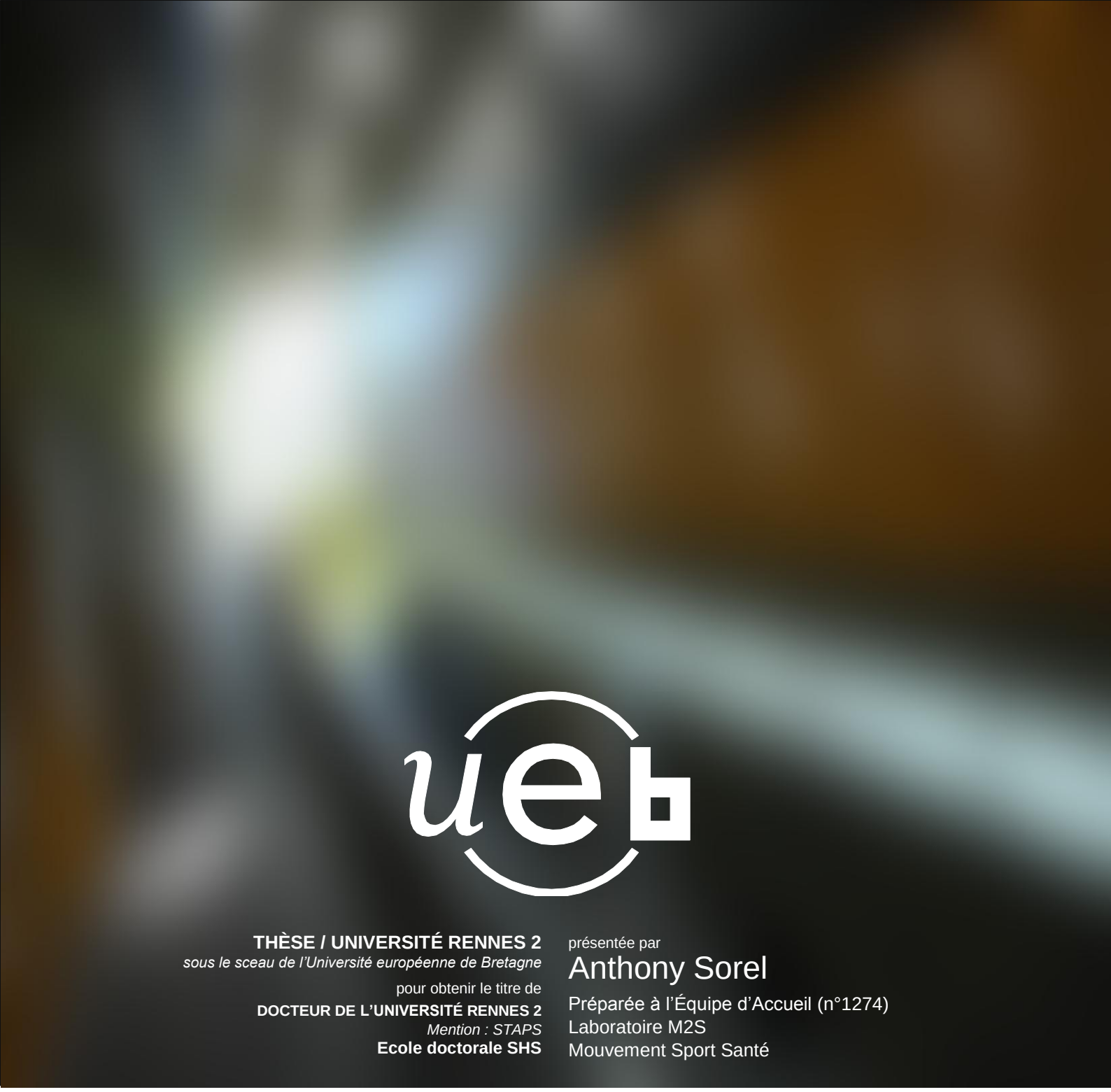
Gestion de la variabilité morphologique pour la reconnaissance de gestes naturels à partir de données 3D

Anthony Sorel



THÈSE / UNIVERSITÉ RENNES 2
sous le sceau de l'Université européenne de Bretagne

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITÉ RENNES 2
Mention : STAPS
Ecole doctorale SHS

présentée par
Anthony Sorel

Préparée à l'Équipe d'Accueil (n°1274)
Laboratoire M2S
Mouvement Sport Santé

Gestion de la variabilité morphologique pour la reconnaissance de gestes naturels à partir de données 3D

Thèse à soutenir le 6 décembre 2012
devant le jury composé de :

Éric Anquetil
Professeur, INSA Rennes / **Rapporteur**

Gilles Dietrich
Maître de Conférence (HDR), Université Paris Descartes / **Rapporteur**

Sylvie Gibet
Professeur, Université Bretagne Sud / **Examinatrice**

Erwan Mahé
Directeur du Développement, Artefacto (Rennes) / **Membre invité**

Richard Kulpa
Maître de Conférence, Université Rennes 2 / **Co-directeur de thèse**

Franck Multon
Professeur, Université Rennes 2 / **Directeur de thèse**

N° d'ordre :

Thèse de doctorat

Sous le sceau de

l'Université Européenne de Bretagne

pour obtenir le grade de

Docteur de l'Université Rennes 2
Discipline STAPS

par

Anthony Sorel

Equipe d'accueil : Laboratoire Mouvement Sport Santé - EA 1274
Ecole Doctorale : Sciences Humaines et Sociales (SHS)

Gestion de la variabilité morphologique pour la reconnaissance de gestes naturels à partir de données 3D

à soutenir le 6 décembre 2012 devant la commission d'examen

Éric	Anquetil	Professeur, INSA Rennes	Rapporteur
Gilles	Dietrich	Maître de Conférence (HDR), Université Paris Descartes	Rapporteur
Sylvie	Gibet	Professeur, Université Bretagne Sud	Examinatrice
Erwan	Mahé	Directeur du Développement, Artefacto (Rennes)	Membre invité
Richard	Kulpa	Maître de Conférence, Université Rennes 2	Co-directeur
Franck	Multon	Professeur, Université Rennes 2	Directeur

"Les attitudes, gestes et mouvements du corps humain sont risibles dans l'exacte mesure où ce corps nous fait penser à une simple mécanique."

Henri Bergson (1859 - 1941)

Table des matières

Introduction

Le mouvement et l'activité physique prennent de plus en plus d'importance dans notre quotidien, particulièrement dans notre société moderne, très propice à la sédentarité. Les campagnes de prévention et d'information se succèdent, aussi bien en France que dans les autres pays du monde, pour lutter contre ce phénomène de sédentarité, une des premières raisons de l'expansion de l'obésité dans les sociétés modernes.

L'évolution de la technologie a pendant longtemps favorisé cette sédentarité en apportant de nouveaux loisirs interactifs à domicile. Le marché des jeux vidéos a connu une véritable explosion et les budgets de développements alloués à ces médias ont atteint ceux des grosses productions de films. Jusqu'aux années 2000, ces loisirs, souvent destinés aux adolescents, se pratiquaient assis, une manette à la main, parfois pour de très longues durées. L'étude [GfK2011] « Les Français et l'Entertainment » de février 2011 montre que la France compte 28 millions de joueurs, amenant le jeu vidéo comme premier loisir des français, particulièrement pour les 13-19 ans qui y consacrent environ 9h par semaine en moyenne.

Les années 2000 ont apporté une vraie révolution en voyant l'arrivée des consoles utilisant des capteurs de mouvements, comme la Wii (Nintendo), la PS Move (Sony) ou la Kinect (Microsoft). En 2010, les ventes de ces capteurs de mouvements intégrés à des consoles ont atteint plus de 500 000 unités, ce qui montre l'engouement du public. Cela a aussi profondément modifié le public, initialement constitué d'adolescents masculins, puisque les femmes représentent maintenant près de 50% des joueurs, et l'âge moyen a augmenté pour atteindre 33 ans, touchant toutes les couches sociales de la population. Les capteurs de mouvements touchent même maintenant les téléphones, les télécommandes... Pour la plupart, ces capteurs sont utilisés pour des applications ludiques afin de permettre une interaction la plus naturelle possible avec des contenus multimédias. Les applications ont donc cherché à tirer le meilleur parti de ces capteurs en proposant des contenus engageant la personne dans une activité physique, profitant des campagnes de promotion à ce sujet. De nombreuses applications sont alors apparues pour pratiquer une activité physique ludique. Plusieurs études ont d'ailleurs montré l'importante motivation que ce type d'applications apportait. Cependant, les principales applications sont restées sur un plan principalement ludique, sans véritable objectif d'amélioration de performance ou de

qualité physique. Améliorer son score au tennis virtuel ne permet pas de s'améliorer face à un vrai joueur.

Cet engouement pour les capteurs de mouvements a aussi intéressé des industriels comme les équipementiers sportifs. Babolat a par exemple développé une raquette de tennis qui peut communiquer au joueur les informations quantifiées de son match. Plusieurs projets, dont le projet Européen SKILLS [1] se sont intéressés à utiliser ce type de technologie pour mesurer, codifier et entrainer des compétences motrices dans différents domaines. S'entrainer dans des environnements immersifs à des tâches complexes, coûteuses ou dangereuses est maintenant largement utilisé dans certains domaines industriels, principalement dans les grands groupes comme EADS. La société Didhaptic [2] développe des solutions et propose des prestations en réalité virtuelle pour la formation appliquée pour les métiers de la santé et pour l'adaptation des postes de travail d'un point de vue ergonomique. Cela repose sur des capteurs de mouvements relativement précis qui permettent d'analyser la performance de l'utilisateur et d'animer un avatar.

La société Artefacto [3], partenaire Cifre de cette thèse, s'est engagée dans ce type d'applications en développant des systèmes et des contenus pour l'entrainement à des tâches métiers. Cette thèse s'inscrit donc dans cette philosophie. Elle s'est partiellement déroulée pendant le projet Biofeedback financé par le plan de relance de l'industrie des « Serious Games », du Ministère des Finances, de 2009 à 2011. L'objectif de ce projet était de démontrer la faisabilité d'un apprentissage de compétences motrices complexes, comme le karaté ou la danse, à partir de systèmes immersifs incluant une captation du mouvement de l'utilisateur.

D'un point de vue scientifique, ce type de projet pose différents problèmes. L'un d'entre eux concerne l'exploitation des données issues des capteurs de mouvements pour reconnaitre l'action effectuée par l'utilisateur et en évaluer la performance. Cette problématique rejoint d'autres domaines de recherche tels que la classification d'activités dans des flux vidéos, ou la définition d'interfaces gestuelles pour des applications interactives (les tablette tactiles en sont un des exemples les plus récents). Une grande partie des travaux dans ce domaine est liée à la vision par ordinateur ou les interfaces homme-machine. Depuis quelques années, le domaine de l'animation graphique s'est aussi intéressé à ce problème avec la particularité d'utiliser des capteurs 3D du mouvement, contrairement aux autres domaines qui, la plupart, analysent le mouvement en 2D.

Dans le domaine des STAPS, l'analyse de la performance motrice à partir de mesures 3D du mouvement est largement répandue. Elle consiste généralement à mesurer des trajectoires articulaires 3D et à comparer des populations avec différents niveaux d'expertise. Ces analyses s'effectuent dans un cadre bien délimité : on étudie exclusivement un mouvement bien connu. Les capteurs délivrent des signaux qu'il est donc facile de découper pour isoler le mouvement à analyser et les critères de performance sont bien définis à l'avance. Ceci n'est pas obligatoirement le cas dans un système automatique et interactif d'entrainement à des tâches motrices complexes. Il est alors nécessaire d'isoler et de reconnaitre le mouvement effectué par l'utilisateur avant d'en évaluer la performance. C'est le principal objectif de cette thèse.

Cette thèse se situe donc à la frontière des différents domaines scientifiques abordés ci-dessus. S'inscrivant dans une problématique de conception d'environnements immersifs d'entrainement, plusieurs problèmes principaux se posent. Le premier concerne la reconnaissance du mouvement effectué par l'utilisateur, alors que les capteurs délivrent un flux continu de données plus ou moins bruitées. Le second problème consiste à gérer la variabilité intra et inter individuelle dans

1. www.skills-ip.eu/
2. www.didhaptic.com
3. www.artefacto.fr

la réalisation d'un mouvement. Cette variabilité peut être intrinsèque au mouvement lui-même (comme saisir un même objet à différents endroits) mais peut aussi être lié à l'utilisateur. Dans ce dernier cas, les dimensions anthropométriques de chaque utilisateur peuvent conduire à des mesures différentes pour un même mouvement. Ces variabilités sont aussi le fruit de styles différents dans la réalisation du mouvement (par exemple lié à l'émotion de l'utilisateur : plus ou moins énervé ou fatigué...). Il est donc nécessaire de proposer une représentation du mouvement qui soit la plus indépendante possible à ces variabilités. Enfin, dans un cadre immersif, l'utilisateur attend un retour à ses actions dans un temps relativement court. L'utilisation d'un avatar rend cette contrainte très forte puisque l'humain virtuel doit reproduire le plus rapidement et fidèlement possible les mouvements de l'utilisateur. L'approche envisagée pour cette thèse doit donc tenir compte de cette contrainte temporelle et ne pas devoir attendre la fin du mouvement pour commencer à l'analyser.

La première partie de cette thèse dresse un état de l'art pluridisciplinaire sur la reconnaissance de mouvements, montrant qu'il est nécessaire d'aborder à la fois les problèmes de représentation du mouvement et de méthode de reconnaissance. La deuxième partie décrit une première contribution qui introduit une nouvelle représentation du mouvement conçue pour limiter l'impact de la variabilité morphologique sur la reconnaissance du mouvement exécuté par l'utilisateur. La troisième partie cherche à démontrer que cette même représentation est robuste même dans des conditions où le système ne dispose pas de la totalité de l'information, la rendant adaptée à la reconnaissance temps réel. Le dernier chapitre donne une conclusion et propose des perspectives à ce travail.

Chapitre 1

Revue de la littérature

Avant de s'intéresser à la reconnaissance de gestes, il apparait essentiel de décrire ce que renferme la notion même de geste. Nous consacrons donc cette première partie 1.1 de l'état de l'art, à décrire ce qu'est un geste et ses différentes taxonomies.

La seconde partie (1.2) aborde les méthodologies employées pour reconnaitre le mouvement. Nous nous intéressons, tout d'abord, aux descriptions du mouvement qui sont utilisées à cet effet. Puis, nous nous consacrons aux approches mathématiques utilisées pour résoudre les problèmes de reconnaissance de mouvements.

1.1 Contexte de la reconnaissance de gestes

Le sens commun attribué au mot geste est plutôt vaste, comme l'observe [Corradini2002] : « Tout le monde prétend savoir ce qu'est un geste, mais personne ne peut vous l'expliquer précisément ». Il apparait dès lors nécessaire de bien définir ce terme équivoque, afin de se donner un vocabulaire minimal cohérent pour la suite du récit. Beaucoup de définitions ont été proposées, sans faire consensus. Selon le dictionnaire [1], c'est un :

> Mouvement du corps (surtout des bras, des mains, de la tête), révélant un état d'esprit, ou visant à exprimer ou à exécuter quelque chose.

[Yang2006] propose en revanche une lecture plus bas niveau :

> Morceau d'une trajectoire spatio-temporelle qui possède une trajectoire stéréotypée autorisant une grande variabilité.

La distinction principale parmi les précédentes définitions réside dans leur niveau d'abstraction sémantique. A partir de ce constat, [Ramstein1991, Cadoz2000, Nielsen2004] proposent de définir le geste selon deux approches :

1. Le Robert de poche, 2011

- une approche fonctionnelle qui se réfère aux fonctions qu'un geste peut exécuter dans des situations spécifiques : la sémantique du geste,

- une approche phénoménologique fondée sur des critères cinématiques (vitesse), spatiaux (l'espace dans lequel le geste s'exécute) et fréquentiels (composantes harmoniques, comme le léger mouvement du poignet au bout du bras).

Cette distinction est très bien adaptée au contexte de la reconnaissance de gestes, l'objectif d'un tel système étant, précisément, de faire correspondre la cinématique d'un geste (le mouvement) à sa sémantique. En effet, dans un cadre immersif, tendre la main vers un avatar devrait être interprété par l'environnement virtuel comme une intention de saluer. Ce dernier proposera une réponse adaptée de l'avatar, qui tendra également la main en retour. Cette différence de niveau conceptuel entre la machine, qui ne connait que les données cinématiques (la trajectoire de la main dans l'espace physique), et l'utilisateur, qui sait les interpréter (l'action « saluer »), est appelée fossé sémantique [Li2004]. Il appartient au système de reconnaissance de combler ce fossé en proposant une interprétation haut-niveau (l'intention du geste) des données bas-niveau (la mesure du geste). La figure 1.1 peut illustrer ce fossé.

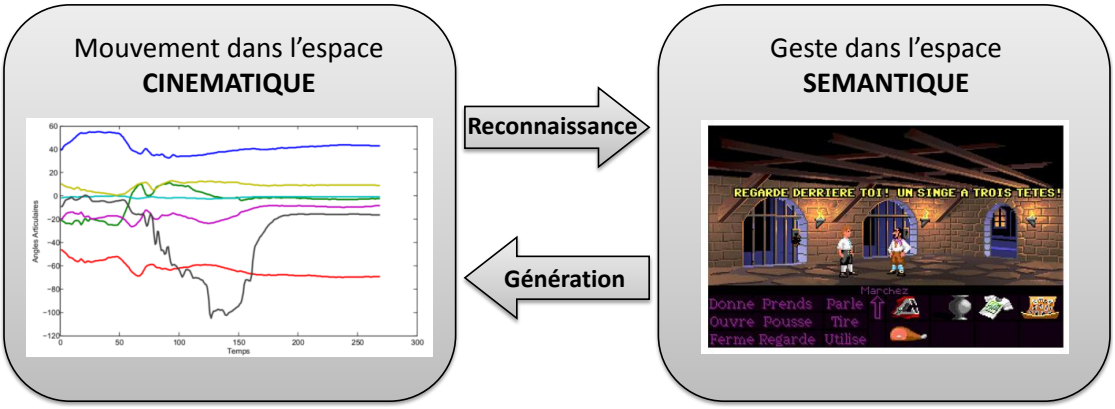


Figure 1.1 - Différence de niveau conceptuel entre une représentation cinématique d'un geste (un ensemble de signaux), à gauche et, à droite, une représentation sémantique des actions réalisables (Donner, Ouvrir, Pousser, Utiliser...) dans un environnement virtuel (image extraite du jeu Monkey Island de LucasArts). L'opération de passage de l'espace cinématique vers l'espace sémantique correspond à la reconnaissance (définir l'action réalisée à partir des signaux mesurés), l'opération inverse représente la génération de mouvement.

Comme le met en évidence [Moeslund2006] dans sa synthèse bibliographique, la littérature utilise, de façon interchangeable, des termes comme mouvement, geste, action, activité, geste complexe, action simple, activité composite, motif de mouvement ou comportement. L'objectif de cette thèse n'est pas de trancher la question. Cependant, afin de se donner un vocabulaire cohérent, nous adoptons le terme *mouvement*, pour se référer aux déplacements des segments corporels du corps humain dans l'espace cinématique. Le terme *geste* se rapporte à la sémantique du mouvement. Dans cette thèse, le but ne sera pas d'aller jusqu'à la sémantique elle-même qui ferait appel à une analyse haut niveau de l'information portée par le mouvement. L'objectif de cette thèse est d'associer une classe de mouvements aux trajectoires mesurées ; classe de mouvements qui pourra ensuite être analysée au niveau sémantique pour déterminer le sens et l'intention cachés de l'utilisateur. Nous parlons donc de reconnaissance de mouvements.

Dans la suite de cette section, nous présentons la notion de geste à travers différentes taxonomies permettant de le caractériser. Ces taxonomies du geste posent certaines bases sur lesquelles vont, par la suite, reposer les méthodologies de reconnaissance. Dans la seconde partie de cette section, nous détaillons la capture et la représentation numérique du mouvement, qui sont les étapes préalables à tout traitement informatique.

1.1.1 Le geste naturel

Pour interagir avec son environnement extérieur, que ce soit manipuler un objet ou communiquer avec un être vivant ou une machine, l'homme dispose de 5 modes de communication (entendre, voir, parler, toucher, agir). Parmi ces 5 modes, le geste est naturellement privilégié dans de nombreuses situations. Le coté naturel du geste est largement validé par de nombreuses études. [Iverson1998], par exemple, démontre que le geste est spontanément associé à la parole auprès d'aveugles congénitaux, démontrant ainsi son coté instinctif. En environnement virtuel, les expériences du type *magicien d'Oz* permettent de mettre en évidence les modes d'interaction naturellement privilégiés avec la machine. Il s'agit d'observer le comportement d'un utilisateur face à un environnement virtuel, les fonctionnalités manquantes étant simulées par un opérateur humain en temps réel. Ce type d'expérience démontre que les gestes peuvent véhiculer des informations pour lesquelles les autres modalités (langage, console de commande...) ne sont pas efficaces ou appropriées [Caridakis2010]. [Dauchy1993] montre ainsi, que la manipulation complexe d'objets virtuels, comme la rotation et la translation en 3D d'un cube, est multimodale. L'étude met également en évidence la redondance des différentes modalités, le geste étant souvent porteur de commande, alors que la parole n'est que commentaire. En revanche, pour des tâches simples, c'est le geste qui est préférentiellement employé de façon monomodale.

A partir de ce constat, il est possible de caractériser le sentiment de présence d'un sujet immergé dans un environnement virtuel en observant le naturel de ses gestes face à des stimuli virtuels. Dans le cadre sportif, où les gestes sont l'essence même du jeu, [Bideau2003] montre ainsi que la performance motrice d'un gardien de but de handball est la même, que le tireur en face de lui soit réel ou virtuel. Le geste est utilisé de manière totalement intuitive. C'est d'ailleurs un intérêt majeur de l'utilisation de gestes naturels en environnement virtuel : la phase d'assimilation de l'interface est réduite. On peut s'en convaincre, en analysant le succès commercial des interfaces de commande tactiles (tablettes, téléphones), qui est, en partie, dû à leur ergonomie. Le geste d'étirement, réalisé avec le pouce et l'index pour zoomer, tout comme le geste de glissement, qui rappelle celui d'une page qu'on tournerait, sont des métaphores de gestes réalisables dans le monde réel.

Les interfaces tactiles démontrent tout l'intérêt que l'homme peut tirer d'un système de reconnaissance sur des gestes 3D. Seulement, les mouvements 3D ne peuvent pas être envisagés simplement comme des mouvements 2D auxquels on ajouterait une autre dimension. La troisième dimension ajoute des complexités nouvelles et inattendues [Gielen2009]. D'autant que les degrés de liberté du corps humain entrainent une importante variabilité du mouvement que la reconnaissance doit prendre en compte.

1.1.1.1 Caractéristiques sémantiques

La classification est une faculté d'abstraction propre à l'homme, qui lui permet d'organiser ses connaissances. Plusieurs classifications, ou taxonomies, ont ainsi vu le jour afin d'étudier le geste sous ses différents aspects (voir [Wexelblat1998, Donovan2005, Kida2005] pour des revues de littérature exhaustives sur la question).

La principale taxonomie est proposée par [Mcneill1992] et précisée par [Cadoz1994, Quek1994]. Elle est représentée en Figure 1.2. Elle oppose deux grandes fonctions :

- la fonction manipulative, qui regroupe les actions matérielles sur l'environnement visant à le modifier (déplacer un objet) ou à le percevoir (toucher).
- la fonction communicative, appelée sémiotique, qui regroupe les gestes porteurs d'informations destinées à l'environnement. Ces gestes servent souvent à préciser le discours oral.

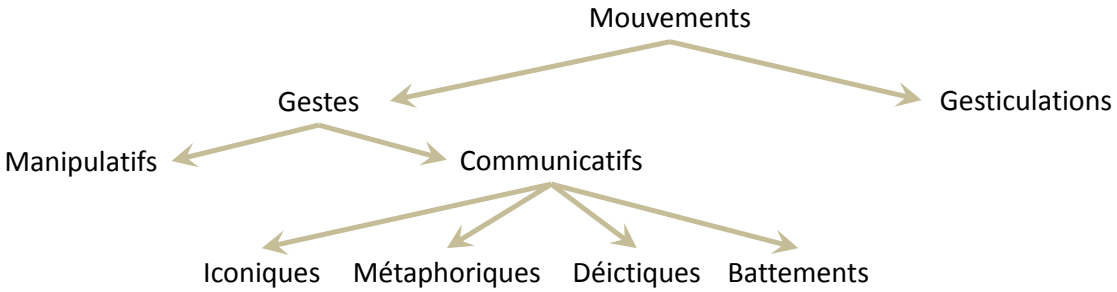


Figure 1.2 - Taxonomie du geste proposée par [Mcneill1992] ; légèrement adaptée par [Pavlovic1997], pour inclure un chapeau « Mouvement », distinguant les gestes, des gesticulations (des mouvements non-signifiants).

Cette taxonomie permet de discrétiser l'espace de représentation sémantique en grandes classes. Elle est reprise dans la littérature sur la reconnaissance automatique de mouvements, notamment pour le pointage ou le dimensionnement [Wilson2002, Corradini2002] ou pour définir un corpus à reconnaitre [Boulabiar2011]. C'est dans ce cadre que se situe ce travail de thèse.

En effet, dans le domaine de la reconnaissance automatique, quelle que soit la taxonomie employée, un geste se résume, du point de vue de la machine, à une classe de mouvements, dont l'étiquette peut être formulée dans un langage de haut niveau. Pour chaque classe, il existe une importante variété de mouvements compatibles, liée au style, à la situation dans laquelle se trouve le sujet, ses dimensions anthropométriques. . . Toute la difficulté de la tâche de reconnaissance consiste à composer avec cette variabilité du mouvement pour isoler des caractéristiques permettant d'identifier la classe à laquelle il appartient. C'est pourquoi, dans la suite de ce mémoire, nous utilisons plutôt la terminologie "reconnaissance de mouvements", ne contribuant pas sur les aspects sémantiques et sémiologiques qui caractérisent le geste.

1.1.1.2 Caractéristiques spatiotemporelles

Dans la section précédente, nous avons introduit l'espace de représentation sémantique, un espace discret, dans lequel les gestes sont catégorisés par fonction, dans un langage de haut

niveau. Cet espace de représentation sémantique est sous-tendu par un espace physique continu, qui compte 3 dimensions spatiales et une dimension temporelle, supportant l'exécution du geste via des mouvements. Cette exécution n'est pas définie de manière unique, elle englobe une grande variété de mouvements compatibles. Il est donc fondamental de considérer les propriétés spatiotemporelles du mouvement.

Un mouvement pouvant être exécuté par différents groupes de segments corporels (doigts, mains, bras, tête, main avec un objet, haut du corps, corps entier, déplacement global...), [Moeslund2006] propose de distinguer 2 catégories : l'approche globale, baptisée holistique, qui considère le corps entier sans tenir compte des segments ; et l'approche locale qui considère les parties du corps exécutant le mouvement. Dans cette thèse, nous nous plaçons clairement dans la deuxième approche, afin de pouvoir catégoriser finement et d'évaluer les mouvements d'un utilisateur impliqué dans un programme d'entrainement en environnement immersif.

L'information spatiale peut aussi fournir des arguments contextuels qui précisent certains mouvements, comme le lieu visé lors d'un pointage, ou la vitesse d'exécution d'un lancer. Certaines approches en reconnaissance automatique permettent de tenir compte de ces informations supplémentaires [Wilson2002].

Un geste peut être considéré à différentes échelles de temps [Ronfard2009]. Par exemple, « prendre un verre » est décomposable en mouvements plus élémentaires (« déplacer sa main au verre », « fermer la main sur le verre », « soulever la main »...). Mais « prendre un verre » peut lui-même faire partie d'une activité plus complexe, comme « manger un repas ». Ces différentes échelles de temps laissent transparaitre la notion de composition temporelle du mouvement, aussi appelée granularité temporelle. Dans cette thèse, nous nous focalisons sur la reconnaissance du mouvement élémentaire.

Les propriétés dynamiques du mouvement peuvent également servir à catégoriser le geste [Huang1995]. Deux catégories sont opposées : les gestes statiques, composés d'une unique posture-clé, et les gestes dynamiques. Ces derniers sont systématiquement décomposables en 3 phases [Quek1994] : l'amorce, le geste, le repli (*preparation*, *stroke*, *retractation* en anglais). Durant la première phase, le corps quitte sa position de repos et se prépare à exécuter le geste. La seconde correspond au geste proprement dit. La troisième ramène le corps à sa position de départ, en équilibre dynamique. [Marr1982] utilise pour la première fois ce type de décomposition, pour segmenter temporellement les périodes d'occurrence d'un geste. Toutefois, cette décomposition théorique est rendue beaucoup plus complexe par le phénomène de coarticulation, qui veut qu'un geste influence le suivant dans une séquence temporelle. L'enchainement des gestes n'implique plus systématiquement un retour à une stricte position de repos entre les gestes. Cette problématique intéresse actuellement les recherches sur la segmentation temporelle du mouvement.

Cette partie montre bien la complexité spatiale et temporelle associée à la réalisation d'un geste. Il est donc nécessaire de définir une représentation spatio-temporelle qui facilite le phase de reconnaissance, en dépit de cette complexité. C'est ce que nous abordons dans la section suivante.

1.1.2 Acquisition et représentation du mouvement humain

De nos jours, la mesure du mouvement humain trouve des applications dans de nombreux domaines et tend ainsi à se démocratiser (biométrie, analyse de performances sportives, monitoring, rééducation, analyse de pathologies, surveillance, production de film, jeux vidéo...).

Du point de vue biomécanique, le mouvement humain est le résultat de contractions musculaires (la cause) entrainant le déplacement de segments corporels (la conséquence). Dans cette section, nous abordons, dans un premier temps, les méthodes de représentation du mouvement humain, qui permettent de l'encoder sous une forme compacte de jeux de paramètres. Puis, nous explorons les outils de mesures disponibles (la capture de mouvements) et les données qu'ils fournissent pour produire ces jeux de paramètres.

1.1.2.1 Représentation du mouvement

Pour le physicien, le mouvement est décrit par l'évolution temporelle de plusieurs paramètres spatiaux que l'on appelle descripteurs (des positions, des angles par exemple). L'ensemble de ces descripteurs est assemblé en un vecteur qui reflète l'état du système à chaque instant. Une observation d'un mouvement est alors décrite par une séquence, de longueur T, d'observations du vecteur descripteur, noté formellement $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$, où chaque vecteur observé $\mathbf{o}(t) = (o_1, \ldots, o_d, \ldots, o_D)^\top (t)$, contient D descripteurs, qui représentent la configuration posturale à l'instant t.

Le corps humain est constitué d'un appareil musculo-squelettique hautement complexe. Il est composé de 206 os mis en mouvement par environ 640 muscles. Une telle complexité structurelle n'est pas accessible aux modèles actuels, même si des recherches s'y emploient activement [Delp2007, Pandy2010] (OpenSim, Anybody). De tels modèles permettent de remonter aux paramètres dynamiques, c.-à-d. aux forces exercées par les muscles et aux contraintes mécaniques subies par les os [Amarantini2004, Bonnefoy2008]. En pratique, modéliser cette complexité n'est pas souhaitable pour la plupart des études. En effet, pour reconnaitre que l'individu A salue B, il n'est pas obligatoirement nécessaire de connaitre la coordination musculaire mise en jeu. L'observation des déplacements du bras est suffisamment porteuse d'information et B reconnait ainsi le salut sans avoir besoin de connaitre les contractions musculaires de A. B n'utilise que des paramètres cinématiques observables.

Cependant, même en laissant de coté l'appareil musculaire, la configuration posturale des 206 os que compte le corps humain n'est pas directement mesurable (cf.1.1.2.3), et ce sont finalement les mouvements de la surface de l'enveloppe corporelle qui sont capturés. De là, il existe globalement deux manières d'appréhender ces mesures pour décrire le mouvement : la description basée capteur, qui utilise directement les données brutes, et la description basée modèle humanoïde, qui, à partir des mesures brutes, utilise une représentation humanoïde simplifiée du corps de l'acteur.

Les descriptions utilisant directement les données fournies par les capteurs de mouvements sont adoptées majoritairement par les études utilisant des systèmes de vision 2D ou des centrales inertielles pour capturer le mouvement en 3D. Ces études extraient des descripteurs issus des données brutes (comme des flux de pixels) qui sont directement utilisés par l'application finale. Nous détaillons comment ces descriptions sont exploitées dans le cadre de la reconnaissance de mouvements dans la section 1.2.2.2. Néanmoins, ce type de description, étant directement dé-

pendant de la mesure, s'avère spécifique au système de mesure, contrairement à l'autre approche utilisant une représentation générique d'humain.

Les descriptions fondées sur un squelette humanoïde polyarticulé utilisent les mesures de surface issues de la capture, pour reconstruire le mouvement de l'acteur sur une structure hiérarchique constituée de parties rigides (les segments), connectées au travers des articulations, dont les degrés de liberté sont contrôlés en qualité (types de liaisons), et en amplitude [Wang1998] (figure 1.3). Elles s'appuient sur des hypothèses simplificatrices pour réduire la complexité structurelle du corps humain, qui passe de 200 degrés de liberté pour l'original, à généralement moins de 50 pour le modèle simplifié. L'objectif des algorithmes de reconstruction est de déterminer, à chaque instant t, la configuration articulaire du squelette polyarticulé (posture) qui soit la plus compatible avec les mesures, afin de conserver les caractéristiques essentielles du mouvement original. La difficulté principale étant que, d'une part, les mesures sont entachées de diverses erreurs (bruit, occultations), d'imprécisions (glissements des capteurs sur la peau) et que, d'autre part, le modèle humanoïde ne rend pas compte de toute la complexité du corps (centres articulaires non ponctuels, degrés de libertés différents). Après reconstruction, le mouvement est entièrement défini par la structure hiérarchique de segments rigides, munie d'un système de coordonnées décrivant les relations spatiales entre ces segments rigides au cours du temps. Plusieurs systèmes de coordonnées ont été proposés dans la littérature.

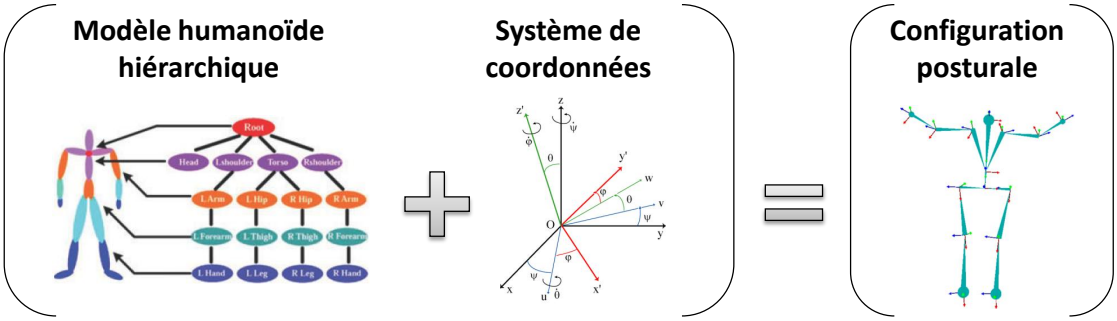


Figure 1.3 - Le modèle humanoïde contient des segments rigides organisés hiérarchiquement, à gauche ([Gu2009]), et le système de coordonnées décrivant la configuration de chaque segment corporel par rapport à son segment parent (angle d'Euler ici), au centre, permettent de décrire entièrement la configuration posturale à chaque instant, à droite ([Kulic2009a]).

Les angles d'Euler, et leurs proches cousins, les angles de Cardan ou les angles de Bryant, sont certainement les plus utilisés, car ils présentent l'avantage d'être intuitifs et compacts (seulement 3 coordonnées). Comme l'illustre la figure 1.4, ils décomposent n'importe quelle rotation R de l'espace en 3 rotations planaires successives $R(\mathbf{x}, \gamma)$, $R(\mathbf{y}, \beta)$, et $R(\mathbf{z}, \alpha)$, autour de 3 axes orthogonaux fixes, tel que :

$$R = R(\mathbf{x}, \gamma)R(\mathbf{y}, \beta)R(\mathbf{z}, \alpha)$$

Les angles de rotation γ, β et α sont souvent nommés tangage, roulis et lacet (*pitch*, *roll*, *yaw*), en référence à leur utilisation historique en navigation maritime.

Pourtant, cette description par les angles d'Euler n'étant pas bâtie sur une structure algébrique [Tournier2011], elle souffre de quelques inconvénients [Faraway2007] : non-commutativité, non-linéarité et apparition de singularités (blocage de cardan) qui rendent impossible la dérivation temporelle ou le calcul d'une norme. La non-commutativité impose d'utiliser une convention dans l'ordre des rotations pour rendre la description unique [Klein2008].

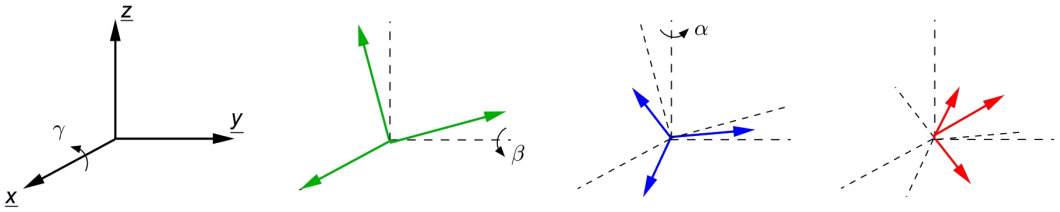


Figure 1.4 - Les angles d'Euler permettent d'exprimer n'importe quelle rotation de l'espace sous la forme de 3 rotations successives d'angles γ, β, α autour de 3 axes orthogonaux $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$. Dans cet exemple, la rotation 3D, qui amène le repère initial (à gauche) dans son orientation finale (à droite), peut être décomposée en 3 rotations successives $R(\mathbf{x}, \gamma)$ ($1^{\text{ère}}$ orientation intermédiaire), $R(\mathbf{y}, \beta)$ (2^{nde} orientation intermédiaire), et enfin $R(\mathbf{z}, \alpha)$ (orientation finale).

En animation, les quaternions sont très utilisés car ils permettent nativement d'interpoler des rotations par des géodésiques. Bien que l'écriture ne soit pas très intuitive, les quaternions peuvent ramener l'expression d'une rotation à un vecteur support, associé à un angle. Ceci donne donc une représentation relativement compacte (4 paramètres). Les quaternions ont également l'avantage d'être construits sur une structure algébrique, ce qui autorise des opérations de dérivation temporelle et de calcul de norme, sans souffrir de singularité. En revanche, il est difficile de réaliser directement des opérations statistiques sur des quaternions [Faraway2007] car seul le sous-espace des quaternions de norme unitaire représente les rotations. Par exemple, la moyenne arithmétique de 2 rotations représentées par des quaternions n'est généralement pas une rotation. Pour remédier aux limitations des quaternions à représenter les rotations, les *exponential maps* ont été proposés pour l'animation d'humanoïdes de synthèses. Il s'agit d'un prolongement des quaternions, utilisant la notion d'espace tangent pour décrire les rotations par un vecteur de 3 dimensions représentant l'axe de rotation, et dont la norme informe sur l'angle [Grassia1998]. Les *exponential maps* offrent donc une compacité d'écriture, mais, de même que les angles d'Euler, ils ne peuvent pas éviter les singularités inhérentes au problème topologique de projection de l'espace des réels sur l'espace des rotations.

Le système de coordonnées polaires/sphériques a aussi été utilisé, notamment par [Raptis2011] pour comparer les performances de danseurs.

1.1.2.2 Variabilité morphologique

Cependant, les méthodologies que nous venons de voir imposent que la reconstruction se fasse sur un modèle humanoïde possédant des caractéristiques morphologiques et anthropométriques identiques à celles de l'acteur. En effet, si on capture un mouvement d'applaudissement sur un acteur adulte pour le reconstruire sur un humanoïde de synthèse enfant en utilisant une description angulaire, on obtiendra nécessairement des problèmes de collision des mains, comme mis en évidence sur la figure 1.5.

Ce problème de variabilité morphologique inter-individuelle est central dans le domaine de l'animation par ordinateurs. Plusieurs réponses ont été proposées pour adapter les mouvements capturés sur un sujet A et permettre d'animer de façon réaliste un avatar B, possédant des dimensions morphologiques différentes. Ces adaptations passent par l'ajout de contraintes sur le mouvement de synthèse et la formulation de lois de comportement sur les chaines cinématiques

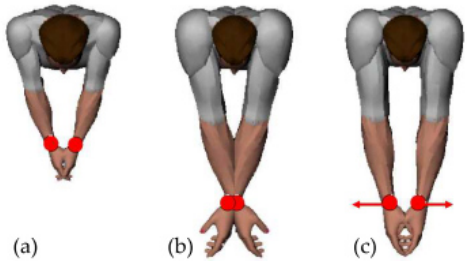


Figure 1.5 - Problème de morphologies différentes [Kulpa2005a]. Les mêmes angles sont appliqués sur deux personnages dont la taille des bras diffère (a) et (b). La partie (c) montre la posture qui respecte le contact initial des mains.

qui composent le modèle humanoïde [Gleicher1998, Ménardais2004, Laumond2005]. Pour encoder intrinsèquement la variabilité inter-individuelle, [Ménardais2003, Kulpa2005b] proposent une représentation indépendante de la morphologie du sujet (figure 1.6). La force de cette représentation réside dans le fait qu'elle intègre des données angulaires et Cartésiennes. Elle s'appuie sur un découpage classique de la hiérarchie en 3 groupes segmentaires, mais dont seul le mouvement de l'extrémité effectrice est contrôlé. Le reste de la chaine segmentaire (les articulations intermédiaires) est adapté automatiquement par des méthodes analytiques à partir des positions Cartésiennes des extrémités.

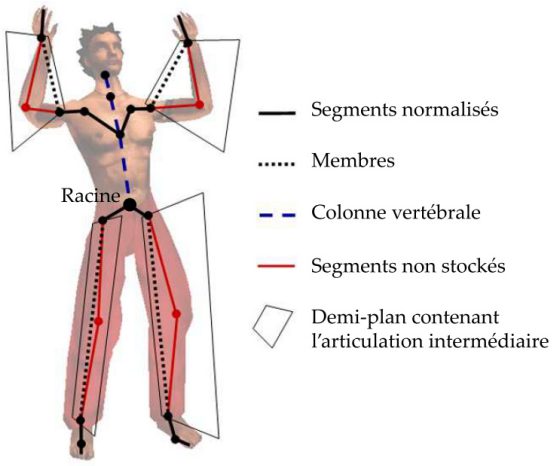


Figure 1.6 - Représentation normalisée d'un modèle humanoïde proposée par [Kulpa2005b].

Dans le domaine de la reconnaissance de mouvements, la variabilité morphologique inter-individuelle n'est pas spécifiquement prise en compte. Un même mouvement peut aussi être exécuté de multiples manières, avec différents styles. Ceci entraine donc deux sources potentielles de variabilité : la morphologie et le style. Ces deux sources de variabilité compliquent considérablement la tâche des méthodes de reconnaissance.

1.1.2.3 Capture de mouvements

Dans la section précédente, nous avons vu que le mouvement humain rend compte de l'évolution de la configuration articulaire du corps à chaque instant. Or, seules des techniques très invasives telles que la radiographie ou le scanner permettent de mesurer directement et avec exactitude la configuration squelettique d'un corps. Les systèmes de capture du mouvements ne peuvent donc mesurer que le mouvement de la surface extérieure du corps. Nous abordons, dans cette section, les techniques les plus utilisées. En dehors de l'obtention de cette description angulaire, certaines approches de reconnaissance de mouvements utilisent directement les données capteurs. Dans cette section, nous montrons les différents types de systèmes de mesure pouvant être utilisés pour ces deux approches.

C'est Etienne-Jules Marey et Eadweard Muybridge qui, dès la fin du 19e siècle, utilisèrent pour la première fois le procédé de chronophotographie afin d'analyser les mouvements humains et déterminer les facteurs intrinsèques de la performance motrice [Marey1894] (figure 1.7, haut). La capture était alors uniquement en 2 dimensions dans le plan de l'appareil photographique. Pour suivre plus facilement les différents segments corporels dans les séquences d'images, Marey perfectionna la technique en positionnant des marqueurs sur les segments corporels d'intérêt (figure 1.7, bas).

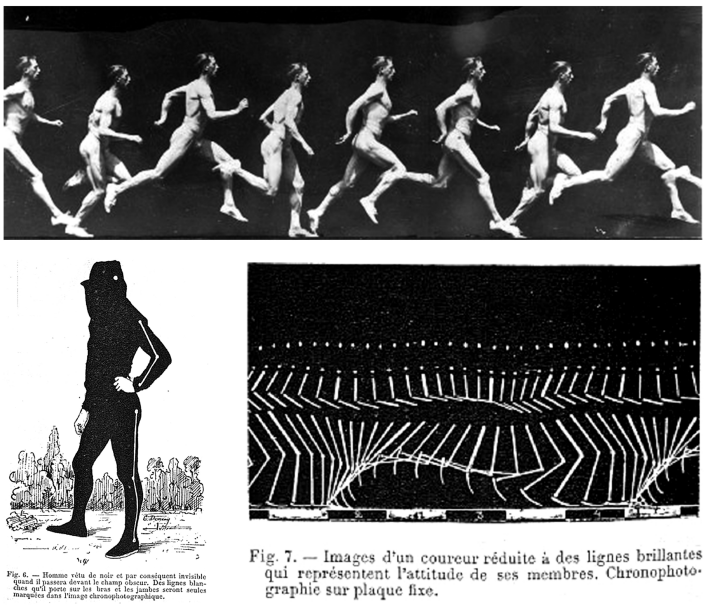


Figure 1.7 - Chronophotographies de Marey [Marey1894]. Analyse de la course, en haut. En bas, la combinaison équipée de marqueurs (à gauche) permet de mettre en évidence la cinématique (à droite).

Si les systèmes de capture de mouvements modernes sont maintenant capables d'établir une reconstruction très fine du mouvement en 3D, le principe, lui, n'a pas tellement évolué depuis Marey et Muybridge. Les articulations et les segments corporels d'intérêt sont repérés par des marqueurs dont le système capte les mouvements. Des algorithmes permettent ensuite de reconstruire le mouvement des marqueurs en 3D puis d'en déduire le mouvement de l'acteur en retournant par exemple un vecteur d'états contenant l'évolution des angles d'Euler ou les quaternions au cours du temps (figure 1.8).

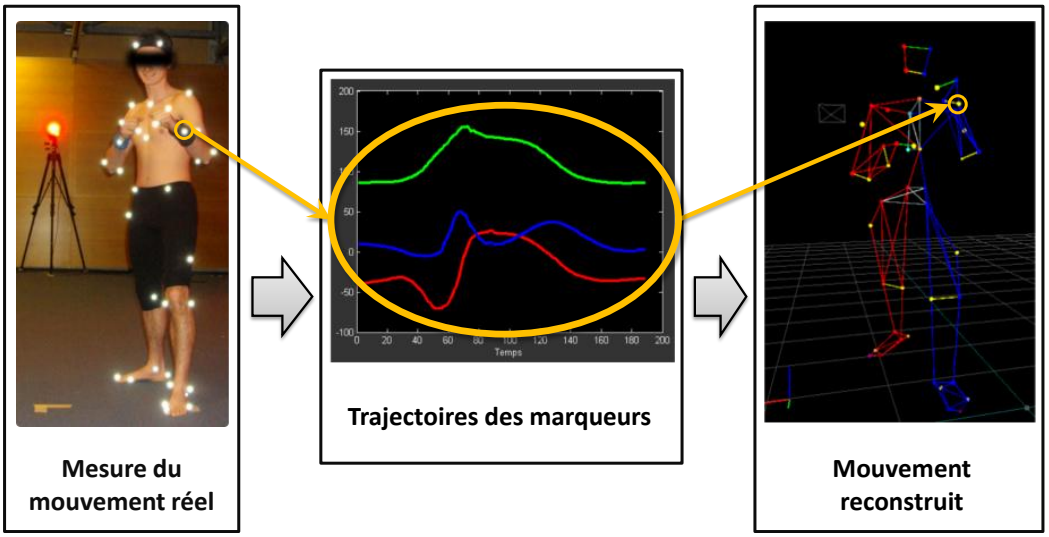


Figure 1.8 - Les 3 étapes de la numérisation du mouvement. Le sujet porte des marqueurs sur des repères anatomiques d'intérêt. Le système de capture enregistre la position de ces marqueurs au cours du temps. Puis il reconstruit les mouvements des marqueurs et en déduit ceux de l'acteur.

À l'heure actuelle, il existe cinq familles de systèmes commerciaux de capture de mouvements. Chacune d'entre elles est fondée sur la mesure d'une grandeur physique de nature différente [Zhou2008b] : les systèmes mécaniques (exosquelettes et goniomètres), les systèmes acoustiques, les systèmes inertiels, les systèmes optoélectroniques et enfin les systèmes magnétiques. Chaque système possède ses avantages et ses inconvénients. Les systèmes mécaniques sont encombrants, surtout lorsqu'il s'agit de capter le corps entier. Les systèmes optoélectroniques sont très précis (erreur de position 3D du marqueur pouvant être inférieure à 1mm pour les meilleurs systèmes [Chiari2005]), ce qui autorise la captation de mouvements des doigts ou d'expressions du visage, sans gêner le sujet. Toutefois, comme ils reposent sur des caméras, ces systèmes sont sujets aux occultations de marqueurs malgré la possibilité de multiplier les points de vues. De plus, ils sont contraints par les conditions lumineuses de capture (reflets, éclairage du jour) et par la dimension réduite du champ de captation. Les systèmes inertiels permettent des captures de mouvement en environnement très peu contraint [Vcelak2006], mais leur nature bruitée et parfois imprécise nécessite de mettre en place des méthodes de compensation [Suthanthira Vanitha2006]. Ces capteurs impliquent d'adapter un mannequin numérique aux mesures afin de reconstruire les descriptions articulaires que nous avons définies précédemment. Les capteurs magnétiques et inertiels ne souffrent d'aucune occultation mais sont sensibles à différentes sources de bruit, comme la présence d'éléments ferromagnétiques.

Récemment, des systèmes vidéo de capture de mouvements 3D, sans marqueur, ont vu le jour. Ils sont, pour la plupart, basés sur une capture vidéo d'une même scène sous plusieurs angles de vue, exigeant une calibration fine. Un algorithme de fusion permet ensuite d'assembler toutes les vues pour reconstruire la scène en 3D et en particulier les acteurs qui s'y trouvent [Corazza2006, Kilner2009]. Il est ensuite possible d'estimer les mouvements d'un squelette humanoïde 3D, soit à partir des volumes reconstruits [Caillette2008, Michoud2009], soit à partir de reprojections du squelette estimé sur les vues 2D [Knossow2008]. La reconstruction 3D du mouvement, à partir de vidéo, peut d'ailleurs être étroitement imbriquée avec des méthodes de reconnaissance de mouvements, en exploitant la granularité temporelle [Peursum2007]. Dans ce cas, la reconnaissance fournit un *a priori* sur le mouvement, qui permet à la reconstruction de réduire l'espace de recherche de la posture courante.

Dernier né des capteurs de mouvement humain, Kinect permet d'obtenir une estimation du mouvement 3D grâce à une caméra de profondeur (PrimeSense) qui extrait la silhouette du sujet mocapé (acteur), pour y projeter un squelette humanoïde. Des algorithmes probabilistes déterminent ensuite la configuration articulaire la plus compatible avec la silhouette courante et avec les précédentes [Shotton2011].

Par ailleurs, même si elle reste en 2D, l'acquisition vidéo monoscopique (descendante directe de la chronophotographie) constitue toujours la méthode de capture de mouvements la plus utilisée, car facile à mettre en œuvre à moindre frais. De plus, ces méthodes proposent parfois une extraction de silhouette et une reconstruction probabiliste du mouvement [Difranco2001, Eian2002]. Le processus de reconstruction peut d'ailleurs tirer parti de la reconnaissance de postures et de mouvements élémentaires (primitives) dans l'image, et inversement [Jenkins2007, Yao2012].

L'apparition récente de la Kinect de Microsoft a démocratisé un nouveau type de capteur, utilisant non seulement l'image mais aussi la profondeur à chaque pixel. Ce nouveau type de système résoud une partie des problèmes liés à la vidéo, en particulier la segmentation du fond et la reconstruction d'une silhouette 3D. De nombreux produits sont apparus, utilisant les mouvements de l'utilisateur, les évaluant pour des systèmes de coaching. Cependant, dans la majorité des cas, les systèmes sont soit en attente d'un mouvement en particulier, soit se limitent à un registre assez restreint et très discriminant.

1.2 Méthodologie de la reconnaissance de mouvements

Cette section est consacrée aux méthodes mises en place dans la littérature afin de reconnaitre les mouvements. Ces méthodes s'appuient généralement sur des descripteurs qui sont chargés de capturer l'information pertinente parmi l'ensemble des informations disponibles. En particulier, les méthodes et leurs descripteurs doivent pouvoir tenir compte des variabilités intra et inter-individuelle inhérentes à la réalistion de chaque classe de mouvements. Nous abordons donc les familles de descripteurs les plus couramment utilisés, soit en lien direct avec le système de mesure, soit en s'appuyant sur un modèle de squelette humain, comme nous l'avons vu précédemment. Nous passons ensuite en revue les outils mathématiques utilisés pour modéliser la variabilité du mouvement à partir de ses descripteurs. La dernière section est entièrement consacrée à l'outil le plus populaire : les modèles de Markov à états cachés.

