Guidage Gestuel pour des Robots Mobiles

Florent Taralle

▶ **To cite this version:**

Florent Taralle. Guidage Gestuel pour des Robots Mobiles. Interface homme-machine [cs.HC]. Université Paris sciences et lettres, 2016. Français. NNT : 2016PSLEM096 . tel-01814263

HAL Id: tel-01814263
https://pastel.hal.science/tel-01814263

Submitted on 13 Jun 2018

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

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

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à MINES ParisTech

Guidage Gestuel pour des Robots Mobiles

École doctorale n°432

SCIENCES DES MÉTIERS DE L'INGÉNIEUR

Spécialité INFORMATIQUE TEMPS-RÉEL, ROBOTIQUE ET MATHÉMATIQUE

COMPOSITION DU JURY :

M Frédéric FOL LEYMARIE
Goldsmiths College, University of London
Président

Mme Indira THOUVENIN
UTC Sorbonne Universites
Rapporteur

M Pierre DE LOOR
Ecole Nationale d'Ingénieurs de Brest
Rapporteur

M Philippe FUCHS
Mines ParisTech
Examinateur, Directeur de thèse

M Alexis PALJIC
Mines ParisTech
Examinateur, Maître de thèse

M Christophe GUETTIER
Safran Electronics Defense
Examinateur, Maître de thèse

M Sotiris MANITSARIS
Mines ParisTech
Examinateur, Maître de thèse

Soutenue par **Florent Taralle**
le 9 novembre 2016

Dirigée par **Philippe Fuchs**
Codirigée par **Alexis Paljic**

MINES ParisTech | PSL RESEARCH UNIVERSITY PARIS

Table des matières

CHAPITRE 1

Introduction

Sommaire

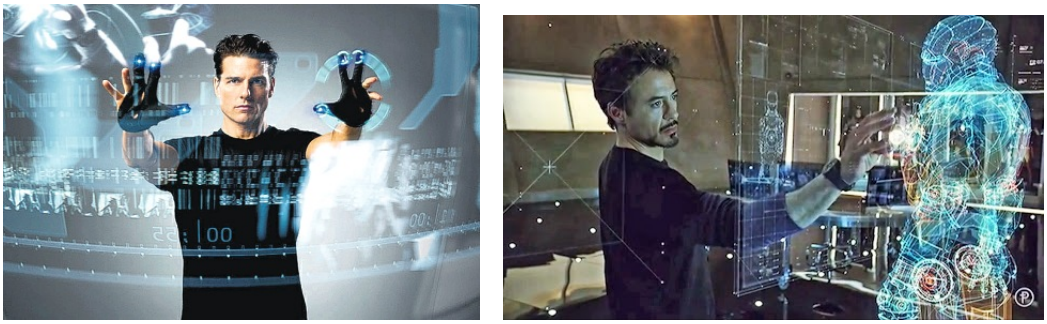
1.1 Gestes et interaction gestuelle

1.1.1 Le geste pour des interfaces naturelles et intuitives

Le geste est une modalité qui, par le simple mouvement du corps humain, permet d'agir et de communiquer. Son apparente simplicité et son universalité (Morris & Ochs 1997) pourraient constituer un formidable moyen pour interagir avec les machines.

Aussi, depuis les années 1970 qui marquent le début de l'*étude moderne* du geste (Kendon 2007), deux domaines scientifiques s'y intéressent particulièrement : d'une part les sciences cognitives qui cherchent à caractériser le geste et y voient un support permettant d'étudier les fondements du langage humain ; et d'autre part, l'informatique qui cherche à doter la machine de la capacité à comprendre les gestes humains et parfois à les utiliser.

Aujourd'hui, aucune définition généraliste et consensuelle ne semble avoir été formulée quant à la notion même de geste. Cependant, il est évident qu'il en existe de différentes sortes. En informatique par exemple, les gestes *manipulatifs* sont en premier lieux opposés



(a) (b)

FIGURE 1.1 – Deux exemples d'interfaces gestuelles imaginaires proposées dans des oeuvres de science-fiction : (a) dans *Minority Report* réalisé par Steven Spielberg en 2002 et (b) dans *Iron Man* réalisé par Jon Favreau en 2008.

aux gestes *communicatifs* (Quek *et al.* 2002). Alors que les premiers sont une manière d'agir physiquement sur l'environnement par l'application de forces mécaniques, les seconds sont un mode d'expression qui permet de transmettre des messages et des intentions.

Des distinctions plus fines peuvent également être faites dans le cas des gestes communicatifs. En effet, selon leur contexte d'emploi, il est aujourd'hui courant de distinguer les gestes *co-verbaux* qui sont les mouvements associés à un énoncé oral, les gestes *sémaphoriques* qui sont des mouvements signifiants standardisés et utilisés seuls, et les *langues des signes* qui sont des systèmes linguistiques complets (Kendon 1988a; McNeill 1992). Ainsi, il ne semble pas possible d'illustrer simplement la distinction entre ces trois types de gestes (avec des images par exemple) puisqu'il ne s'agit pas ici d'une différence de forme, ni de sémantique, mais bien d'une différence de contexte. Le même mouvement portant le même sens pourra aussi bien être un co-verbal, un sémaphorique ou un signe ; tout dépend de comment il est employé.

Ainsi, le geste est une modalité très riche qui, sous différentes formes, pourrait permettre une interaction naturelle et intuitive avec les machines.

Cette vision tend par ailleurs à apparaître dans l'imaginaire collectif ; notamment grâce à la science fiction comme avec les films *Minority Report* réalisé par Steven Spielberg en 2002 et *Iron Man* réalisé par Jon Favreau en 2008 dont les interfaces gestuelles imaginaires sont illustrées en Figure 1.1 ; et à l'apparition de nombreux capteurs de mouvement bas coûts comme par exemple la *Kinect* développée par *Microsoft* en 2008, le *Leap-Motion* développé par la société du même nom en 2012, ou le *Myo* plus récemment produit par la start-up canadienne Thalmic Labs.

1.1.2 Différents niveaux de maturité

L'idée de l'interaction gestuelle devient donc commune, mais qu'en est-il concrètement de son usage ? Il apparaît dans les faits, alors que les gestes manipulatifs se démocratisent, que les gestes communicatifs semblent bien moins utilisés.

En effet, l'usage des gestes manipulatifs se démocratise pour les environnement virtuels, pour certains jeux vidéos, et devient même un standard pour les interfaces mobiles avec l'utilisation d'interfaces tactiles. Ce nouveau mode d'interaction s'est révélé particulièrement simple d'utilisation au point d'être accessible aux enfants en bas âge, et extrêmement efficace si bien que refaire l'expérience d'une interaction non tactile peut se révéler frustrant. Le succès de ce type d'interaction et l'évolution des technologies conduit progressivement à étendre leur usage et notamment de passer d'une manipulation 2D à une manipulation 3D (Widmer *et al.* 2014). Cependant, le phénomène de fatigue et l'adéquation entre une interaction gestuelle 3D pour un écran 2D restent encore des points de discussion.

Contrairement aux gestes manipulatifs, les gestes communicatifs restent quant à eux peu représentés. Alors que la reconnaissance et la génération automatique des langues des signes constituerait une aide significative pour la communauté des sourds et muets, la complexité technique semble demeurer un obstacle. En effet, la taille très importante du vocabulaire (quelques 6000 signes pour la langue des signes américaine), la rapidité et la variabilité de l'exécution sont des contraintes fortes.

Pour les gestes co-verbaux, leur reconnaissance et leur interprétation pourrait permettre une interaction multi-modale plus riche et plus naturelle. En effet il permettraient de compléter les informations exprimées par la parole et parfois d'aider à lever certaines ambiguïtés linguistiques. A la suite de l'article de référence *Put that there* (Bolt 1980), de nombreux travaux ont été conduits. Cependant, la complexité technique semble à nouveau un frein. Comme pour les langues des signes, la rapidité et variabilité de l'exécution sont problématiques. A cela s'ajoute également le besoin de synchroniser les informations vocales et gestuelles et une absence totale de vocabulaire standardisé.

Enfin, les gestes sémaphoriques pourraient être utilisés pour activer à distance des fonctionnalités automatiques d'un système : que ce soit dans le cadre de l'interaction avec un robot, d'un environnement intelligent ou d'un ordinateur. Ce type de geste est souvent considéré comme techniquement plus simple que les gestes co-verbaux ou les langues des signes (Karam 2006). En effet, il s'agit en général de simplement détecter et reconnaître un nombre réduit de gestes standardisés et temporellement isolés pour activer des fonctions autonomes prédéfinies. Alors que de nombreux travaux portent sur ce type d'interaction gestuelle, à nouveau, très peu d'applications réelles existent. Mais qu'est ce qui pour ce troisième type de geste communicatif explique son faible usage ? Deux éléments semblent à considérer :

— Tout d'abord, de manière théorique, ce type de geste correspond-t-il à un besoin réel ? En effet, certains auteurs dénoncent le caractère artificiel d'une interaction symbolique ; qu'ils ne sont ni naturels ni intuitifs tant ils représentent en réalité un pourcentage minuscule de la communication non-verbale humaine (Quek *et al.* 2002). Ainsi, quel apport pourrait avoir une telle modalité en comparaison d'interfaces classiques telles que les claviers et les joysticks (Wexelblat 1997) ?

— Ensuite, de manière pratique, quel est le niveau de maturité technologique ? Pour les utilisateurs, le niveau de robustesse est-il suffisant ; et, pour les concepteurs, les technologies sont-elles suffisamment accessibles : simples à mettre en oeuvre ?

Aussi, il peut sembler que seule l'interaction avec des gestes manipulatifs soit aujour-

d'hui possible. Cela est-il vrai et faut-il s'en contenter en attendant les prochaines évolutions en matière de reconnaissance des gestes communicatifs ? Nous pensons que non.

1.1.3 Hypothèse quant à l'interaction gestuelle sémaphorique

> If users must make one fixed gesture to, for example, move forward in a system then stop, then make another gesture to move backward, I find myself wondering why the system designers bother with gesture in the first place. Why not simply give the person keys to press : one for forward and one for backward ? [...] One of the major points of gesture modes of operation is their naturalness. If you take away that advantage, it is hard to see why the user benefits from a gestural interface at all.
>
> *Alan Wexelblat* (Wexelblat 1997)

> Results also suggest that in a desktop scenario, traditional input methods are more appropriate than gestures.
>
> *Maria Karam* (Karam *et al.* 2006)

En effet, le caractère naturel et intuitif de certains gestes est intéressant ; et cela constitue certainement l'argument principal en faveur de l'interaction gestuelle. Certes, il est raisonnable de penser que tous les types de gestes ne sont pas naturels ; notamment les gestes sémaphoriques. Il apparaît également que la modalité gestuelle ne convient pas nécessairement à toutes les applications.

Cependant, nous pensons que le geste est bien plus riche que cela ; qu'il possède selon les types, d'autres caractéristiques qui peuvent être tour à tour des avantages ou des inconvénients. Et, que c'est de ces caractéristiques que peut dépendre l'adéquation ou l'incompatibilité avec un besoin spécifique.

Aussi, en souhaitant dépasser cette limitation conceptuelle, nous nous sommes intéressé aux gestes sémaphoriques : Puisque techniquement ils semblent aujourd'hui les plus accessibles, quelles pourraient être leurs particularités et pour quelle application auraient-ils un avantage ?

Nous avons fait l'hypothèse que les gestes sémaphoriques sans support physique , donc en faisant abstraction des gestes de tracé (*strokes* (Karam 2006)), permettent une interaction expressive, rapide, et surtout qui requiert peu d'attention.

Ainsi, nous pensons que ce type de geste est particulièrement adapté aux applications qui nécessitent une interaction tout en maintenant un certain niveau de conscience de l'environnement. Et cela nous semble convenir au cas de la commande des drones militaires de contact sur le champs de bataille.

1.2 Cas d'application : la commande des drones militaires de contact

1.2.1 Pourquoi des robots militaires ?

Au cours de l'histoire, les stratégies militaires ont évolué. Selon certains auteurs (Gazette 1989), l'histoire des stratégies militaires peut être découpée en quatre grandes phases également appelées générations :

1. La première génération qui a connu sont apogée lors des campagnes napoléoniennes, consistait en un ensemble de tactiques de lignes et de colonnes pour organiser les fantassins.
2. Puis, la seconde génération a été marquée par un usage massif de l'artillerie. Celle-ci a brisé les lignes statiques de fantassins comme lors de la *Première guerre mondiale*.
3. La *Seconde guerre mondiale* a quant à elle été particulièrement représentative de la troisième génération. Les combats n'ont plus reposé sur des positions statiques et l'usure, mais sur des manoeuvres rapides qui ont permis de contourner les lignes ennemies pour en anéantir la puissance de feu.
4. Enfin, la quatrième génération est apparue pendant la *Guerre Froide* et la décolonisation. Dès lors, des puissances militaires conventionnelles font face à des acteurs non-étatiques.

Bien que la notion de guerre de quatrième génération soit controversée (Gray 2012), il ne fait aucun doute que les combats revêtent un visage nouveau. Dans ces conflits dits asymétriques, les guérillas et le terrorisme sont de nouveaux modes d'action. L'environnement est chaotique : il est piégé et évolue rapidement.

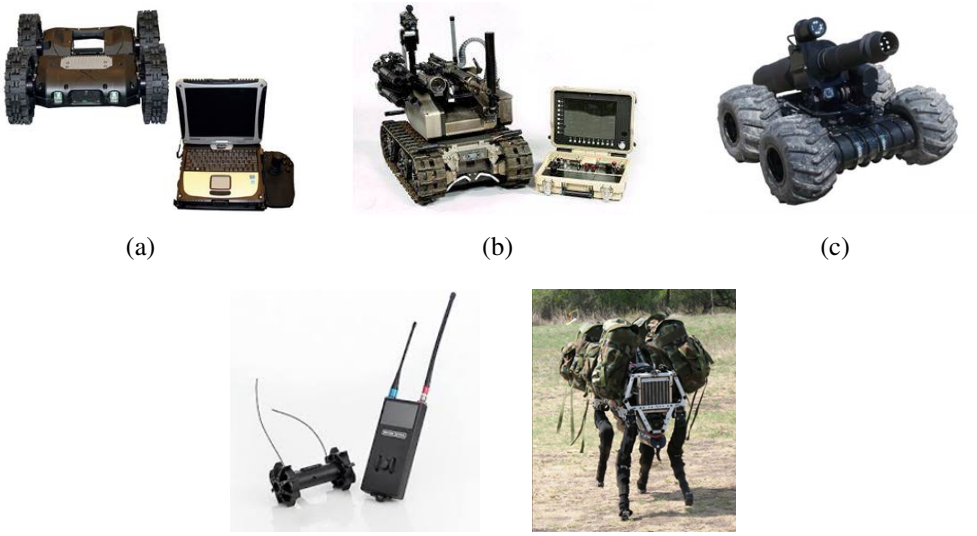
Y déployer des forces ne peut donc plus se faire sans information précise. Or, recueillir de l'information requiert nécessairement une présence sur le terrain. Présence qui, si elle est humaine, conduit à un paradoxe sécuritaire évident.

Cependant, les nouvelles technologies semblent apporter une solution : la robotique. En effet, les robots mobiles permettent d'apporter une présence sur le terrain pour acquérir de la donnée sans toute fois directement compromettre la sécurité des hommes.

Dans ce contexte d'environnement fortement hostile, la robotique constitue donc un atout majeur. Aussi, de nombreux états ont entrepris d'améliorer les outils en dotation de leurs forces armées. Le programme américain *Future Combat Systems* (FCS) ou le programme français *Fantassin à Équipements et Liaisons INtégrés* (FELIN) en sont des exemples.

1.2.2 Qu'est ce qu'un robot mobile ?

Les robots militaires sont aujourd'hui des machines spécialisées et leur usage est par conséquent limité à des missions spécifiques (Marec 2013). Ainsi, il en existe différents types pour différents environnements (Martinic 2014).



(a) (b) (c)

(d) (e)

FIGURE 1.2 – Quelques exemples d'UGV militaires : (a) Le Nerva LG est un robot léger de la société Nexter Robotics ; (b) le MAARS est un robot modulaire dévelopé par Qinetiq ; (c) le Cobra MK2 est un mini-UGV du groupe ECA ; (d) le Scout XT est un micro-UGV de la société ReconRobotics ; et (e) le BigDog est un robot mule quadrupède développé par Boston Dynamics .

En mer, les navires de surface sans équipage (USV) ou les véhicules sous-marins sans équipage (UUV) jouent un rôle important dans la lute contre les mines marines ou sous-marines. En détectant et détruisant ces engins explosifs à distance, la sécurité des marins et de leurs bâtiments est augmentée (Carreiro & Burke 2006).

Sur terre, les véhicules sans pilote (UGV) permettent de recueillir de l'information dans des bâtiments, de détecter et manipuler à distance des engins explosifs improvisés (IED), ou encore d'apporter de la puissance de feu déportée. Les robots mules, quant à eux, apportent un soutien logistique en transportant du matériel ou des blessés. Quelques modèles d'UGV sont présentés en Figure 1.2.

Enfin, dans les airs, les véhicules sans pilote (UAV) sont particulièrement utilisés. Ils sont également connus sous l'appellation d'origine anglaise *drone* par comparaison à l'insecte lent et bruyant du même nom. Quelques modèles d'UAV sont présentés en Figure 1.3.

1.2.3 Les différents types de drones

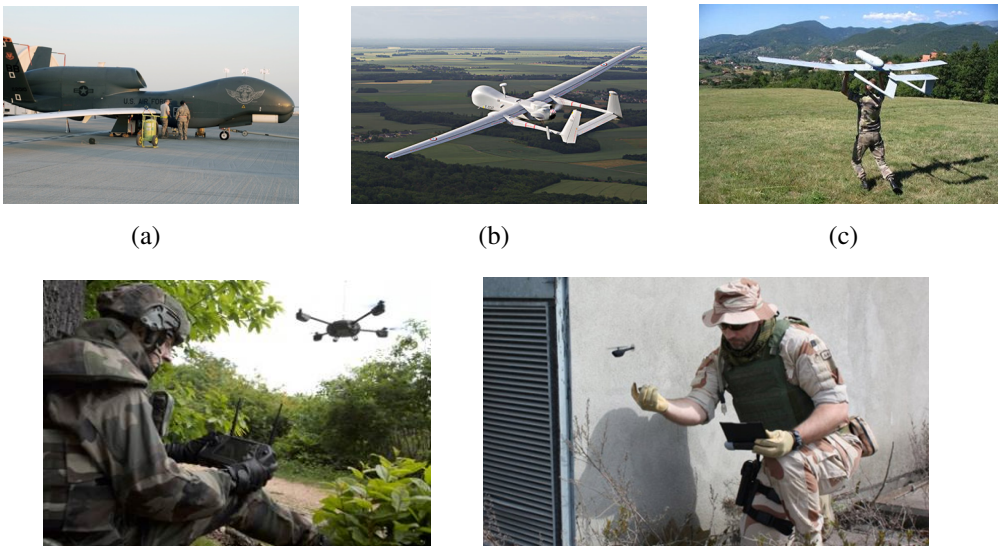
Il existe différents types de drones adaptés aux besoins de missions spécifiques. Au regard d'une classification proposée par l'OTAN (Organisation du traité de l'Atlantique nord) la famille des drones militaires est subdivisée en trois catégories :

— **Classe I (jusqu'à 150 kg)** : Les drones de contact permettent aux unités au contact de l'adversaire de recueillir du renseignement de proximité, direct et immédiat, en

terrain ouvert ou en zone urbaine. Ils sont principalement destinés aux troupes de mêlée ou aux forces de sécurité Ces drones sont en général légers, peu encombrants, et relativement simples à mettre en œuvre. Suivant le poids on pourra parler de mini-drone, de micro-drone ou de nano-drone.

— **Classe II (150 à 600 kg)** : Les drones tactiques servent lors de missions de renseignement et d'acquisition de cibles avec une portée d'une centaine de kilomètres. En général, ils privilégient des modes de lancement et de récupération ne nécessitant pas d'infrastructures fixes.

— **Classe III (plus de 600 kg)** : Les drones de grande endurance et à liaison par satellite sont généralement mis en oeuvre à partir d'infrastructures fixes. Les drones de classe III peuvent être dits *de théatre* pour les drones à moyenne altitude et longue endurance (MALE), *stratégiques* pour les drones à très haute altitude avec une grande endurance (HALE), ou *d'attaque* lorsqu'ils sont armés pour des missions offensives ou défensives.



(a) (b) (c)

(d) (e)

FIGURE 1.3 – Quelques exemples d'UAV militaires : (a) Le Global Hawk est un drone HALE (haute altitude longue endurance) construit par Northrop Grumman pour l'US Air Force ; (b) Le Harfang est un drone MALE (moyenne altitude longue endurance) produit par EADS pour l'armée de l'air française ; (c) Le DRAC est un drone de reconnaissance léger fabriqué par EADS ; (d) Le Nx110m, micro-drone de reconnaissance construit par Novadem, a été évalué par les forces françaises ; (e) Le Black Hornet, nano-drone construit par Proxdynamics, a été utilisé en Afghanistan par des soldats britanniques .

1.2.4 Méthodes de pilotage : téléopération et supervision

Aujourd'hui, pour commander ces différents types de robots, deux méthodes peuvent être utilisées. Souvent complémentaires, ces deux méthodes sont la téléopération et la supervision.

1.2.4.1 La téléopération

Bien que l'automatisation des robots évolue vite, l'intelligence artificielle de ces derniers reste encore assez limitée. Et ce, au regard des besoins et risques inhérents au domaine militaire. Aussi, les robots militaires sont-ils encore majoritairement télé-opérés (Fong & Thorpe 2001; Vertut 2013; Martinic 2014).

En télé-opération, le pilote dirige le robot à distance grâce à un joystick couplé à un retour visuel provenant de caméras embarquées. Souvent, l'objectif est de permettre à l'opérateur de se sentir comme présent à l'intérieur du robot qu'il pilote (*inside-out piloting*). L'effet de télé-présence peut être augmenté en créant un poste de pilotage simulant un cockpit d'avion (Canan 1999) ou en asservissant l'orientation des caméras sur l'orientation de la tête du pilote (Gage 1995).

La téléopération est utilisée avec succès pour les drones de classe III (de théâtre, stratégiques ou d'attaque). Protégés dans un bâtiment ou dans un véhicule, les opérateurs peuvent être totalement dédiés au pilotage. Pour une efficacité maximale, le pilotage est souvent réalisé par deux opérateurs spécialisés : un pilote gère le déplacement du robot pendant qu'un analyste oriente la caméra et recherche les menaces en temps réel dans le flux vidéo. Un exemple de poste de pilotage de drone de classe III est présenté Figure 1.4(a).

Ce mode de pilotage est également souvent utilisé à courte portée par des opérateurs sur le terrain. Ils utilisent alors des outils de pilotage plus simples et moins encombrants. Deux exemples d'interfaces de drones de contact sont présentées Figure 1.4(c) et 1.4(d).

En résumé, ce type de pilotage est particulièrement adapté au contrôle précis et temps réel. Une bande passante très importante, de très faibles délais de communication et un opérateur totalement disponible sont requis. Lorsque ces conditions ne sont pas respectées, le pilotage devient alors difficile, fatigant et des erreurs peuvent êtres commises (McGovern 1991; Sheridan 1992).

1.2.4.2 La supervision : utiliser un plan de navigation

Lorsqu'un opérateur ne peut être dédié au pilotage ou que les conditions techniques ne permettent pas d'assurer une boucle de contrôle stable, la supervision s'avère être une solution acceptable (Sheridan 1992).

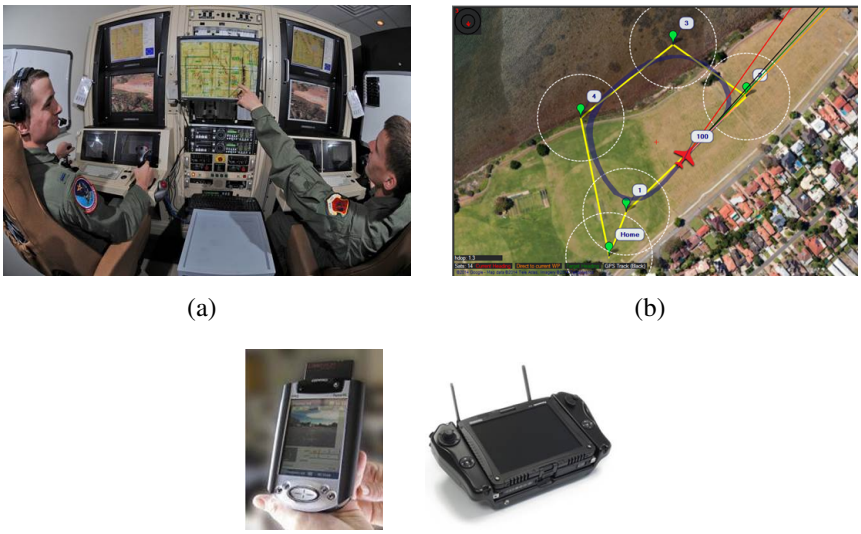
De manière générale, le principe du contrôle supervisé est de diviser un problème à résoudre en une séquence de sous-tâches qu'un robot peut exécuter de manière autonome. Alors que la téléopération requiert un effort constant et soutenu tout le long de la mission, la supervision se limite à une préparation de mission, à un contrôle de l'exécution, et éventuellement à une synchronisation des actions au cours de l'exécution.

Les interfaces utilisant un plan de navigation sont une application du principe de supervision à la problématique du pilotage de robot (Cameron *et al.* 1987; Kay 1995). Il s'agit de définir, en préparation de mission, un itinéraire composé de points de passages géo-référencés (GPS) et successifs sur une vue de la zone à explorer. Un point stratégique sécurisé, appelé base, peut également être défini pour indiquer au drone où retourner en cas d'avarie.

