.15 Matrice de confusion par trame pour le corpus UTKinect-HumanDetection pour ϱ=4,5 (dictionnaire de 19 poses).

Matrice de confusion	porter	taper des mains	ramasser	tirer	pousser	s'asseoir	se lever	jeter	marcher	faire signe avec les mains
porter	**78,3**	0,8	5,3	0,5	0,2	4,0	0,7	3,8	6,3	0,0
taper des mains	7,2	**82,5**	1,0	0,0	0,0	0,1	0,0	2,8	0,9	5,4
ramasser	9,1	0,0	**50,5**	0,0	0,0	18,0	11,5	1,1	9,5	0,3
tirer	0,2	0,6	0,0	**84,8**	12,7	0,0	0,0	1,6	0,0	0,1
pousser	0,0	0,2	0,0	**47,7**	46,5	0,0	0,0	4,5	0,9	0,2
s'asseoir	22,0	0,0	10,8	0,0	0,0	**51,0**	8,9	1,3	6,0	0,0
se lever	14,6	0,0	3,1	0,0	0,0	25,0	**51,9**	3,2	2,2	0,0
jeter	2,7	11,2	0,0	8,2	13,2	1,0	2,0	**58,7**	3,1	0,0
marcher	30,1	0,1	8,1	0,0	1,4	6,4	4,3	1,9	**46,4**	1,3
faire signe avec les mains	0,0	1,1	1,7	0,0	0,3	0,0	0,5	6,5	3,7	**86,1**

Tableau VII.16 Matrice de confusion par trame pour le corpus UTKinect-HumanDetection pour ϱ=4,0 (dictionnaire de 24 poses).

Matrice de confusion	porter	taper des mains	ramasser	tirer	pousser	s'asseoir	se lever	jeter	marcher	faire signe avec les mains
porter	**81,2**	0,8	4,0	0,0	0,1	4,4	0,6	3,5	5,5	0,0
taper des mains	4,9	**85,1**	0,6	0,0	0,0	0,9	0,0	2,6	0,7	5,1
ramasser	10,1	0,1	**51,8**	0,0	0,0	16,1	6,9	1,1	13,5	0,2
tirer	0,0	0,0	0,0	**84,8**	12,7	0,0	0,0	2,1	0,4	0,0
pousser	0,0	0,4	0,0	45,4	**48,2**	0,0	0,0	4,9	0,9	0,2
s'asseoir	22,9	0,0	9,6	0,0	0,0	**46,8**	11,2	1,0	8,4	0,1
se lever	15,0	0,0	2,5	0,0	0,0	22,8	**52,6**	2,4	4,6	0,1
jeter	1,9	10,7	0,1	7,8	14,0	2,2	2,2	**58,6**	2,5	0,0
marcher	23,9	0,0	7,6	0,0	1,0	5,4	3,1	2,8	**54,3**	1,8
faire signe avec les mains	0,0	0,9	2,2	0,0	0,3	0,6	0,3	7,3	1,2	**87,2**

Tableau VII.17 Matrice de confusion par trame pour le corpus UTKinect-HumanDetection pour ϱ=3,5 (dictionnaire de 32 poses).

Matrice de confusion	porter	taper des mains	ramasser	tirer	pousser	s'asseoir	se lever	jeter	marcher	faire signe avec les mains
porter	**81,5**	0,4	3,9	0,0	0,4	5,0	0,6	3,2	4,9	0,0
taper des mains	9,1	**81,4**	0,4	0,0	0,6	0,6	0,2	3,2	0,9	3,5
ramasser	9,0	0,1	**51,7**	0,0	0,0	15,9	8,7	1,2	12,7	0,6
tirer	0,0	0,0	0,0	**84,3**	14,2	0,0	0,0	1,0	0,4	0,0
pousser	0,0	0,1	0,0	44,7	**49,8**	0,0	0,0	4,4	0,8	0,2
s'asseoir	22,6	0,0	9,3	0,0	0,0	**48,7**	9,4	1,5	8,4	0,1
se lever	15,3	0,0	2,3	0,0	0,0	24,2	**50,9**	3,2	3,7	0,3
jeter	2,0	9,7	0,0	10,1	12,5	1,9	2,0	**59,4**	2,3	0,0
marcher	24,5	0,0	8,5	0,0	1,0	5,8	3,2	3,5	**52,0**	1,5
faire signe avec les mains	0,0	0,5	1,5	0,0	0,6	0,7	0,1	6,8	1,4	**88,3**

Tableau VII.18 Matrice de confusion par trame pour le corpus UTKinect-HumanDetection pour ϱ=3,0 (dictionnaire de 39 poses).

Matrice de confusion	porter	taper des mains	ramasser	tirer	pousser	s'asseoir	se lever	jeter	marcher	faire signe avec les mains
porter	**83,4**	2,1	3,6	0,0	0,5	4,0	0,6	2,9	3,0	0,0
taper des mains	2,0	**86,4**	0,7	0,0	0,9	0,6	0,3	5,1	0,4	3,3
ramasser	8,7	0,1	**54,0**	0,0	0,0	13,9	8,1	1,1	13,8	0,2
tirer	0,0	0,0	0,0	**86,4**	11,9	0,0	0,0	1,2	0,4	0,0
pousser	0,0	0,1	0,2	38,4	**57,2**	0,0	0,0	3,4	0,7	0,0
s'asseoir	19,4	0,0	10,4	0,0	0,0	**47,8**	11,2	1,0	10,1	0,0
se lever	12,0	0,0	1,8	0,0	0,0	17,9	**58,0**	1,3	8,8	0,3
jeter	1,7	7,3	1,2	8,8	12,7	1,6	1,9	**60,6**	3,9	0,3
marcher	11,8	0,0	9,0	0,0	1,2	5,1	2,4	3,1	**66,2**	1,2
faire signe avec les mains	0,0	0,6	1,6	0,1	0,3	0,6	0,2	5,5	0,2	**91,0**

VII.3.4. Résultats sur le corpus HTI 2014-2015

Pour finir, nous présentons les résultats obtenus pour le corpus HTI 2014-2015 que nous avons expressément construit pour cette expérimentation (*cf.* section V.2.2). Après observation des séquences et de la variabilité des réalisations de chaque type de geste, nous avons décidé de garder K = 10 poses représentatives par classe. Les Figure VII.20 et Figure VII.21 montrent différentes poses retenues dans le corpus pour les gestes respectifs *intercepter un objet* et *se mettre à genoux*. Ces poses sont les configurations de squelette qui correspondent aux poses extraites dans l'espace de nos descripteurs de

Laban. Soulignons que cela est possible grâce à l'utilisation de l'algorithme de *clustering* k-medians, qui fournit des poses de références correspondant aux poses réelles rencontrées dans les séquences gestuelles.

On observe que ces clés du mouvement semblent échantillonner l'espace des poses 3D d'une manière satisfaisante tout en réduisant la variabilité des poses parcourues dans un geste donné.

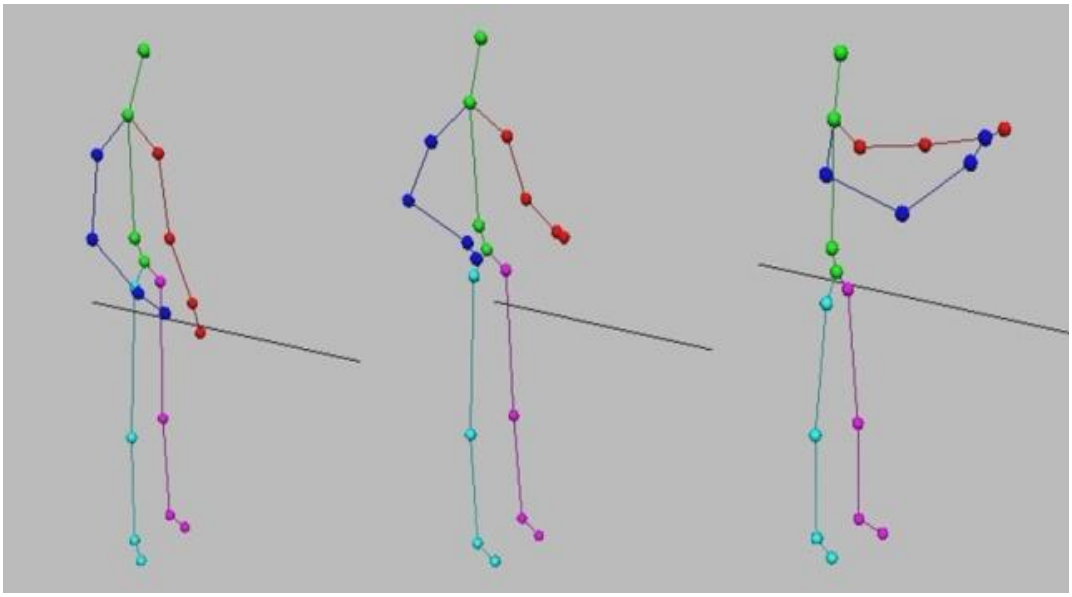


Figure VII.20 Exemples de poses-clés retenus pour la catégorie *intercepter un objet*.