1.2.1 Méthodologie classique de la reconnaissance

La plupart des approches en reconnaissance de mouvements utilisent des méthodes d'apprentissage automatique, qui consistent en 2 étapes. Le système de reconnaissance est tout d'abord entrainé à partir de mouvements dont la classe est connue. Cette étape permet au système de construire un modèle interne de chaque classe de mouvements. Une fois entrainé, le système est capable de reconnaitre un mouvement inconnu en le comparant avec les modèles de classe dont il dispose.

La figure 1.9 présente les différents modules, qui constituent la chaine de traitements utilisée généralement pour reconnaitre le mouvement. Les deux étapes sont assez similaires en début

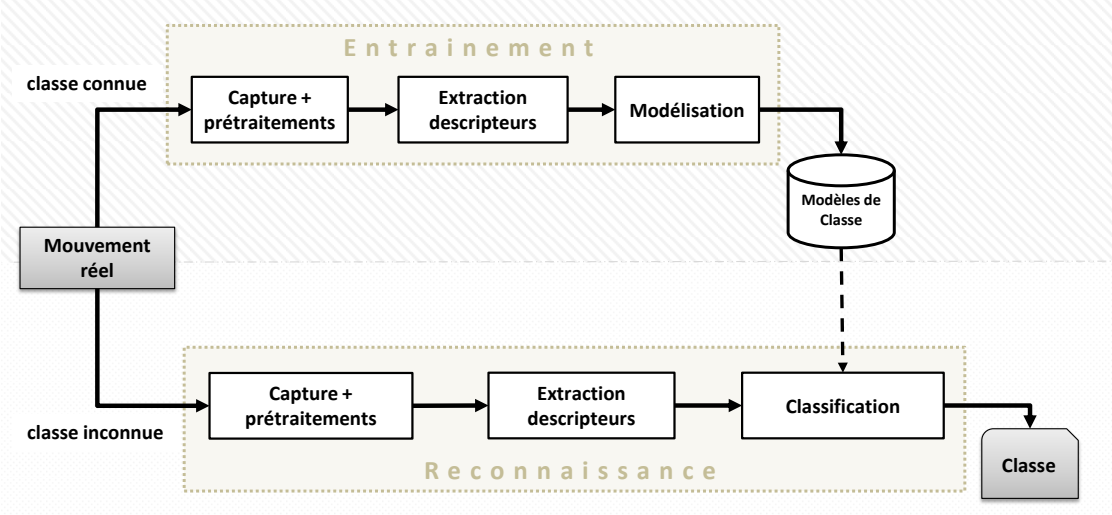


Figure 1.9 - chaine de traitement de la reconnaissance de mouvements. Un mouvement observé est proposé soit pour entrainer le système (chemin au dessus) soit pour être reconnu par le système (chemin au dessous). L'entrainement permet au système de générer ses modèles internes de classe. Ces modèles sont utilisés en phase de reconnaissance pour déterminer la classe d'un mouvement observé.

de chaine. Il s'agit d'extraire des informations numériques grâce à un système de capture de mouvements. Le module suivant est chargé d'extraire les descripteurs du mouvement, à partir des données brutes issues des capteurs. Il peut s'agir de données angulaires issues de la reconstruction 3D du mouvement capturé, des pixels composant la silhouette de l'acteur dans une vidéo, ou encore des trajectoires de certains points caractéristiques.

Après ces modules chargés de la description des données, les étapes d'entrainement et de reconnaissance divergent. L'objectif de l'entrainement est de caractériser chaque classe de mouvements pour générer des modèles capables d'encoder la variabilité intrinsèque à chacune d'entre elles. L'étape de reconnaissance, proprement dite, peut ensuite exploiter ces modèles pour déterminer à quelle classe appartient un mouvement inconnu.

A ce stade de la discussion, il est utile d'aborder la question de la segmentation temporelle. En effet, les événements que l'on cherche à reconnaitre peuvent intervenir à n'importe quel instant dans le flux de données, et avoir une durée variable. Il est donc nécessaire de découper le flux de données entrant en segments temporels, avant de pouvoir reconnaitre les mouvements qui le composent.

Tout comme l'extraction des descripteurs du mouvement, en amont, et le module de classification, en aval, le module de segmentation est déterminant dans le processus qui mène à la reconnaissance [Guenterberg2009]. C'est encore plus vrai en temps réel. Cependant, si le rôle de chaque module est bien défini en théorie, en pratique, la frontière est plus difficile à situer. En effet, extraction, segmentation et classification sont souvent étroitement imbriquées. De nombreux aller-retours permettent d'affiner le travail de chaque module. [Keogh2001a], dans sa revue de la question, recense 3 catégories d'algorithmes de segmentation :

- la fenêtre glissante, qui consiste à observer le signal dans une fenêtre temporelle de durée

fixée, puis à décaler la fenêtre au fur et à mesure que de nouvelles valeurs du signal sont disponibles. Sur chaque fenêtre, l'algorithme de reconnaissance détermine la classe du mouvement observé.

- l'approche descendante (ou *top-down*), qui consiste à partitionner l'intégralité du signal en morceaux de plus en plus petits jusqu'à atteindre un critère d'arrêt (seuil sur l'erreur, nombre de segments). La partition consiste à détecter des frontières naturelles dans les signaux (discontinuités, extrema sur des accélérations, changements de courbure...).

- l'approche *bottom-up*, qui consiste à agréger les séquences de configurations posturales semblables en primitives de mouvements.

De nombreux auteurs ont travaillé sur la question de la segmentation temporelle [Kim2002, Fod2002, Barbic2004, Li2005, Lv2006, Kadone2006, Nakata2007, Kwon2007, Kulic2009b, Spriggs2009, Gu2009, Ren2009, Alon2009, Schulz2010]. Cependant, dans le cadre de la reconnaissance de mouvements, la majorité des travaux traite uniquement des mouvements pré-segmentés, comme le fait remarquer [Weinland2011], ou considère, comme [Marr1982], qu'il existe des phases systématiques de repos entre les mouvements. Dans des environnements interactifs, des déclencheurs permettent de mettre l'environnement en alerte pour le préparer à recevoir une information gestuelle (pression d'un bouton avant ou pendant l'exécution du mouvement, passage par une posture de mise en alerte). C'est le parti que nous prenons dans cette thèse dont l'objectif est d'aborder le problème de variabilité dans la reconnaissance de mouvements à proprement parler et pas de travailler sur la segmentation du flux de données.

1.2.2 Extraction des descripteurs

Quel que soit le type de capteur utilisé pour mesurer le mouvement, l'élément le plus important pour reconnaitre sa classe reste de déterminer les descripteurs qui la représentent le mieux, et qui lui soient le plus spécifique possible. Une fois extraits des mesures, ces descripteurs forment un vecteur qui renseigne sur l'état du système (la configuration posturale du corps, ou du segment corporel considéré). On le nomme donc explicitement vecteur descripteur.

Comme nous l'avons évoqué en section 1.1.2.1, les types de descripteurs utilisés pour la reconnaissance de mouvements peuvent être classés en deux grandes catégories :

- les descripteurs reposant sur un modèle de squelette humanoïde, qui décrivent la configuration posturale de l'acteur ou de certains de ses segments corporels,

- les descripteurs extraits directement de l'espace capteur, comme des silhouettes ou des contours de l'acteur sur une image, l'amplitude maximum d'un signal temporel...

Néanmoins, dans bien des cas, le vecteur descripteur qui découle de cette extraction se trouve être de grande dimensionnalité et souvent bruité. Extraire un sous-ensemble de descripteurs-clés permet, alors, d'exprimer le mouvement dans une représentation plus compacte et plus robuste, tout en restant suffisamment expressive [Poppe2010]. C'est là tout l'enjeu de la réduction de dimension évoquée en section 1.2.2.4.

1.2.2.1 Analogie avec la perception visuelle humaine

Le cerveau est un formidable module de traitement de l'information. [Tenenbaum2000] fait remarquer que :

> « *(...) pour percevoir le monde quotidien, le cerveau humain extrait de ses entrées sensorielles d'extrêmement haute dimensionnalité (30 000 nerfs auditifs,* 10^6 *fibres nerveuses optiques) seulement un petit nombre plus acceptable de descripteurs perceptuellement pertinents.* »

Les études en Point Light Display (PLD) initiées par [Johansson1973] attestent en effet qu'à partir d'indices réduits sur le mouvement, des vidéos de points lumineux sur fond noir situés au niveau des articulations d'un acteur vu de profil (figure 1.10.a), le système visuel humain était capable d'associer immédiatement ces points mobiles à un mouvement dit *biologique* (issus d'un être vivant). Mieux, il peut déterminer la classe des mouvements réalisés, des locomotions en l'occurrence. Selon Lange et coll. [Lange2006], ce mouvement biologique nous renseigne à la fois sur la forme globale et sur la forme locale du mouvement (figure 1.10.b et c). Le PLD permet donc

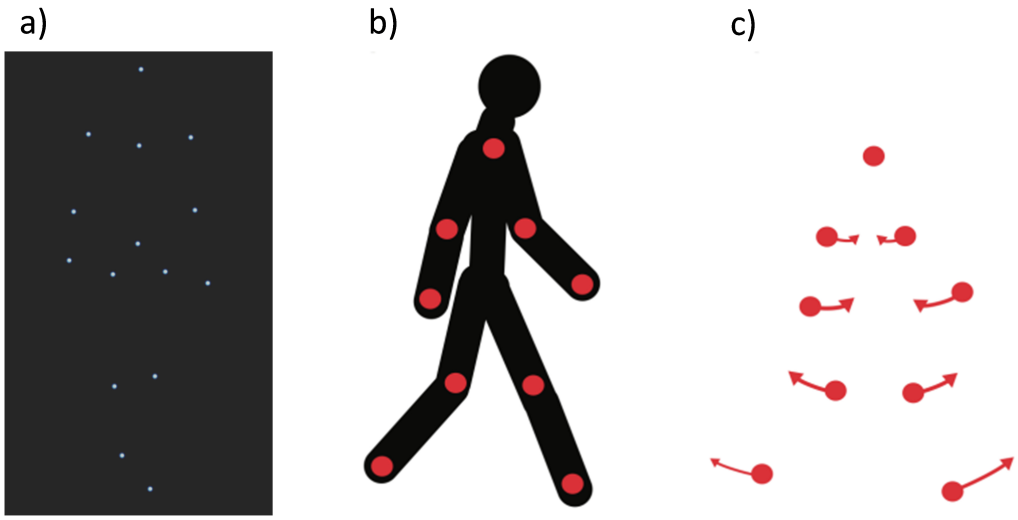


Figure 1.10 - a) Représentation d'un marcheur en PLD. b) Forme globale du mouvement. c) Forme locale du mouvement [Lange2006].

de déterminer le minimum d'information nécessaire au système visuel humain pour reconnaitre une classe de mouvements. Ainsi, cette méthodologie a pu mettre en évidence l'aptitude de l'humain à reconnaitre le genre [Barclay1978, Hill2001, Kozlowski1977, Runeson1994, Troje2002, Pollick2002], l'identité [Beardsworth1981, Cutting1977, Hill2000, Stevenage1999, Jokisch2006] ou l'état émotionnel [Dittrich1996, Pollick2001, Walk1984, Clarke2005] d'un acteur uniquement à partir de sa représentation en PLD. Cependant, la reconnaissance du mouvement biologique peut être perturbée par différents facteurs. C'est ainsi que le taux de reconnaissance du genre s'échelonne de 46% à 86% en fonction de la classe de mouvements, de l'âge et de l'angle de vue [Barclay1978, Hirashima1999, Kozlowski1977, Montepare1988, Runeson1981, Troje2002]. La variabilité du mouvement induite par certains facteurs trouble donc les modèles internes permettant à l'humain de reconnaitre les mouvements qu'il perçoit.

Cette capacité de reconnaissance par le cerveau est également attestée sur des vidéos dégradées [Cédras1995, Bobick1996]. En sport, des tâches de jugement en environnement virtuel

ont par ailleurs démontré que le niveau de détail de la représentation graphique d'un tireur de handball n'influencait pas la faculté des gardiens immergés à déterminer la trajectoire du ballon virtuel [Bideau2003, Vignais2009]. Tous les indices nécessaires et suffisants sont donc déjà présents dans la représentation en PLD.

Ce rapide tour d'horizon des aptitudes du système visuel humain nous apprend qu'un nombre restreint de descripteurs est suffisant pour reconnaitre le mouvement. Si la complexité du cerveau est évidemment loin d'être accessible aux machines, il n'en fournit pas moins de précieux indices sur la nature des descripteurs d'intérêt pour la reconnaissance automatique de mouvements.

1.2.2.2 Descripteurs de l'espace capteur

Ce type d'approches se réfère aux représentations du mouvement sans modèle explicite, directement à partir des mesures [Polana1994]. Elles utilisent le plus fréquemment des données d'entrées vidéo, qui mesurent un flux d'image 2D, dans le plan d'une caméra (monoscopique) fixe. Les capteurs inertiels sont également de plus en plus utilisés. A l'inverse des captures vidéo, ils sont embarqués directement sur le corps de l'acteur pour en mesurer les mouvements.

Les descripteurs extraits à partir de vidéos 2D visent principalement à modéliser la dynamique spatio-temporelle des pixels en mouvement dans l'image. Ils poursuivent plusieurs objectifs :

- être robustes aux conditions de prise de vue (éclairage variable, couleur de peau, point de vue, occultations propres, occultations liées à l'encombrement de la scène),
- permettre le suivi des mouvements de l'acteur,
- être le plus discriminant possible vis-à-vis des classes de mouvements à reconnaitre.

La prise en compte de la variabilité liée aux conditions de prise de vue entraine des problématiques d'extraction très spécifiques. Les descripteurs obtenus en sortie de ce processus ne permettent pas de dissocier cette première source de variabilité, de celle qui est intrinsèque au mouvement lui-même.

La miniaturisation récente des centrales inertielles, alliée à leur précision croissante et leur coût réduit ont permis l'émergence de nouveaux outils de capture de mouvements 3D [Sakaguchi1996, Mayagoitia2002, Giansanti2005]. Une centrale inertielle est composée d'une combinaison de plusieurs instruments de mesure, généralement un accéléromètre et un gyroscope, qui sont souvent couplés à un magnétomètre (associé au terme « centrale inertielle » par abus de langage). Il est donc possible d'exploiter chacun de ces instruments à des fins de reconnaissance de mouvements.

Les accélérations sont le plus souvent exploitées. En condition statique, la seule accélération subie étant celle de la pesanteur, l'accéléromètre fournit une mesure relative à l'inclinaison du segment corporel sur lequel il est fixé. En condition dynamique, l'accéléromètre renseigne sur la direction du mouvement du segment. Partant de là, de nombreuses études ont cherché à reconnaitre des mouvements naturels tels que le maintien d'une posture, la locomotion, ou certains sports [Bussmann2001, Bao2004, Gallagher2004, Bailador2007, Junker2008, Amft2008, Sorel2009, Altun2010]. Le mouvement segmenté est décrit sur une fenêtre temporelle en traitant en parallèle chaque centrale inertielle de manière à extraire des caractéristiques spécifiques du signal (moments, coefficients d'autocorrélation, transformée de Fourier, décomposition en ondelette [Mantyjarvi2001, Najafi2003, Bao2004]...). Cependant en condition dynamique, les signaux inertiels bruts sont très dépendants du sujet et du segment corporel qui porte les cap-

teurs. Par exemple, les accélérations aux chevilles lors d'une locomotion varient du simple au double entre un enfant et un adulte. Cela rend complexe la reconnaissance de mouvements similaires. En conséquence, sur des mouvements dynamiques courts et similaires, la plupart des applications se focalisent sur la reconnaissance de courtes trajectoires stéréotypées de type : ligne droite, cercle, chiffres... [Benbasat2002, Kallio2006, Liang2009].

1.2.2.3 Descripteurs utilisant un modèle de squelette humanoïde

Les modes de représentation du mouvement utilisant un squelette humanoïde polyarticulé ont déjà été abordés en section 1.1.2.1, dans le cadre général de l'analyse et de la représentation du mouvement. Les descripteurs utilisés plus spécifiquement en reconnaissance sont passés en revue dans cette section.

Pour décrire une configuration posturale à chaque instant, les descripteurs utilisant des angles d'Euler sont certainement les plus répandus. Le corps est représenté par une structure hiérarchique de segments corporels rigides, connectés par des articulations mécaniques parfaites, et dont l'orientation est définie par rapport aux segments parents, par l'intermédiaire d'angles d'Euler (figure 1.3). Par exemple, [Sukthankar2005, Sminchisescu2006, Peursum2007, Wang2008, Natarajan2008] utilisent un vecteur descripteur comportant 23 à 56 angles d'Euler plus la position Cartésienne de la racine de la hiérarchie pour décrire la configuration posturale du corps entier. [Fod2002, Calinon2004] utilisent un modèle de bras réduit à 4 angles d'Euler décrivant les rotations des segments bras et avant-bras dans la sphère articulaire de l'épaule et du coude respectivement. En plus des angles, [Ben-Arie2002, Inamura2004] adjoignent leurs dérivées temporelles. Pourtant, ces angles d'Euler posent de nombreux problèmes liés à leur construction non algébrique (cf. 1.1.2.1). [Tournier2011] montre qu'il est incorrect d'utiliser des opérations matricielles linéaires, telle que l'analyse en composante principale, réalisées chez [Fod2002, Sukthankar2005]. De plus, beaucoup d'études reposent sur l'idée fausse qu'une description angulaire de la posture permet d'encoder intrinsèquement les différences de morphologie inter-individuelles (cf.1.1.2.1). Ces différences viennent donc accroitre artificiellement la variabilité intrinsèque de chaque classe de mouvements.

D'autres travaux ont recours aux données angulaires pour décrire une posture. [Wu2009] a recours aux quaternions. [Brand1997] utilise les coordonnées polaires des mains dans le repère de la tête. Grâce à différentes considérations géométriques, [Raptis2011] décrit l'intégralité de la posture par des coordonnées polaires. [Yang2006] décrit une posture par les projections des angles articulaires sur le plan sagittal, frontal et horizontal. Cependant, aucune de ces études ne propose d'adapter la description pour réduire les différences de morphologie inter-individuelles inhérentes aux données angulaires.

[Liang2008] discrétise chaque sphère articulaire en grandes zones, où peuvent se trouver les segments corporels, comme illustré sur la figure 1.11. Cependant, ce type de description risque d'échouer à différencier des classes de mouvements possédant des propriétés spatiotemporelles similaires.

Les positions Cartésiennes sont nettement moins utilisées, alors qu'elles peuvent totalement contraindre le mouvement effectué, comme toucher un objet, lancer quelque chose dans une direction donnée... [Wilson1998, Park2011] recourent aux positions Cartésiennes des mains dans le repère de la tête pour déterminer la direction pointée lors de mouvements de pointage, mais

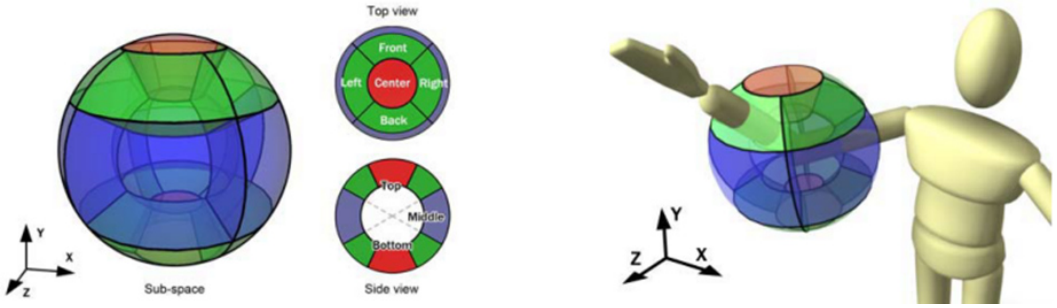


Figure 1.11 - Discrétisation de la sphère articulaire en zones de l'espace [Liang2008]. À gauche, la définition des zones selon différents angles de vue. À droite, un exemple dans lequel le segment avant-bras est attribué à la zone verte.

la reconnaissance n'est pas abordée. [Chai2005, Ott2008] utilisent directement les coordonnées cartésiennes de marqueurs réfléchissants placés sur les articulations d'un sujet pour reconnaitre des configurations posturales. Ces approches permettent un pilotage interactif d'un robot ou d'un personnage virtuel, mais ne proposent pas de reconnaitre une classe de mouvements particulière. [Jin2011] utilise un découpage en 3 groupes segmentaires (bras, jambes et tronc) d'un vecteur descripteur utilisant des positions Cartésiennes. La méthode proposée permet cette fois de reconnaitre des mouvements, mais issus de classes assez discriminées. Cette méthode est basée sur une réduction de dimension des vecteurs descripteurs, mais n'aborde pas le problème de variabilité morphologique pendant l'étape de description. De la même manière, [Lv2006] s'appuie sur un découpage d'un vecteur descripteur issu de positions Cartésiennes, sans traiter le problème de variablilité morphologique. Pourtant, les descripteurs Cartésiens sont très impactés par la variabilité morphologique, puisque la taille des segments corporels influe directement sur l'amplitude des déplacements des extrémités des segments corporels. En ne traitant pas cette variabilité dès l'étape de description, on accroit encore une fois la variabilité de chaque classe de mouvements de façon artificielle.

[Campbell1995] a proposé une description hybride, qui utilise à la fois des angles d'Euler et des positions cartésiennes pour extraire des descripteurs de plus haut niveau, basés sur la courbure des trajectoires articulaires. Seulement, les mouvements à reconnaitre sont très contraints (danse classique) et la population ne contenait que 2 sujets. Toujours sur des mouvements très stéréotypés (Tai-chi), [Campbell1996] montre que les vitesses linéaires permettent d'obtenir un meilleur taux de reconnaissance que les positions et que l'invariance de l'orientation d'un segment corporel permettait à elle seule d'obtenir une performance satisfaisante. Ces résultats demandent à être confirmer sur des mouvements plus naturels et plus dynamiques, qui sont propices à plus de variabilité.

[Caridakis2010] compare les performances de reconnaissance de mouvements 2D d'une description basée sur la position de la main à celle basée sur la direction angulaire de sa trajectoire. L'étude aborde la variabilité inter-individuelle et conclut que, bien que chaque type de descripteur permette une reconnaissance satisfaisante, la plupart des mouvements nécessitent une combinaison appropriée pour obtenir une reconnaissance qui soit robuste et fiable. Malgré tout, ces résultats sont difficiles à extrapoler au mouvement 3D, dans la mesure où, dans ce cas, plusieurs articulations participent au mouvement et que la complexité des trajectoires articulaires est décuplée par cette dimension supplémentaire [Gielen2009].

Vu la multitude de descripteurs disponibles, [Kulic2009a] propose de comparer les capacités de descripteurs Cartésiens, utilisant des quaternions et des angles d'Euler, à segmenter temporelle-

ment le mouvement. Cette étude démontre que les coordonnées Cartésiennes, composées de 39 DdL [2], correspondent mieux aux critères de segmentation humains, atteignant 91% de précision, alors que les angles d'Euler (41 DdL) et les quaternions (51 DdL) fournissent respectivement des taux de 83% et 81%. Cependant ces résultats sont obtenus pour un seul sujet et on peut s'interroger sur leur application à de multiples sujets ayant des morphologies et des styles très différents.

A partir de reconstruction 3D du mouvement, [Müller2006] dérive 39 descripteurs fondés sur des relations géométriques entre différents segments corporels, comme « la main droite est au-dessus du cou », « le pied droit est derrière la jambe », « la main gauche se déplace vers l'avant » ou encore « l'humérus droit est en abduction ». La figure 1.12 illustre quelques uns de ces descripteurs. Pour chaque classe de mouvements, les méthodes de reconnaissance permettent ensuite de déterminer les valeurs attendues pour chaque descripteur, sous la forme d'un modèle (appelé *motion template*) qui sert à indexer le mouvement dans la base de données. Ce type de descripteur présente l'avantage de décrire le mouvement de façon totalement indépendante de la morphologie du sujet. Cependant, définir un tel ensemble de descripteurs qui soit générique à toute classe de mouvements semble impossible. En particulier, ces descripteurs doivent être choisis avec soin si les classes de mouvements possèdent des caractéristiques spatiotemporelles similaires.

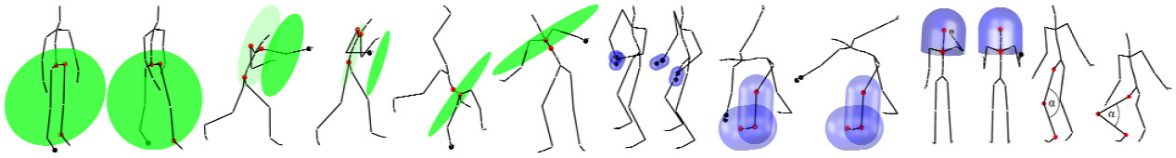


Figure 1.12 - Descripteurs qualifiant les relations géométriques entre différentes articulations (marquées par les points rouges et noirs) [Müller2005].

En définitive, très peu d'études en reconnaissance de mouvements ont abordé la problématique de la variabilité morphologique, comme le fait remarquer [Turaga2008]. Cela vient s'ajouter à la variabilité intrinsèque du mouvement, qui peut être déjà importante dans de nombreux exemples. Au final, c'est la tâche de reconnaissance qui se trouve complexifiée.

1.2.2.4 Réduction de dimension du vecteur descripteur

Le corps humain disposant d'un très grand nombre de degrés de liberté, l'extraction de descripteurs du mouvement mène souvent à un vecteur de très grande dimension. Il est alors nécessaire de réduire l'espace des descripteurs à un sous-espace de plus faible dimension, tout en conservant un maximum d'informations utiles. Or, compte-tenu des synergies et des coordinations motrices, le sous-espace effectif dans lequel évolue le mouvement humain est potentiellement de faible dimension. [Bernstein1967] montre en effet que l'homme contrôlerait des combinaisons de degrés de liberté de façon simultanée, réduisant ainsi le nombre de variables de contrôle. Toutefois, si ce constat nous montre que le sous-espace existe, il n'est pas trivial de l'expliciter. Des méthodes mathématiques de traitement de l'information permettent cependant de définir des sous-espaces de dimension réduite dans lesquels projeter l'espace originel.

2. Degrés de Liberté

L'approche la plus simple consiste à sélectionner les descripteurs en fonction de leur pertinence. De nombreuses méthodes ont été proposées dans la littérature et sont recensées par [Kohavi1997]. Mais cette approche implique la suppression d'une partie des descripteurs jugés non pertinents et par la même des informations qu'ils contiennent. C'est pourquoi, des approches moins destructives sont préférées. L'analyse en composante principale (ACP) est une méthode linéaire de réduction de la dimension très largement répandue en reconnaissance de mouvements [Wu2001, Fod2002, Masoud2003, Daffertshofer2004, Barbic2004, Bashir2005, Calinon2005, Lu2006, Coogan2006, Carvalho2007, Taniguchi2011, Jin2011]. Il s'agit d'observer les corrélations entre les descripteurs et de déterminer l'espace propre possédant les directions de variances maximales. La décomposition de la matrice de covariance en valeurs et vecteurs propres permet de déterminer cet espace de manière unique. Les données sont alors réexprimées dans cet espace que l'on réduit aux premières composantes qui supportent la majeure partie de l'information.

Le mouvement étant un processus essentiellement non linéaire, les techniques de la famille des ACP peuvent échouer à détecter la dimensionnalité intrinsèque du vecteur de descripteurs [Tenenbaum2000]. Pire, l'ACP est théoriquement inapplicable sur des descripteurs aussi répandus que les angles d'Euler [Tournier2011]. De plus, l'ACP ne tient pas non plus compte de la séquentialité des poses qui composent le mouvement. Des méthodes de projection non linéaires ont donc été proposées.

Ces méthodes consistent pour la plupart à discrétiser l'espace de description. Elles font appel à des méthodes de quantification vectorielle classiquement utilisées pour la compression de données. L'idée est de projeter l'espace de description initial sur un espace discret de plus faible dimensionnalité, tout en conservant les propriétés fondamentales de l'espace initial. La figure 1.13 illustre ce concept selon la méthode des cartes auto-organisatrices (Self-Organising Map) [Caridakis2010].

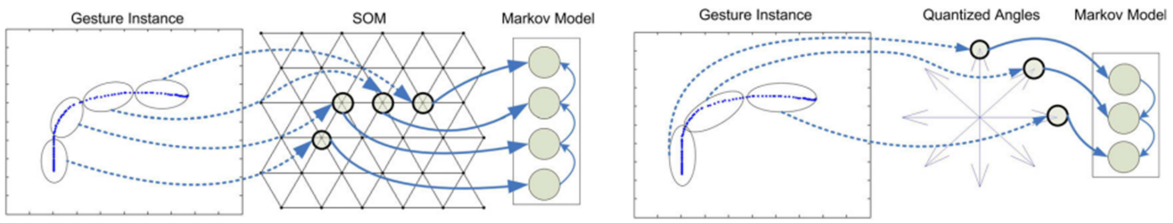


Figure 1.13 - Deux exemples de projection d'un mouvement 2D sur des cartes auto-adaptatives afin de discrétiser la description [Caridakis2010]. A gauche, la projection s'appuie sur des données de position (continues) qui sont projetées sur un maillage (discret), alors qu'à droite, elle s'appuie sur des angles. Ensuite ces descriptions seront reconnues par des modèles Markoviens (voir 1.2.4).

1.2.3 Reconnaissance

Nous avons vu en section 1.1.2 que le mouvement était encodé sous la forme d'une séquence temporelle d'observations d'un vecteur descripteur. Nous abordons à présent les méthodes de reconnaissance de ces séquences. Elles consistent à modéliser la dynamique du vecteur descripteur, pour en établir une signature qui lui soit propre.

Il existe globalement deux catégories de méthodes pour modéliser et reconnaitre une séquence : les méthodes symboliques et les méthodes statistiques [Ramasso2007]. Les méthodes symboliques, proches des modèles de raisonnement humain, reposent sur un langage formel permettant de retranscrire les connaissances d'experts d'un domaine en un ensemble d'heuristiques (connaissances expertes de haut niveau sémantique). Elles gèrent les contraintes logiques et temporelles entre les événements observables. Avec ce type de méthode, la distinction entre une marche et une course se ferait alors explicitement sur l'observation ou non de la phase de double appui au sol, car un expert humain aurait fourni ce critère. Leur principal défaut est de mal gérer l'incertitude. D'autre part, l'explicitation des heuristiques devient vite fastidieuse, dès lors que le nombre de classes à reconnaitre devient grand. Sans compter que pour des mouvements proches, le niveau de précision de ces heuristiques doit être très important.

Les méthodes statistiques ne disposent pas de connaissances *a priori* sur le modèle qui gouverne la séquence, mais seulement de séquences observées. Les séquences observées sont alors utilisées comme données d'entrainement pour déterminer le modèle qui préside à l'évolution du système. La distinction entre marche et course peut alors être construite sur l'observation de critères aussi variés que la vitesse de déplacement globale, la fréquence de pas, ou encore l'absence de baskets aux pieds du coureur. Toute la force de ces méthodes réside dans leur capacité à lier l'état du système à l'observation de ses descripteurs, en tenant compte des incertitudes, tant sur les mesures, que sur l'estimation de l'état.

Comme nous l'avons vu, l'une des principales problématiques de la reconnaissance du mouvement naturel, réside dans sa variabilité spatiotemporelle intrinsèque. Le lieu, l'amplitude et la vitesse d'exécution d'une classe de mouvements naturelle sont variables, d'une réalisation à l'autre. Une classe *salut de la main*, par exemple, va garder son statut de *salut* que le mouvement soit destiné à une personne face à soi ou sur le coté, qu'il soit rapide ou lent, que l'auteur du mouvement ait les jambes croisées ou qu'il soit assis. Contrairement aux mouvements synthétiques de commandes utilisés en Interfaces Homme-Machine ou aux mouvements très codifiés (que ce soit par un art martial, une danse ou un code militaire), le mouvement naturel optimal n'existe pas. En effet, le nombre et l'extrême redondance des degrés de liberté du squelette humain autorise de nombreuses variantes dans l'exécution d'une classe de mouvements. Les critères d'attribution d'un mouvement observé à une classe sont donc flous et les trajectoires empruntées par deux mouvements de classes différentes peuvent se chevaucher, même si les descripteurs sont judicieusement choisis. S'il peut éventuellement exister une notion d'efficience du mouvement, la classe, elle, est principalement associée à la tâche à effectuer (manipulative ou communicative), indépendamment de l'efficacité avec laquelle elle est réalisée. Ainsi, un mouvement naturel est rarement effectué deux fois de la même manière et ce, même par un unique individu.

La deuxième grande source de variabilité est inter-individuelle. Le mouvement s'appuie en effet sur un appareil musculo-squelettique et un système pyramidal de contrôle reposant sur une morphologie et un style naturellement différents d'un individu à l'autre. Cette variabilité amène chaque individu à mobiliser différemment son corps pour effectuer un même mouvement. Au final, ces différences dans les dimensions anthropométriques et dans le style entrainent une dispersion importante dans l'espace des descripteurs que recouvre une classe de mouvements. Cela tend à augmenter encore l'étendue du chevauchement entre des classes de mouvements, les rendant difficilement différentiables.

Ces considérations orientent naturellement les travaux de reconnaissance de mouvements vers des modélisations statistiques, qui sont plus aptes à encoder ces variabilités que les méthodes symboliques. Dans la section suivante, nous passons brièvement en revue la littérature des outils utilisés en reconnaissance de mouvements, avant de détailler, en section 1.2.4, la méthode de référence que constituent les modèles de Markov à états cachés.

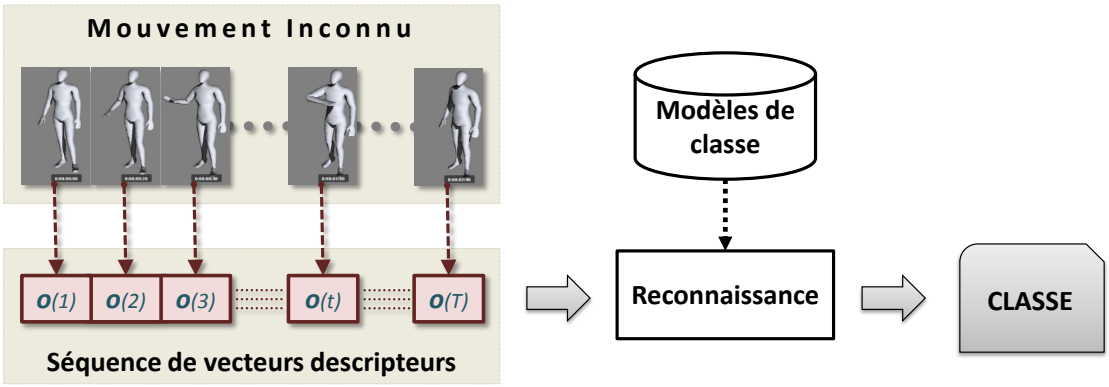


Figure 1.14 - Principe générique de la reconnaissance de mouvements. L'étape d'entrainement préalable permet de générer les modèles de classes.

1.2.3.1 Méthodes de reconnaissance automatique

La reconnaissance de mouvements utilise des méthodes statistiques qui permettent de modéliser les classes à partir d'observations de mouvements qui leur appartiennent. Le déroulement classique suit 2 étapes, que nous avons déjà évoquées sur la figure 1.9.

Tout d'abord une phase d'entrainement, pendant laquelle on présente au système des observations types de chacune des classes de mouvements que l'on cherche à reconnaitre. Pour chaque classe, l'algorithme d'entrainement détermine automatiquement les caractéristiques-clés du vecteur descripteur à partir des observations, puis ajuste en conséquence les paramètres internes du modèle statistique de la classe. Les modèles de chaque classe de mouvements sont ensuite assemblés pour former le système de reconnaissance, souvent nommé classifieur. Dans un second temps, la phase de reconnaissance proprement dite utilise le classifieur afin de déterminer l'adéquation d'une observation de mouvement inconnue $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$ avec le modèle de chaque classe. La figure 1.14 illustre cette phase de reconnaissance.

Les méthodes de modélisation statistique se divisent en deux familles [Bouveyron2006], dont la différence fondamentale est illustrée sur la figure 1.15 :

- les méthodes génératives, dans lesquelles une classe est décrite par les propriétés caractéristiques des objets qui la composent. Ces approches ont l'avantage de bien modéliser la structure intrinsèque des données, ce qui autorise théoriquement la synthèse d'un objet, mais leur prédiction peut s'avérer fortement biaisée [Moeslund2006].
- les méthodes discriminantes, dans lesquelles une classe est décrite uniquement par la frontière qui la sépare des autres classes. Elles ont l'avantage de fournir des règles de décisions optimales, mais sont plus difficilement interprétables [Ikizler2007].

Parmi les méthodes discriminantes les plus utilisées en reconnaissance de mouvements, on trouve les machines à vecteurs de support (Support Vector Machine), aussi nommées séparateurs à vaste marge (SVM), développées par Vladimir Vapnik dans les années 90 [Cortes1995]. Les SVM s'attaquent tout particulièrement au problème des données inséparables dans l'espace de description. En effet, un problème majeur tient au fait que les vecteurs descripteurs décrivant deux classes aux propriétés spatiotemporelles similaires sont très difficiles à discriminer, d'autant que la variabilité accroit la confusion. La figure 1.16.a illustre ce problème de séparabilité des

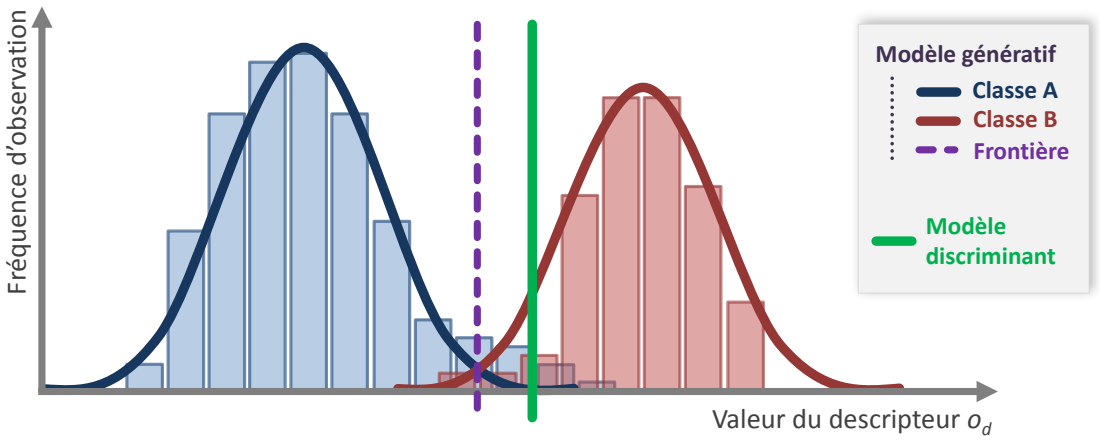


Figure 1.15 - Différence entre les approches générative et discriminante dans le cas d'un descripteur 1D. Les histogrammes représentent la fréquence d'observation de chaque valeur de o_d, pour des mouvements de classe A (bleu) et de classe B (rouge). Les densités de ces valeurs sont modélisées chacune par une fonction de distribution (courbes bleue et rouge). La frontière entre ces 2 modèles de distribution (ligne violette) est moins optimale que la frontière du modèle discriminant, puisqu'elle place beaucoup d'observations de la classe A dans le modèle de la classe B.

données. L'apport majeur des SVM réside dans leur aptitude à déterminer automatiquement un nouvel espace de description dans lequel les classes de mouvements soient séparables de façon optimale [Burges1998, Byun2002]. Ils reposent sur une fonction noyau (fonction radiale, polynomiale...) qui projette l'espace de description original vers le nouvel espace, possiblement de plus grande dimensionnalité. De nombreux travaux en reconnaissance de mouvements recourent à cette méthode [Sukthankar2005, Ikizler2007, Laptev2007, Niebles2007, Fleury2009, Rekha2011]

D'autres approches classiques ont également été utilisées : les arbres de décision [Wu2001, Mathie2004, Huang2007, Francke2007, Ramadoss2008, Lu2009, Lin2009] (figure 1.16.b) et les forêts [Yu2010], l'algorithme K-moyennes et ses dérivés [Zhou2008a], les modèles gaussiens [Ren2005], ou encore les modélisations issues de la logique floue [Ramasso2007, Chan2009]. De même, les systèmes connexionnistes comme les perceptrons [Mantyjarvi2001, Stephan2010] et les réseaux de neurones artificiels [Murakami1991, Yang1999, Yoon2001, Corradini2002, Laxmi2002, Kubota2005, Bailador2007, Yang2008, Arsic2010, Maraqa2012] ont été exploités en reconnaissance de mouvements.

La méthode des *K plus proches voisins* est aussi très populaire en reconnaissance de mouvements [Liu2003, Efros2003, Xi2006, Laptev2007, Vicente2007b, Spriggs2009, Shotton2011]. Pour déterminer la classe d'un mouvement inconnu, l'idée est de regarder le voisinage de son vecteur descripteur dans l'espace de description, puis de lui attribuer la classe correspondant à la majorité de ses K plus proches voisins connus. Comme dans toute méthode de reconnaissance, la principale difficulté est de déterminer une métrique de similarité dans l'espace de description permettant de quantifier les proximités entre les observations de chaque mouvement. Cette question est abordée dans la section suivante.

Les automates à états ont également connu un grand succès dans le domaine [Yeasin2000, Hong2000], pour leur capacité à encoder la séquentialité du mouvement. Dans cette approche générative, chaque classe de mouvements est découpée temporellement en différentes phases

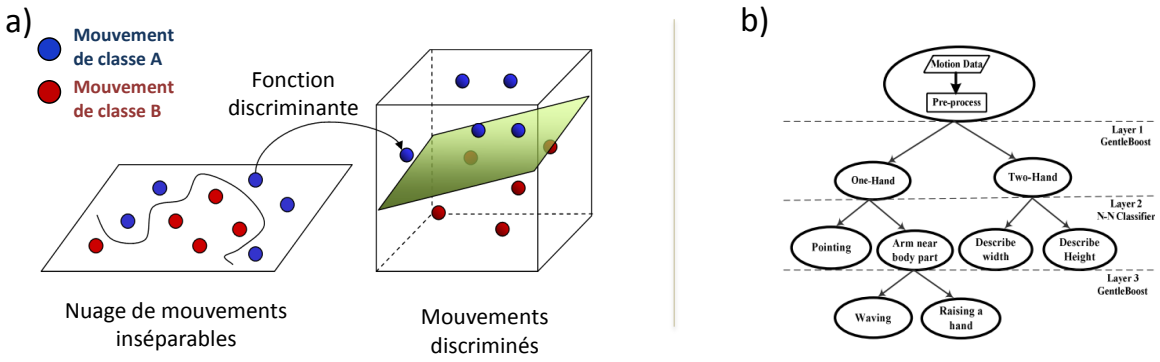


Figure 1.16 - a) Illustration du principe des SVM, qui projettent les données vers un espace où elles deviennent séparables (adapté de `http://www.imtech.res.in/raghava/rbpred/svm.jpg`). b) Arbre de classification utilisé pour reconnaitre des mouvements de bras [Lu2009].

d'initiation, d'exécution et de conclusion qui sont codées sous la forme d'états explicites, comme l'illustre la figure 1.17. L'identification d'une séquence d'états successifs indique la présence d'un mouvement de la classe correspondante. Celle-ci est donc reconnue comme une trajectoire prototypée, un gabarit, dans l'espace de description. Cependant, ces automates sont assimilables à des méthodes symboliques, puisqu'ils fondent leurs décisions sur des heuristiques que le concepteur du système de reconnaissance doit expliciter. Ce côté déterministe limite l'utilisabilité de ces méthodes sur des classes de mouvements de plus grande variabilité.

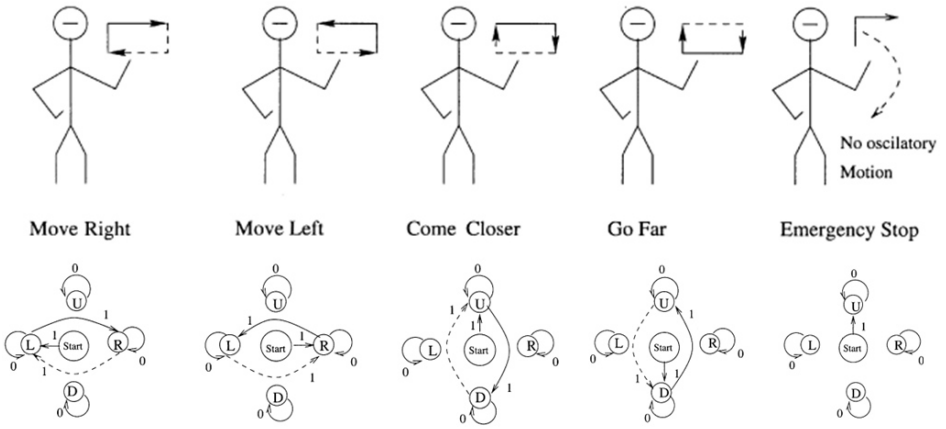


Figure 1.17 - Automate à états finis. En haut, les mouvements de commandes (trajectoire de la main) sont capturés par une caméra vidéo. En bas, l'encodage des trajectoires par un automate à 4 états permet de discriminer les commandes.

Les automates à états peuvent toutefois profiter de l'approche Bayésienne pour gommer cette nature déterministe. C'est ce que proposent les modèles de Markov à états cachés (*Hidden Markov Models*, HMM) [Ghahramani2001], qui sont de loin la méthode la plus utilisée en reconnaissance de mouvements. Les aptitudes de ces derniers à modéliser la séquentialité du mouvement et à tenir compte de l'incertitude sur les observations du vecteur descripteur en font l'outil idéal. C'est pourquoi nous consacrons entièrement la section 1.2.4 à détailler cette approche générative. Plusieurs méthodes dérivent de ce formalisme, comme les Conditional Random Fields [Sutton2006] ou les Maximum Entropy Markov Model qui sont apparues récemment en reconnaissance de mouvements [Sminchisescu2006, Wang2006, Wang2007].

Les filtres particulaires, aussi connus sous l'appellation d'algorithme de *condensation*, sont une alternative au filtre de Kalman [Welch2001] apparus dans les années 80 [Murphy2002, Kwok2004]. Il s'agit de filtres bayésiens récursifs basés sur un processus de rééchantillonnage. Ils n'ont été que récemment appliqués à l'analyse et au suivi de mouvement [Green2004, Peursum2007, Thome2008, Park2011], mais aussi à la reconnaissance [Zhou2004, Gonzalez2006, Kim2007, Yao2012]. Leur succès en fait une approche très prometteuse.

Des approches hybrides combinant plusieurs méthodes de classification pré-citées ont été proposées, raffinant toujours plus les algorithmes et améliorant significativement le taux de reconnaissance. La combinaison de méthodes la plus récurrente consiste à mêler HMM et SVM [Castellani2004, Sukthankar2005, Vicente2007a, Hu2009, Rashid2009]. L'intérêt de ce type d'approche est de coupler l'aptitude des HMM à modéliser la temporalité du mouvement avec le pouvoir discriminant des SVM qui réduit les confusions entre les différentes classes de mouvements. Malheureusement, si ces combinaisons de méthodes s'avèrent plus performantes, elles souffrent d'un manque d'interprétabilité qui peut les rendre complexes à implémenter et à maintenir.

Plus récemment, les méthodes ensemblistes [Sewell2011] ont été appliquées avec succès à la reconnaissance de mouvements. Là encore, il s'agit de combiner des méthodes de classification, mais de manière plus parallèle : les prédictions de plusieurs classifieurs faibles, c.-à-d. dont les prédictions sont au moins meilleures que le hasard, sont pondérées afin d'établir un classifieur fort, c.-à-d. dont les prédictions sont meilleures. Les classifieurs faibles peuvent être issus de méthodes de modélisation différentes ou bien utiliser des vecteurs descripteurs différents. Par exemple, [Ben-Arie2002, Chakraborty2008] entrainent un classifieur par membre, voire par degré de liberté [Lv2006], puis combinent les décisions par vote (figure 1.18), en cascade [Park2011] ou encore par pondération [Xiang2006, Yu2010, Tran2010]. Les algorithmes de dopage (boosting), tel AdaBoost, sont très utilisés, soit pour sélectionner les descripteurs les plus discriminants [Lv2006, Koch2010], soit comme classifieur proprement dit [Liu2010]. Bien que souvent efficaces, ces méthodes souffrent toutefois du même manque d'interprétabilité que les approches hybrides. En outre, leur bon fonctionnement est fondé sur l'hypothèse que les erreurs des classifieurs faibles ne sont pas corrélées entre elles, ce qui en pratique est rarement le cas.

1.2.3.2 Métrique de similarité

De nombreuses métriques gravitent autour des méthodes de reconnaissance afin de quantifier les similarités entre différentes observations d'une classe de mouvements. Les plus élémentaires sont les métriques purement spatiales comme la distance Euclidienne (norme $L2$) ou sa généralisation aux normes Lp de Minkowski. Si le vecteur descripteur représentant le mouvement est assimilable à un point dans l'espace de description, ce type de métrique est bien adapté. Dans le cas d'espaces de description discrets, des métriques du type *edit distance* permettent de calculer la similarité entre des séquences de symboles [Fihl2006, Amft2007].

Bien souvent cependant, le mouvement est défini par une série temporelle à valeurs continues. Pour prendre en compte cette nature temporelle, il est possible d'intégrer les normes de Minkowski sur la durée du mouvement. La figure 1.19, à gauche, présente la distance euclidienne, intégrée instant par instant, entre deux séries temporelles. Toutefois, comme les durées de 2 mouvements sont rarement identiques, il est nécessaire de synchroniser leurs débuts et fins, via

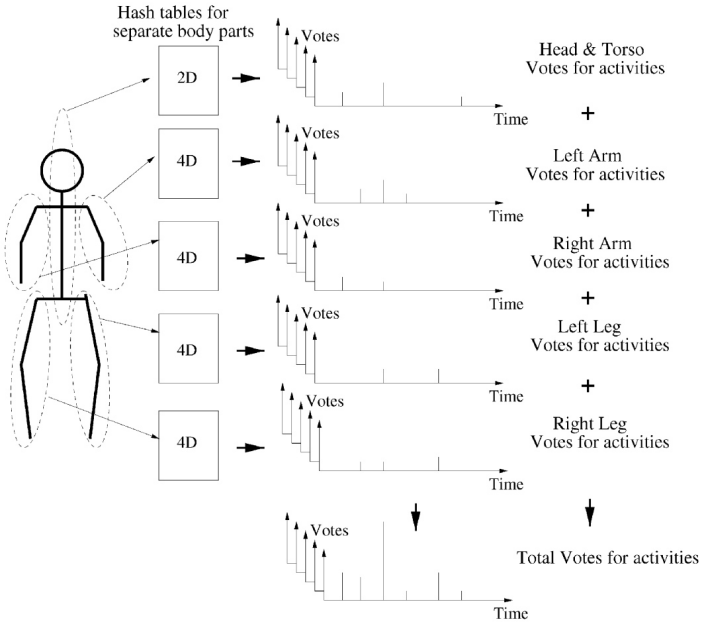


Figure 1.18 - Classifieur fort issu du vote de classifieurs faibles établis sur chaque membre [Ben-Arie2002]

des méthodes d'alignement par dilatation linéaire [Page2007]. Mieux, les méthodes d'alignement temporel non linéaire (*Dynamic Time Warping*, DTW), sont capables de synchroniser les principaux événements à l'intérieur des 2 séries temporelles dont les exécutions varient principalement en vitesse. La figure 1.19, à droite, illustre ce principe de synchronisation préalable à la mesure de similarité.

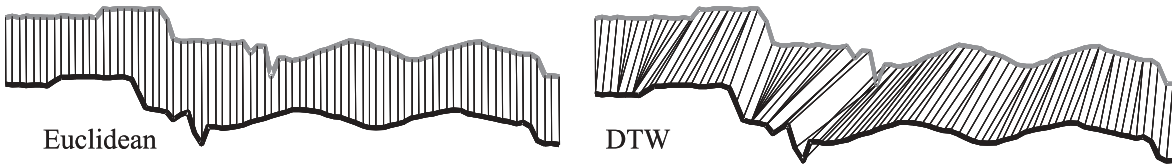


Figure 1.19 - Comparaison de deux métriques de similarité sur des séries temporelles. À gauche, la distance Euclidienne, qui mesure la distance instant par instant, est une métrique qui reflète mal la similarité entre les deux séries temporelles. À droite, l'alignement temporel non linéaire (DTW) est une métrique plus intuitive, qui mesure la similarité entre des instants correspondant aux mêmes événements [Keogh2002].

La similarité peut ensuite être calculée entre les 2 séries alignées, produisant une métrique plus intuitive que la distance euclidienne intégrée. Le DTW est extrêmement utilisé en reconnaissance de mouvements [Müller2006, Blackburn2007, Cherla2008, Paiyarom2009, Zhang2010, Raptis2011]. Des extensions du DTW permettent de tenir compte des dérivées (Derivative Dynamic Time Warping [Keogh2001b]), d'étendre l'approche à des signaux multidimensionnels [Wollmer2009] ou de synchroniser des signaux incomplets au fil de l'eau [Tormene2009].