Par la suite, au cours de la mission, l'activation de quelques fonctions de haut niveau devient alors suffisant. Les fonctions de haut niveau classiques sont de faire décoller le drone, de le faire atterrir, de lui demander d'aller au point suivant sur l'itinéraire, de revenir au point précédent, de rejoindre le point de repli sécurisé, et d'interrompre immédiatement le déplacement en cours.

L'itinéraire peut également être augmenté par des actions contextuelles ; par exemple, réaliser une prise de vue panoramique à une position et à un instant précis. Ces interfaces, dont la Figue 1.4(b) est une illustration, peuvent être extrêmement simples et rapides à utiliser.

De plus, les outils modernes de planification et d'analyse permettent d'assister l'opérateur dans sa tâche de construction du plan. Un plan de navigation peut également être modifié en cours de mission pour répondre à des problématiques de consommation énergétique ou de nouvelles menaces, et intégrer dynamiquement de nouveaux objectifs.



(a) (b)

(c) (d)

FIGURE 1.4 – Différents outils pour la commande de drones : (a) un poste de pilotage de drone de classe III ; (b) l'application ArduPilot proposée par DIY Drones ; (c) PdaDriver (Fong *et al.* 2003) pour la téléopération à courte portée ; et (d) la station de contrôle au sol de la société Novadem pour la téléopération ou la supervision.

1.2.5 Le geste pour la commande des drones de contact

Il existe donc différents types de robots militaires, pouvant être pilotés avec différents outils et de différentes manières.

Le cas de la commande des drones de contact nous intéresse particulièrement ici. En effet, ces drones de petite taille et de faible endurance permettent d'acquérir de l'information par et pour le combattant pendant sa progression sur le terrain. Ils permettent par exemple de contrôler des intersections avant franchissement ou d'éclairer derrière une colline ou un bâtiment.

Paradoxalement, alors que le rôle de ces petits drones est d'aider à la progression, force est de constater que leurs interfaces de pilotage peuvent constituer un frein à cette même progression. En effet, elles imposent des contraintes fortes à la fois visuelles et physiques : l'opérateur doit manipuler du matériel et focaliser son attention sur un restituteur (écran). Aussi, pendant les phases de commande, l'opérateur n'est plus en mesure d'observer son environnement pour y déceler la présence d'éventuelles menaces, de se déplacer librement, ni dans le cas extrême, de faire usage de son arme. L'intérêt pratique de ces drones en est donc réduit.

Or, si, comme nous le pensons, les gestes sémaphoriques permettent effectivement de commander sans perturber l'attention visuelle et sans intermédiaire matériel, ils pourraient donc constituer une modalité avantageuse pour les drones de contact. Ainsi, en comparaison des interfaces classiques, ils permettraient de libérer les yeux et les mains du pilote et conduiraient à une meilleure réactivité et fluidité dans la progression sur le terrain.

1.3 Résumé des objectifs et organisation du manuscrit

1.3.1 Objectifs

Dans ce travail, nous souhaitons donc montrer que les gestes sémaphoriques sont effectivement utilisables pour interagir avec une machine.

Aussi, nous voulons

— d'une part, valider notre hypothèse quant à une plus faible influence de la commande au geste sur la capacité à percevoir l'environnement en comparaison d'interfaces standards et plus globalement à libérer l'utilisateur ;

— et d'autre part, montrer qu'il est actuellement possible de mettre en oeuvre un système technique relativement simple.

Pour ce faire, nous avons choisi de nous placer dans le contexte de la commande des drones militaires de contact qui, nous le pensons, bénéficierait de l'usage du geste sémaphorique et constitue donc un cadre d'étude privilégié.

1.3.2 Organisation du manuscrit

Après cette introduction, le manuscrit est organisé en cinq chapitres principaux. Dans un premier temps, les chapitres 2, 3 et 4 adressent la problématique de l'interaction ges-

tuelle d'un point de vue théorie : le modèle d'interaction (2), la méthode pour choisir des gestes (3) et l'impact sur l'attention visuelle (4). Puis, dans un second temps, les chapitres 5 et 6 concrétisent les apports théoriques en proposant une plate-forme technique et en l'appliquant pour conduire un test utilisateur auprès d'opérationnels militaires. Finalement, le chapitre 7 propose en conclusion une revue des contributions et une présentation de différentes perspectives.

Chapitre 2 : Après une revue de la littérature quant aux types de gestes et leurs usages en interaction homme-machine, un modèle d'interaction spécifique aux gestes sémaphoriques et à la commande de robots mobiles est proposé. Celui-ci repose sur l'usage d'un agent conversationnel qui crée une passerelle sécurisée entre l'homme et le robot. L'interaction consiste alors en un dialogue : l'humain utilise la modalité gestuelle sémaphorique pour exprimer ses intentions et l'agent répond et informe l'utilisateur via une modalité sonore iconique ou synthétique vocale.

Chapitre 3 : Une interaction gestuelle sémaphorique repose sur l'usage d'un vocabulaire gestuel standardisé qu'il faut donc définir. Après une revue des différentes méthodes de la littérature permettant de constituer un vocabulaire gestuel, nous constatons que deux d'entre elles sont complémentaires et nous semblent particulièrement pertinentes. En les combinant, nous proposons alors une nouvelle méthodologie qui fait participer des utilisateurs finaux pour proposer, élire puis valider un ensemble de gestes qui expriment le mieux possible les commandes d'un système. Finalement, nous appliquons notre méthodologie dans le cadre de la commande d'un drone militaire de contact.

Chapitre 4 : Après avoir proposé un modèle d'interaction et défini méthodiquement un vocabulaire de gestes, nous conduisons une étude en laboratoire pour évaluer l'impact de la commande au geste sémaphorique sur l'attention visuelle en comparaison d'une commande plus standard par appui sur des boutons tactiles. Cette étude réalisée avec un protocole de double-tâche et en magicien d'Oz (en simulant un système), d'une part, confirme que notre vocabulaire gestuel est simple à apprendre et à utiliser ; et d'autre part semble valider notre hypothèse quant-à une plus grande liberté permise par le geste et donc son adéquation avec la problématique militaire.

Chapitre 5 : Sur la base des résultats de la précédente évaluation, une plate-forme technique générique est développée. Nous décrivons la composition de celle-ci : une brique de détection pour les gestes et un automate fini configurable pour le modèle d'interaction. Pour la brique de détection de gestes, une revue des différentes problématiques, capteurs et algorithmes classiquement considérés dans la littérature relative aux gestes 3D révèle une certaine complexité de mise en oeuvre ; et particulièrement pour l'apprentissage automatique. Pour plus de simplicité dans la mise en oeuvre, nous proposons un système différent reposant sur la définition déclarative des gestes sous la forme d'expressions rationnelles ; méthode jusqu'alors rencontrée uniquement dans le cadre des interfaces tactiles. Finalement, nous montrons comment cette plate-forme permet la mise en œuvre du modèle d'interaction

et du vocabulaire gestuel proposés pour l'interaction avec un drone militaire de contact.

Chapitre 6 : Finalement, un *test utilisateur* est réalisé auprès d'opérationnels militaires. On demande à ceux-ci de se déplacer, d'observer leur environnement et de conserver leur arme disponible tout en envoyant des commandes à un drone virtuel dans différents scénarios. Pour cela, les participants expérimentent la commande gestuelle sémaphorique ainsi qu'une interface tactile simple afin d'avoir un outil de comparaison. Nous concluons sur la facilité d'apprentissage et d'usage de ces modalités.

CHAPITRE 2

Positionnement : design de l'interaction

Sommaire

Le domaine de l'interaction homme-machine par geste est extrêmement vaste. En dehors même de la multitude des problématiques et solutions techniques, il existe différents types de gestes et autant de manières de les utiliser. Aussi il convient, lorsque l'on veut parler de geste, de préciser d'abord ce que l'on entend par là. C'est ce que nous nous proposons de faire dans la première partie de ce chapitre.

Dans un premier temps, nous définissons la notion de geste. Pour cela, nous présentons une définition généraliste ainsi que les classifications les plus influentes de la littérature.

Dans un second temps, nous présentons une revue de la littérature spécifique aux interactions gestuelles homme-robot.

Enfin, et dans un troisième temps, nous nous positionnons en choisissant un type de geste ainsi qu'une manière de l'utiliser. Avec une approche centrée utilisateur, nous construisons un modèle d'interaction, reposant sur le geste et adapté à notre cas d'application : la commande des drones militaires de contact.

2.1 Geste et interaction gestuelle

2.1.1 Définition généraliste

Dans la plupart des premiers documents scientifiques de l'étude moderne du geste, le concept n'y est pas explicitement défini. Cela laisse donc à penser que la notion de geste est évidente pour tous et qu'y revenir est inutile.

Pourtant, lorsque l'on compare ce qui est alors implicite, il n'est pas évident que les auteurs parlent de la même chose : certains parlent de mouvements qui accompagnent la parole (McNeill 1992; Rimé & Schiaratura 1991), d'autres, de signes (Kendon 1988b; Godenschweger & Strothotte 1998).

Ainsi, il ne semble pas y avoir d'évidence ; et plusieurs auteurs ont affirmé le besoin de définir, ou de redéfinir, ce qu'est le geste (Kendon 1996; Wexelblat 1997) :

> In everyday discussion we all think we know what we mean by 'gesture.' The problem is to make explicit on what this knowledge is based.
>
> (Kendon 1996)

Finalement, différentes propositions ont été faites. Nous en citons cinq avant de préciser ce qu'elles ont en commun : un mouvement, une signification et un observateur ; et ce qui semble relier ces trois éléments de manière implicite : une trace sur un support et une norme.

> For an action to be treated as a 'gesture', it must have features wich make it stand out as such.
>
> (Kendon 1986)

> Gestures are not just movements and can never be fully explained in purely kinesics terms. [...] but symbols that exhibits meanings [...].
>
> (McNeill 1992)

> A gesture is a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on its way to hitting the key is neither observed nor significant. All that matters is which key was pressed. [..] Beckoning with your index finger is a gesture. Handwriting is not a gesture because the motion of the hand expresses nothing ; it is only the resultant words that convey the information.
>
> (Kurtenbach & Hulteen 1990)

> A gesture is that thing which distinguishes itself from the background motions by virtue of something - a movement, a shape, etc. - that catches our attention." [...] Gestures are like the proverbial tree falling in the forest : if no one sees the gesture it is lost.
>
> (Wexelblat 1994)

> A gesture is a movement of one's body that conveys meaning to oneself or to a partner in communication. That partner can be a human or a computer. Meaning is information that contributes to a specific goal.
>
> (Hummels *et al.* 1997)

Un mouvement : Le premier point, assez explicite est la notion de mouvement que l'on retrouve à la fois sous l'appellation *motion* (Kurtenbach & Hulteen 1990; Wexelblat 1994) et *movement* (Wexelblat 1994; Hummels *et al.* 1997). Il convient de préciser que ce mouvement est celui d'un corps humain : *body* (Kurtenbach & Hulteen 1990; Hummels *et al.* 1997), et l'emploi du mot *action* (Kendon 1986) suggère que le mouvement doit être propre au corps et non imposé par une force extérieure.

Une signification : Le mouvement contient de l'*information* (Kurtenbach & Hulteen 1990; Hummels *et al.* 1997) et possède une signification : *meaning* (Hummels *et al.* 1997). Dans les deux cas il s'agit d'une forme d'abstraction : les paramètres physiques du mouvement ne sont pas à considérer au premier degré mais font l'objet d'une analyse qui permet d'en déterminer la sémantique.

Un observateur : Si les deux premiers éléments sont assez évidents, la notion d'observateur est quant à elle d'avantage cachée. Pourtant, si un mouvement est perçu et si une signification est donnée, c'est nécessairement par quelque-chose ou quelqu'un : *human or a computer* (Hummels *et al.* 1997). Parfois, il peut y avoir plusieurs observateurs et la personne qui réalise le geste peut être sont propre observateur (notamment dans les phases d'apprentissage). Enfin, dans le cadre de l'étude du geste, le chercheur est nécessairement un observateur ; ce qui explique que dès lors qu'il y a étude du mouvement, il peut être raisonnable de parler de geste.

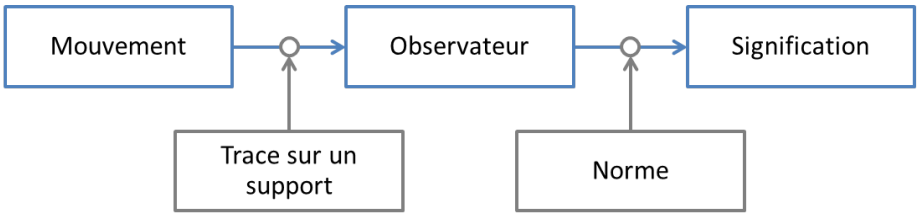


FIGURE 2.1 – Les cinq éléments d'une définition généraliste du geste.

Une trace sur un support :

En archéologie, différents auteurs dont Leroi-Gourhan (Leroi-Gourhan 2013) ont étudié des gestes d'hommes du passé ; et pour ce faire, ils ont observé et analysé les traces laissées sur des objets fossilisés. Ainsi, les marques présentes sur un silex peuvent révéler : la manière dont cet outil a été construit ou comment il a été utilisé.

Au regard de ces recherches, nous proposons que ce que l'observateur perçoit du mouvement c'est une trace laissée sur un support ; et cette trace n'est pas nécessairement directe ou immédiate. Ainsi, il est possible de juger de la précision d'un peintre au regard d'une toile ou de la virtuosité d'un musicien à l'écoute d'un morceau joué.

Dans le cas des gestes réalisés en l'air et donc purement visuels, la trace et le support semblent absents. On peut cependant considérer deux possibilités : soit la trace est temporaire et immatérielle, soit, le support est la mémoire visuelle de l'observateur et la trace est un souvenir. Finalement, ce qui rend évident les notions de trace et de support, c'est que l'on puisse réaliser un enregistrement pour revivre l'expérience d'un geste : la vidéo en est depuis longtemps un exemple en permettant de capturer le mouvement.

Une norme : Enfin, le dernier point qui nous semble important est la notion de norme ou de modèle. Ce sont eux qui permettent de décrypter une signification à partir d'une trace voire même, décrivent ce qu'est une trace. Ainsi, pour communiquer par geste, il est fondamental que le locuteur et l'interlocuteur partagent la même norme : le même vocabulaire. Lorsqu'ils considèrent des normes différentes, cela peut conduire à des incompréhensions. Finalement, la notion de norme est importante car elle montre qu'un mouvement ne se suffit pas pour être un geste. Il est ainsi raisonnable de penser que bien que l'on puisse capturer une trace, on ne peut pas capturer un geste.

Ainsi, nous proposons la définition généraliste suivante pour le geste : tout **mouvement** d'un corps humain généré par celui-ci, dont la **trace** laissée sur un support de manière temporaire ou permanente, est perçue par un **observateur** qui, au regard d'une **norme**, lui attribue une **signification**. Les cinq éléments de cette définition sont repris en Figure 2.1.

Finalement, au regard de cette définition généraliste, le domaine de recherche qui porte sur le geste est donc bien un tout cohérent mais vaste, dans lequel il nous semble nécessaire, pour une plus grande clarté, de considérer un niveau de détail plus important.

2.1.2 Classifications et modèles classiques

Les domaines de la psychologie cognitive et de la psycholinguistique ont posé les bases de l'étude théorique du geste. Le geste est alors considéré comme un élément linguistique visuo-kinésique (visuo-moteur). On cherche, entre autres, à en déterminer les caractéristiques discriminantes et à constituer des catégories. Nous allons présenter les classifications et modèles les plus influents.

2.1.2.1 La classification d'Efron

Bien qu'Efron n'emploie pas explicitement le mot geste, ses travaux sont considérés comme les premiers du domaine (Efron 1941).

A New York, il observe les comportements non-verbaux d'immigrants européens juifs et italiens. Il entreprend alors de différencier les gestes suivant le rapport aux concepts qu'ils représentent : les objets référés (ou référents). Alors que certains gestes ne semblent avoir aucun référent explicite, certains renvoient à des objets de l'environnement (externes au locuteur), et d'autres à des objets invisibles, fruits de la pensée du locuteur (internes au locuteur). Sur cette base, Efron propose alors différentes catégories :

1. Les **discursive gestures** n'ont pas de signification propre et servent à ponctuer un énoncé oral.
2. Les **objective gestures** possèdent une signification propre. Il peuvent être employés avec la parole mais n'ont pas besoin d'elle pour être compréhensibles.
   (a) Les **deictic gestures** désignent physiquement le référent lorsque celui-ci est présent ; généralement en pointant du doigt.
   (b) Les **physiographic** décrivent le référent lorsque celui-ci est absent.
   — Les **iconographic gestures** décrivent directement l'objet référé.
   — Les **kinetographic gestures** décrivent l'objet référé par le biais de son usage. Par exemple tourner un volant pour signifier une voiture.
   (c) Les **symbolic gestures** ou **emblematic gestures** désignent un référent de manière totalement abstraite avec une forme standardisée.

Bien que difficile à appliquer systématiquement, cette première classification sémiotique a montré qu'il est possible de discriminer certains types de gestes en fonction de la manière dont ils expriment un concept. Par la suite, cette approche deviendra un standard et sera adoptée par de nombreux auteurs.

2.1.2.2 Le continuum de Kendon

Rapidement la classification d'Efron a été reprise, et dans la lignée de ces travaux, une classification particulièrement influente a été proposée par McNeill (McNeill 1992). Nommée *continuum de Kendon* en référence à l'auteur du même nom, cette classification a la particularité d'organiser différentes catégories selon plusieurs continuums. Le nombre de catégories et de continuums ont par ailleurs évolué au cours du temps : composée à l'origine

de quatre catégories et de deux continua, elle est par la suite enrichie de deux nouveaux continua (McNeill 2000), ainsi que d'une cinquième catégorie (Kendon 2004).

Le continuum de Kendon, dans sa version la plus récente est représenté en Figure 2.2 et propose donc l'utilisation de quatre axes pour caractériser cinq catégories de gestes. Ces quatre axes sont :

1. **Relation à la parole** : Le geste est-il utilisé seul ou avec la parole ?
2. **Propriétés linguistiques** : Le geste est il structuré ?
3. **Existence d'une convention** : Les gestes reposent-ils sur un standard ?
4. **Sémiotique** : De quelle manière le sens d'un geste est-il relié à sa forme ?

Les cinq classes de gestes sont les suivantes :

1. Les gestes **co-verbaux** : Également appelés *gesticulations* ou *gestes co-occurents*, ces gestes sont toujours en relation avec le contenu de l'énoncé oral qu'ils accompagnent. Les co-verbaux étudiés seuls, c'est à dire sans considérer le canal vocal, ne présentent aucune propriété linguistique. Il apparaît également que ces gestes ne sont pas standardisés. Ils sont au contraire créés à la volée en fonction du besoin et selon l'image mentale du locuteur. Toutefois, il s'avère que lors d'un échange entre deux interlocuteurs, une convention temporaire s'établisse (Gerwing & Bavelas 2004; Holler & Wilkin 2011; Clark 1996). Enfin, les gestes co-verbaux sont globaux et synthétiques : la signification des gestes est définie par l'ensemble instantané de ses différentes composantes (forme de la main, position, mouvent, ...).

   Différentes sous-catégories de gestes co-verbaux ont par ailleurs été définies par McNeill (McNeill 1992). Illustrée en Figure 2.3, elles sont les suivantes :

   (a) Les gestes **imagés** : Ils sont non contextuels dans la mesure où leur sens est décrit directement ou indirectement par la forme du geste. On trouve alors deux types de gestes imagés :

      i. Les gestes **iconiques** représentent par leur exécution un objet, une action ou un évènement concret.

      ii. Les gestes **métaphoriques** sont semblables aux gestes iconiques dans leur forme, mais représentent des éléments abstraits. Il sont généralement plus complexes que les gestes iconiques.

   (b) Les gestes **non-imagés** : Ils s'inscrivent dans un contexte spatial ou temporel dont la connaissance est nécessaire à l'interprétation.

      i. Les gestes **déictiques** sont l'action de pointer, généralement du doigt, l'élément référé. Les objets pointés peuvent être concrets ou abstraits mais également présents ou absents.

      ii. Les **battements** n'ont pas de signification dans leur forme ni dans leur position, mais dans leur usage. Ils peuvent, entre autres, signaler les éléments importants d'un énoncé.

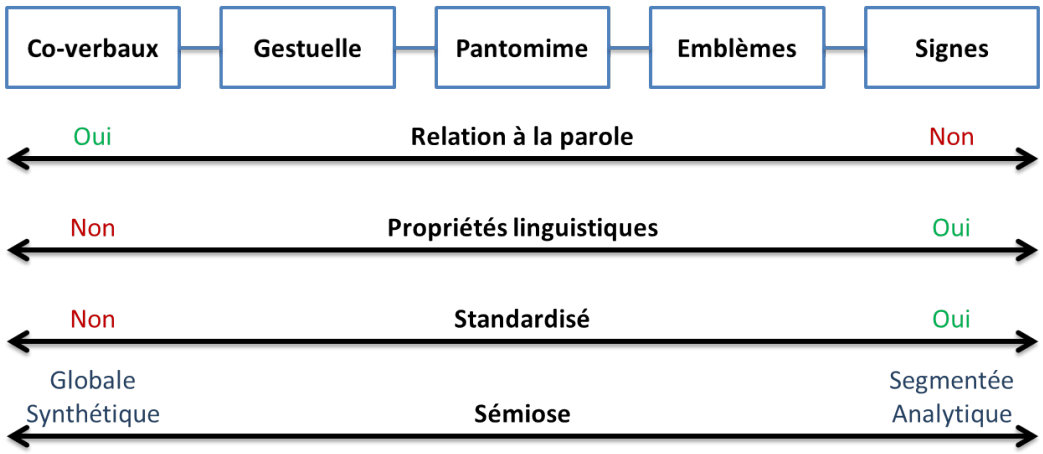


FIGURE 2.2 – Le continuum de Kendon, dans sa version de 2004 (Kendon 2004), propose cinq catégories de gestes organisées selon les propriétés de quatre continuums.

2. Les **gestuelles** : Également appelés *Speech-Framed gestures*, ces gestes sont analogues aux gestes co-verbaux. Cependant, les canaux verbaux et gestuels ne sont pas utilisés en parallèle mais en alternance. Les gestes servent à remplacer des mots ou des concepts que la parole échoue à exprimer. Par exemple, ils surviennent très fréquemment lors de l'expression dans une langue étrangère, lorsque le vocabulaire du locuteur est plus pauvre que dans sa langue natale.
3. Les **emblèmes** (ou sémaphoriques) : Ce sont des gestes conventionnels dont la forme et la signification sont partagés par la communauté qui les utilise. Ils sont donc fortement dépendants de la culture et peuvent parfois avoir différentes significations.
4. Les gestes de **pantomime** : Sont un type de geste ou une séquence de gestes ayant pour but de transmettre une histoire en suivant un fil narratif.
5. Les **signes** : Enfin, les signes sont les gestes utilisés au sein de langues normalisées et structurées. Les travaux de Stokoe (Stokoe 1960; Stokoe 2005) ainsi que ceux de Bellugi (Bellugi 1979) ont initié l'étude de leurs processus de création et de standardisation (Goldin-Meadow 1993; Shun-Chiu 1992; Dos Santos Souza 1999), ainsi que de leurs aspects linguistiques tels que la phonologie (Friedman 1976) ou la syntaxe (Liddell 1977; Grosjean & Lane 1977).

   Par ailleurs, il a été montré que des langues des signes, plus ou moins complexes, sont utilisés dans toute situation où la parole est contrainte. De manière évidente lors d'un handicap (sourds et muets), mais également dans des environnements particuliers (pompiers, militaires, plongeurs, astronautes, chefs d'orchestre ...). Enfin, les langues des signes ont la propriété d'être segmentées et analytiques : les éléments sont présentés de manière linéaire et le sens de chacun permet de déterminer le sens de l'ensemble.

Le continuum de Kendon propose donc différentes catégories claires de gestes qui peuvent êtres organisées selon différents axes de lecture. Accompagnée d'une méthode pour

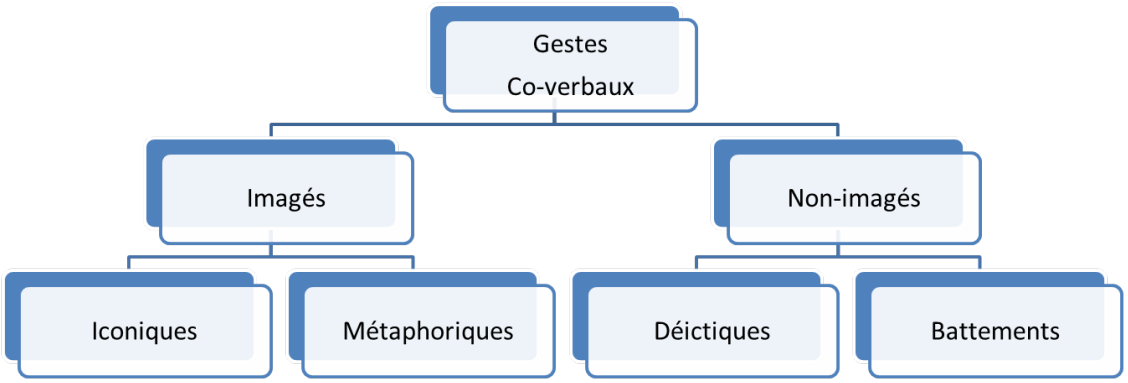


FIGURE 2.3 – Catégorisation sémiotique des gestes co-verbaux du continuum de Kendon (McNeill 1992).

classer les gestes, cette classification est beaucoup employée et constitue une référence du domaine.