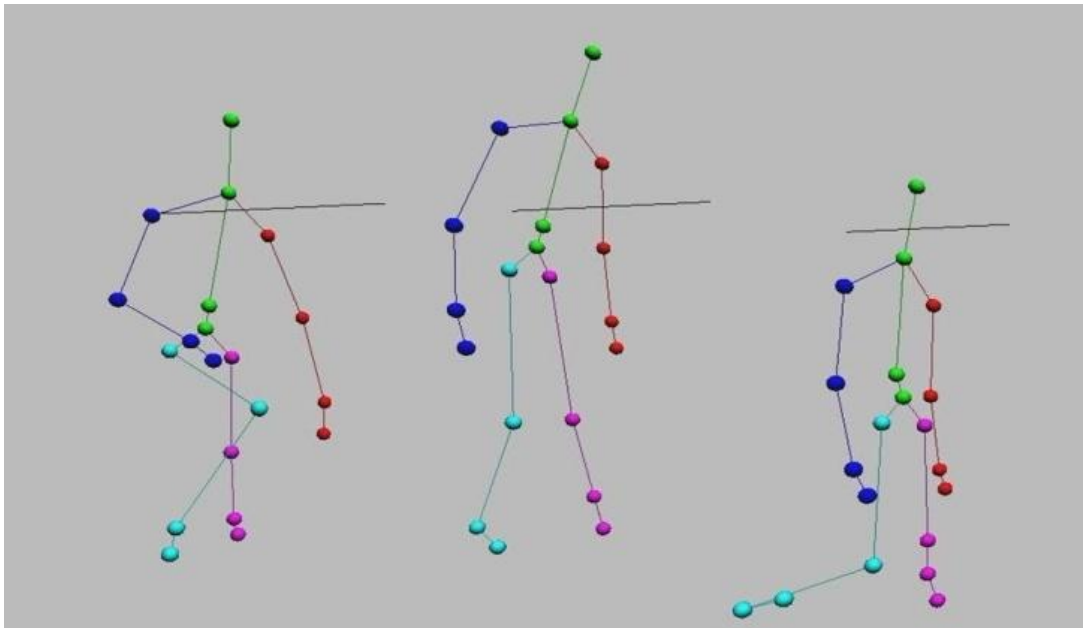


Figure VII.21 Exemples de poses-clés retenus pour la catégorie *se mettre à genoux*.

Le seuil ϱ (*cf.* section VII.2.1) utilisé pour fusionner les dictionnaires obtenus pour chacune des catégories gestuelles a été déterminé expérimentalement. Le Tableau VII.19 fournit la taille M du dictionnaire global que nous avons obtenu en fonction de ce seuil de fusion ϱ.

Pour évaluer l'influence du paramètre ϱ sur les performances de reconnaissance, les Figure VII.22, Figure VII.23 et Figure VII.24 présentent les résultats obtenus en termes de taux de reconnaissance par catégorie d'action, pour les différentes valeurs de ce seuil de fusion ϱ.

Globalement, les résultats sont stables. Une amélioration de la performance est toutefois obtenue lorsque l'on augmente le nombre de poses utilisées. Ainsi, les meilleurs résultats sont donc obtenus un seuil ϱ = 3.5, qui conduit à un nombre de poses-clés égal à 40 (Figure VII.24). Dans ce cas, le taux de reconnaissance moyen en première position $TR^{moyen}(1)$ (équation VII.62) est de 67.6%. Les taux de reconnaissance cumulatifs aux deux premières positions et aux trois premières positions sont présentés dans le Tableau VII.20. En moyenne, la catégorie gestuelle est correctement reconnue aux deux premières (respectivement trois premières) positions dans 89.5% (respectivement 95.2%) des cas. De tels scores démontrent la pertinence de l'approche de reconnaissance dynamique proposée.

Nous pouvons observer que trois gestes obtiennent des taux de reconnaissance $TR^{G}(1)$ supérieurs à 92% : *jongler*, *se boucher les yeux* et *s'étirer*. De très bonnes performances sont également obtenues pour les catégories *faire ses lacets* et *se mettre à genoux*, qui atteignent des scores supérieurs à 82%. Dans les cas de *dire merci en langage des signes*, *faire un tour sur soi-même* et *lancer un objet devant soi*, les taux de reconnaissance sont légèrement inférieurs (75.5%, 57.1% et 63.3%, respectivement). Les taux $TR^{G}(1)$ les plus faibles sont obtenus pour les gestes *faire un cercle avec le bras droit*, *intercepter un objet* et *se frotter les yeux*. Néanmoins, y compris dans le cas de ces classes, la catégorie est correctement retournée dans plus de 52.1% des cas parmi les deux premiers candidats, et dans plus de 64.6% des cas aux trois premières positions.

Les matrices de confusion sont représentées dans les Tableau VII.21, Tableau VII.22 et Tableau VII.23 pour les différentes valeurs du seuil de fusion ϱ. Comme dans le cas des taux de reconnaissance cumulés précédemment introduits (*e.g.*, $\{TR^{G}_{cum}(i)\}_{\forall\, i\in\{1,2,3\}}$, équation VII.65), on remarque que l'agrandissement du dictionnaire tend à améliorer les taux de reconnaissance par trame, hormis pour quatre gestes : *faire un tour sur soi-même*, *intercepter un objet*, *jongler* et *se mettre à genoux*.

Globalement, les meilleurs résultats sont donc obtenus pour le dictionnaire de 40 poses (Tableau VII.23). Une confusion relativement importante concerne les gestes *dire merci en langage des signes* et *lancer un objet devant soi*, qui peut s'expliquer par les similarités entre les mouvements corporels impliqués dans leur réalisation et notamment le mouvement du bras. Une forte confusion se produit également entre *se frotter les yeux* et *se boucher les yeux*, pour des raisons similaires. De même, les erreurs de classification entre *faire un cercle avec le bras droit* et *dire merci en langage des signes* sont sans doute liées au fait que seul le bras droit est utilisé pour ces deux actions. Enfin, nous supposons que les confusions entre *faire un tour sur soi-même* et *dire merci en langage des signes* s'expliquent par la procédure de normalisation des poses qui rend une partie de nos descripteurs insensibles à la position absolue des articulations dans l'espace.

Les matrices de confusion confirment les tendances exprimées par les taux de reconnaissance cumulés (Figure VII.22, Figure VII.23 et Figure VII.24). Néanmoins, dans la majorité des cas, les catégories reconnues sont correctes.

Tableau VII.19 Taille *M* du dictionnaire global en fonction du seuil de fusion ϱ pour le corpus HTI 2014-2015.

seuil ϱ	**Taille *M* du dictionnaire global**
4.5	20
4.0	31
3.5	40

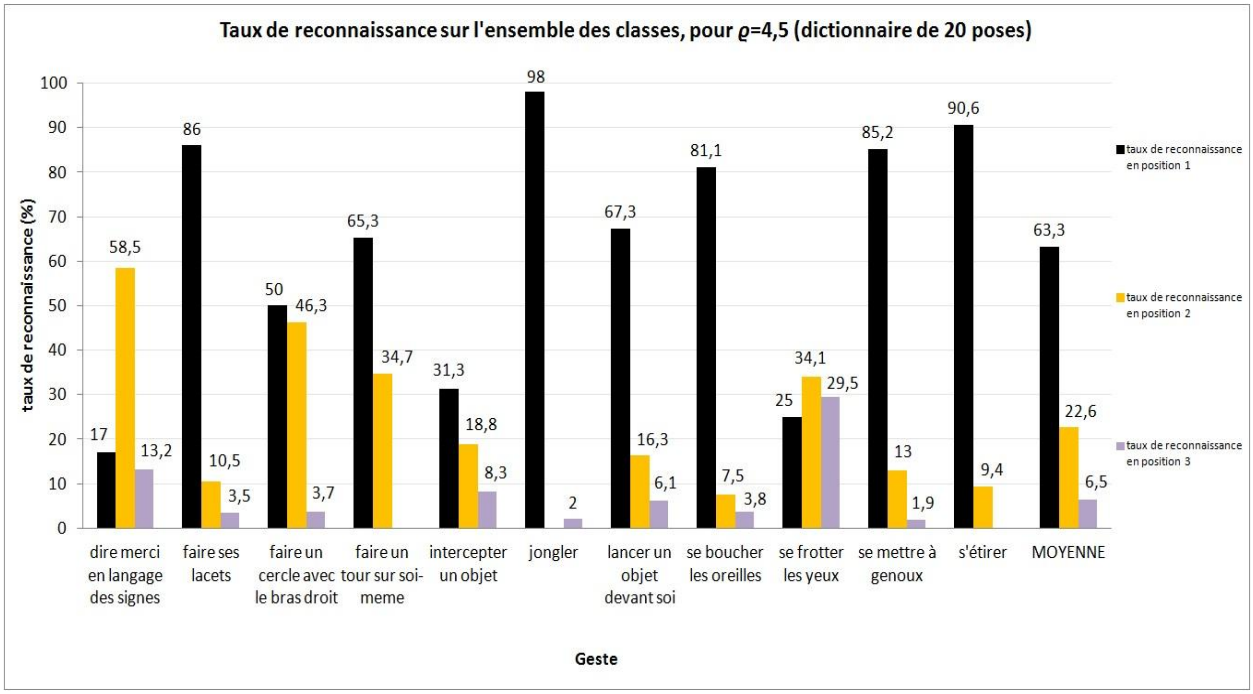


Figure VII.22 Taux de reconnaissance par classe pour le corpus HTI 2014-2015 pour un dictionnaire de 20 poses.