Pour pallier à la charge calculatoire liée au calcul du DTW, [Keogh2002] introduit la *LB_Keogh*, une borne inférieure du DTW, qui permet par exemple de segmenter une trajectoire spatio-temporelle par des cubes englobants [Anagnostopoulos2006]. L'égalité de Parseval autorise également une mesure de distance dans le domaine fréquentiel [Agrawal1993], via des décom-

positions de Fourier ou en ondelettes.

Des distances entre matrices de covariance, telle la mise à l'échelle multidimensionnelle (*Multidimensional Scaling*, MDS), sont également très prisées pour comparer les corrélations entre les descripteurs qui portent la signature de la coordination inter-segmentaire propre à une classe de mouvements.

1.2.4 Reconnaissance par modèles de Markov à états cachés (HMM)

Dans le cadre de la reconnaissance de séries temporelles, les automates à états en général, et ceux s'appuyant sur le formalisme Markovien en particulier, ont déjà démontré toute leur puissance en matière de reconnaissance vocale. Ils se sont alors naturellement imposés dans le domaine de la reconnaissance de mouvements, seuls ou combinés avec les méthodes précédentes. La nature Markovienne de cette approche générative suppose que les données à un instant t sont uniquement dépendantes des données observées les plus récentes ($t - \Delta t$). Elle permet de modéliser par un processus stochastique les liens spatio-temporels et les liens causaux entre les états ainsi qu'entre les états et les observations.

Une classe de mouvements m est généralement modélisée par un HMM λ_m. La structure globale du classifieur $\Lambda = \{\lambda_1, ..., \lambda_M\}$ est ensuite construite en connectant en parallèle chaque HMM λ_m précédemment entrainé. La classification d'un mouvement inconnu est ensuite effectuée en la confrontant à chaque λ_m. La classe du modèle ayant la vraisemblance la plus importante est choisie comme étant celle du mouvement inconnu.

1.2.4.1 Bases Théoriques

Les HMM sont des modèles stochastiques qui ont été largement utilisés pour encoder des séries temporelles. En effet, leur nature Markovienne, qui lie les observations passées aux observations futures, en fait des méthodes très bien adaptées à la modélisation de données séquentielles. Plus concrètement, un vecteur descripteur variant dans le temps est modélisé par un automate à états dans lequel chaque état correspond à un ensemble de valeurs observables de ce vecteur, tandis que les transitions entre les états permettent de modéliser la temporalité. Les valeurs observables et les transitions entre les états sont gouvernées par des probabilités, ce qui rend les HMM très robustes à la variabilité spatiotemporelle. Les imprécisions de mesure, la variabilité dans l'exécution d'un mouvement ou les occultations sont ainsi gérées intrinsèquement par le HMM.

La figure 1.20 introduit graphiquement les différents paramètres d'un HMM, qui sont exposés dans la suite de cette section.

Rappelons que chaque mouvement est décrit par une séquence de vecteur comprenant D descripteurs mono-dimensionnels variant dans le temps de manière continue (typiquement des trajectoires). Un tel mouvement est une séquence de longueur T, formellement notée $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$, où chaque vecteur descripteur est noté $\mathbf{o}(t) = (o_1, \ldots, o_d, \ldots, o_D)^\top (t)$, à chaque instant t.

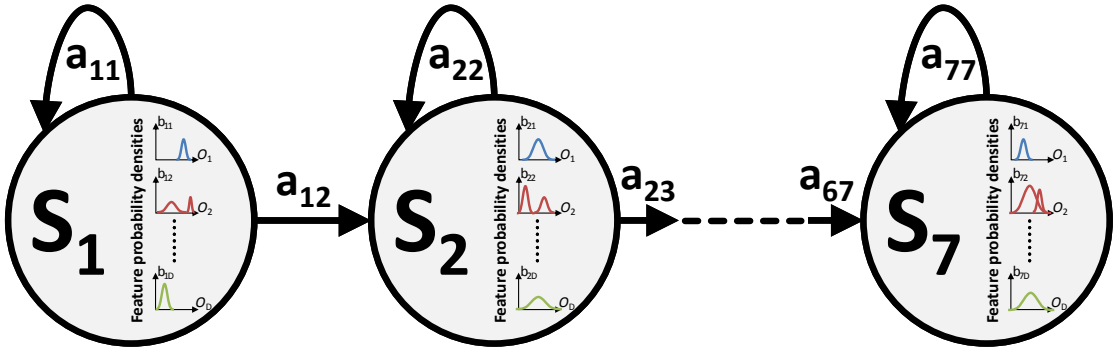


Figure 1.20 - Représentation graphique d'un HMM linéaire à 7 états $\{S_i\}$. Les densités de probabilité b_{ik} de chaque descripteur o_k sont représentées à l'intérieur de chaque état S_i. Un HMM linéaire autorise uniquement les auto-transitions et les transitions de $\{S_i\}$ à $\{S_{i+1}\}$. Ces probabilités de transition d'état sont données par a_{ij}.

Un HMM modélise $\mathbf{O}$ par un processus stationnaire par morceaux, les états, en fournissant des probabilités de transitions entre les états et des probabilités d'observer une certaine valeur du vecteur descripteur dans chaque état [Bishop2006]. Nous notons $q(t)$ l'état qui sous-tend le vecteur descripteur à l'instant t. Formellement, un HMM consiste en un ensemble de N_S états $S = \{S_i\}$ accompagnés d'une matrice de probabilités de transition $A = \{a_{ij}\}$, où a_{ij} désigne la probabilité de transiter vers l'état $q(t+1) = S_j$ depuis l'état courant $q(t) = S_i$:

$$a_{ij} = P(q(t+1) = S_j | q(t) = S_i)$$

avec $a_{ij} \geq 0 \ \forall i, j \in [1, N_S]$ et $\sum_{j=1}^{N_S} a_{ij} = 1$.

Dans chaque état S_i, les valeurs que peut prendre un descripteur o_k sont associées à une probabilité d'observation b_{ik} modélisée par une densité de mélange gaussien, puisque o_k prend des valeurs continues réelles ($o_k \in \mathbb{R}$).

$$\begin{aligned} b_{ik}(o_k(t)) &= P(o_k(t)|q(t) = S_i) \\ &= \sum_{g=1}^{N_g} \omega_g \mathcal{N}(o_k(t)|\mu_g, \sigma_g) \end{aligned}$$

avec

$$\mathcal{N}(o_k(t)|\mu_g, \sigma_g) = \frac{1}{\sigma_g \sqrt{2\,\pi}} \, \mathrm{e}^{-\frac{(o_k(t) - \mu_g)^2}{2\sigma_g^2}}$$

et où N_g est le nombre de composantes gaussiennes, ω_g, μ_g et σ_g sont respectivement le poids, la moyenne et la variance de la g^{ime} composante gaussienne. Comme les données de mouvement possèdent une dimensionnalité élevée, ils génèrent un vecteur d'observations $\mathbf{o}(t) = (o_1, \ldots, o_d, \ldots, o_D)^\top (t)$à chaque instant t. La densité de probabilité globale $b_i = (b_{i1}, \ldots, b_{ik}, \ldots, b_{iD})$ est donc modélisée par une densité de mélange gaussien multivarié à covariance

diagonale.

$$b_i(\mathbf{o}(t)) = P(\mathbf{o}(t)|q(t) = S_i) = \sum_{k=1}^{D} b_{ik} = \sum_{k=1}^{D}\sum_{g=1}^{N_g} \omega_g \mathcal{N}(o_k(t)|\mu_{kg}, \Sigma_{kg})$$

Notons qu'en réalité, des corrélations existent entre les dimensions du vecteur descripteur. Cependant, dans cette thèse, la matrice de covariance Σ est considérée diagonale par commodité calculatoire, comme dans la majorité des travaux de la littérature. Au final, les densités de probabilités $b_i(\mathbf{o}(t))$ d'observation d'un vecteur $\mathbf{o}(t)$ dans l'état S_i sont consignées dans la matrice $B = \{b_i(\mathbf{o}(t))\}$.

Le dernier paramètre nécessaire pour définir un HMM est la distribution de probabilité des états initiaux $\Pi = (\pi_1, \ldots, \pi_i, \ldots, \pi_{N_S})$ avec

$$\pi_i = P(q(0) = S_i)$$

Pour résumer, un HMM est parfaitement défini par l'ensemble de paramètres $\lambda = \{A, B, \Pi\}$.

En pratique, pour reconnaitre une classe de mouvements m, sa dynamique spatiotemporelle doit être modélisé par un HMM λ_m. Il s'agit de la phase d'entrainement du modèle. Pour englober toute la variabilité de la classe gestuelle, l'algorithme d'apprentissage doit disposer d'un grand nombre d'échantillons de cette classe, incluant le plus possible de variabilité (en position, vitesse, amplitude et morphologie).

Au cours de la phase de classification, une séquence gestuelle inconnue $\mathbf{O}$ est classée parmi les modèles entrainés $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$. Pour cela, les vraisemblances de cette séquence avec chacun des modèles de mouvement λ_m sont calculées en parallèle. La classe de mouvement finalement attribuée à la séquence observée $\mathbf{O}$ est celle du modèle qui fournit le maximum de vraisemblance :

$$GestureClass(\mathbf{O}) = \underset{m\in\{1\ldots M\}}{\arg\max}\, P(\mathbf{O}|\lambda_m)$$

où $P(\mathbf{O}|\lambda_m)$, la vraisemblance de la séquence sachant le modèle, exprime la similitude entre la séquence gestuelle observée $\mathbf{O}$ et le modèle λ_m. Ce calcul est résolu par l'algorithme de Viterbi qui, pour un modèle donné, fournit la séquence d'états maximisant la vraisemblance de la séquence observée. La figure 1.21 illustre ce processus de classification.

La théorie que nous venons d'introduire est la plus générique. Elle présente les HMM continus, c.-à-d. utilisant des données à valeurs continues. Toutefois, nous avons vu que beaucoup d'études en reconnaissance de mouvements se ramènent à des vecteurs descripteurs discrets, prenant un nombre fini de valeurs appelées *symboles*. Les HMM discrets permettent de gérer ces symboles par des probabilités d'observation discrètes, et non plus par des densités de probabilités. Cependant, [Caridakis2010] fait remarquer que l'approche discrète, notamment utilisée par [Coogan2006] sur des mouvements dynamiques de la main ne semble pas apte à gérer les variabilités intra- et inter-individuelles.

Dans la pratique, la modélisation par HMM requiert une paramétrisation manuelle du nombre d'états (correspondant à la dimension de la matrice A), du nombre de symboles ou de composantes gaussiennes par état (correspondant à la dimension de chaque matrice B_i relative à l'état S_i), ainsi que de la topologie des transitions autorisées entre ces états (répartition des 0 dans la matrice A).

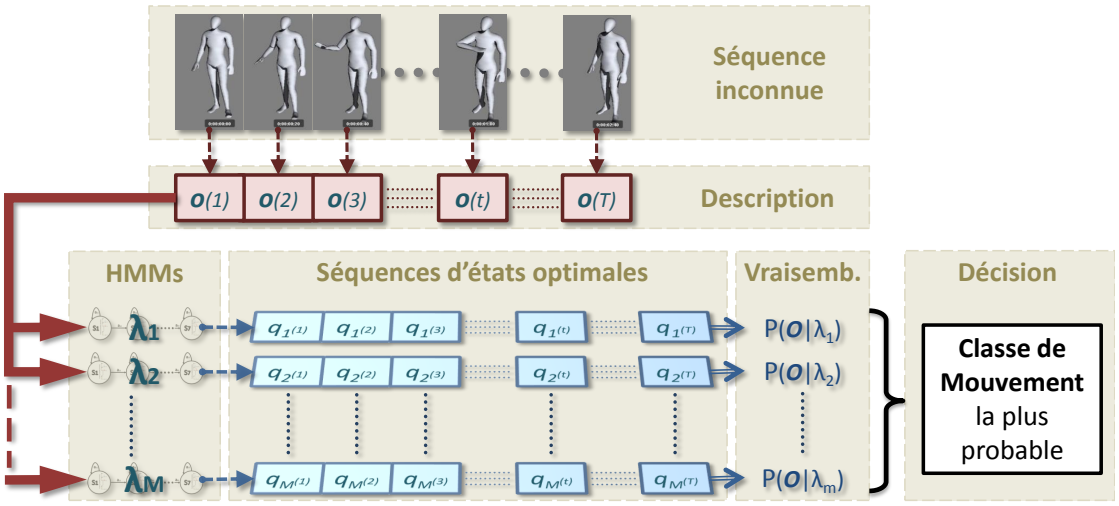


Figure 1.21 - Processus de classification (de haut en bas et de gauche à droite) : 1) Les données brutes correspondant à la capture d'un mouvement inconnu entrent dans le processus de classification. 2) Les descripteurs sont extraits. 3) Pour chaque HMM λ_m, l'algorithme de Viterbi détermine la séquence d'état $(q_m(1, ..., T))$ qui correspond le mieux à la séquence d'observation, et fournit la vraisemblance correspondante. 4) La classe associée au HMM ayant la plus grande vraisemblance est attribuée au mouvement inconnu.

Décider *a priori* du nombre optimal d'états et de composantes gaussiennes par état n'est pas trivial [Bhowmik2011]. Si, récemment, [Cholewa2011] propose une méthode basée sur l'observation du nombre de points critiques (extrema locaux) dans la capture de mouvements, ces choix sont, le plus souvent, déterminés empiriquement par compromis entre complexité calculatoire et efficacité de la reconnaissance [Lv2006].

De même, la topologie des transitions autorisées n'est pas évidente à déterminer automatiquement. La structure ergodique, dans laquelle tous les états sont interconnectés, est la plus générique. Cette liberté de transition entraine une grande complexité calculatoire. C'est pourquoi l'architecture est simplifiée dès que possible en imposant la topologie de la matrice de transition. Les topologies causales, telles les modèles gauche-droite ou Bakis, autorisent uniquement les transitions vers l'avant, c.-à-d. d'un état S_i à un état $S_{j>i}$ (figure 1.22). Elles sont largement privilégiées en reconnaissance du mouvement car elles correspondent bien à sa nature séquentielle. De plus, [Romaszewski2011] démontre que les résultats sont équivalents à ceux de la topologie ergodique pour reconnaitre des mouvements de la main. Cette topologie présente l'avantage d'imposer l'état S_1 comme état initial, rendant connue la distribution d'états initiale $\pi = \{1, 0, \ldots, 0\}$.

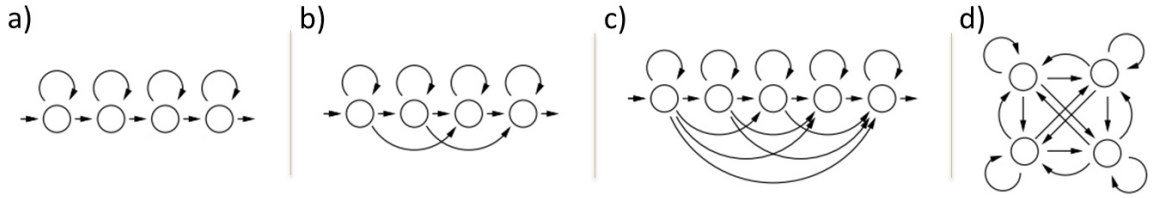


Figure 1.22 - Représentation schématique de différentes topologies d'un HMM [Fink2008]. a) modèle linéaire ; b) modèle Bakis ; c) modèle gauche-droite. Ces 3 topologies sont causales. d) modèle ergodic, où toutes les transitions sont disponibles.

1.2.4.2 Utilisation en reconnaissance de mouvements

Les tous premiers travaux sur la reconnaissance de mouvements par HMM sont à mettre à l'actif de [Yamato1992]. A partir d'une vidéo, l'objectif étant de déterminer un type de frappe de balle au tennis parmi 6 possibles. La silhouette de l'acteur est extraite puis les ratios de pixels appartenant à cette silhouette dans chaque partie de l'image (figure 1.23, gauche) sont transformés en symboles (figure 1.23, droite) et utilisés comme descripteurs. Après entrainement, un HMM discret est capable de reconnaitre à 96% les nouvelles frappes d'un sujet. Cependant, ce taux de succès est uniquement obtenu si les données du sujet ont été utilisées pour constituer la base de données d'entrainement. Ce taux de reconnaisance chute à 61% quand les sujets utilisés pour l'entrainement et la classification sont différents. Il remonte à 71% quand un deuxième sujet est ajouté à la base d'entrainement. Malheureusement, le faible effectif (3 sujets) de cette étude ne permet pas d'établir la significativité statistique de ces résultats. Cependant, ce constat empirique souligne la difficulté de la tâche de reconnaissance liée à la variabilité inter-individuelle. L'étude semble également montrer que plus le système connait de morphologies et de styles différents mieux il serait apte à reconnaitre un mouvement effectué par un acteur inconnu. Cette hypothèse reste cependant à démontrer.

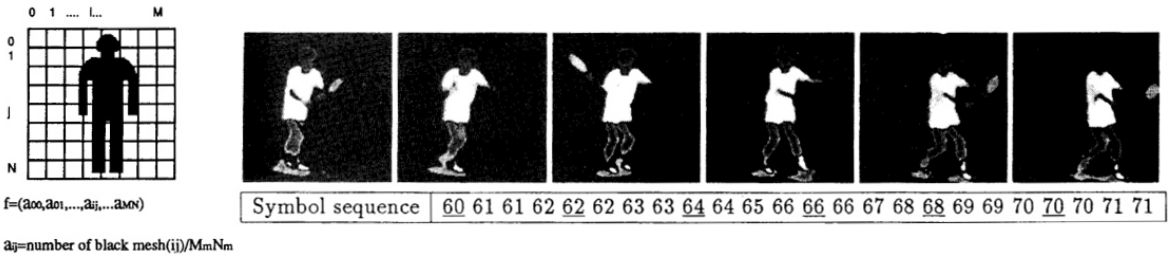


Figure 1.23 - Méthode d'extraction de descripteurs de [Yamato1992]. À gauche, le ratio de pixel contenant la silhouette de l'acteur dans chaque partie de l'image est extrait. À droite, ces ratios sont transformés en symboles, dont les HMM vont modéliser la probabilité d'observation dans chaque classe de mouvements.

La problématique de la variabilité inter-individuelle est encore mise en évidence dans les travaux de [Romaszewski2011]. Différentes paramétrisations (topologie, taille de la base d'apprentissage, qualité de la présegmentation temporelle) y sont expérimentées pour reconnaitre 22 gestes communicatifs et manipulatifs à partir d'un gant de données équipé d'une centrale inertielle.

Alors que le taux de reconnaissance atteint 96% pour des mouvements réalisés par un acteur appartenant déjà à la base d'entrainement, il tombe à 62% quand l'acteur n'en fait pas partie.

Or, comme le souligne [Turaga2008], malgré la quantité de travaux en reconnaissance de mouvements ayant recours au HMM [Stoll1995, Feng2002, Inamura2004, Ilg2004, Oliver2004, Peursum2005, Hossain2005, Lv2006, Jhuang2007, Vicente2007a, Aarno2008, Caillette2008, Chung2008, Kulic2008, Elmezain2009, Caridakis2010], très peu se sont explicitement attelés à réduire la variabilité inter-individuelle. Pourtant, étant donné qu'il est impossible de d'entrainer un système avec l'intégralité de la population, c'est bien l'une des conditions nécessaires pour pouvoir déployer des applications de reconnaissance de mouvements le plus largement possible.

Pour tenir compte des variations dans les conditions de capture (un changement de point de vue par exemple) ou dans les informations contextuelles véhiculées par une classe de mouvement (la direction d'un mouvement de pointage par exemple), [Wilson2002] introduit les HMM paramétriques, aussi repris par [Herzog2008, Axenbeck2008]. Ce type de modèle est le seul capable de modéliser explicitement la variabilité intrinsèque à une classe de mouvement. En plus de la classe, chaque mouvement de la base d'entrainement est labellisé avec la valeur d'un paramètre θ (la direction pointée pour un mouvement de pointage, l'écartement des mains pour un mouvement de dimensionnement...). Ainsi, au lieu de modéliser la variation systématique du vecteur descripteur due au paramètre θ sous forme de bruit comme le font les HMM classiques, les HMM paramétriques sont capables de distinguer les deux variantes de la classe de mouvements et d'en produire un modèle. La figure 1.24 illustre cette méthode pour un mouvement de la classe dimensionnement (représenté à gauche), dont les modèles sont présentés pour différents écartements des mains. Cependant ce type d'approche n'est pas dédié à gérer la variabilité inter-individuelle.

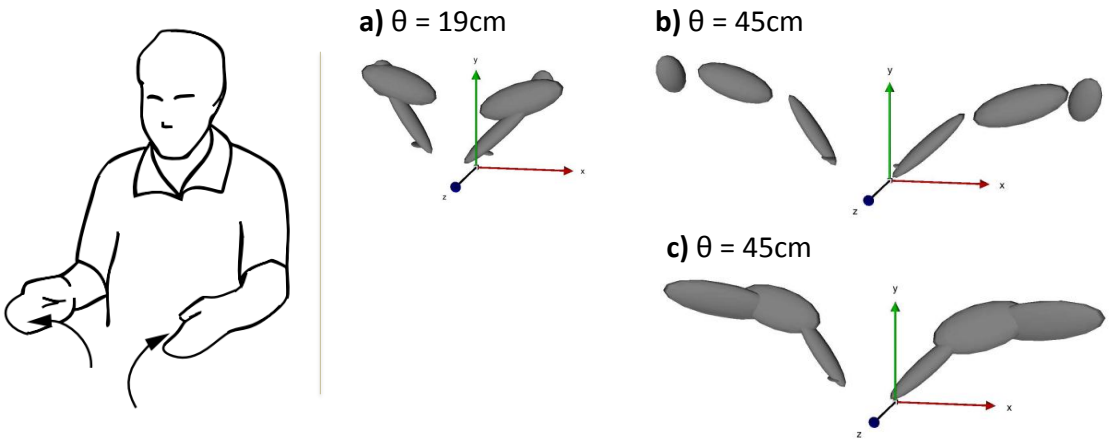


Figure 1.24 - Densités de probabilité de position des mains droite et gauche (ellipsoïdes gris) pour les 4 premiers états d'une classe de mouvements de dimensionnement (représenté à gauche) [Wilson2002]. a) HMM paramétrique avec $\theta = 19cm$ (écart entre les mains) ; b) HMM paramétrique avec $\theta = 45cm$; c) HMM non paramétrique. L'espace recouvert par les HMM paramétrique est plus faible ce qui les rend plus spécifiques.

Dans cette section, nous avons détaillé la chaine de traitement permettant la reconnaissance de mouvements, depuis l'extraction de descripteurs, jusqu'aux méthodes de modélisations des

classes de mouvement. Nous avons vu que de nombreux outils sont utilisés, mais que les HMM constituent la méthode la plus répandue en raison de leur aptitude à gérer de front la variabilité et la séquentialité des mouvements. Cependant, malgré une littérature abondante sur le sujet, très peu d'études ont abordé explicitement la problématique de la variabilité morphologique inter-individuelle.

Synthèse et objectifs

Ce travail s'inscrit dans un projet plus large cherchant à définir et à évaluer des méthodes d'entrainement à des tâches motrices complexes, à partir de systèmes immersifs. Comme nous l'avons vu, cela nécessite de mesurer la performance de l'utilisateur, de reconnaitre le mouvement en cours et de l'évaluer. Dans cette thèse, nous abordons le problème de la reconnaissance de mouvements à partir de mesures 3D, telles que les positions et les angles des segments corporels.

Cette problématique a été principalement abordée pour la reconnaissance d'activités à partir de vidéos, dans le domaine de la vision par ordinateur. Un autre domaine en pleine évolution est la gestion d'interfaces gestuelles s'appuyant, par exemple sur des données tactiles 2D. Avec l'apparition et la diffusion de systèmes de capture de mouvements 3D, le domaine de l'animation par ordinateur s'est lui aussi intéressé à cette question, afin de pouvoir interagir de manière naturelle avec des environnements numériques et de piloter des avatars de l'utilisateur.

Dans tous ces domaines, l'un des problèmes clés reste le prétraitement des données capteurs, qui doit permettre de déterminer un vecteur descripteur le plus représentatif possible du mouvement à reconnaitre. Ces descripteurs peuvent être très proches des données 3D disponibles, telles que les positions Cartésiennes des segments corporels ou les angles articulaires. Cependant, ces descripteurs sont très sensibles aux problèmes de variabilité inter-individuelle. Ainsi deux mouvements identiques peuvent donner lieu à des angles articulaires différents s'ils sont exécutés par deux personnes de morphologie différente. D'autres auteurs ont proposé des descripteurs représentant les principales propriétés géométriques du mouvement, comme le fait d'avoir un bras devant ou derrière le corps. Ces descripteurs sont inspirés des notations de Laban introduites en danse et qui se veulent très descriptives. Cependant, le manque de précision de ces descripteurs ne permet pas de discriminer des classes de mouvements dont les dynamiques spatiotemporelles sont très proches. Ils ont donc principalement été utilisés pour retrouver un sous-ensemble de mouvements d'une base de données, compatibles avec ces descripteurs (appelé *motion retrieval* en animation par ordinateur). De plus, le choix de ces descripteurs peut être totalement lié à l'application et nécessite donc une expertise spécifique à chaque condition d'utilisation. Définir des descripteurs robustes à la variabilité intra et inter-individuelles pour reconnaitre des mouvements reste donc un problème ouvert.

Une fois les descripteurs choisis, il est nécessaire de définir une métrique associée, afin de calculer la distance entre un mouvement à reconnaitre et un ensemble de classes candidates. Cette métrique est, elle aussi, source de problèmes. En effet, les descripteurs sont généralement regroupés au sein de vecteurs de grande dimension, qui rendent difficile le calcul d'une valeur unique de distance. Cette métrique doit permettre de discriminer deux mouvements de classes différentes même s'ils ont des descripteurs pouvant être relativement similaires. La métrique est donc fortement liée à la nature des descripteurs, mais aussi à la méthode mise en jeu pour effectuer la reconnaissance.

Malgré de nombreuses propositions récentes, utilisant des méthodes d'apprentissage automatique, les chaines de Markov à états cachés ou HMM restent les plus utilisées pour leur capacité à capturer l'aspect temporel dans le mouvement. Quel que soit le vecteur descripteur utilisé, la métrique intrinsèque aux HMM repose sur un même formalisme probabiliste. Une fois les HMM choisis, la performance du système de reconnaissance est donc principalement liée aux descripteurs utilisés. Dans le cadre général, ces descripteurs doivent être insensibles aux variabilités intra et inter-individuelles. Dans cette thèse, nous nous focalisons plus spécifiquement sur les variabilités dues aux différences morphologiques entre les utilisateurs. En effet, un système de reconnaissance ne peut pas être entrainé avec les mouvements de tous ses futurs utilisateurs potentiels. Il est donc important que les descripteurs permettent de s'abstraire le plus possible des différences morphologiques entre les utilisateurs. C'est l'objet de la première étude, dans laquelle nous définissons une représentation du mouvement indépendante de la morphologie. Nous évaluons sa capacité à reconnaitre les mouvements d'un utilisateur, qui n'a pas participé à la phase d'entrainement du HMM.

Une autre contrainte de notre travail est le temps interactif. Le mouvement de l'utilisateur doit être reconnu suffisamment rapidement pour animer un avatar ou simplement pour évaluer sa performance et lui retourner ses erreurs afin qu'il progresse. Or, les HMM s'appuient sur l'observation de l'intégralité du mouvement effectué par l'utilisateur. Attendre la fin d'un mouvement avant de lancer sa reconnaissance n'est généralement pas possible dans un système d'entrainement interactif. La deuxième étude s'intéresse donc à évaluer la performance de ce nouveau descripteur pour des méthodes compatibles avec ces contraintes, comme les mixtures de Gaussiennes.

Chapitre 2

Gestion de la variabilité morphologique pour reconnaître les mouvements naturels

2.1 Introduction

De nos jours, une grande variété de dispositifs matériels permettent à l'utilisateur d'interagir avec un environnement virtuel d'une manière naturelle. Des systèmes à faible coût ont été largement utilisés dans l'industrie du jeu vidéo pour interagir directement avec le jeu grâce à des mouvements naturels, tels que le déplacement d'un dispositif tenu dans la main (Nintendo Wii ou Sony PS Move) ou la capture des mouvements du corps entier (Microsoft kinect ou SoftKinetic iisu). Quel que soit le dispositif, l'un des principaux défis consiste à reconnaître le mouvement de l'utilisateur afin de pouvoir calculer une réaction appropriée de l'environnement virtuel en temps interactif. C'est particulièrement le cas dans notre contexte de coach virtuel devant reconnaître et évaluer la performance de l'utilisateur de manière automatique.

La plupart des études antérieures se sont attachées à reconnaitre des mouvements très discriminés, comme la locomotion (marcher, trottiner, courir), des postures particulières (allongé, accroupi, assis) ou des mouvements de bras (coup de poing) [Weinland2011]. Ces mouvements ne mobilisent pas les mêmes parties du corps et semblent donc relativement simples à discriminer car ils présentent de fortes différences dans les contraintes spatiotemporelles [Raptis2011]. Toutefois, la navigation et l'interaction dans des environnements immersifs peuvent conduire à traiter des gestes proches qui engagent le haut du corps, tels que les mouvements liés à des tâches de manipulation, de pointage, de saisie, ou à des actions comme taper, pousser, tirer, frapper... Dans ce cas, les propriétés spatiotemporelles/cinématiques des mouvements sont très similaires puisqu'elles mobilisent les mêmes degrés de liberté du corps. Les descriptions de

chaque type de mouvement occupent alors le même sous-espace, ce qui peut les rendre difficiles à différencier. Dans ce contexte, déterminer les descripteurs les plus pertinents est donc un point crucial pour la reconnaissance. Or, les données de capture de mouvements d'où sont extraits les descripteurs sont fortement liées aux caractéristiques anthropométriques de l'utilisateur : de longs bras entrainent de grands déplacements des capteurs, alors même que la sémantique du mouvement est censée être identique. En plus de cette variabilité morphologique, le vecteur descripteur doit pouvoir tenir compte d'un même mouvement effectué dans différentes parties de l'espace, à différentes vitesses et avec différents styles.

La plupart des travaux antérieurs en matière de reconnaissance de mouvements sont fondés sur des modèles de Markov à états cachés (HMM). Cependant, ces derniers s'appuient sur des descripteurs qui affectent fortement les performances de reconnaissance, en particulier pour des types de mouvements très similaires. Dans cette étude, nous proposons une alternative originale aux descripteurs classiques, qui utilisent des positions Cartésiennes [Kovar2002] ou des angles d'Euler [Kulic2009a], afin de limiter l'impact des variations morphologiques entre les utilisateurs.

2.2 État de l'art

La reconnaissance de mouvements 3D est principalement utilisée pour piloter des humains virtuels (avatars) ou pour interagir avec des mondes simulés. Pour le premier type d'application, de nombreux chercheurs ont travaillé sur des données précises (directement à partir de capture de mouvements) afin d'animer un avatar [Bodenheimer1997, Molet1999]. Ces méthodes visent à calculer les informations requises pour animer chaque articulation de l'humain virtuel, tout en corrigeant certaines imprécisions de mesure. Cependant, dans de nombreux cas, et en particulier dans l'industrie du jeu vidéo, des dispositifs de capture de mouvements moins onéreux et peu précis doivent être utilisés. Avec ce type de données, l'animation d'un avatar est impossible directement et des solutions alternatives sont proposées. La principale alternative consiste à chercher dans une base de données prétraitée, le mouvement qui ressemble le plus à la performance de l'utilisateur. Le problème de reconnaissance de mouvements se substitue ainsi à celui de l'animation directe d'avatars.

Des auteurs ont proposé de contrôler un avatar de manière métaphorique à l'aide des déplacements d'une poupée [Johnson1999]. Ce type d'interface d'animation permet une interaction naturelle de l'utilisateur avec son avatar mais les mouvements de l'utilisateur restent très codifiés et faciles à reconnaitre. [Chai2005] propose d'animer un avatar à partir de quelques caméras et d'un nombre restreint de marqueurs réfléchissants judicieusement positionnés sur le corps d'un acteur. Les signaux de contrôle de basse dimensionnalité sont transformés en mouvements du corps entier en construisant une série de modèles locaux depuis une base de données de mouvements. Certaines de ces techniques prennent en compte les contraintes dynamiques que l'avatar rencontre dans son environnement virtuel pour sélectionner la pose qui les satisfait le mieux [Ishigaki2009]. Ces méthodes offrent des résultats impressionnants , mais elles ont été principalement appliquées à des mouvements très différents les uns des autres, facilitant ainsi la reconnaissance. D'autres auteurs ont introduit une métrique de similarité pour fouiller une base de données de mouvements afin d'en extraire celui le plus proche de la performance de l'utilisateur [Slyper2008], mais sans tenir compte de la sémantique. Ainsi deux mouvements différents mais ayant des propriétés cinématiques proches sont confondus. A l'opposé, d'autres approches permettent de différencier deux mouvements ayant une sémantique différente, au

delà des aspects cinématiques. Une fois la classe du mouvement identifiée, une requête sémantique peut être lancée dans une base de données afin d'appliquer le mouvement correspondant à l'avatar [Shiratori2008, Liang2009]. Cependant, ce type de système est généralement limité à des mouvements simples et très différents tels que des coups de poing, de pied ou des locomotions. Certaines de ces approches proposent des descripteurs géométriques dont le but est d'être indépendant des problèmes de variabilité. Par exemple, un mouvement peut être caractérisé par la position de chaque segment dans les grandes directions de l'espace. Cependant, cette imprécision sur l'état de chaque segment apporte une confusion entre mouvements ayant des propriétés géométriques similaires.

Quel que soit le type d'application interactive, concevoir un système de reconnaissance automatique capable de traiter une large gamme de morphologies d'utilisateurs et de situations immersives est toujours difficile. Un des problèmes majeurs réside dans la nature bruitée et la grande dimensionnalité des données issues de la capture de mouvement. Des méthodes de réduction de dimension, comme l'analyse en composantes principales [Lu2006], des modèles de Markov cachés [Chakraborty2008, Lv2006], des machines à états finis [Hong2000, Ikizler2008], des filtres de Kalman [Ramamoorthy2003], ou encore des filtres à particules [Kim2007, Kwok2004, Gillies2009] ont été utilisées pour remédier à ce problème.

A partir des informations 3D liées au mouvement, les approches décrivent généralement chaque pose par l'intermédiaire des orientations relatives entre les segments du squelette. Le squelette humain est décrit comme une hiérarchie de corps rigides reliés entre eux par des articulations mécaniques parfaites. L'état de l'articulation est alors donné par la transformation géométrique qui relie les deux segments qui lui sont attachés : proximal et distal. En général, cette transformation géométrique est limitée à une séquence de rotations autour des axes principaux, comme le suggère l'International Society of Biomechanics [Wu2005]. Au final, cela revient à fournir 3 angles d'Euler à chaque articulation, pour chaque instant. Cependant, cette représentation, en plus d'être non-linéaire, est aussi fortement contrainte par la morphologie du sujet, en particulier pour les mouvements impliquant des contraintes Cartésiennes avec l'environnement (contacts par exemple). Deux personnes de taille différente auront des angles articulaires différents pour une même contrainte Cartésiennes avec l'environnement, comme le montre [Kulic2009a]. Par exemple, les angles issus d'une personne petite ne correspondent pas à ceux d'un grand lorsqu'il est question de frapper dans les mains, ou de toucher un objet dans l'espace. Une alternative consiste à définir des descripteurs géométriques pertinents, qui captent une partie des informations sémantiques. La plupart de ces techniques réduisent l'espace de description à un sous-espace de plus faible dimensionnalité [Bashir2005, Wang2008], le discrétisent en zones [Müller2005, Liang2008, Deng2009, Liang2010], appelées *symboles*, ou utilisent la notation (type Laban [Yu2005] pour la danse) pour encoder des mouvements complexes d'une manière compacte et efficace.

Certains auteurs ont proposé des descripteurs géométriques indépendants de la morphologie de l'utilisateur. [Müller2006], par exemple utilise des descripteurs binaires associés à la vérification qu'une main est au dessus de l'épaule ou encore qu'un pied est situé devant le corps de l'acteur. Ces descripteurs sont efficacement combinés pour récupérer un ensemble de séquences de poses dans une base de données de mouvements préenregistrés, par l'intermédiaire de requêtes explicitées sous forme de contraintes géométriques. Toutefois, la forte dimensionnalité du vecteur descripteur peut produire des classifications contradictoires de mouvements. En travaillant à un haut niveau d'abstraction géométrique, ces descripteurs ne permettent pas de différencier deux mouvements ayant des propriétés géométriques proches, comme *lancer un objet* et *donner un coup de poing*. Il est alors nécessaire d'optimiser la méthode de classification pour pouvoir distinguer ces subtilités, en particulier en sélectionnant le sous-ensemble de descripteurs

le plus signifiant pour le mouvement étudié [Lv2006]. Plus exactement, cela permet au système de sélectionner automatiquement les descripteurs qui fournissent les meilleures performances. Ces techniques s'avèrent cependant coûteuses en temps de calcul et elles impliquent de longs processus d'apprentissage automatique sur des bases de données conséquentes.

Définir un ensemble générique de descripteurs qui conduisent à des performances et des résultats fiables en dépit des variations de morphologie et de style est toujours une tâche difficile. À notre connaissance, très peu de travaux ont explicitement abordé le problème de la réduction de la variabilité inter-individuelle en reconnaissance de mouvements. Tuaraga et al. [Turaga2008] précise dans son état de l'art que traiter ces variations anthropométriques est toujours un défi important et nécessite une attention particulière en reconnaissance de mouvements. En animation par ordinateur, il existe un problème similaire lorsqu'il est question d'appliquer un mouvement effectué par un acteur à un mannequin virtuel ayant une morphologie différente. Ce problème, appelé *retargetting* est souvent résolu en prenant en compte des contraintes cinématiques qui caractérisent le mouvement exécuté par l'acteur [Gleicher1998, jin Choi1999]. L'erreur due à la variation morphologique est en quelque sorte compensée en adaptant chaque pose afin de respecter ces contraintes cinématiques. Une autre approche consiste à changer de représentation. Plutôt que de travailler sur des angles articulaires, fortement dépendantes de la morphologie, certains auteurs ont proposé une autre représentation, indépendante de la morphologie [Kulpa2005b, Hecker2008]. Cette approche a été appliquée uniquement au problème de *retargetting* mais nous semble très prometteuse pour gérer les problèmes de variabilité morphologique en reconnaissance de mouvements. Dans cette étude, nous proposons donc d'évaluer si ce type de représentation pourrait répondre au problème de la reconnaissance de mouvements multi-utilisateurs.

2.3 Méthodologie générale

2.3.1 Descripteurs du mouvement

Dans cette étude, nous avons décidé d'évaluer la pertinence de trois vecteurs de descripteurs du mouvement pour résoudre le problème de reconnaissance de mouvements naturels. En particulier, nous cherchons à évaluer si une représentation amorphologique, inspirée de celle de [Kulpa2005b, Hecker2008], permet réellement de s'abstraire de la variabilité morphologique.

En effet, l'information la plus importante dans un mouvement est généralement liée à la position de l'articulation distale qui est censée interagir avec les objets de l'environnement. Les articulations intermédiaires dépendent plutôt de la morphologie et du style propres à chaque sujet. Pour cette raison, la représentation amorphologique que nous proposons d'utiliser ne tient pas compte des articulations intermédiaires. D'autre part, pour encoder explicitement la variabilité morphologique, il est nécessaire de s'abstraire de la longueur de la chaîne cinématique contenant l'effecteur. Il s'en suit que les vecteurs 3D $\mathbf{r}^{Bras}$, liant directement l'épaule au poignet, doivent être normalisés par leur extension maximale, c'est à dire par la longueur du bras.

$$\mathbf{r}^{Bras}_{Amorpho}(t) = \frac{\mathbf{r}^{Bras}(t)}{\max \|\mathbf{r}^{Bras}\|}$$

Cette représentation devrait permettre de réduire l'influence de la morphologie : lorsque le bras

est complètement tendu $\left\|\mathbf{r}^{Bras}_{Amorpho}\right\| = 1$ et quand il est à moitié plié $\left\|\mathbf{r}^{Bras}_{Amorpho}\right\| = 0.5$, quelles que soient les données anthropométriques du sujet. Les dérivées correspondantes sont également ajoutées à cette représentation, à gauche $\dot{\mathbf{r}}^{BrasG}$ et à droite $\dot{\mathbf{r}}^{BrasD}$. Au final, à chaque instant t, le descripteur utilisé pour effectuer la reconnaissance s'appuie sur cette représentation afin d'obtenir le vecteur suivant :

$$\mathbf{o}_{Amorpho}(t) = \left(\mathbf{r}^{BrasG}_{Amorpho}, \dot{\mathbf{r}}^{BrasG}_{Amorpho}, \mathbf{r}^{BrasD}_{Amorpho}, \dot{\mathbf{r}}^{BrasD}_{Amorpho}\right)^{\top}(t)$$

L'acquisition de mouvements nous offre des données brutes encodées dans des fichiers BVH. Un descriptif de ce format est disponible dans [Meredith2000]. Ils décrivent les mouvements sous la forme d'une hiérarchie de 19 segments corporels de longueur donnée, pouvant chacun se déplacer autour de 3 degrés de liberté (DDL) de rotation orthogonaux. Soit un total de 57 DDL. A partir de cet encodage, il est possible d'obtenir n'importe quelle autre représentation par l'intermédiaire d'opérations de géométrie. Nous présentons maintenant les descripteurs que nous avons confrontés à la représentation amorphologique pour évaluer un système de reconnaissance fondé sur des HMM :

- Le descripteur angulaire $\mathbf{o}_{Euler}$ composé de 12 angles d'Euler. Dans ce type de descripteur, chaque segment corporel est attaché à son propre repère local $\mathcal{R}_j$ et lié à un segment parent. Dans la posture de référence (appelée *T-Pose*, les bras en croix, à l'horizontale, de part et d'autre du corps), tous les repères sont alignés par rapport au monde (l'axe x orienté vers la gauche, y vers le haut, z vers l'avant). Dans notre étude, seuls les angles définissant l'orientation locale de l'avant bras dans le repère du bras et du bras dans le repère du torse sont pris en compte. Soit 6 angles pour chaque bras, donc 12 en tout. Comme ces orientations sont locales, le descripteur est indépendant de l'orientation globale du sujet.
- Le descripteur Cartésien $\mathbf{o}_{Cartsien}$ est composé de 12 coordonnées cartésiennes de centres articulaires. Dans ce type de descripteur, les coordonnées Cartésiennes de chaque articulation dans le repère du monde permettent de décrire une posture. Dans notre étude, nous retenons uniquement les 3 coordonnées de position du poignet et les 3 coordonnées de position du coude pour chaque bras. Toutes ces coordonnées sont exprimées dans le repère local attaché aux hanches du sujet $\mathcal{R}_{Hips}$, de sorte que le descripteur est invariant par rapport à la direction de l'espace à laquelle le sujet fait face.
- Le descripteur amorphologique $\mathbf{o}_{Amorpho}$ qui repose sur la représentation amorphologique introduite ci-dessus. Comme pour le descripteur Cartésien, les coordonnées du vecteur sont exprimées dans le repère local attaché aux hanches du sujet $\mathcal{R}_{Hips}$, de sorte que son orientation globale n'intervient pas.

La figure 2.1 illustre ces différents descripteurs.

Enfin, les dérivées temporelles de chaque paramètre sont aussi ajoutées à chaque descripteur. En effet, comme le souligne [Campbell1996], les paramètres de vitesse jouent un rôle majeur en reconnaissance. Le nombre de paramètres contenus dans le vecteur descripteur est ainsi doublé, portant le nombre à 24 pour $\mathbf{o}_{Euler}$ (12 angles + 12 taux de rotation de chaque angle) et $\mathbf{o}_{Cartsien}$ (12 coordonnées de position + 12 dérivées correspondantes) et à 12 descripteurs pour $\mathbf{o}_{Amorpho}$ (6 coordonnées normalisées + 6 dérivées correspondantes).

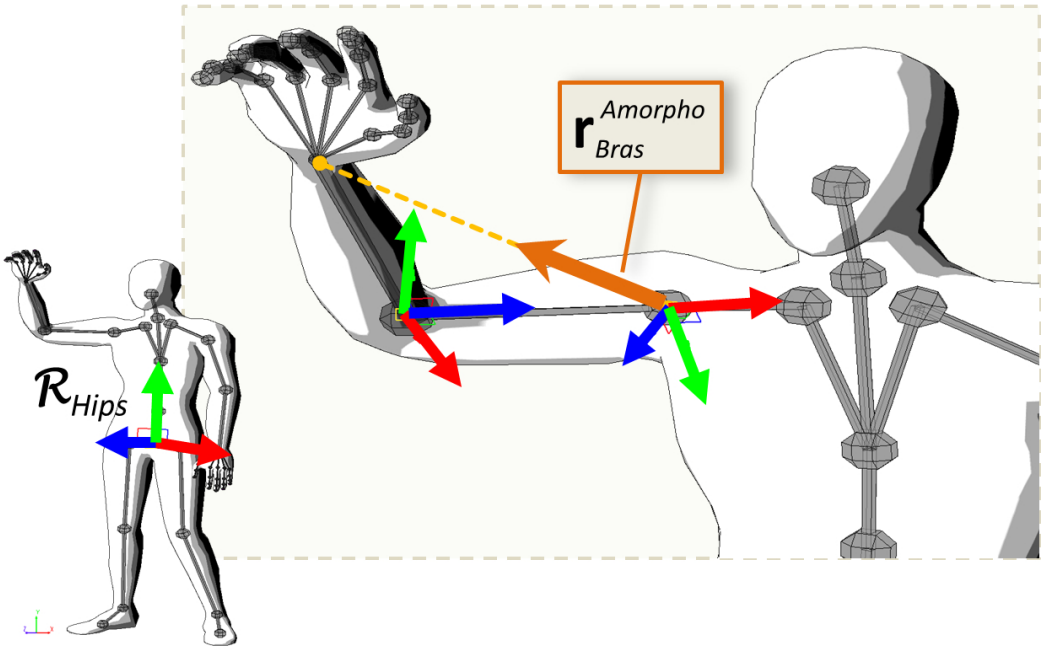


Figure 2.1 - Descripteurs amorphologique, angulaire et Cartésien utilisés en entrée de la classification par HMM. Le descripteur amorphologique englobe deux vecteurs 3D $\mathbf{r}_{Bras}^{Amorpho}$ reliant chaque épaule à son poignet exprimés dans le repère local $\mathcal{R}_{Hips}$, dont la norme est divisée par la longueur totale du bras (pointillés orange clair). La dérivée temporelle de chaque vecteur est également incluse. Le descripteur Cartésien comprend les positions des poignets et des coudes exprimées dans le repère local $\mathcal{R}_{Hips}$, ainsi que leurs dérivées temporelles. Enfin, le descripteur angulaire correspond aux 3 rotations autour du repère local de l'épaule auxquelles on ajoute les 3 rotations autour du coude (les repère locaux sont indiqués par des flèches rouge / vert / bleu).

2.3.2 Protocole

Dans cette étude, nous proposons d'évaluer l'impact de la morphologie sur les taux de reconnaissance de trois classifieurs HMM, utilisant chacun l'un des trois descripteurs proposés ci-dessus. Notre hypothèse est que le descripteur amorphologique est plus performant pour reconnaître des mouvements naturels assez similaires, effectués par des sujets aux morphologies différentes. À cette fin, plusieurs expérimentations ont été menées. Elles consistent à subdiviser la base de données en deux parties : une pour l'apprentissage, l'autre pour l'évaluation. Nous avons ainsi testé deux subdivisions, ou échantillonnages :

- l'échantillonnage aléatoire, qui a pour objectif d'évaluer le taux de reconnaissance global de chaque descripteur en séparant aléatoirement la base de données en deux parties, l'une utilisée pour l'entrainement du classifieur HMM et la partie restante pour la reconnaissance.

- l'échantillonnage *leave-one-out* (L1O), pour lequel l'évaluation s'effectue sur un utilisateur dont aucun exemple n'a servi à la phase d'apprentissage. L'objectif de cet échantillonnage est d'évaluer l'impact de la morphologie sur le taux de reconnaissance associé à chaque descripteur. Cette approche est étendue en isolant successivement 2 sujets (*leave-2-out*, L2O), puis 3 (L3O)... jusqu'à 9 (L9O). Dans ce dernier cas, extrême, les mouvements d'un seul sujet sont utilisés pendant l'apprentissage, et l'évaluation est effectuée sur les mouvements des 9 sujets restant.

Comme ce travail s'effectue dans le contexte d'une application interactive en environnement immersif, nous montrons ensuite un exemple de réalisation de ce type impliquant les trois descripteurs.

2.3.2.1 Base de données

Afin de mener à bien cette évaluation, nous avons conçu une base de données de mouvements naturels réalisés par plusieurs sujets présentant des morphologies différentes. Bien que sémantiquement différents, ces mouvements ont été choisis car ils présentaient une difficulté particulière pour la classification : ils possédaient des propriétés cinématique relativement proches. Nous avons ainsi choisi 15 différentes classes de mouvements naturels du haut du corps (fig. 2.2) : *applaudir*, *croiser les bras*, *gifler avec la paume*, *gifler avec le revers de la main*, *se gratter le menton*, *lancer quelque chose*, *poser les mains sur les hanches*, *mettre les mains dans les poches*, *saisir quelque chose au niveau des hanches*, *saisir quelque chose en hauteur*, *saisir quelque chose au niveau de la poitrine*, *donner un coup de poing*, *saluer ostensiblement (avec la main au-dessus de la tête)*, *saluer discrètement (avec la main à hauteur de tête)* et *donner un uppercut*. Ces classes de mouvements sont représentés en figure 2.2 et l'annexe A en propose une chronophotographie.

En définitive, on peut constater que la base de données comporte des classes de mouvements impliquant les deux bras, de façon symétrique ou latéralisée, et que certains d'entre eux possèdent des propriétés cinématiques très similaires, tels que les *punchs* et les *lancers*. Dans les travaux publiés précédemment, la plupart des auteurs se sont concentrés sur des gestes aux propriétés spatiotemporelles très différentes, tels que marcher, attraper, ramper, courir... Or dans les applications interactives, l'utilisateur est souvent amené à effectuer différentes mani-

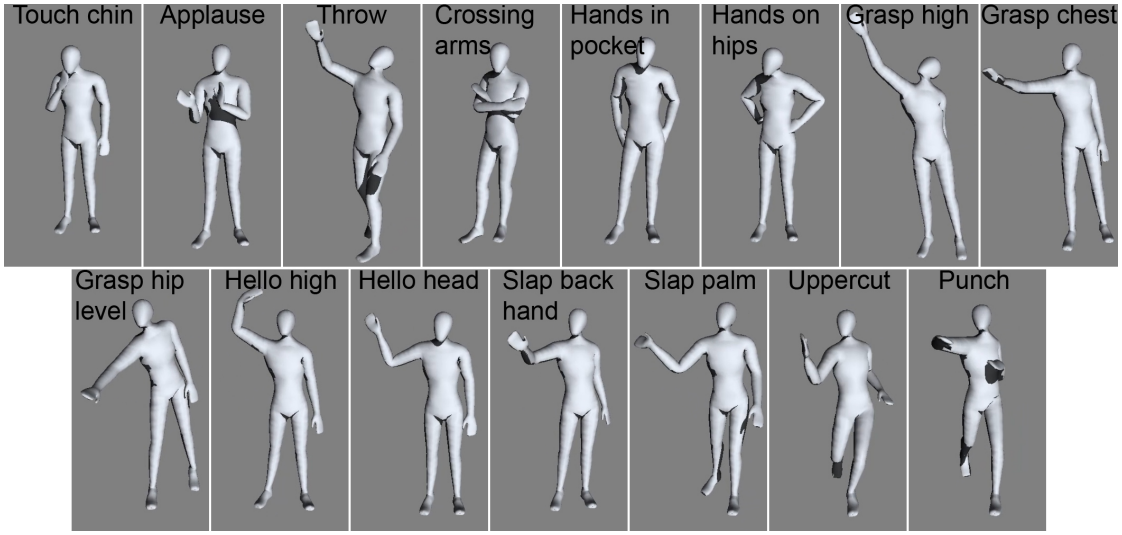


Figure 2.2 - Aperçu des 15 mouvements naturels du haut du corps capturés pour évaluer la performance du système de reconnaissance. Certains de ces mouvements ont des caractéristiques spatio-temporelles très similaires.

pulations avec les membres supérieurs, dont certaines peuvent être très similaires d'un point de vue cinématique. Par exemple, dans le contexte du coach virtuel, il sera important de différencier différentes formes de coup de poing réalisées par l'utilisateur afin de l'évaluer.

Afin de créer la base de données, 10 sujets, dont les caractéristiques anthropométriques sont exposées dans le tableau 2.3.2.1, ont réalisé chaque type de mouvement au moins 5 fois de chaque côté, avec pour consigne d'inclure de la variabilité dans leurs performances (vitesse, position et amplitude). Ces mouvements ont été capturés à l'aide d'un système Optitrack (Natural Point) permettant de mesurer les déplacements de 34 marqueurs réfléchissants disposés sur le corps du sujet, à l'aide de caméras infra-rouge placées autour de la scène (fig. 2.3). L'acquisition a été réalisée à 100Hz. Après le post-traitement (interpolation des données manquantes...) et la reconstruction 3D des mouvements, les données obtenues ont été stockées sous la forme de fichiers BVH. Au final, la base de données contient environ 2300 fichiers correspondant chacun à la réalisation d'un mouvement par un sujet.