2.1.2.3 Modèles anatomiques

En parallèle des classifications sémiotiques qui héritent des travaux d'Efron, différents modèles anatomiques ont également été proposés. Ces modèles décrivent, non pas la sémantique des gestes, mais leurs structuration temporelle.

Kendon (Kendon 1980) a proposé une décomposition de l'exécution des gestes en différents éléments temporels : un geste, qu'il soit un signe ou un co-verbal est une unité gestuelle (*G-Unit*), composée d'une ou plusieurs phrases gestuelles (*G-Phrase*) qui, à son tour, est composée d'une succession de phases gestuelles élémentaires (*G-Phase*) :

1. La **préparation** est la phase pendant laquelle la main, ou la partie du corps qui sert le geste, quitte sa position de repos et rejoint la position où la partie sgnifiante du geste sera réalisée.
2. Le **stroke** est la phase du geste qui contient par sa forme et sa dynamique, la sémantique du geste.
3. La **rétraction**, facultative, est la phase pendant laquelle la main retourne dans une position de repos ; éventuelle différentes de la position de repos d'origine.

De manière concrète, un geste survient donc entre deux positions de repos de la main. Il peut être considéré, soit comme geste dynamique si le stroke comporte du mouvement, soit statique si le stroke n'en comporte pas. On parlera également de *pause* pour les gestes statiques. A noter que ce qui importe c'est la présence de mouvement dans la phase de stroke uniquement. Ainsi, la plupart du temps un geste statique comporte tout de même une part mouvement. Enfin, la distinction entre dynamique et statique ne dépend absolument pas de la durée du geste : une pause peut être plus longue qu'un geste dynamique et inversement.

A ce modèle proposé par Kendon, deux nouvelles phases élémentaires et facultatives ont été ajoutées (Kita *et al.* 1997) :

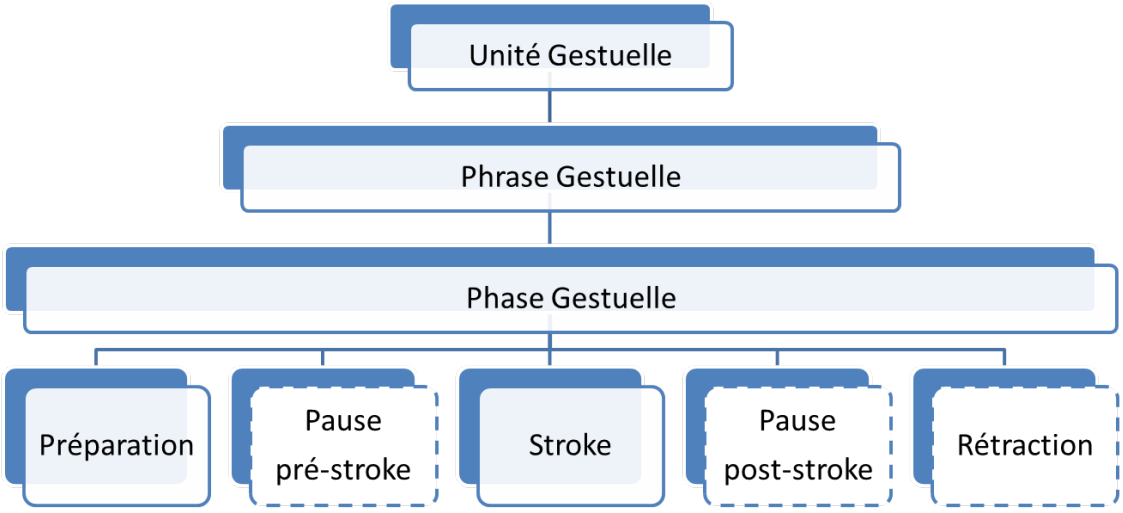


FIGURE 2.4 – Modèle de description anatomique des gestes proposé par Kita et al. (Kita *et al.* 1997) sur la base des travaux initiaux de Kendon (Kendon 1980). Les blocs discontinus sont facultatifs.

1. Une pause dite **pré-stroke** qui est un arrêt du déplacement avant une phase de stroke.
2. Une pause dite **post-stroke** qui est un arrêt du déplacement après une phase de stroke.

La version complète de ce modèle anatomique est illustré en Figure 2.4. Elle a par ailleurs été fréquemment utilisée et différentes manière de segmenter les phases ont été proposées (Seyfeddinipur 2006; Bressem & Ladewig 2011) ; par exemple, dans un flux vidéo, l'apparition ou la disparition d'un flou de mouvement est caractéristique d'un changement de phase.

2.1.3 Classifications en IHM

Dans le domaine des interfaces homme-machine, de nombreux auteurs ont proposé d'utiliser le geste comme modalité pour interagir avec un système. Dans ce contexte, les différentes classifications proposées en psychologie cognitive et en psycholinguistique ont naturellement été reprises.

Certains auteurs se positionnent directement au sein de ces classifications classiques ; par exemple, Sparell (Sparrell 1993) s'intéresse spécifiquement aux gestes *co-verbaux iconiques* de la classification de McNeill (McNeill 1992).

D'autres (Quek 1995; Quek *et al.* 2002; Karam 2006) s'appuient sur les classifications classiques mais les modifient. Deux changements semblent alors notables : d'une part, la catégorie des gestes manipulatifs est ajoutée, et d'autre part, le niveau de détail est réduit conduisant à des classifications de plus haut niveau.

2.1.3.1 Ajout des gestes manipulatifs

Queck (Quek 1995), en informatique, s'intéresse à l'application de techniques de vision par ordinateur pour permettre à une machine d'interpréter automatiquement les gestes d'un utilisateur. Il pose alors la question de savoir si il existe différents modes de communication gestuelle et, dans l'affirmative, lequel serait profitable à une interface homme-machine. Pour y répondre, il s'appuie sur différentes typologies classiques dont le continuum de Kendon mais constate qu'elles semblent ne traiter que de gestes servant la communication (*communicatifs*) et ainsi omettre une catégorie de gestes : les gestes *manipulatifs*.

Les gestes manipulatifs sont alors définis comme l'établissement d'une relation qui lie la configuration et les mouvements du corps à un objet. Contrairement aux gestes communicatifs, ils ne sont pas sémantiques mais physiques et leur effet est continu et instantané. De plus, puisque ces gestes ne sont pas destinés à être interprétés par quelqu'un, leurs caractéristiques ne sont pas nécessairement visibles.

2.1.3.2 Réduction du niveau de détail

Parallèlement à l'ajout de la catégorie des gestes manipulatifs, le niveau de détail considéré semble diminuer. En effet, alors que dans les typologies classiques on cherche à distinguer les subtilités sémantiques de chaque geste, en informatique, des catégories de haut niveau semblent suffisantes.

Ainsi, pour Quek (Quek *et al.* 2002), les trois seules catégories des gestes manipulatifs, sémaphoriques et co-verbaux suffisent pour classer les gestes utilisés en informatique.

Par la suite, dans le cadre d'une revue de la littérature des IHM gestuelles, Karam (Karam 2006) reprend et modifie la classification de Quek. Elle y ajoute alors les deux catégories des gestes déictiques et des langues des signes.

Cette dernière classification, pourtant reprise par d'autres auteurs (Fikkert 2010), nous semble cependant poser question : en particulier la considération des gestes déictiques en tant que catégorie de haut niveau. En effet, lorsque l'on applique une classification pour comparer différents gestes ou systèmes gestuels, il est usuel de considérer des catégories mutuellement exclusives. Or ici, la catégorie des gestes déictiques semble compatible de l'ensemble des quatre autres catégories : à l'évidence des différents gestes communicatifs mais également des gestes manipulatifs.

Pour illustrer ce questionnement relatif à la catégorie des gestes déictiques, on peut prendre pour exemple les travaux de Bolt (Bolt 1980) considérés comme l'un des premiers dans le domaine de l'interaction gestuelle homme-ordinateur. Dans l'article de référence *Put-That-There*, Bolt propose d'interagir avec un écran numérique de grande taille en employant la modalité verbale couplée à l'usage des gestes. Il est alors possible de créer, détruire et déplacer des objets virtuels simples (forme et couleur). Le geste de désigner du doigt une zone de l'écran permet soit de sélectionner un objet existant, soit d'indiquer la destination d'un déplacement ou de la création d'un nouvel objet. Au sein de la classification qu'elle considère, Karam propose de classer les travaux de Bolt dans la catégorie des gestes déictiques. Ce choix est cohérent, mais il semble également possible de classer ces travaux

dans la catégorie des gestes co-verbaux. Que faut-il alors considérer ?

Nous choisissons de ne conserver que les quatre catégories des gestes manipulatifs, co-verbaux, sémaphoriques et des langues des signes, et de considérer la forme déictique comme un paramètre commun qui peut être précisé au besoin.

2.1.4 Revue de la littérature en interaction gestuelle Homme-Robot

Après avoir présenté une définition généraliste de la notion de geste ainsi que différentes typologies permettant de les classer, nous proposons maintenant une revue de la littérature spécifique aux interactions gestuelles Homme-Robot.

Pour cela, nous avons choisi de ne conserver que les documents scientifiques présentant concrètement un modèle d'interaction et ne se limitant pas à la seule évocation d'un use-case en introduction. Ainsi, sur cette base, nous avons retenu 25 documents.

Pour chacun, nous avons recensé différentes caractéristiques : le type de robot, la proximité entre le robot et l'utilisateur, le mode d'interaction, le type de geste et enfin, le type de feedback.

2.1.4.1 Type de robot

Nous avons relevé trois grands types de robots :

— **Robots mobiles :** La grande majorité des travaux adressent l'usage des robots mobiles. Qu'il s'agisse de plate-formes roulantes (UGV) ou de drones réels et virtuels, l'interaction a pour objectif de faire se déplacer le robot.

— **Essaims de robots mobiles :** Dans le cas de l'interaction avec un essaim de drones (Monajjemi *et al.* 2013), les auteurs proposent de pouvoir gérer un groupe de drones actifs : pour chaque drone, l'utilisateur peut l'inclure ou l'exclure du groupe actif. Une fois le groupe constitué, les commandes de déplacement passées par l'utilisateur sont exécutées par l'ensemble des robots du groupe actif. Les auteurs indiquent par ailleurs, que cette méthode est applicable pour n'importe quel essaim de robot mobile, quel que soit leur type (UGV ou UAV). On peut également envisager un essaim mixte composé à la fois de robots roulants et volants.

— **Bras articulés :** Enfin, trois articles seulement adressent l'usage d'un bras articulé (Coupeté *et al.* 2016; Triesch & Von Der Malsburg 1998; Rogalla *et al.* 2002). Il s'agit alors d'indiquer au robot un objet à déplacer ou à quel moment le faire.

2.1.4.2 Proximité

Nous avons relevé deux types de configuration :

— **Local :** Dans la grande majorité des travaux, le robot et l'utilisateur sont face à face.

— **Distant :** Dans seulement quatre travaux, l'utilisateur commande le drone à distance.

2.1.4.3 Mode d'interaction

Les deux modes classiques d'interaction ont été rencontrés :

— **La commande :** Le robot exécute de manière autonome des actions de plus ou moins haut niveau allant du décollage à l'exécution complètement automatique d'une mission pré-programmée (Monajjemi *et al.* 2013).
— **Le contrôle :** Une boucle d'asservissement est établie entre le drone et l'utilisateur. Ce dernier gère alors certains paramètres du robot comme sa vitesse ou son orientation, de manière continue ou avec des commandes de très bas niveau (augmentation vitesse dans une direction déterminée).

2.1.4.4 Type de geste

Nous avons donc choisi de classer les gestes rencontrés dans les quatre catégories de haut niveau (manipulatif, co-verbal, sémaphorique et langue des signes) et de préciser, le cas échéant, si il s'agit seulement de gestes déictiques.

— **Gestes manipulatifs :** Dans seulement deux références, les gestes manipulatifs sont utilisés.
  — Dans un cas (Fong *et al.* 2001), la main droite de l'utilisateur est maintenue en l'air et fait office de joystick. Par ce moyen, il est alors possible de contrôler les déplacements du drone.
  — Dans l'autre cas (Coupeté *et al.* 2016), l'utilisateur manipule des pièces de moteur sur une chaîne de montage. Le robot, ici un bras articulé, observe ces gestes et présente en conséquence la pièce de moteur dont l'utilisateur aura ensuite besoin. Ce qui est intéressant dans ce second cas, c'est que les gestes manipulatifs ne sont pas destinés au robot ; et malgré cela, l'utilisateur n'a pas besoin d'exprimer explicitement de commandes. La machine réalise un suivi de l'activité de l'opérateur humain et sa perception des gestes peut être considérée comme une forme d'interface transparente : les commandes sont invisibles pour l'utilisateur qui pourtant les utilise.
— **Gestes co-verbaux et co-verbaux déictiques :** Seulement trois références utilisent des gestes co-verbaux : c'est à dire combinent une modalité vocale et gestuelle.
  — Dans deux cas (Rogalla *et al.* 2002; Perzanowski *et al.* 1998), il s'agit uniquement de gestes co-verbaux déictiques. Pour un bras articulé, à la manière des travaux de Bolt (Bolt 1980) ils permettent de désigner un objet à déplacer et où le déplacer. Pour un drone, ils permettent d'indiquer une position où le drone doit se rendre. Dans les deux cas, la modalité vocale est la modalité principale et le geste permet seulement de compléter avec des informations spatiales.
  — Dans le troisième cas (Jones *et al.* 2010), les auteurs décrivent une étude d'élicitation en environnement virtuel et en magicien d'Oz (comportement du drone simulé par un opérateur humain caché). Il est demandé à des participants d'essayer d'interagir naturellement avec le drone virtuel ; aucune contrainte ne leur est imposée et cela permet, après analyse, de dresser une liste des gestes co-verbaux utilisables. Comme précédemment, les gestes déictiques permettent de désigner une position que le drone doit atteindre. Cependant, des gestes non déicitiques sont également proposés : ils permettent d'exprimer des commandes

telles que le décollage ou l'atterrissage.

— **Gestes sémaphoriques et sémaphoriques déictiques :**
  — Dans la très grande majorité des cas, les gestes utilisés sont de type sémaphorique. Différents gestes abstraits mais standardisés permettent d'activer une commande du robot. Pour certaines commandes relativement autonomes (décollage, atterrissage), l'exécution a lieu une seule fois et démarre lorsque le geste est terminé ; c'est toujours le cas lorsque les gestes sont dynamiques, et parfois avec les gestes statiques. Pour d'autres commandes de plus bas niveau (augmenter la vitesse ou l'altitude) et avec des gestes statiques, la commande peut être exécutée de manière répétitive tant que la pause est maintenue. Ce cas d'exécution répétitif s'apparente à de la manipulation ; cependant, il n'en est rien puisqu'aucun lien spatial et temporel continu n'est établi entre le corps de l'utilisateur et le robot.
  — Enfin, dans deux cas seulement (Van den Bergh *et al.* 2011; Bauer *et al.* 2009), il s'agit de gestes uniquement déictiques : il s'agit d'indiquer à un drone une position à atteindre. Cependant, contrairement aux travaux précédemment évoqués, la parole n'est pas utilisée. L'usage de l'information spatiale exprimée par le geste est toujours implicitement interprétée comme une commande de déplacement.

Aucune référence ne semble utiliser de langue des signes pour la commande d'un robot.

2.1.4.5 Feedback

Enfin, la dernière caractéristique que nous avons relevée est le type de feedback proposé à l'utilisateur pour interagir avec le robot. Quatre types ont été recensés :

— **Naturel :** Dans la très grande majorité des cas, le feedback est l'interprétation du comportement du robot. L'utilisateur observe la machine et détermine si ses mouvements correspondent bien à ce qui est demandé.
— **Vidéo :** Dans les cas où l'interaction est distante, le feeback est l'affichage, sur un écran, du flux vidéo d'une ou plusieurs caméras embarquées par le drone (Iba *et al.* 1999; Fong *et al.* 2003; Nahapetyan & Khachumov 2015; Zhang & Liu 2015).
— **Voyants lumineux :** Dans un unique cas (Monajjemi *et al.* 2013), des drones informent l'utilisateur par l'intermédiaire de voyants lumineux colorés. Le code couleur explicite alors l'état de chaque drone.
— **Messages vocaux :** Enfin, dans seulement deux cas (Van den Bergh *et al.* 2011; Bauer *et al.* 2009), le robot renseigne l'utilisateur à l'aide de messages sonores vocaux. Il s'agit soit de messages pré-enregistrés, soit de synthèse vocale (TTS-*Text To Speech*).

2.1.4.6 Synthèse et discussion

L'ensemble des références et les caractéristiques recensées pour chacune sont résumées en Table 2.1. Assez majoritairement, les gestes sémaphoriques sont utilisés en gardant

le drone à vue (local). La commande et le contrôle peuvent tous deux être utlisés, et les feedback privilégiés semblent être l'observation directe du comportement du drone, ou un retour vidéo lorsque le drone est distant.

Le feedback

Les types de feedback privilégiés semblent donc être l'observation du comportement du drone lorsque celui-ci est à vue, et un retour vidéo lors d'une interaction distante. Ces deux solutions requièrent une attention soutenue de la part de l'utilisateur et l'empêchent en conséquence de rester conscient de son environnement ; ce qui ne semble pas compatible du contexte militaire. Pour ce cas d'usage, un feedback sonore semble d'avantage adapté.

Ainsi, pour commander par geste un robot, différentes solutions existent mais aucune ne semble correspondre parfaitement à notre cas d'application. Aussi, un design de l'interaction spécifique est nécessaire et a été réalisé. Le détail des éléments qui le composent ainsi que la justification du choix de chacun vont maintenant être présentés.

TABLE 2.1 – Synthèse de la revue de la littérature des interactions gestuelles homme-robot.

Référence	Robot	Proximité	Mode	Geste	Feedback
(Lee & Xu 1996)	Générique	Non spécifié	Commande	Sémaphorique	Non spécifié
(Kortenkamp *et al.* 1996)	Mobile	Local	Commande	Sémaphorique	Naturel
(Triesch & Von Der Malsburg 1998)	Bras articulé	Local	Commande	Sémaphorique	Naturel
(Boehme *et al.* 1998)	Mobile	Local	Commande	Sémaphorique	Naturel
(Perzanowski *et al.* 1998)	Mobile	Local	Commande	Co-verbal déictique	Naturel
(Iba *et al.* 1999)	Mobile	Distant Local	Contrôle	Sémaphorique	Vidéo Naturel
(Waldherr *et al.* 2000)	Mobile	Local	Commande	Sémaphorique	Naturel
(Fong *et al.* 2001)	Mobile	Distant	Contrôle	Manipulation	Vidéo
(Rogalla *et al.* 2002)	Bras articulé	Local	Commande	Co-verbal déictique	Naturel
(Urban *et al.* 2004)	Mobile	Local	Commande	Sémaphorique	Non spécifié
(Guo & Sharlin 2008)	Mobile	Local	Contrôle	Sémaphorique	Naturel
(Bauer *et al.* 2009)	Mobile	Local	Commande	Sémaphorique déictique	Vocal
(Jones *et al.* 2010)	Drone virtuel	Local	Commande	Co-verbal	Naturel
(Van den Bergh *et al.* 2011)	Mobile	Local	Commande	Sémaphorique déictique	Vocal et naturel
(Ng & Sharlin 2011)	Drone	Local	Commande	Sémaphorique	Naturel
(Monajjemi *et al.* 2013)	Essaim de drones	Local	Commande	Sémaphorique	Naturel et voyants
(Pfeil *et al.* 2013)	Drone	Local	Contrôle	Sémaphorique	Naturel
(Sanna *et al.* 2013)	Drone	Local	Contrôle	Sémaphorique	Naturel
(Nahapetyan & Khachumov 2015)	Drone	Distant	Contrôle	Sémaphorique	Vidéo
(Mashood *et al.* 2015)	Drone	Local	Contrôle	Sémaphorique	Naturel
(Zhang & Liu 2015)	Drone virtuel	Distant	Commande	Sémaphorique	Vidéo
(Thinh *et al.* 2016)	Mobile	Local	Contrôle	Sémaphorique	Naturel
(Coupeté *et al.* 2016)	Bras articulé	Local	Commande	Manipulation	Naturel

2.2 Design spécifique de l'interaction

2.2.1 Le choix d'une interaction basée agent conversationnel

L'objectif principal est de proposer un modèle d'interaction qui permette à l'opérateur humain, d'utiliser un robot mobile tout en restant conscient de l'environnement qui l'entoure. On cherche donc à réduire le *coût* de l'interaction, c'est à dire à limiter l'usage de ressources physiques et cognitives qu'elle implique.

Pour cela, classiquement, deux grandes approches peuvent être considérées : la manipulation directe et les agents conversationnels (Shneiderman 1982; Shneiderman & Maes 1997).

La manipulation directe est un mode d'interaction nommé et décrit par Shneiderman en 1982 (Shneiderman 1981; Shneiderman & Plaisant 1987; Javed *et al.* 2011). Également appelé interaction basé monde, basé action (Frohlich) ou encore interaction à la première personne (Laurel 1986), la manipulation directe est caractérisée par :

— une représentation continue des objets d'intérêt,
— une action physique sur ces objets d'intérêt et
— les actions de l'utilisateur sont incrémentales (additives) et leurs effets sont immédiatement visibles.

Comme représenté en Figure 2.5(a), avec la manipulation directe, l'utilisateur agit directement sur l'environnement et on part du principe et accepte que ce soit de manière continue.

Ainsi, pour réduire le coût de l'interaction, l'enjeu est de proposer des métaphores et shèmes qui présentent les objets et les actions de manière cohérente et immédiatement compréhensible afin de les rendre évidents et donc d'en simplifier l'usage.

Hutchins et al. (Hutchins *et al.* 1985; Hutchins 1987) ont indiqué que ce mode d'interaction est particulièrement attractif. En effet, en réduisant la distance (au sens de Norman) entre l'utilisateur et la machine, l'engagement de l'utilisateur et la satisfaction de pouvoir agir directement sont augmentés. Il en résulte un sentiment de maîtrise du système.

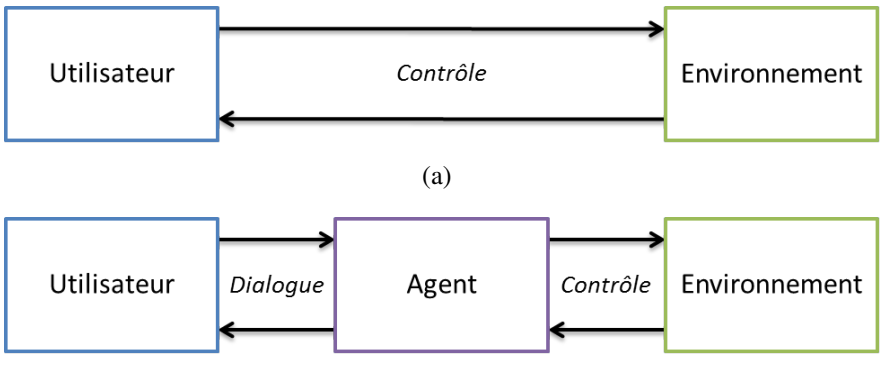


FIGURE 2.5 – Deux modes classiques d'interaction : (a) la manipulation directe crée un sentiment d'engagement et de contrôle et, (b) les agents conversationnels libèrent l'utilisateur en partie en prenant à leur charge la réalisation de certaines actions .

Toutefois certains auteurs (Frohlich & Luff 1990; Claassen *et al.* 1990) indiquent que la manipulation directe est particulièrement impactée par les phénomènes de latence perçue ; que ce mode s'applique difficilement aux objets invisibles, métaphoriques, aux groupes d'objets ou lorsque les actions n'ont pas un effet immédiat ; et que ce mode d'interaction est également peu adapté aux actions longues, répétitives et ne permet pas une planification ou une parallélisation des tâches.

Les agents conversationnels (Franklin & Graesser 1996) sont le second mode d'interaction. Également appelé mode langage (Frohlich & Luff 1990) ou interaction à la seconde personne (Laurel 1986), il s'agit de libérer l'utilisateur de tout ou partie de sa charge de travail par une délégation de celle-ci à un intermédiaire virtuel. Cet intermédiaire est appelé agent et est défini comme étant une entité virtuelle qui peut agir à la place de l'utilisateur et en son nom (Franklin & Graesser 1996).

Comme représenté en Figure 2.5(b), ce second mode d'interaction repose sur l'établissement d'un dialogue entre l'utilisateur et l'agent. Ce dialogue est réalisé grâce à un langage qui peut éventuellement être naturel. Les supports de ce langage peuvent être multiples : écrit, oral, gestuel. L'utilisateur écrit dans ce qui doit être fait ; la machine le réalise puis, informe du résultat.

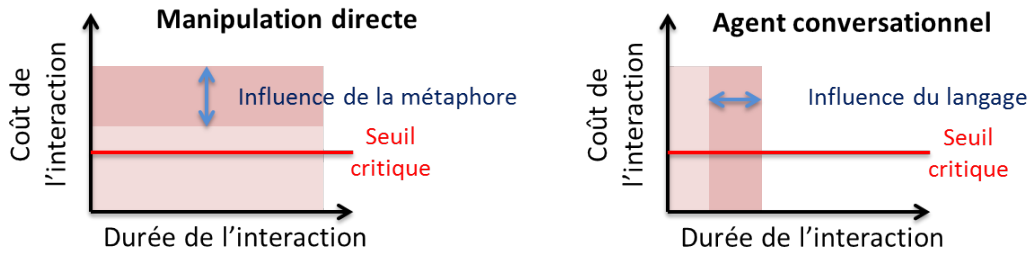
Différents types d'agents peuvent être définis. Par exemple une distinction est faite entre les agents cognitifs et les agents réactifs (Coutaz *et al.* 1996; Demazeau & Müller 1991). Les agents cognitifs cherchent à atteindre un but. Ils sont capables d'analyser un ensemble de données et de raisonner dessus pour prendre des décisions. Les agents réactifs, quant à eux, ont une capacité de raisonnement limitée. Ils n'ont pas de but précis mais suivent un comportement prédéfini par des règles explicitement implémentées lors du développement.