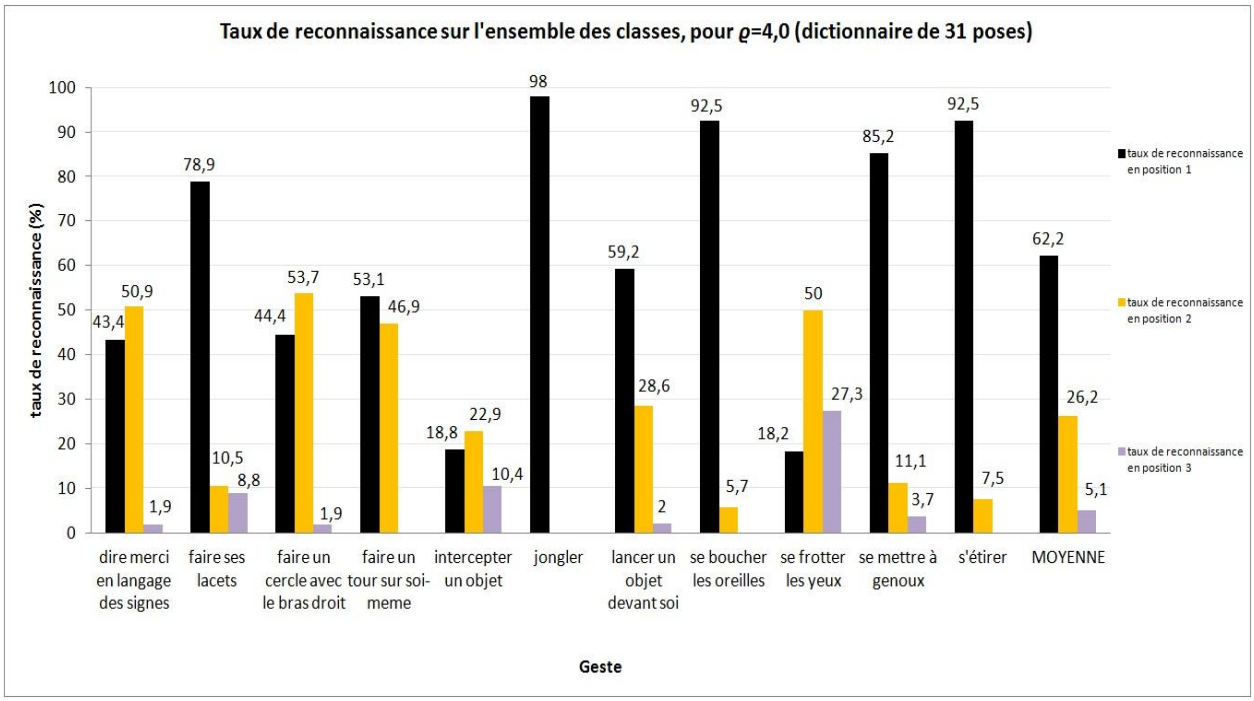


Figure VII.23 Taux de reconnaissance par classe pour le corpus HTI 2014-2015 pour un dictionnaire de 31 poses.

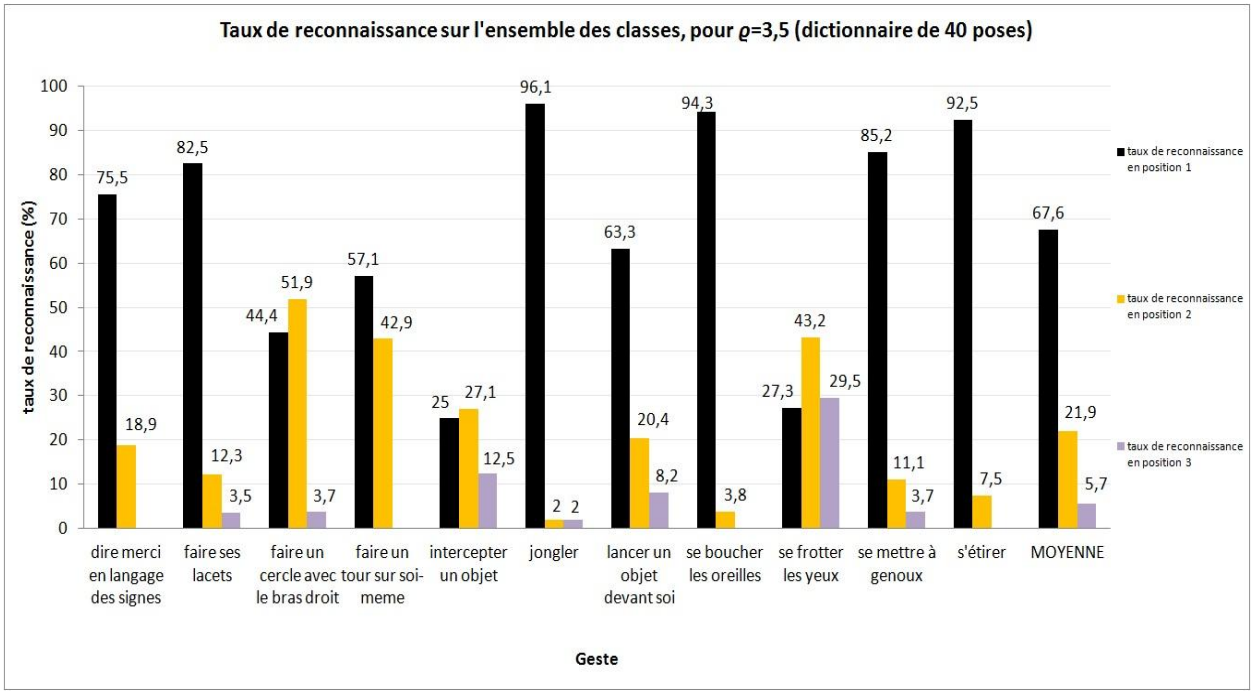


Figure VII.24 Taux de reconnaissance par classe pour le corpus HTI 2014-2015 pour un dictionnaire de 40 poses.

Tableau VII.20 Taux de reconnaissance cumulés par classe pour le corpus HTI 2014-2015 pour un dictionnaire de 40 poses.

Action	$TR^{G}_{cum}(1)$	$TR^{G}_{cum}(2)$	$TR^{G}_{cum}(3)$
dire merci en langage des signes	75,5	94,4	94,4
faire ses lacets	82,5	94,8	98,3
faire un cercle avec le bras droit	44,4	96,3	100
faire un tour sur soi-même	57,1	100	100
intercepter un objet	25	52,1	64,6
jongler	96,1	98,1	100,1
lancer un objet devant soi	63,3	83,7	91,9
se boucher les oreilles	94,3	98,1	98,1
se frotter les yeux	27,3	70,5	100
se mettre à genoux	85,2	96,3	100
s'étirer	92,5	100	100
MOYENNE	67.6	89.5	95.2

Tableau VII.21 Matrice de confusion par trame pour le corpus HTI 2014-2015 pour ϱ = 4,5 (dictionnaire de 20 poses).

Matrice de confusion	dire merci en langage des signes	faire ses lacets	faire un cercle avec le bras droit	faire un tour sur soi-même	intercepter un objet	jongler	lancer un objet devant soi	se boucher les oreilles	se frotter les yeux	se mettre à genoux	s'étirer
dire merci en langage des signes	29,3	0	12,2	1,8	0	3,2	**51,8**	0	0	1,7	0
faire ses lacets	18,4	**50,9**	2,6	3	0,3	0	0,2	0	0,2	24,4	0
faire un cercle avec le bras droit	33,7	0	**35,1**	3,8	0,8	2,6	14,1	0	0	0,1	9,8
faire un tour sur soi-même	36,9	1	7,8	**44,2**	0,7	1,9	0,3	0	0,6	6,6	0
intercepter un objet	**24,1**	0	16,9	3	24	4,3	12	0,2	9,4	1,2	4,9
jongler	17,9	0	5,8	1,8	3,2	**62,5**	3,3	0	4,7	0,6	0,2
lancer un objet devant soi	27,3	1	13,8	4,3	6,3	3,9	**38,2**	0	1	0,5	3,7
se boucher les oreilles	16,2	0,1	2	2,3	9	1,5	1,2	**48,8**	18,2	0,7	0
se frotter les yeux	19,7	0,1	3,1	2,1	9	3,7	0,5	**35,7**	25,4	0,7	0
se mettre à genoux	14,5	20,3	2,3	2,1	1,7	0,5	0,5	0	0,1	**58**	0
s'étirer	14,4	0	4,2	3,2	10,8	1,9	2	1	6	0,5	**56**

Tableau VII.22 Matrice de confusion par trame pour le corpus HTI 2014-2015 pour ϱ = 4,0 (dictionnaire de 31 poses).