2.3.2.2 Définition des HMM utilisés pour la reconnaissance

Dans cette section, nous détaillons comment les descripteurs introduits ci-dessus ont été utilisés par les HMM pour reconnaitre les mouvements de l'utilisateur. Toutes les méthodes décrites dans cette section ont été mises en œuvre en langage C++, au sein d'une unique application. La bibliothèque Ogre3D [1] a été utilisée pour les rendus visuels des mouvements, tels qu'on peut l'observer sur la figure 2.2. La bibliothèque LTI-Lib [2] a été utilisée pour l'entrainement des classifieurs HMM par la méthode segmental k-means [Fink2008].

Les HMM sont des modèles stochastiques qui ont été largement utilisés pour encoder des

1. http://www.ogre3d.org/
2. http://ltilib.sourceforge.net/doc/homepage/index.shtml

Sujets	Sexe	Âge (an)	Taille (cm)
Suj_A	F	16	154
Suj_B	M	27	176
Suj_C	F	27	166
Suj_D	M	23	186
Suj_E	M	26	183
Suj_F	F	21	163
Suj_G	F	22	160
Suj_H	F	30	162
Suj_I	M	28	177
Suj_J	M	29	180
Moyenne		25	171
Écart type		4	11

Table 2.1 - Caractéristiques anthropométriques de la population d'étude.

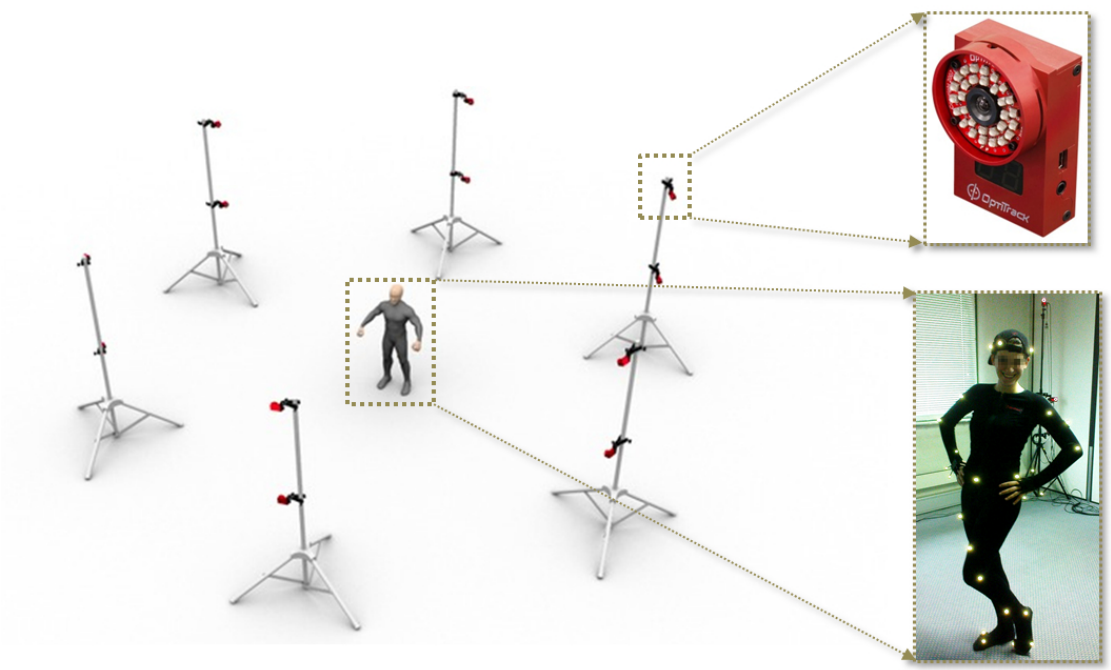


Figure 2.3 - Mise en œuvre de la capture de mouvement. Le sujet dont les mouvements sont capturés est équipé de marqueurs (en bas à droite), qui réfléchissent la lumière infra-rouge émise par des caméras Optitrack V100 :R2 (en haut à droite). Chacune de ces caméras, disposées autour de la scène, capte la lumière IR réfléchie par les marqueurs. En combinant les points de vue de toutes les caméras, il est possible de reconstruire en 3D les déplacements de chaque marqueur et par suite le mouvement du sujet.

séries temporelles. Ils sont totalement définis par l'ensemble de paramètres $\lambda = \{A, B, \Pi\}$ Comme nous l'avons vu, formellement, un HMM consiste en un ensemble de N_S états $S = \{S_i\}$ accompagnés d'une matrice de probabilités de transition $A = \{a_{ij}\}$, où a_{ij} qui désigne la probabilité de transiter vers l'état $q(t+1) = S_j$ depuis l'état courant $q(t) = S_i$:

Dans chaque état S_i, les valeurs que peut prendre un descripteur o_k sont associées à une probabilité d'observation b_{ik} modélisée par une densité de mélange gaussien, puisque o_k prend des valeurs continues réelles ($o_k \in \mathbb{R}$).

Comme les données de mouvement possèdent une dimensionnalité élevée, elles génèrent un vecteur d'observations $\mathbf{o}(t) = (o_1, \ldots, o_d, \ldots, o_D)^\top$ (t)à chaque instant t. La densité de probabilité globale $b_i = (b_{i1}, \ldots,$
$b_{ik}, \ldots, b_{iD})$ est donc modélisée par une densité de mélange Gaussien multivarié à covariance diagonale. Au final, les densités de probabilités $b_i(\mathbf{o}(t))$ d'observation d'un vecteur $\mathbf{o}(t)$ dans l'état S_i sont consignées dans la matrice $B = \{b_i(\mathbf{o}(t))\}$.

Le dernier paramètre nécessaire pour définir un HMM est la distribution de probabilité des états initiaux $\Pi = (\pi_1, \ldots, \pi_i, \ldots, \pi_{N_S})$ avec

$$\pi_i = P(q(0) = S_i)$$

Dans cette étude, chaque classe de mouvement m est modélisée par un HMM continu d'ordre 1, composé de 7 états. Chaque état modélise la probabilité d'observer un vecteur descripteur par une densité de mélange gaussien multivarié, comptant au maximum 6 composantes.

La topologie choisie est de type causale linéaire (fig. 1.20). Elle autorise uniquement les auto-transitions (pas de changement d'état) et les transitions d'un état $\{S_i\}$ à un état $\{S_{i+1}\}$. Cette topologie impose deux contraintes pour tous les modèles de mouvement. Sur la matrice Π, d'abord, l'état initial est nécessairement S_1, ce qui conduit à une uniformisation de la distribution initiale des états pour chaque modèle, qui prend la forme $\Pi = (1, 0, \ldots, 0)$. Ensuite, sur la matrice de transition A, qui se trouve réduite à la forme suivante :

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \ldots & 0 \\ 0 & a_{22} & a_{23} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & a_{66} & a_{67} \\ 0 & \ldots & \ldots & 0 & a_{77} \end{bmatrix}$$

Les matrices b_i, qui composent B, ne subissent pas de contrainte liée à la topologie, puisqu'elles n'interviennent qu'à l'intérieur de chaque état.

La topologie linéaire semble bien adaptée au mouvement car elle modélise correctement sa nature séquentielle, surtout lorsqu'aucun mouvement cyclique n'apparait dans le geste. De plus, le choix d'une telle topologie permet de réduire de façon conséquente la charge calculatoire qu'amènerait une topologie ergodique (entièrement connectée), tout en conservant des performances qui lui sont comparables en reconnaissance de mouvements [Romaszewski2011].

Décider automatiquement du nombre d'états et du nombre de composantes du mélange est difficile dans la pratique. Comme il est suggéré dans la littérature [Lv2006], ces paramètres de modélisation du HMM ont été empiriquement déterminés par un compromis entre la complexité de calcul et la performance de reconnaissance du système.

2.3.2.3 Paramétrisation des HMMs

Pour établir ce compromis, nous avons d'abord déterminé le nombre optimal d'états. A cette fin, nous avons utilisé le descripteur amorphologique et fixé arbitrairement le nombre de Gaussiennes à 5. Puis nous avons fait varier le nombre d'état de 1 à 30, et entrainer parallèlement 15 HMM (un par classe de mouvements), sur la moitié de la base de données. Nous avons ensuite effectué une validation en aveugle sur l'autre moitié de la base de données. Trois critères guident le choix du meilleur compromis : le taux de reconnaissance moyen obtenu, la log-vraisemblance moyenne et le temps de calcul pour l'entrainement. La figure 2.4 présente ces critères. Il ressort que le temps d'entrainement se comporte comme une fonction quasi linéaire du nombre d'états. Ce critère pousse à choisir un nombre d'états le plus faible possible. La log-vraisemblance et le taux de reconnaissance présentent des coudes entre 5 et 9 états, avant de se stabiliser sur un plateau. Le choix du nombre d'états s'est donc porté empiriquement sur 7, offrant un compromis entre les différentes contraintes à respecter.

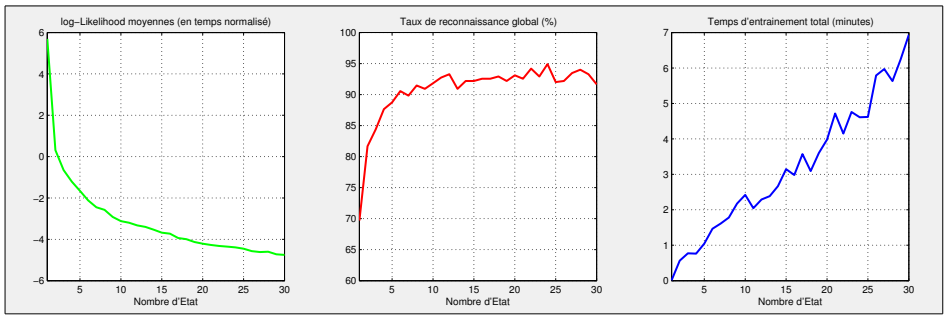


Figure 2.4 - Évolution des critères de décision en fonction du nombre d'états. A gauche, la log-vraisemblance moyenne sur l'ensemble des classes. Au centre, le taux de reconnaissance moyen. A droite, la durée totale de la phase d'entrainement.

À partir de ces 7 états, nous avons réalisé la même expérimentation, en faisant varier, cette fois-ci, le nombre de composantes Gaussiennes entre 1 et 20. La figure 2.5 présente les critères de décision. Encore une fois, le temps d'entrainement se comporte comme une fonction quasi-linéaire du nombre de Gaussiennes. Les coudes sur les deux autres critères sont moins flagrants que sur la figure 2.4, mais se situent entre 5 et 8 composantes. En observant le taux de

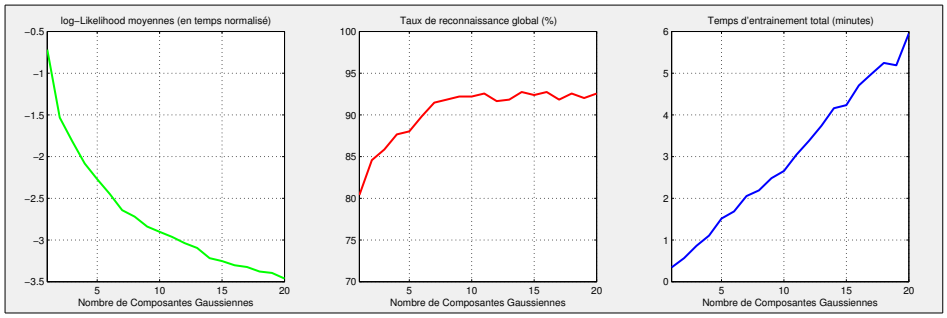


Figure 2.5 - Évolution des critères de décision en fonction du nombre de composantes Gaussiennes.

reconnaissance par mouvement, présenté sur la figure 2.6, on distingue cependant un coude sur certaines classes, tels que *claque de la paume*, *uppercut*, ou *lance*, qui sont mieux reconnus à

partir de 6 Gaussiennes. C'est pourquoi le nombre de 6 composantes Gaussiennes a été retenu.

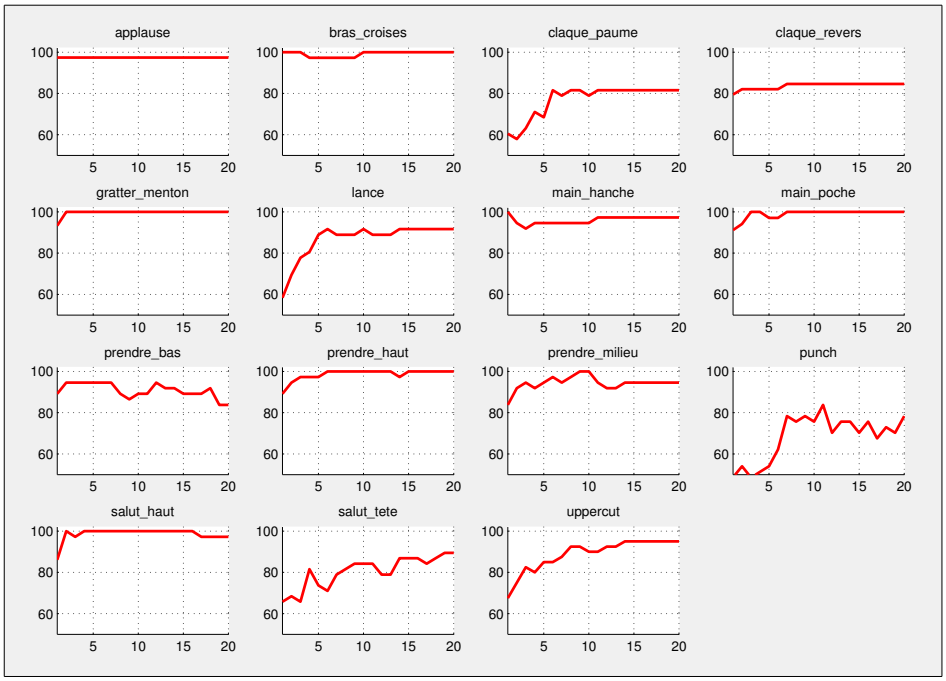


Figure 2.6 - Évolution du taux de reconnaissance de chaque classe de geste en fonction du nombre de composantes Gaussiennes.

2.4 Résultats

2.4.1 Répartition aléatoire

Classiquement, les méthodes de reconnaissance de mouvements attribuent 50% des échantillons composants chaque classe de la base de données pour entrainer leur système, et les 50% restants pour évaluer le taux de reconnaissance en aveugle. De la sorte, aucun des gestes utilisés pour l'évaluation n'est connu du système. Nous adoptons cette méthodologie d'évaluation, à ceci près que nous faisons également varier la proportion de gestes attribués à l'entrainement et à la reconnaissance. En effet, en situation réelle, le nombre d'échantillons disponibles pour entrainer le système est nécessairement limité, ou, en tout cas, largement inférieur au panel de gestes qui seront amenés à être reconnus dans le cadre de l'application finale. Ainsi, les expérimentations successives attribuent 10% puis 30%, 50%, 70% et enfin 90% des échantillons (fichiers BVH) de chaque classe de la base de données à l'entrainement, et les 90% puis, 70%, 50%, 30% et enfin 10% restants sont consacrés à l'évaluation du taux de reconnaissance.

Pour chaque partitionnement de la base de données, nous avons répété dix fois l'expérimentation avec, à chaque fois, une répartition aléatoire différente des échantillons entre l'entrainement et l'évaluation, afin de s'assurer qu'il n'y ait pas d'influence des gestes sélectionnés sur les résultats.

2.4.2 Répartition 50/50

Nous nous intéressons, tout d'abord, à l'approche classique, où 50% des échantillons de mouvements de chaque classe sont utilisés pour l'entrainement du système. Les résultats par mouvement et par descripteur sont présentés dans le tableau 2.2. Il en ressort que le descripteur

Geste	**Amorph. (%)**	**Cart. (%)**	**Euler (%)**
1. Applaudir	98.3	96.7	92.5
2. Bras croisés	98.8	98.8	99.9
3. Claque paume	85.8	85.9	83.0
4. Claque revers	90.4	92.2	83.7
5. Gratter menton	97.7	96.7	94.2
6. Lancer	89.0	85.5	84.4
7. Mains hanches	97.0	98.5	96.5
8. Mains poches	96.8	98.2	94.1
9. Prendre bas	88.4	78.4	66.8
10. Prendre haut	91.1	91.5	78.2
11. Prendre milieu	89.3	87.7	70.9
12. Punch	74.5	81.9	76.6
13. Salut haut	97.6	95.8	90.7
14. Salut tête	84.0	79.2	75.7
15. Uppercut	83.8	87.2	80.6
Moyenne $\pm$ Écart-type	90.8 ± 0.88	90.2 ± 1.0	84.5 ± 1.7

Table 2.2 - Taux de reconnaissance pour chaque mouvement et chaque descripteur avec partitionnement aléatoire (50% des échantillons d'une classe pour l'entrainement, les 50% restants pour l'évaluation). Les moyennes et écart-types sont calculés sur les différents tirages aléatoires (et non sur les taux de reconnaissance par mouvement).

amorphologique ne se comporte globalement ni mieux, ni moins bien que les deux autres pour reconnaitre les 15 classes de mouvement de notre étude. Les taux de reconnaissance moyens des systèmes utilisant ces descripteurs oscillent entre 90% et 91%. Avec 84.5% de classifications correctes, la descripteur angulaire se trouve légèrement, mais significativement, plus en retrait, comme le confirme un test des rangs signés de Wilcoxon ($T = 0.890, p < 0.05$ par rapport aux descripteurs amorphologiques, et $T = 0.780, p < 0.05$ par rapport aux Cartésiens).

Construites à partir des résultats du tableau 2.2, les figures 2.7 permettent de se rendre compte graphiquement des classes de mouvements les mieux et les moins bien reconnues. L'axe vertical représente le taux de reconnaissance entre 0 et 1, et l'axe horizontal représente le taux de "faux

positifs" (mouvements reconnus avec ce modèle HMM de manière erronée) reconnus par le modèle HMM étudié. Chaque point représente une classe de mouvements. L'idéal serait d'avoir tous les points regroupés en haut à gauche de la figure, indiquant un fort taux de reconnaissance et un faible taux de "faux positifs". Les figures permettent de voir à quel point un modèle est spécifique de la classe de mouvement qu'il modélise. Plus un modèle se voit attribuer de mouvements appartenant à d'autres classes (faux positifs), moins il est spécifique.

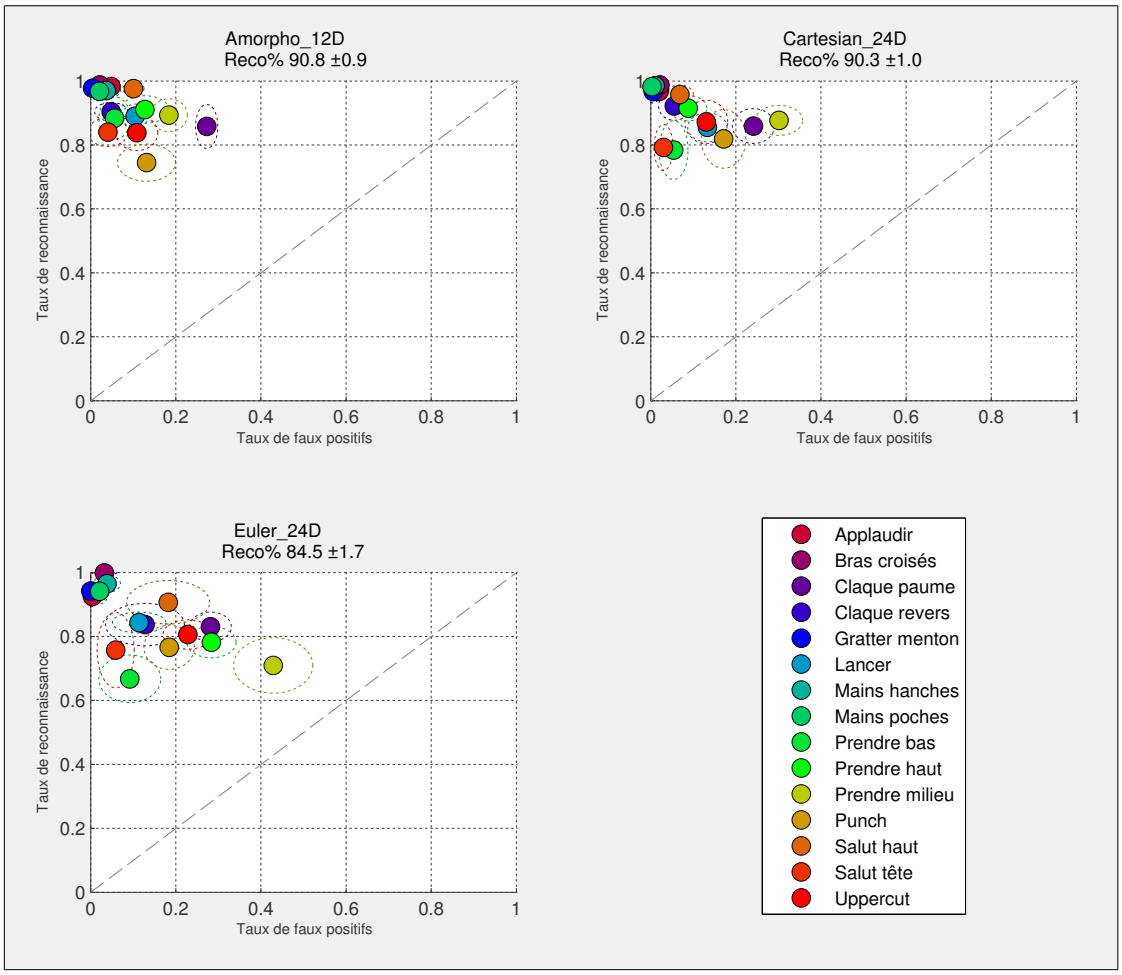


Figure 2.7 - Diagramme *sensibilité / anti-spécificité* pour la répartition aléatoire 50/50. En haut à gauche, description amorphologique ; à droite, cartésienne ; en bas, eulérienne. La légende, en bas à droite, fournit la correspondance entre un modèle et la couleur qui le représente. Le taux de reconnaissance de chaque modèle HMM de mouvement apparait en ordonnées, le taux de faux positif en abscisses. Plus un mouvement se trouve en haut du graphe, mieux il est reconnu, plus il est sur la gauche, plus le modèle HMM est spécifique et rejette les autres mouvements. Les pointillés autour de chaque point représentent les écart-type verticaux et horizontaux.

2.4.3 Autres répartitions

Ces précédents résultats tendent à montrer qu'il n'existe pas de différence entre les descripteurs Cartésiens et amorphologiques. Cependant, l'échantillonnage aléatoire d'une moitié des mouvements de chaque classe pour l'entrainement du système implique deux choses. D'abord, il existe une probabilité non négligeable pour que tous les sujets soient représentés dans chaque classe de la base d'entrainement. Leurs morphologies peuvent alors être prises en compte dans le modèle HMM. Ensuite, les styles de mouvements, qui peuvent être partagés par plusieurs sujets, ont également une probabilité importante d'être tous représentés dans la base d'entrainement, et donc modélisés par les HMM. Or, dans le cadre d'une application interactive, il n'est pas envisageable de recueillir une base de données exhaustive qui engloberait toutes ces variabilités. Pour observer l'influence de la taille de la base d'apprentissage, nous avons donc fait varier sa part de 10% à 90%, par pallier de 20%. Les résultats sont présentés dans le tableau 2.3 et représentés graphiquement sur la figure 2.8.

Ratio (%) entr./valid.	**Amorpho. (%)**	**Cart. (%)**	**Euler (%)**
10 / 90	79.0 ± 1.2	65.7 ± 2.6	57.5 ± 3.0
30 / 70	88.3 ± 1.3	86.8 ± 0.7	80.2 ± 1.6
50 / 50	90.8 ± 0.9	90.3 ± 1.0	84.5 ± 1.7
70 / 30	91.8 ± 0.9	92.4 ± 1.3	86.4 ± 0.5
90 / 10	93.6 ± 2.1	93.5 ± 2.0	89.5 ± 1.4

Table 2.3 - Taux de reconnaissance en fonction du pourcentage d'échantillon de chaque classe de mouvement alloué à l'entrainement et à la validation. Ces taux sont moyennés sur 10 tirages aléatoires.

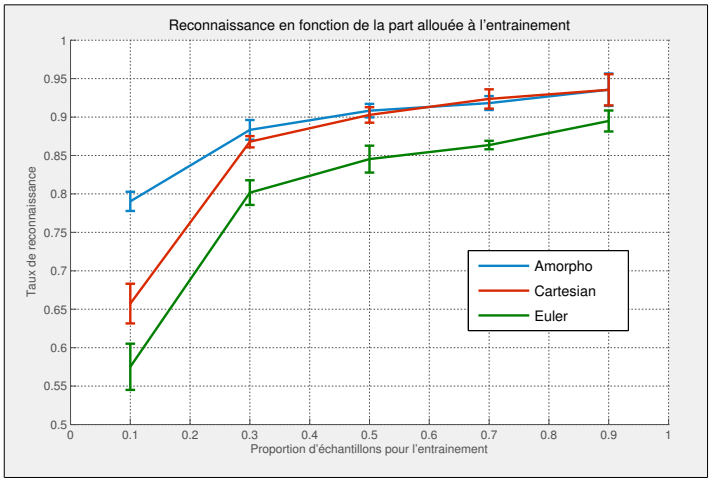


Figure 2.8 - Taux de reconnaissance des systèmes utilisant chaque descripteur, en fonction de la proportion de la base de données attribuée à l'entrainement.

La figure 2.8 montre l'évolution du taux de reconnaissance en fonction de la répartition de la

base de données utilisées pour l'apprentissage. Elle met en évidence que lorsque 10% seulement des données sont utilisées pour l'apprentissage, le taux de reconnaissance du modèle utilisant le descripteur amorphologique chute beaucoup moins que pour les deux autres descripteurs. Le taux de reconnaissance du descripteur amorphologique atteint 79.0% alors que ceux des autres descripteurs atteignent 65,7% et 57,5% pour respectivement le Cartésien et l'angulaire. Une ANOVA par rang d'ordre 2 de Friedman démontre que cette différence est signifcative, $\chi^2(2, N = 150) = 139.31$, $p < 0.0001$. Un test post-hoc des rangs signés de Wilcoxon confirme la significativité de la différence par rapport aux descripteurs Cartésien ($T = 0.793, p < 0.05$) et angulaire ($T = 1.347, p < 0.05$). La figure 2.9 présente les taux de reconnaissance et de faux positifs pour chaque descripteur dans ce cas de figure.

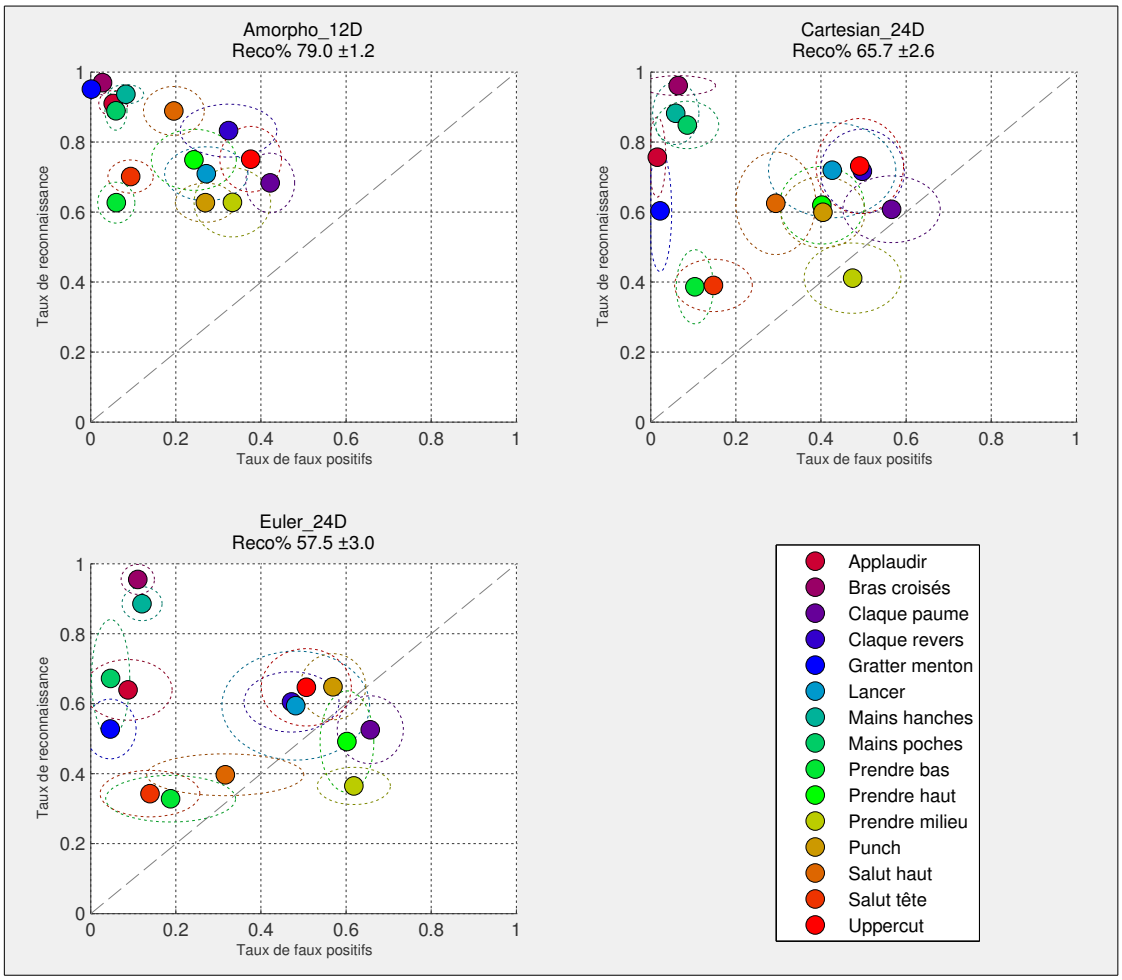


Figure 2.9 - Diagramme *sensibilité / anti-spécificité* pour la répartition aléatoire 10% entrainement / 90% validation. Voir la légende de la figure 2.7 pour plus d'explication.

On notera, en particulier, que les mouvements symétriques sont les mieux reconnus quelle que soit le descripteur. Cela tient pour partie au fait que les mouvements de la base de données sont réalisés à droite comme à gauche, ce qui complique la tâche de modélisation des HMM pour les mouvements latéralisés (disposant ainsi de deux fois moins d'information pour l'entrainement que les mouvements symétriques entrainant les deux bars en même temps).

En plus de la symétrie, les propriétés dynamiques du mouvement semblent avoir une influence sur le taux de reconnaissance. Mise à part la classe de mouvements *applaudir*, les mouvements

symétriques sont assez peu dynamiques comparés aux gifles, punchs, uppercuts, saluts et lancers, qui sont d'ailleurs moins bien reconnus. Or, une classe de mouvements dynamiques implique une variabilité spatiotemporelle importante liée au style (trajectoires des articulations, vitesse d'exécution), qui complique la tâche de modélisation. Il est donc logique de trouver un taux de reconnaissance moins bon pour ces mouvements dynamiques.

En dehors de ces constats globaux, on peut remarquer que le taux de reconnaissance du descripteur amorphologique est significativement meilleur pour cette répartition 10% / 90%. La dispersion est importante sur les graphiques représentant les résultats des descripteurs Cartésien et angulaire. Cette dispersion est moindre pour le descripteur amorphologique : au dessus des 60% de reconnaissances correctes et en dessous des 40% de faux positifs (fig.2.9). En outre, chaque classe de mouvements est systématiquement mieux reconnue avec le descripteur amorphologique. Enfin, l'écart-type autour du taux de reconnaissance moyen est au moins deux fois plus important sur les descripteurs Cartésien ($\pm$2.6%) et angulaire ($\pm$3.0%) que sur le descripteur amorphologique ($\pm$1.2%). Ce constat tend à démontrer qu'il existe des tirages aléatoires plus favorables que d'autres et que cela impacterait plus les descripteurs Cartésien et angulaire. En d'autres termes, le descripteur amorphologique serait plus stable vis-à-vis des différences liées à l'échantillonnage des mouvements servant à l'apprentissage ; différences que l'on peut principalement imputer à la variabilité inter-individuelle.

A partir de la répartition 30% entrainement / 70% évaluation, la différence des taux de reconnaissance n'est plus significative entre les descripteurs Cartésiens et amorphologiques. Le descripteur angulaire obtient des taux de reconnaissance de 4 à 6% inférieurs aux autres descripteurs. On peut toutefois noter que la classe de mouvements *punch* est la seule qui soit systématiquement mieux reconnue en utilisant le descripteur Cartésien à partir de la répartition 30% / 70%.

2.4.4 Discussion

Si l'on s'en tient uniquement à des répartitions où au moins 30% de la base de données est utilisée pour l'apprentissage, le descripteur amorphologique n'apporte pas d'amélioration significative sur le taux de reconnaissance global par rapport à un descripteur Cartésien. Dans le cas inverse, lorsque cette répartition dédiée à l'apprentissage descend au-dessous de 30%, la probabilité qu'un sujet ne soit absolument pas représenté pour l'apprentissage devient importante. La morphologie de ce sujet ne peut donc pas être apprise par le système, qui a donc plus de chance d'échouer à le reconnaitre. Le descripteur amorphologique est justement conçu pour résoudre ce type de problème, ce qui peut expliquer sa meilleure performance dans ce cas de figure. Pour s'en assurer, nous proposons dans la section suivante d'évaluer directement l'influence de la morphologie par l'intermédiaire d'un échantillonnage par sujet, dit *leave-one-out*, dans lequel on exclut totalement le sujet à reconnaitre de la base d'apprentissage.

2.5 Répartition par sujet

2.5.1 Répartition Leave-One-Out

Dans la méthode précédente d'échantillonnage, bien que les mouvements à reconnaitre ne fassent pas partie de la base de données d'entrainement, le système peut avoir déjà vu des mouvements effectués par les mêmes sujets. Toutes les caractéristiques morphologiques et les styles peuvent donc être modélisés par le HMM. Notons, de plus, que les conditions expérimentales peuvent varier légèrement d'un sujet à l'autre, entrainant des biais systématiques lors de le capture et de la reconstruction du mouvement (placement des marqueurs non rigoureusement identique, initialisation du squelette en pose T, paramètres de reconstruction...).

Pour limiter ces possibles biais, l'échantillonnage *Leave-One-Out* consiste à évaluer le taux de reconnaissance du système à partir des mouvements d'un sujet dont aucun fichier n'appartient à la base de données d'entrainement. Cette situation correspond bien aux conditions réelles d'utilisation d'un système de reconnaissance en environnement virtuel : l'utilisateur final ne participe pas à la phase d'élaboration du système, en revanche ce sont bien ses mouvements qui doivent être reconnus.

Concrètement, pour chaque sujet Suj_i, l'entrainement du HMM est réalisé à partir des mouvements de l'ensemble de tous les autres sujets $\{Suj_{j\neq i}\}$. Puis, le taux de reconnaissance du HMM est évalué à partir des mouvements réalisés par Suj_i. Les résultats de cette validation *leave-one-out* montre que le descripteur amorphologique fournit des performances significativement meilleures que les deux autres : 85% de bonnes classifications contre 71% pour le Cartésien et 55% pour l'angulaire. Une ANOVA de Friedman démontre l'influence des descripteurs sur le taux de reconnaissance, $\chi^2(2, N = 150) = 52.91,\ p < 0.0001$. Un test post-hoc des rangs signé de Wilcoxon montre que le taux de reconnaissance est significativement plus élevé pour le descripteur amorphologique par rapport aux descriptions utilisant les positions Cartésiennes ($T = 0.343, p < 0.05$) ou les angles d'Euler ($T = 0.657, p < 0.05$).

Le tableau 2.4 fournit le taux de reconnaissance moyenné entre les sujets pour chaque mouvement et pour les trois descripteurs testées. On peut voir que l'utilisation du descripteur amorphologique conduit à de meilleures performances (un minimum de 63% pour la classe *punch*) par rapport au descripteur Cartésien (minimum de 47% pour *saisir un objet en hauteur*) ou angulaire (minimum de 17% pour *saisir un objet en hauteur*). De manière générale, le descripteur amorphologique obtient systématiquement un meilleur score sur tous les types de mouvements sauf l'*uppercut* et dans une moindre mesure le *punch* qui sont mieux reconnus par le descripteur Cartésien.

Les matrices de confusion 2.10, 2.11 et 2.12,fournissent les classes affectées à l'intégralité des échantillons de mouvements par les classifieurs utilisant chaque type de descripteur. Les 10 expérimentations L1O y sont additionnées. L'effectif de chaque classe est rapporté à 100 échantillons de mouvements par classe afin de clarifier la lecture. Notons que tous les sujets n'ont pas effectué le même nombre de mouvements, ce qui entraine une légère différence avec les données du tableau 2.4 (qui sont moyennées sur les sujets). Ces matrices de confusion permettent de se rendre compte des erreurs de reconnaissance pour chaque classe de mouvements. Ainsi, le descripteur amorphologique confond fréquemment la classe n° 15 (*uppercut*) avec la classe n° 3 (*claque de la paume*) ou avec la classe n° 6 (*lancer*). En outre, les mouvements symétriques et

Geste entr./valid.	**Amorph. (%)**	**Cart. (%)**	**Euler (%)**
1. Applaudir	90.9	88.9	69.9
2. Bras croisés	100	93.5	80.0
3. Claque paume	77.4	61.0	52.3
4. Claque revers	80.2	75.4	45.1
5. Gratter menton	96.4	68.0	53.0
6. Lancer	77.5	70.8	69.2
7. Mains hanches	96.7	87.3	88.3
8. Mains poches	97.2	90.0	68.0
9. Prendre bas	84.9	46.7	17.1
10. Prendre haut	90.6	70.9	89.3
11. Prendre milieu	84.2	57.3	35.5
12. Punch	68.6	70.7	51.0
13. Salut haut	94.5	80.7	26.5
14. Salut tête	77.7	33.4	24.2
15. Uppercut	67.2	76.7	55.7
Moyenne $\pm$ Écart-type	85.5 $\pm$ 11.8	71.4 $\pm$ 16.2	55.0 $\pm$ 22.7

Table 2.4 - Taux de reconnaissance pour chaque mouvement et chaque descripteur avec l'approche L1O. Il s'agit de moyennes sur tous les sujets, elles diffèrent donc des sommes sur tous les mouvements que l'on retrouve dans les matrices de confusion. L'analyse statistique confirme un taux de reconnaissance significativement plus élevé du descripteur amorphologique par rapport aux autres. Les moyennes et écart-types sont calculés en considérant les différents sujets (et non les différents mouvements).

peu dynamiques sont, encore une fois, les mieux reconnus quel que soit le descripteur.

Par ailleurs, l'écart-type plus faible entre les taux de reconnaissance obtenus en considérant les différents sujets avec le descripteur amorphologique tend à montrer qu'il serait moins sensible à de nouveaux utilisateurs contrairement aux autres descripteurs dont l'écart-type est plus important. Par exemple, le sujet ayant le score le plus faible obtient un taux de reconnaissance moyen de 35.3% avec le descripteur angulaire. Le sujet ayant le plus haut score obtient un taux de reconnaissance moyen de 76.2% avec le même descripteur. Dans le même temps, le taux de reconnaissance est de 66.0% et 99.3% respectivement pour les sujets ayant les scores les plus bas et le plus élevé lorsque le modèle utilise des données amorphologiques.

Enfin, lorsqu'on regarde individuellement les évolutions des taux de reconnaissance en fonction de la description utilisée (2.13), il apparait clairement que le descripteur amorphologique est systématiquement meilleur pour chaque sujet.

2.5.2 Répartition Leave-k-Out

En conditions réelles, le système de reconnaissance ne peut pas être entrainé avec presque tous les utilisateurs potentiels. L'entrainement est, au contraire, réalisé à partir d'une petite base de données par rapport à l'ensemble de toutes les personnes (et de leurs morphologies propres) susceptibles d'utiliser le système final. Il est donc important d'aller plus loin dans l'approche Leave-One-Out pour observer le comportement du système de reconnaissance lorsque deux sujets sont extraits de la base d'apprentissage (L2O), puis trois (L3O)... Nous avons donc prolongé l'expérimentation L1O jusqu'à extraire neuf sujets (L9O), ce qui signifie que seuls les mouvements d'un unique sujet sont utilisés pour la phase d'entrainement, pendant que ceux des neuf autres servent à tester les performances de reconnaissance.

Ce type d'échantillonnage pose cependant un problème combinatoire si l'on veut réaliser un test exhaustif. Pour L2O, il existe $C_{10}^{2} = 45$ combinaisons possibles pour choisir 2 sujets parmi les 10. Il en existe $C_{10}^{3} = 120$ pour L3O, $C_{10}^{4} = 210$ pour L4O... Traiter tous les cas n'est pas envisageable en pratique. Nous avons donc réalisé un échantillonnage de ces combinaisons $C_{n=10}^{k}$ en n'en conservant que 10 pour chaque LkO. Une première idée serait d'effectuer une série de 10 tirages aléatoires purs pour chaque LkO. Le problème de cette méthode est qu'elle ne peut pas garantir que chaque sujet apparaisse un même nombre de fois dans chaque LkO successif. Cela pourrait biaiser la comparaison des différents LkO. Plutot que de tirer aléatoirement 10 combinaisons pour chaque LkO, nous avons choisi d'effectuer un tirage pseudo-aléatoire par permutation circulaire. Le but est d'assurer une répartition homogène des sujets dans chaque LkO afin d'obtenir des taux de reconnaissance comparables à chaque incrémentation de LkO. La méthode est la suivante. On tire au hasard un n° d'ordre pour chaque sujet. Puis on tire au hasard un chiffre $a \in [1, 9]$ et on associe par permutation circulaire le sujet n° 1 avec le sujet n° $(1 + a)$, le sujet n° 2 avec le sujet n° $(2 + a)$... A la fin de ce tour, tous les sujets apparaissent bien 2 fois parmi les 10 tirages qui forment notre échantillon de combinaison pour la validation L2O. La méthode est la même pour les LkO suivants.

La figure 2.14 montre le taux de reconnaissance global pour toutes les classes de mouvements

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	93	6	0	0	0	0	0	0	0	0	0	1	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	79	0	0	13	0	0	0	0	3	1	0	1	3
4	0	0	1	84	0	3	0	0	0	3	0	0	0	5	4
5	0	0	4	0	96	0	0	0	0	0	0	0	0	0	0
6	0	0	7	1	0	81	0	0	0	0	0	4	0	0	7
7	0	1	0	0	0	0	97	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	3	97	0	0	0	0	0	0	0
9	0	0	7	1	0	0	0	0	83	0	7	1	0	0	0
10	0	0	0	0	0	0	0	0	0	88	0	0	12	0	0
11	0	0	1	10	0	0	0	0	1	3	81	3	0	0	0
12	7	0	7	0	1	3	0	0	0	0	4	74	0	0	3
13	0	0	2	0	0	2	0	0	0	0	0	0	95	0	2
14	0	0	7	0	0	0	0	0	0	4	1	0	10	77	0
15	4	0	16	0	0	8	0	0	0	0	0	1	0	0	71

Figure 2.10 - Matrice de confusion pour le descripteur amorphologique en L1O. Les véritables classes apparaissent en ligne, les classes attribuées en colonne. Toutes les classes sont ramenées à 100 gestes pour faciliter la lecture. Sur la première ligne, par exemple, la classe n° 1 est reconnue correctement 93 fois (sur 100), attribuée 6 fois à la classe n° 2 et 1 fois à la classe n° 12. La coloration de certaines cases permet de mettre en relief les erreurs les plus fréquentes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	89	3	0	0	0	0	0	0	0	0	0	3	0	0	5
2	3	97	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	64	0	0	18	0	0	0	0	11	5	0	0	3
4	0	0	6	75	0	7	0	0	0	0	6	1	0	0	4
5	0	0	2	10	66	0	0	0	12	0	6	0	0	0	4
6	0	0	15	0	0	82	0	0	0	0	0	2	0	0	2
7	0	2	0	0	0	0	89	9	0	0	0	0	0	0	0
8	0	0	0	0	0	0	9	91	0	0	0	0	0	0	0
9	0	0	19	3	0	3	0	0	48	0	17	6	0	0	3
10	0	0	0	0	0	3	0	0	0	71	16	0	8	0	2
11	0	0	16	5	0	6	0	0	6	3	55	5	3	0	2
12	6	0	5	0	0	9	0	0	0	0	2	77	0	0	2
13	0	0	0	10	0	0	0	0	0	7	2	0	82	0	0
14	0	0	8	22	0	2	0	0	2	2	11	0	18	35	2
15	7	0	9	0	0	3	0	0	0	0	0	1	0	0	80

Figure 2.11 - Matrice de confusion pour le descripteur Cartésien en L1O.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	76	3	0	0	0	0	0	0	0	0	0	18	0	0	3
2	13	87	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	53	0	0	9	0	0	0	15	3	9	0	0	11
4	0	0	9	41	0	6	0	0	0	12	4	4	3	3	18
5	0	0	4	0	58	2	0	0	0	6	24	6	0	0	0
6	0	0	5	0	0	73	0	0	0	0	0	13	0	0	8
7	16	2	0	0	0	0	83	0	0	0	0	0	0	0	0
8	10	0	0	0	0	0	10	71	0	0	2	3	0	0	3
9	0	0	5	0	0	9	0	0	16	20	30	13	0	0	8
10	0	0	0	0	0	10	0	0	0	90	0	0	0	0	0
11	0	0	3	0	0	6	0	0	0	52	33	6	0	0	0
12	8	0	2	0	0	11	0	0	0	0	3	61	0	0	16
13	0	0	8	13	0	8	0	0	0	48	0	0	23	0	0
14	0	0	26	9	0	9	0	0	0	9	8	0	12	26	0
15	9	0	6	0	0	6	0	0	0	1	0	23	0	0	55

Figure 2.12 - Matrice de confusion pour le descripteur angulaire en L1O.

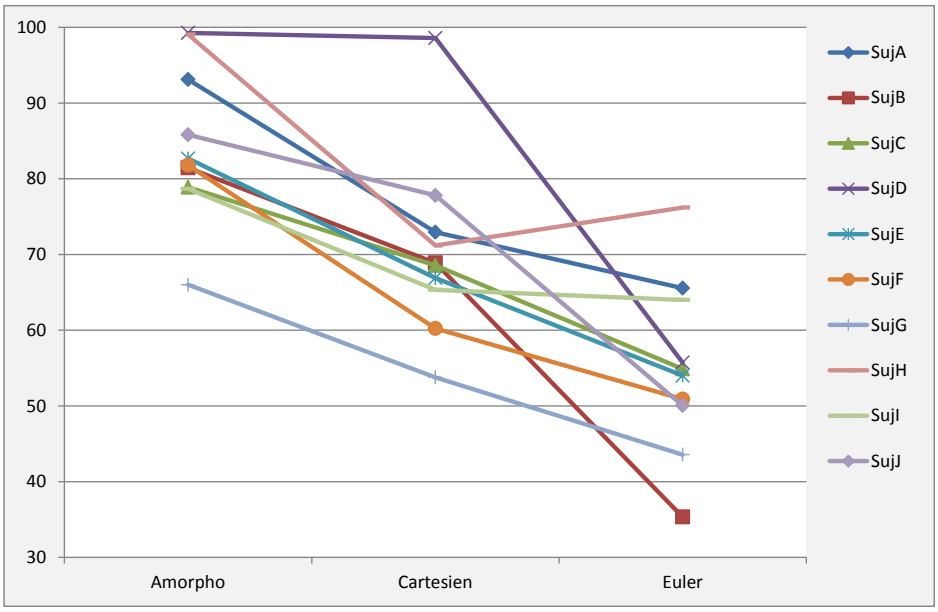


Figure 2.13 - Taux de reconnaissance par sujet pour l'échantillonnage L1O. Le descripteur amorphologique est systématiquement meilleur.

en fonction de la méthode d'évaluation (à partir de L1O jusqu'à L9O). Lorsque le nombre de sujets utilisés pour l'apprentissage diminue (ce qui revient à augmenter le nombre d'utilisateurs nouveaux), les taux de reconnaissance des descripteurs Cartésien et angulaire décroissent plus rapidement que celui du descripteur amorphologique. Le score de ce dernier chute en effet de 85% à 68% dans le pire des cas, où les mouvements d'un seul sujet sont utilisés pour l'entrainement. Le taux de reconnaissance des autres descripteurs chute à moins de 25% pour cette même situation extrême.

La meilleure performance du descripteur amorphologique, qui était déjà significativement importante pour l'approche L1O est encore une fois significative pour tous les autres échantillonnages (de L2O à L9O). Des ANOVA de Friedman ont, de nouveau, démontré l'influence des descripteurs sur le taux de reconnaissance et des test post-hoc des rangs signés de Wilcoxon ont confirmé, à chaque fois, que les taux de reconnaissance étaient significativement plus élevés pour le descripteur amorphologique par rapport aux autres.

Les mauvais scores obtenus avec les autres descripteurs ne peuvent pas être attribués uniquement aux changements de morphologie. En effet, un nouvel utilisateur n'est pas seulement une nouvelle morphologie à prendre en compte, mais c'est aussi un nouveau style, une nouvelle manière d'effectuer les mouvements. Prenons deux sujets Suj_F et Suj_G, de même sexe, de tailles et de poids similaires (cf. tableau 2.3.2.1) et intéressons nous aux résultats en L1O. Le taux de reconnaissance à partir du descripteur Cartésien sur des *uppercut* est de 20% pour Suj_F, alors qu'il est de 100% pour Suj_G. Cela semble démontrer que le style de Suj_F est très différent de celui des autres sujets qui ont été utilisés pour entrainer le système. Ce n'est pas le cas pour Suj_G. Ce genre de constat évolue aussi en fonction du mouvement. Par exemple, les *gifler avec la paume de la main* sont reconnus à 100% pour Suj_F et seulement 29% des fois pour Suj_G, en utilisant ces mêmes descripteurs Cartésien. En conséquence, l'utilisation du descripteur Cartésien conduit à des taux de reconnaissance très différents, qui dépendent à la fois du sujet et du mouvement.

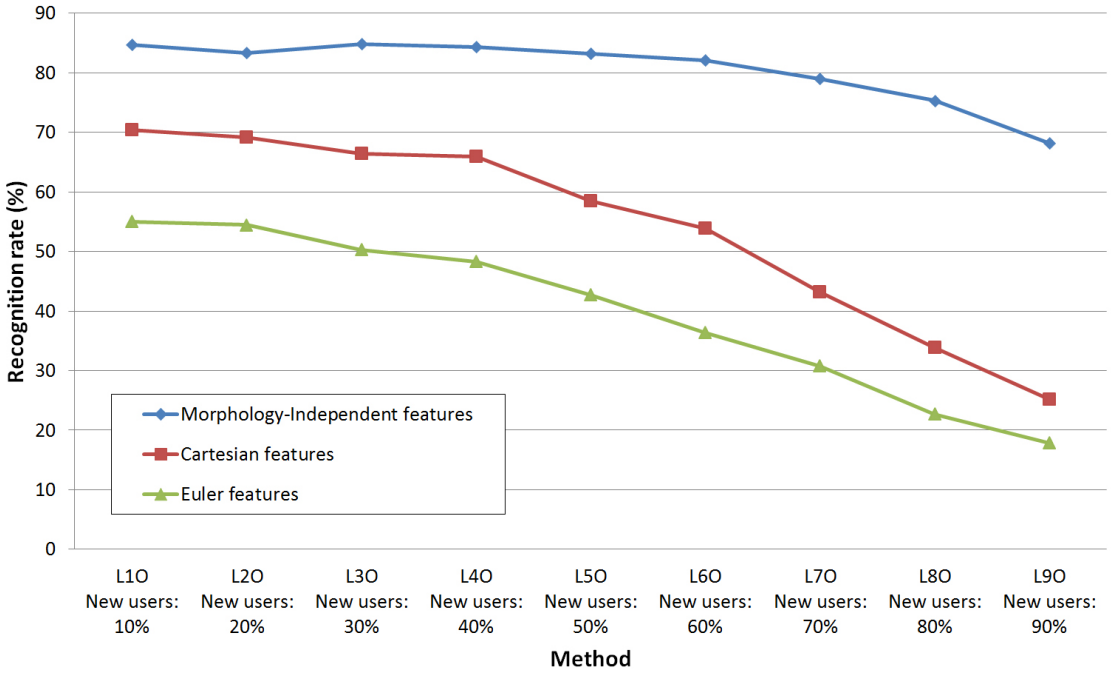


Figure 2.14 - Évolution du taux de reconnaissance en fonction de la méthode d'évaluation. Plus on avance vers la droite du graphique, plus grande est la proportion de sujets n'appartenant qu'à la base de validation, c.-à-d. de nouveaux utilisateurs dont les mouvements doivent être reconnus. Le taux de reconnaissance du descripteur amorphologique commence à infléchir légèrement lorsqu'au moins 60% des utilisateurs ne sont pas connus du système. La rupture de pente est plus brutale et plus précoce (plutot autour de 40%) pour les autres descripteurs.

2.6 Application dans un démonstrateur interactif

Afin d'évaluer les trois descripteurs en conditions réelles, nous avons conçu un jeu interactif : un utilisateur est équipé de 34 marqueurs réfléchissants disposés aux mêmes endroits que lors de l'acquisition de la base de données. Un optitrack permet de capturer et de reconstruire ses mouvements en temps réel. L'utilisateur est placé devant un grand écran où trois humains virtuels sont affichés, comme l'illustre la figure 2.15). Chaque humain virtuel est associé à un système de reconnaissance utilisant un descripteur différent (amorphologique, Cartésien ou angulaire).