Ces deux grands modes d'interaction ont fait l'objet de nombreux débats à la fin des années 1990 (Shneiderman & Maes 1997). Finalement, il est apparu que chacun présente des spécificités complémentaires qui peuvent être utiles à des applications différentes. Aussi, dans le cadre des robots mobiles, ces deux modes d'interaction semblent chacun convenir à un type de pilotage : la manipulation directe est particulièrement pertinente pour la téléopération et les agents conversationnels sont adaptés à la supervision.

Alors, de ces deux modes, lequel utiliser ? De manière simpliste, la manipulation directe et les agents conversationnels sont deux modes d'interaction qui influencent le coût d'une interface. Cependant, comme présenté en Figure 2.6, chacun le fait de manière différente : la manipulation cherche à réduire la charge imposée par l'interface à chaque instant alors que les agents permettent de réduire le temps total de l'interaction.

Quel serait l'impact de l'usage de ces deux modes en environnement hostile ? Le choix de la manipulation directe conduit à charger et immobiliser un opérateur, certes moins mais pendant toute la durée du déplacement du drone. Tactiquement, cela impose à l'opérateur de tomber à l'arrêt ce qui requiert une préparation et une protection de l'ensemble de son équipe. Le choix de l'agent conversationnel, quant à lui, rend la commande rapide et peu fréquente. Bien qu'en théorie, une communication constitue un élément de distraction, si la durée est suffisamment courte, il n'est pas alors nécessaire d'immobiliser toute l'équipe.

Le mode d'interaction basé agent conversationnel semble donc le mieux adapté. En effet, l'utilisation d'un plan de navigation (supervision) par l'intermédiaire d'un simple



(a) (b)

FIGURE 2.6 – La manipulation directe et les agents conversationnels influencent de manière différente le coût de l'interaction : (a) la manipulation permet une réduction de la charge à chaque instant alors que (b) les agents cherchent à réduire la durée de l'interaction.

dialogue, permet une interaction moins coûteuse pour l'opérateur au cours de la mission, et une progression plus fluide de toute l'équipe sur le champs de bataille.

L'opérateur peut donner des consignes de manière réactive et retourner rapidement à sa mission première. De cette manière, le robot mobile peut ne plus être un outil contraignant mais devenir un coéquipier partiellement autonome supervisé par un, voir plusieurs opérateurs.

2.2.2 Le choix des gestes sémaphoriques comme type de geste

Il a donc été choisi d'utiliser un mode d'interaction basé agent conversationnel. Ce mode repose sur une communication, un dialogue, entre l'utilisateur et la machine. Dans ce dialogue, l'utilisateur exprime donc l'intention d'activer des commandes de haut niveau à la machine ; et pour cela, il requiert une modalité d'entrée.

Puisque les gestes sont employés par les militaires et semblent présenter différentes propriétés intéressantes, il a été proposé d'explorer cette piste. Aussi, c'est par ce moyen que l'utilisateur va agir sur l'état du système.

Or, différentes catégories de gestes existent : les geste manipulatifs, co-verbaux, sémaphoriques et les langues des signes. Alors, quel type de geste choisir ?

En premier lieu, et de manière évidente, ce sont les gestes de type communicatif qui vont être considérés. En effet, la fonction sémiotique, que remplissent ces gestes, permet à l'utilisateur d'exprimer son intention. Les gestes manipulatifs quant à eux, ne semblent pas convenir dans la mesure où ils ne permettent pas d'exprimer une intention ou tout du moins pas directement. Ils semblent d'avantage convenir à un mode d'interaction de type manipulation directe (téléopération) (Fong *et al.* 2001; Pfeil *et al.* 2013)

Parmi les gestes communicatifs, les co-verbaux sont souvent considérés comme les plus intuitifs étant donné que ce sont les plus utilisés par l'homme au quotidien (McNeill 1992). Depuis l'article de référence "Put that there" de Bolt (Bolt 1980), ce type de geste a beaucoup été proposé comme modalité (ou multi-modalité) permettant à la fois la modification de l'état du système (Perzanowski *et al.* 2001) et la perception de l'évolution de cet état

(Stiefelhagen *et al.* 2004). En redondance ou complément de la parole, cela permet de préciser des informations visuo-spatiales et, par métaphore, des informations temporelles. Les co-verbaux sont donc une multi-modalité particulièrement intéressante qui, cependant, ne convient pas au contexte des drones de contact du fait de l'usage de la parole. En effet, la modalité verbale nuit à la discrétion et perturbe les canaux de communication au sein de l'équipe. Il n'est donc pas possible, à priori, de les utiliser pour cette application.

Restent alors les gestes sémaphoriques et les langues des signes. Tous deux permettent l'expression volontaire d'intentions par la réalisation de formes standardisées prenant le corps pour seul support. La distinction majeure qui existe entre ces deux catégories est que les gestes sémaphoriques sont des éléments isolés alors que les langues des signes sont structurées. Les langues des signes, de fait, possèdent une grande expressivité. Cependant, cela va de pair avec une plus grande complexité. Les co-verbaux sont quant à eux simples et immédiats : la réalisation d'un unique geste permet d'exprimer une intention précise dans un temps très court. Or, l'usage d'un plan de navigation requièrent seulement l'activation d'un nombre réduit de fonctions de haut niveau. Aussi, la complexité des langues des signes ne semble pas requise et, de fait, les gestes symboliques sont le type de geste qui semble le mieux adapté.

La modalité d'interaction proposée consiste donc, pour commander un drone de contact, de réaliser un geste sémaphorique particulier. Ce geste est détecté et reconnu par l'agent conversationnel qui active alors et exécute la fonction de haut niveau associée.

2.2.3 Ajout d'un feedback et d'un mécanisme de sécurisation

Un mode d'interaction de type agent conversationnel et l'usage d'une modalité gestuelle sémaphorique ont été choisis. On pourrait alors penser que cela est suffisant et adapté. Cependant, une *évaluation heuristique* révèle que deux éléments importants semblent manquer : un feedback et un mécanisme de sécurisation.

2.2.3.1 Évaluation heuristique

Pour déterminer si une interface, ou une proposition d'interface est adaptée, on peut effectuer une recherche des éventuels problèmes d'ergonomie en réalisant une évaluation heuristique, également appelée audit d'ergonomie (Cockton *et al.* 2008). Il s'agit alors de contrôler le respect d'un ensemble de règles pratiques : les heuristiques.

Différentes listes de règles ont été proposées par différents auteurs ; par exemple les 8 *Golden rules* de Shneiderman (Shneiderman & Plaisant 1987), les 7 principes de Norman (Norman 1988) issus de la *théorie de l'action* du même auteur, les 16 principes de Tognazzini (Tognazzini 1993), les critères ergonomiques de Bastien et Scapin (Bastien & Scapin 1993) ou encore, les 10 heuristiques de Nielsen (Nielsen 1995).

Parmi ces différentes possibilités, pour sa simplicité, nous avons choisi d'utiliser la liste proposée par Nielsen, composée des 10 règles suivantes :

1. **Visibilité de l'état du système.**
   Le système doit toujours tenir l'utilisateur informé de ce qui se passe, par un retour

approprié.

2. **Correspondance du système avec le monde réel.**
   Le système doit parler le langage de l'utilisateur, avec des mots, des phrases et des concepts qui lui sont familiers, plutôt que d'utiliser un langage propre au système.
3. **Liberté et contrôle pour l'utilisateur.**
   Les utilisateurs peuvent faire des erreurs. Le système doit donc proposer des *sorties de secours* claires telles que l'annulation, défaire/refaire.
4. **Cohérence et standards.**
   L'utilisateur ne doit pas avoir à se poser des questions pour savoir si différents mots, situations ou actions signifient la même chose. Il faut suivre les conventions liées à la plate-forme.
5. **Prévention des erreurs.**
   La conception doit anticiper les problèmes que pourrait rencontrer l'utilisateur.
6. **Reconnaître plutôt que se souvenir.**
   Rendre visibles les objets, les actions et les options. Pour éviter une surcharge de la mémoire de travail, il est préférable d'éviter de demander à l'utilisateur de se souvenir d'une information, d'une séquence de dialogue à une autre.
7. **Flexibilité dans l'utilisation.**
   La conception du système doit tenir compte des différents profils d'utilisateurs. Notamment des utilisateurs experts et novices en proposant des contrôles adaptés à chacun. Les raccourcis par exemple, permettent de gagner en performance pour les utilisateurs experts.
8. **Esthétique et design minimaliste.**
   Les dialogues ne doivent pas proposer d'informations qui ne sont pas pertinentes au risque de perdre l'utilisateur. Chaque information étant en concurrence avec les autres, un trop grand nombre diminue la lisibilité globale.
9. **Faciliter l'identification, le diagnostic et la récupération des erreurs par l'utilisateur.**
   Les messages d'erreur doivent indiquer, en langage clair, le problème et suggérer une solution pour le résoudre.
10. **Aide et documentation.**
    Lorsqu'il n'est pas possible d'afficher toutes les informations dans l'interface, de l'aide peut être apportée via une documentation. Celle-ci doit être facile à trouver, centrée sur les tâches de l'utilisateur, indiquer concrètement les étapes à suivre et ne pas être trop longues.

2.2.3.2 Un feedback audio pour informer de l'état du système

— (heuristique 1) Visibilité de l'état du système
— (heuristique 8) Esthétique et design minimaliste

Dans le modèle d'interaction proposé, seule une modalité en entrée du système a été présentée. Comment savoir alors, pour l'utilisateur, si le système comprend bien les commandes et si il les exécute ? La revue de la littérature des interactions homme-robot a montré que dans la majorité des cas, c'est l'observation et la compréhension du comportement du drone qui servent de feedback.

Cependant, cela suppose que le drone reste à vue de l'utilisateur ce qui n'est pas nécessairement le cas sur le champs de bataille puisque son usage est d'acquérir de l'information dans les zones hors du champs de vision. De plus, cela impose à l'utilisateur de rester attentif au comportement du drone pendant toute la durée de son déplacement ce qui ne convient pas à la problématique de permettre une bonne conscience de l'environnement.

En réponse à ce problème, et puisque le mode d'interaction repose sur un agent conversationnel, il semble pertinent de munir celui-ci d'une modalité événementielle sonore. Par des messages audios (sons iconiques et synthèse vocale), l'utilisateur peut être averti par l'agent des informations pertinentes comme la bonne prise en compte d'une commande et de l'évolution de son exécution. Ainsi, l'utilisateur est complètement déchargé des tâches de contrôles à la fois en entrée et en sortie.

On peut cependant s'interroger sur la discrétion d'un retour sonore. En réalité, ceci ne constitue pas un réel problème puisque différents dispositifs techniques permettent un retour sonore individuel. Par exemple, les casques ostéophoniques (à conduction osseuse) permettent à un utilisateur d'entendre des informations audio sans qu'elles ne soit audibles par d'autres, et sans masquer l'environnement sonore.

2.2.3.3 Un mécanisme de confirmation des commandes

— (heuristique 3) Liberté et contrôle pour l'utilisateur
— (heuristique 5) Prévention des erreurs
— (heuristique 9) Faciliter l'identification, le diagnostic et la récupération des erreurs par l'utilisateur

De manière générale, il est fondamental qu'un système soit fiable. Cela est d'autant plus vrai lorsque la sécurité de l'utilisateur et de son équipe est engagée. Ainsi, il semble pertinent considérer l'existence d'erreurs, c'est à dire de les anticiper et de permettre un traitement simple et rapide de celles-ci.

Le modèle d'interaction basé agent, à modalité gestuelle en entrée et sonore en sortie semble présenter 3 types d'erreur :

— Le premier type d'erreur est lorsque l'utilisateur, veut exprimer une intention par geste, mais que celle si n'est pas comprise par l'agent. Soit parce que le geste employé est inconnu du système ou mal reconnu, soit parce que le geste est mal réalisé. L'utilisateur se retrouve alors confronté à un système qui ne répond pas et ce sans élément de compréhension.

— Le second type d'erreur est lorsque l'utilisateur exprime par geste une intention, que l'agent comprend, mais dont l'exécution échoue. L'exécution peut échouer pour des raisons environnementales imprévisibles (vent, obstacle), ou pour des raisons contextuelles (énergie faible ou commande incohérente).
— Enfin, le troisième type d'erreur est lorsque le système détecte une intention non exprimée par l'utilisateur, ou une intention qui n'est pas la bonne. Cela peut provenir d'une défaillance technique ou d'une réalisation involontaire d'un geste.

Les deux premiers types d'erreur semblent peu critiques et simples à traiter. En effet, dans ces deux cas, un message sonore peut être généré pour indiquer la présence d'une erreur et, lorsque cela est possible, expliquer sa cause.

Le troisième type d'erreur est cependant bien plus contraignant. En effet, que le drone ne réagisse pas est une source de frustration évidente, mais qu'il débute un déplacement de manière inattendue, ou un mauvais déplacement, peu avoir de lourdes conséquences. Le drone peut être trop proche d'un individu ou d'un bâtiment et ainsi risquer de blesser ou d'endommager. Il peut également trahir la présence de l'utilisateur en ayant un comportement stratégiquement inadéquat.

Une commande involontaire doit donc impérativement être évitée. Et pour cela, nous proposons l'utilisation d'un mécanisme de confirmation dont le fonctionnement est le suivant :

Lorsque l'agent détecte une intention de la part de l'utilisateur, il émet en premier lieu, un message audio qui indique quelle commande a été comprise. Puis, l'agent demande à l'utilisateur de confirmer (par geste) cette commande. Si l'utilisateur confirme, l'agent entreprend alors d'exécuter celle-ci et l'indique grâce à un élément sonore. Si l'utilisateur annule la commande, l'agent se contente d'émettre un élément sonore différent. Enfin, si l'utilisateur ne confirme ni n'annule la commande après un délai de quelques secondes, celle-ci est automatiquement abandonnée. En effet, l'utilisateur peut ne pas être disponible et donc ignorer les messages en provenance de l'agent. L'agent considère alors que la détection était erronée ou que l'activation de la commande n'est plus une priorité.

Un schéma synthétique de ce mécanisme de confirmation qui permet la sécurisation de l'activation des commandes, est présenté en Figure 2.7.

2.2.4 Synthèse du modèle d'interaction proposé

Avec pour objectif de proposer un modèle d'interaction qui permette d'utiliser un robot mobile tout en restant conscient de l'environnement qui l'entoure et plus spécifiquement pour le cas des drones de contact, différentes possibilités ont été étudiées.

Les choix réalisés, ont conduit à la composition d'un modèle d'interaction qui libère l'utilisateur en délégant une grande part de sa charge de travail à un agent. Cette délégation est réalisée grâce à l'établissement d'un dialogue simple et sécurisé.

Le modèle proposé repose donc sur 5 éléments fondamentaux. Ces derniers, résumés en Figure 2.8, sont les suivants :

— **Usage d'un plan de navigation** :
Le pilotage reposant sur l'usage d'un plan de navigation permet, au cours d'une

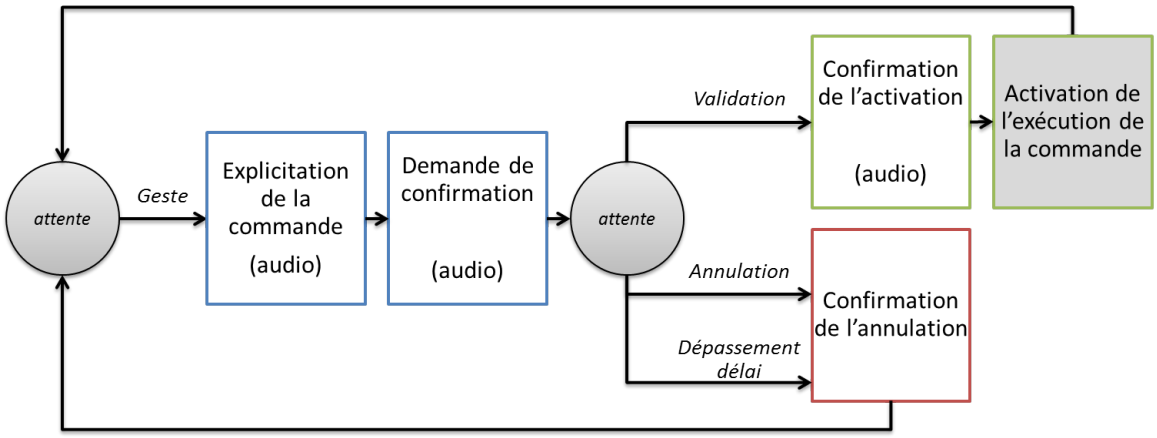


FIGURE 2.7 – Mécanisme de confirmation des commandes qui protège des activations involontaires.

mission, de seulement superviser la progression du drone sur un itinéraire. Un plan de vol détaillé est défini au cours de la préparation de mission. Puis, au cours de la mission, seule l'activation de fonctions de haut niveaux (Décoller, Atterrir, Suivant, Précédent, Base et Stop) permet rapidement de faire progresser le drone.

— **Dialogue avec un agent logiciel** :
L'exécution détaillée de ces fonctions de haut niveau, pour ne plus être à la charge de l'utilisateur, est déléguée à un agent logiciel. L'interaction entre cet agent et l'utilisateur est alors réduit à un dialogue rapide, formel et sécurisé. Ce dialogue se résume à une expression de l'intention d'activer une fonction de haut niveau par l'utilisateur, et à une description des éléments pertinents de l'état du système par l'agent.

— **Modalité gestuelle symbolique en entrée** :
Dans le but de libérer l'utilisateur de tout périphérique le contraignant physiquement et cognitivement, il a été fait le choix de la commande gestuelle. Parmi les différents types de gestes, la catégories des gestes symboliques semble la mieux adaptée. C'est donc par ce moyen que la communication de l'utilisateur à l'agent est permise.

— **Modalité sonore en sortie** :
L'agent informe donc l'utilisateur des éléments de l'état du système qui sont importants pour lui. Pour cela, l'agent est doté d'une modalité sonore à la fois vocale (synthétique) et iconique.

— **Sécurisation de l'activation des commandes** :
Afin d'empêcher la prise en compte de l'expression involontaire d'une commande ou d'une mauvaise commande, un mécanisme de sécurité est proposé. L'agent, lors de la perception d'une intention de l'utilisateur, initie une courte phase de dialogue. Il indique à l'utilisateur pour quelle commande il a détecté une intention, puis, requiert une confirmation de celle-ci. Si une réponse négative ou qu'aucune réponse n'est donnée par l'utilisateur dans les secondes qui suivent, la commande est abandonnée. En cas de réponse positive, l'agent entreprend l'exécution de la commande. Dans

les deux cas, l'agent signale par un son distinctif le résultat de la confirmation.

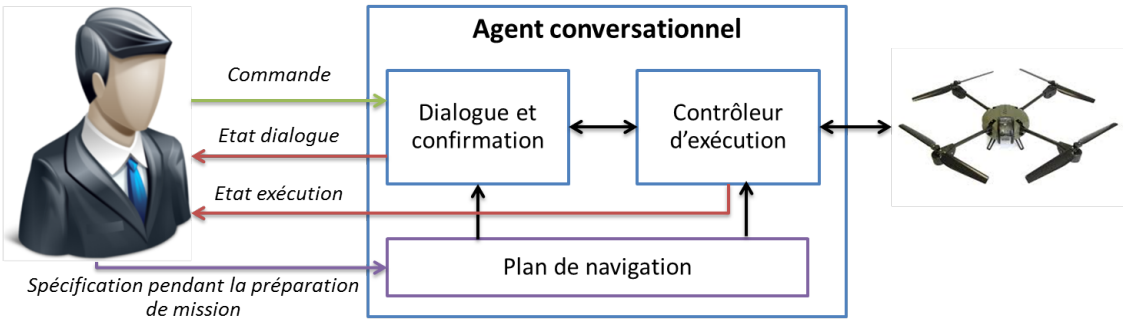


FIGURE 2.8 – Synthèse du modèle d'interaction proposé. Il est basé sur un agent conversationnel utilisant une modalité gestuelle symbolique en entrée et sonore en sortie.

CHAPITRE 3

Protocole pour la construction d'un vocabulaire gestuel consensuel et non ambigu

Sommaire

3.1 L'importance du choix des gestes

Dans le contexte d'une interaction gestuelle sémaphorique appliquée à la commande d'un système, des fonctionnalités d'une machine sont activées par la réalisation consciente et volontaire d'un geste par l'utilisateur. Comme les raccourcis claviers des interfaces logicielles, ce sont des commandes expertes qui doivent leur efficacité à la rapidité de leur emploi. Elles doivent être utilisables de manière immédiate et en toute circonstance.

Pour cela, il est nécessaire, que les gestes et les fonctions qu'ils commandent soient appris. A notre connaissance, aucune preuve n'existe quant à une limite du nombre de gestes que l'on peut apprendre. Toute fois, un grande nombre de couples geste-commande

et la complexité de l'environnement dans lequel ils sont utilisés impactent la durée de l'apprentissage et la facilité de leur rappel.

Le caractère sémantique des gestes est également un élément important à considérer. En effet, moins l'association entre la forme d'un geste et le concept qu'il représente est logique pour l'utilisateur, plus il est difficile de la mémoriser (Jégo *et al.* 2013).

Pour que l'apprentissage et l'utilisation d'une commande symbolique soit possible, il est donc fondamental que les associations entre les gestes et les fonctions soient logiques. Cependant, il a été montré que les gestes ne sont pas universels. Ils dépendent de la culture et du contexte (Archer 1997; McNeill 1985; Calbris 1990). Aussi, il est nécessaire de choisir les gestes ; et la manière dont cela est fait est particulièrement important.

3.2 Méthodes existantes de choix des gestes

Classiquement, trois approches existent pour le choix d'un dictionnaire de gestes : les méthodes techno-centrées, méthodes centrées utilisateur et les méthodes mixtes.

Les méthodes techno-centrées sont certainement les plus employées. Dans la plupart des cas, les gestes sont proposés par les développeurs des systèmes en cherchant à maximiser les taux de reconnaissance. Les gestes, quelque soit leur forme, sont donc ainsi adaptés à la machine et non à l'utilisateur. Ils sont, pour cette raison, souvent plus difficiles à utiliser : fatiguants et complexes à mémoriser.

L'approche opposée des méthodes techno-centrées est l'approche centrée utilisateur. Celle-ci consiste à choisir des gestes qui soient sémantiquement les plus intuitifs et logiques via des expériences utilisateurs. L'objectif est donc de définir des gestes pertinents pour un utilisateur quitte à négliger dans un premier temps la faisabilité technique de leur reconnaissance. Pour procéder, il est possible de proposer naïvement un ensemble de gestes puis, de manière itérative, de les faire évaluer et modifier par des participants (Pfeil *et al.* 2013; Ng & Sharlin 2011). Il est également possible d'utiliser des méthodes d'élicitation : c'est à dire de faire proposer les gestes directement par des utilisateurs (Nielsen *et al.* 2003; Choi *et al.* 2012). On cherche alors à maximiser le consensus (Wobbrock *et al.* 2005; Wobbrock *et al.* 2009; Vatavu & Wobbrock 2015), c'est à dire le fait que les utilisateurs s'accordent sur les gestes à utiliser ; ce qui semble être un gage d'intuitivité.

Enfin, certains auteurs (Stern *et al.* 2004; Stern *et al.* 2008a; Stern *et al.* 2008b) ont proposé une méthode intermédiaire. Cette méthode *mixte* prend à la fois en compte la problématique humaine (intuitivité et fatigue) et la problématique technique (robustesse). Dans un premier temps, un ensemble de participants élicitent un geste par fonction. Ainsi, on obtient un ensemble de gestes utilisables associés à un score d'intuitivité : le nombre de fois qu'un geste a été proposé pour chaque fonction. Puis, pour chacun des gestes recueillis, un score de fatigue est calculé à l'aide d'une formule bio-mécanique (proposée par les auteurs) et un taux de reconnaissance estimé par un système d'apprentissage automatique. Finalement, un algorithme d'optimisation permet de déterminer le vocabulaire gestuel qui, globalement, maximise l'intuitivité et la robustesse et minimise la fatigue. Cette approche utilisable pour

des gestes élémentaires, semble toutefois difficile à mettre en oeuvre pour des gestes plus complexes ; en effet, le critère de fatigue ne semble pas applicable et difficiles à adapter. De plus, cette méthode impose de disposer d'un système de reconnaissance générique.

Finalement, les méthodes techo-centrées ne permettant pas de constituer des vocabulaires gestuels sémantiquement satisfaisants, et les méthodes mixtes ne semblant pas applicables simplement, nous avons fait le choix d'une approche centrée utilisateur.

3.3 Deux approches centrées utilisateur à priori complémentaires

Nous avons donc fait le choix d'une méthode centrée utilisateur, et c'est ce que proposent Nielsen et al. (Nielsen *et al.* 2003) et Choi et al. (Choi *et al.* 2012) dans leurs travaux respectifs.

Nielsen et al. (Nielsen *et al.* 2003) ont proposé une approche en trois phases.