Matrice de confusion	dire merci en langage des signes	faire ses lacets	faire un cercle avec le bras droit	faire un tour sur soi-même	intercepter un objet	jongler	lancer un objet devant soi	se boucher les oreilles	se frotter les yeux	se mettre à genoux	s'étirer
dire merci en langage des signes	41,9	0	5,4	2,3	0,2	2,4	**45,2**	0	0,5	2,1	0
faire ses lacets	21,1	**48,7**	0,5	4,4	0,2	0,2	0,2	0	0,2	24,5	0
faire un cercle avec le bras droit	**40,5**	0	37,4	5,3	1,1	0,3	11,2	0	0	1,8	2,4
faire un tour sur soi-même	**41,5**	1,2	4,6	39,4	0,6	1,4	1,1	0	0,5	9,7	0
intercepter un objet	**28,1**	0,1	15,7	3,4	21,4	4,2	10,8	0,4	9	2,1	4,8
jongler	19,8	0,2	2,9	2,1	5,5	**59,3**	3,8	0,5	4,2	1,6	0,1
lancer un objet devant soi	32,1	0,5	7,9	5,4	5,9	2,8	**39,5**	0	1,1	2,2	2,6
se boucher les oreilles	18	0,3	0,2	2,7	3,4	0,9	1,5	**55,6**	16	1,1	0,3
se frotter les yeux	21,4	0,6	0,5	3	2,2	1,8	0,8	**40,6**	27,1	2	0
se mettre à genoux	16,7	14,6	0,5	2,2	2,2	0,3	0,8	0	0,1	**62,5**	0,1
s'étirer	16,4	0	0,8	3,7	7,5	1,2	2,4	2,4	7,6	1,5	**56,5**

Tableau VII.23 Matrice de confusion par trame pour le corpus HTI 2014-2015 pour ϱ = 3,5 (dictionnaire de 40 poses).

Matrice de confusion	dire merci en langage des signes	faire ses lacets	faire un cercle avec le bras droit	faire un tour sur soi-même	intercepter un objet	jongler	lancer un objet devant soi	se boucher les oreilles	se frotter les yeux	se mettre à genoux	s'étirer
dire merci en langage des signes	**56,6**	0	2,4	2,6	0,1	3,5	33,8	0	0,1	0,9	0
faire ses lacets	21,7	**51,4**	0,4	4,5	0,1	0,1	0,9	0	0,1	20,8	0
faire un cercle avec le bras droit	38,2	0	**40,5**	5,1	1,1	0,4	12,4	0	0	0,2	2,1
faire un tour sur soi-même	**42,4**	1,7	5,2	40,8	0,2	1	2,8	0,1	0,2	5,6	0
intercepter un objet	**29,2**	0,1	12,4	3,2	23,7	5	12	0,5	7,9	0,8	5,2
jongler	20,4	0,3	2,2	2,5	4,8	**60,5**	4,7	0	3,8	0,6	0,2
lancer un objet devant soi	32,3	0,5	6,4	5,6	7,1	2,6	**41,2**	0	0,7	0,7	2,9
se boucher les oreilles	18,1	0,4	0,2	3,4	1,9	1,4	1,8	**59,9**	11,6	0,5	0,8
se frotter les yeux	21,8	0,6	0,4	3,7	1,3	1,8	1,7	**39,6**	28,1	0,8	0,2
se mettre à genoux	17,1	14,9	0,6	2,6	1,8	0,5	1,3	0	0	**61,1**	0,1
s'étirer	16,4	0,1	1,6	4,3	5	1,2	3,5	3,2	4,3	0,7	**59,7**

Globalement, les résultats de notre approche locale sont inférieurs à ceux que nous avions obtenus au chapitre VI où le mouvement corporel était décrit dans son intégralité et a posteriori (*cf.* Tableau VI.5). Une telle différence de performance peut s'expliquer par les capacités de nos descripteurs globaux à caractériser l'évolution d'indices de référence du mouvement de manière statistique et sur l'entièreté de la période de vie du geste. Outre le fait de se référer à de tels paramètres statistiques de séries de référence, ces descripteurs incorporent par ailleurs une dimension temporelle, pour par exemple saisir l'instant d'atteinte de l'extremum d'une série. Au contraire, l'approche dynamique que nous avons présentée dans ce chapitre consiste à étiqueter chaque trame de geste par estimation dynamique, sous réserve des hypothèses propres aux HMM énumérées à la section VII.2.3.1 (*cf.* équations VII.17, VII.18 et VII.19). De telles hypothèses, bien que permettant de modéliser les dépendances statistiques de nos signaux corporels, ne peuvent suffire à égaler les résultats d'une méthode de reconnaissance basée sur une description du geste faisant appel à son intégralité. Dans la mesure où nos taux de reconnaissance globaux ont été calculés à partir des signaux décodés à chaque trame et par un simple décompte par classe (*cf.* section VII.2.4, équations VII.62, VII.63 et VII.64), il nous paraissait bien difficile d'égaler les niveaux de performance de l'ordre de ceux obtenus au chapitre VI.

Bien que moins discriminante que l'approche globale présentée au chapitre VI, notre méthode d'analyse dynamique du geste, basée sur des descripteurs par trame, ainsi que sur du *soft assignment* dans un espace réduit de poses-clés, semble tout de même restituer les classes de gestes de façon convenable. Notre approche, que nous avons évaluée à l'aide de quatre corpus de gestes, dont trois de référence, obtient des résultats proches de ceux présentés dans l'état de l'art.

Dans la mesure où les taux de reconnaissance $TR^G(1)$, $TR^G(2)$ et $TR^G(3)$ dépendent des résultats de classification par trame dans les différentes classes de geste, l'évolution des taux de reconnaissance, à mesure que la taille de dictionnaire grandit, suit généralement celle des taux de reconnaissance par trame, tout en étant généralement plus ample que cette dernière, du fait de la troncature en laquelle consiste le

fait de ne conserver pour chaque geste que la catégorie la plus reconnue (respectivement les deux et trois catégories les plus reconnues) pour les calculs de $TR^G(1)$ (respectivement $TR^G(2)$ et $TR^G(3)$).

Nous constatons par ailleurs que les erreurs de reconnaissance qu'expriment les matrices de confusion par trame tendent à s'atténuer avec l'augmentation de la taille des dictionnaires de poses.

De tels résultats prouvent la légitimité de notre modèle de l'expressivité du geste et de son application à chaque instant du mouvement corporel, tout en soulignant que notre niveau de précision dans la quantification de l'expressivité ne permet pas de différencier clairement des gestes aussi proches que *se frotter les yeux* et *se boucher les oreilles* (*cf.* corpus HTI 2014-2015), ou encore *dessiner un cercle*, *cocher une case* et *dessiner une croix* (*cf.* corpus MSR Action 3D).

VIII. Conclusion et perspectives

Au cours de ce travail de thèse, nous avons proposé une méthodologie d'analyse de l'expressivité du geste corporel. Nous nous sommes directement inspirés de concepts abstraits appelés « qualités de mouvement », issus de l'analyse du mouvement dansé inaugurée par Rudolf Laban au siècle précédent sous le nom de *Laban Movement Analysis* (LMA [**26**] [**27**]). Une brève présentation de la LMA est proposée à la section II.3, où nous avons par ailleurs expliqué en quoi le modèle théorique de Laban, bien qu'originellement dédié à la danse, nous paraît à ce jour être le plus pertinent et suffisamment générique pour permettre une caractérisation expressive de gestes quelconques. Notre approche a abouti à la définition d'un ensemble de séries numériques temporelles dédiées à la quantification desdites qualités de Laban, pour caractériser un geste donnée selon les diverses dimensions de son expressivité. Ces divers éléments descriptifs ont été détaillés au chapitre IV. Ils sont calculés à chaque instant dans une fenêtre temporelle de quelques trames, à partir de positions 3D de référence du corps fournies par une Kinect tout au long du mouvement.

Nous avons émis l'hypothèse que notre représentation intermédiaire du mouvement, de par sa capacité à intégrer un certain degré d'information sémantique (*e.g.*, expressivité, intentionnalité, aspects communicationnels), pourrait être aisément couplée à l'usage d'algorithme d'apprentissages supervisés, en vue de la reconnaissance de contenus haut-niveau (actions, émotions…).

Suivant cette perspective, nous avons constitué un premier corpus de gestes, *ORCHESTRE-3D*, consacré à la gestuelle de chefs d'orchestre. Plus de 850 gestes acquis au cours de diverses répétitions ont été pré-segmentés manuellement, et proposés à l'annotation de praticiens de la musique selon un lexique de catégories d'émotions musicales défini en concertation avec des musiciens, chefs d'orchestre, et musicologues. L'objectif de cette première approche était de tenter d'appréhender un lexique émotionnel sous-jacent à la direction orchestrale.

Le corpus *ORCHESTRE-3D* a donc fait l'objet d'une première étude expérimentale, présentée au chapitre VI, où pour chacun de ses segments gestuels, nous avons calculé les séries numériques descriptives de l'expressivité du geste. Ces séries ont été utilisées de façon à établir une description globale de geste : un vecteur descripteur « global » de 81 valeurs dédiées à l'entièreté de la durée de vie du geste.

Dans ce cadre, nous avons tout d'abord testé et évalué les capacités de nos descripteurs à discriminer des contenus plus « classiques », à savoir les actions prédéfinies, à la fois iconiques et métaphoriques, du corpus *Microsoft Research Cambridge-12 Kinect gesture* (*MSRC-12*) [**36**]. Les actions du corpus MSRC-12 ainsi décrites ont été utilisées en entrée de différents algorithmes d'apprentissage, essentiellement à base de calssifieurs SVM (*Support Vector Machine*) et des forêts d'arbres décisionnels (*Random decision forest*). Les résultats obtenus sur ce corpus sont nettement supérieurs à ceux de l'état de l'art avec des taux de reconnaissances moyens de 99% pour les gestes iconiques et métaphoriques.