L'utilisateur doit effectuer l'un des 15 mouvements de notre étude et revenir à une posture de repos, sans autre forme de contrainte.

Les trois systèmes de reconnaissance sont alors exécutés en même temps sur le mouvement exécuté par l'utilisateur. Une fois que chaque système a déterminé la classe correspondante, un mouvement type de la base de données associé à cette classe est joué par l'humain virtuel correspondant. Cette partie du jeu est répétée 10 fois (l'utilisateur choisit 10 mouvements au hasard parmi les 15) et les points sont additionnés pour chaque humain virtuel qui reconnait correctement la classe choisie par l'utilisateur, aboutissant à un score global compris entre 0 et 10 à la fin du jeu.

Toute l'expérimentation a été répétée 10 fois, en condition L1O (utilisateur nouveau). Les résultats indiquent un score moyen de 8.8 ± 0.6 pour le descripteur amorphologique, 7.2 ± 0.8

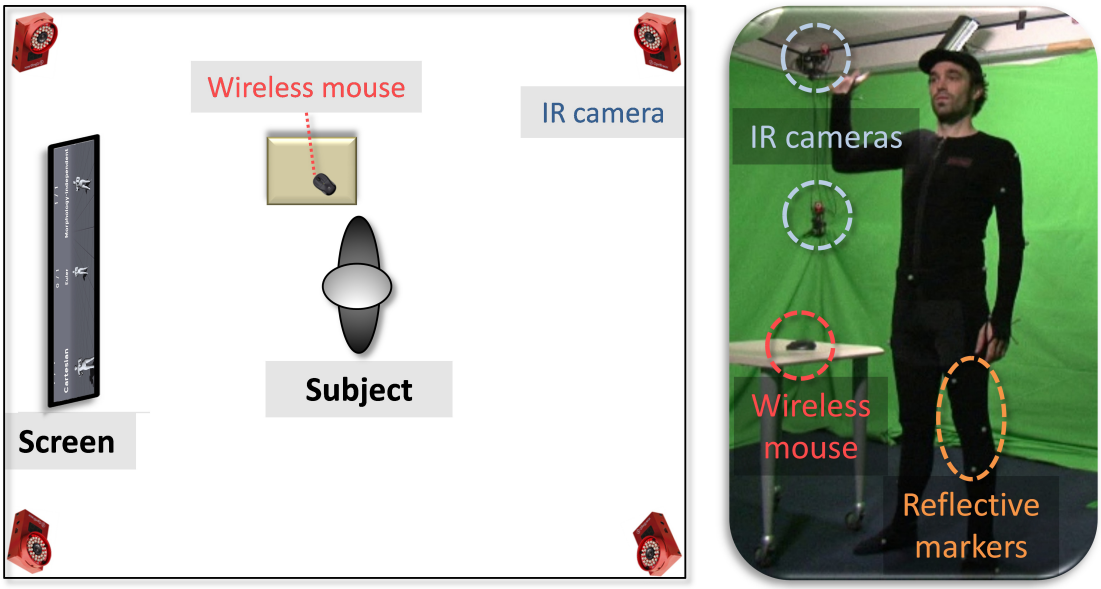


Figure 2.15 - Mise en œuvre expérimentale de l'application démonstrateur interactive. Le sujet est placé face à un écran qui affiche 3 avatars, un pour chaque descripteur. Ses gestes sont capturés et reconstruits en temps réel grâce au système Optitrack. Une souris est disponible à portée de sa main pour attribuer des scores aux avatars, après qu'ils aient déterminé la classe des mouvements effectués.

pour le Cartésien et 5.0 ± 1.1 pour Euler. Ces résultats montrent en condition réelle le bon comportement du descripteur amorphologique (test Q de Cochran $\chi^2(2, N = 102) = 44.98$, $p < 0.001$).

Le processus de reconnaissance a été exécuté sur un PC standard disposant d'un processeur Intel Core 2 Quad Q8400 cadencé à 2,66 GHz et de 4Go de mémoire. L'application est entièrement codée en langage C++. Les temps de calcul moyen par mouvement sont présentés dans la figure 2.16. Quel que soit le descripteur, ce temps de calcul est nettement inférieur à la durée du mouvement et donc tout à fait compatible avec des applications interactives. En outre, on peut voir que le temps d'exécution du processus de reconnaissance est nettement plus faible lorsqu'on utilise les descripteurs amorphologiques, plutôt que les autres. Ce constat est conforme au fait que le descripteur amorphologique est de dimension 12, tandis que les deux autres sont de dimension 24.

2.7 Conclusion

La principale contribution de cette étude est la définition et l'évaluation d'un nouveau type de descripteurs amorphologiques de mouvements humains. La représentation amorphologique permet de maintenir un taux élevé de reconnaissance, quand bien même le système n'a jamais observé de mouvements réalisés par l'utilisateur final. Nous avons clairement démontré que les descripteurs classiques fondés sur l'utilisation d'angles d'Euler ou de positions Cartésiennes

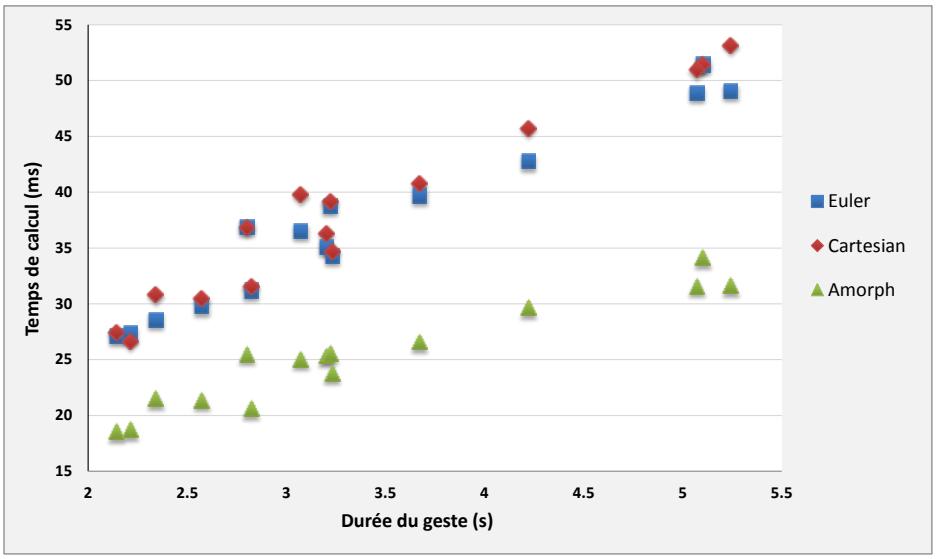


Figure 2.16 - Temps de calcul moyen des 3 classifieurs HMM utilisant chacun l'une des descriptions du mouvement (moyenné pour chacune des 15 classes de geste). L'axe des abscisses correspond à la durée moyenne des classes de gestes étudiées.

peinent à régler ce problème pour des mouvements possédant des caractéristiques spatiotemporelles très similaires. Dans le même temps, l'utilisation de descripteurs amorphologiques conduit à un taux de reconnaissance tout à fait correct.

Un autre apport de ce descripteur réside dans sa faible dimensionnalité, qui laisse à penser qu'une capture de mouvements sommaire pourrait suffire à reconnaitre des gestes naturels, même très similaires. Un outil de capture grand public, tel que Microsoft kinect ou SoftKinetic iisu, pourrait alors être envisagé. Cette hypothèse mériterait naturellement de nouvelles expérimentations.

Dans ce travail, nous avons volontairement limité notre descripteur aux mouvements des deux bras, qui sont les plus utilisés dans de nombreuses activités. Ce travail pourrait cependant être étendu à une représentation des mouvements du corps entier, comme le suggère [Kulpa2005b] dans le domaine de l'animation par ordinateur. D'autres expériences seraient nécessaires pour vérifier si cette représentation est appropriée pour la reconnaissance de mouvements du corps entier.

Nous avons sélectionné les HMM pour modéliser les mouvements, car ils constituent l'approche la plus populaire. Pour nos évaluations, nous avons fixé la topologie à partir de considérations empiriques, et déterminé le nombre d'états et de composantes Gaussiennes par état à partir d'un compromis entre le taux de reconnaissance et la complexité calculatoire. Pour guider le choix de la topologie en reconnaissance d'écriture, [Bhowmik2011] propose d'utiliser le critère d'information Bayésien. Pour décider du nombre optimal d'états en reconnaissance de mouvement, [Cholewa2011] propose de regarder le nombre de points critiques dans les données de capture. Il pourrait être intéressant d'évaluer les apports de ces méthodes de paramétrisation des HMM sur la performance globale de reconnaissance.

En outre, la description amorphologique pourrait également être appliquée à d'autres familles de modélisation. Les machines à vecteurs de support, par exemple [Laptev2007], sont très populaires en classification automatique. Récemment, les méthodes ensemblistes, et plus particulièrement les méthodes de dopage, comme AdaBoost [Lv2006], ont été très utilisées en reconnaissance de

mouvements. Elles sont fondées sur l'utilisation d'une énorme quantité de descripteurs du mouvement. Le processus d'apprentissage identifie alors les combinaisons les plus pertinentes, pour optimiser le taux de reconnaissance global. Ces méthodes pourraient tirer parti de descripteurs amorphologiques pour résoudre les problèmes dus à la variabilité anthropométrique.

Enfin, un défi majeur des systèmes de reconnaissance de mouvements reste la segmentation temporelle, en particulier dans un cadre temps-réel. En effet, dans les applications interactives, il est impossible d'attendre la fin du geste pour déterminer l'action effectuée par l'utilisateur, car l'environnement doit réagir de façon continue. Il est également impossible d'inviter l'utilisateur à se mouvoir uniquement pendant quelques fenêtres de temps imposées, en particulier dans des applications d'entrainement virtuel. Ces deux problèmes sont des points clés pour les développements futurs.

Chapitre 3

Vers une reconnaissance précoce des mouvements naturels

3.1 Introduction

Dans la précédente étude, nous avons mis en évidence qu'une normalisation de la morphologie permettait de réduire la variabilité inter-individuelle et apportait un gain dans le taux de reconnaissance de mouvements naturels ayant des caractéristiques spatiotemporelles similaires. Ce gain est particulièrement substantiel lorsque le sujet, dont les mouvements sont à reconnaitre, n'a encore jamais été observé par le système de reconnaissance (c.-à-d. qu'il ne fait pas partie de la base de données d'entrainement). La méthodologie mise en place pour cette première étude faisait appel aux modèles de Markov à états cachés (HMM). Elle a été sélectionnée car c'est la méthode la plus répandue dans la littérature en reconnaissance de mouvements. Elle se base sur une architecture parallèle de HMM modélisant chacun une classe de mouvements. Lors de la phase de reconnaissance, un mouvement inconnu est classé en calculant en parallèle sa vraisemblance vis-à-vis de chaque modèle HMM et en lui attribuant la classe du HMM ayant la vraisemblance la plus grande. Les temps de traitement que nous avons constatés sont tout à fait compatibles avec une application interactive, à ceci près, que la séquence doit être disponible dans son intégralité. Cette condition est imposée par l'algorithme de Viterbi, qui détermine la séquence d'état la plus probable pour une séquence d'observation donnée, selon un processus séquentiel de propagation-rétropropagation.Cela devient rédhibitoire pour une véritable utilisation interactive, dans laquelle l'environnement virtuel doit être en mesure de réagir au plus vite à la gestuelle de l'utilisateur, de préférence avant que le mouvement ne soit totalement achevé. Dans le cadre d'un combat virtuel en karaté, il n'est par exemple pas concevable qu'un personnage virtuel recevant un coup de poing de l'utilisateur ne réagisse pas avant que l'utilisateur ait tranquillement terminé son mouvement.

Dans cette étude, nous proposons donc une méthode permettant de reconnaitre le mouvement de façon précoce, c.-à-d. lorsque peu de données sont encore disponibles, tout en conservant un taux de reconnaissance équivalent à ce qu'on obtiendrait avec le mouvement entier. Nous nous appuyons pour cela sur les descripteurs amorphologiques, qui ont déjà démontré leur robustesse, ainsi que sur les descripteurs plus classiques (Cartésiens et angulaires).

Dans la section suivante nous dressons un bref état de l'art sur la problématique de la reconnaissance précoce. Puis, nous présentons, en section 3.3, la méthodologie mise en place pour y répondre. Nous évaluons ensuite cette méthodologie à travers deux sous-études, l'une basée sur un échantillonnage aléatoire de la base de données d'entrainement, l'autre basée sur un échantillonnage par sujet. Enfin, la dernière section (3.6) est consacrée à la conclusion sur les performances de la méthodologie.

3.2 Etat de l'art

Les HMM constituent la méthode de référence en reconnaissance de mouvements, mais pose un problème majeur dans le cadre d'applications temps réel : l'intégralité de la séquence de mouvement doit être observée avant qu'une décision sur sa classe d'appartenance puisse être prise. En effet, la phase de reconnaissance repose sur l'algorithme de Viterbi, qui permet, grâce à des méthodes d'espérance-maximisation, de déterminer la séquence d'état la plus probable pour une séquence observée. Seulement, l'algorithme de Viterbi utilise la procédure *forward-backward* (propagation-rétropropagation), qui nécessite de disposer de la séquence entière [Bishop2006]. Cela rend impossible la reconnaissance précoce du mouvement en cours d'exécution.

Le véritable problème est posé par la procédure de rétropropagation. Elle consiste à calculer les probabilités $P(\mathbf{o}(t+1)\dots\mathbf{o}(T)|q(t)=S_i,\lambda_m)$ d'observer la fin de la séquence $\mathbf{o}(t+1)\dots\mathbf{o}(T)$ (depuis un instant $t+1$ jusqu'à la fin (T)), en partant, à l'instant t, d'un état S_i de λ_m. Les méthodes de décodage incrémental, capables de fournir une estimation de ces probabilités au fil de l'eau, ne permettent généralement que de déterminer une solution sous-optimale [Fink2008].

Dans le cadre d'un suivi continu en temps réel d'une performance motrice, [Bevilacqua2010] conserve le formalisme des HMM, mais propose de supprimer la procédure de rétropropagation, afin de ne conserver que la propagation. Mais l'objectif est de synchroniser les instants clés de la performance avec une source extérieure (musicale en l'occurrence), et non de reconnaitre un mouvement parmi plusieurs classes.

Relativement peu d'études scientifiques se sont intéressées à la reconnaissance précoce du mouvement. Les travaux de [Mori2006] sont certainement la première tentative explicite pour aborder cette problématique. Une méthode de programmation dynamique permet d'établir une correspondance temporelle non linéaire entre un mouvement de référence et une observation inconnue. Un mouvement de la main peut ainsi être reconnu après que le système ait observé le premier quart de sa réalisation. Malheureusement la méthode et les résultats ne sont pas suffisamment détaillés.

À partir de données vidéo, [Axenbeck2008] exploite le formalisme HMM pour modéliser 6 classes de mouvements selon des topologies différentes, sélectionnées à partir d'heuristiques. La connaissance explicite des états permet de restreindre la recherche de la séquence optimale uniquement aux premiers états et aux tous premiers instants du mouvement. Cette approche autorise une

reconnaissance des mouvements dès la fin de la première seconde. Elle souffre toutefois d'un manque de généricité lié à l'explicitation de la topologie pour chaque modèle de classe. En outre, seules 2 classes de mouvements (pointer et dimensionner) peuvent réellement porter à confusion car elles sont paramétrées par un argument variable (la direction pointée, et la taille de l'écart entre les mains). Toujours sur des données vidéo, [Schindler2008] n'aborde pas réellement la question de la reconnaissance précoce, mais cherche plus précisément à déterminer la sous-séquence de longueur minimale qui permette une reconnaissance correcte de la séquence entière. Il détermine ainsi qu'une séquence d'observations de 0.3 à 0.5 secondes est suffisante. Néanmoins, cette séquence caractéristique n'a que peu de chance de se trouver au début du mouvement. De plus, les classes de mouvements sont encore une fois très discriminées (locomotions, coup de poing, coup de pieds...)

[Shimada2010] utilise des cartes auto-organisatrices pour encoder des postures à partir de positions 3D de marqueurs. Une métrique dédiée (distance de Hausdorf) est ensuite utilisée pour reconnaitre précocement 10 classes de mouvements du corps entier réalisés par plusieurs individus qui interagissent les uns avec les autres. Cette approche rend possible une reconnaissance précoce, à partir du moment où la moitié du mouvement est observée par le système, et cela avec un excellent taux de reconnaissance. Cependant parmi les 10 classes, 3 sont symétrisées à droite ou à gauche (coup de pied, coup de poing et claque) pour en obtenir 6 distinctes et seulement 3 classes peuvent être considérées comme assez similaires (punch, claque et uppercut).

En définitive, aucune étude ne s'est réellement penchée sur la reconnaissance précoce de classes de mouvements possédant des propriétés spatiotemporelles similaires, donc très sujettes à confusion. Au niveau méthodologique, la plupart de ces études ont cherché à se ramener au formalisme HMM, à travers diverses adaptations pour le rendre compatible avec une reconnaissance précoce. De plus, aucune étude n'a cherché à tester l'influence de la variabilité morphologique sur les performances de leurs méthodes.

Dans cette étude, nous simplifions le formalisme HMM et proposons d'évaluer plusieurs méthodologies fondées sur des modèles de mélange Gaussien, pour reconnaitre, de façon précoce, des classes de mouvements possédant des propriétés spatiotemporelles plus similaires que les travaux sus-cités. Dans le cadre de la reconnaissance précoce, les modèles de mélange Gaussien permettent de s'affranchir du point bloquant que constitue le décodage des HMM par propagation-rétropropagation, tout en conservant leurs propriétés d'encodage statistique de la variabilité du vecteur descripteur. En contrepartie, les modèles de mélange Gaussien sont incapables de modéliser la temporalité du mouvement, qui fait la grande force des HMM. De plus, l'influence de la variabilité inter-individuelle est explicitement abordée, en adoptant des répartitions de la base de données de mouvements tenant compte de l'identité des sujets. Nous étudions l'impact de cette variabilité sur les performances de reconnaissance des trois types de modèles GMM proposées, ainsi que sur les trois types de descripteurs du chapitre 2 (amorphologiques, Cartésiens et angulaires).

3.3 Méthodologie générale

3.3.1 Capture et description du mouvement

Nous utilisons la même base de données que dans l'étude précédente. La segmentation temporelle est réalisée manuellement, de sorte que le début et fin de chaque acquisition coïncident parfaitement avec le début et fin du mouvement. Chaque mouvement est ensuite intégré en tant qu'échantillon de la base de données.

Les vecteurs de descripteurs sont également conservés par rapport à l'étude du chapitre 2, c.-à-d. :

- une description angulaire à 24 descripteurs, comprenant, pour chaque côté, 3 angles d'Euler décrivant l'orientation du bras dans le repère de son épaule, et 3 angles d'Euler décrivant l'orientation de l'avant-bras dans le repère de son coude. La dérivée temporelle de chaque angle est également ajoutée.
- une description Cartésienne à 24 descripteurs, comprenant, pour chaque côté, les 3 coordonnées Cartésiennes de position du poignet et du coude dans le repère des hanches. La dérivée temporelle de chaque position est également ajoutée à la description.
- une description amorphologique, qui compte 12 descripteurs, comprenant, pour chaque bras, les 3 coordonnées de position normalisée du vecteur liant l'épaule au poignet. Ces coordonnées sont exprimées dans le repère local des hanches du sujet. Les 3 coordonnées de vélocité (dérivée temporelle) de ce vecteur font également partie du vecteur de descripteurs.

3.3.2 Modélisation par mélange Gaussien

Dans cette étude, nous cherchons à reconnaitre le mouvement le plus précocement possible, alors même que celui-ci n'est pas encore complètement réalisé. Cet objectif nous interdit d'utiliser les HMM pour la phase de classification, à cause de leur algorithme de décodage qui implique une rétropropagation sur la séquence intégrale [Fink2008]. Nous proposons donc de contourner cette limitation, en supprimant la nature séquentielle des HMM.

La première alternative que nous avons choisie, consiste à modéliser les mouvements par des modèles de mélanges Gaussiens (dénotés GMM, pour Gaussian Mixture Model). Ces modèles étaient utilisés par les HMM de l'étude précédente pour modéliser les probabilités d'observation du vecteur de descripteurs dans chaque état.

3.3.2.1 Entrainement

La phase d'entrainement des modèles consiste à estimer la distribution de la densité des échantillons de chaque classe de mouvement m, sous la forme d'un mélange de Gaussiennes multivariées. Intuitivement, cette estimation permet de « résumer » les échantillons composant la

classe m, à un ensemble restreint de paramètres caractérisant leur densité dans l'espace des descripteurs. Ces paramètres sont le poids ω_g, la moyenne μ_g et la variance Σ_g de chaque composante Gaussienne multivariée g. L'ensemble des paramètres $\omega = \{\omega_g\}$, $\mu = \{\mu_g\}$ $\Sigma = \{\Sigma_g\}$ pour toutes les N_G composantes Gaussiennes définit entièrement le GMM. Pour une classe de mouvements m, on note ce vecteur de paramètres $\theta_m = \{\omega, \mu, \Sigma\}$. La densité de probabilité d'un vecteur descripteur quelconque $\mathbf{o}(t)$ selon la distribution de densité de mélange du modèle $\theta_m$est alors donnée par :

$$P(\mathbf{o}(t)|\theta_m) = \sum_{g=1}^{N_G} \omega_g \mathcal{N}(\mathbf{o}(t)|\mu_g, \Sigma_g)$$

sous contrainte $\omega_g > 0 \ \forall g$ et $\sum_{g=1}^{N_g} \omega_g = 1$, et avec

$$\mathcal{N}(\mathbf{o}(t)|\mu_g, \Sigma_g) = \frac{1}{(2\pi)^{D/2} |\Sigma_g|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{o}(t) - \mu_g)^\top \Sigma_g^{-1} (\mathbf{o}(t) - \mu_g)\right)$$

En raison, à la fois de la grande dimensionnalité des données et d'une volonté de simplification, les Gaussiennes sont considérées à covariance diagonale. Cela implique qu'on ne tient pas compte des corrélations entre les descripteurs. La matrice de covariance prend la forme $\Sigma_g = \sigma_g I$, où I est la matrice identité.

Il n'existe pas de méthode de résolution analytique permettant d'ajuster une distribution GMM sur des données. L'estimation d'un modèle de classe est réalisée par une méthode de maximisation de la vraisemblance de l'ensemble des échantillons selon la distribution GMM de ce modèle. La distribution obtenue correspond à un maximum local. Plus concrètement, un ensemble d'entrainement de N_m séquences de mouvements $\{\mathbf{O}\}^{1:N_m}$ appartenant à une classe m est extrait de la base de données. Cet ensemble de séquences contient un grand nombre d'observations du vecteur descripteur $\mathbf{o}(t)$. Nous notons $\mathcal{O}_m = \{\{\mathbf{o}(t)\}^{1:N_m}\}$ l'ensemble de ces observations. À partir de l'ensemble d'entrainement $\mathcal{O}_m$, l'objectif de l'estimation est de déterminer les paramètres θ_m qui maximisent la vraisemblance :

$$P(\mathcal{O}_m|\theta_m) = \prod_{j=1}^{N_m} \prod_{t=1}^{T_j} P(\mathbf{o}^j(t)|\theta_m)$$

En pratique, N_m est supérieur à 10, T_j supérieur à 100 et $P(\mathbf{o}^j(t)|\theta_m)$ souvent très petit devant 1. Cette expression risque donc de générer rapidement un nombre infiniment petit. Pour éviter tout dépassement induit par la précision machine, on utilise préférentiellement le logarithme de cette vraisemblance.

$$\ln(P(\mathcal{O}_m|\theta_m)) = \sum_{\substack{1 \leq j \leq N_m \\ 1 \leq t \leq T_j}} \ln(P(\mathbf{o}^j(t)|\theta_m))$$

Les détails de cette maximisation de la vraisemblance sont disponibles dans le chapitre 9 de [Bishop2006]. Elle fait appel à l'algorithme espérance-maximisation (EM), une méthode itérative illustrée sur la figure 3.1 et dont les principales étapes sont les suivantes :

1. Initialiser les paramètres θ_m. L'algorithme K-moyennes fournit une bonne initialisation.
2. Étape *espérance* : calculer la log-vraisemblance $\ln(P(\mathcal{O}_m|\theta_m))$ avec ces nouveaux paramètres.

3. Étape *maximisation* : ajuster chaque paramètre de θ_m de sorte qu'il maximise la log-vraisemblance. Cet ajustement fait appel au calcul des dérivées partielles de la log-likelihood par rapport aux paramètres μ et Σ ainsi qu'à l'utilisation d'un multiplicateur de Lagrange pour affiner les coefficients de mélange ω.

4. Retourner à l'étape 2 (*espérance*) jusqu'à convergence.

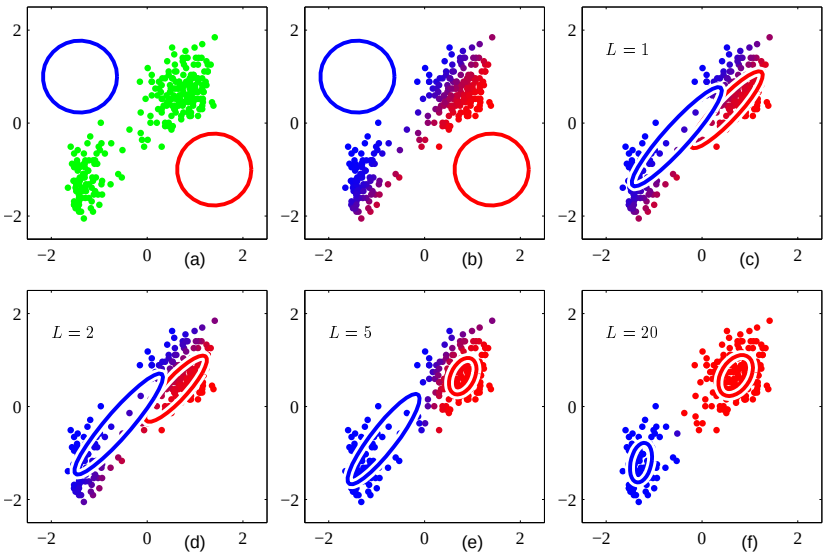


Figure 3.1 - Illustration des itérations successives de l'algorithme EM sur des données 2D que l'on cherche à estimer par un mélange de 2 composantes Gaussiennes à matrice de covariance complète (d'après [Bishop2006]). En haut, une initialisation aléatoire est proposée (à gauche), puis les données sont affectées à chaque composante (au centre) avant que les paramètres soient ajustés (à droite). En bas, de gauche à droite, les ajustements de la distribution après 1, 5 et 20 itérations de l'algorithme EM.

Pour éviter le phénomène de singularité, lié au fait qu'une composante Gaussienne puisse être ajustée sur un seul $\mathbf{o}(t)$, l'algorithme EM est relancé, le cas échéant. Si une singularité apparait 10 fois de suite, le nombre de composantes Gaussiennes est décrémenté jusqu'à obtention d'une distribution sans singularité. Comme la phase d'entrainement est exécutée entièrement hors ligne, le temps d'entrainement n'est pas crucial et le nombre de composantes Gaussiennes peut donc être initialement grand. Afin de garder une cohérence avec l'étude du chapitre 2, le nombre de composante Gaussienne est fixé à 42, soit l'équivalent des 6 Gaussiennes fois 7 états. En cas d'échec de l'algorithme EM, ce nombre de composantes est réduit. La figure 3.2 montre l'ajustement de la distribution pour la classe de mouvements *claque paume* superposée aux observations de 6 descripteurs amorphologiques.

3.3.2.2 Classification

Une fois la distribution de chaque classe m estimée selon un GMM θ_m, la classification d'une séquence inconnue $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$ est réalisée en calculant en parallèle sa log-

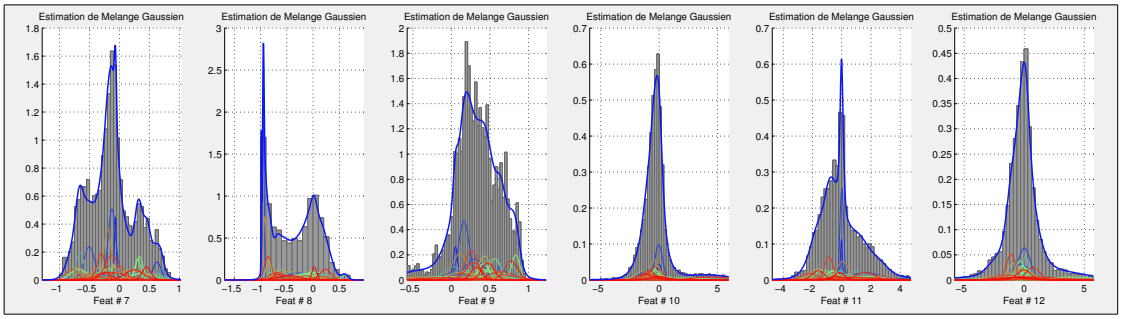


Figure 3.2 - Exemple d'ajustement d'un modèle GMM pour 6 descripteurs amorphologiques. De gauche à droite, les descripteurs sont les suivants : coordonnée x de $\mathbf{r}^{droit}_{amorpho}$ (feat #7), puis y (feat #8), puis z (feat #9), idem pour les coordonnées $\dot{x}$(feat #10), $\dot{y}$ (feat #11) et $\dot{z}$ (feat #12) de $\dot{\mathbf{r}}^{droit}_{amorpho}$. Les histogrammes gris représentent la fréquence d'observation de chaque valeur du descripteur. Les courbes épaisses bleues sont les ajustements des modèles GMM sur les observations du descripteur. Les courbes d'autres couleurs sont les composantes Gaussiennes du GMM.

vraisemblance selon chaque θ_m :

$$\ln(P(\mathbf{O}|\theta_m))=\sum_{t=1}^{T}\ln(P(\mathbf{o}(t)|\theta_m))$$

La classe attribuée à la séquence inconnue est celle dont la distribution GMM possède la plus grande log-vraisemblance avec la séquence.

$$Classe(\mathbf{O}) = \underset{m\in\{1...M\}}{\arg\max}\ \ln(P(\mathbf{O}|\theta_m))$$

L'intérêt des GMM en phase de classification est de pouvoir nativement calculer la vraisemblance sur n'importe quel sous-segment temporel du mouvement inconnu. Chaque nouvelle observation du vecteur descripteur vient modifier la vraisemblance cumulée de la séquence en la multipliant par un facteur correspondant à sa vraisemblance individuelle propre (en l'additionnant en log-vraisemblance). Il est donc possible de classer la séquence alors qu'elle n'est pas complètement terminée. Les sections 3.4 et 3.5 sont consacrées à la détermination du seuil de temps nécessaire à une reconnaissance efficace.

3.3.3 Modélisation par mélange Gaussien à états

Dans cette section, nous proposons une alternative aux modèles de mélange Gaussien, fondée sur la réutilisation des états HMM calculés au chapitre 2. Afin de différencier clairement les modèles, nous appelons GMM naïfs les modèles présentés en section précédente, et nous nommons GMM à états les modèles présentés dans cette section.

L'objectif est de conserver une partie de l'information séquentielle extraite par les HMM et introduite dans les états. En effet, la topologie linéaire *gauche-droite* des HMM implique que les observations du vecteur descripteur dans chaque état ont un lien séquentiel fort, en plus

d'avoir un lien de proximité dans l'espace de description. Les densités qui en résultent sont donc caractéristiques de différentes phases temporelles du mouvement. À l'inverse, les GMM naïfs modélisent les densités d'observations uniquement en fonction de leurs similarités dans l'espace de description, indépendamment de la proximité temporelle. En préservant ces états, on fait l'hypothèse que les GMM à états conservent une partie de l'information séquentielle qui échappe aux GMM naïfs.

Concrètement, les états des HMM sont conservés, mais les transitions, elles, sont supprimées, puisqu'elles rendent impossible la reconnaissance précoce. Les matrices A et Π sont donc écartées pour ne conserver que les distributions de densité du vecteur descripteur dans chaque état, stockées dans le paramètre $B = \{b_i\}$. On dispose à présent d'une collection d'états sans lien séquentiel, que nous nommons GMM à états. La figure 3.3 rend compte de la disparition des paramètres A et Π.

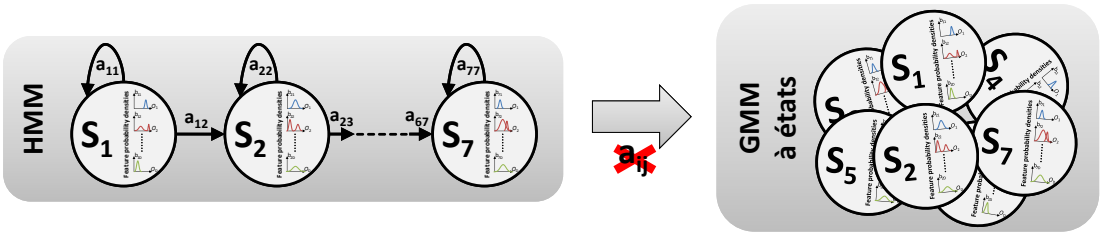


Figure 3.3 - Illustration du passage des modèles HMM utilisés pour l'étude du chapitre 2, aux modèles GMM à états de la présente étude. La suppression des liens de transition qui modélisaient la temporalité du processus Markovien transforme le HMM en une collection d'états disparates.

3.3.3.1 Entrainement

Afin de visualiser la qualité de l'ajustement des distributions avec les vecteurs descripteurs, nous avons déterminé les chemins de Viterbi pour chaque échantillon gestuel appartenant aux M différentes classes de la base de données d'entrainement du chapitre 2. Le chemin de Viterbi fournit la séquence d'états la plus probable pour un échantillon $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$ et un λ_m donnés. En observant ces chemins d'états sur les échantillons d'entrainement de la classe m, il est possible d'obtenir l'ensemble des observations du vecteur descripteur correspondant à chaque état de λ_m. La figure 3.4 représente les distributions GMM dans chaque état, superposées aux données sur lesquelles elles sont ajustées, pour un modèle HMM issu du chapitre 2. Pour simplifier la lecture, le modèle représenté sur la figure utilise les descripteurs amorphologiques pour modéliser un mouvement de claque réalisé exclusivement avec la paume de la main droite. Le descripteur #7 (1ère colonne, fig. 3.4) correspond au déplacement latéral x du poignet droit par rapport à l'épaule droite dans le repère local des hanches, le descripteur #8 à son déplacement vertical y et le descripteur #9 à son déplacement antéro-postérieur z. Le repère est direct, avec x défini positivement vers la gauche, y vers le haut et z vers l'avant. On constate que la temporalité du mouvement a été correctement et automatiquement encodée par le HMM :

- dans l'état S_1 (1ère ligne) la main est dans une position de repos le long du corps ;
- dans l'état S_2, elle se dirige vers la droite, le haut et l'avant ;
- en S_3, la main balaie l'espace de droite à gauche et la vitesse vers la gauche est importante

($\dot{x}$ en 4ème colonne) ;

- en S_4, la main est arrivée sur la gauche à hauteur d'épaule ;
- lors des états S_5 et S_6, la main regagne sa position de repos ;
- état S_7, la main est à nouveau au repos le long du corps.

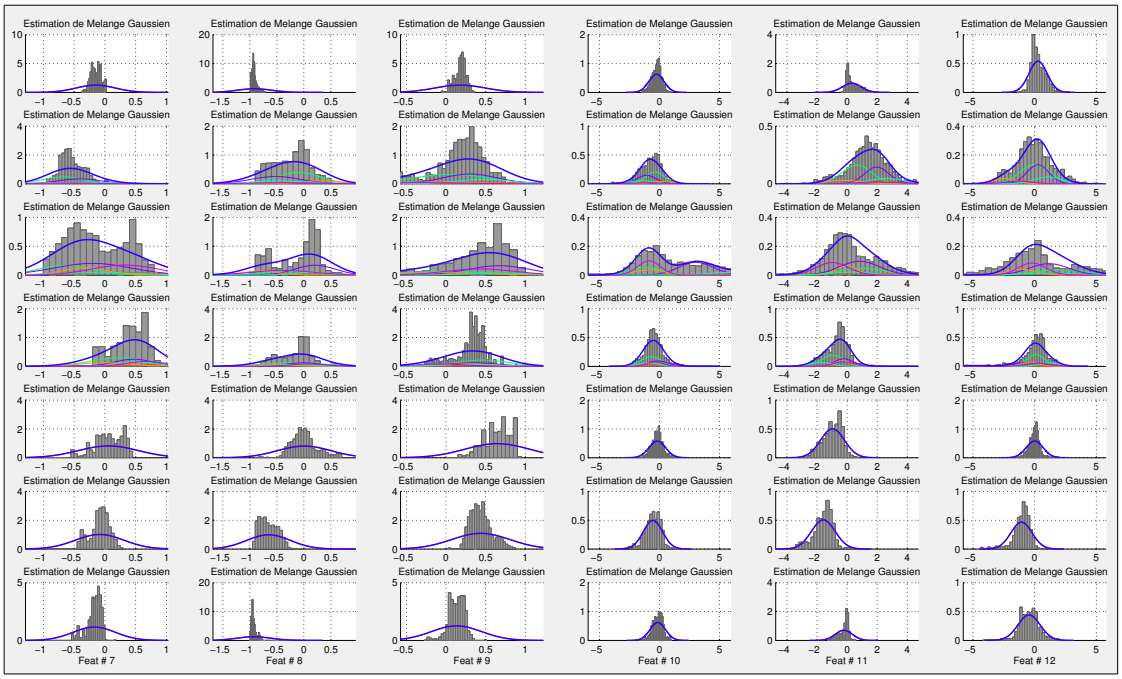


Figure 3.4 - Exemple d'ajustement des distributions du vecteur descripteur dans chaque état du HMM (de S_1 en haut à S_7 en bas) relativement aux données d'entrainement (le mouvement *claque de la paume*, effectué de la main droite ici). Ce modèle HMM est issu du chapitre 2. Les histogrammes gris représentent la fréquence de chaque descripteur observé. Les courbes épaisses bleues sont les ajustements GMM, composés des gaussiennes représentées par les courbes d'autres couleurs. Pour simplifier la lecture, seuls les descripteurs relatifs au bras droit sont indiqués sur les différents graphes. De gauche à droite, les descripteurs sont les suivants : coordonnée x de $\mathbf{r}^{droit}_{amorpho}$ (feat #7), puis y (feat #8), puis z (feat #9), idem pour les coordonnées $\dot{x}$, $\dot{y}$ et $\dot{z}$ de $\dot{\mathbf{r}}^{droit}_{amorpho}$

Néanmoins, l'ajustement présenté par la figure 3.4 manque de spécificité sur certains descripteurs et certains états. Par exemple, le descripteur de la 3ème colonne dans l'état 4 (4ème ligne) est modélisé par un GMM qui ne retranscrit pas la fréquence d'observation plus importante pour des valeurs entre 0.3 et 0.5. La distribution GMM (courbe bleue) n'est pas suffisamment resserrée autour des valeurs observées (histogrammes gris). Elle n'est donc pas très spécifique aux observations attendues dans cet état.

Par conséquent, nous ré-estimons toutes les distributions GMM qui modélisent chaque état, à partir des données de mouvements qui leur sont attribuées par l'algorithme de Viterbi. La ré-estimation est effectuée par l'algorithme EM que nous avons présenté en section précédente, à la nuance près, que les données d'entrainement d'une classe m sont regroupées en fonction de leur appartenance à un état. Ainsi, chaque observation du vecteur descripteur $\mathbf{o}(t)$ appartenant à l'ensemble d'entrainement, noté précédemment $\mathcal{O}_m$, est redistribuée vers un sous-ensemble d'entrainement $\mathcal{O}_{m/S_i}$, en fonction de l'état S_i qui lui a été attribué par l'algorithme de Viterbi.

Nous nommons θ_{m/S_i} le modèle GMM ré-estimé à partir de l'ensemble $\mathcal{O}_{m/S_i}$. Nous conservons la même paramétrisation qu'au chapitre 2, à savoir 7 états composés de 6 Gaussiennes chacun.

Pour résumer, à partir de tous les échantillons d'entrainement $\{\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)\}^m$ d'une classe de mouvements m, les étapes de la procédure de ré-estimation sont les suivantes :

1. Pour chaque $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$, déterminer la séquence d'états optimale par rapport au modèle λ_m ([Fink2008]). Ce qui équivaut à maximiser la vraisemblance

$$\mathbf{q}_{viterbi}(\mathbf{O}, \lambda_m) = \underset{q(1),\ldots,q(T)\in\{S_i\}^T}{\arg\max} P(q(1), \ldots, q(T)|\mathbf{o}(1), \ldots, \mathbf{o}(T), \lambda_m)$$

2. Assigner chaque $\mathbf{o}(t)$ à l'ensemble d'entrainement $\mathcal{O}_{m/S_i}$ de l'état $q(t) = S_i$ attribué par l'algorithme de Viterbi.
3. Pour chaque $\mathcal{O}_{m/S_i}$, déterminer grâce à l'algorithme EM le modèle GMM θ_{m/S_i}, qui caractérise la densité des vecteurs descripteurs le composant.

À la sortie de cette procédure, chaque classe de mouvements m est modélisée par un modèle $\theta_m = \{\theta_{m/S_1}, \ldots, \theta_{m/S_i}, \ldots, \theta_{m/S_7}\}$.La figure 3.5 présente le modèle GMM correspondant à la classe *claque paume* effectué avec la main droite. Par comparaison avec la figure 3.4, on peut observer que l'estimation des densités par les distributions GMM sont mieux ajustées aux données d'entrainement.

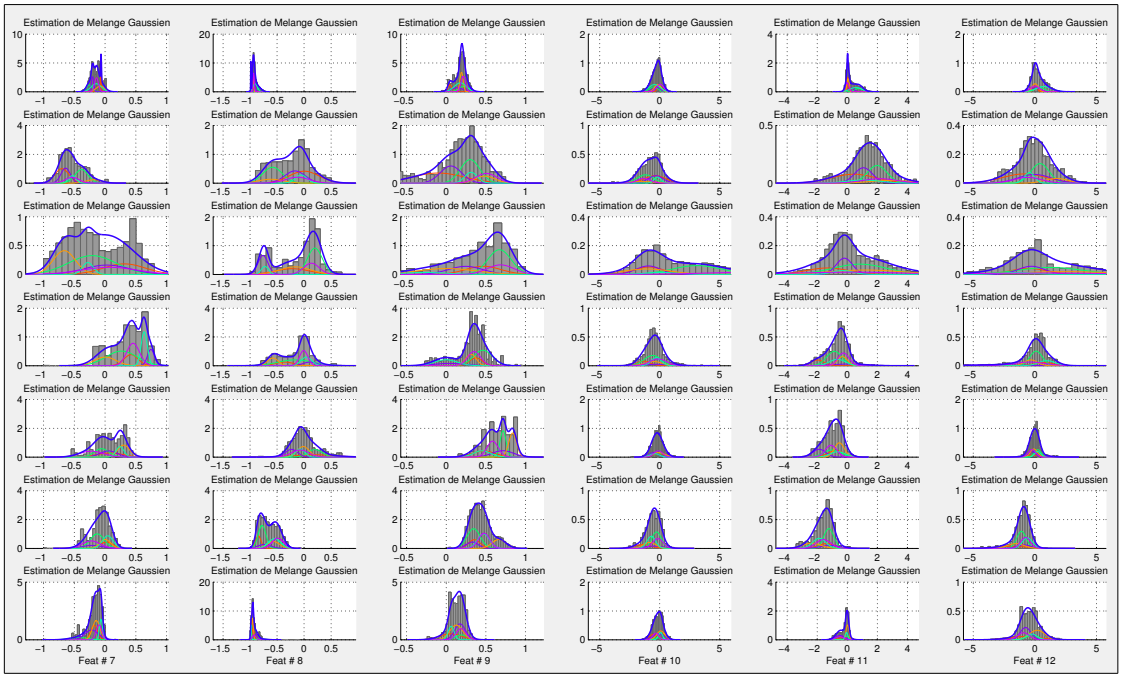


Figure 3.5 - Exemple d'ajustement des distributions du vecteur descripteur dans chaque état du GMM fondé sur les états du HMM. Les distributions sont sensiblement plus affinées autour des données d'entrainement, comparées à celles du HMM (fig. 3.4).

3.3.3.2 Classification

D'une manière analogue aux modèles GMM, les mélanges Gaussiens à états sont ensuite directement utilisables en classification. À partir d'une séquence gestuelle inconnue $\mathbf{O} = \mathbf{o}(1), \ldots, \mathbf{o}(t), \ldots, \mathbf{o}(T)$,

on détermine, pour chaque modèle θ_m, la vraisemblance individuelle de chaque $\mathbf{o}(t)$ avec les états θ_{m/S_i} qui composent θ_m. La vraisemblance cumulée globale de la séquence avec le modèle θ_m est calculée en sélectionnant l'état de vraisemblance maximale pour chaque à $\mathbf{o}(t)$. Ce calcul est formalisé par l'expression suivante :

$$P(\mathbf{O}|\theta_m) = \prod_{t=1}^{T} \max_{S_i} P(\mathbf{o}(t)|\theta_{m/S_i})$$

En pratique, on utilise le logarithme de cette expression pour éviter les dépassements de précision numérique. Finalement, on attribue à la séquence gestuelle inconnue la classe du modèle ayant la vraisemblance cumulée la plus grande.

3.3.4 Pondération temporelle des modèles de mélange Gaussien à états

En supprimant les matrices de transition des HMM, nous abandonnons l'aptitude de ces derniers à modéliser la variabilité dans la temporalité du mouvement. Pour compenser cette disparition, nous proposons une variante des GMM à états, dans laquelle nous introduisons une pondération temporelle explicite des états. L'idée d'une telle pondération est que la vraisemblance d'un état peu probable pendant certaines phases du mouvement doit être pondérée négativement lors de ces phases. Considérons, par exemple, l'état S_4 d'un modèle de mouvement *claque paume* de la main droite, dont on a vu précédemment qu'il correspondait à la phase où la main se trouve à hauteur d'épaule légèrement sur la gauche. Cette phase, qui intervient en fin d'exécution du mouvement, a très peu de chance d'être observée au début du mouvement. En revanche, un état ressemblant à celui-ci intervient au début du modèle *claque revers*. Une pondération temporelle doit donc permettre de sanctionner cet état au début du modèle *claque paume*, mais pas au début du modèle *claque revers*.

Concrètement, ces pondérations temporelles prennent la forme d'un coefficient $\rho_{m/S_i}(t)$, qui vient modifier la vraisemblance de chaque état θ_{m/S_i} en fonction de l'instant t auquel cet état est attendu dans un mouvement de classe m. Le calcul de ces pondérations s'appuie à nouveau sur les HMM du chapitre 2 et sur l'algorithme de Viterbi.

3.3.4.1 Entrainement

L'entrainement de la modélisation GMM à états est strictement identique à celui décrit précédemment. Nous détaillons dans cette partie les nouveautés introduites par les pondérations temporelles.

La détermination de ces pondérations temporelles s'effectue en plusieurs étapes. Tout d'abord, chaque mouvement de la base de données est exprimé en pourcentage de sa durée totale. Ensuite, pour chaque échantillon gestuel $\mathbf{O}^j$ appartenant à une classe m, le chemin de Viterbi $q^j(1), \ldots, q^j(t), \ldots, q^j(T = 100)$ fournit les instants t où chaque état S_i du HMM λ_m est « activé ». La pondération temporelle de chaque état θ_{m/S_i} du GMM à état θ_m est obtenue en établissant la moyenne des activations de l'état S_i sur l'ensemble des échantillons de la classe

m.

$$\rho_{m/S_i}(t) = \frac{1}{N_m}\sum_{j=1}^{N_m}\delta_{ij}(t), \text{ avec } \delta_{ij}(t) = \begin{cases} 0, \text{si } q^j(t) \neq S_i \\ 1, \text{si } q^j(t) = S_i \end{cases}$$

Ces fonctions de pondération temporelle sont ensuite normalisées entre 0 et 1 et lissées par un filtre passe bas.Le tracé de ces fonctions est présenté sur la figure 3.6 pour chaque mouvement (graphes individuels) et chaque état (courbes de couleur). On peut constater que les mouvements peu dynamiques, tels que *bras croisés*, *gratter menton*, *mains hanches* ou *mains poches* voient leur états centraux (S_3 à S_5) activés plus ou moins simultanément en milieu de mouvement. Les mouvements présentant des phases cycliques, tels que *applaudir*, *salut haut* et *salut tête* ont un comportement assez similaire. Les états S_3 à S_5 semblent correspondre aux phases du mouvement pendant lesquelles interviennent les mouvements cycliques qui sont porteurs de la sémantique de ces mouvements (les mains qui se clapent, ou l'oscillation de la main qui salue). Enfin, les mouvements dynamiques présentent un découpage séquentiel relativement strict, les états se succédant dans l'ordre attendu. Quel que soit le type de mouvement, on remarque également que les états S_1 et S_7 sont bien activés pendant les phases de repos en début et fin de mouvement.

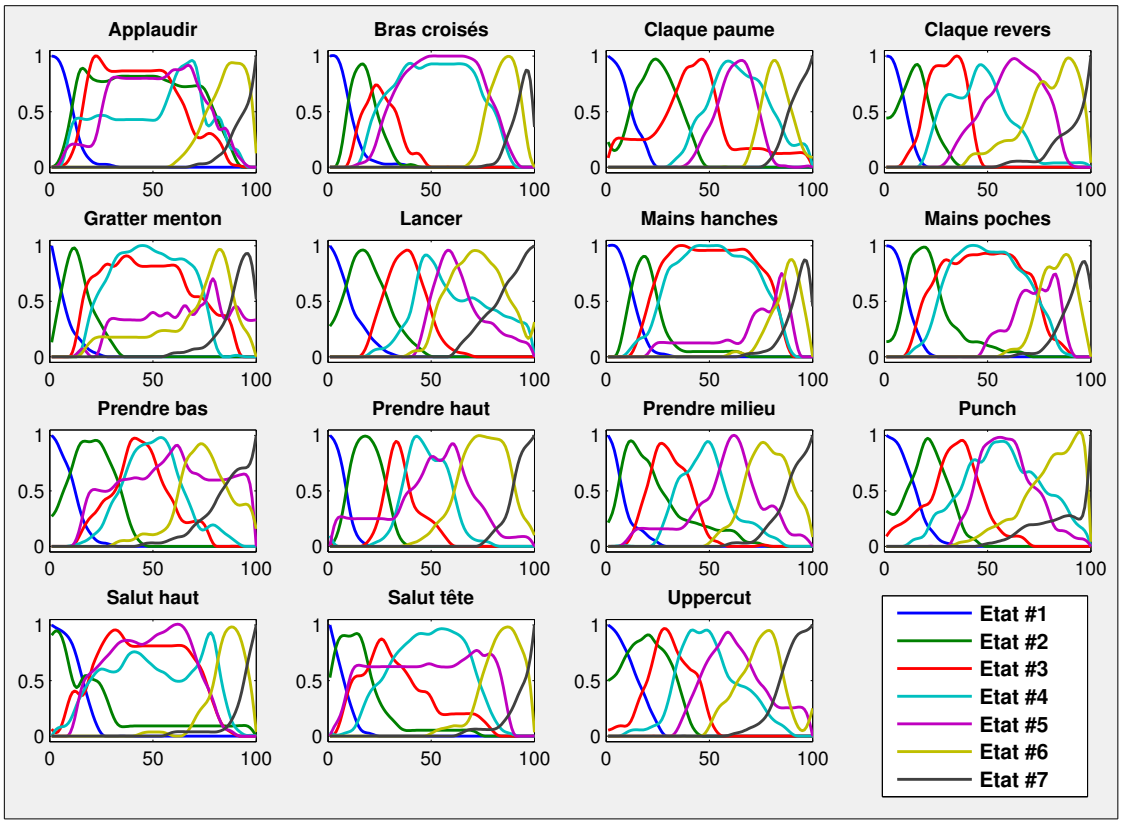


Figure 3.6 - Tracés (courbes de couleur) des fonctions de pondération temporelle $\rho_{m/S_i}(t)$ de chaque état S_i ($i \in [1,7]$) pour chaque classe de mouvement. Le temps (en abscisse) est normalisé à 100% de la durée d'exécution du mouvement.

3.3.4.2 Classification

En phase de classification, la fonction de pondération temporelle rend la vraisemblance d'une observation dépendante de l'instant où celle-ci intervient dans la séquence gestuelle. La nouvelle vraisemblance d'une observation du vecteur descripteur avec le modèle pondéré θ_m^{pond} à un instant t (normalisé) prend la forme :

$$P(\mathbf{o}(t)|\theta_m^{pond}) = \max_{S_i \in [S_1 ... S_7]} [\rho_{m/S_i}(t).P(\mathbf{o}(t)|\theta_{m/S_i})]$$

Ainsi, si un état est peu probable à un certain instant t, il voit sa vraisemblance réduite d'un facteur $\rho_{m/S_i}(t)$. La vraisemblance cumulée d'une séquence gestuelle inconnue $P(\mathbf{O}|\theta_m)$ est encore obtenue par le produit des $P(\mathbf{o}(t)|\theta_m)$. En pratique, les calculs sont réalisés avec les logarithmes de toutes ces grandeurs. Notons qu'il est nécessaire de ré-échantillonner la fonction de pondération temporelle de sorte que la durée de celle-ci corresponde à celle de la séquence gestuelle inconnue.