1. Collecter des gestes de la part d'utilisateurs finaux : Dans un premier temps, des utilisateurs participent individuellement ou en groupe à des séance de proposition de gestes. Ces séances peuvent être soit de simples entretiens verbaux, soit des simulations scénarisées. Toutes les propositions de gestes sont filmées pour être analysées.
2. Création du dictionnaire de gestes qui comporte exactement un geste par fonction : Des experts analysent les gestes proposées lors de la première phase. Les gestes identiques proposés par différents participants sont regroupés pour former un unique geste candidat. Une fois tous les gestes candidats constitués, les plus fréquemment proposés sont retenus pour le dictionnaire gestuel.
3. Évaluer le dictionnaire : Un petit nombre de participants essaye le dictionnaire pour en évaluer l'utilisabilité.

Choi et al. (Choi *et al.* 2012) ont proposé une approche similaire qui, cependant, ajoute une phase d'élection après la phase d'élicitation et retire la phase d'évaluation de la qualité du dictionnaire constitué. Il y a donc trois phases :

1. Des gestes sont élicités par des utilisateurs finaux.
2. Les gestes identiques sont regroupés pour former des candidats qui sont ajoutés à un catalogue.
3. Les participants procèdent ensuite à l'élection des gestes du dictionnaire sur la base du catalogue.

Un score de consensus est calculé après la seconde et la troisième phase. Il apparaît que le niveau de consensus augmente et est donc supérieur après élection. Les auteurs justifient ce résultat en proposant l'hypothèse suivante : les participants ne seraient pas nécessairement capables de proposer seuls et instantanément les meilleurs gestes.

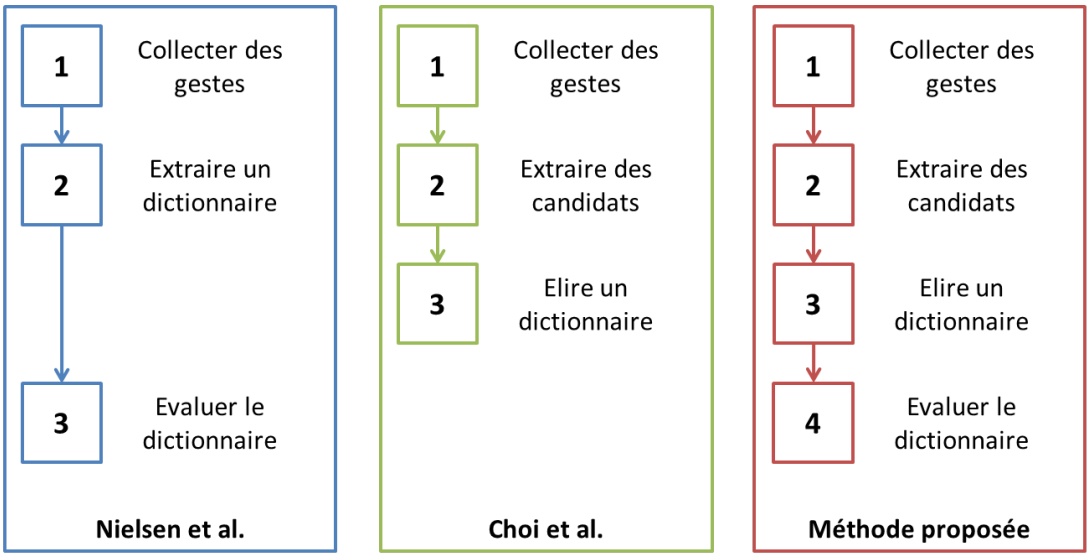


FIGURE 3.1 – Trois approches pour la construction d'un dictionnaire de gestes. De gauche à droite, l'approche présentée par Nielsen et al. (Nielsen *et al.* 2003), une expérimentation présentée par Choi et al. (Choi *et al.* 2012) et la méthode proposée ici.

En effet, la phase d'élicitation ne permet pas un échange d'idée entre les participants ce qui limite la créativité et l'expérience globale. Aussi, en ajoutant une phase d'élection, on permet de manière implicite à chacun de profiter des proposition des autres.

Cependant, augmenter le consensus ne protège pas contre la présence d'ambiguïtés au sein du dictionnaire élu. Par conséquent il semble important de conserver une phase de validation du dictionnaire en fin de protocole.

Ces deux méthodes semblent particulièrement adaptées et complémentaires. Aussi, nous proposons une nouvelle méthode combinant ces deux approches. La Figure 3.1 présente un résumé graphique de ces trois méthodes et met en évidence leurs similitudes et différences.

3.4 Méthode proposée

3.4.1 Plan du protocole

Sur la base des travaux complémentaires de Nielsen et al. (Nielsen *et al.* 2003) et de Choi et al. (Choi *et al.* 2012), la méthode proposée comporte donc les quatre phases suivantes :

1. Collection des gestes proposés par un premier groupe de participants.
2. Extraction, par des experts qui identifient les gestes identiques, des gestes candidats pour former un catalogue.
3. Élection du dictionnaire par un nouveau groupe de participants.
4. Évaluation de la non-ambiguïté des gestes élus par un troisième groupe de participants.

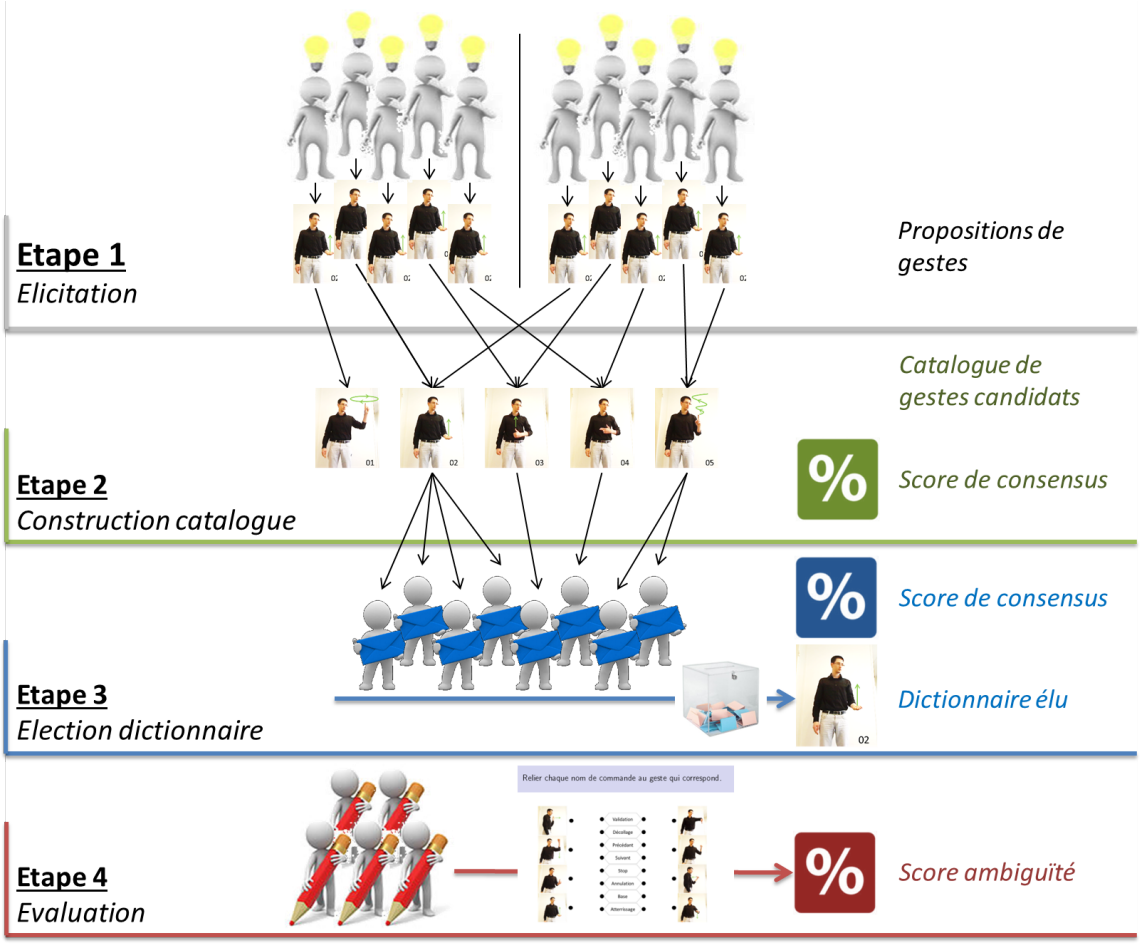


FIGURE 3.2 – Résumé des quatre étapes qui composent le protocole proposé pour construire un dictionnaire de gestes consensuel et non ambiguë.

Nous avons appliqué cette méthodologie au contexte de la commande sémaphorique d'un drone de contact militaire. Les trois groupes de participants, les différentes phases et les résulats obtenus vont maintenant être présentés. Un résumé schématique de l'ensemble de ce protocole est également proposé en Figure 3.2.

3.4.2 Participants

Un total de 39 participants a pris part à ce protocole de choix de gestes. Ils ont été répartis en trois groupes distincts :

1. Le premier groupe était composé de 20 volontaires d'âge moyen 38.6 ans (S.D.=14.4). Ce premier groupe comportait 10 participants civils, tous étudiants à Mines-Paristech, et 10 participants militaires, conseillers opérationnels de SAGEM.
2. Le second groupe était composé de 9 participants militaires d'âge moyen 42.2 ans (S.D.=4.7).

3. Le troisième groupe était composé de 10 participants militaires d'âge moyen 45.9 ans (S.D.=4.1).

Tous les participants militaires des deuxième et troisième groupes sont des volontaires de la Section Technique de l'Armée de Terre Francaise (STAT).

3.4.3 Étape 1 - Collecter des propositions de gestes

La première étape du protocole a pour but de collecter un ensemble de gestes proposé par des participants et pour chaque fonction. Pour cela, chaque volontaire du premier groupe participe individuellement à une séance d'élicitation dont l'organisation est illustrée en Figure 3.3.

Dans un premier temps, le contexte est rapidement présenté : *collecter des gestes intuitifs pour commander un UAV*. Ensuite, le système, ses 8 fonctions et quelques consignes pour l'élicitation sont expliqués oralement.

Les différentes fonctions sont ensuite rappelées, une par une, et dans un ordre aléatoire afin d'éviter un éventuel effet d'ordre (d'ancrage). Pour chaque fonction, un unique geste est collecté. Il est demandé aux participants de proposer un geste utilisant une seule main, discret et surtout, le plus intuitif possible ; c'est à dire qui semble le plus logique et immédiat.

Pour cela, ils disposent d'autant de temps qu'ils souhaitent. Ils sont invités à imaginer différents gestes et à les essayer afin de bien se les approprier.

Une fois un geste choisi, le participant est invité à le réaliser en étant filmé par deux caméras (une de face, et une de profil). Les flux de ces deux caméra sont mixés et enregistrés avec le logiciel libre *Open Broadcaster Software* (OBS). Enfin, il est demandé au participant d'expliquer les éléments du geste choisi : quels sont les éléments importants et quelle peut être leur signification (métaphore sous-jacente).

Lors de cette première étape, un groupe de participants civils est ajouté au groupe de participants militaires. L'objectif principal est d'augmenter la taille de la population globale pour pouvoir recueillir plus de propositions de gestes. Un objectif secondaire est également de faire proposer des gestes par une population à priori non représentative des utilisateurs finaux (civils vs militaires) afin de recueillir des gestes ne reposant pas sur des représentations expertes. Ces propositions pourraient être différentes, voir plus simples.

3.4.4 Étape 2 - Constitution du catalogue des gestes candidats

Une fois la première phase d'élicitation terminée, l'ensemble des enregistrements est analysé par un opérateur expert pour identifier les gestes candidats. Tous les gestes présentant la même configuration spatiale sont considérés comme identiques et ajoutés au catalogue en tant que geste candidat.

Pour chaque geste candidat, une vidéo et une image portant le numéro du geste candidat sont également créés. Il serviront lors des étapes 4 et 5.

Enfin, le nombre de propositions d'un même geste candidat est calculé ainsi que le score de consensus moyen du catalogue (A1) à partir des valeurs de consensus de chaque

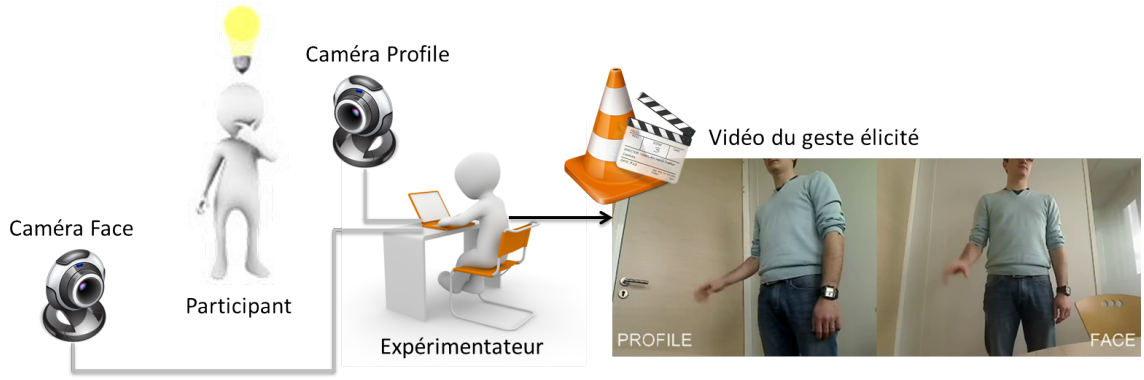


FIGURE 3.3 – Illustration de la configuration pour les séances d'élicitation de gestes.

fonction. Le score de consensus de chaque fonction est calculé avec la formule proposée par Vatavu et Wobbrok (Vatavu & Wobbrock 2015) :

$$AR_r = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 - \frac{1}{|P|-1} \tag{3.1}$$

Lors de l'élicitation, pour chaque fonction (ou référent : décoller, atterrir, ...) r, un ensemble de propositions de gestes sont recueillies P. Le nombre de ces propositions est $|P|$ et parmi elles, certaines sont identiques et forment des sous-groupes P_i de taille $|P_i|$. Finalement pour calculer le score de consensus du dictionnaire, on réalise la moyenne des scores de consensus sur l'ensemble des fonctions.

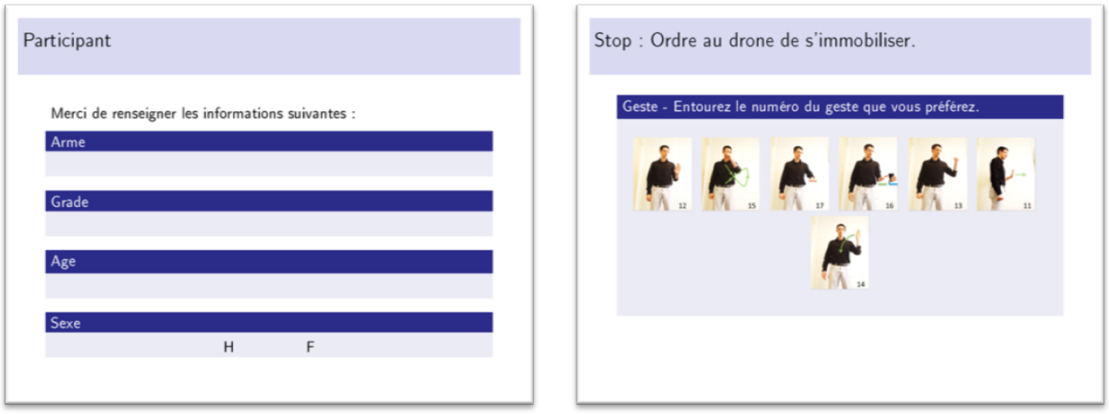
3.4.5 Étape 3 - Élection du dictionnaire

Le but de cette troisième étape est d'élire, dans le catalogue, un geste candidat par fonction. Pour ce faire, les participants sont réunis dans une salle de réunion disposant d'un vidéo-projecteur.

Après une brève introduction de l'objectif de la séance, du drone, des fonctions adressées et de l'origine du catalogue, chaque participant renseigne individuellement un questionnaire papier composé de 9 pages. La première page est un questionnaire informel précisant l'âge le sexe et le grade du participant. Sur les 8 autres pages, une par fonction, le nom de la fonction, une courte description (textuelle), ainsi que les différents gestes candidats sont présentés. Le questionnaire papier utilisé est présenté en Figure ‘3.4.

Chaque geste candidat est représenté par une image associée à un numéro. Sur chaque feuille, et pour chaque participant, les gestes sont ordonnés différemment afin d'éviter un éventuel effet d'ordre. Pour indiquer leur préférence, les participants doivent entourer l'image correspondante.

Pour s'assurer que les gestes candidats sont correctement compris par les participants, un film de leur exécution par un acteur est projetée en continu et, pour une bonne compréhension, chaque vidéo est référencée avec le numéro présent sur les questionnaires. Les participants sont invités à pratiquer les différents gestes candidats afin de se les approprier.



(a) (b)

FIGURE 3.4 – Questionnaire papier renseigné par les participants pour élire le dictionnaire de gestes. Il était demandé aux participants de compléter les informations personnelles (a) et, pour chaque commande, d'indiquer le geste préféré parmi les gestes du catalogue (b).

Une fois la session terminée, les questionnaires sont dépouillés : le nombre de sélection (vote) de chaque geste candidat est comptabilisé, un score moyen de consensus (A2) est calculé et comparé à celui de la seconde étape (A1). Enfin, pour chaque fonction, le geste candidat ayant été le plus sélectionné est retenu pour la constitution du dictionnaire de gestes.

3.4.6 Étape 4 - Évaluation du dictionnaire

La dernière étape a pour objectif d'évaluer la bonne compréhension des gestes élus et que leur association avec les fonctions est logique pour des personnes n'ayant pas participé, ni aux propositions, ni à l'élection.

Pour ce faire, les participants sont réunis dans une salle de réunion disposant d'un vidéo-projecteur. Après une présentation du contexte et des objectifs de la séance, il est demandé aux participants de renseigner individuellement un questionnaire papier.

Le questionnaire présente les noms des fonctions et une représentation des gestes élus dans un ordre aléatoire. Il est demandé a chaque participant de relier (par un trait) chaque geste au nom de la fonction qui semble le mieux lui correspondre. Un exemple de ce questionnaire est présenté en Figure 3.5.

Après la séance, le nombre de chaque association geste-fonction est comptabilisé et le taux d'association correcte est calculé (le nombre de bonne association divisé par le nombre total d'associations). Un fort taux de bonne association est considéré comme indicateur d'un dictionnaire peu ambigu.

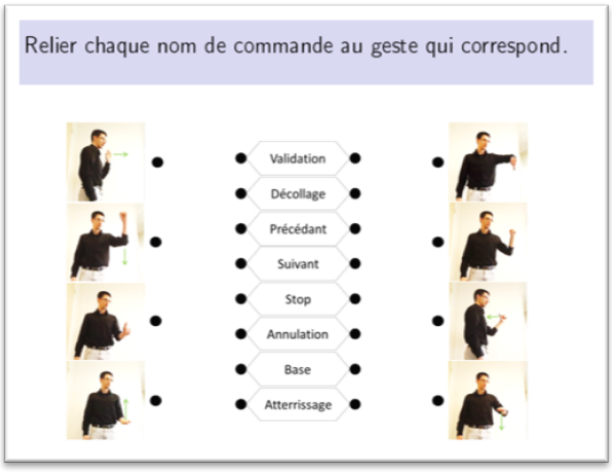


FIGURE 3.5 – Questionnaire papier renseigné par les participants lors de l'étape de validation du dictionnaire gestuel.

3.5 Résultats

3.5.1 Étapes 1 et 2 : Catalogue

Lors de la première étape, 160 propositions de gestes (8 fonctions fois 20 participants) ont été collectées. Après analyse lors de la seconde étape, 46 gestes candidats ont été identifiés et ajoutés au catalogue. L'ensemble de ce catalogue, organisé par fonction, est présenté en Annexe A.

La Figure 3.7(a) présente le nombre de gestes candidats différents par fonction, ainsi que le nombre de participants ayant proposé chaque geste candidat. A partir de ces données, le niveau de consensus moyen calculé est $A_1 = 24\%$. Au regard de l'échelle de lecture proposée par Vatavu et al. (Vatavu & Wobbrock 2015), ce niveau de consensus est *moyen* (de 10% à 30%).

La Figure 3.7(c) présente l'origine de chaque geste candidats :

— Les gestes candidats *civils* sont les gestes proposés uniquement par des participants civils.

— Les gestes candidats *militaires* sont les gestes proposés uniquement par des participants militaires.

— Les gestes candidats *mixtes* sont les gestes proposés à la fois par des participants civils et militaires.

3.5.2 Étape 3 : Dictionnaire

Lors de la troisième étape, un geste candidat a été élu à la majorité relative pour chaque fonction. Le dictionnaire des gestes élus est présenté Figure 3.6.

La Figure 3.7(b) présente le nombre de voix allouées à chaque geste candidat. A partir de ces données, le niveau de consensus moyen calculé est $A_2 = 32\%$. A nouveau au regard

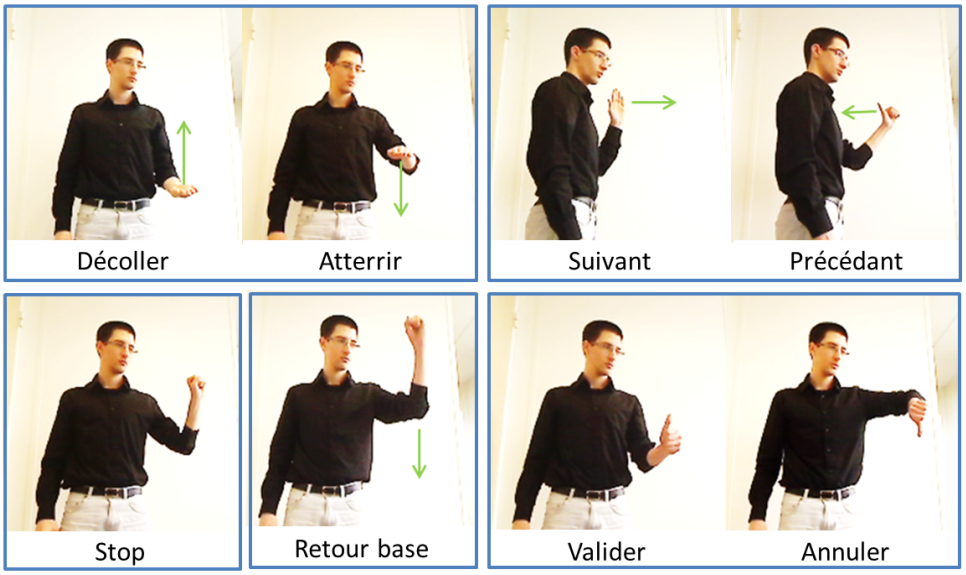


FIGURE 3.6 – Ensemble des gestes du dictionnaire élu.

de l'échelle de lecture proposée par Vatavu et al. (Vatavu & Wobbrock 2015), ce niveau de consensus est *haut* (à partir de 30%).

La Figure 3.7(c) permet de déterminer l'origine de chaque geste élu. Il apparaît que la totalité des gestes élus sont des gestes mixtes : à l'origine, ils ont été proposés à la fois par des participants militaires et des participants civils.

3.5.3 Étape 4 : Evaluation

Pour la dernière phase : lors de l'évaluation de la bonne compréhension du dictionnaire, 8 des 10 participants ont reconstitué l'ensemble des associations. Un participant a inversé les associations des fonctions *suivant* et *précédent* et un participant a mélangé les associations des fonctions *précédent*, *annuler* et *retour à la base*. Aussi, 94% des associations faites correspondent aux associations attendues.

3.6 Conclusion

Le nombre de gestes candidats est trois fois moins important que le nombre des gestes proposés. Ceci montre une certaine convergence ; et puisque plusieurs participants ont proposé le même geste, il est possible que le nombre de gestes pouvant exprimer une fonction (un concept), pour un groupe d'utilisateur donné, puisse être fini. Bien qu'avec une vingtaine de participants *seulement* ont ne puisse pas considérer que tous les gestes qui pourraient être proposés l'ont été, il semble raisonnable de penser que les plus importants ont bien été recueillis.

Une disparité apparaît cependant lorsque l'on regarde le nombre de gestes candidats différents selon les fonctions (dernière colonne en Figure 3.7(a)). Seulement 5 gestes

ont été proposés pour les fonctions *Atterrir* et *Décoller* alors que 11 gestes différents ont été proposés par la fonction *Base*. Une interprétation de ceci peut être, que certaines fonctions sont *simples* ou *classiques*. Les participants posséderaient déjà des symboles standardisés, ou l'imagerie mentale produite par la fonction serait assez uniforme. A l'inverse, les fonctions présentant une forte disparité, comme la commande *Base*, peuvent être considérées comme complexes (éventuelles floues) et ainsi provoquer différentes métaphores ou représentations. C'est particulièrement pour ces fonctions qui présentent une forte variabilité lors de l'élicitation qu'une phase d'élection semble primordiale.

Dans le dictionnaire élu, il est intéressant de constater que les fonctions complémentaires telles que décoller-atterrir, suivant-précédent et valider-annuler ont été associées à des gestes symétriques. Ceci laisse à penser que le dictionnaire sera d'autant plus logique et donc simple à comprendre et à apprendre.