Dans un second temps et toujours dans une perspective d'analyse/reconnaissance globale de gestes, nous avons tenté d'inférer les différentes catégories émotionnelles que met en jeu notre base de chefs d'orchestre. Sans surprise, les scores de reconnaissance obtenus dans ce cas sont inférieurs à ceux correspondant à la base MSRC-12, en raison de la complexité sémantique évidente que met en jeu ce type de caractérisation émotionnelle ainsi que de la multiplicité des expressions possibles de chacune des émotions. Néanmoins, le taux global de reconnaissance pour ce corpus est de 57%. En outre, les résultats ont été particulièrement élevés pour une bonne moitié des classes d'émotions, incluant les catégories *calme* (76.6%), *agité* (59.1%), *éveillé* (76.5%), *serein* (66.7%) et *facile* (58.7%), pour lesquelles le modèle de Laban est particulièrement discriminant.

Par la suite, nous avons souhaité étendre les possibilités qu'offraient les quantifications proposées des qualités de mouvement à un contexte de reconnaissance dynamique, en temps réel des gestes.

Comme nous l'expliquons au chapitre V, la majorité des corpus de gestes 3D publiquement disponibles ne proposent que des séquences de mouvements corporels composées d'une seule action. A notre connaissance, seul le corpus *UTKinect-HumanDetection* [**37**] propose des successions d'actions au sein une même séquence. Néanmoins ces successions sont systématiquement composées des mêmes actions. Dans un premier temps nous avons donc élaboré une base de gestes qui puisse être adaptée à des objectifs de reconnaissance dynamique, et qui propose des successions variées d'actions variées. Ce corpus, appelé *HTI 2014-2015*, inclut 11 actions de la vie courante et a été constitué avec l'aide des étudiants de la majeure *High-Tech Imaging* (HTI) de Télécom SudParis pour la saison 2014-2015. Pour ce corpus, des séquences gestuelles comprenant quatre à six actions successives ont été réalisées par une dizaine de participants et enregistrées à l'aide d'une caméra Kinect de façon similaire à la constitution du corpus ORCHESTRE-3D. Le protocole d'élaboration de ce corpus est décrit en détail au chapitre V.

Le corpus HTI 2014-2015 a ensuite été utilisé à des fins de reconnaissance dynamique d'actions. Les différents indices sur lesquels reposent les séries numériques quantifiant les différents concepts de la LMA ont alors été utilisés pour constituer un vecteur descripteur par trame, où chaque instant du mouvement est décrit par une série de valeurs locales. A partir du nouvel espace vectoriel dans lequel chaque instant gestuel était représenté, nous avons construit un dictionnaire de poses-clés, extraites d'une sous-partie du corpus dédiée à l'apprentissage, et qui nous ont servi de référence pour échantillonner le mouvement. Chaque instant gestuel peut dès lors être projeté sur un certain nombre de ces poses représentatives de la base de données, et acquérir ainsi une représentation réduite, selon un procédé d'affectation douce (*soft assignment*). Cette représentation des gestes a été ensuite utilisée pour nourrir des chaînes de Markov cachées (*Hidden Markov Models - HMM*), afin de pouvoir assigner une action à chaque instant du geste.

L'application de notre méthode dynamique aux actions de la base HTI 2014-2015 nous a permis d'obtenir des taux de reconnaissance d'action supérieurs à 82% pour cinq gestes : *jongler*, *se boucher les yeux*, *s'étirer*, *faire ses lacets* et *se mettre à genoux*. Son utilisation sur d'autres corpus gestes (MSRC-12, *UTKinect-HumanDetection* [**37**] et *MSR Action 3D* [**38**]) montre la pertinence de nos descripteurs et de leur utilisation dans l'extraction de poses de référence.

Les résultats de cette approche « locale » sont logiquement inférieurs à ceux que nous avions obtenus avec une méthode globale (*e.g.*, descripteur défini pour la durée de vie entière du geste). En effet, dans le cas de l'approche locale, les taux de reconnaissance par classe ont été calculés à partir des signaux décodés à chaque trame et par un simple décompte par classe (*cf.* section VII.2.4). Il semblait dès lors bien difficile d'égaler les niveaux de performance de l'ordre de ceux obtenus avec l'approche globale, où les descripteurs caractérisaient l'évolution d'indices de référence du mouvement de manière statistique, et incorporaient par ailleurs une dimension temporelle en saisissant des instants relatifs d'atteinte d'extrema particuliers.

L'ensemble de nos résultats démontre globalement la légitimité de notre modèle de geste expressif. Les principales difficultés qui restent à lever concernent la différenciation des gestes structurellement proches, comme *se frotter les yeux* et *se boucher les oreilles* (*cf.* corpus HTI 2014-2015), ou encore *dessiner un cercle*, *cocher une case* et *dessiner une croix* (*cf.* corpus MSR Action 3D), pour lesquelles une description plus fine se doit d'être élaborée.

Ces dernières remarques nous permettent d'envisager de premières pistes d'avenir pour une meilleure description du contenu des gestes. En effet, l'idée d'une caractérisation du geste basée sur la LMA ne

suggère en rien que l'on doive réduire la description du mouvement corporel à des indices expressifs. On rappellera à cet égard les remarques effectuées à la fin du chapitre III consacré à l'état de l'art, où il était précisé que la plupart des approches établissant des descripteurs basés sur la LMA le font pour reconnaître des émotions – et non des actions. Il semble donc qu'au vu de la littérature, l'idée d'un modèle intermédiaire du geste se limitant à l'expressivité ne puisse valoir que selon un objectif de reconnaissance d'émotions ou d'états affectifs.

Nous avons pour notre part examiné en quoi la LMA pouvait s'avérer pertinente à discriminer aussi bien des émotions que des actions. Au contraire de descripteurs strictement dédiés à la structure du geste ou à ses aspects visuels (respectivement à l'expressivité) qui ne seraient aptes à caractériser que des actions définies avec précision (respectivement des émotions), nous avons prouvé que notre modèle descriptif est générique, capable de discriminer aussi bien des actions et que émotions.

Nous avons opté pour cette stratégie parce que les diverses composantes de la LMA nous paraissaient recouvrir les aspects les plus importants du geste. Il faut rappeler que des diverses qualités et sous-qualités de Laban, seules deux d'entre-elles, à savoir les composantes d'*Effort* et de *Forme*, réfèrent au rapport de l'exécutant du geste à son corps et à la *manière* de réaliser son geste. Ces deux qualités répondent ainsi à la question : « *Comment* le mouvement est-il exécuté ? ». Les composantes de *Corps* et d'*Espace* sont quant à elle davantage liées à l'aspect *structurel* du mouvement, à ce qui est directement visible. Elles répondent à la question : « *Quel* est le mouvement exécuté ? ». Cette référence à la stricte structure du mouvement devra donc être approfondie à l'avenir dans nos descripteurs, de façon à mieux pouvoir différencier entre des gestes aussi proches que *dessiner un cercle*, *cocher une case* et *dessiner une croix* (*cf.* corpus MSR Action 3D). Tout en restant dans le cadre d'une analyse inspirée par l'expressivité, il serait alors intéressant d'étudier une description plus exhaustive des éléments du mouvement.

Une autre perspective est également possible : il s'agit d'envisager des descriptions du mouvement respectivement en termes structuraux et expressifs, mais de façon séparée, afin d'entrainer des classifieurs indépendants, déterminant chacun les contenus d'intérêt (*e.g.*, actions, émotions). Une fusion des résultats, similaire au cas des approches multimodales, pourrait alors apporter la solution de reconnaissance recherchée (exemple : fusion son-geste [**150**] [**151**]). Une telle méthodologie pose néanmoins la question du statut de l'expressivité au regard de l'intégralité du geste. Est-elle un élément à la marge d'un profil « moyen » du geste ? Est-elle partie intégrante de celui-ci ?

Nous voudrions également évoquer d'autres problématiques associées à notre corpus ORCHESTRE-3D et aux difficultés que représente l'annotation d'un tel corpus.

Dans la mesure où la taille de notre corpus de direction orchestrale n'excédait pas le millier de gestes, et où le lexique d'émotions musicales ne comprenait pas moins de neuf émotions (malgré sa réduction initiale), toute tentative d'examen relatif à des mélanges d'émotions musicales ou tout projet de classification multi-étiquetage nous sont parus inutiles et inconséquents, malgré un projet initial qui tendait dans ce sens. C'est bien la raison pour laquelle lors de notre expérience d'analyse émotionnelle présentée dans la section VI.3, nous avons construit un classifieur pour chaque émotion. Implicitement, un tel choix suggérait l'hypothèse relativement restrictive que chaque classe peut être considérée comme indépendante des autres. En effet, les consignes données aux annotateurs relativement à l'usage d'un maximum de trois catégories émotionnelles par segment gestuel laissaient libre choix aux participants de combiner les émotions. Une étude future pourrait se pencher sur ces choix d'usage du lexique, en proposant plutôt des « lots » d'émotions comme étiquetage des échantillons.

Enfin, nos travaux futurs pourraient également être consacrés à l'analyse de corrélations entre les valeurs prises par les descripteurs de l'expressivité et les émotions choisies par les annotateurs, comme cela fut le cas dans [**75**] ou [**76**], ou des analyses factorielles permettent de relier un profil expressif à une émotion. Ici encore, la solidité des liens déduits ne peux que reposer sur une annotation unifiée et légitime.