Pour vérifier l'efficience des corrections apportées par les pondérations temporelles sur la vraisemblance cumulée, un indicateur est proposé. Il consiste, pour chaque observation du vecteur descripteur $\mathbf{o}(t)$, à déterminer l'état S_{iMax} appartenant au modèle de classe M_{Max} qui maximise la vraisemblance $P(\mathbf{o}(t)|\theta_{M_{Max}/S_{iMax}})$ de $\mathbf{o}(t)$ avec la modélisation GMM à états. Si la fonction de pondération associée $\rho_{M_{Max}/S_{iMax}}(t)$ est inférieure à un seuil fixé à 0.1 (on rappelle que $\rho \in [0, 1]$), pour cet instant t, l'état et le modèle attribués à l'observation sont dits incompatibles avec la fonction de pondération temporelle. Sur l'intégralité de la séquence gestuelle, on relève ensuite le pourcentage d'observations $\mathbf{o}(t)$ incompatibles avec les fonctions de pondération temporelle pour la modélisation GMM à état d'une part et pour la modélisation GMM à états pondérés d'autre part.

3.3.5 Synthèse

Dans cette section nous avons présenté trois déclinaisons de modèles de mélange Gaussien, compatibles avec une reconnaissance précoce du mouvement. Le premier GMM, dit naïf, hérite du formalisme classique. Le second GMM, dit à états, est une adaptation des HMM du chapitre 2 débarrassé des matrices de transitions, qui permet de modéliser par des états les observations présentant un fort lien de séquentialité. Le GMM à états pondérés est une adaptation du précédent, qui permet de pondérer les états en fonction de leur probabilité d'apparaitre à chaque instant.

L'objectif principal de cette étude est d'évaluer les descripteurs amorphologiques dans des conditions permettant une reconnaissance précoce du mouvement. Ces conditions sont réunies grâce aux modèles GMM naïfs et GMM à états, dont nous venons de détailler le formalisme dans cette section.

Comme dans le chapitre 2, nous mettons en place une méthodologie d'évaluation, qui permet de mettre en évidence le comportement du système de reconnaissance face à différentes formes de variabilité du mouvement naturel. La sous-étude 1 aborde de manière classique la variabilité, alors que la sous-étude 2 se focalise sur la variabilité inter-individuelle.

3.4 Sous-étude 1 : Répartition aléatoire

3.4.1 Méthode d'évaluation

Les modèles probabilistes, dont font partie les 3 types de GMM que nous cherchons à évaluer, utilisent les données pour ajuster leurs paramètres (les poids, moyennes et variances entre autres). Pour les évaluer, il est donc nécessaire de répartir notre base de données en un sous-ensemble d'entrainement, qui permet aux modèles de s'ajuster, et un sous-ensemble de validation, qui permet d'apprécier la qualité de la reconnaissance du système sur des données nouvelles (non connues du système).

Dans cette sous-étude, la méthode d'évaluation des performances du système de reconnaissance consiste à répartir aléatoirement les échantillons de chaque classe de la base de données entre un sous-ensemble d'entrainement et sous-ensemble de validation. La proportion d'échantillons allouée à l'entrainement varie de 10% à 90% de l'effectif total de la base de données de mouvements, par pallier de 20%. Les échantillons restants constituent la base de validation. Afin de s'assurer qu'il n'existe pas d'effet lié à la répartition aléatoire des échantillons, 10 tirages aléatoires sont effectués pour chaque ratio entrainement/validation.

La phase d'entrainement consiste, pour chaque répartition aléatoire de la base de données, à entrainer des systèmes de reconnaissances fondés sur les 3 modèles proposés : GMM naïfs, GMM à états et GMM à états pondérés. En entrée de chaque modèle, les 3 types de descripteurs du mouvement (amorphologiques, Cartésiens et angulaires) sont utilisés, donnant lieu à 3 systèmes de reconnaissance par modèle. À la sortie de la phase d'entrainement, on dispose ainsi de 9 systèmes de reconnaissance.

La phase d'évaluation consiste à déterminer le taux de reconnaissance de chaque système à partir des échantillons de validation. Les résultats sont présentés sous forme de tableaux et de graphiques commentés. Des tests statistiques sont réalisés pour déterminer la significativité des résultats. Comme 3 conditions différentes sont généralement testées (les 3 modèles ou les 3 types de descripteurs), les tests consistent en une ANOVA par rang de Friedman sur le taux de reconnaissance moyen de chacune des 15 classes de mouvement et pour chacun des 10 tirages aléatoires de la base de donnée. Si l'ANOVA met en évidence un lien entre les modèles ou les descripteurs et la performance du système, un test posthoc des rangs signés de Wilcoxon permet de déterminer quels modèles ou quels descripteurs possèdent une performance significativement supérieure ou inférieure aux autres.

D'autre part, l'objectif principal de cette étude est d'évaluer la capacité de chaque système à reconnaitre les mouvements le plus précocement possible. Nous évaluons donc le taux de reconnaissance pour plusieurs instants après le début de l'exécution de chaque échantillon de validation. Ces instants sont définis en pourcentage de la durée totale de l'échantillon. Nous relevons les taux de reconnaissance après 10%, 20%, 30%,..., 100% de la durée totale d'exécution du mouvement. Pour la clarté de l'exposé, ces conditions sont référencées respectivement par les abréviations $t_{10\%}$, $t_{20\%}$, $t_{30\%}$,...,$t_{100\%}$.

La contrainte temps réel implique également que le système de reconnaissance ait des temps de calculs relativement faibles. Le temps de calcul, en Matlab, correspond au temps moyen mis par le système pour reconnaitre l'intégralité d'une séquence un gestuelle ($t_{100\%}$), en conditions sur une architecture matérielle composée d'un processeur Intel Core 2 Quad (Q8400) cadencé à 2.66GHz avec 4Go Ram. Cependant, l'algorithme n'exploite qu'un seul des 4 processeurs.

3.4.2 Résultats pour la répartition 50% entrainement / 50% validation

Dans cette section, les bases de données d'entrainement et de validation comptent le même nombre d'échantillons. Il s'agit de la méthode de validation la plus classique. Les tracés de la figure 3.7 présentent le taux de reconnaissance en fonction de la fraction du mouvement observée (de $t_{10\%}$ à $t_{100\%}$) pour le système de reconnaissance fondé sur les GMM naïfs (à gauche) et pour le système fondé sur les GMM à états (à droite). Cette figure permet de dégager quelques constats globaux.

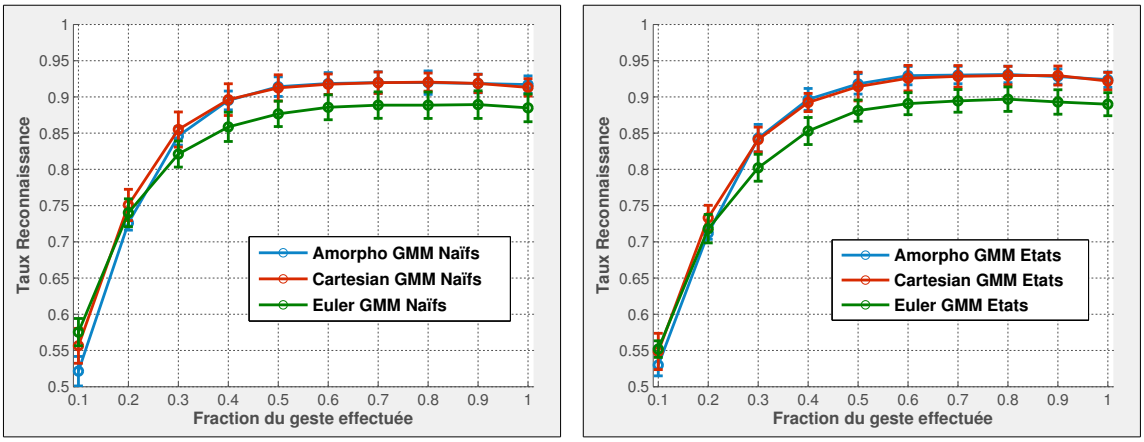


Figure 3.7 - Taux de reconnaissance moyen en fonction de la fraction de mouvement observée. À gauche pour des modèles GMM naïfs, à droite pour des modèles GMM à état. Les types de descripteurs utilisés apparaissent dans différentes couleurs (cf. légende interne à la figure).

Premièrement, un plateau asymptotique se dégage après $t_{50\%}$ et se maintient jusqu'à $t_{100\%}$ (mouvement complètement exécuté), quels que soient le type de descripteurs et de modèles choisis. Cela tend à démontrer qu'une reconnaissance précoce est envisageable à partir de l'instant où la moitié du mouvement est observée par le système, puisque le taux de reconnaissance n'évolue plus après.

Le second constat tient aux taux de reconnaissance asymptotiques très similaires obtenus par chaque type de modèles. Ils atteignent 88% à 93% en fonction des descripteurs utilisés. L'intérêt des GMM à états peut être remis en cause, étant donné l'absence d'amélioration significative par rapport aux GMM naïfs. Pour aller plus loin, en comparant ces performances à celles obtenues par les HMM du chapitre précédent (tableau 2.2), qui sont également de l'ordre de 90-91% pour les descripteurs amorphologiques et Cartésiens, on pourrait s'interroger sur l'apport réel des HMM, puisque des modèles de mélange Gaussien naïfs obtiennent des résultats équivalents.

Notons que les résultats obtenus avec les pondérations temporelles des états ne sont pas présentés ici, car ils sont très similaires (courbe du taux de reconnaissance atteignant un plateau asymptotique autour de 90-93% à partir de $t_{50\%}$). Pour chaque type de descripteurs, une ANOVA par rang de Friedman ne montre aucune différence entre les GMM naïfs, les GMM à états et les GMM à états pondérés.

Enfin, le dernier constat tient aux performances des différents descripteurs. D'abord, les taux de reconnaissance des descripteurs amorphologiques et Cartésiens se confondent à partir de $t_{40\%}$. Un test post-hoc des rangs signés de Wilcoxon réalisé sur les taux de reconnaissance moyens par

mouvement et par sujet, pour les fractions d'observation de $t_{40\%}$ à $t_{100\%}$, démontre à chaque fois l'absence de différence significative. Ensuite, le taux de reconnaissance des descripteurs angulaires accuse un certain retrait. Pour chaque fraction de mouvement au delà de $t_{40\%}$, une ANOVA par rang d'ordre 2 de Friedman démontre l'influence de chaque type de descripteurs sur les taux de reconnaissance moyen. Des tests post-hoc des rangs signés de Wilcoxon confirment, à chaque fois, la supériorité significative de la performance des descripteurs amorphologiques et Cartésiens sur les descripteurs angulaires.

La figure 3.8 présente graphiquement ces résultats par classe, pour des mouvements intégralement observés. Sur cette figure, chaque graphe présente les résultats de reconnaissance d'un type de modèles (un par ligne) utilisant un type de descripteurs (un par colonne). Les taux de reconnaissance pour chaque classe apparaissent en ordonnées. Plus un modèle est capable de reconnaitre correctement des mouvements appartenant à sa classe, plus il se trouve haut sur le graphe. Les taux de faux positifs apparaissent en abscisses. Ils caractérisent la spécificité de chaque modèle et correspondent au pourcentage de mouvements attribués à tort à chaque classe. Plus le modèle est spécifique à la classe de mouvement qu'il modélise, plus il apparait sur la gauche. Les ellipses en pointillés autour de chaque point figurent l'écart-type constaté entre les 10 répétitions de l'évaluation.

Au delà des taux de reconnaissance très similaires en moyenne, cette figure atteste que chaque classe de mouvements individuelle est reconnue de manière sensiblement équivalente, quels que soient les modèles utilisés. Elle permet également de mettre en évidence le comportement relativement équivalent des résultats individuels par classe pour les descripteurs amorphologiques et Cartésiens.

D'autre part, la figure 3.9 reprend la charte graphique de la figure 3.8 pour présenter les résultats pour des mouvements dont seule la première moitié est observée ($t_{50\%}$). L'extrême similarité de ces deux figures met en avant la stabilisation des résultats sur le plateau asymptotique.

Pour chaque type de descripteurs et pour chaque modèle, les temps de calculs moyens pour reconnaitre un mouvement ont été relevés (cf. tableau 3.1). Ces temps de traitement, quels que soient les modèles et le type de descripteurs, sont compatibles avec le temps réel. La durée moyenne d'un mouvement est de 3.39s, quand le temps de calcul du processus de reconnaissance atteint au maximum 195ms pour la modélisation GMM à états pondérés. Soit moins de 6% de la durée d'un mouvement. Enfin, les descripteurs amorphologiques permettent les calculs les plus rapides, quels que soient les modèles, ce qui est normal puisqu'ils comptent moins de descripteurs (12, contre 24 pour les deux autres).

3.4.3 Résultats pour les autres répartitions

Nous venons de démontrer qu'il n'existe pas de différence notable entre les 3 modèles GMM proposés, lorsque la base de données de mouvement est répartie aléatoirement à parts égales entre le sous-ensemble d'entrainement et le sous-ensemble de validation. Dans ces mêmes conditions, les descripteurs amorphologiques et Cartésiens montrent également des performances similaires.

Dans cette section, nous faisons varier les effectifs de la répartition de la base de données

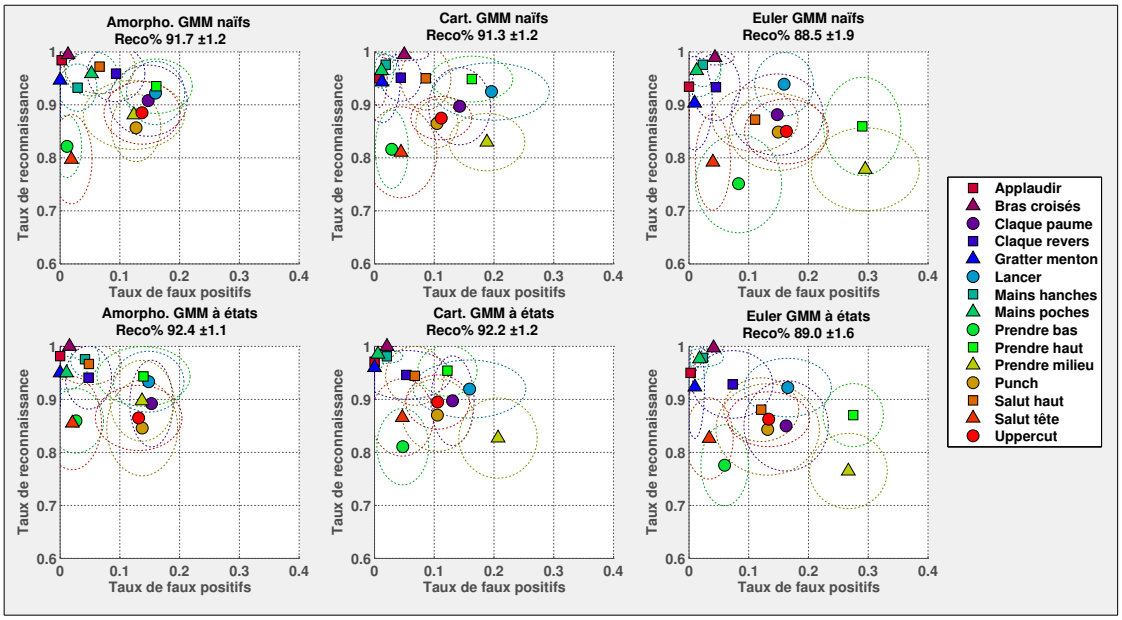


Figure 3.8 - Taux de reconnaissance (en ordonnées) et de faux positifs (en abscisses) pour chaque classe (points de couleur) sur des mouvements entièrement observés ($t_{100\%}$). Les graphes de la ligne du haut présentent les résultats des GMM naïfs, et ceux du bas, les GMM à état. De gauche à droite, les graphes des colonnes présentent les performances des modèles utilisant les descripteurs amorphologiques, Cartésiens puis angulaires. Notons bien que l'échelle est agrandie sur les taux de reconnaissance > 60% et les taux de faux positifs < 40%.

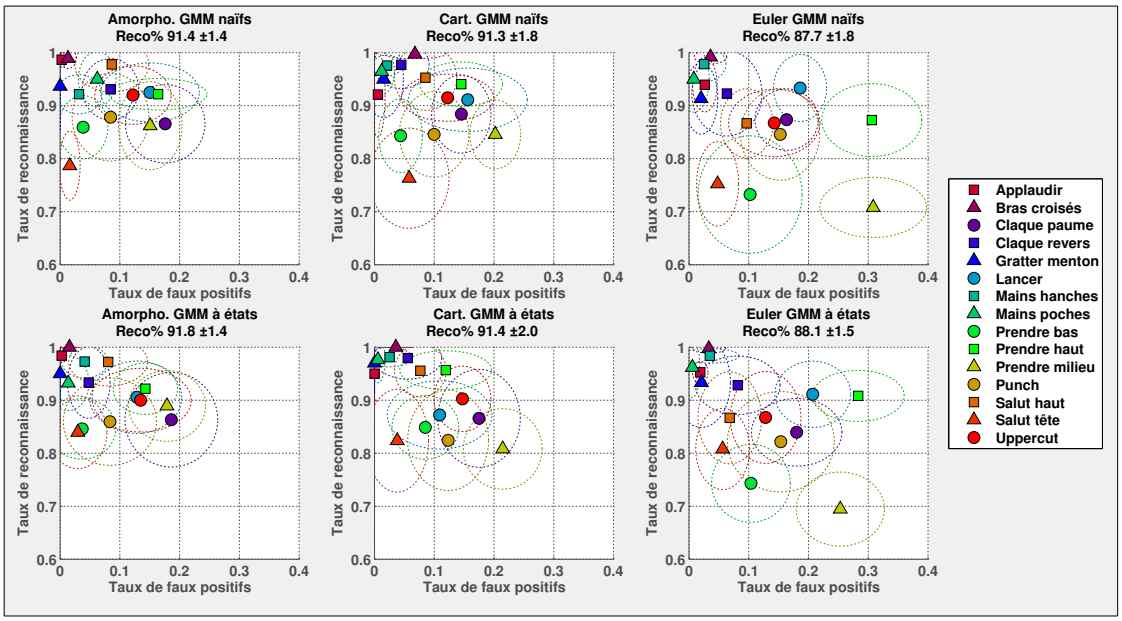


Figure 3.9 - Taux de reconnaissance et de faux positifs par classe sur des mouvements partiellement observés ($t_{50\%}$) (voir légende figure 3.8).

	Amorpho. (ms)	**Cart. (ms)**	**Euler (ms)**
GMM naïf	74	92	99
GMM à états	88	106	195
GMM à états pondérés	147	195	181
HMM	25	38	37

Table 3.1 - Temps de calcul moyen en fonction des modèles et des descripteurs, pour reconnaitre une séquence gestuelle de durée moyenne (3.39s). La dernière ligne (HMM) correspond aux temps de calcul optimisés, obtenus au chapitre précédent.

alloués à l'entrainement et à la validation. Ces variations permettent d'étudier l'influence de la taille de la base de données sur les performances du système de reconnaissance. Chaque type de descripteurs et de modèles sont à nouveau évalués. Les résultats généraux sont d'abord présentés. Puis, une attention particulière est portée sur les répartitions dans lesquelles l'effectif du sous-ensemble d'entrainement est le plus faible. En effet, ces conditions représentent mieux les cas d'usages concrets, dans lesquels peu d'échantillons sont disponibles pour l'entrainement du système comparé à la masse potentielle des mouvements que l'application devra reconnaitre durant son utilisation finale.

3.4.3.1 Cas général

Le tableau 3.2 expose les performances des GMM naïfs pour reconnaitre les 15 classes, lorsque seule la première moitié des mouvements est observée ($t_{50\%}$). Les taux de reconnaissance sont moyennés sur 10 tirages aléatoires pour chaque répartition de la base de données. Le tableau 3.3 expose les mêmes résultats pour la modélisation GMM à états. Ces résultats sont repris graphiquement dans la figure 3.10, à gauche pour $t_{50\%}$, à droite pour $t_{100\%}$.

Effectifs (%) entr./valid.	**Amorpho. (%)**	**Cart. (%)**	**Euler (%)**
10 / 90	61.9 ± 2.6	52.5 ± 3.9	47.7 ± 3.2
30 / 70	85.0 ± 1.9	84.0 ± 1.1	80.6 ± 2.3
50 / 50	91.8 ± 1.5	91.7 ± 1.4	88.6 ± 1.7
70 / 30	95.3 ± 1.1	95.4 ± 1.0	93.9 ± 1.3
90 / 10	97.2 ± 1.9	96.9 ± 1.4	96.5 ± 1.8

Table 3.2 - Moyennes des taux de reconnaissance pour des mouvement observés à 50% ($t_{50\%}$), en fonction de la répartition entrainement/validation des échantillons, pour une modélisation GMM naïfs.

Effectifs (%) entr./valid.	Amorpho. (%)	Cart. (%)	Euler (%)
10 / 90	68.0 ± 3.5	63.5 ± 4.0	57.2 ± 4.6
30 / 70	87.2 ± 2.0	86.9 ± 1.6	81.3 ± 1.7
50 / 50	92.9 ± 1.3	92.6 ± 1.8	89.1 ± 1.5
70 / 30	94.6 ± 1.4	94.8 ± 1.3	92.6 ± 1.5
90 / 10	96.7 ± 1.5	96.8 ± 1.4	94.7 ± 1.5

Table 3.3 - Idem tableau 3.2, pour une modélisation GMM à états.

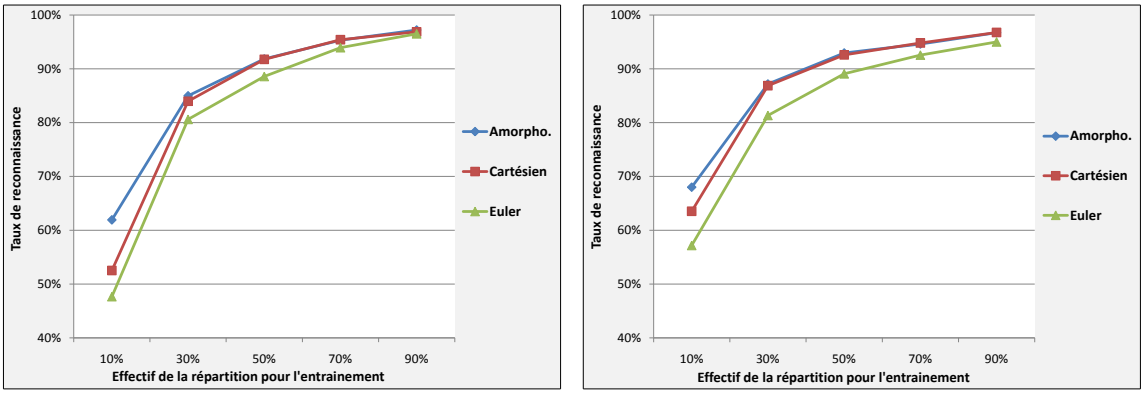


Figure 3.10 - Moyennes des taux de reconnaissance pour des mouvement observés à 50% ($t_{50\%}$) en fonction de la répartition des échantillons entre l'entrainement et la validation, pour une modélisation GMM naïf (à gauche) et GMM à états (à droite).

Pour les répartitions où l'effectif d'entrainement est supérieur à 50% de la base de données, des ANOVA par rang de Friedman ne montrent aucune différence significative entre les résultats des GMM naïfs et des GMM à états. De même, il n'existe aucune différence significative entre les résultats des différents types de descripteurs.

En revanche, comme le suggère la figure 3.10, les répartitions, où l'effectif d'entrainement est inférieur à 50% de la base de données, présentent des différences de performances notables entre les différents modèles et les types de descripteurs. Les figures 3.11 et 3.12 fournissent, pour chaque répartition et chaque fraction de mouvement, les différences moyennes (en %) entre les taux de reconnaissance obtenus par les GMM à états et par les GMM naïfs (les différences positives sont en faveur des GMM à états). Que ce soit pour les descripteurs amorphologiques (fig. 3.11) ou Cartésiens (fig. 3.12), qui présentent les meilleurs taux de reconnaissance, ces figures mettent en évidence la supériorité de la performance des GMM à états devant les GMM naïfs, à mesure que l'effectif de la base d'entrainement diminue (c.-à-d. en se déplaçant vers le haut du tableau). Elles rendent également compte de la stabilité de ces différences tout au long du plateau asymptotique $t_{50\%}$ $t_{100\%}$, sur lequel tous les résultats semblent définitivement stabilisés.

De la même manière, les descripteurs amorphologiques obtiennent de meilleures performances que les descripteurs Cartésiens uniquement pour la répartition 10/90, c.-à-d. lorsque l'effec-

	$t_{10\%}$	$t_{20\%}$	$t_{30\%}$	$t_{40\%}$	$t_{50\%}$	$t_{60\%}$	$t_{70\%}$	$t_{80\%}$	$t_{90\%}$	$t_{100\%}$
10/90	1	2	3	5	6	6	6	7	6	6
30/70	1	0	1	1	2	2	2	2	2	2
50/50	1	-1	-0	0	0	1	1	1	1	1
70/30	2	-1	-1	-1	-1	-1	-0	-0	-0	-0
90/10	4	-3	-2	0	-0	-0	-1	-1	-0	-0

Figure 3.11 - Différences (en %) entre les taux de reconnaissance des GMM à état et des GMM naïfs pour les descripteurs amorphologiques. Les répartitions de la base de données apparraissent en lignes, les fractions de mouvements en colonnes.

	$t_{10\%}$	$t_{20\%}$	$t_{30\%}$	$t_{40\%}$	$t_{50\%}$	$t_{60\%}$	$t_{70\%}$	$t_{80\%}$	$t_{90\%}$	$t_{100\%}$
10/90	1	4	7	9	10	11	11	11	11	11
30/70	0	-1	1	3	3	3	3	3	3	3
50/50	-1	-2	-1	-0	0	1	1	1	1	1
70/30	-2	-3	-2	-2	-1	-1	-1	0	-0	-0
90/10	-2	-1	-1	-1	-0	-0	-0	0	0	-0

Figure 3.12 - Différences (en %) entre les taux de reconnaissance des GMM à état et des GMM naïfs pour les descripteurs Cartésiens.

tif d'entrainement est réduit à 10% de la base de données. La figure 3.13 met en relief ces différences pour les GMM à états.

	$t_{10\%}$	$t_{20\%}$	$t_{30\%}$	$t_{40\%}$	$t_{50\%}$	$t_{60\%}$	$t_{70\%}$	$t_{80\%}$	$t_{90\%}$	$t_{100\%}$
10/90	-0	1	3	5	5	5	5	5	5	5
30/70	-3	-2	-0	-0	0	0	0	0	1	1
50/50	-2	-2	0	0	0	0	0	0	-0	0
70/30	-2	-3	-1	0	-0	-0	-0	-1	-0	-0
90/10	-1	-6	-3	0	-0	-0	-0	-1	-1	0

Figure 3.13 - Différences (en %) entre les taux de reconnaissance obtenus pour les descripteurs amorphologiques et Cartésiens avec une modélisation GMM à états. Les répartitions de la base de données apparraissent en lignes, les fractions de mouvements en colonnes.

Tous ces constats sur les performances des différents modèles et descripteurs pour les faibles effectifs d'entrainement sont discutés plus exhaustivement dans la section suivante.

3.4.3.2 Focus sur les faibles effectifs d'entrainement

Pour les répartitions où le sous-ensemble d'entrainement contient seulement 10% des échantillons de la base de données, l'évolution du taux de reconnaissance en fonction de la fraction de mouvement observée par le système, présente toujours le même type de plateau asymptotique entre $t_{50\%}$ et $t_{100\%}$. Quels que soient les descripteurs, la valeur atteinte par le taux est cependant bien inférieure aux résultats pour les répartitions comptant une base d'entrainement plus importante (tab. 3.2 et 3.3).

Pour cette répartition 10/90, les résultats montrent que les descripteurs amorphologiques obtiennent systématiquement des taux de reconnaissance significativement meilleurs que les descripteurs Cartésiens et angulaires. Ce constat renforce la conclusion de l'étude du chapitre 2, selon laquelle les descripteurs amorphologiques sont plus propices à encoder la variabilité morphologique dans les mouvements naturels.

De plus, les GMM à états sont significativement plus performants que les GMM naïfs. Ce constat est valable quels que soient les descripteurs choisis, même si les descripteurs amorphologiques sont ceux qui profitent le moins de cette modélisation (+6.1% contre +11% pour les Cartésiens et +9.5% pour les angulaires). On constate de fortes disparités inter-classes dans l'amélioration des performances proposée par les GMM à états. Certaines classes sont reconnues de manière très supérieure : +25% pour *gratter menton*, +15% pour *salut avec la main au niveau de la tête* ou +10% pour *applaudir*, quand les autres stagnent ou gagnent seulement quelques points (les présents taux de reconnaissance sont donnés pour les descripteurs amorphologiques). Aucune classe n'est toutefois moins bien reconnue avec les GMM à états.

Cependant le taux de 68% de reconnaissance obtenu par les GMM à états avec les descripteurs amorphologiques est significativement inférieur à celui obtenu avec les HMM du chapitre 2, qui atteignait 79%. La prise en compte de la temporalité du mouvement semble donc fondamentale pour encoder la variabilité du mouvement lorsque le nombre d'échantillons disponibles pour entrainer le système est faible. L'introduction des pondérations temporelles des états n'apporte toutefois pas d'amélioration notable, et ce, quel que soit le type de descripteurs utilisés comme le présente le tableau 3.4 et comme le confirme une ANOVA par rang de Friedman.

Modélisation	**Amorpho. (%)**	**Cart. (%)**	**Euler (%)**
HMM	79.0 ± 1.2	65.7 ± 2.6	57.5 ± 3.0
GMM naïfs	61.9 ± 2.6	52.5 ± 3.9	47.7 ± 3.2
GMM à états	68.0 ± 3.5	63.5 ± 4.0	57.2 ± 4.6
GMM à états pondérés	67.6 ± 4.0	64.0 ± 3.7	57.5 ± 3.4

Table 3.4 - Taux de reconnaissance en fonction de la modélisation pour la répartition 10% entrainement / 90% validation (à $t_{50\%}$ pour les modélisations de la famille GMM).

Enfin, lorsque l'on passe à la répartition 30/70, les taux de reconnaissance des GMM à états restent significativement plus élevés que ceux des GMM naïfs, pour les descripteurs amorphologiques et Cartésiens, bien que les écarts soient nettement plus resserrés que pour la répartition 10/90. En revanche la différence n'est pas significative pour les descripteurs angulaires.

3.4.4 Résumé

Les résultats présentés dans cette sous-étude 1 montrent qu'il est possible de reconnaitre correctement un mouvement, parmi les 15 classes de notre base de données, à partir du moment où

le système de reconnaissance a pu observer la première moitié de l'exécution de ce mouvement. L'observation de la seconde moitié n'apporte pas d'amélioration notable des performances du système.

Pour une répartition aléatoire 50% entrainement / 50% validation de la base de données, les systèmes fondés sur des GMM naïfs ou des GMM à états, pondérés ou non, offrent des performances équivalentes, de l'ordre de 92% de reconnaissance à partir de descripteurs amorphologiques ou Cartésiens. L'utilisation de descripteurs angulaires détériore significativement les taux de reconnaissance de l'ordre de 2 à 3%. Ces résultats sont conformes à ceux obtenus avec les HMM du chapitre 2.

Le taux de reconnaissance chute quand l'effectif du sous-ensemble d'entrainement se réduit. Pour une répartition entrainement/validation de 10/90, on constate que les GMM à états présentent un taux de reconnaissance de 68% (67.6% avec pondérations temporelles), alors qu'il descend à 61.9% pour les GMM naïfs. La pondération temporelle n'apporte pas d'amélioration significative. Dans les mêmes conditions, la modélisation HMM maintenait un taux de reconnaissance de l'ordre de 79%. Bien qu'on ait cherché à conserver une partie de l'information temporelle en gardant les états, la disparition de la nature purement séquentielle de la modélisation semble avoir un effet dramatique sur les systèmes utilisant les descripteurs amorphologiques (-11%).

En outre, comme pour l'étude du chapitre 2, il semble que l'influence des descripteurs sur la méthodologie de reconnaissance ne se ressente que pour de faibles effectifs de la base de données d'entrainement (10/90). Les descripteurs amorphologiques possèdent alors les meilleures performances.

3.5 Sous-étude 2 : Répartition par sujet

3.5.1 Méthode d'évaluation

Comme nous l'avons déjà précisé, un système de reconnaissance de mouvements ne peut pas, en pratique, être entrainé avec l'ensemble de la population. Pour tester le comportement du système face à un utilisateur nouveau, nous proposons, dans cette sous-étude 2, d'évaluer le système de reconnaissance en isolant un des 10 sujets dont l'intégralité des mouvements est utilisée pour la validation. Ce type de répartition de la base de données est appelé *Leave-One-Out* (L1O). Tous les sujets, à tour de rôle, sont évalués de la sorte.

Dans un second temps, nous étendons l'expérimentation en isolant 2 sujets, puis 3, etc... jusqu'à isoler 9 sujets, ce qui revient à entrainer le système avec un seul sujet. Par extension, on appelle ce type de répartition *Leave-k-Out* (LkO), on sort k sujets du sous-ensemble d'entrainement. En pratique, il serait très fastidieux d'obtenir la totalité des combinaisons qu'une telle évaluation suppose (252 combinaisons pour L5O). Pour réduire ce nombre à un sous ensemble raisonnablement calculable de 10 combinaisons, nous recourons à une méthode d'échantillonnage pseudo-aléatoire, que nous avons déjà introduite en section 2.5.2. Cette méthode est sensée mieux représenter le comportement général qu'un échantillonnage purement aléatoire, et permet, de plus, de pouvoir comparer les résultats des LkO successifs.

À l'exception de ces modifications dans la répartition de la base de données, la méthode d'évaluation des résultats est identique à celle de la sous-étude 1 (3.4.1). Les modèles GMM naïfs, GMM à états et GMM à états pondérés sont évalués pour chaque répartition, en condition de reconnaissance précoce, c.-à-d. pour des mouvements observés partiellement (de $t_{10\%}$ à $t_{100\%}$).

3.5.2 Résultats pour la répartition Leave-One-Out

La figure 3.14 présente les performances moyennes de reconnaissance pour les modélisations GMM naïfs et GMM à états en fonction de la fraction de début de mouvement observée. Comme pour l'échantillonnage aléatoire de la section précédente, les taux de reconnaissance de chaque modélisation atteignent un plateau à partir de $t_{50\%}$, démontrant qu'une reconnaissance précoce est non seulement envisageable, mais qu'en plus le fait d'observer l'intégralité du mouvement n'apporte pas d'amélioration aux performances de chaque modélisation.

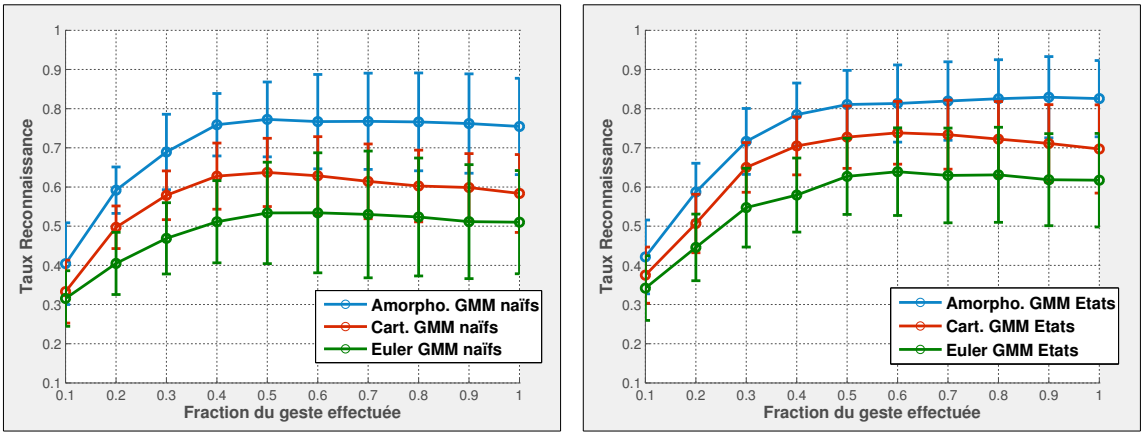


Figure 3.14 - Taux de reconnaissance moyen en fonction de la fraction observée en début de mouvement. À gauche pour un modèle GMM naïf, à droite pour un modèle GMM à états. Les types de descripteurs utilisés apparaissent dans différentes couleurs.

Pour chaque type de descripteurs, une ANOVA par rang de Friedman suivie d'un test posthoc des rangs signés de Wilcoxon démontrent que les GMM à états proposent des performances de reconnaissance significativement meilleures que les GMM naïfs, à partir du moment où au moins la première moitié du mouvement est observée ($t_{50\%}$).

De même, dès lors qu'au moins la moitié du mouvement est observée, l'influence des descripteurs utilisés en entrée du système sur la performance est statistiquement attestée par une ANOVA par rang de Friedman ($\chi^2(2, N = 150) = 24.005$, $p < 0.0001$, à $t_{50\%}$). Les descripteurs amorphologiques présentent un taux de reconnaissance de 81.1% ($\pm$8.7), significativement plus élevé que les 72.7% ($\pm$8.0) des descripteurs Cartésiens (test post hoc des rangs signés de Wilcoxon pour 50% d'exécution : $N = 150, T = 0.443, p < 0.05$) et les 62.7% ($\pm$9.7) des descripteurs angulaires ($N = 150, T = 0.318, p < 0.05$).

L'écart-type important dans les taux de reconnaissance souligne la disparité des performances en fonction des sujets utilisés pour valider le système. La figure 3.15 présente les résultats par sujet pour la modélisation GMM à état en fonction des descripteurs choisis. La fraction de mouvement observée est $t_{50\%}$. Cette figure confirme la meilleure tenue des descripteurs amorphologiques.

La dispersion des résultats ne peut s'expliquer par la morphologie des sujets. Le sujet Suj_D est le mieux reconnu (94%), alors qu'il est pourtant le plus grand (186 cm) et que la taille moyenne des sujets de l'ensemble d'entrainement tombe à 169 $\pm$ 10 cm dans ce cas. De même, Suj_A est le plus petit (154 cm) et ses mouvements ne sont pas moins bien reconnus que la moyenne, alors que la taille moyenne des sujets de l'ensemble d'entrainement monte alors à 173$\pm$ 10. À l'inverse, les mouvements de Suj_G sont les moins bien reconnus (64%) alors qu'il est de même taille et même sexe que Suj_H, dont les mouvements sont correctement reconnus (85%). Ces résultats soulignent l'importance du style dans la variabilité gestuelle.

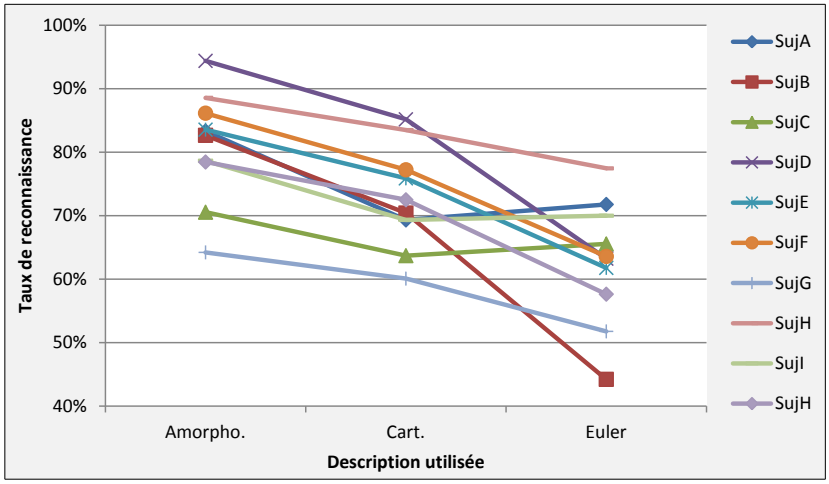


Figure 3.15 - Évolution du taux de reconnaissance par sujet en fonction des descripteurs utilisés. La modélisation est un GMM à états et les échantillons sont observés à $t_{50\%}$. Les descripteurs amorphologiques sont systématiquement meilleurs, confirmant ainsi l'étude précédente (fig.2.13).

Pour simplifier la suite de l'exposé, nous présentons principalement les résultats pour les GMM à états et les descripteurs amorphologiques, dont on vient de montrer qu'ils possèdent statistiquement les meilleures performances de reconnaissance. L'introduction d'une information temporelle explicite dans le modèle GMM, sous la forme de pondérations temporelles des états n'apporte pas d'amélioration notable, comme le montre la figure 3.16. Une ANOVA par rang de Friedman indique néanmoins que les différences entre les 3 modélisations ne sont jamais statistiquement significatives, quelle que soit la fraction de mouvement observée au-dessus de $t_{50\%}$ ($\chi^2(2, N = 150) = 0.974$, $p = 0.614$, à $t_{60\%}$ par exemple).

L'influence des pondérations temporelles sur la performance des GMM à états apparait très minime au regard de la figure 3.16. Pourtant, ces pondérations temporelles ont un rôle effectif sur la vraisemblance des états. L'indicateur, proposé en section 3.3.4, en atteste : sans pondération temporelle, 16% ($\pm$3) des vecteurs descripteurs observés $\mathbf{o}(t)$ sont attribuées à un état temporellement incompatible avec la phase du mouvement dans laquelle cet état est attendu. Cette proportion d'incompatibilité tombe à 3% ($\pm$3) lorsqu'on ajoute les pondérations temporelles.

La figure 3.17 reprend les résultats de la figure 3.16 en détaillant individuellement les taux de reconnaissance par mouvement. Si le plateau asymptotique du taux de reconnaissance apparait bien pour tous les mouvements, son instant d'apparition est toutefois plus dispersé que ce que la figure 3.16 ne le laisse présager.

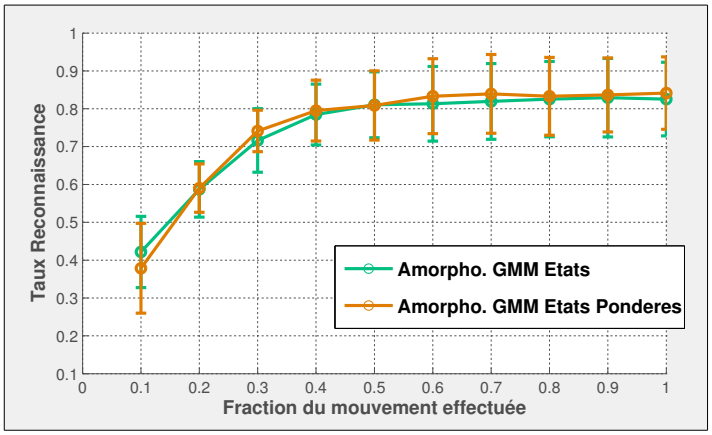


Figure 3.16 - Influence de la pondération temporelle sur le taux de reconnaissance en fonction de la fraction de mouvement observée.

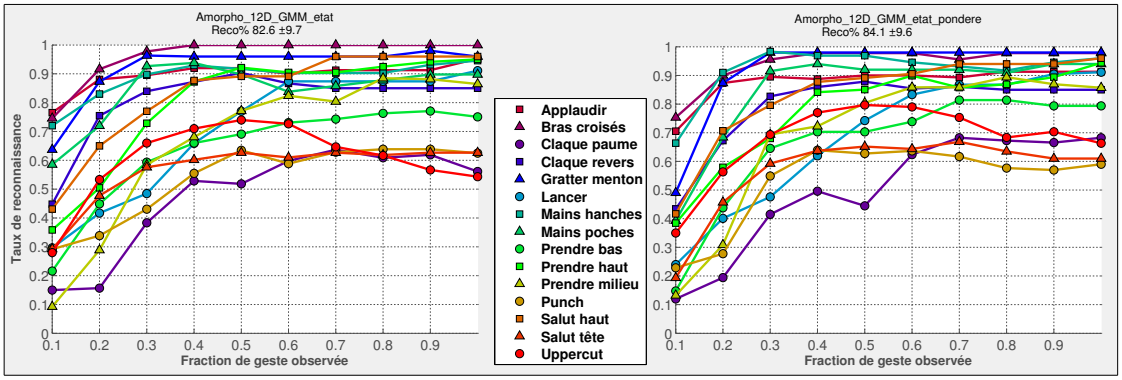


Figure 3.17 - Taux de reconnaissance par mouvement individuel en fonction de la fraction de mouvement observée à partir des descripteurs amorphologiques. À gauche, pour la modélisation GMM à états, à droite pour la modélisation GMM à états pondérés temporellement.

3.5.3 Répartition Leave-k-Out

Dans cette section, l'approche *Leave-One-Out* est étendue à une approche *Leave-k-Out* (LkO) en isolant, 2 sujets pour le sous-ensemble de validation (L2O), puis 3 (L3O), etc... Le reste des sujets est utilisé pour entrainer le système. Ces répartitions permettent d'évaluer le comportement du système de reconnaissance face à de nouveaux utilisateurs (les k sujets de validation), à mesure que l'on réduit le nombre de sujets dans le sous-ensemble d'entrainement.

Les taux de reconnaissance moyens pour les LkO successifs, sur des mouvements observés entièrement ($t_{100\%}$), sont présentés sur la figure 3.18, pour les GMM naïfs (à gauche) et pour les GMM à états (à droite). Quels que soient les modèles, les descripteurs amorphologiques se

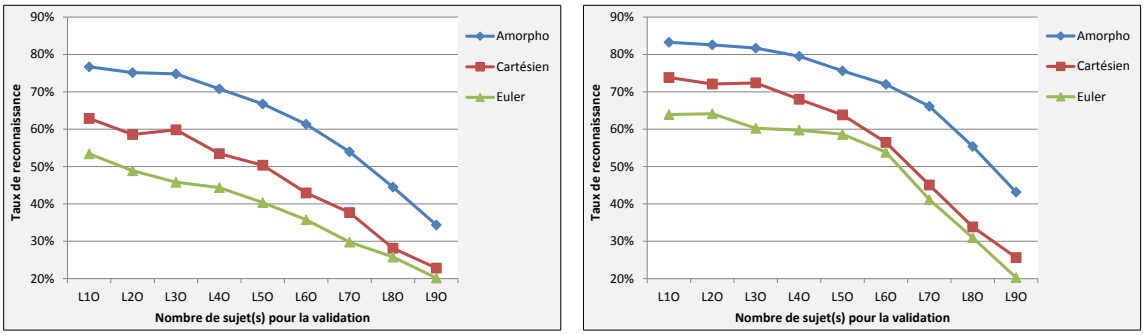


Figure 3.18 - Évolution du taux de reconnaissance en fonction du nombre de sujets réservés à l'évaluation, pour les GMM naïfs (à gauche) et pour les GMM à états (à droite), sur des mouvements entièrement observés ($t_{100\%}$). Plus on avance vers la droite du graphique, plus grande est la proportion de sujets appartenant uniquement au sous-ensemble de validation.

comportent mieux que les descripteurs Cartésiens et angulaires. Pour chaque LkO successif, une ANOVA par rang de Friedman suivie d'un test posthoc de Wilcoxon confirment que, à modélisation fixée, le taux de reconnaissance des descripteurs amorphologiques est significativement supérieur aux autres à $t_{100\%}$.

De même, à type de descripteurs fixé, les GMM à états proposent des performances de reconnaissance significativement supérieures aux GMM naïfs, pour tous les LkO à $t_{100\%}$. En revanche, la pondération temporelle des états ne permet d'améliorer le taux de reconnaissance que de 1 à 2% au maximum par rapport aux GMM à états, sans que ces améliorations ne soient significatives.

Dans le cadre de la reconnaissance précoce, les tests statistiques démontrent qu'à partir d'une fraction de mouvement observée supérieure à $t_{30\%}$, les GMM à états avec et sans pondérations temporelles obtiennent des taux significativement meilleurs que les GMM naïfs. Ce résultat est illustré par la figure 3.19, qui met en relief les différences de taux de reconnaissance entre les GMM à états et les GMM naïfs, en utilisant les descripteurs amorphologiques (qui fournissent les meilleurs résultats).

De même, les descripteurs amorphologiques mènent à un taux de reconnaissance significativement meilleur que les deux autres descripteurs à partir de $t_{30\%}$, pour tous les LkO. La figure 3.20 permet de s'en convaincre. Elle recense les différences des taux de reconnaissance obtenus par les descripteurs amorphologiques et Cartésiens pour les GMM à états. Notons que la dernière colonne correspond à la distance euclidienne entre chaque point des courbes « Amorpho. » et

	$t_{10\%}$	$t_{20\%}$	$t_{30\%}$	$t_{40\%}$	$t_{50\%}$	$t_{60\%}$	$t_{70\%}$	$t_{80\%}$	$t_{90\%}$	$t_{100\%}$
L1O	2	0	3	4	5	7	6	6	7	7
L2O	1	3	6	6	7	7	8	8	8	9
L3O	1	1	6	6	7	7	7	7	8	8
L4O	2	2	5	6	8	9	9	9	9	10
L5O	2	3	8	10	9	9	9	9	9	10
L6O	1	2	8	10	11	11	10	10	10	11
L7O	2	2	8	11	12	12	12	12	12	13
L8O	1	2	7	8	11	11	11	11	12	13
L9O	0	1	6	7	9	9	9	10	10	11

Figure 3.19 - Différences (en %) entre les taux de reconnaissance des GMM à états et des GMM naïfs pour les descripteurs amorphologiques (les valeurs positives dénotent une supériorité des GMM à états, et inversement). Les différentes répartitions de la base de données (LkO) apparraissent en lignes, les fractions de mouvements observées en colonnes.

	$t_{10\%}$	$t_{20\%}$	$t_{30\%}$	$t_{40\%}$	$t_{50\%}$	$t_{60\%}$	$t_{70\%}$	$t_{80\%}$	$t_{90\%}$	$t_{100\%}$
L1O	-3	9	8	9	10	10	10	10	12	13
L2O	1	9	12	10	11	10	11	11	12	13
L3O	-1	10	10	8	9	9	10	11	11	12
L4O	0	9	12	11	12	12	11	11	13	13
L5O	2	10	13	14	12	12	13	14	14	15
L6O	3	11	16	17	15	16	14	16	17	19
L7O	5	14	19	22	22	21	21	21	22	23
L8O	3	11	17	20	22	22	21	21	21	23
L9O	3	9	14	17	18	18	18	18	18	19

Figure 3.20 - Différences (en %) entre les taux de reconnaissance obtenus pour les descripteurs amorphologiques et Cartésiens avec les GMM à états (les valeurs positives dénotent une supériorité des descripteurs amorphologiques, et inversement). Les différentes répartitions de la base de données (LkO) apparraissent en lignes, les fractions de mouvements observées en colonnes.

Cartésien tracées sur la figure 3.18, à droite.

Comme la sous-étude 1 semblait le montrer, les figures 3.19 et 3.20 confirment que les différences de taux de reconnaissance moyens, entre les différents modèles (fig. 3.19) et types de descripteurs (fig. 3.20), s'établissent précocement (vers $t_{30\%}$-$t_{40\%}$), puis sont stabilisées par la suite sur le plateau asymptotique ($t_{50\%}$ à $t_{100\%}$).

Pour conclure, regardons plus en détail les performances des GMM à états utilisant les descripteurs amorphologiques, dont nous venons de démontrer la supériorité.La figure 3.21 montre les taux de reconnaissance moyens en fonction de la fraction de mouvement observée par le système, pour les LkO successifs. Elle confirme très clairement la présence d'un plateau asymptotique à partir de $t_{50\%}$, voire dès $t_{40\%}$. La différence fondamentale entre ces courbes réside dans la hauteur du plateau asymptotique, qui atteint, par exemple un taux de reconnaissance de 81 à 82% pour L2O, alors qu'il parvient à environ 75% pour L5O.

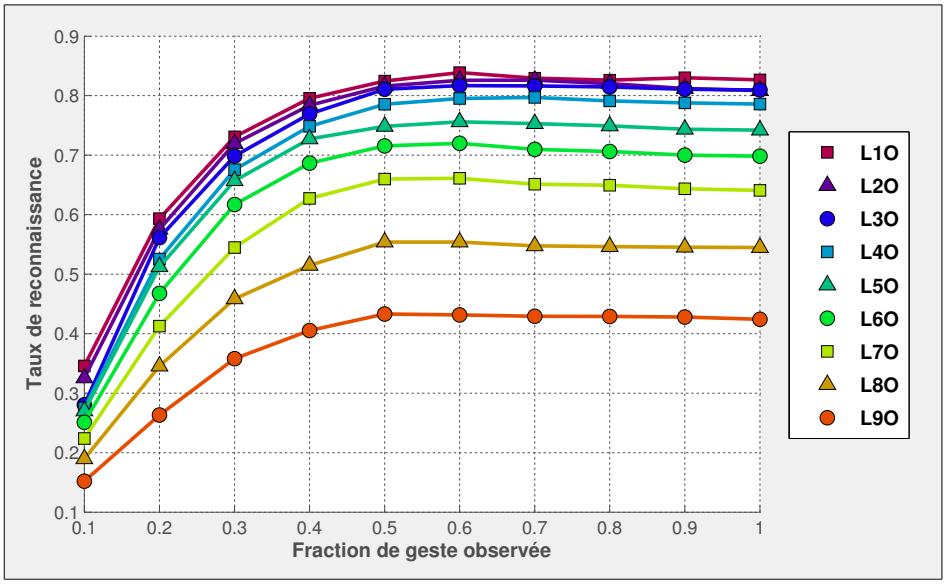


Figure 3.21 - Évolution du taux de reconnaissance en fonction de la fraction de mouvement observée pour les répartitions LkO successives.