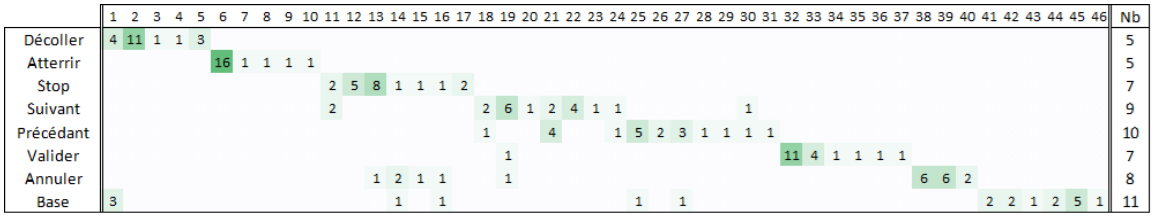
Par ailleurs, puisque l'ensemble des gestes élus ont à l'origine été proposés à la fois par des participants militaires et des participants civils, il peut sembler qu'ils ne reposent pas sur des concepts experts. Ces gestes semblent donc d'autant plus simples à apprendre et à comprendre, et pourraient être utilisables de manière plus large que pour le seul contexte des drones militaires. La forme des gestes de décollage et d'atterrissage décrivent un mouvement vertical. Ce vocabulaire gestuel pourrait donc convenir à toute commande d'un drone à décollage et atterrissage de cette forme (VTOL) dans un contexte civil et militaire.

Les scores de consensus moyen du catalogue (24%) et du dictionnaire (32%) sont cohérents avec ceux présentés par Choi et al. (Choi *et al.* 2012) et mettent en évidence une augmentation du niveau de consensus lors de la phase d'élection comparativement à la phase d'élicitation. Ce résultat confirme que permettre à des participants de considérer les gestes proposés par d'autres permet de constituer un dictionnaire plus consensuel et justifie donc bien l'usage d'une phase d'élection.

Concernant le score de consensus obtenu lors de la phase d'évaluation, un taux de 94% suggère que le dictionnaire constitué est suffisamment compréhensible pour pouvoir être appris et utilisé facilement.

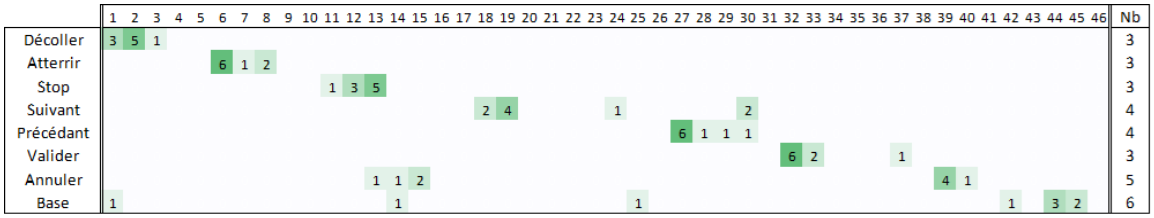
Enfin, les participants militaires ont indiqué avoir apprécié d'être consultés pour le choix des gestes tant cela leur semble un élément personnel. Cela leur a également permis de se projeter dans l'utilisation d'une commande au geste, et a confirmé la pertinence des questions posées dans nos recherches.

En conclusion, il semble important de réaliser une phase d'élection à la suite d'une phase d'élicitation pour choisir des gestes. La méthode que nous avons proposée permet donc la construction d'un dictionnaire adapté, et l'implication d'utilisateurs finaux dans ce processus est primordial. Par ailleurs, cette méthode peut être appliquée dans d'autres contextes et pour toute application qui requiert un dictionnaire de gestes sémaphoriques.



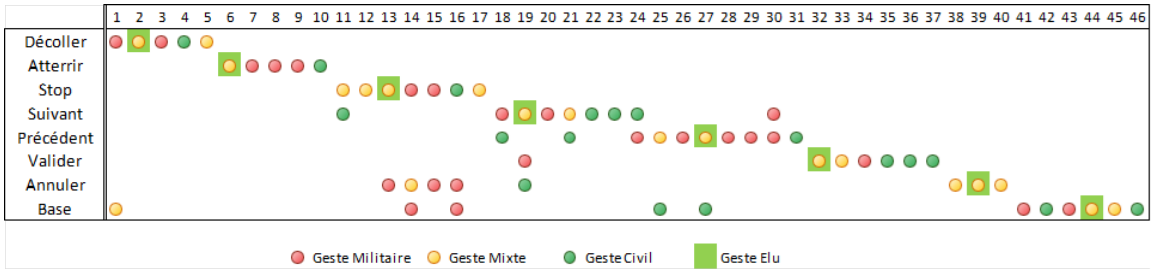
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	Nb
Décoller	4	11	1	1	3																																									5
Atterrir						16	1	1	1	1																																						5
Stop											2	5	8	1	1	1	2																														7
Suivant											2							2	6	1	2	4	1	1						1																9
Précédant																		1			4			1	5	2	3	1	1	1	1																10
Valider																			1													11	4	1	1	1	1									7
Annuler													1	2	1	1			1																			6	6	2							8
Base	3													1		1									1		1														2	2	1	2	5	1	11

(a) Nombre de participant ayant proposé chaque geste candidat lors de la phase d'élicitation.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	Nb
Décoller	3	5	1																																											3
Atterrir						6	1	2																																							3
Stop											1	3	5																																	3
Suivant																		2	4					1						2																4
Précédant																										6	1	1	1																			4
Valider																														6	2				1											3
Annuler													1	1	2																									4	1							5
Base	1													1											1																		1		3	2		6

(b) Nombre de voix allouées à chaque candidat lors de la phase d'élection.



(c) Origine de chaque geste candidat et gestes élus.

FIGURE 3.7 – Tables des résultats des étapes 2 et 3. Dans chacune de ces tables, une ligne correspond à une fonction et une colonne correspond à un geste candidat désigné par son numéro. Ces numéros sont ceux utilisés dans le catalogue présenté en Annexe A. La table (a) présente le nombre de participants différents ayant proposé chaque geste candidat lors de la phase d'élicitation. La table (b) présente le nombre de voix obtenues par chaque geste candidat lors de la phase d'élection. La table (c) présente l'origine de chaque geste candidat et ceux qui ont été élus.

CHAPITRE 4

Impact du vocabulaire proposé sur l'attention visuelle

Sommaire

4.1 Gestes et attention visuelle

Un modèle d'interaction et un dictionnaire de gestes ont donc été proposés. Ces deux éléments ont été constitués en considérant l'hypothèse que la commande gestuelle sémaphorique permet de libérer l'utilisateur ; c'est à dire, qu'elle permet, en comparaison d'interfaces classiques, d'être d'avantage en mesure de percevoir et d'agir sur l'environnement. A ce stade : avant l'ajout de tout facteur technique, il convient de contrôler la validité d'une telle hypothèse.

L'attention visuelle en tant qu'aptitude à percevoir et analyser pour détecter un stimulus, nous intéresse particulièrement. En effet, en plus d'être représentatif d'une action critique,

que ce soit dans le domaine militaire ou de la conduite de véhicule, la perception et l'analyse sont deux éléments qui peuvent être influencés par la commande gestuelle.

D'une part, la capacité à percevoir l'environnement peut bénéficier de la commande par geste puisque celle-ci affranchit de tout support visuel. En effet, contrairement aux interfaces classiques, les gestes n'imposent pas un mouvement de la tête ni une focalisation des yeux vers un objet pour y trouver les éléments activables, ce qui constitue donc, à priori, un avantage.

D'autre part, l'analyse des données perçues, c'est à dire le traitement mental de celles-ci, pourrait être influencé négativement par le fait d'avoir à mémoriser et se rappeler les gestes à utiliser.

Ainsi, l'effort mental à fournir pour utiliser un vocabulaire gestuel sémaphorique, pourrait éventuellement annuler l'avantage de ne pas avoir de support visuel. A notre connaissance, il n'existe aucune étude portant sur ce phénomène. Aussi, une étude en laboratoire a été conduite pour explorer l'impact de la commande gestuelle sur l'attention visuelle, en comparaison de l'impact d'une interface tactile usuelle.

4.2 Utilisation d'un protocole de type double tâche

La question de l'interaction entre deux tâches concurrentes est donc posée : la commande au geste est-elle en conflit avec l'attention visuelle ? Et si oui, dans quelle mesure ?

Au quotidien, il est commun de réaliser plusieurs tâches en même temps, mais il arrive parfois que la réalisation de l'une des tâches, voir des deux, échoue ou soit dégradée. Ceci se révèle problématique lorsque l'une des actions impacte la sécurité de personnes comme c'est le cas de la conduite de véhicule, le pilotage d'avion ou la commande de drones.

Ce phénomène est connu sous le nom de *Dual-Task Interference*. Il révèle une certaine forme de limitation du corps humain et permet d'émettre quelques hypothèses quant au fonctionnement de celui-ci (Sala *et al.* 1995).

Différentes théories existent pour expliquer ce phénomène de baisse de performance lorsque deux tâches sont réalisées de manière concurrente (Pashler 1994) : le partage de capacité (*capacity sharing*) et le goulot d'étranglement (*bottleneck*) :

— Le partage de capacité fait l'hypothèse que la réalisation d'une tâche requiert une certaine quantité de ressources (physiques et ou cognitives) et que les ressources disponibles sont globalement limitées. Aussi, lorsque deux tâches sont réalisées en même temps, elles se partagent les ressources. Et lorsque les ressources deviennent insuffisantes pour l'une ou pour l'autre, les performances sont alors dégradées. Pour certains auteurs, il existe un seul type de ressources (Kahneman 1973). Pour d'autres, il en existe différents types (Navon & Gopher 1979; Wickens 1980) ce qui expliquerait que certaines actions ne rentrent pas en conflit.

— Le goulot d'étranglement fait quant-à lui l'hypothèse que ce ne sont pas des ressources qui sont limitées mais l'utilisation de processus cognitifs. Si deux tâches requièrent l'usage d'un même et unique processus cognitif au même moment, elles sont réalisées de manière alternée ou alors, l'une est traitée en priorité et la seconde

est mise en attente ou abandonnée.

Ainsi, deux tâches peuvent ne pas être correctement réalisées, de manière simultanée, si elles consomment une même ressource (physique et cognitive) en trop grande quantité ou si elles utilisent simultanément une même infrastructure (processus cognitif).

L'interaction entre deux tâches concurrentes peut être évaluée avec un protocole dit de *double tâche*. Ce type de protocole, comporte deux grandes phases impliquant une population de participants : Dans un premier temps, les participants réalisent les tâches seules. On mesure alors les performances de chaque participant pour cette situation témoin. Puis, dans un second temps, les performances de chaque participant sont mesurées lorsque les tâches sont réalisées simultanément. Les mesures effectuées lors des deux phases sont comparées pour chaque participant (correction de la variance intra-sujet) puis, sont analysées de manière globale pour déterminer si un effet significatif est présent ou non au sein de la population.

Lors de la seconde phase, il est important que les participants cherchent à maintenir les performances atteintes en situation témoin pour s'assurer que d'éventuels résultats soient bien dus à une limitation de capacité et non à une baisse de volonté ou d'implication.

Ce type de protocole a par exemple été employé pour évaluer la dangerosité de différentes activités réalisées au volant d'un véhicule ; et notamment l'usage des téléphones portables (Strayer & Johnston 2001). De manière analogue, nous souhaitons évaluer l'interaction entre deux tâches concurrentes : la perception consciente de stimuli visuels et la commande gestuelle sémaphorique. Aussi, pour cette étude, nous avons fait le choix d'un protocole de type double tâche.

4.3 Méthode

4.3.1 Plan du protocole

Un protocole de type *double tâche* a donc été mis en oeuvre pour évaluer et comparer l'impact de la commande gestuelle et de la commande tactile.

Différents participants ont pris part à deux sessions identiquement organisées ; une par modalité de commande. Pour chaque participant, le choix de l'ordre des sessions est rendu aléatoire afin de compenser un éventuel effet d'ordre.

Chaque session comportait quatre grandes phases : (1) une phase de présentation de la modalité de commande et un entraînement à son usage, (2) une phase de simple tâche : détection de stimuli visuels, (3) une phase de double tâche : détection de stimuli visuels et commande d'un drone virtuel, et (4) une phase de questionnaire post-session.

Au cours de ces quatre phases, différentes mesures ont été réalisées. Certaines étaient *objectives* : des durées d'entraînement et des temps de réactions ; et d'autres étaient *subjectives* : des résultats de questionnaires.

L'organisation du protocole mis en oeuvre est résumé en Figure 4.1. Les participants, le matériel et les différentes phases sont détaillées dans les sous-sections qui suivent.

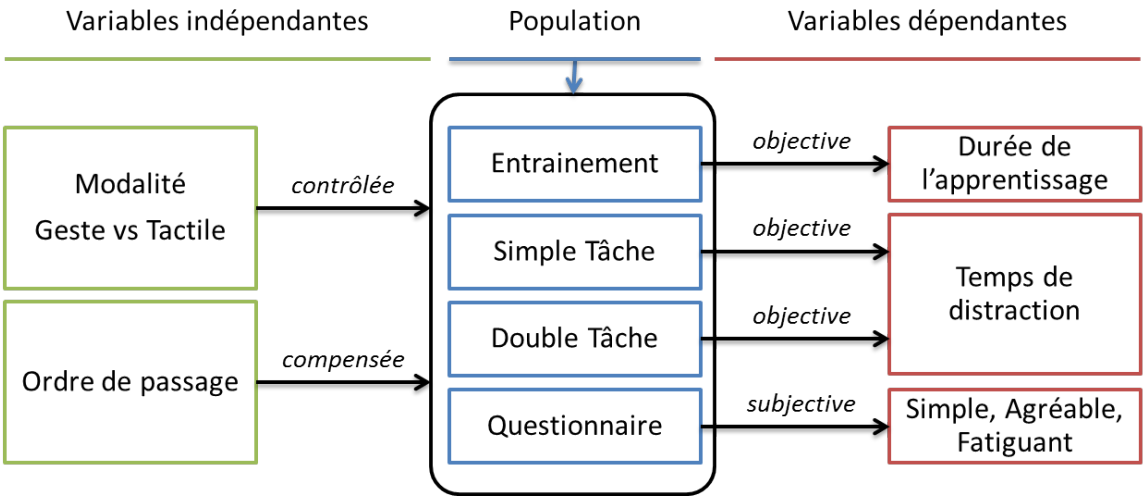


FIGURE 4.1 – Synthèse du protocole mis en oeuvre pour évaluer l'impact de la commande gestuelle sur l'attention visuelle.

4.3.2 Participants

Un total de 20 participants a pris part à cette étude. Tous étaient étudiants, chercheurs ou personnel administratif à Mines-Paristech. L'âge moyen de cette population était de 29 ans (S.D : 9.6).

4.3.3 Matériel

Les participants étaient placés debout face à un écran rétro-projeté. Cet écran possédait des dimensions de 3,1m par 1,74m, une résolution de 1920 par 1080 et une fréquence d'affichage de 60Hz.

Deux personnes accompagnaient les participants : un expérimentateur et un magicien d'Oz (WoZ) Le magicien d'Oz avait pour rôle de simuler la reconnaissance des gestes faits par le participant lors des phases de commande gestuelle. Le rôle de ce magicien d'Oz n'était pas connu des participants.

Les participants étaient équipés d'un microphone. Celui-ci connecté à une application simple permettait de générer des messages informatiques précisément datés lorsque le participant prononçait le mot "Vu".

Une interface de commande tactile était également mise en oeuvre. Elle était composée d'une matrice de 8 boutons logiciels (touches tactiles) représentées par des icônes simples. Chacune correspondait à une commande du modèle d'interaction précédemment proposé (Décoller, Atterrir, Suivant, Précédent, Base, Stop, Valider, Annuler). Cette matrice était affichée sur l'écran d'une tablette : un *Samsung Galaxy Tab* possédant un écran 7 pouces et pesant 380g. Cette interface est présentée en Figure 4.2(c).

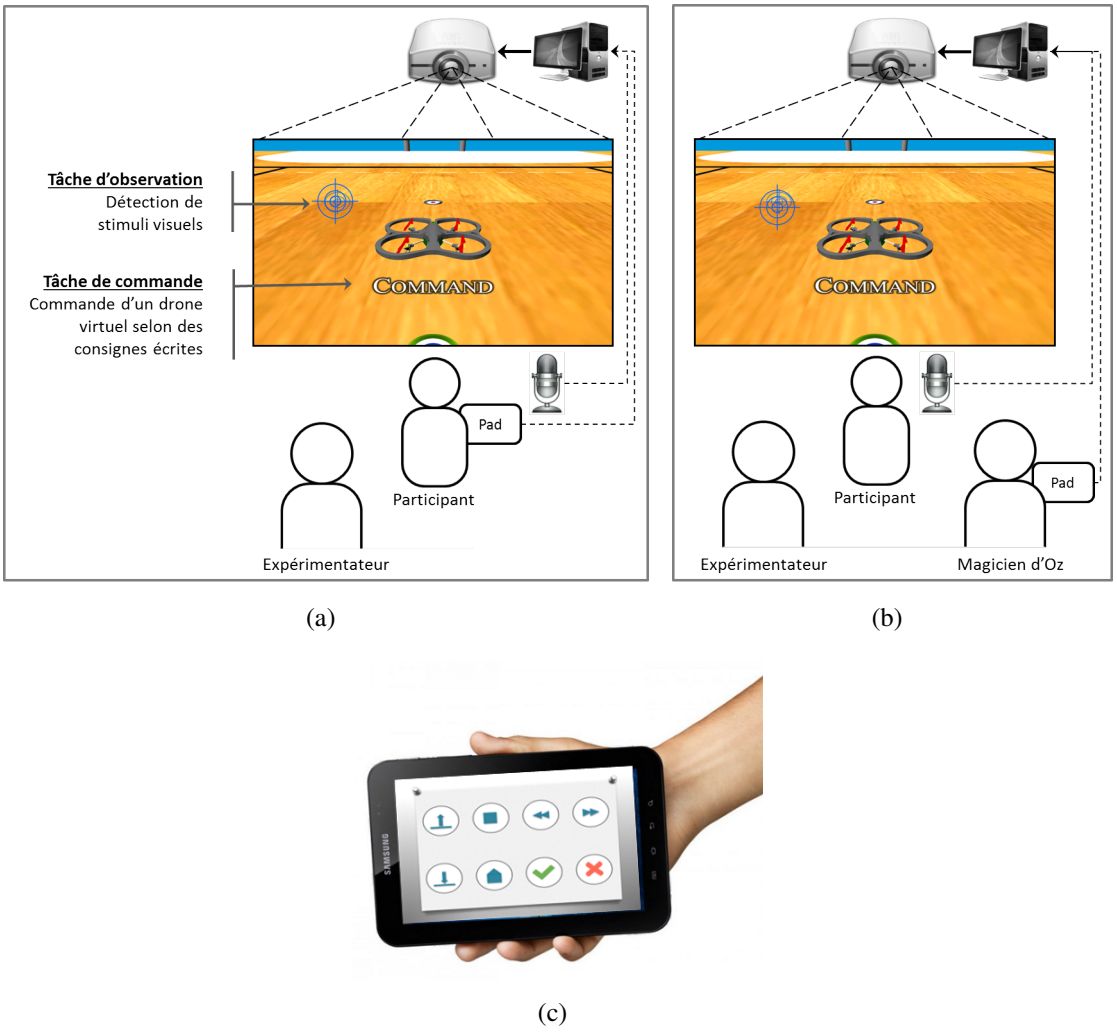
Le microphone, la tablette et le vidéo-projecteur étaient tous connectés à un ordinateur sur lequel une application 3D temps réel affichait un diaporama pour les phases d'explications, un drone virtuel et des consignes pour la tâche de commande, et des stimuli

visuels pour la tâche d'observation. Cette application avait également pour rôle d'enregistrer automatiquement les différents évènements temporels dans un fichier, pour permettre leur analyse. L'ensemble de cette organisation est résumée en Figure 4.2.

4.3.4 Description du protocole

4.3.4.1 Phase 1 : Entraînement à la modalité de commande

Chaque participant prend donc part à deux sessions ; une pour la modalité tactile et une pour la modalité gestuelle. Au début de chacune, le participant est accueilli et le contexte



(a) (b)

(c)

FIGURE 4.2 – Matériel et organisation pour les sessions de double tâche. (a) Pour les sessions imposant la modalité tactile, le participant prenait en main une tablette. (b) Pour les sessions imposant la modalité gestuelle, un magicien d'Oz utilisait la tablette de manière discrète pour simuler une reconnaissance automatique. (c) Dans ces deux situations, la tablette utilisée présentait une simple interface composée d'une matrice de touches logicielles. .

de la commande de drone est rapidement exposé.

Puis, la tâche de commander un drone via le modèle d'interaction proposé est présentée : d'une part l'autonomie du drone et les différentes actions qu'ils est capable de réaliser automatiquement, et d'autre part, l'agent conversationnel et les différentes phases de dialogue (validation de commande).

Ensuite, la modalité utilisée pour la session en cours est choisie par l'expérimentateur et présentée au participant. Le choix de la modalité est fait aléatoirement afin de compenser un éventuel effet de l'ordre des sessions sur les mesures.

Si la modalité imposée est la commande tactile, le participant est alors équipé de la tablette. Il lui est demandé de la maintenir avec sa main non dominante et d'activer les touches logicielles avec sa main dominante. Si en revanche, la modalité imposée est la commande gestuelle, il est demandé au participant de réaliser les gestes en direction de l'écran. De son côté, c'est le magicien d'Oz qui utilise la tablette. Sans être vu du participant, il observe les gestes réalisés et active les touches correspondantes sur l'interface.

Une fois le participant (ou le magicien d'Oz) équipé, l'activation des différentes commandes est démontrée une par une. Dans le cas de l'interface tactile, l'icône du bouton correspondant est décrite et le participant est invité a l'activer. Dans le cas de l'interface gestuelle, le geste est présenté par un texte et une image sur l'écran (face au participant). Le participant est alors invité à réaliser le geste correctement au moins une fois. Cette phase est contrôlée par l'expérimentateur qui juge si le geste est correctement fait et apporte des précisions en cas de besoin.

Lorsque toutes les commandes on été présentées une première fois, le participant démarre un cycle d'entraînement : sur l'écran, face au participant, les noms des huit commandes sont affichés un par un. Pour chacun, il est demandé au participant d'activer la commande Si le participant ne se souvient pas de l'action à réaliser, il peut demander à l'examinateur de passer à la commande suivante. Que le participant réalise la bonne action, se trompe ou passe, le système enregistre le résultat et passe à la commande suivante.

Lorsque le cycle complet des huit commandes se termine, l'ordinateur comptabilise le nombre d'erreur. Si au cours du cycle, le participant s'est trompé ou a passé au moins une fois, un résumé de l'ensemble des gestes ou des boutons est présenté et le participant est invité à se les remémorer. Suite à quoi, un nouveau cycle d'entraînement est débuté.

Lorsqu'un cycle est réalisé sans aucune erreur, le participant est félicité et la phase d'entraînement prend fin. Le nombre de cycles effectués par le participant est un indicateur de la difficulté d'apprentissage. Il est par conséquent enregistré.

4.3.4.2 Phase 2 : Tâche d'observation seule

Dans ce protocole, la tâche d'observation permet d'évaluer la capacité du participant à détecter un stimuli visuel dans son environnement et à adopter le comportement adéquat face à celui-ci. C'est une tâche en soi très générique qui s'applique parfaitement à de nombreuses situations de la vie courante comme la détection et le traitement des signalisations sur la route. Cette tâche constitue un acte élémentaire fondamental pour le combattant sur le champ de bataille.

Pour simuler cette tâche, des cibles visuelles apparaissent sur l'écran face au participant, une à la fois et à une position aléatoire. Un exemple est présenté en Figure 4.2. Cette cible reste visible pendant une durée de 3 secondes puis disparaît. Après une pause de durée aléatoire allant de 3 à 5 secondes, une nouvelle cible apparaît à un nouvel emplacement. Le positionnement et la cadence aléatoire des cibles force le participant à rester vigilant sur l'ensemble de l'écran et à ne pas anticiper les apparitions.

Au début, la forme et la couleur du stimulus visuel est présenté au participant. Il sait donc ce qu'il attend. Puis, il lui est demandé de rester vigilant à l'apparition de ces stimuli pendant une durée de 5 minutes. Lorsque le participant détecte l'apparition d'une cible, il doit le signaler oralement en prononçant distinctement le mot "Vu" et ce, le plus rapidement possible. Pour indiquer au participant que sa réaction a bien été prise en compte, la cible est automatiquement effacée.

Lors ce cette tâche, le système enregistre l'ensemble des informations temporelles des évènements d'apparition de cible et de leurs signalements.

4.3.4.3 Phase 3 : Double tâche

Après quelques instants de repos, il est demandé au participant de réaliser simultanément deux tâches pendant une durée de 5 minutes. La première est la tâche d'observation que le participant vient de réaliser. La seconde tâche est la commande d'un drone virtuel 3D selon des consignes données par le système.

Pour cette tâche de commande, un drone virtuel est affiché au centre de l'écran et réagit lorsque des commandes sont passées. Le système génère automatiquement une consigne (une commande à réaliser) lorsque le drone est dans un état d'attente. Cette consigne, compatible avec l'état actuel du drone, est affichée sur l'écran, en dessous du drone. Un exemple est présenté en Figure 4.2.

Lorsque le nom d'une commande apparaît, il est demandé au participant de compléter le dialogue avec l'agent conversationnel pour activer cette commande. Pour ce faire donc, le participant réalise le geste ou presse le bouton de la commande ; attend la demande de confirmation vocale de l'agent ; puis valide à nouveau en gestuel ou en tactile. Lorsqu'une commande est activée, le drone virtuel simule l'exécution automatique de la fonction et la consigne disparaît. A la fin de de l'action en cours, une nouvelle consigne est générée et présentée au participant.

A nouveau, les informations temporelles de chaque évènement sont enregistrés : apparition d'une cible, signalement d'une cible, apparition d'une consigne et validation d'une consigne.