Liste des publications

- Arthur Truong, Hugo Boujut and Titus Zaharia, "Laban descriptors for gesture recognition and emotional analysis", *CGI'14, 31st Computer Graphics International*, 10-13/06/14, Sydney, Australia
- Arthur Truong, Hugo Boujut and Titus Zaharia, "Laban movement analysis for action recognition", *Measuring Behavior 2014, 9th International Conference on Methods and Techniques in Behavioral Research*, 27-29/08/14, Wageningen, The Netherlands
- Arthur Truong, Hugo Boujut and Titus Zaharia, "A gesture expressive model based on Laban qualities", *IEEE 2014 ICCE Berlin, 4th IEEE International Conference on Consumer Electronics*, 07-10/09/14, Berlin, Germany (**cet article a reçu le prix du *Best Paper*: http://www.icce-berlin.org/2014/**)
- Arthur Truong, Hugo Boujut, and Titus Zaharia, "Laban descriptors for gesture recognition and emotional analysis", *The Visual Computer*, 2016, vol. 32, no 1, p. 83-98
- Arthur Truong and Titus Zaharia, "Laban Movement Analysis for real-time 3D gesture recognition", *Measuring Behavior 2016, 10th International Conference on Methods and Techniques in Behavioral Research*, 25-27/05/16, Dublin, Ireland
- Arthur Truong and Titus Zaharia, "Dynamic Gesture Recognition with Laban Movement Analysis and Hidden Markov Models", *CGI'16, 33st Computer Graphics International*, June 28-July 01, 2016, Heraklion, Greece
- Arthur Truong and Titus Zaharia, "Laban Movement Analysis and Hidden Markov Models for Dynamic 3D Gesture Recognition", submitted to *The Visual Computer* in 2016

Références

[1] Gordon Kurtenbach and Eric A. Hulteen, "Gestures in human-computer communication," *The art of human-computer interface design*, pp. 309-317, 1990.

[2] Platon, *The Republic*.

[3] Jean-Jacques Rousseau, *Reveries of a Solitary Walker*., 1782.

[4] René Descartes, *Méditations métaphysiques*.: Robert-Marc d'Espilly, 1724.

[5] Antonio R. Damasio, *Spinoza avait raison; joie et tristesse, le cerveau des émotions*.: Odile Jacob, Paris, 2003, vol. 318.

[6] Emmanuel Kant, *Critique de la raison pure*.: A. Trémesaygues et B. Pacaud, Paris, puf, 1781, vol. 4.

[7] Edmund Husserl, "Phenomenology," *Encyclopaedia Britannica*, vol. 14, pp. 699-702, 1927.

[8] Jean-Luc Petit, "For a theory of performance (in application to orchestra conducting)," in *Conscious Body 2: Performance and spectating (Paris 30/09/2013-06/10/2013)*, Villejuif.

[9] Gérard Jorland and Bérangère Thirioux, "Note sur l'origine de l'empathie," *Revue de métaphysique et de morale*, no. 58, pp. 269-280, février 2008.

[10] Alain Berthoz and Jean-Luc Petit, "Nouvelles propositions pour une physiologie de l'action," in *Education et formation*.: Presses Universitaires de France, 2006, pp. 253-259.

[11] Maurice Merleau-Ponty, *Résumé de cours au Collège de France*, Gallimard ed. Paris, 1968.

[12] Maurice Merleau-Ponty, *The world of perception*.: Cambridge University Press, 2004.

[13] Mark Billinghurst and Bill Buxton, "Gesture based interaction," in *Haptic input*.: Cambridge University Press, 2011, vol. 24.

[14] David McNeill, *Language and gesture*.: Cambridge University Press, 2000, vol. 2.

[15] Francisco Varela, Evan Thompson, and Eleanor Rosch, *L'inscription corporelle de l'esprit*. Paris: Seuil, 1993.

[16] Francisco Varela, *Autonomie et connaissance*.: Edition Seuil, 1989.

[17] Patrick Shove and Bruno H. Repp, "Musical motion and performance: Theoretical and empirical perspectives," in *The practice of performance*, J. Rink, Ed.: Cambridge University Press, 1995, pp. 55-83.

[18] Xin Wei Sha, Michael Fortin, and Jean-Sébastien Rousseau, "Calligraphic video: a phenomenological approach to dense visual interaction," in *Proceedings of the 17th ACM international conference on Multimedia*.: ACM, 2009, pp. 1091-1100.

[19] Thecla Schiphorst, "soft (n): Toward a Somaesthetics of Touch," in *CHI'09 Extended Abstracts on Human Factors in Computing Systems*.: ACM, 2009, pp. 2427-2438.

[20] Guillaume Garreta, Patricia Osganian, and Richard Shusterman, "Esthétique pragmatiste et conscience du corps," *Mouvements*, no. 1, pp. 71-76, 2009.

[21] Marianne Graves Petersen, Ole Sejer Iversen, Peter Gall Krogh, and Martin Ludvigsen, "Aesthetic Interaction: a pragmatist's aesthetics of interactive systems," in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*.: ACM, 2004, pp. 269-276.

[22] Richard Shusterman, *Pragmatist aesthetics: Living beauty, rethinking art*.: Cambridge University Press, 1992, vol. 27.

[23] Françoise Lejeune, *Corps à corps oeuvre-public: l'expérience des installations interactives*.: l'Harmattan, 2015.

[24] Barbara Formis, "Richard Shusterman, Conscience du corps. Pour une soma-esthétique," *Mouvements*, no. 1, pp. 155-157, 2009.

[25] Sarah Fdili Alaoui, "Analyse du geste dansé et retours visuels par modèles physiques: apport des qualités de mouvement à l'interaction avec le corps entier," Université Paris Sud-Paris XI, Thèse de doctorat 2012.

[26] Rudolf Laban, *La Maîtrise du Mouvement*. Arles: Actes Sud, 1994.

[27] Rudolf Laban, *Espace Dynamique*, Contredanse, Ed. Bruxelles, 2003.

[28] Robert Francès, *La perception de la musique*. Paris: Vrin, 1984.

[29] Xavier Hautbois, "Les Unités Sémiotiques Temporelles : de la sémiotique musicale vers une sémiotique générale du temps dans les arts," in *ICMS 8, Huitième Congrès International sur la Signification Musicale : Gestes, formes et processus signifiants en musique et sémiotique interarts*, Paris, 2004.

[30] Claude Cadoz, "Musique, geste, technologie," *Les nouveaux gestes de la musique*, pp. 47-92, 1999.

[31] Jean-Marie Adrien, "Une approche polyvalente: Direction Musicale Dansée, Captation Gestuelle Causale," in *Actes des Journées d'Informatique Musicale*, Rennes, 2010.

[32] Jean-Marie Adrien, "D'un phrasé en amont de la musique et de la danse," in *Questions de phrasé*, Paris, 2010.

[33] Baptiste Caramiaux, "Etudes sur la relation geste-son en performance musicale," Paris 6, Thèse de doctorat 2011.

[34] Jacques Mandelbrojt. (2003) La peinture s'exprime-t-elle, comme la musique, en UST? http://www.labo-mim.org/site/index.php?2008/08/22/49-publications-du-mim.

[35] Oliver Grewe, Reinhard Kopiez, and Eckart Altenmüller, "L'évaluation des sentiments musicaux: une comparaison entre le modèle circomplexe et les inventaires d'émotions à choix forcé,"

Musique, langage, émotion. Approche neuro-cognitive, pp. 49-73, 2010.

[36] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*.: ACM, 2012, pp. 1737-1746.

[37] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.: IEEE, 2012, pp. 20-27.

[38] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.: IEEE, 2010, pp. 9-14.

[39] Claude Cadoz, "Le geste canal de communication homme/machine: la communication" instrumentale"," *Technique et science informatiques*, vol. 13, no. 1, pp. 31-61, 1994.

[40] Adam Kendon, *Gesture: Visible action as utterance*.: Cambridge University Press, 2004.

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

[42] Mark Johnson, *The body in the mind: The bodily basis of meaning, imagination, and reason*.: University of Chicago Press, 2013.

[43] Penny Boyes Braem and Thüring Bräm, "A pilot study of the expressive gestures used by classical orchestra conductors," *Journal of the Conductor's Guild*, vol. 22, no. 1-2, pp. 14-29, 2001.

[44] Dominique Boutet, "Une morphologie de la gestualité: structuration articulaire," *Cahiers de linguistique analogique*, no. 5, pp. 81-115, 2008.

[45] Dominique Boutet, "Structuration physiologique de la gestuelle: modèle et tests," *Lidil. Revue de linguistique et de didactique des langues*, no. 42, pp. 77-96, 2010.

[46] Klaus R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[47] Eva Hudlicka, "What are we modeling when we model emotion?," in *AAAI spring symposium: emotion, personality, and social behavior*., 2008, pp. 52-59.

[48] Paul Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169-200, 1992.

[49] Paul Ekman and Wallace V. Friesen, *Facial Action Coding System Investigator's Guide*.: Consulting Psychologists Press, 1978.

[50] Charles Darwin, *The expression of the emotions in man and animals*. London, UK: John Muray, 1872.

[51] Grégory Beller, "Analyse et Modèle génératif de l'expressivité: Application à la parole et à l'interprétation musicale," Université Paris VI - Pierre et Marie Curie, Ecole Doctorale d'Informatique, Télécommunications et Electronique (EDITE) de Paris, spécialité Informatique.

IRCAM, Thèse de doctorat 2009.

[52] Wilhelm Max Wundt, *Grundriss der psychologie*.: W. Engelmann, 1896.

[53] Marc Schröder, "Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," University of Saarland, PhD thesis 2003.

[54] Albert Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261-292, 1996.