Les taux de reconnaissance des LkO successifs restent très proches entre L1O et L4O. Les résultats de reconnaissance de chaque tirage pseudo-aléatoire LkO sont alors approximativement égaux à la moyenne des résultats L1O des sujets qui composent le tirage. Par exemple, le premier tirage L2O obtient un taux de reconnaissance de 81.3%. Or, ce tirage contient les sujets Suj_A et Suj_C, dont les taux de reconnaissance en L1O sont respectivement 92% et 73%, soit une moyenne de 81.5%, très proche du résultat obtenu pour le L2O associant ces 2 sujets. Ces résultats se vérifient avec une erreur de l'ordre de 3% sur l'ensemble des tirages pris individuellement entre L1O et L4O, quelle que soit la fraction de mouvement de $t_{50\%}$ à $t_{100\%}$. Cela tendrait à montrer que, à L4O (6 sujets pour l'entrainement, 4 pour la validation), la base de données d'entrainement n'a encore pas subi de perte majeure d'information. À partir de L5O, la réduction du nombre de sujets dans la base de données d'entrainement devient décisive et les taux de reconnaissance se dégradent de plus en plus. Le problème est alors double. D'abord, les GMM à états disposent de moins de variabilité pour s'entrainer. Les Gaussiennes se trouvent très centrées et resserrées, autour de données d'entrainement peu dispersées. Elles leur sont trop spécifiques, c'est le phénomène de sur-apprentissage. À l'inverse, la variabilité de la base

de données de validation est bien plus importante. La probabilité d'observer des données qui correspondent mal à un modèle auquel elles sont sensées appartenir augmente. La confusion augmente de paire.

3.6 Conclusion

La première contribution de cette étude est de montrer qu'il est possible de reconnaitre un mouvement de façon précoce, à partir du moment où le système a pu observer la première moitié de son exécution. Les 3 types de modèles proposés (GMM, GMM à états et GMM à états pondérés) ont été testés avec les descripteurs amorphologiques, Cartésiens et angulaires. Toutes ces associations modèles/descripteurs ont montré un plateau asymptotique du taux de reconnaissance à partir de l'instant où la première moitié du mouvement a été observée par le système. Après cet instant, les résultats sont définitivement stabilisés jusqu'à la fin du mouvement.

Malgré des algorithmes non optimisés et l'utilisation de Matlab, les temps de calculs sont compatibles avec une utilisation en temps réel, puisque, dans le pire des cas, moins de 6% de la durée du mouvement suffisent au processus de reconnaissance pour établir sa décision.

En outre, cette étude montre que l'appréciation de la performance d'une méthode de reconnaissance de mouvements ne peut pas être réalisée uniquement à la lumière d'une répartition aléatoire 50/50 de la base de données, comme c'est souvent le cas dans la littérature. Les différents modèles, mis en œuvre dans le cadre de cette étude, démontrent clairement qu'il n'existe pas de différence notable entre les performances des GMM naïfs et des GMM à états pour une répartition de la base de données 50% entrainement / 50% validation. Pourtant, cette différence devient statistiquement significative, dès lors que la part de la base de données allouée à l'entrainement se réduit, ou qu'on se place dans un cadre L1O, voire LkO. Les performances de reconnaissance sont alors meilleures pour la modélisation GMM à états. De la même manière, les descripteurs utilisés ont peu d'influence sur les performances du système de reconnaissance en 50/50. Or, les descripteurs amorphologiques deviennent significativement plus performants que les descripteurs Cartésiens et angulaires, dès lors que les sujets, dont les mouvements constituent la base de données de validation, ne se trouvent pas dans la base de données d'entrainement (conditions LkO).

Les pondérations temporelles, telles qu'elles ont été mises en œuvre dans cette étude, n'ont pas montré d'amélioration significative de la performance de reconnaissance globale. Elles permettent néanmoins de rattraper quelques classes de mouvement moins bien reconnues en L1O. Celles-ci peuvent gagner de 4 à 10%. Or dans un cadre interactif, ce gain peut s'avérer décisif car il est alors fondamental que l'ensemble des classes de mouvements soient correctement reconnues sans exception. Dans le cas du karaté, par exemple, si un coup de poing ne peut pas être différencié d'une parade, c'est l'application toute entière qui devient inopérante. Toutefois, si les pondérations temporelles permettent de rattraper quelques mouvements, elles posent à nouveau un problème d'incompatibilité avec la reconnaissance précoce. Pour chaque mouvement à reconnaitre, il est nécessaire d'adapter les pondérations temporelles à la durée de ce mouvement, ce qui suppose d'en connaître au moins la durée totale à l'avance.

En outre, les résultats de cette étude suggèrent des optimisations et offrent quelques perspec-

tives de travail.

Une première piste d'optimisation apparait au regard de la figure 3.6, qui présente les instants d'activation privilégiés de chaque état. On constate que les états S_1 et S_7 se trouvent en début et fin de mouvement de façon assez brève, et ce, sur toutes les classes de mouvement. Comme le suggère [Marr1982], il s'agit d'états de repos, qui précèdent et suivent la phase « active » du mouvement. Une mutualisation de ces états pour toutes les classes permettrait de gagner un précieux temps de calcul, sans pour autant pénaliser la performance de la modélisation GMM à états. En effet, ces états de repos sont en principe peu spécifiques de la classe de mouvement à laquelle ils appartiennent.

Ensuite, les méthodes de modélisation des classes de mouvement soulèvent plusieurs perspectives d'amélioration. Malgré l'introduction d'états et de pondérations sur ces états, les GMM n'atteignent pas le niveau de performance des HMM pour les effectifs d'entrainement contenant très peu de sujets (LkO avec k > 6). Pour combiner la puissance des HMM avec la reconnaissance précoce, une modélisation des classes réduite à la première moitié du mouvement pourrait être envisagée, comme le suggère [Axenbeck2008]. D'autre part, l'influence positive très nette des HMM sur les résultats en LkO suggère d'étudier le comportement d'autres types de modèles. En particulier, des approches discriminantes, tels que les SVM ([Laptev2007]), permettraient certainement de minimiser le taux d'erreurs entre les classes les plus fréquemment confondues.

Pour être utilisable en environnement virtuel, le système de reconnaissance doit maintenant être capable de segmenter temporellement le flux de données entrant. La présente étude montre qu'il est possible de reconnaitre précocement un mouvement en se concentrant sur son début. En prolongeant cette approche, nous pourrions isoler, pour chaque classe de mouvements, la sous-séquence qui lui est le plus caractéristique. La détermination de ces sous-séquences, généralement appelées primitives, est un champ de recherche très actif en reconnaissance de mouvement [Reng2006, Ogawara2009]. La recherche de ces primitives, dans le flux de données, permettrait de guider efficacement la segmentation. Il s'agit là d'une des prochaines étapes dans la quête menant à un système de reconnaissance de mouvements complètement temps réel.

Un problème fondamental des systèmes de reconnaissance utilisant des modélisations probabilistes, telles que les GMM ou les HMM, réside dans leurs difficultés à fournir des indicateurs de confiance sur leurs prédictions. La reconnaissance se base sur la vraisemblance cumulée d'une séquence inconnue, relativement à différents modèles de classe. Mais la vraisemblance n'est pas une métrique absolue interprétable en terme de ressemblance entre la séquence et le modèle. Or, une telle métrique peut s'avérer précieuse si le mouvement à reconnaitre est trop mal réalisé, ou, pire, s'il ne fait partie d'aucune classe connue par le système, ce qui peut arriver en cas de segmentation temporelle erronée. Fournir un indicateur de confiance sur la reconnaissance apparait donc crucial. Cette problématique trouve encore assez peu d'échos dans le domaine de la reconnaissance de mouvement, même si [Yang2006] propose une première approche.

Enfin, la reconnaissance des mouvements naturels, propices à une grande variabilité, a permis de démontrer la supériorité de l'approche amorphologique. La base de données de mouvement doit maintenant être étendue, notamment pour y intégrer des mouvements sportifs, très dynamiques, et des mouvements paramétriques (pointer, dimensionner...), qui, en plus de leur sémantique, véhiculent un argument supplémentaire (la direction pointée, la taille de l'objet dimensionné...), qui accroit la variabilité.

Conclusion et perspectives

Cette thèse s'inscrit dans le cadre du projet industriel Biofeedback, qui vise à concevoir un outil interactif et immersif dédié à l'apprentissage de tâches motrices complexes. Un tel système trouve des débouchés dans la formation de professionnels à des mouvements métier, ou, dans le cadre sportif, à l'apprentissage de mouvements techniques. La mise en œuvre de ce système soulève toutefois plusieurs problèmes scientifiques. L'un d'entre eux concerne l'exploitation des données issues des capteurs de mouvements, afin de reconnaitre l'action de l'utilisateur et d'en évaluer la performance.

Dans cette thèse, nous avons spécifiquement abordé la problématique de la reconnaissance de mouvements, à partir de données 3D. La densité de la littérature scientifique atteste d'un engouement certain pour ce domaine de recherche, depuis une quinzaine d'années. Le problème central réside dans la variabilité du mouvement, qui découle de l'étendue des degrés de liberté dont dispose le corps humain pour exécuter une tâche motrice. Les applications en vision par ordinateur, en interfaces homme-machine ou en animation graphique ont participé à l'émergence d'une importante diversité de méthodes de reconnaissance, qui permettent de prendre en compte la variabilité dans l'exécution du mouvement. La majorité de ces méthodes dérivent de modèles Markoviens (HMM), très bien adaptés à la modélisation de la variabilité spatiotemporelle. Cependant, comme le souligne [Turaga2008], peu d'études ont explicitement abordé la question de la variabilité morphologique inter-individuelle. Or, encoder celle-ci dès l'étape de description du mouvement, permettrait à la modélisation d'être déchargée de la part de variabilité associée à la morphologie.

Dans la première étude, nous avons proposé de nouveaux descripteurs du mouvement, nommés *amorphologiques*, qui se veulent le plus indépendants possible de la morphologie du sujet capturé. En les confrontant avec des descripteurs classiques (Cartésiens ou angulaires), nous avons clairement démontré qu'ils amélioraient sensiblement les performances d'un système de reconnaissance s'appuyant sur des modèles HMM. Ce résultat est encore plus net dans le cas où la morphologie du sujet n'était pas connue du système. Les mouvements sélectionnés se voulaient naturels, donc empreints de variabilité, et issus de 15 classes possédant des propriétés spatiotemporelles similaires, donc propices à la confusion. Mieux, l'utilisation des descripteurs

amorphologiques permet de maintenir un taux de reconnaissance élevé, même lorsque très peu de sujets ont été observés par le système durant sa phase de calibration. Cette méthodologie d'évaluation met en évidence la nette supériorité des descripteurs amorphologiques à encoder intrinsèquement la variabilité morphologique inter-individuelle dans les systèmes de reconnaissance de mouvements.

Par ailleurs, les algorithmes développés lors de cette première étude ont été optimisés pour mettre en œuvre un démonstrateur, dans lequel les mouvements d'un utilisateur, inconnu du système, sont reconnus en temps interactif. Notons, toutefois, deux limitations à notre approche. D'abord, la segmentation temporelle des mouvements est ici éludée en demandant à l'utilisateur de revenir en position de repos entre chaque réalisation. Ce type de segmentation, bien qu'efficace techniquement, se révèle contraignant dans un cadre d'immersion naturelle en environnement virtuel, puisqu'il impose à l'utilisateur un retour dans la posture de repos entre chaque mouvement. La seconde limitation tient à la nuance existant entre temps interactif et temps réel. En effet, même si les temps de calcul sont largement compatibles avec le temps réel, puisqu'inférieur d'un facteur 100 à la durée du mouvement, les modèles HMM ne peuvent fournir une décision qu'après avoir observé le mouvement dans son intégralité, c.-à-d. en temps légèrement différé. Cette nuance a son importance pour une véritable utilisation interactive, dans laquelle l'environnement virtuel doit proposer une réponse adaptée au mouvement de l'utilisateur au plus vite, c.-à-d. avant même que le mouvement soit complètement terminé.

À partir de ce constat, la seconde étude a abordé la problématique de la reconnaissance précoce du mouvement. Pour cela, nous avons proposé et évalué trois méthodes fondées sur des modèles de mélange Gaussien (noté GMM) :

- les GMM, reprenant le formalisme classique ;
- les GMM à états, qui sont une adaptation des GMM visant à conserver une partie de l'information séquentielle contenue dans les états des HMM de la première étude ;
- les GMM à états pondérés, qui sont une modification des GMM à états visant à introduire explicitement une pondération temporelle des états suivant la probabilité qu'ils interviennent à chaque instant dans le mouvement.

Les résultats de cette étude ont démontré que le taux de reconnaissance n'évoluait plus, dès lors que le système a observé la première moitié du mouvement. À partir de cet instant, la décision du système est stabilisée, quels que soient les modèles et les descripteurs adoptés. L'observation de la seconde moitié du mouvement n'apporte pas d'amélioration notable. En revanche, la valeur atteinte par le taux de reconnaissance dépend des modèles et des descripteurs choisis pour le système. Pour les faibles effectifs de la base de données d'entrainement, ou pour les bases de données de validation ne comportant que des nouveaux sujets (non utilisés pour l'entrainement), les descripteurs amorphologiques confirment à nouveau leur nette supériorité devant les descripteurs classiques. Dans ces mêmes conditions, les GMM à états présentent également des performances bien supérieures aux GMM. La pondération temporelle des états n'a en revanche pas démontré de franche amélioration ou détérioration globale des performances des GMM à états. En définitive, cette étude a démontré que les descripteurs amorphologiques sont robustes aux contraintes du temps réel et à des modèles alternatifs aux HMM, qui permettent une reconnaissance précoce du mouvement.

Outre ces conclusions spécifiques à chaque étude, les travaux menés au cours de cette thèse permettent de tirer plusieurs constats généraux et d'esquisser quelques perspectives à plus ou moins brève échéance.

Tout d'abord, la méthode d'évaluation classiquement utilisée dans la littérature et consistant à

répartir aléatoirement la base de données, à parts égales, entre un sous-ensemble d'entrainement et un sous-ensemble de validation, n'a pas permis de montrer de différence entre les performances des différents modèles et descripteurs. Or, nous avons démontrer que ces différences existent bien, dès lors que la répartition de la base de données entre l'entrainement et la validation tient compte de l'identité des sujets, ou que l'effectif de la base de données d'entrainement devient faible. Ces répartitions, qui permettent de rendre compte des performances d'un système de reconnaissance face à de nouveaux utilisateurs, devraient être évaluées plus systématiquement dans les études en reconnaissance de mouvements.

Les modèles utilisés soulèvent quelques points à explorer. Pour commencer, on note qu'à mesure que le nombre de sujets diminue dans la base de données d'entrainement et augmente dans celle de validation, les descripteurs amorphologiques permettent aux HMM de conserver une performance de reconnaissance correcte. Dans ces mêmes conditions, la performance des GMM à états finit par chuter abruptement (au delà de L5O). Deux causes complémentaires pourraient expliquer ce constat : l'encodage intrinsèque de la temporalité du mouvement par les HMM (qui améliore leur performance), et l'ajustement trop resserré (sur-apprentissage) des GMM à états sur des données d'entrainement en trop faible effectif (performance des GMM à états détériorée). Avec le taux de reconnaissance, nous ne disposons que d'un indicateur général. Seule une étude comparative de l'évolution des paramètres des deux types de modèles, en fonction du nombre de sujets dans l'effectif d'entrainement, permettrait de quantifier l'influence de chaque cause.

D'autre part, en démontrant qu'une reconnaissance précoce est possible à partir de la moitié du mouvement, la seconde étude a ouvert la perspective d'une utilisation des HMM. L'idée serait de modéliser uniquement la première moitié du mouvement, ce qui permettrait aux HMM de fournir une décision plus précoce.

Par ailleurs, le comportement des descripteurs amorphologiques en entrée d'autres types de modèles reste à évaluer. Des modèles discriminants, telles que les machines à vecteur support (SVM), permettraient de réduire les confusions les plus fréquentes entre des classes de mouvements similaires. Des modèles paramétriques sont également à tester. Ces modèles présentent le double intérêt de reconnaitre la classe d'un mouvement, tout en déterminant un attribut de celui-ci, comme la direction pointée, ou l'écartement des mains, lors de mouvements de pointage ou de dimensionnement.

Pour conclure sur la modélisation, les bons résultats obtenus avec les descripteurs amorphologiques en entrée des HMM, lorsque très peu d'échantillons d'entrainement sont disponibles (68% en L9O par exemple), suggèrent une perspective nouvelle. Ils permettent d'envisager des méthodes incrémentales d'entrainement du système de reconnaissance. Ces méthodes sont connues sous le nom d'*apprentissage incrémental*. L'objectif de ce type de méthode est de proposer des modèles de mouvement, à partir de peu de données, puis d'adapter automatiquement et continuellement ces modèles au profil de l'utilisateur final. Dans le cadre du projet Biofeedback, qui sous-tend ces travaux de thèse, cette perspective a des implications déterminantes. Dans le scénario d'utilisation type, l'utilisateur effectue les mouvements demandés par le système, qui les reconnait et les évalue par rapport à une base de données pré-enregistrée. Le système est figé après sa phase de conception. L'apprentissage incrémental autorise un scénario moins restrictif, dans lequel la base de données peut être enrichie de nouveaux mouvements par l'utilisateur final. Ce dernier pourrait être un entraineur sportif, produisant lui-même de nouveaux contenus qu'il intègrerait dans des séances personnalisées pour ses apprenants. Ce scénario évolutif suppose de disposer d'un système de reconnaissance capable de reconnaitre les mouvements des apprenants uniquement à partir des mouvements d'un seul sujet (l'entraineur sportif), donc avec très peu de données d'entrainement. Or, ces conditions correspondent exactement à l'évaluation L9O, pour laquelle les descripteurs amorphologiques ont montré une reconnaissance de 68%

des mouvements. D'autre part, la performance motrice de l'utilisateur apprenant change avec l'entrainement, son style se précise. Des méthodes d'apprentissage incrémental permettraient de suivre l'évolution de l'utilisateur apprenant au fur et à mesure de ses progrès.

A plus brève échéance, la robustesse de l'encodage de la variabilité inter-individuelle par les descripteurs amorphologiques doit être évaluée selon plusieurs modalités. D'abord les descripteurs amorphologiques doivent être étendus au corps entier pour inclure les jambes et la tête. D'autres types de mouvements doivent venir compléter la base de données. Les mouvements sportifs, en particulier, sont de toute première importance dans le cadre de l'apprentissage de performances motrices. Les descripteurs amorphologiques doivent démontrer leur capacité à encoder la variabilité morphologique sur ce type de mouvements, où la vitesse d'exécution peut devenir importante et où l'absence de données sur l'articulation du coude peut éventuellement poser problème. Ensuite, les capacités d'encodage de la variabilité inter-individuelle par les descripteurs amorphologiques doivent être validées sur une population plus importante et plus diversifiée. La morphologie est une source de variabilité certaine, mais le style l'est tout autant. Observer le comportement du système de reconnaissance sur des mouvements d'enfant, de personnes âgées ou encore de sujets de corpulences variées, apporterait un éclairage plus fin sur les aptitudes des descripteurs amorphologiques à encoder la part de variabilité associée au style.

De même, la robustesse des descripteurs amorphologiques doit être évaluée sur des captures de mouvements 3D moins précises et plus bruitées, telles que celles proposées par les systèmes grand public récemment apparus (Microsoft Kinect, SoftKinetic Iisu). Dans le cadre du projet Biofeedback, obtenir un système de reconnaissance, qui soit robuste à ce type de capture, permet d'envisager un déploiement à grande échelle de l'application. La faible dimensionnalité des descripteurs amorphologiques plaide en leur faveur, mais les descripteurs associés à la vitesse du mouvement risquent de se trouver très impactés par le bruit.

Par ailleurs, l'utilisabilité d'un système de reconnaissance en environnement virtuel suppose une segmentation temporelle efficace du flux continu de données de mouvement entrant. Le recours à une posture de repos n'est pas une solution valable, puisqu'en conditions naturelles, les mouvements s'enchainent de façon continue, dans un phénomène appelé co-articulation. Concrètement, dans un environnement virtuel d'apprentissage de karaté, l'apprenant ne peut et ne doit pas observer de posture de repos intermédiaire lors d'un enchainement. Le problème de la segmentation est un véritable défi dans le domaine de la reconnaissance de mouvements. Néanmoins, les résultats de la reconnaissance précoce ouvrent une piste potentielle, puisqu'elle montre qu'une sous-séquence suffit à reconnaitre le mouvement intégral. En poursuivant cette approche, une sous-séquence optimale caractérisant chaque classe de mouvements pourrait être isolée. On se réfère généralement à ce type de sous-séquence sous le terme de *primitive gestuelle*. Ainsi, segmenter et reconnaitre le mouvement reviendrait à rechercher les primitives dans le flux de données. Les descripteurs amorphologiques proposés dans cette thèse apparaissent comme un candidat idéal pour décrire ces primitives.

La segmentation temporelle pose un autre problème majeur : la capacité du système à fournir un indicateur de confiance sur sa décision. Les conditions d'utilisation peuvent amener le système de reconnaissance à traiter des situations inattendues. Un mouvement peut être interrompu ou modifié en cours d'exécution, la segmentation peut proposer un mouvement ne faisant pas partie des classes connues par le système... Pourtant, dans l'état actuel, notre système fournirait tout de même une décision, nécessairement erronée. Doter le système d'un indicateur de confiance, lui permettrait d'éviter d'attribuer à tort une classe à un mouvement inconnu lorsqu'un doute survient.

Outre les implications scientifiques dans le domaine de la reconnaissance de mouvement, les

travaux menés durant cette thèse ont débouché sur la conception d'outils algorithmiques dédiés à l'analyse et à la reconnaissance de mouvements. Ces outils sont maintenant à la disposition du laboratoire M2S et de la société Artefacto, partenaire industriel de cette thèse dans le cadre d'une convention industrielle (CIFRE). Artefacto a développé une expertise autour de la réalité virtuelle, de la réalité augmentée et de la visualisation 3D. Dans ces secteurs en perpétuelle mutation, l'innovation est nécessairement au cœur de l'activité. La société consacre d'ailleurs 15% de son effectif au département R&D et participe activement à des projets de recherche au sein de partenariats nationaux et internationaux. La collaboration avec le laboratoire M2S autour du projet Biofeedback en est un exemple. Les développements réalisés au cours de cette thèse ont notamment abouti à une bibliothèque C++, qui s'intègre pleinement aux outils d'Artefacto. Ces fonctionnalités sont les suivantes :

- gérer des flux de capture de mouvement, en temps réel ou stockés dans des fichiers ;
- transformer les flux en descripteurs cinématiques ;
- produire un rendu graphique des mouvements ;
- entrainer des modèles de mouvement ;
- reconnaitre des mouvements observés à la volée.

D'autre part, ces fonctionnalités s'intègrent également, au sein du projet Biofeedback, avec les méthodes d'évaluation de la performance motrice, qui ont été développées en étroite collaboration avec Emmanuel Badier, ingénieur au laboratoire M2S. Ces développements visent à quantifier les erreurs dans la performance des utilisateurs en fonction de critères d'expertise attendus. Ils s'appuient pour partie sur les descripteurs amorphologiques. L'ensemble de ces travaux est actuellement exploité dans le cadre de la thèse d'Anne-Marie Burns au laboratoire M2S, dont l'objectif est de démontrer, en pratique, les apports de la réalité virtuelle dans l'apprentissage de mouvements de karaté.

Pour conclure, ces travaux de recherche s'inscrivent dans une perspective plus large d'évolution de notre rapport à l'environnement extérieur, qui se veut de plus en plus dématérialisé. Les modes d'interaction naturels améliorent grandement l'interactivité avec les environnements virtuels, en réduisant la phase d'appropriation de l'interface par l'utilisateur. Le geste étant naturellement un mode d'interaction privilégié, sa reconnaissance est essentielle dans la mise en œuvre d'Interfaces Homme-Machine (IHM) efficaces.

Apport au domaine

Dans le domaine des STAPS, la reconnaissance de mouvements ouvre des perspectives importantes dans le domaine de l'entraînement ou dans l'étude des stratégies motrices du sportif.

Il existe de plus en plus d'applications interactives dans lesquelles les utilisateurs sont amenés à pratiquer une activité physique. Ces applications grand public sont pour l'heure restées circonscrites à une approche purement ludique, sans objectif d'amélioration de performance ou de qualité physique. Cette thèse s'inscrit dans le cadre du projet Biofeedback, qui vise justement à concevoir un environnement interactif d'entraînement dédié à l'apprentissage de performances motrices en sport (karaté et danse aérobic). Une des problématiques majeures posée par un tel système réside dans sa capacité à évaluer automatiquement la performance de l'utilisateur apprenant par rapport aux critères d'expertise requis par la discipline. À terme, cette évaluation doit permettre de guider l'utilisateur dans son apprentissage, en fournissant un retour sur ses erreurs et en adaptant les séances d'entraînement à son niveau. La figure 3.22 présente la philosophie de cette approche. Cependant, en condition interactive, le système ne peut pas savoir à l'avance quel type de mouvement est effectué par l'utilisateur. La reconnaissance automatique des mouvements de l'utilisateur, qui fait l'objet de cette thèse, est donc un pré-requis essentiel.

De plus, les études sur l'évaluation automatique de la performance ont été menées en étroite collaboration avec les travaux de cette thèse. En effet, un aspect essentiel pour évaluer le mouvement est de pouvoir comparer des performances de plusieurs utilisateurs. Réduire les différences morphologiques est donc fondamental. La description amorphologique du mouvement proposée dans cette thèse montre que cette réduction est possible dès l'étape de description des données. Les méthodes développées pour la reconnaissance de mouvements ainsi que celles de l'évaluation de la performance ont été réalisées sur un socle commun afin d'avoir une bibliothèque homogène de calcul sur le mouvement humain. Cette bibliothèque est déjà utilisée dans le cadre d'un autre projet et va permettre de fédérer les travaux du laboratoire dans ce domaine.

Par ailleurs, cette thèse a été réalisée au laboratoire M2S qui possède une solide expertise en matière d'analyse du mouvement à partir d'outils immersifs. En immergeant des sportifs dans

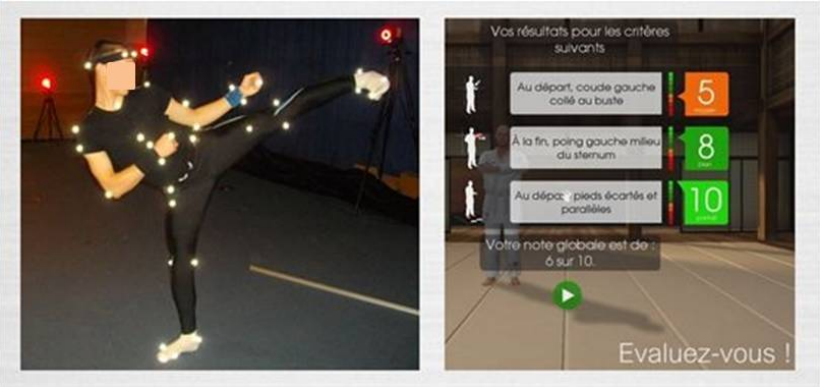


Figure 3.22 - Illustration du projet Biofeedback. À gauche, des mouvements ont été capturés sur des experts de karaté. À droite, l'environnement virtuel estime les performances motrices de l'utilisateur apprenant et le guide dans son apprentissage.

des situations totalement standardisées, la réalité virtuelle permet de mettre en évidence des processus naturels de perception, de décision et d'action [Bideau2010]. Des travaux menés au laboratoire ont déjà permis d'analyser les stratégies motrices mises en place lors de situations de duel au handball [Vignais2009] ou au rugby [Brault2011], face à des personnages virtuels. La figure 3.23 présente une méthodologie mise en place dans ce cadre. Ces études mettent en évidence de précieux indices sur les processus de prise de décision de l'utilisateur immergé dans ces duels.

Cependant, les personnages virtuels ne possèdent pas encore de capacité d'adaptation aux mouvements de l'utilisateur. Le champ d'étude est donc limité à des situations où les personnages virtuels ont un comportement prédéterminé. La reconnaissance des mouvements de l'utilisateur permettrait de lever en partie cette limitation. À terme, l'environnement virtuel pourra adapter les comportements des sportifs virtuels en fonction des actions du sujet immergé. Le cadre strict du duel peut alors être dépassé pour étudier de nouvelles situations de jeu. Au-delà des perspectives de travail qu'ouvre une telle adaptation, ce sont les possibilités d'interactions naturelles de l'utilisateur avec l'environnement virtuel qui apportent une plus-value majeure. Elles renforcent le sentiment de présence, permettant à l'utilisateur d'agir plus naturellement. En définitive, pour observer des comportements naturels, il est nécessaire de proposer des modalités d'interactions naturelles.

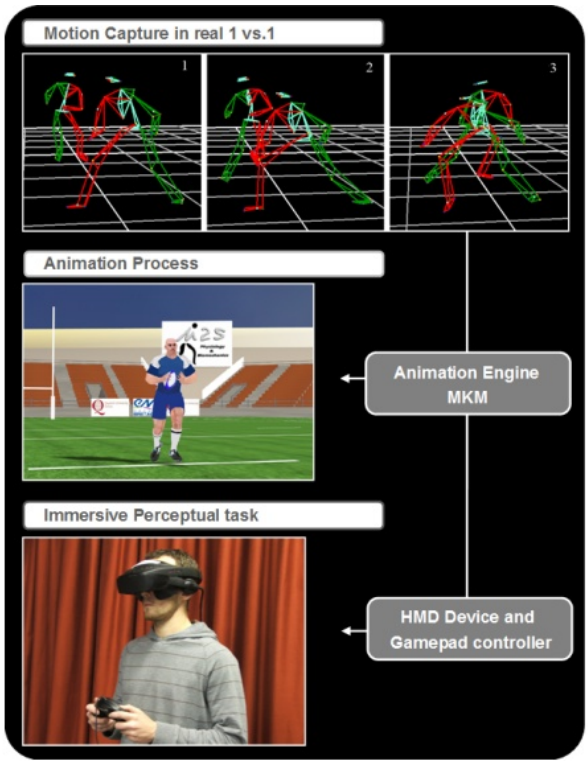


Figure 3.23 - Méthodologie générale de [Brault2011] : de la capture de mouvement à la tâche de jugement en environnement virtuel.

Annexes

Annexe A

Chronophotographies des mouvements

Dans cette annexe sont présentés des exemples types de chaque classe de mouvements utilisée dans les études de cette thèse.

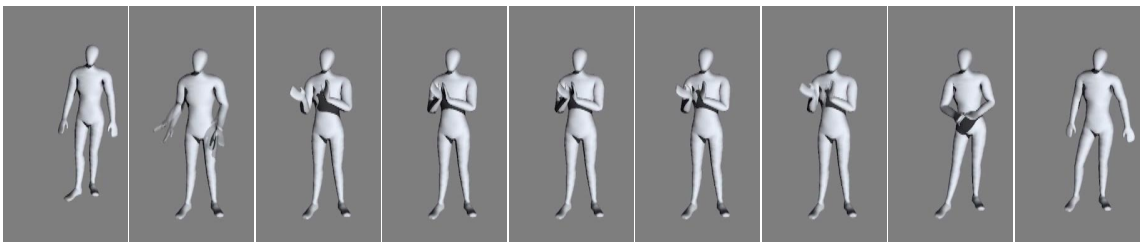


Figure A.1 - Applaudir

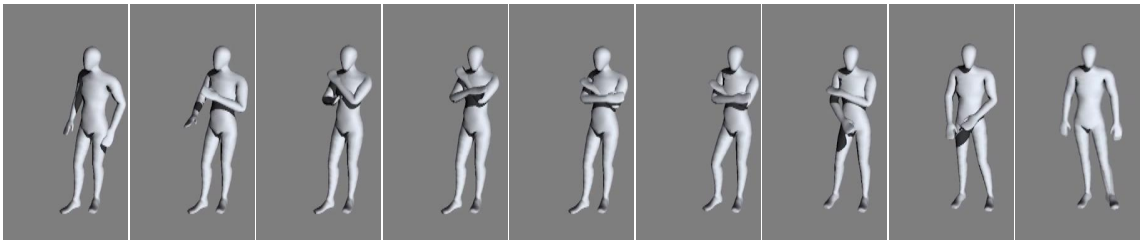


Figure A.2 - Bras croisés

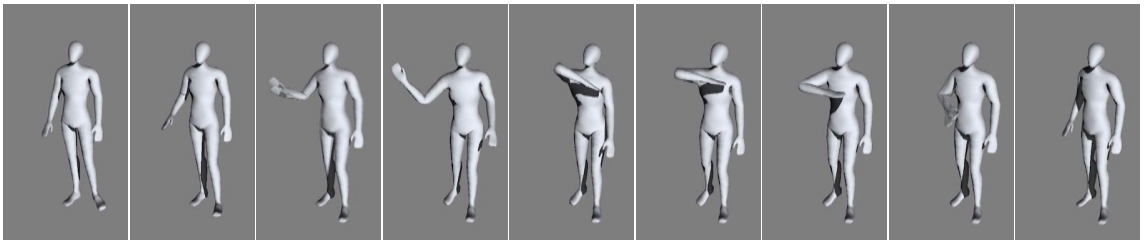


Figure A.3 - Claque paume

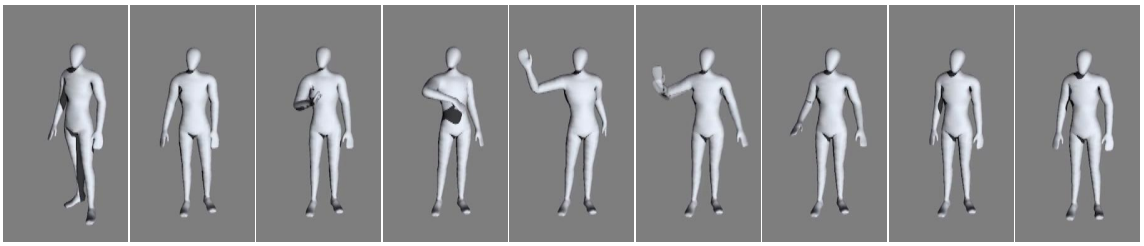


Figure A.4 - Claque revers

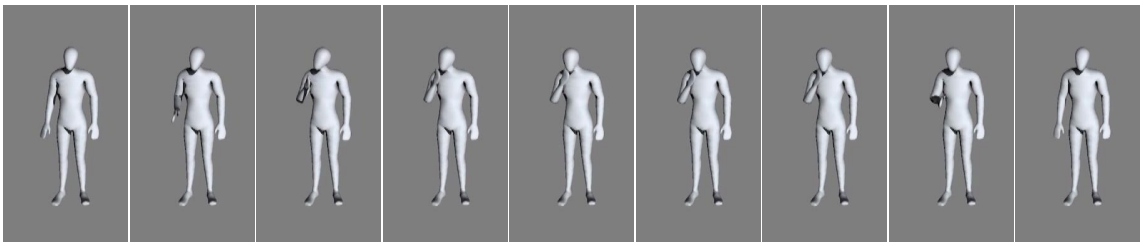


Figure A.5 - Gratter menton

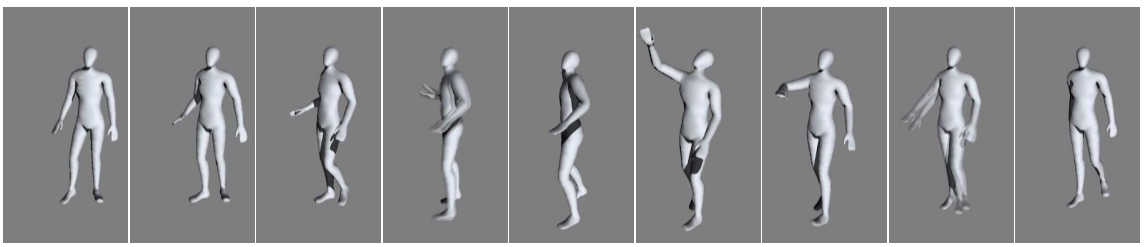


Figure A.6 - Lancer

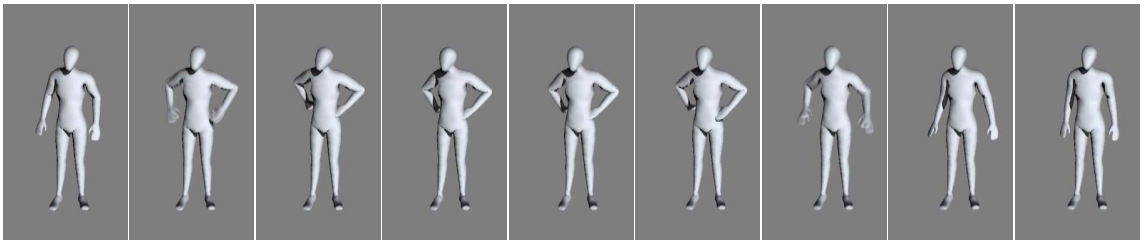


Figure A.7 - Mains hanches

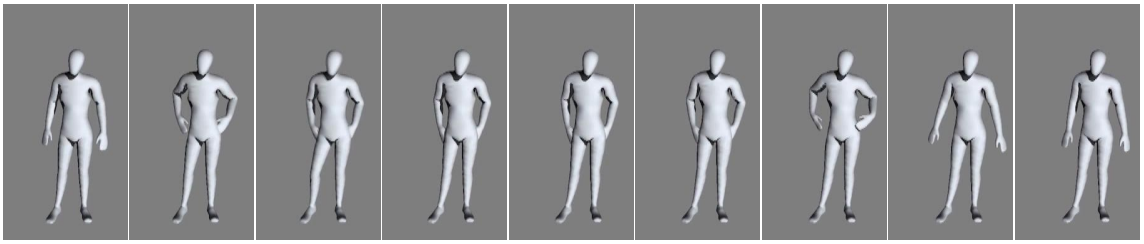


Figure A.8 - Mains poches

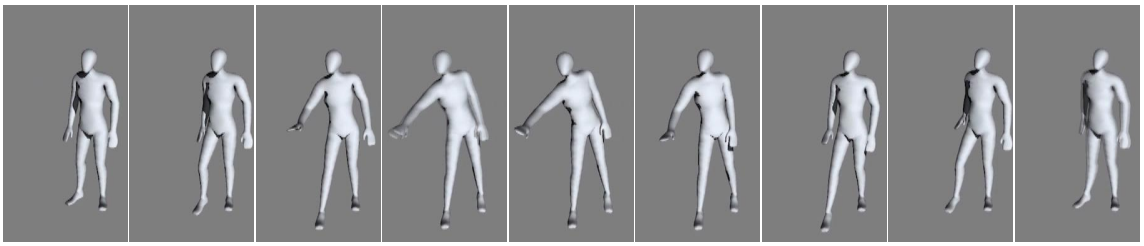


Figure A.9 - Prendre bas

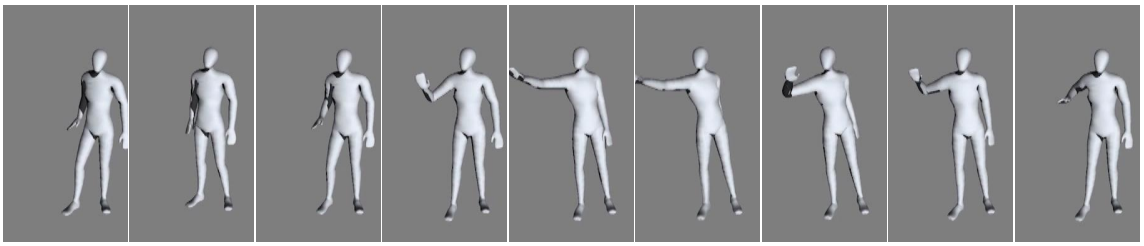


Figure A.10 - Prendre haut

Figure A.11 - Prendre milieu

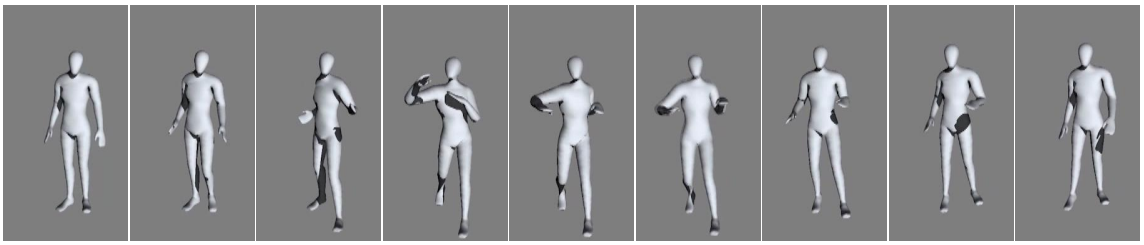


Figure A.12 - Punch

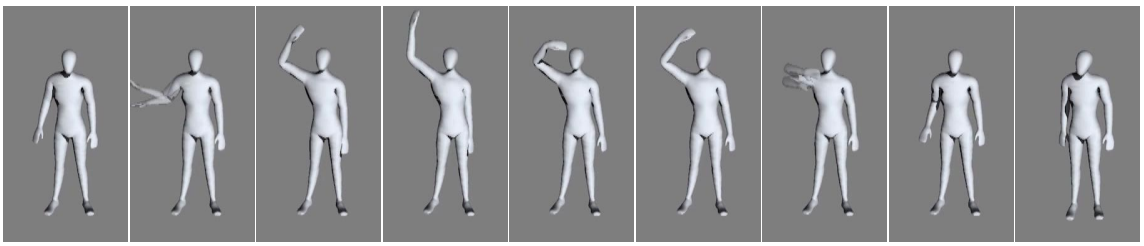


Figure A.13 - Salut haut

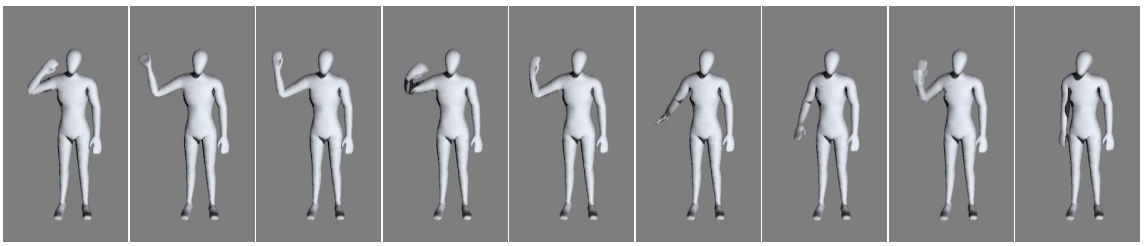


Figure A.14 - Salut tête

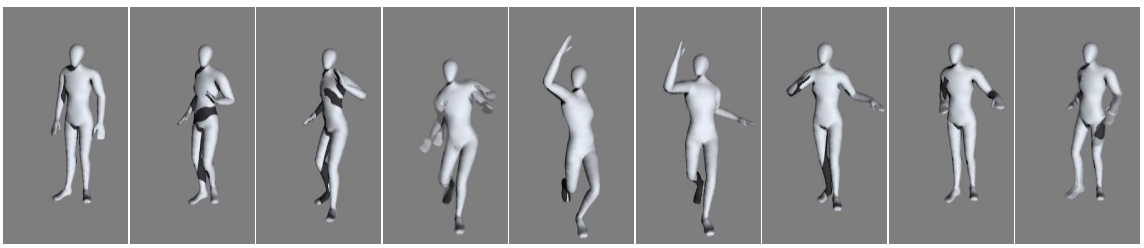


Figure A.15 - Uppercut

Bibliographie

[Aarno2008] Aarno, D. et Kragic, D. *Motion intention recognition in robot assisted applications*. Robotics and Autonomous Systems, vol. 56, n° 8, pages 692 – 705, 2008.

[Agrawal1993] Agrawal, R., Faloutsos, C. et Swami, A. *Efficient similarity search in sequence databases*. Foundations of Data Organization and Algorithms, vol. 730, pages 69–84, 1993.

[Alon2009] Alon, J., Athitsos, V., Yuan, Q. et Sclaroff, S. *A unified framework for gesture recognition and spatiotemporal gesture segmentation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, n° 9, pages 1685–1699, 2009.

[Altun2010] Altun, K., Barshan, B. et Tunçel, O. *Comparative study on classifying human activities with miniature inertial and magnetic sensors*. Pattern Recogn., vol. 43, n° 10, pages 3605–3620, 2010.

[Amarantini2004] Amarantini, D. et Martin, L. *A method to combine numerical optimization and EMG data for the estimation of joint moments under dynamic conditions*. Journal of biomechanics, vol. 37, n° 9, pages 1393–1404, 2004.

[Amft2007] Amft, O., Lombriser, C., Stiefmeier, T. et Tröster, G. *Recognition of user activity sequences using distributed event detection*. In Proceedings of the 2nd European conference on Smart sensing and context, EuroSSC'07, pages 126–141, Kendal, England, 2007. Springer-Verlag.

[Amft2008] Amft, O. et Troster, G. *Recognition of dietary activity events using on-body sensors*. Artificial Intelligence in Medicine, vol. 42, n° 2, pages 121 – 136, 2008. Wearable Computing and Artificial Intelligence for Healthcare Applications.

[Anagnostopoulos2006] Anagnostopoulos, A., Vlachos, M., Hadjieleftheriou, M., Keogh, E. et Yu, P. S. *Global distance-based segmentation of trajectories*. In KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 34–43, Philadelphia, PA, USA, 2006. ACM.

[Arsic2010] Arsic, D., Roalter, L., Wöllmer, M., Eyben, F., Schuller, B., Kaiser, M., Kranz, M. et Rigoll, G. *3d gesture recognition applying long short-term memory and contextual knowledge in a CAVE*. In

Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis, MPVA '10, pages 33–36, Firenze, Italy, 2010. ACM.

[Axenbeck2008] Axenbeck, T., Bennewitz, M., Behnke, S. et Burgard, W. *Recognizing complex, parameterized gestures from monocular image sequences*. In Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pages 687–692. IEEE, 2008.

[Bailador2007] Bailador, G., Roggen, D., Tröster, G. et Trivi no, G. *Real time gesture recognition using continuous time recurrent neural networks*. In BodyNets '07 : Proceedings of the ICST 2nd international conference on Body area networks, pages 1–8, Florence, Italy, 2007. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[Bao2004] Bao, L. et Intille, S. *Activity Recognition from User-Annotated Acceleration Data*. In Ferscha, A. et Mattern, F., éditeurs, Pervasive Computing, volume 3001 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin / Heidelberg, 2004.

[Barbic2004] Barbic, J., Safonova, A., Pan, J., Faloutsos, C., Hodgins, J. et Pollard, N. *Segmenting motion capture data into distinct behaviors*. In Proceedings of Graphics Interface 2004, pages 185–194. Canadian Human-Computer Communications Society School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004.

[Barclay1978] Barclay, C. D., Cutting, J. E. et Kozlowski, L. T. *Temporal and spatial factors in gait perception that influence gender recognition*. Perception & Psychophysics, vol. 23(2), pages 145–152, 1978.

[Bashir2005] Bashir, F., Qu, W., Khokhar, A. et Schonfeld, D. *HMM-based motion recognition system using segmented PCA*. In Image Processing, 2005. ICIP 2005. IEEE International Conference on, volume 3, pages III–1288. IEEE, 2005.

[Beardsworth1981] Beardsworth, T. et Buckner, T. *The ability to recognize oneself from a video recording of one's movements without seeing one's body.* Bulletin of the Psychonomic Society, vol. 18, pages 19–22, 1981.

[Ben-Arie2002] Ben-Arie, J., Wang, Z., Pandit, P. et Rajaram, S. *Human Activity Recognition Using Multidimensional Indexing*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, pages 1091–1104, August 2002.

[Benbasat2002] Benbasat, A. et Paradiso, J. *An Inertial Measurement Framework for Gesture Recognition and Applications*. pages 77–90. 2002.

[Bernstein1967] Bernstein, N. The co-ordination and regulation of movements. Oxford, UK : Pergamon, 1967.

[Bevilacqua2010] Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F. et Rasamimanana, N. *Continuous Realtime Gesture Following and Recognition*. In Kopp, S. et Wachsmuth, I., éditeurs, Gesture in Embodied Communication and Human-Computer Interaction, volume 5934 of *Lecture Notes in Computer Science*, pages 73–84. Springer Berlin / Heidelberg, 2010.

[Bhowmik2011] Bhowmik, T. K., van Oosten, J.-P. et Schomaker, L. *Segmental K-means learning with mixture distribution for HMM based handwriting recognition*. In Proceedings of the 4th international conference on Pattern recognition and machine intelligence, PReMI'11, pages 432–439, Moscow, Russia, 2011. Springer-Verlag.

[Bideau2003] Bideau, B., Kulpa, R., Ménardais, S., Fradet, L., Multon, F., Delamarche, P. et Arnaldi, B. *Real handball goalkeeper vs. virtual handball thrower*. Presence : Teleoperators & Virtual Environments, vol. 12, n° 4, pages 411–421, 2003.

[Bideau2010] Bideau, B., Kulpa, R., Vignais, N., Brault, S., Multon, F. et Craig, C. *Using virtual reality to analyze sports performance*. Computer Graphics and Applications, IEEE, vol. 30, n° 2, pages 14–21, 2010.

[Bishop2006] Bishop, C. Pattern recognition and machine learning, volume 4. Springer New York, 2006.

[Blackburn2007] Blackburn, J. et Ribeiro, E. *Human motion recognition using Isomap and dynamic time warping*. In Proceedings of the 2nd conference on Human motion : understanding, modeling, capture and animation, pages 285–298, Rio de Janeiro, Brazil, 2007. Springer-Verlag.

[Bobick1996] Bobick, A. et Davis, J. *An appearance-based representation of action*. In Pattern Recognition, 1996., Proceedings of the 13th International Conference on, volume 1, pages 307–312. IEEE, 1996.

[Bodenheimer1997] Bodenheimer, B., Rose, C., Rosenthal, S. et Pella, J. *The process of motion capture : Dealing with the data*. In Computer Animation and Simulation, volume 97, page 2. Citeseer, 1997.

[Bonnefoy2008] Bonnefoy, A., Robert, T., Dumas, R. et Cheze, L. *Advanced biomechanical methods for the computation of joint moments and muscular forces*. Imagerie et Recherche Biomédicale, 2008.

[Boulabiar2011] Boulabiar, M.-I., Burger, T., Poirier, F. et Coppin, G. *A low-cost natural user interaction based on a camera hand-gestures recognizer*. In Proceedings of the 14th international conference on Human-computer interaction : interaction techniques and environments - Volume Part II, HCII'11, pages 214–221, Orlando, FL, 2011. Springer-Verlag.

[Bouveyron2006] Bouveyron, C. *Modélisation et classification des données de grande dimension : application à l'analyse d'images*. These, Université Joseph-Fourier - Grenoble I, Septembre 2006.

[Brand1997] Brand, M., Oliver, N. et Pentland, A. *Coupled hidden Markov models for complex action recognition*. In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pages 994–999. IEEE, 1997.

[Brault2011] Brault, S. *La feinte de corps au rugby : déterminants biomécaniques, processus de détection et action de défense : pourquoi l'expert est-il meilleur ?* Thèse, Université Rennes 2, 2011.

[Burges1998] Burges, C. *A tutorial on support vector machines for pattern recognition*. Data mining and knowledge discovery, vol. 2, n° 2, pages 121–167, 1998.

[Bussmann2001] Bussmann, J., Martens, W., Tulen, J., Schasfoort, F., van den Berg-Emons, H. et Stam, H. *Measuring daily behavior using ambulatory accelerometry : the Activity Monitor*. Behavior Research Methods, vol. 33, n° 3, pages 349–356, 2001.

[Byun2002] Byun, H. et Lee, S.-W. *Applications of Support Vector Machines for Pattern Recognition : A Survey*. In Lee, S.-W. et Verri, A., éditeurs, Pattern Recognition with Support Vector Machines, volume 2388 of *Lecture Notes in Computer Science*, pages 571–591. Springer Berlin / Heidelberg, 2002.

[Cadoz1994] Cadoz, C. *Le geste canal de communication homme/machine*. page 31. Afcet, 1994.

[Cadoz2000] Cadoz, C. et Wanderley, M. *Gesture-music*. Trends in Gestural Control of Music, pages 71–93, 2000.

[Caillette2008] Caillette, F., Galata, A. et Howard, T. *Real-time 3-D human body tracking using learnt models of behaviour*. Computer Vision and Image Understanding, vol. 109, n° 2, pages 112–125, 2008.

[Calinon2004] Calinon, S. et Billard, A. *Stochastic gesture production and recognition model for a humanoid robot*. In Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, volume 3, pages 2769–2774. IEEE, 2004.

[Calinon2005] Calinon, S. et Billard, A. *Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM*. In Proceedings of the 22nd international conference on Machine learning, ICML '05, pages 105–112, Bonn, Germany, 2005. ACM.

[Campbell1995] Campbell, L. et Bobick, A. *Recognition of human body motion using phase space constraints*. In International Conference on Computer Vision, pages 624–630. Citeseer, 1995.

[Campbell1996] Campbell, L. W., Becker, D. A., Azarbayejani, A., Bobick, A. F. et Pentland, A. *Invariant features for 3D gesture recognition*. In Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on, pages 157–162, 1996.

[Caridakis2010] Caridakis, G., Karpouzis, K., Drosopoulos, A. et Kollias, S. *SOMM : Self organizing Markov map for gesture recognition*. Pattern Recognition Letters, vol. 31, n° 1, pages 52 – 59, 2010.