4.3.4.4 Phase 4 : Questionnaire subjectif

A la fin de chaque session, il est demandé au participant de remplir un court questionnaire. Sur une échelle de Likert à 5 valeurs allant de (1) *pas du tout d'accord* à (5) *tout a fait d'accord*, le participant doit indiquer son accord ou son désaccord avec les 3 affirmations suivantes :

— Concernant la modalité que vous venez d'utiliser pour commander le drone virtuel, vous l'avez trouvé **simple à utiliser**.
— Concernant la modalité que vous venez d'utiliser pour commander le drone virtuel, vous l'avez trouvé **amusant**.
— Concernant la modalité que vous venez d'utiliser pour commander le drone virtuel, vous l'avez trouvé **fatiguant**.

A la fin de la seconde session, lorsque les deux modalités ont donc été utilisées par le participant, une nouvelle affirmation est présentée :

— Concernant les deux modalités que vous avez utilisées pour commander le drone virtuel, vous avez **préféré le geste**.

4.3.5 Temps de réaction et de distraction

Une fois les deux sessions passées par tous les participants, les données enregistrées par l'ordinateur lors des phases de simple et de double tâche sont analysées. Différents indicateurs objectifs sont alors calculés.

Tout d'abord, pour chaque participant, différents temps de réaction sont calculés. Le temps de réaction est le temps écoulé entre l'apparition d'une cible et son signalement par le participant.

Ce type d'indicateur est représentatif de la capacité du participant à détecter et traiter des évènement visuels dans son environnement. Un temps de réaction court indique que le participant est visuellement attentif alors qu'un temps de réaction long indique que le participant est distrait soit physiquement, soit cognitivement.

Pour chaque participant p et pour chaque modalité m, différents temps de réactions sont mesurés :

— TR_1^{pm} : le temps de réaction moyen en situation de simple tâche (situation témoin).
— TR_2^{pm} : le temps de réaction moyen en situation de double tâche.

Aussi, sur l'ensemble de la population, un temps de réaction moyen peut être calculé par modalité : TR_1^m

On cherche à déterminer si il existe une différence significative entre les situations de commande gestuelle et de commande tactile. Cependant, les temps de réaction ne peuvent pas être directement utilisés dans la mesure où ils dépendantes de facteurs différentiels :

— D'une part, le temps de réaction dépend de chaque participant et de l'état physique et émotionnel dans lequel il se trouve au moment d'une session.
— D'autre part, les valeurs mesurées sont impactées par les temps de latence du système qui les enregistre. La latence du microphone et du logiciel de détection associé sont certainement les éléments les plus impactants.

Aussi, une comparaison directe des temps de réaction des situations de double tâche pourraient être influencés par ces deux biais.

Pour annuler ces deux effets on soustrait aux temps de réaction moyens en double tâche, les temps de réaction moyens en simple tâche de la même session. En effet, nous considérons que l'état émotionnel et de fatigue d'un participant est constant au cours d'une même session.

Ainsi, de nouveaux indicateurs non biaisés sont calculés : pour chaque participant p et pour chaque modalité m, on calcule la moyenne des différences de temps de réaction entre les situations de double tâche et de simple tâche : TD^{pm}. Une moyenne sur la population de ces temps de distraction moyens est également calculée : TD^{m}.

Finalement, les indicateurs de temps de distraction permettent d'étudier si la commande au geste et la commande tactile perturbent différemment la détection de cibles visuelles. Un temps de distraction faible indique un impact faible de la commande sur l'attention visuelle, et un temps de distraction plus important indique un impact négatif plus fort.

4.4 Résultat

4.4.1 Impact sur l'attention visuelle

Les différents temps de réaction (TR) et de distraction (TD) mesurés sont présentés en Table 4.1 et un histogramme des temps de réaction moyens est également proposé en Figure 4.3.

	TR_1^m	TR_2^m	TD^m
Tactile	1028 mS	1304 mS	275 mS
Gestuel	1032 mS	1122 mS	90 mS

TABLE 4.1 – Temps de réaction moyens (TR) et des temps de distraction moyens (TD).

Tout d'abord, on constate que les temps de réaction moyens sont relativement identiques quelque soit la modalité lors des phases en simple tâche (Anova F=0,12 ; p=91%). Ce résultat est logique puisque lors de cette phase du protocole, il n'y a pas de commande et donc la modalité ne rentre pas en jeu. Cependant, ce constat confirme qu'il ne semble pas y avoir d'effet non voulu (externe) qui influencerait la tâche d'observation.

Ensuite, on constate une différence significative entre les situations de simple tâche et de double tâche quelque soit la modalité (Anova F=14.37 ; p<0.1%). Ce résultat montre que de manière générale, le fait même d'avoir à activer une commande, impacte l'attention visuelle.

Enfin, l'analyse des temps de distraction (TD) montre que la commande au geste impacte significativement moins la capacité à percevoir et signaler l'apparition d'un stimuli visuel (Anova F=8,27 ; p<0,1%).

Ces différents résultats montrent que la charge cognitive imposée à l'utilisateur par l'apprentissage et le rappel d'un vocabulaire gestuel a un effet moins important que la charge cognitive et la focalisation du regard imposées par une interface tactile.

Or, la tablette utilisée lors de cette étude était relativement petite et l'interface n'était composée que des seuls boutons nécessaires (simples à trouver et à activer). Les interfaces standards sont en réalité beaucoup plus encombrantes et affichent un nombre très important de contrôles et d'indicateurs. Le résultat obtenu, en faveur du geste, semble donc d'autant plus encourageant.

4.4.2 Facilité d'apprentissage

Lors de la phase d'apprentissage, le nombre de cycles d'entraînement a donc été comptabilisé pour chaque participant et selon la modalité. Pour la modalité tactile, une moyenne de 1,25 cycles a été nécessaire à la bonne maîtrise de l'ensemble des commandes, et 1,70 cycle en gestuel. Pour chacune de ces deux modalités et sur l'ensemble de la population, 1 cycle a été le minimum et 3 cycles le maximum.

Il apparaît donc que la modalité gestuelle requiert plus d'apprentissage. Cependant cette différence est relativement faible : de l'ordre de quelques minutes seulement. Ce résultat montre que le dictionnaire de gestes constitué est effectivement simple à apprendre et à se rappeler.

4.4.3 Résultats du questionnaire

Les résultats du questionnaire de fin de session sont présentés en Figure 4.4.

Les participants on en moyenne trouvé l'interaction gestuelle moins fatigante. En effet, le port de la tablette à la main en continu engendre une fatigue qui a été ressentie par les participants. En revanche, les gestes sont réalisés de manière ponctuelle et permettent naturellement le retour du bras dans une position de repos lorsqu'il ne sont pas utilisés.

L'interaction par geste a également parut globalement plus agréable. Les participants ont indiqué se sentir plus libres, et le fait de ne pas avoir à bouger la tête en permanence pour regarder l'interface tactile a été décrit comme un bénéfice significatif par presque tous les participants.

Les participants ont toute fois indiqué que la tablette tactile était relativement plus simple à utiliser. En effet, alors que les commandes sur l'interface tactile étaient relativement intuitives, les gestes devaient être appris. Ce n'est pas nécessairement une difficulté à apprendre les gestes qui a été exprimée par les participants mais qu'une phase de formation avec un support (un manuel) soit incontournable. Aussi, au premier abord, le participant se sent perdu. Au contraire, pour l'interface tactile, la présence et la forme des boutons semblaient suffisantes pour une première prise en main. Toute fois, la majorité des participants a relativisé ce phénomène : les gestes restaient relativement simples et peu nombreux.

Finalement, après avoir expérimenté les deux modalités, les participants ont globalement préféré l'interaction par geste. La moindre fatigue et surtout le fait de ne pas avoir à tourner la tête ont été des éléments déclarés comme étant déterminants.

4.5 Conclusion

En conclusion, un groupe de 20 participants a expérimenté le modèle d'interaction proposé et le dictionnaire de gestes construit. Ils se sont avérés très simples à comprendre, à apprendre et à utiliser.

Plus spécifiquement, le dictionnaire de gestes conçu pour être consensuel et non ambigu a été maîtrisé très rapidement. Aussi, la méthodologie présentée dans ces travaux impliquant

différentes population en plusieurs phases semble particulièrement concluante. De plus, l'usage des gestes semble pouvoir gagner en fluidité avec un usage prolongé.

Cependant, une phase d'apprentissage avec un support (externe) reste incontournable ce qui n'est pas nécessairement le cas avec une interface tactile simple et explicite. Cette nécessité d'un apprentissage peut décourager certains utilisateurs et pourrait ne pas convenir pour des interfaces grand public ; éventuellement pour des particuliers intéressés mais assurément pas pour du libre service.

Il apparaît également que si un utilisateur ne parvient pas à se rappeler d'un geste, il se retrouve alors bloqué. Il est alors forcé de se reporter à une documentation externe ou d'utiliser une autre interface complémentaire. Ce constat montre à nouveau toute l'importance de la phase d'apprentissage et confirme qu'une telle modalité est plutôt adaptée pour des systèmes experts.

Deux solutions semblent envisageables pour répondre à cette problématique de blocage :

— La première solution qui a déjà été évoquée, est que la modalité gestuelle peut être une redondance experte à une interface standard qui reste rangée mais disponible en cas de besoin. Une interface gestuelle peut donc être un ajout éventuellement facultatif pour des interfaces existantes.

— La seconde solution, est d'intégrer à l'agent conversationnel une documentation audio activable avec un nouveau geste spécifique. Ce geste serait alors le seul à impérativement retenir et, éventuellement, plusieurs gestes pourraient êtres permis pour cette fonction critique. Cette seconde option ouvre différentes possibilités intéressantes comme la création d'un tutoriel interactif, mais pause de nombreuses questions sur comment décrire un geste oralement (bien souvent les gestes sont transmis par démonstration-imitation), de manière progressive et contextuelle (synchroniser l'explication avec la réalisation de l'utilisateur novice).

Cette étude a également montré que ce manque avéré de support qui impose un apprentissage des gestes, en constitue également l'atout majeur. En effet, les yeux et la main sont libérés : le regard n'a pas besoin d'être porté constamment sur un écran et aucun matériel ne nécessite plus d'être manipulé.

On peut cependant relativiser ce second point, qui est ici suggéré par le constat de gestes moins fatigants que le port d'une tablette ; différentes solutions techniques sont évidemment envisageables pour positionner une interface tactile de manière plus adéquate. Toutefois, alors que l'expressivité des gestes semble infinie (à minima très grande), les possibilités d'une interface tactile (boutons logiciels ou mécaniques) restent relativement contraintes par ses dimensions et son poids.

Cette étude a confirmé que l'usage des gestes associés à une modalité de retour audio favorise l'attention visuelle. Des stimuli dans l'environnement proche sont plus rapidement perçus et traités. Cette plus grande liberté du regard, et plus globalement de la tête, a par ailleurs été consciemment perçue par l'ensemble des participants. Et ceci a conduit à une relative, mais globale, préférence pour la modalité gestuelle. Aussi, une modalité gestuelle sémaphorique semble particulièrement adaptée pour toute situation où la perception de l'environnement est requise ; Dans le contexte de la commande des drones de contact ou cela est critique, l'usage des gestes est donc assurément pertinent.

L'ensemble des constats réalisés lors de cette étude, valident donc globalement le modèle d'interaction proposé, l'usage des gestes, le dictionnaire gestuel et indirectement la méthodologie de construction de ce dernier. Ceci conclu de manière encourageante cette première itération de développement centré utilisateur ou l'*interaction* était au coeur de la réflexion.

Dans les trois travaux qui viennent d'être présentés : (a) la définition du modèle d'interaction, (b) la construction d'un dictionnaire de geste et (c) l'évaluation de l'impact de la modalité gestuelle sur l'attention visuelle, aucune considération technique n'a encore été prise en compte. Il a jusque là été supposé que la technologie actuelle permet effectivement de mettre en oeuvre tout ce qui a été proposé (interaction et gestes).

L'usage d'un magicien d'Oz a permis dans un premier temps, de simuler un système de reconnaissance *parfait*, affranchissant de tout artefact technique (latence, erreurs de détection ou de reconnaissance). Différents participants ont d'ailleurs indiqué être très surpris de la qualité du système qu'ils croyaient utiliser ; et que cela importait beaucoup dans leur ressenti global. Un système gestuel présentant des défauts n'aurait peut-être rien changé à la libération du regard et de la main, mais aurait été globalement moins acceptable. Surtout aujourd'hui, alors que les interfaces tactiles de dernière génération font preuve d'une très grande fluidité.

Il importe donc de développer une interface réalisant effectivement la reconnaissance de geste de façon automatique et d'évaluer son utilisabilité concrète. Aussi ces deux éléments ont fait l'objet de travaux qui vont être présentés dans les chapitres qui suivent.

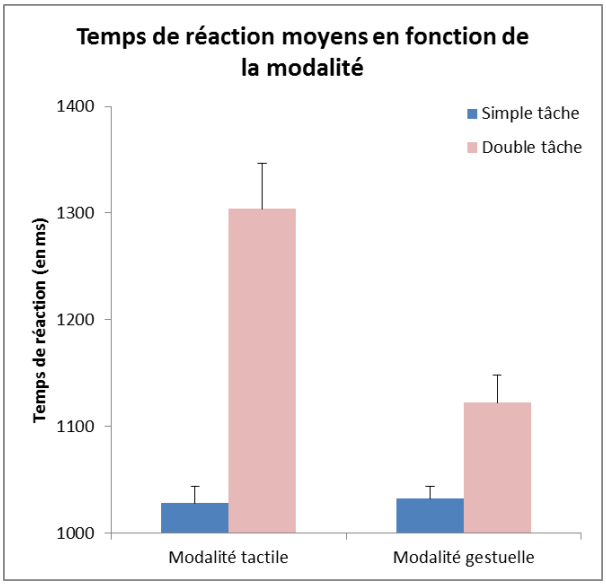


FIGURE 4.3 – Temps de réaction moyens (TR) des participants pour la détection et le signalement de stimuli visuels.

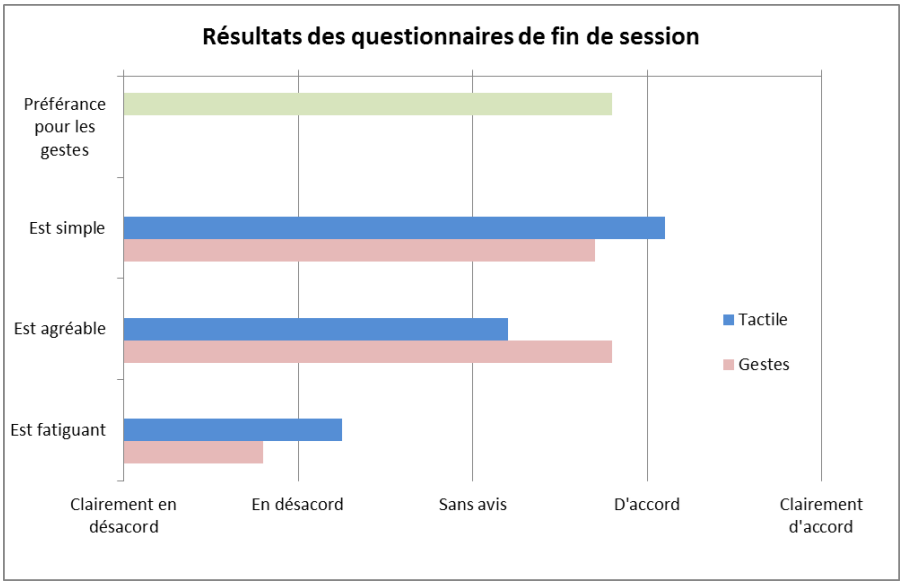


FIGURE 4.4 – Résumé des résultats du questionnaire subjectif de fin de session.

CHAPITRE 5

Reconnaissance gestuelle et plate-forme interactive

Sommaire

Dans ce chapitre, nous nous intéressons à la mise en oeuvre technique des deux éléments précédemment proposés : d'une par, le dictionnaire de gestes et d'autre part, le modèle d'interaction.

Ainsi, dans un premier temps, une brique de reconnaissance gestuelle complète est proposée et implémentée. : en marge des systèmes classiques qui reposent généralement sur l'apprentissage automatique d'exemples, elle repose ici sur la description formelle des gestes par un opérateur humain.

Puis, dans un second temps, la brique de reconnaissance gestuelle est intégrée au sein d'une architecture qui permet l'activation des commandes d'un drone et la génération des différents feedbacks sonores associés, conformément au modèle d'interaction proposé au Chapitre 2.

5.1 Brique de reconnaissance gestuelle utilisant les expressions régulières

5.1.1 Fonctions et architecture standards

Une *brique de reconnaissance gestuelle* est une appellation relativement générique qui désigne un système technique réalisant automatiquement une analyse des mouvements et de la configuration d'un corps humain sur une plage temporelle pour en extraire du sens.

Une brique gestuelle remplit et repose sur différentes fonctions élémentaires : (1) la captation et la représentation des mouvements, (2) la sélection temporelle, (3) la détection et la classification, et (4) dans certains cas, le suivi de la progression.

Les deux premières fonctions permettent de préparer les données alors que les deux fonctions suivantes constituent à proprement parler, l'analyse de ces données.

5.1.1.1 Captation et représentation

Le premier composant de tout système gestuel est nécessairement une brique de captation et de représentation des mouvements.

Il s'agit donc de mettre en oeuvre un ensemble de capteurs, quelle que soit leur technologie, qui permettent d'acquérir les informations corporelles déterminantes pour les gestes considérés. Un compromis doit être trouvé entre le nombre de capteurs et leur utilité. En effet, les capteurs pouvant être chers et encombrants, il est important de limiter leur nombre.

Lorsqu'il s'agit de geste, les informations à capter sont généralement spatiales : on s'intéresse à des positions relatives ou absolues, à des orientations, à des déplacements (vitesse ou accélération), ou à des configurations ; c'est à dire des angles entre des segments articulés.

Les capteurs fournissent des données brutes qu'il faut, au besoin, transformer dans une représentation intéressante. Cela suppose d'une part, de détruire les informations

redondantes ou non pertinentes (ce qui révèle en soi, un nombre de capteurs trop important), et d'autre part, d'opérer des changement de repère pour exprimer les données dans un nouvel espace d'avantage représentatif du corps et des gestes à reconnaître (donc sémantiquement plus élevé). Un exemple de changement de représentation, est la conversion d'un flux vidéo en un squelette : par exemple du corps ou de la main.

Ce qui est implicite, c'est que les données deviennent contextualisées : ce qui avant pouvait n'être qu'une accélération, peut devenir une accélération spécifique à une partie du corps, par exemple de la main.

Le changement de représentation peut également consister en une discrétisation de l'espace. Des données jusqu'alors continues sont labélisées comme appartenant à des catégories spécifiques. Par exemple, à l'échelle d'un doigt, la courbure qui est une information continue peut être transformée en deux catégories : doigt plié ou doigt tendu. A nouveau le niveau sémantique de la représentation augmente.

Enfin, il est raisonnable d'associer à ce premier composant de tout système gestuel, la possibilité d'un enregistrement des données captées et transformées. Cela permet, au besoin, de simuler les capteurs en *re-jouant* les données .

5.1.1.2 Sélection temporelle

Les capteurs et la représentation des données fournissent un flux temporel continu. Cependant, il est important pour réaliser une analyse, de découper ce flux temporel continu en blocs de données de taille raisonnable. En effet, il serait trop coûteux de conserver et d'analyser l'ensemble des données, d'autant plus que la réalisation d'un geste est borné dans le temps : un geste possède un début et une fin. Il s'agit donc d'extraire des segments temporels intéressants à analyser.

Pour ce faire, deux grandes stratégies existent : la segmentation qui est une stratégie *intelligente* s'appuyant sur une connaissance à priori des données, et le fenêtrage glissant qui est une stratégie *aveugle* ne nécessitant aucune connaissance experte.

Segmentation

Également appelée *spotting*, la segmentation consiste à localiser dans le flot de données, le début et la fin de ce qui pourrait être un geste pour l'extraire dans son intégralité. Il s'agit, d'une certaine manière, d'un filtre qui ne va conserver pour analyse, que des séquences intéressantes : des gestes potentiels.

Il est possible de procéder de manière manuelle : le découpage est réalisé par un opérateur humain. Cette méthode est évidemment non applicable pour une application temps-réel, mais permet de réaliser une *vérité-terrain* lorsque l'on souhaite comparer les performances de différentes méthodes d'analyse.

La segmentation peut également être réalisée de manière automatique. Un ensemble de règles simples peut parfois suffire : par exemple, pour les gestes sur une surface tactile l'apparition et la disparition d'un point de contact peuvent être considérés comme des bornes naturelles. Un principe similaire peut être appliqué pour des gestes réalisés face à une caméra : l'apparition et la disparition de la main dans le flux vidéo sont considérés

comme les bornes implicites d'un geste.

L'utilisation d'un méta-modèle, c'est à dire d'un modèle générique de geste, peut également être utilisé (Zhu *et al.* 2002; Davis & Shah 1994). Les modèles génériques proposés semblent relativement proches des modèles anatomiques de Kendon et de Kita et al. présentés au Chapitre 3 (préparation, stroke, rétraction), mais cependant, restreints à des gestes statiques. Un exemple de modèle générique est alors : (1) déplacement de la main, (2) immobilisation de la main et (3) déplacement de la main ; le tout étant borné temporellement.

Enfin, une troisième méthode pour une segmentation automatique est la détection de points caractéristiques. Deux exemples sont la détection des minima locaux d'une fonction de l'énergie des déplacements du corps (Kahol *et al.* 2004), ou des maxima locaux d'une fonction d'erreur de l'approximation linéaire des orientations de l'avant bras (Junker *et al.* 2008).

Il est à noter que les méthodes automatiques ne sont généralement pas parfaites. En effet, elles conservent parfois des séquences qui ne correspondent pas à un geste, ou pas à l'intégralité d'un geste. Cependant, cela est peu important dans la mesure ou les séquences retenues sont ensuite analysées (détection) pour déterminer si effectivement il s'agit bien d'un geste. En revanche, il est fondamental qu'une segmentation automatique ne filtre pas par erreur un geste réel car il serait alors définitivement perdu.

Fenêtre glissante

Une seconde stratégie pour sélectionner des données à analyser, et certainement la plus commune bien que peu souvent explicitée, est l'utilisation d'une fenêtre temporelle glissante (*sliding window*). Il s'agit alors simplement de conserver l'ensemble des données consécutives sur un intervalle de temps pour les proposer ensuite aux fonctions d'analyse. Dans ce type d'approche, deux paramètres sont à définir : la taille de la fenêtre et son déplacement.

La taille de la fenêtre détermine assez naturellement la durée des données à analyser. Son choix doit être un compromis entre une fenêtre suffisamment grande pour capturer le geste le plus lent, mais pas trop grande car cela serait par la suite, trop coûteux à analyser.

Le déplacement de la fenêtre est la fréquence à laquelle les données sont proposées pour l'analyse. Cela détermine également le recouvrement de deux fenêtres consécutives : c'est à dire les données communes et donc analysées plusieurs fois. Pour ce paramètre, à nouveau, un compromis doit être trouvé. En effet, un déplacement trop lent conduit à un recouvrement important, et donc à analyser de nombreuses fois les mêmes données ; ce qui est inutilement coûteux et un déplacement trop rapide conduit à ne pas analyser toutes les données et donc éventuellement, à rater un geste.

5.1.1.3 Détection et classification

La détection et la classification sont les deux éléments d'analyse fondamentaux d'une brique de reconnaissance gestuelle.

La détection consiste formellement à déterminer si un geste connu a effectivement été

réalisé. Cela revient à rechercher dans les données sélectionnées pour analyse, la présence de ce qui s'apparente suffisamment au geste théorique, ou à une modélisation de celui-ci. En effet, ce qui est caractéristique des gestes, c'est qu'ils varient d'une exécution à une autre, que ce soit pour des personnes différentes ou pour une même personne. De plus, les variations sont aussi bien spatiales : un geste pourra être plus ou moins ample, que temporelles : un geste pourra être réalisé plus lentement ou plus rapidement. Il ne s'agit donc pas d'une recherche exacte.

La classification consiste quant à elle, à déterminer quel geste a été réalisé parmi les différentes possibilités du vocabulaire. Il s'agit donc de mettre les gestes théoriques en concurrence.

A priori donc, il serait logique de commencer par détecter un geste, puis de le catégoriser. Cependant, la problématique est plus complexe puisque parfois, du fait de la variabilité des gestes, on ne sait pas exactement ce que l'on cherche à détecter. Il est alors difficile de trouver un point de départ pour l'analyse.

Pour cette raison, la détection et la classification sont généralement réalisées de manière conjointe. Pour chaque geste du dictionnaire, un modèle est défini à priori, et lors d'une nouvelle analyse, on estime le niveau de correspondance entre les données et chacun des modèles (suivant les algorithmes, on parlera de distance ou de probabilité). Ensuite, une détection survient lorsque le niveau de correspondance avec au moins un des modèles est suffisant (dépasse un seuil fixe ou adaptatif), et la classification se fait en choisissant le modèle avec lequel la correspondance est la plus forte.

En fin d'analyse, la sortie de l'association de ces deux fonction est en général le triplet suivant :

— Détection : une valeur binaire qui indique si un geste a été détecté.
— Classification : l'indice du geste qui a été détecté.
— Confiance : un indicateur de la fiabilité de l'analyse.

5.1.1.4 Suivi de la progression

Le suivi de la progression est un autre type d'analyse qui peut être réalisé. En faisant l'hypothèse que l'on sait qu'un geste est en cours et duquel il s'agit, il est alors possible de suivre, en continu, la progression de son exécution. Il est également possible d'anticiper son évolution au regard d'un modèle. La sortie de cette fonction est donc le singleton :

— Progression : l'indice temporel dans le modèle du geste suivi.

Le suivi de geste, bien que assez peu représenté dans la littérature (Bevilacqua *et al.* 2011), est particulièrement intéressant et notamment dans le domaine de l'art où cela permet, par exemple, de synchroniser des évènements sonores ou visuels avec le déroulement d'une chorégraphie.