[55] Laurence Devillers et al., "Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches," *5th international conference on Language Resources and Evaluation (LREC 2006)*, p. 22, 2006.

[56] Eva Hudlicka and Hatice Gunes, "Benefits and limitations of continuous representations of emotions in affective computing: introduction to the special issue," *Journal of Synthetic Emotions*, vol. 3, no. 1, 2012.

[57] William James, "What is an emotion?," *Mind*, vol. 9, no. 34, pp. 188-205, April 1884.

[58] Andrew Ortony, Gerald L. Clore, and Allan Collins, *The cognitive structure of emotions*.: Cambridge university press, 1990.

[59] Alessandro Valitutti, "Interfacing Wordnet-affect with OCC model of emotions," in *The Workshop Programme*., 2010, p. 16.

[60] David Sander, Didier Grandjean, and Klaus R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural networks*, vol. 18, no. 4, pp. 317-352, 2005.

[61] M. David Sadek, Philippe Bretier, and Franck Panaget, "ARTIMIS: Natural dialogue meets rational agency," *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI'97), Nagoya, Japon*, pp. 1030-1035, 1997.

[62] Jonathan Gratch and Stacy Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269-306, 2004.

[63] Helmut Prendinger, Sylvain Descamps, and Mitsuru Ishizuka, "Scripting affective communication with life-like characters in web-based interaction systems," *Applied Artificial Intelligence*, vol. 16, no. 7-8, pp. 519-553, 2002.

[64] Fiorella De Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina De Carolis, "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent," *International journal of human-computer studies*, vol. 59, no. 1, pp. 81-118, 2003.

[65] Jean-Claude Martin, Radoslaw Niewiadomski, Laurence Devillers, Stéphanie Buisine, and Catherine Pelachaud, "Multimodal complex emotions: Gesture expressivity and blended facial expressions," *International Journal of Humanoid Robotics*, vol. 3, no. 3, pp. 269-291, 2006.

[66] Magalie Ochs, Radoslaw Niewiadomski, Catherine Pelachaud, and David Sadek, "Expressions intelligentes des émotions," *Revue d'intelligence artificielle*, vol. 20, no. 4-5, pp. 607-620, 2006.

[67] Amandine Grizard, Marco Paleari, and Christine Lisetti, "Adaptation d'une théorie psychologique pour la génération d'expressions faciales synthétiques pour des agents d'interface," in *WACA 2006, 2eme Workshop sur les Agents Conversationnels Animés, 26-27 octobre 2006, Toulouse, France*., 2006.

[68] Evthymios Papataxiarchis, "Emotions et stratégies d'autonomie en Grèce égéenne," *Terrain, revue d'ethnologie de l'Europe*, no. 22, pp. 5-20, 1994.

[69] Vincent Crapanzano, "Réflexions sur une anthropologie des émotions," *Terrain, revue d'ethnologie de l'Europe*, no. 22, pp. 109-117, 1994.

[70] Catherine A. Lutz and Lila Abu-Lughod, *Language and the politics of emotion*.: Editions de la Maison des Sciences de l'Homme, 1990.

[71] Olivier Roueff, "Musique et émotions," *Terrain*, no. 37, 2001.

[72] Annie Paradis, "Lyriques apprentissages. Les métamorphoses de l'émotion," *Terrain*, no. 37, 2001.

[73] Jacques Lacan, "Au delà du « Principe de réalité »," *L'Evolution Psychiatrique*, no. 3, pp. 67-86, Août-Octobre 1936.

[74] Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, and Joseph T Kider Jr, "Efficient motion retrieval in large motion databases," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*.: ACM, 2013, pp. 19-28.

[75] Toru Nakata, Taketoshi Mori, and Tomomasa Sato, "Analysis of impression of robot bodily expression," *Journal of Robotics and Mechatronics*, vol. 14, no. 1, pp. 27-36, 2002.

[76] Ali-Akbar Samadani, Sarahjane Burton, Rob Gorbet, and Dana Kulic, "Laban effort and shape analysis of affective hand and arm movements," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*.: IEEE, 2013, pp. 343-348.

[77] Kozaburo Hachimura, Katsumi Takashina, and Mitsu Yoshimura, "Analysis and evaluation of dancing movement based on LMA," in *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005*.: IEEE, 2005, pp. 294-299.

[78] Andreas Aristidou and Yiorgos Chrysanthou, "Feature Extraction for Human Motion Indexing of Acted Dance Performances," *GRAPP 2014 - International Conference on Computer Graphics Theory and Applications*, 2014.

[79] Dilip Swaminathan et al., "A dynamic bayesian approach to computational laban shape quality analysis," *Advances in Human-Computer Interaction*, vol. 2009, pp. 1-17, 2009.

[80] Liwei Zhao and Norman I. Badler, "Acquiring and validating motion qualities from live limb gestures," *Graphical Models*, vol. 67, no. 1, pp. 1-16, 2005.

[81] Durell Bouchard and Norman Badler, "Semantic segmentation of motion capture using laban movement analysis," in *Intelligent Virtual Agents*.: Springer Berlin Heidelberg, 2007, pp. 37-44.

[82] Diane Chi, Monica Costa, Liwei Zhao, and Norman Badler, "The EMOTE model for effort and shape," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*.: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 173-182.

[83] Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 213-225, 2003.

[84] Katharine Parton and Guy Edwards, "Features of conductor gesture: Towards a framework for analysis within interaction," in *The Second International Conference on Music Communication Science, 3-4 December 2009*., 2009.

[85] Donald Glowinski et al., "Multi-scale entropy analysis of dominance in social creative activities," *Proceedings of the international conference on Multimedia*, pp. 1035-1038, 2010.

[86] Donald Glowinski, Leonardo Badino, Alessandro D'Ausilio, Antonio Camurri, and Luciano Fadiga, "Analysis of Leadership in a String Quartet," *Third International Workshop on Social Behaviour in Music at ACM ICMI 2012*, 2012.

[87] Alessandro D'Ausilio et al., "Leadership in orchestra emerges from the causal relationships of movement kinematics," *PloS one*, vol. 7, no. 5, 2012.

[88] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*.: IEEE, 2004, vol. 3, pp. 32-36.

[89] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*.: IEEE, 2005, vol. 2, pp. 1395-1402.

[90] Meinard Müller et al., *Documentation mocap database hdm05*, Citeseer ed., 2007.

[91] Yale Song, David Demirdjian, and Randall Davis, "Tracking body and hands for gesture recognition: Natops aircraft handling signals database," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*.: IEEE, 2011, pp. 500-506.

[92] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free Viewpoint Action Recognition using Motion History Volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249-257, 2006.

[93] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*.: IEEE, 2008, pp. 1-8.

[94] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre, "HMDB51: A large video database for human motion recognition," in *High Performance Computing in Science and Engineering '12*.: Springer, 2013, pp. 571-582.

[95] Hamed Pirsiavash and Deva Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.: IEEE, 2012, pp. 2847-2854.

[96] Khurram Soomro and Amir R. Zamir, "Action Recognition in Realistic Sports Videos," *Computer Vision in Sports*, 2014.

[97] Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen, and Suh-Yin Lee, "Human action

recognition using star skeleton," in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*.: ACM, 2006, pp. 171-178.

[98] Lawrence R. Rabiner and Biing-Hwang Juang, "An introduction to hidden Markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4-16, 1986.

[99] Ying Wang, Kaiqi Huang, and Tieniu Tan, "Human activity recognition based on r transform," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*.: IEEE, 2007, pp. 1-8.

[100] Feiyue Huang and Xu Guangyou, "Viewpoint insensitive action recognition using envelop shape," in *Computer Vision-ACCV 2007*.: Springer Berlin Heidelberg, 2007, pp. 477-486.

[101] Imran N. Junejo, Khurrum Nazir Junejo, and Zaher Al Aghbari, "Silhouette-based human action recognition using SAX-Shapes," *The Visual Computer*, vol. 30, no. 3, pp. 259-269, 2014.

[102] Leo Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.

[103] Mohamed Bécha Kaâniche and François Brémont, "Recognizing gestures by learning local motion signatures of HOG descriptors," *IEEE Transactions on Pattern and Machine Intelligence*, 2012.

[104] Johan A.K. Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293-300, 1999.

[105] Vivek Kumar Singh and Ram Nevatia, "Simultaneous tracking and action recognition for single actor human actions," *The Visual Computer*, vol. 27, no. 12, pp. 1115-1123, 2011.

[106] Chi-Wei Chu and Isaac Cohen, "Posture and gesture recognition using 3D body shapes decomposition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops*.: IEEE, 2005, pp. 69-78.

[107] Stéphane G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397-3415, 1993.

[108] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*.: BMVA Press, 2009, pp. 124.1-124.11.

[109] Thanh Phuong Nguyen and Antoine Manzanera, "Action recognition using bag of features extracted from a beam of trajectories," in *2013 IEEE International Conference on Image Processing*.: IEEE, 2013, pp. 4354-4357.

[110] Alfons Juan and Enrique Vidal, "Comparison of four initialization techniques for the k-medians clustering algorithm," in *Advances in Pattern Recognition*.: Springer, 2000, pp. 842-852.

[111] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram, "A novel approach for recognition of human actions with semi-global features," *Machine Vision and Applications*, vol. 19, no. 1, pp. 27-34, 2008.

[112] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 3, pp. 710-719, 2005.

[113] Michael E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211-244, 2001.