[Carvalho2007] Carvalho, S., Boulic, R. et Thalmann, D. *Motion Pattern Preserving IK Operating in the Motion Principal Coefficients Space*. In The 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2007), pages 97–104, Plzen - Czech Republic, 2007.

[Castellani2004] Castellani, A., Botturi, D., Bicego, M. et Fiorini, P. *Hybrid HMM/SVM model for the analysis and segmentation of teleoperation tasks*. In Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on, volume 3, pages 2918–2923. IEEE, 2004.

[Cédras1995] Cédras, C. et Shah, M. *Motion-based recognition a survey*. Image and Vision Computing, vol. 13, n° 2, pages 129–155, Mars 1995.

[Chai2005] Chai, J. et Hodgins, J. K. *Performance Animation from Low-dimensional Control Signals*. ACM Transactions on Graphics, vol. 24, pages 686–696, 2005.

[Chakraborty2008] Chakraborty, B., Pedersoli, M. et Gonzàlez, J. *View-Invariant Human Action Detection Using Component-Wise HMM of Body Parts*. In Perales, F. et Fisher, R., éditeurs, Articulated Motion and Deformable Objects, volume 5098 of *Lecture Notes in Computer Science*, pages 208–217. Springer Berlin / Heidelberg, 2008.

[Chan2009] Chan, C. S. et Liu, H. *Fuzzy qualitative human motion analysis*. Trans. Fuz Sys., vol. 17, pages 851–862, August 2009.

[Cherla2008] Cherla, S., Kulkarni, K., Kale, A. et Ramasubramanian, V. *Towards fast, view-invariant human action recognition*. Computer Vision and Pattern Recognition Workshop, vol. 0, pages 1–8, 2008.

[Chiari2005] Chiari, L., Croce, U. D., Leardini, A. et Cappozzo, A. *Human movement analysis using stereophotogrammetry : Part 2 : Instrumental errors*. Gait &amp ; Posture, vol. 21, n ° 2, pages 197 – 211, 2005.

[Cholewa2011] Cholewa, M. et Glomb, P. *Deciding of HMM parameters based on number of critical points for gesture recognition from motion capture data*. Rapport technique, 2011.

[Chung2008] Chung, P.-C. et Liu, C.-D. *A daily behavior enabled hidden Markov model for human behavior understanding*. Pattern Recognition, vol. 41, n ° 5, pages 1572 – 1580, 2008.

[Clarke2005] Clarke, T., Bradshaw, M., Field, D., Hampson, S., Rose, D.*et al. The perception of emotion from body movement in point-light displays of interpersonal dialogue*. PERCEPTION-LONDON-, vol. 34, n ° 10, page 1171, 2005.

[Coogan2006] Coogan, T., Awad, G., Han, J. et Sutherland, A. *Real Time Hand Gesture Recognition Including Hand Segmentation and Tracking*. In Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Nefian, A., Meenakshisundaram, G., Pascucci, V., Zara, J., Molineros, J., Theisel, H. et Malzbender, T., éditeurs, Advances in Visual Computing, volume 4291 of *Lecture Notes in Computer Science*, pages 495–504. Springer Berlin / Heidelberg, 2006.

[Corazza2006] Corazza, S., Mündermann, L., Chaudhari, A., Demattio, T., Cobelli, C. et Andriacchi, T. *A Markerless Motion Capture System to Study Musculoskeletal Biomechanics : Visual Hull and Simulated Annealing Approach*. Annals of Biomedical Engineering, vol. 34, pages 1019–1029, 2006.

[Corradini2002] Corradini, A. et Cohen, P. *Multimodal speech-gesture interface for handfree painting on a virtual paper using partial recurrent neural networks as gesture recognizer*. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, volume 3, pages 2293–2298. IEEE, 2002.

[Cortes1995] Cortes, C. et Vapnik, V. *Support-vector networks*. Machine Learning, vol. 20, pages 273–297, 1995. 10.1007/BF00994018.

[Cutting1977] Cutting, J. et Kozlowski, L. *Recognizing friends by their walk : Gait perception without familiarity cues*. Bulletin of the Psychonomic Society, vol. 9, n ° 5, pages 353–356, 1977.

[Daffertshofer2004] Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G. et Beek, P. J. *PCA in studying coordination and variability : a tutorial*. Clinical Biomechanics, vol. 19, n° 4, pages 415 – 428, 2004.

[Dauchy1993] Dauchy, P., Mignot, C. et Valot, C. *Joint speech and gesture analysis some experimental results on multimodal interface.* In EUROSPEECH. ISCA, 1993.

[Delp2007] Delp, S., Anderson, F., Arnold, A., Loan, P., Habib, A., John, C., Guendelman, E. et Thelen, D. *OpenSim : open-source software to create and analyze dynamic simulations of movement*. Biomedical Engineering, IEEE Transactions on, vol. 54, n° 11, pages 1940–1950, 2007.

[Deng2009] Deng, Z., Gu, Q. et Li, Q. *Perceptually consistent example-based human motion retrieval*. In I3D '09 : Proceedings of the 2009 symposium on Interactive 3D graphics and games, pages 191–198, Boston, Massachusetts, 2009. ACM.

[Difranco2001] Difranco, D. E. et jen Cham, T. *Reconstruction of 3-d figure motion from 2-d correspondences*. In In Computer Vision and Pattern Recognition, pages 307–314, 2001.

[Dittrich1996] Dittrich, W., Troscianko, T., Lea, S. et Morgan, D. *Perception of emotion from dynamic point-light displays represented in dance*. Perception-London, vol. 25, n° 6, pages 727–738, 1996.

[Donovan2005] Donovan, J. et Brereton, M. *Movements in gesture interfaces*. In Larssen, A., Robertson, T., Brereton, M., Loke, L. et Edwards, J., éditeurs, Critical Computing 2005 - Between Sense and Sensibility, the Fourth Decennial Aarhus Conference. Proceedings of the Worskshop : Approaches to Movement- Based Interaction. IDWoP, Interaction Design and Work Practice Lab, Faculty of Information Technology, UTS, 2005.

[Efros2003] Efros, A., Berg, A., Mori, G. et Malik, J. *Recognizing action at a distance*. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pages 726–733. IEEE, 2003.

[Eian2002] Eian, J. et Poppele, R. E. *A single-camera method for three-dimensional video imaging*. Journal of Neuroscience Methods, vol. 120, n° 1, pages 65 – 83, 2002.

[Elmezain2009] Elmezain, M., Al-Hamadi, A., Pathan, S. et Michaelis, B. *Spatio-temporal feature extraction-based hand gesture recognition for isolated American Sign Language and Arabic numbers*. In Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on, pages 254–259. IEEE, 2009.

[Faraway2007] Faraway, J. et Reed, M. *Statistics for digital human motion modeling in ergonomics*. Technometrics, vol. 49, n° 3, pages 277–290, 2007.

[Feng2002] Feng, X. et Perona, P. *Human Action Recognition By Sequence of Movelet Codewords*. 3D Data Processing Visualization and Transmission, International Symposium on, vol. 0, page 717, 2002.

[Fihl2006] Fihl, P., Holte, M., Moeslund, T. et Reng, L. *Action Recognition Using Motion Primitives and Probabilistic Edit Distance*. In Springer, B., éditeur, Articulated motion and deformable objects (4th international conference), pages 375–384, 2006.

[Fink2008] Fink, G. Markov models for pattern recognition. Springer, 2008.

[Fleury2009] Fleury, A., Noury, N., Vacher, M.*et al. Application des SVM à la classification des Activités de la Vie Quotidienne d'une personne à partir des capteurs d'un Habitat Intelligent pour la Santé*, 2009.

[Fod2002] Fod, A., Matarić, M. J. et Jenkins, O. C. *Automated Derivation of Primitives for Movement Classification*. Auton. Robots, vol. 12, n° 1, pages 39–54, 2002.

[Francke2007] Francke, H., Ruiz-del Solar, J. et Verschae, R. *Real-Time Hand Gesture Detection and Recognition Using Boosted Classifiers and Active Learning*. In Mery, D. et Rueda, L., éditeurs, Advances in Image and Video Technology, volume 4872 of *Lecture Notes in Computer Science*, pages 533–547. Springer Berlin / Heidelberg, 2007.

[Gallagher2004] Gallagher, A., Matsuoka, Y. et Ang, W.-T. *An efficient real-time human posture tracking algorithm using low-cost inertial and magnetic sensors*. volume 3, pages 2967–2972 vol.3, 2004.

[GfK2011] *Les Français et l'Entertainment*, 2011.

[Ghahramani2001] Ghahramani, Z. *An introduction to hidden Markov models and Bayesian networks*. IJPRAI, vol. 15, n° 1, pages 9–42, 2001.

[Giansanti2005] Giansanti, D., Maccioni, G. et Macellari, V. *The development and test of a device for the reconstruction of 3-D position and orientation by means of a kinematic sensor assembly with rate gyroscopes and accelerometers*. IEEE transactions on biomedical engineering, vol. 52, n° 7, pages 1271–1277, 2005.

[Gielen2009] Gielen, S. *Review of Models for the Generation of Multi-Joint Movements in 3-D*. Progress in Motor Control, vol. 629, pages 523–550, 2009.

[Gillies2009] Gillies, M. *Learning Finite State Machine Controllers from Motion Capture Data*. IEEE transactions on computational intelligence and AI in games, vol. 1, n° 1, pages 63–72, 2009.

[Gleicher1998] Gleicher, M. *Retargetting motion to new characters*. In Proceedings of the 25th annual conference on Computer graphics and interactive techniques, SIGGRAPH '98, pages 33–42. ACM, 1998.

[Gonzalez2006] Gonzalez, G. Kinematic tracking and activity recognition using motion primitives. Master's thesis, Royal Institute of Technology (Department of Computer and Systems Sciences), 2006.

[Grassia1998] Grassia, F. *Practical parameterization of rotations using the exponential map*. Journal of graphics tools, vol. 3, pages 29–48, 1998.

[Green2004] Green, R. D. et Guan, L. *Quantifying and recognizing human movement patterns from monocular video Images-part i : a new framework for modeling human motion*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, n° 2, pages 179–190, 2004.

[Gu2009] Gu, Q., Peng, J. et Deng, Z. *Compression of Human Motion Capture Data Using Motion Pattern Indexing*. Computer Graphics Forum, vol. 28, pages 1–12(12), march 2009.

[Guenterberg2009] Guenterberg, E., Ghasemzadeh, H., Loseu, V. et Jafari, R. *Distributed Continuous Action Recognition Using a Hidden Markov Model in Body Sensor Networks*. Distributed Computing in Sensor Systems, pages 145–158, 2009.

[Hecker2008] Hecker, C., Raabe, B., Enslow, R. W., DeWeese, J., Maynard, J. et van Prooijen, K. *Real-time motion retargeting to highly varied user-created morphologies*. ACM Trans. Graph., vol. 27, pages 27 :1–27 :11, August 2008.

[Herzog2008] Herzog, D., Krüger, V. et Grest, D. *Parametric hidden Markov models for recognition and synthesis of movements*. In Proceedings of the British Machine Vision Conference, 2008.

[Hill2000] Hill, H. et Pollick, F. *Exaggerating temporal differences enhances recognition of individuals from point light displays*. Psychol Sci., vol. 11, pages 223–8, 2000.

[Hill2001] Hill, H. et Johnston, A. *Categorizing sex and identity from the biological motion of faces*. Current Biology, vol. 11, n° 11, pages 880–885, 2001.

[Hirashima1999] Hirashima, S. *Recognition on the gender of point-light walkers moving in different directions].* Shinrigaku kenkyu : The Japanese journal of psychology, vol. 70, n° 2, page 149, 1999.

[Hong2000] Hong, P., Huang, T. S. et Turk, M. *Gesture Modeling and Recognition Using Finite State Machines*. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, FG '00, pages 410–. IEEE Computer Society, 2000.

[Hossain2005] Hossain, M. et Jenkin, M. *Recognizing Hand-Raising Gestures using HMM*. Computer and Robot Vision, Canadian Conference, vol. 0, pages 405–412, 2005.

[Hu2009] Hu, D., Zhang, X., Yin, J., Zheng, V. et Yang, Q. *Abnormal activity recognition based on hdp-hmm models*. In Proceedings of the 21st international jont conference on Artifical intelligence, pages 1715–1720. Morgan Kaufmann Publishers Inc., 2009.

[Huang1995] Huang, T. et Pavlovic, V. *Hand gesture modeling, analysis, and synthesis*. In In Proc. of IEEE International Workshop on Automatic Face and Gesture Recognition. Citeseer, 1995.

[Huang2007] Huang, C.-H., Wang, C.-S. et Yu, M.-L. *Automatic 3D CBIR on kinematical human motion*. In Proceedings of the 2007 annual Conference on International Conference on Computer Engineering and Applications, pages 242–247, Gold Coast, Queensland, Australia, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[Ikizler2007] Ikizler, N. et Duygulu, P. *Human Action Recognition Using Distribution of Oriented Rectangular Patches*. pages 271–284. 2007.

[Ikizler2008] Ikizler, N. et Forsyth, D. A. *Searching for Complex Human Activities with No Visual Examples*. Int. J. Comput. Vision, vol. 80, pages 337–357, December 2008.

[Ilg2004] Ilg, W., Bakir, G. et Mezger, J.and Giese, M. *On the Representation, Learning and Transfer of Spatio-Temporal Movement Characteristics*. In Humanoids Proceedings, 2004.

[Inamura2004] Inamura, T., Toshima, I., Tanie, H. et Nakamura, Y. *Embodied symbol emergence based on mimesis theory*. The International Journal of Robotics Research, vol. 23, n° 4-5, page 363, 2004.

[Ishigaki2009] Ishigaki, S., White, T., Zordan, V. et Liu, C. *Performance-based Control Interface for Character Animation*. ACM Trans on Graphics, vol. 28, n° 3, pages 61 :1–61 :8, 2009.

[Iverson1998] Iverson, J. M. et Goldin-Meadow, S. *Why people gesture when they speak*. Nature, vol. 396, n° 6708, pages 228–228, Novembre 1998.

[Jenkins2007] Jenkins, O. C. et Serrano, G. G. *Interactive human pose and action recognition using dynamical motion primitives*. International Journal of Humanoid Robotics, vol. 4, n° 2, pages 365–386, 2007.

[Jhuang2007] Jhuang, H., Serre, T., Wolf, L. et Poggio, T. *A Biologically Inspired System for Action Recognition*. Computer Vision, IEEE International Conference on, vol. 0, pages 1–8, 2007.

[jin Choi1999] jin Choi, K. et seok Ko, H. *On-line Motion Retargetting*. Journal of Visualization and Computer Animation, vol. 11, pages 223–235, 1999.

[Jin2011] Jin, Y. et Prabhakaran, B. *Knowledge discovery from 3D human motion streams through semantic dimensional reduction*. ACM Trans. Multimedia Comput. Commun. Appl., vol. 7, pages 9 :1–9 :20, March 2011.

[Johansson1973] Johansson, G. *Visual perception of biological motion and a model for its analysis*. Attention, Perception & Psychophysics, vol. 14, pages 201–211, 1973.

[Johnson1999] Johnson, M. P., Wilson, A., Blumberg, B., Kline, C. et Bobick, A. *Sympathetic interfaces : using a plush toy to direct synthetic characters*. In Proceedings of the SIGCHI conference on Human factors in computing systems : the CHI is the limit, CHI '99, pages 152–158, Pittsburgh, Pennsylvania, United States, 1999. ACM.

[Jokisch2006] Jokisch, D., Daum, I. et Troje, N. *Self recognition versus recognition of others by biological motion : Viewpoint-dependent effects*. Perception, vol. 35, pages 911–920, 2006.

[Junker2008] Junker, H., Amft, O., Lukowicz, P. et Troster, G. *Gesture spotting with body-worn inertial sensors to detect user activities*. Pattern Recognition, vol. 41, n° 6, pages 2010 – 2024, 2008.

[Kadone2006] Kadone, H. et Nakamura, Y. *Segmentation, memorization, recognition and abstraction of humanoid motions based on correlations and associative memory*. In Humanoid Robots, 2006 6th IEEE-RAS International Conference on, pages 1–6. IEEE, 2006.

[Kallio2006] Kallio, S., Kela, J., Korpipää, P. et MÄNtyjÄRvi, J. *User independent gesture interaction for small handheld devices*. International Journal of Pattern Recognition and Artificial Intelligence, vol. 20, n° 4, pages 505–524, 2006.

[Keogh2001a] Keogh, E., Chu, S., Hart, D. et Pazzani, M. *Segmenting time series : A survey and novel approach*. Data mining in time series databases, pages 1–22, 2001.

[Keogh2001b] Keogh, E. J. et Pazzani, M. J. *Derivative Dynamic Time Warping*. In In First SIAM International Conference on Data Mining (SDM 2001), 2001.

[Keogh2002] Keogh, E. *Exact indexing of dynamic time warping*. In VLDB '02 : Proceedings of the 28th international conference on Very Large Data Bases, pages 406–417, Hong Kong, China, 2002. VLDB Endowment.

[Kida2005] Kida, T. *Appropriation du geste par les étrangers : le cas d'étudiants japonais apprenant le français*. Thèse, Laboratoire Parole et Langage (LPL), CNRS : UMR6057 - Université de Provence - Aix-Marseille I, 2005.

[Kilner2009] Kilner, J., Starck, J., Guillemaut, J. et Hilton, A. *Objective quality assessment in free-viewpoint video production*. Signal Processing : Image Communication, vol. 24, n° 12, pages 3 – 16, 2009. Special issue on advances in three-dimensional television and video.

[Kim2002] Kim, J., Park, K., Bang, W. et Bien, Z. *Continuous gesture recognition system for Korean sign language based on fuzzy logic and hidden Markov model*. In Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on, volume 2, pages 1574–1579. IEEE, 2002.

[Kim2007] Kim, H., Lee, Y. et Lee, C. *A Study on the Gesture Recognition Based on the Particle Filter*. In Apolloni, B., Howlett, R. et Jain, L., éditeurs, Knowledge-Based Intelligent Information and Engineering Systems, volume 4692 of *Lecture Notes in Computer Science*, pages 429–438. Springer Berlin / Heidelberg, 2007.

[Klein2008] Klein, P. et Sommerfeld, P. Biomécanique des membres inférieurs : bases et concepts, bassin, membres inférieurs. Elsevier, 2008.

[Knossow2008] Knossow, D., Ronfard, R. et Horaud, R. *Human motion tracking with a kinematic parameterization of extremal contours*. International Journal of Computer Vision, vol. 79, n° 3, pages 247–269, 2008.

[Koch2010] Koch, P., Konen, W. et Hein, K. *Gesture recognition on few training data using slow feature analysis and parametric bootstrap*. In Neural Networks (IJCNN), The 2010 International Joint Conference on, pages 1–8. IEEE, 2010.

[Kohavi1997] Kohavi, R. et John, G. H. *Wrappers for feature subset selection*. Artif. Intell., vol. 97, pages 273–324, December 1997.

[Kovar2002] Kovar, L., Gleicher, M. et Pighin, F. *Motion graphs*. ACM Trans. Graph., vol. 21, n° 3, pages 473–482, 2002.

[Kozlowski1977] Kozlowski, L. et Cutting, J. *Recognizing the sex of a walker from a dynamic point-light display*. Attention, Perception, & Psychophysics, vol. 21, pages 575–580, 1977. 10.3758/BF03198740.

[Kubota2005] Kubota, N. et Abe, M. *Computational Intelligence for Cyclic Gestures Recognition of a Partner Robot*. In Khosla, R., Howlett, R. et Jain, L., éditeurs, Knowledge-Based Intelligent Information and Engineering Systems, volume 3681 of *Lecture Notes in Computer Science*, pages 159–159. Springer Berlin / Heidelberg, 2005.

[Kulic2008] Kulic, D., Lee, D., Ott, C. et Nakamura, Y. *Incremental learning of full body motion primitives for humanoid robots*. In Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pages 326–332. IEEE, 2008.

[Kulic2009a] Kulic, D. et Nakamura, Y. *Comparative study of representations for segmentation of whole body human motion data*. In Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems, IROS'09, pages 4300–4305, St. Louis, MO, USA, 2009. IEEE Press.

[Kulic2009b] Kulic, D., Takano, W. et Nakamura, Y. *Online segmentation and clustering from continuous observation of whole body motions*. Trans. Rob., vol. 25, pages 1158–1166, October 2009.

[Kulpa2005a] Kulpa, R. *Adaptation interactive et performante des mouvements d'humanoïdes synthétiques : aspects cinématique, cinétique et dynamique*. Thèse, PhD thesis, INSA, Rennes-France, 2005.

[Kulpa2005b] Kulpa, R., Multon, F. et Arnaldi, B. *Morphology-independent representation of motions for interactive human-like animation*. In for Computer Graphics, E. A., éditeur, Eurographics. M. Alexa, J. Marks, AoÃżt 2005. http ://www.eg.org/.

[Kwok2004] Kwok, C., Fox, D. et Marina, M. *Real-time particle filters*. Proceedings of the IEEE, vol. 92, n° 3, pages 469–484, 2004.

[Kwon2007] Kwon, D. et Gross, M. *A Framework for 3D Spatial Gesture Design and Modeling Using a Wearable Input Device*. In Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers-Volume 00, pages 23 – 26. IEEE Computer Society Washington, DC, USA, 2007.

[Lange2006] Lange, J., Georg, K. et Lappe, M. *Visual perception of biological motion by form : A template-matching analysis*. Journal of Vision, vol. 6, n° 8, 2006.

[Laptev2007] Laptev, I., Caputo, B., Schüldt, C. et Lindeberg, T. *Local velocity-adapted motion events for spatio-temporal recognition*. Comput. Vis. Image Underst., vol. 108, n° 3, pages 207–229, DÃľcembre 2007.

[Laumond2005] Laumond, J., Ferré, E., Arechavaleta, G. et Esteves, C. *Mechanical part assembly planning with virtual mannequins*. In Assembly and Task Planning : From Nano to Macro Assembly and Manufacturing, 2005.(ISATP 2005). The 6th IEEE International Symposium on, pages 132–137. IEEE, 2005.

[Laxmi2002] Laxmi, V., Carter, J. et Damper, R. *Biologically-inspired human motion detection*. In 10th European Symposium on Artificial Neural Networks. Citeseer, 2002.

[Li2004] Li, L., Huang, W., Gu, I. et Tian, Q. *Statistical modeling of complex backgrounds for foreground object detection*. Image Processing, IEEE Transactions on, vol. 13, n° 11, pages 1459–1472, 2004.

[Li2005] Li, H. et Greenspan, M. *Continuous time-varying gesture segmentation by dynamic time warping of compound gesture models*. In HAREM 2005 : BMVC Workshop on Human Activity Recognition and Modelling, 2005.

[Liang2008] Liang, X., Zhang, S., Li, Q., Pronost, N., Geng, W. et Multon, F. *Intuitive motion retrieval with motion sensors*. In Proceedings of Computer Graphics International (CGI), Istambul - Turkey, jun 2008.

[Liang2009] Liang, X., Li, Q., Zhang, X., Zhang, S. et Geng, W. *Performance-driven motion choreographing with accelerometers*. Comput. Animat. Virtual Worlds, vol. 20, n° 2&dash ;3, pages 89–99, 2009.

[Liang2010] Liang, X., Hoyet, L., Geng, W. et Multon, F. *Responsive action generation by physically-based motion retrieval and adaptation*. In Proceedings of the Third international conference on Motion in games, MIG'10, pages 313–324, Utrecht, The Netherlands, 2010. Springer-Verlag.

[Lin2009] Lin, Z., Jiang, Z. et Davis, L. *Recognizing actions by shape-motion prototype trees*. In Computer Vision, 2009 IEEE 12th International Conference on, pages 444–451. IEEE, 2009.

[Liu2003] Liu, F., Zhuang, Y., Wu, F. et Pan, Y. *3D motion retrieval with motion index tree*. Computer Vision and Image Understanding, vol. 92, n° 2-3, pages 265–284, Novembre 2003.

[Liu2010] Liu, C. et Yuen, P. C. *Human action recognition using boosted EigenActions*. Image and Vision Computing, vol. 28, n° 5, pages 825 – 835, 2010. Best of Automatic Face and Gesture Recognition 2008.

[Lu2006] Lu, W.-L. et Little, J. *Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor*. In The 3rd Canadian Conference on Computer and Robot Vision, 2006. University of British Columbia, Canada, IEEE, June 2006.

[Lu2009] Lu, W., Li, W., Wang, L. et Pan, C. *Gestures Classification Based on Semantic Classification Tree*. In Image and Signal Processing, 2009. CISP'09. 2nd International Congress on, pages 1–5. IEEE, 2009.

[Lv2006] Lv, F. et Nevatia, R. *Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost*. In Lecture Notes in Computer Science, pages 359–372. Springer Berlin / Heidelberg, 2006.

[Mantyjarvi2001] Mantyjarvi, J., Himberg, J. et Seppanen, T. *Recognizing human motion with multiple acceleration sensors*. In Systems, Man, and Cybernetics, 2001 IEEE International Conference on, volume 2, pages 747–752. Ieee, 2001.

[Maraqa2012] Maraqa, M., Al-Zboun, F., Dhyabat, M. et Abu Zitar, R. *Recognition of Arabic Sign Language (ArSL) Using Recurrent Neural Networks*. Journal of Intelligent Learning Systems and Applications, vol. 4(1), pages 41–52, 2012.

[Marey1894] Marey, E. Le mouvement. Editions J. Chambon, 1894.

[Marr1982] Marr, D. et Vaina, L. *Representation and recognition of the movements of shapes*. Proceedings of the Royal Society of London. Series B. Biological Sciences, vol. 214, n° 1197, pages 501–524, 1982.

[Masoud2003] Masoud, O. et Papanikolopoulos, N. *A method for human action recognition*. Image and Vision Computing, vol. 21, n° 8, pages 729 – 743, 2003.

[Mathie2004] Mathie, M., Celler, B., Lovell, N. et Coster, A. *Classification of basic daily movements using a triaxial accelerometer*. Medical and Biological Engineering and Computing, vol. 42, n° 5, pages 679–687, Septembre 2004.

[Mayagoitia2002] Mayagoitia, R. E., Nene, A. V. et Veltink, P. H. *Accelerometer and rate gyroscope measurement of kinematics : an inexpensive alternative to optical motion analysis systems*. Journal of Biomechanics, vol. 35, n° 4, pages 537 – 542, 2002.

[Mcneill1992] Mcneill, D. Hand and mind : What gestures reveal about thought. University Of Chicago Press, August 1992.

[Meredith2000] Meredith, M. et Maddock, S. *Motion capture file formats explained*. Rapport technique, Department of Computer Science, University of Sheffield, 2000.

[Michoud2009] Michoud, B. *Reconstruction 3D à partir de séquences vidéo pour la capture de mouvements de personnages en temps réel et sans marqueur*. Thèse, Laboratoire d'InfoRmatique en Images et Systèmes d'information UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon, 2009.

[Müller2005] Müller, M., Röder, T. et Clausen, M. *Efficient content-based retrieval of motion capture data*. ACM Trans. Graph., vol. 24, n° 3, pages 677–685, 2005.

[Müller2006] Müller, M. et Röder, T. *Motion templates for automatic classification and retrieval of motion capture data*. In SCA '06 : Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation, pages 137–146, Vienna, Austria, 2006. Eurographics Association.

[Ménardais2003] Ménardais, S. *Fusion et adaptation temps réel de mouvements acquis pour l'animation d'humanoïdes synthétiques*. Thèse, Université Rennes 1 (IRISA), 2003.

[Ménardais2004] Ménardais, S., Kulpa, R., Multon, F. et Arnaldi, B. *Synchronization for dynamic blending of motions*. In Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation, SCA '04, pages 325–335, Grenoble, France, 2004. Eurographics Association.

[Moeslund2006] Moeslund, T. B., Hilton, A. et Krüger, V. *A survey of advances in vision-based human motion capture and analysis*. Computer Vision and Image Understanding, vol. 104, n° 2-3, pages 90 – 126, 2006. Special Issue on Modeling People : Vision-based understanding of a person's shape, appearance, movement and behaviour.

[Molet1999] Molet, T., Aubel, T., Gaspin, T., Carion, S. et Lee, E. *Anyone for tennis ?* Presence, vol. 8, n° 2, pages 140–156, 1999.

[Montepare1988] Montepare, J. et Zebrowitz-McArthur, L. *Impressions of people created by age-related qualities of their gaits*. Journal of Personality and Social Psychology, vol. 55, n° 4, page 547, 1988.

[Mori2006] Mori, A., Uchida, S., Ryo, K., Taniguchi, R.-i., Hasegawa, T. et Sakoe, H. *Early Recognition and Prediction of Gestures*. In Proceedings of the 18th International Conference on Pattern Recognition - Volume 03, ICPR '06, pages 560–563. IEEE Computer Society, 2006.

[Murakami1991] Murakami, K. et Taguchi, H. *Gesture recognition using recurrent neural networks*. In Proceedings of the SIGCHI conference on Human factors in computing systems : Reaching through technology, CHI '91, pages 237–242, New Orleans, Louisiana, United States, 1991. ACM.

[Murphy2002] Murphy, K. P. *Dynamic Bayesian Networks : Representation, Inference and Learning*. Thèse, University of California, Berkeley, 2002.

[Najafi2003] Najafi, B., Aminian, K., Paraschiv-Ionescu, A., Loew, F., Bula, C. et Robert, P. *Ambulatory system for human motion analysis using a kinematic sensor : monitoring of daily physical activity in the elderly*. Biomedical Engineering, IEEE Transactions on, vol. 50, n° 6, pages 711–723, 2003.

[Nakata2007] Nakata, T. *Temporal segmentation and recognition of body motion data based on inter-limb correlation analysis*. In Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on, pages 1383–1388. IEEE, 2007.

[Natarajan2008] Natarajan, P. et Nevatia, R. *Online, real-time tracking and recognition of human actions*. In Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on, pages 1–8. IEEE, 2008.

[Niebles2007] Niebles, J. et Fei-Fei, L. *A hierarchical model of shape and appearance for human action classification*. In IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, pages 1–8, 2007.

[Nielsen2004] Nielsen, M., Störring, M., Moeslund, T. B. et Granum, E. *A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI*. In Gesture-Based Communication in Human-Computer Interaction, volume 2915 of *Lecture Notes in Computer Science*, pages 105–106. Springer Berlin / Heidelberg, 2004.

[Ogawara2009] Ogawara, K., Tanabe, Y., Kurazume, R. et Hasegawa, T. *Detecting repeated motion patterns via dynamic programming using motion density*. In Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, pages 1743–1749. IEEE, 2009.

[Oliver2004] Oliver, N., Garg, A. et Horvitz, E. *Layered representations for learning and inferring office activity from multiple sensory channels*. Computer Vision and Image Understanding, vol. 96, n° 2, pages 163–180, 2004.

[Ott2008] Ott, C., Lee, D. et Nakamura, Y. *Motion capture based human motion recognition and imitation by direct marker control*. In Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pages 399–405. IEEE, 2008.

[Page2007] Page, A. et Epifanio, I. *A simple model to analyze the effectiveness of linear time normalization to reduce variability in human movement analysis*. Gait & posture, vol. 25, n° 1, pages 153–156, 2007.

[Paiyarom2009] Paiyarom, S., Tungamchit, P., Keinprasit, R. et Kayasith, P. *Activity monitoring system using dynamic time warping for the elderly and disabled people*. In Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on, pages 1–4. IEEE, 2009.

[Pandy2010] Pandy, M. et Andriacchi, T. *Muscle and joint function in human locomotion*. Annual review of biomedical engineering, vol. 12, pages 401–433, 2010.

[Park2011] Park, C.-B. et Lee, S.-W. *Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter*. Image and Vision Computing, vol. 29, n° 1, pages 51 – 63, 2011.

[Pavlovic1997] Pavlovic, V. I., Sharma, R. et Huang, T. S. *Visual Interpretation of Hand Gestures for Human-Computer Interaction : A Review*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, n° 7, pages 677–695, Juillet 1997.

[Peursum2005] Peursum, P., Bui, H. H., Venkatesh, S. et West, G. *Robust recognition and segmentation of human actions using HMMs with missing observations*. EURASIP J. Appl. Signal Process., vol. 2005, pages 2110–2126, Janvier 2005.

[Peursum2007] Peursum, P., Venkatesh, S. et West, G. *Tracking-as-recognition for articulated full-body human motion analysis*. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[Polana1994] Polana, R. et Nelson, R. *Low level recognition of human motion (or how to get your man without finding his body parts)*. In Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on, pages 77 – 82, 1994.

[Pollick2001] Pollick, F. E., Paterson, H. M., Bruderlin, A. et Sanford, A. J. *Perceiving affect from arm movement*. Cognition, vol. 82, n° 2, pages B51 – B61, 2001.

[Pollick2002] Pollick, F., Lestou, V., Ryu, J. et Cho, S. *Estimating the efficiency of recognizing gender and affect from biological motion*. Vision research, vol. 42, n° 20, pages 2345–2355, 2002.

[Poppe2010] Poppe, R. *A survey on vision-based human action recognition*. Image and Vision Computing, vol. 28, n° 6, pages 976 – 990, 2010.

[Quek1994] Quek, F. K. H. *Toward a vision-based hand gesture interface*. In VRST '94 : Proceedings of the conference on Virtual reality software and technology, pages 17–31, Singapore, Singapore, 1994. World Scientific Publishing Co., Inc.

[Ramadoss2008] Ramadoss, B. et Rajkumar, K. *Modeling the Dance Video Semantics using Regular Tree Automata*. Fundamenta Informaticae, vol. 86, n° 1, pages 175–189, 2008.

[Ramamoorthy2003] Ramamoorthy, A., Vaswani, N., Chaudhury, S. et Banerjee, S. *Recognition of dynamic hand gestures*. Pattern Recognition, vol. 36, n° 9, pages 2069 – 2081, 2003. Kernel and Subspace Methods for Computer Vision.

[Ramasso2007] Ramasso, E. *Reconnaissance de séquences d'états par le Modèle des Croyances Transférables. Application à l'analyse de vidéos d'athlé-*

tisme. Thèse, Université Joseph-Fourier - Grenoble I GIPSA-lab - Grenoble Images Parole Signal Automatique, december 2007.

[Ramstein1991] Ramstein, C. *Analyse, représentation et traitement du geste instrumental.* Thèse, IMAG - Institut d'Informatique et de Mathématiques Appliquées de Grenoble, 1991.

[Raptis2011] Raptis, M., Kirovski, D. et Hoppe, H. *Real-time classification of dance gestures from skeleton animation*. In Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '11, pages 147–156, Vancouver, British Columbia, Canada, 2011. ACM.

[Rashid2009] Rashid, O., Al-Hamadi, A. et Michaelis, B. *A framework for the integration of gesture and posture recognition using HMM and SVM*. In Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on, volume 4, pages 572–577. IEEE, 2009.

[Rekha2011] Rekha, J., Bhattacharya, J. et Majumder, S. *Shape, texture and local movement hand gesture features for Indian Sign Language recognition*. In Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference on, pages 30–35. IEEE, 2011.

[Ren2005] Ren, L., Patrick, A., Efros, A. A., Hodgins, J. K. et Rehg, J. M. *A data-driven approach to quantifying natural human motion*. ACM Trans. Graph., vol. 24, n° 3, pages 1090–1097, Juillet 2005.

[Ren2009] Ren, W., Singh, S., Singh, M. et Zhu, Y. *State-of-the-art on spatio-temporal information-based video retrieval*. Pattern Recognition, vol. 42, n° 2, pages 267 – 282, 2009. Learning Semantics from Multimedia Content.

[Reng2006] Reng, L., Moeslund, T. et Granum, E. *Finding Motion Primitives in Human Body Gestures*. In Gibet, S., Courty, N. et Kamp, J.-F., éditeurs, Gesture in Human-Computer Interaction and Simulation, volume 3881 of *Lecture Notes in Computer Science*, pages 133–144. Springer Berlin / Heidelberg, 2006.

[Romaszewski2011] Romaszewski, M. et Gomb, P. *The Effect of Multiple Training Sequences on HMM Classification of Motion Capture Gesture Data*. In Burduk, R., Kurzynski, M., Wozniak, M. et Zolnierek, A., éditeurs, Computer Recognition Systems 4, volume 95 of *Advances in Intelligent and Soft Computing*, pages 365–373. Springer Berlin / Heidelberg, 2011.

[Ronfard2009] Ronfard, R. *Analyse automatique de film - Des séquences d'images aux séquences d'actions.* Hdr, Université de Grenoble, Dec 2009.

[Runeson1981] Runeson, S. et Frykholm, G. *Visual perception of lifted weight*. Journal of Experimental Psychology : Human Perception and Performance, vol. 7, n° 4, page 733, 1981.

[Runeson1994] Runeson, S. *Perception of biological motion : The KSD-principle and the implications of a distal versus proximal approach*. Perceiving events and objects, pages 383–405, 1994.

[Sakaguchi1996] Sakaguchi, T., Kanamori, T., Katayose, H., Sato, K. et Inokuchi, S. *Human motion capture by integrating gyroscopes and accelerometers*. pages 470–475, 1996.

[Schindler2008] Schindler, K. et Van Gool, L. *Action snippets : How many frames does human action recognition require ?* In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[Schulz2010] Schulz, S. et Woerner, A. *Automatic motion segmentation for human motion synthesis*. In Proceedings of the 6th international conference on Articulated motion and deformable objects, AMDO'10, pages 182–191, Mallorca, Spain, 2010. Springer-Verlag.

[Sewell2011] Sewell, M. *Ensemble learning*, 2011.

[Shimada2010] Shimada, A., Kawashima, M. et Taniguchi, R.-i. *Early Recognition Based on Co-occurrence of Gesture Patterns*. In Wong, K., Mendis, B. et Bouzerdoum, A., éditeurs, Neural Information Processing. Models and Applications, volume 6444 of *Lecture Notes in Computer Science*, pages 431–438. Springer Berlin / Heidelberg, 2010.

[Shiratori2008] Shiratori, T. et Hodgins, J. K. *Accelerometer-based user interfaces for the control of a physically simulated character*. ACM Trans. Graph., vol. 27, n ° 5, pages 123 :1–123 :9, DÃľcembre 2008.

[Shotton2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. et Blake, A. *Real-time human pose recognition in parts from single depth images*. In CVPR, volume 2, page 3, 2011.

[Slyper2008] Slyper, R. et Hodgins, J. *Action Capture with Accelerometers*. In 2008 ACM SIGGRAPH / Eurographics Symposium on Computer Animation, pages 193–199, Juillet 2008.

[Sminchisescu2006] Sminchisescu, C., Kanaujia, A. et Metaxas, D. *Conditional models for contextual human motion recognition*. Comput. Vis. Image Underst., vol. 104, pages 210–220, November 2006.

[Sorel2009] Sorel, A., Nicolas, G., L'Hours, L., Prioux, J. et Quinton, P. *A light computing method for real-time activity recognition*. Computer Methods in Biomechanics and Biomedical Engineering, n ° 12, pages 231–232, 2009.

[Spriggs2009] Spriggs, E., De La Torre, F. et Hebert, M. *Temporal segmentation and activity classification from first-person sensing*. In Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, pages 17 – 24, 2009.

[Stephan2010] Stephan, J. J. et Khudayer, S. *Gesture Recognition for Human-Computer Interaction (HCI)*. International Journal of Advancements in Computing Technology, vol. 2(4), pages 30–35, 2010.

[Stevenage1999] Stevenage, S., Nixon, M. et Vince, K. *Visual analysis of gait as a cue to identity*. Applied Cognitive Psychology, vol. 13, n ° 6, pages 513–526, 1999.

[Stoll1995] Stoll, P. et Ohya, J. *Applications of HMM modeling to recognizing human gestures in image sequences for a man-machine interface*. In Robot and Human Communication, 1995. RO-MAN'95 TOKYO, Proceedings., 4th IEEE International Workshop on, pages 129–134. IEEE, 1995.

[Sukthankar2005] Sukthankar, G. et Sycara, K. *A cost minimization approach to human behavior recognition*. In AAMAS '05 : Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, pages 1067–1074, The Netherlands, 2005. ACM.

[Suthanthira Vanitha2006] Suthanthira Vanitha, N., Mani, M. et Palanisamy, V. *Magnetic and Inertial Orientation Tracking for Inserting Humans into Networked Synthetic Environment*. International Journal of Soft Computing, vol. 1, n° 4, pages 271–278, 2006.

[Sutton2006] Sutton, C. et Mccallum, A. *An Introduction to Conditional Random Fields for Relational Learning*. In Getoor, L. et Taskar, B., éditeurs, Introduction to Statistical Relational Learning. MIT Press, 2006.

[Taniguchi2011] Taniguchi, T., Hamahata, K. et Iwahashi, N. *Unsupervised Segmentation of Human Motion Data Using a Sticky Hierarchical Dirichlet Process-Hidden Markov Model and Minimal Description Length-Based Chunking Method for Imitation Learning*. Advanced Robotics, vol. 25, n° 17, pages 2143–2172, 2011.

[Tenenbaum2000] Tenenbaum, J. B., Silva, V. d. et Langford, J. C. *A Global Geometric Framework for Nonlinear Dimensionality Reduction*. Science, vol. 290, n° 5500, pages 2319–2323, 2000.

[Thome2008] Thome, N., Merad, D. et Miguet, S. *Learning articulated appearance models for tracking humans : A spectral graph matching approach*. Signal Processing : Image Communication, vol. 23, n° 10, pages 769 – 787, 2008.

[Tormene2009] Tormene, P., Giorgino, T., Quaglini, S. et Stefanelli, M. *Matching incomplete time series with dynamic time warping : an algorithm and an application to post-stroke rehabilitation*. Artificial Intelligence in Medicine, vol. 45, n° 1, pages 11 – 34, 2009.

[Tournier2011] Tournier, M. *Réduction de dimension pour l'animation de personnages*. Thèse, INRIA Grenoble Rhône-Alpes, 2011.

[Tran2010] Tran, K., Kakadiaris, I. A. et Shah, S. K. *Fusion of Human Posture Features for Continuous Action Recognition*. In International Workshop on Sign Gesture Activity, 2010.

[Troje2002] Troje, N. *Decomposing biological motion : A framework for analysis and synthesis of human gait patterns*. Journal of vision, vol. 2, n° 5, pages 371–387, 2002.

[Turaga2008] Turaga, P., Chellappa, R., Subrahmanian, V. et Udrea, O. *Machine recognition of human activities : A survey*. Circuits and Systems for Video Technology, IEEE Transactions on, vol. 18, n° 11, pages 1473–1488, 2008.

[Vcelak2006] Vcelak, J., Ripka, P., Platil, A., Kubik, J. et Kaspar, P. *Errors of AMR compass and methods of their compensation*. Sensors and Actuators A : Physical, vol. 129, n° 1-2, pages 53 – 57, 2006. EMSA 2004 - Selected Papers from the 5th European Magnetic Sensors & Actuators Conference - EMSA 2004, Cardiff, UK, 4-6 July 2004.

[Vicente2007a] Vicente, I. S., Kyrki, V., Kragic, D. et Larsson, M. *Action recognition and understanding through motor primitives*. Advanced Robotics, vol. 21, pages 1687–1707(21), 2007.

[Vicente2007b] Vicente, S., Kragic, D. et Eklundh, J. *Learning and recognition of object manipulation actions using linear and nonlinear dimensionality reduction*. In Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, pages 1010–1015. IEEE, 2007.

[Vignais2009] Vignais, N. *Mise en oeuvre et évaluation d'une méthodologie fondée sur la réalité virtuelle pour l'analyse de la prise d'informations visuelles du gardien de but de handball*. Thèse, Université Rennes 2, 2009.

[Walk1984] Walk, R. et Homan, C. *Emotion and dance in dynamic light displays.* Bulletin of the Psychonomic Society ; Bulletin of the Psychonomic Society, 1984.

[Wang1998] Wang, X., Maurin, M., Mazet, F., Maia, N. D. C., Voinot, K., Verriest, J. P. et Fayet, M. *Three-dimensional modelling of the motion range of axial rotation of the upper arm*. Journal of Biomechanics, vol. 31, n° 10, pages 899 – 908, 1998.

[Wang2006] Wang, S. B., Quattoni, A., Morency, L.-P. et Demirdjian, D. *Hidden Conditional Random Fields for Gesture Recognition*. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 1521–1527. IEEE Computer Society, 2006.

[Wang2007] Wang, L. et Suter, D. *Recognizing human activities from silhouettes : Motion subspace and factorial discriminative graphical model*. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[Wang2008] Wang, J., Fleet, D. et Hertzmann, A. *Gaussian process dynamical models for human motion*. IEEE transactions on pattern analysis and machine intelligence, vol. 30, n° 2, pages 283–298, 2008.

[Weinland2011] Weinland, D., Ronfard, R. et Boyer, E. *A survey of vision-based methods for action representation, segmentation and recognition*. Comput. Vis. Image Underst., vol. 115, pages 224–241, February 2011.

[Welch2001] Welch, G. et Bishop, G. *An Introduction to the Kalman Filter*. In Proc of SIGGRAPH, Course, volume 8, page 81, 2001.

[Wexelblat1998] Wexelblat, A. *Research challenges in gesture : Open issues and unsolved problems*. In Wachsmuth, I. et Fröhlich, M., éditeurs, Gesture and Sign Language in Human-Computer Interaction, volume 1371 of *Lecture Notes in Computer Science*, pages 1–11. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0052984.

[Wilson1998] Wilson, A. D. et Bobick, A. F. *Recognition and Interpretation of Parametric Gesture*. In Proceedings of the Sixth International Conference on Computer Vision, ICCV '98, pages 329–. IEEE Computer Society, 1998.

[Wilson2002] Wilson, A. D. et Bobick, A. F. Hidden markov models for modeling and recognizing gesture under variation, pages 123–160. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[Wollmer2009] Wollmer, M., Al-Hames, M., Eyben, F., Schuller, B. et Rigoll, G. *A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams*. Neurocomputing, vol. 73, n° 1-3, pages 366–380, 2009.

[Wu2001] Wu, H. et Sutherland, A. *Dynamic gesture recognition using PCA with multi-scale theory and HMM*. Proc. SPIE, vol. 4550, pages 132–139, 2001.

[Wu2005] Wu, G., Van der Helm, F., Veeger, H., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A. et McQuade, K. *ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion–Part II : shoulder, elbow, wrist and hand*. Journal of Biomechanics, vol. 38, n° 5, pages 981–992, 2005.

[Wu2009] Wu, S., Xia, S., Wang, Z. et Li, C. *Efficient motion data indexing and retrieval with local similarity measure of motion strings*. The Visual Computer, vol. 25, pages 499–508, 2009.

[Xi2006] Xi, X., Keogh, E., Shelton, C., Wei, L. et Ratanamahatana, C. A. *Fast time series classification using numerosity reduction*. In ICML '06 : Proceedings of the 23rd international conference on Machine learning, pages 1033–1040, Pittsburgh, Pennsylvania, 2006. ACM.

[Xiang2006] Xiang, J., Weng, J.-g., Zhuang, Y.-t. et Wu, F. *Ensemble learning HMM for motion recognition and retrieval by Isomap dimension reduction*. Journal of Zhejiang University - Science A, vol. 7, pages 2063–2072, 2006.

[Yamato1992] Yamato, J., Ohya, J. et Ishii, K. *Recognizing human action in time-sequential images using hidden Markov model*. In Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on, pages 379–385. IEEE, 1992.

[Yang1999] Yang, M. et Ahuja, N. *Recognizing hand gesture using motion trajectories*. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., volume 1, 1999.

[Yang2006] Yang, H.-D., Park, A.-Y. et Lee, S.-W. *Robust Spotting of Key Gestures from Whole Body Motion Sequence*. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, FGR '06, pages 231–236. IEEE Computer Society, 2006.

[Yang2008] Yang, J.-Y., Wang, J.-S. et Chen, Y.-P. *Using acceleration measurements for activity recognition : An effective learning algorithm for constructing neural classifiers*. Pattern Recognition Letters, vol. 29, n° 16, pages 2213 – 2220, 2008.

[Yao2012] Yao, A., Gall, J. et Van Gool, L. *Coupled Action Recognition and Pose Estimation from Multiple Views*. International Journal of Computer Vision, vol. 99, pages 1–22, 2012.

[Yeasin2000] Yeasin, M. et Chaudhuri, S. *Visual understanding of dynamic hand gestures*. Pattern Recognition, vol. 33, n° 11, pages 1805 – 1817, 2000.

[Yoon2001] Yoon, H.-S., Soh, J., Bae, Y. J. et Yang, H. S. *Hand gesture recognition using combined features of location, angle and velocity*. Pattern Recognition, vol. 34, n° 7, pages 1491 – 1501, 2001.

[Yu2005] Yu, T., Shen, X., Li, Q. et Geng, W. *Motion retrieval based on movement notation language*. Computer Animation and Virtual Worlds, vol. 16, n° 3-4, pages 273–282, 2005.

[Yu2010] Yu, T., Kim, T. et Cipolla, R. *Real-time action recognition by spatiotemporal semantic and structural forests*. In Proceedings of the British Machine Vision Conference, page 56. Citeseer, 2010.

[Zhang2010] Zhang, H., Liu, Z., Zhao, H. et Cheng, G. *Recognizing human activities by key frame in video sequences*. Journal of Software, vol. 5, n° 8, pages 818–825, 2010.

[Zhou2004] Zhou, S., Chellappa, R. et Moghaddam, B. *Visual tracking and recognition using appearance-adaptive models in particle filters*. Image Processing, IEEE Transactions on, vol. 13, n° 11, pages 1491–1506, 2004.

[Zhou2008a] Zhou, F., la Torre, F. D. et Hodgins, J. K. *Aligned Cluster Analysis for temporal segmentation of human motion.* In FG, pages 1–7. IEEE, 2008.

[Zhou2008b] Zhou, H. et Hu, H. *Human motion tracking for rehabilitation–A survey*. Biomedical Signal Processing and Control, vol. 3, n° 1, pages 1 – 18, 2008.

Table des figures

Liste des tableaux

Publications liées à la thèse

Les travaux liés à cette thèse ont donné lieu aux publications suivantes :

- Sorel A., Kulpa R., Badier E. et Multon F. *Dealing with variability when recognizing user's performance in natural gesture interfaces*. International Journal of Pattern Recognition and Artificial Intelligence, soumis.
- Sorel A., Kulpa R., Badier E. et Multon F. *Dealing with Variability When Recognizing User's Performance in Natural Gesture Interfaces*. In : M. Kallmann and K. Bekris (Eds.) : MIG 2012, LNCS 7660, Springer, pp. 370-373, 2012.
- Sorel A., Nicolas G., L'Hours L., Prioux J. et Quinton P. *A light computing method for real-time activity recognition*. Computer Methods in Biomechanics and Biomedical Engineering, n° 12, pages 231-232, 2009.

Ces travaux ont également donné lieu à la réalisation d'un court-métrage de vulgarisation scientifique intitulé « Allô docteur, sur quel pied danser ? », dans le cadre de l'édition 2012 du festival *Sciences en cour[t]s*[1].

1. www.sciences-en-courts.fr

Gestion de la variabilité morphologique pour la reconnaissance de gestes naturels à partir de données 3D

La reconnaissance de mouvements naturels est de toute première importance dans la mise en œuvre d'Interfaces Homme-Machine intelligentes et efficaces, utilisables de manière intuitive en environnement virtuel. En effet, elle permet à l'utilisateur d'agir de manière naturelle et au système de reconnaitre les mouvements corporel effectués tels qu'ils seraient perçu par un humain. Cette tâche est complexe, car elle demande de relever plusieurs défis : prendre en compte les spécificités du dispositif d'acquisition des données de mouvement, gérer la variabilité cinématique dans l'exécution du mouvement, et enfin gérer les différences morphologiques inter-individuelles, de sorte que les mouvements de tout nouvel utilisateur puissent être reconnus. De plus, de part la nature interactive des environnements virtuels, cette reconnaissance doit pouvoir se faire en temps-réel, sans devoir attendre la fin du mouvement. La littérature scientifique propose de nombreuses méthodes pour répondre aux deux premiers défis mais la gestion de la variabilité morphologique est peu abordée. Dans cette thèse, nous proposons une description du mouvement permettant de répondre à cette problématique et évaluons sa capacité à reconnaitre les mouvements naturels d'un utilisateur inconnu. Enfin, nous proposons une nouvelle méthode permettant de tirer partie de cette représentation dans une reconnaissance précoce du mouvement.

Mots clés : reconnaissance automatique, mouvements naturels, variabilité morphologique, vecteur descripteur, modèles de Markov cachés (HMM).

Addressing morphological variability for natural gesture recognition from 3D data

Recognition of natural movements is of utmost importance in the implementation of intelligent and effective Human-Machine Interfaces for virtual environments. It allows the user to behave naturally and the system to recognize its body movements in the same way a human might perceive it. This task is complex, because it addresses several challenges : take account of the specificities of the motion capture system, manage kinematic variability in motion performance, and finally take account of the morphological differences between individuals, so that actions of any new user can be recognized. Moreover, due to the interactive nature of virtual environments, this recognition must be achieved in real-time without waiting for the motion end. The literature offers many methods to meet the first two challenges. But the management of the morphological variability is not dealt. In this thesis, we propose a description of the movement to address this issue and we evaluate its ability to recognize the movements of an unknown user. Finally, we propose a new method to take advantage of this representation in early motion recognition.

Keywords : automatic recognition, natural gesture, morphological variability, feature vector, hidden Markov model (HMM).