[114] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.

[115] Vili Kellokumpu, Guoyin Zhao, and Matti Pietikäinen, "Human activity recognition using a dynamic texture based method," in *BMVC*., 2008, pp. 1-10.

[116] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.

[117] Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*.: IEEE, 2009, pp. 1454-1461.

[118] Ivan Laptev and Tony Lindeberg, "Interest point detection and scale selection in space-time," in *Scale Space Methods in Computer Vision*.: Springer, 2003, pp. 372-387.

[119] Chris Harris and Mike Stephens, "A combined corner and edge detector," in *Alvey vision conference*.: Citeseer, 1988, vol. 15, p. 50.

[120] Jingen Liu and Shah Mubarak, "Learning human actions via information maximization," in *CVPR 2008. IEEE Conference on Computer Vision and Pattern Recognition, 2008*.: IEEE, 2008, pp. 1-8.

[121] Yang Li, Junyong Ye, Tongqing Wang, and Shijian Huang, "Augmenting bag-of-words: a robust contextual representation of spatiotemporal interest points for action recognition," *The Visual Computer*, 2014.

[122] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299-318, 2008.

[123] Jianzhai Wu, Dewen Hu, and Fanglin Chen, "Action recognition by hidden temporal models," *The Visual Computer*, vol. 30, pp. 1395-1404, 2013.

[124] Nicolas Ballas et al., "Space-time robust representation for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*., 2013, pp. 2704-2711.

[125] Jonathan Alon, Vassilis Athitsos, Quan Yuan, and Stan Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1685-1699, 2009.

[126] Fabio Martinez, Antoine Manzanera, Michèle Gouiffès, and Annelies Braffort, "A Gaussian mixture representation of gesture kinematics for on-line Sign Language video annotation," in *International Symposium on Visual Computing*.: Springer, 2015, pp. 293-303.

[127] Fabrizio Pedersoli, Sergio Benini, Nicola Adami, and Riccardo Leonardi, "XKin: an open source framework for hand pose and gesture recognition using kinect," *The Visual Computer*, vol. 30, pp. 1107-1122, 2014.

[128] Dorin Comaniciu and Peter Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, 2002.

[129] Orasa Patsadu, Chakarida Nukoolkit, and Bunthit Watanapa, "Human gesture recognition using Kinect camera," in *2012 International Joint Conference on Computer Science and Software Engineering (JCSSE)*.: IEEE, 2012, pp. 28-32.

[130] Yale Song, Louis-Philippe Morency, and Randall Davis, "Distribution-Sensitive Learning for Imbalanced Datasets," in *2013 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2013)*.: IEEE, 2013.

[131] Farhood Negin, Firat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil, "A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras," in *Image Analysis and Recognition*.: Springer, 2013, pp. 648-657.

[132] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the IEEE International Conference on Computer Vision*.: IEEE Computer Society, 2013, pp. 2752-2759.

[133] Georgios Th. Papadopoulos, Apostolos Axenopoulos, and Petros Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *MultiMedia Modeling*.: Springer, 2014, pp. 473-483.

[134] Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban, "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*.: Press, AAAI, 2013, pp. 2466-2472.

[135] Xinbo Jiang, Fan Zhong, Qunsheng Peng, and Xueying Qin, "Online robust action recognition based on a hierarchical model," *The Visual Computer*, vol. 30, pp. 1021-1033, 2014.

[136] S. Ali Etemad and Ali Arya, "Correlation-optimized time warping for motion," *The Visual Computer*, 2014.

[137] CMU Graphics Lab Motion Capture Database. [Online]. mocap. cs.cmu.edu

[138] Xin Zhao, Xue Li, Chaoyi Pang, and Xi Zhu, "Online human gesture recognition from motion data streams," in *Proceedings of the 21st ACM international conference on Multimedia*.: ACM, 2013, pp. 23-32.

[139] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849-856, 2002.

[140] Xiaodong Yang and YingLi Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *2012 IEEE computer society conference on Computer vision and pattern recognition workshops (CVPRW)*.: IEEE, 2012, pp. 14-19.

[141] Yu Zhu, Wenbin Chen, and Guodong Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.: IEEE, 2013, pp. 486-491.

[142] Baptiste Caramiaux, Marcelo M. Wanderley, and Frederic Bevilacqua, "Segmenting and Parsing Instrumentalists' Gestures," *Journal of New Music Research*, vol. 41, no. 1, pp. 13-29, 2012.

[143] Gutemberg Guerra-Filho and Yiannis Aloimonos, "A language for human action," *Computer*, vol. 40, no. 5, pp. 42-51, 2007.

[144] Antoine Liutkus, Angélique Drémeau, Dimitrios Alexiadis, Slim Essid, and Petros Daras, "Analysis of dance movements using Gaussian processes," in *Proceedings of the 20th ACM international conference on Multimedia*.: ACM, 2012, pp. 1375-1376.

[145] Joshua S. Richman and J. Randall Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039-H2049, 2000.

[146] Gokcen Cimen, Hacer Ilhan, Tolga Capin, and Hasmet Gurcay, "Classification of human motion based on affective state descriptors," *Computer Animation and Virtual Worlds*, vol. 24, no. 3-4, pp. 355-363, 2013.

[147] Daniel Bernhardt and Peter Robinson, "Detecting affect from non-stylised body motions," in *Affective Computing and Intelligent Interaction*.: Springer Berlin Heidelberg, 2007, pp. 59-70.

[148] Yingliang Ma, Helena M. Paterson, and Frank E. Pollick, "A motion capture library for the study of identity, gender, and emotion perception from biological motion," *Behavior research methods*, vol. 38, no. 1, pp. 134-141, 2006.

[149] Themis Balomenos et al., "Emotion analysis in man-machine interaction systems," *Machine learning for multimodal interaction*, pp. 318-328, 2005.

[150] Hatice Gunes and Massimo Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334-1345, 2007.

[151] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92-105, 2011.

[152] Ellen Douglas-Cowie et al., "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*.: Springer Berlin Heidelberg, 2007, pp. 488-500.

[153] Phillip Ian Wilson and John Fernandez, "Facial feature detection using Haar classifiers," *Journal of Computing Sciences in Colleges*, vol. 21, no. 4, pp. 127-133, 2006.

[154] Danijela Vukadinovic and Maja Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *2005 IEEE International Conference on Systems, Man and Cybernetics*.: IEEE, 2005, vol. 2, pp. 1692-1698.

[155] Martin Wöllmer et al., "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH*.: Citeseer, vol. 2008, pp. 597-600. [Online].

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[156] Mike Schuster and Kuldip K. Paliwal, "Bidirectional recurrent neural networks," vol. 45, no. 11, pp. 2673-2681, 1997.

[157] Alex Graves and Jürgen Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602-610, 2005.

[158] Anand S. Rao and Michael P. Georgeff, "Modeling rational agents within a BDI-architecture," *KR*, vol. 91, pp. 473-484, 1991.

[159] Magy Seif El-Nasr, Thomas R. Ioerger, and John Yen, "Learning and emotional intelligence in agents," in *Proceedings of AAAI (American Association for Artificial Intelligence) Fall Symposium on Emotional Intelligence*. Floride, 1998, pp. 1017-1025.

[160] Stefano Pasquariello and Catherine Pelachaud, "Greta: A simple facial animation engine," in *Soft Computing and Industry*.: Springer, 2002, pp. 511-525.

[161] Paul Ekman and Wallace V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*.: Englewood Cliffs, NJ: Prentice Hall Trade, 1975.

[162] Donald Glowinski et al., "Toward a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106-118, 2011.

[163] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe, "Multimodal analysis of expressive gesture in music and dance performances," *Gesture-based communication in human-computer interaction*, pp. 20-39, 2004.

[164] Norman I. Badler and Stephen W. Smoliar, "Digital representations of human movement," *ACM Computing Surveys (CSUR)*, vol. 11, no. 1, pp. 19-38, 1979.

[165] Charles E. Clauser, John T. McConville, and John W. Young, *Weight, Volume and Center of Mass of Segments of the Human Body*. Wright-Patterson Air Force Base, Ohio (AMRL-TR-69-70): ANTIOCH COLL YELLOW SPRINGS OH, 1969.

[166] scikit-learn. [Online]. http://scikit-learn.org/stable/

[167] Jonathan Milgram, Mohamed Cheriet, and Robert Sabourin, ""One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs?," in *Tenth International Workshop on Frontiers in Handwriting Recognition*., 2006.

[168] Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3-42, 2006.

[169] Payam Refaeilzadeh, Lei Tang, and Huan Liu, "Cross-validation," in *Encyclopedia of database systems*.: Springer, 2009, pp. 532-538.

[170] George Hripcsak and Adam S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296-298, 2005.

[171] Yunqiang Chen, Xiang Sean Zhou, and Thomas S. Huang, "One-class SVM for learning in image retrieval," in *2001 International Conference on Image Processing, 2001. Proceedings*.: IEEE, 2001, vol. 1, pp. 34-37.

[172] Eulalia Szmidt and Janusz Kacprzyk, "Distances between intuitionistic fuzzy sets," *Fuzzy sets and systems*, vol. 114, no. 3, pp. 505-518, 2000.

[173] Douglas Reynolds, "Gaussian mixture models," *Encyclopedia of Biometrics*, pp. 827-832, 2015.

[174] Tobias P. Mann, "Numerically stable hidden Markov model implementation," in *An HMM scaling tutorial*., 2006, pp. 1-8.