Cependant, ce procédé semble surtout applicable à des gestes longs. De plus, l'hypothèse que l'on sait qu'un geste est en cours, dès le début de son exécution, est assez forte. En effet, cela suppose soit que l'on ne considère qu'un unique geste, soit que les gestes analysés sont très différents, de sorte à ce qu'aucune ambiguïté ne puisse subsister.

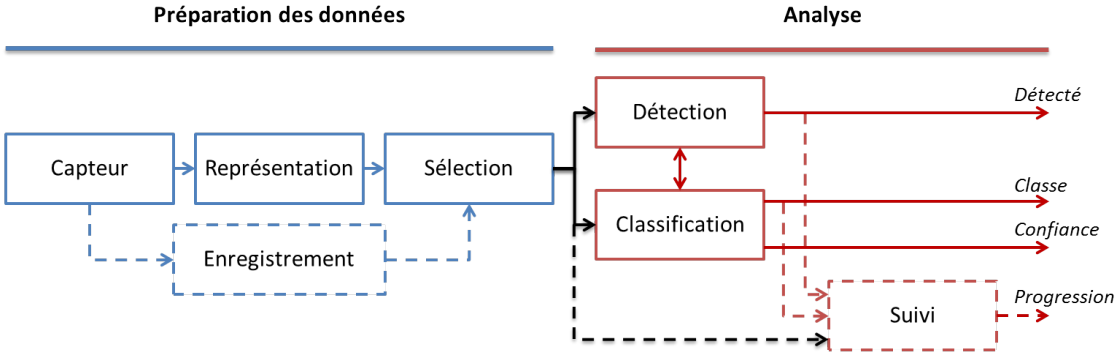


FIGURE 5.1 – Synthèse des fonctions élémentaires de tout système gestuel.

5.1.1.5 Synthèse de l'architecture standard

Ainsi, comme illustré par la Figure 5.1, une brique de reconnaissance gestuelle est composée de différents éléments.

Nous avons donc implémenté un système de reconnaissance comprenant ces différentes parties, hormis le suivi de geste.

5.1.2 Approches classiques avec apprentissage automatique

L'analyse du mouvement pour la reconnaissance gestuelle a de particulier de devoir traiter des données multi-dimentionnelles qui présentent non seulement une forte variabilité spatiale, mais également temporelle. En effet, un geste peut être réalisé plus ou moins rapidement, sans que cela ne change sa signification (sauf pour de rares cas comme en langue des signes).

De plus, cette variabilité temporelle n'est ni constante ni linéaire ce qui impose l'usage d'algorithmes spécifiques ; parmi lesquels l'alignement temporel dynamique et les modèles de Markov cachés sont les plus utilisés.

5.1.2.1 Alignement temporel dynamique

De manière générale, l'alignement temporel (*Time Warping*) consiste à modifier la progression d'une série de données pour la faire correspondre le plus possible à une seconde série à laquelle on souhaite la comparer. Les modifications applicables sont la compression et la dilatation du temps. Ces deux opérations préservent néanmoins l'ordre des éléments.

Une fois deux séries alignées, on peut estimer leur niveau de correspondance en utilisant la formule

$$D(A, B, W) = \frac{\sum_t d(A_t, B_{W(t)})}{M + N} \tag{5.1}$$

où A et B sont les deux séries comparées. Respectivement de taille M et N, leurs éléments sont du même type (appartiennent à un même espace). $W(t)$ est la fonction d'alignement qui associe à chaque élément de la série A, l'élément qui lui correspond dans la série B.

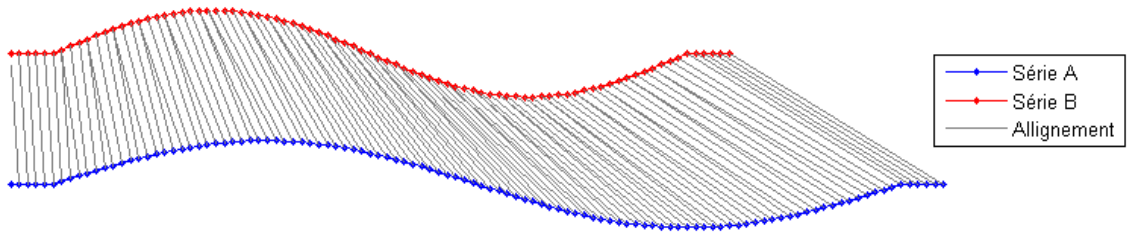


FIGURE 5.2 – Exemple d'alignement temporel entre deux séries de données.

Enfin, $d(a, b)$ est une fonction qui détermine le niveau de correspondance entre un point de la série A et un point de la série B. Généralement, on utilise une distance euclidienne, ce qui implique que $D(A, B, W)$ est nul lorsque les deux séries alignées sont parfaitement identiques et augmente lorsqu'elles différent. La division par $M + N$ est une normalisation pour que le niveau de correspondance estimé ne dépende pas de la taille des séries (Myers & Rabiner 1981).

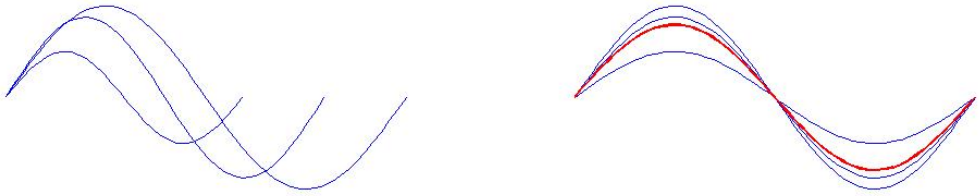
L'utilisation de cette formule suppose donc que l'on connaisse la fonction d'alignement $W(t)$ à utiliser. Or, parmi de nombreuses possibilités, celle à utiliser est celle qui optimise le niveau de correspondance. Autrement dit, pour calculer $D(A, B, W)$, il faut connaître $W(t)$; mais $W(t)$ est déterminé par le $D(A, B, W)$ optimal. Ainsi, les deux doivent être déterminés conjointement.

L'approche naïve consistant à tester toutes les fonctions d'alignement possibles pour ne conserver que celle qui optimise le résultat n'est pas applicable car trop coûteuse (trop de possibilités à tester). Une autre approche a donc été proposée (Sakoe & Chiba 1978) : l'utilisation d'un algorithme de programmation dynamique ; logiquement nommé *Dynamic Time Warping* (DTW).

Cet algorithme permet donc d'aligner de manière optimale deux séries temporelles et de calculer leur niveau de correspondance. Par ailleurs, différentes variantes de cet algorithme existent pour optimiser les temps de calcul : l'espace de recherche peut être réduit de manière statique (Sakoe & Chiba 1978; Itakura 1975) ou dynamique (Candan *et al.* 2012) ou en compressant les séries de données avant de chercher à les aligner (Keogh & Pazzani 1999; Dupont & Marteau 2015).

Dans le cadre de la reconnaissance gestuelle, des mouvements peuvent être décrits sous forme de séries temporelles ; la DTW permet donc d'estimer le niveau de correspondance entre deux mouvements et son usage est alors assez intuitif : un ou plusieurs exemples de chaque geste à reconnaître sont enregistrés et fournis à la machine. Ensuite, lors d'une analyse, la DTW est utilisée pour comparer les nouvelles données à l'ensemble des exemples connus. Si les données sont relativement semblables à l'un des exemples, la machine sait alors qu'un geste a été réalisé et duquel il s'agit.

Ainsi, un score de DTW est calculé entre la série de données à analyser et chacun des exemples enregistrés. Si au moins un score est inférieur à un seuil qui aura été fixé à priori, il y a détection ; et, pour la classification, c'est le meilleur score qui l'emporte. Dans le cas où un geste est décrit par plusieurs exemples, un système classique de vote est utilisé (plus



(a) (b)

FIGURE 5.3 – Exemple de construction d'un geste prototype à partir de différents exemples, en utilisant la DTW. (a) : avant alignement, et (b) après alignement et calcul du geste prototype (en rouge).

proches voisins).

Finalement, un vocabulaire gestuel peut être *modélisé* par une base d'exemples enregistrés, la détection est réalisée par seuillage et la classification par vote. Seule reste la question de l'apprentissage : comment construire la base d'enregistrements ? La méthode la plus simple est d'ajouter autant d'exemples qu'on le souhaite : à minima un exemple par geste.

Cependant, il est important de considérer que lors d'une analyse, on réalise une DTW pour chaque exemple de la base. En conséquence, plus la base est grande, plus les temps de calcul sont importants. Une base trop grande peut alors ajouter de la latence, voir ne pas être utilisable en temps réel. Ainsi, des stratégies doivent êtres utilisées pour en limiter la taille : il peut être fait le choix de ne conserver qu'un ou deux exemples pour chaque geste en ne sélectionnant que les plus récents (Liu *et al.* 2009) ou les plus représentatifs : le niveau de correspondance est calculé entre chaque couple d'exemple d'un même geste et seuls les exemples ayant en moyenne les meilleurs niveaux de correspondance sont conservés. Plutôt que de réaliser une sélection, un *exemple médian* peut aussi être construit (Wilson & Wilson 2004; Choe *et al.* 2010) : comme représenté en Figure 5.3, pour chaque geste, différents exemples sont alignés sur une même échelle de temps et la série temporelle médiane (ou moyenne) est calculée.

5.1.2.2 Modèle de Markov caché

Les modèles de Markov cachés (MMC) de l'anglais Hidden Markov Models (HMM), sont une autre approche pour la modélisation et la reconnaissance des gestes et comme pour les DTW, leur usage est largement hérité du domaine de la reconnaissance de la parole (Rabiner & Juang 1986).

De manière générale, un HMM sert à modéliser statistiquement le comportement dynamique d'un processus :

— Le processus comporte un nombre déterminé d'états : différents sous-processus ou différentes configurations, par lesquels il passe successivement au cours du temps.
— A un instant donné, le processus ne peut être que dans un seul état à la fois.

— A un instant donné, le choix de changer d'état (effectuer une transition) est complètement déterminé par l'état actuel.
— L'état dans lequel le processus se trouve à un instant donné n'est pas observable ; d'où l'appellation d'états *cachés*.
— Le processus possède des sorties qui elles, sont observables, et dont les valeurs dépendent uniquement de l'état actuel du processus.
— Les valeurs des observations (sorties) peuvent être des variables continues ou discrètes ; et dans ce second cas, on parle de symboles.

De manière formelle, un HMM λ est caractérisé par $\lambda = (N, M, A, B, \pi)$:
— N le nombre d'état cachés.
— M le nombre de symboles observables.
— A la matrice de taille N^2 des probabilités de transition entre chaque état.
— B la matrice de taille $N * M$ des probabilités d'observation de chaque symbole pour chaque état.
— π le vecteur de taille N des probabilités pour chaque état d'être l'état actif au démarrage du processus.

Associés à cette modélisation, différents algorithmes standards permettent de réaliser les trois fonctions fondamentales suivantes (Rabiner & Juang 1986) :

1. **Estimation :** En considérant une séquence d'observations et un modèle dont on connaît les paramètres, l'algorithme *Forward-Backward* permet de déterminer le niveau de correspondance entre les deux ; c'est à dire de calculer la probabilité que la séquence d'observation ait été générée par le modèle en question (Baum *et al.* 1967; Baum & Sell 1968).
2. **Décodage :** En considérant une séquence d'observations et le modèle qui a généré cette séquence, l'algorithme de *Viterbi* permet d'estimer la séquence des états cachés la plus probable (Viterbi 1967; Forney Jr 1973).
3. **Apprentissage :** En considérant un ensemble de séquences d'observations, l'algorithme de *Baum–Welch* permet d'estimer et d'optimiser les paramètres d'un modèle afin que celui-ci explique le mieux possible ces séquences (Dempster *et al.* 1977).

Ainsi, pour un dictionnaire gestuel à reconnaître, un HMM est appris pour chaque geste. Puis, lors d'une nouvelle analyse, on estime la probabilité entre chaque modèle et les données sélectionnées. Si parmi les probabilités estimées, celle d'au moins un modèle dépasse un seuil, il y a détection. Le seuil peut être une valeur fixe (Amft *et al.* 2005) choisie à priori et ne dépendant pas des données analysées, ou adaptatif (Lee & Kim 1999). Dans ce second cas, un HMM supplémentaire, souvent appelé *garbage model* (Wilcox & Bush 1992), est construit de sorte à modéliser le mieux possible les phases de non-geste. C'est alors l'estimation faite pour ce modèle qui sert de seuil de détection. Pour la classification, c'est le modèle présentant la probabilité la plus forte qui est sélectionné et cette probabilité peut être conservée comme indicateur du niveau de confiance de l'analyse.

L'apprentissage consiste quant à lui, à choisir le nombre d'état N des modèles, estimer les différentes probabilités d'observation et de transition, et, dans le cas d'un seuillage statique, de définir la valeur du seuil. L'apprentissage des paramètres numériques des modèles

est réalisé avec l'algorithme de *Baum–Welch* auquel on présente des exemples qui auront été enregistrés au préalable. Il n'existe cependant pas de manière évidente pour estimer le nombre d'état et la valeur du seuil. Pour ces deux paramètres, une procédure itérative est généralement employée : avec une base d'enregistrements contenant des exemples de chaque geste, on réalise des tests d'apprentissage et de reconnaissance en faisant varier le nombre d'état et le seuil et on conserve la configuration qui donne les meilleurs résultats.

5.1.2.3 Comparaison entre ces deux approches

Les HMM et la DTW permettent donc de modéliser et de reconnaître des gestes. Quelques différentes sont cependant notable entre ces deux technologies :

— Tout d'abord, la DTW est une approche déterministe alors que les HMM sont une approche probabiliste. Dans le premier cas, on parle de distance ou de similitude par rapport à un exemple, alors que dans le second cas, on parle de probabilité d'occurrence selon un modèle. Cette distinction, principalement d'ordre conceptuelle importe en soit assez peu mais participe néanmoins à rendre le comportement des systèmes basés HMM plus difficile à appréhender.

— Une seconde distinction est le traitement qui est fait de la dimension temporelle. Bien que ces deux approches permettent de traiter des gestes de durée variable, seul le résultat des HMM en est numériquement impacté. En effet, alors que la DTW procède en deux étapes : aligne les données, puis seulement, estime le niveau de similitude sur les dimensions spatiales ; les HMM modélisent l'aspect temporel avec la matrice des probabilités de transition et l'utilise lors de l'estimation. Ainsi, la probabilité calculée par un HMM décroît lorsque le geste analysé est réalisé avec une dynamique différente de celle modélisée. En revanche l'estimation faite avec une DTW reste invariante. Cette différence pourrait constituer un critère de choix entre HMM et DTW suivant si la sémantique des gestes à reconnaître dépend de leur temporalité ou non. Cependant, et à notre connaissance, ce critère ne semble pas considéré dans la littérature.

— Une autre différence entre HMM et DTW est le nombre d'exemples requis pour effectuer l'apprentissage. Alors que dans le cas de la DTW un unique exemple est suffisant (*one-shot learning*), il en faut en revanche bien d'avantage pour estimer statistiquement l'ensemble des paramètres pour un HMM (Carmona & Climent 2012). Il a certes été montré que du *one-shot learning* est également possible avec un HMM (Bevilacqua *et al.* 2009), mais il faut pour cela réduire le nombre de paramètres en contraignant fortement le modèle, ce qui réduit théoriquement son pouvoir de description. Ainsi, d'un point de vu pratique, la DTW semble avantageuse puisque ne nécessitant pas de campagne d'enregistrement particulièrement chronophage.

— En matière de performances, il ne semble exister aucune évidence quant à la supériorité d'une méthode sur l'autre. En effet, alors que certains auteurs indiquent les HMM comme significativement meilleurs avec des taux de reconnaissance de l'ordre de 90% pour les HMM contre seulement 70% pour la DTW (Wilson & Wilson 2004; Corradini & Gross 2000), un résultat plus récent (Carmona & Climent 2012) se

porte lui, en faveur de la DTW mais avec une différence relativement faible : 96% pour les HMM contre 99% pour la DTW.

— Finalement, le dernier éléments de comparaison entre ces deux technologies est le temps de calcul ; et sur ce plan, à performances égales, les HMM semblent plus avantageux (Carmona & Climent 2012). En effet, dans le cas des HMM, on utilise un unique modèle pour chaque geste alors que dans le cas de la DTW on utilise généralement plusieurs exemples par geste. Or, le temps de calcul est directement proportionnel au nombre d'exemples, ce qui explique la différence en faveur des HMM ; une différence qui s'accroît d'autant plus que le nombre de geste à reconnaître est important. Cependant, dans le cadre d'un vocabulaire gestuel de faible dimension, ce qui est notre cas, la différence reste minime.

En conclusion, il apparaît que ces deux approches semblent relativement équivalentes en ce qui concerne leurs performances. Ainsi n'importe laquelle de ces deux technologies pourrait être utilisée. La table 5.1 présente un résumé des quelques points de comparaison que nous venons d'aborder.

	DTW	HMM
Type d'approche	déterministe	probabiliste
Résultat impacté par la dynamique du geste	non	oui
Nombre d'exemples requis pour l'apprentissage	peu (1)	beaucoup (>=50)
Performances	équivalentes	
Temps de calcul	en faveur des HMM	

TABLE 5.1 – Synthèse des comparaison entre les deux approches majeures utilisée pour la reconnaissance gestuelle : DTW et HMM.

5.1.3 Expressions régulières pour la reconnaissance gestuelle

5.1.3.1 De l'apprentissage automatique à la description formelle

HMM et DTW reposent sur un principe d'apprentissage automatique : d'apprentissage par l'exemple. Il s'agit de fournir au système un certain nombre d'exemples de réalisation de chaque geste à reconnaître pour en extraire automatique des caractéristiques discriminantes. Cela revient donc à demander au système de trouver par lui-même la norme qui régit chaque geste.

Or, dans le cadre de gestes sémaphoriques standardisés, on peut s'interroger sur la pertinence d'une telle approche. En effet, nous connaissons cette norme qui a été définie lors du choix des gestes et que nous imposons par ailleurs, autant à la machine qu'à son utilisateur.

Ainsi, utiliser une approche d'apprentissage automatique revient, dans notre cas, à expliquer à des participants comment réaliser chaque geste avant d'en enregistrer des exemples ; exemples que l'on fourni ensuite à la machine pour qu'elle en construise une

modélisation. Pourquoi alors, ne pas directement donner à la machine les règles que l'on donne à l'utilisateur ?

Cette approche semble d'autant plus pertinente que l'apprentissage automatique dépend fortement de la qualité des exemples que l'on utilise ; notamment leur justesse et le niveau de couverture (Fothergill *et al.* 2012) qui sont des paramètres qui doivent être contrôlés.

Cependant, une approche par description formelle soulève de nouvelles questions : comment décrire à la machine un geste, dans quel langage, et quel algorithme utiliser pour la reconnaissance ?

Une première idée serait de trouver une manière de construire directement un HMM à partir d'une connaissance experte du geste ou, plus simplement, d'une description en langage naturel de celui-ci. Cependant, les paramètres numériques d'un HMM n'ont pas tous de correspondance physique ou sémantique évidente avec le geste qu'ils modélisent : en particulier le nombre d'états et le seuil de détection. Créer un modèle de toute pièce ne semble donc pas possible de manière simple.

Ainsi, pour répondre à cette problématique, nous nous sommes intéressé à un autre outil : les *expressions régulières* qui sont bien connues dans les domaines de l'informatique théorique et de la théorie des langages.

5.1.3.2 Introduction aux expressions régulières

Dans la théorie des langages, un *langage* est un ensemble de *mots* valides dans un contexte spécifique. Ces mots sont construits en combinant des *éléments atomiques* (insécables) d'un *alphabet* dont la taille est finie. Les combinaisons valides sont définies par énumération ou par des règles. Enfin l'ensemble de ce qui définit un langage constitue une *grammaire*.

Classiquement, différentes familles de grammaires (et donc de langages) existent. La hiérarchie de Chomsky (Chomsky 1956) en distingue quatre imbriquées qui, par ordre de complexité décroissante, sont :

— (L0) les grammaires générales,
— (L1) les grammaires contextuelles,
— (L2) les grammaires non contextuelles ou algébriques et
— (L3) les grammaires régulières.

En informatique, les grammaires sont beaucoup utilisées pour l'analyse de documents ; en particulier les grammaires algébriques pour l'analyse syntaxique (structure) et les grammaires régulières pour l'analyse lexicale (vocabulaire).

Deux types d'analyse sont utilisées : la *validation* qui consiste à vérifier si une chaîne de caractère appartient dans son intégralité à un langage défini, et la *recherche* qui consiste à trouver et extraire tous les éléments qui appartiennent à un langage donné dans une chaîne de caractères.

Ces fonctions sont assurées par différents algorithmes (machines) suivant le type de grammaire : les machines de Turing pour les grammaires générales, les automates linéairement bornés pour les grammaires contextuelles, les automates à pile non déterministe pour les grammaires algébriques et les automates finis pour les grammaires régulières.

Finalement, dans ce contexte, une *expression régulière*, ou *expression rationnelle*, est une manière de décrire une grammaire régulière. Il s'agit d'une chaîne de caractères respectant une syntaxe précise. Après écriture, cette chaîne est transformée, par un compilateur, en automate fini constitué d'un ensemble d'états et de transitions logiques. L'automate généré permet ensuite l'analyse de documents qu'on lui fourni.

Les bases de la syntaxe reposent sur l'usage de symboles atomiques combinés avec différents opérateurs. Les symboles atomiques sont les caractères eux mêmes ('a','A', ...) ou leur code ASCII avec la forme '\xFF' ('a'='\x61').

Parmi les opérateurs principaux, on trouve :

— la concaténation : 'ab' se traduit par *'a' suivit de 'b'*,

— l'union : 'ab|bc' qui signifie *'ab' ou 'bc'*, et

— le regroupement : '[ab]' qui signifie *un parmi 'a' ou 'b'*.

S'ajoutent également les opérateurs de quantification qui permettent d'indiquer le nombre de fois qu'un symbole ou qu'un groupe de symboles est attendu :

— ' ?' se traduit par *zéro ou une fois*,

— '+' se traduit par *au moins une fois*,

— '*' se traduit par *zéro ou plusieurs fois*, et

— '{n,m} se traduit par *de n à m fois*.

Les opérateurs ont des ordre de priorité dans leur application ; cependant, l'usage de parenthèse permet de modifier cet ordre. Par exemple '(ab){2}' est équivalent à 'abab', alors que 'ab{2}' est équivalent à 'abb'.

En résumé, les expressions régulières sont un outil informatique élégant permettant de décrire formellement une grammaire régulières. Une fois compilées ces chaînes de caractères sont transformées en automates qui permettent la validation et la recherche de pattern spécifiques.

5.1.3.3 Application à la reconnaissance gestuelle

Les expressions régulières permettent donc la description formelle de motifs recherchés dans du texte ; mais comment alors les utiliser dans le contexte de la reconnaissance gestuelle ?

D'une part, elles permettent une analyse lexicale, ce qui correspond à la reconnaissance des gestes isolés. Chaque geste est décrit par une expression régulière et les fonctions de validation et de recherche sont alors utilisables :

— La validation permet de contrôler si des données sélectionnées correspondent ou non au geste décrit. Cela suppose cependant une sélection de type *segmentation* qui ne conserve que les données du geste et exclut tout résidu parasite en début ou fin de celui-ci ; en effet, ces résidus de non-geste invalideraient le motif.

— La recherche quant à elle, extrait la ou les occurrences de geste dans les données sélectionnées et ignore les résidus éventuels. La recherche est donc adaptée à une sélection de type fenêtre glissante.

D'autre part, il est usuel que les expressions régulières soient utilisées pour de l'analyse de texte, mais ce n'est aucunement une restriction. En effet, il est possible de redéfinir les

symboles atomiques de l'alphabet et ainsi étendre leur usage à n'importe quelle séquence de données. Pour utiliser les expressions régulières dans notre contexte, il est donc nécessaire de définir un nouvel alphabet de taille fini qui soit représentatif des composantes d'un geste.

Au regard du dictionnaire gestuel constitué selon le protocole présenté au chapitre 3, la sémantique des gestes dépend de l'évolution temporelle de la configuration, du mouvement et de l'orientation de la main. Les symboles atomiques devront donc être une combinaison de ces trois éléments.

Ainsi, sur la base de ce qui semble sémantiquement pertinent, nous avons fait les choix suivants :

— Pour la configuration, on caractérise l'état de chaque doigt : (0)plié ou (1)tendu ; hormis pour le pouce dont l'extension n'est pas nécessairement marquée du fait du stress que cela provoque. Ainsi, la configuration est définie par le quadruplet de booléens : D_i pour l'index, D_m pour le majeur, D_r pour l'annulaire et D_l pour l'auriculaire.

— Pour l'orientation, trois dimensions sont disponibles : le lacet (angle autour de l'axe vertical), le roulis (angle autour de l'axe avant-arrière) et le tangage (angle autour de l'axe droite-gauche). Cependant, le lacet dépend de l'orientation, à la fois de la main et à la fois de la personne ; il peut alors être raisonnable de négliger cette dimension. Pour chacune des deux dimensions restantes, quatre valeurs semblent sémantiquement pertinentes : paume de la main (0)vers le haut, (1)le bas, (2)l'intérieur et (3)l'extérieur pour le roulis ; et direction des doigts vers (0)l'avant, (1)l'arrière, (2)le haut ou (3)le bas pour le tangage. Ici, les termes intérieur et extérieur sont repris du vocabulaire de l'escrime et ont la particularité d'être symétriques : ils ne dépendent pas de la main utilisée (droitier / gaucher). Ainsi, *intérieur* désigne la direction de la main vers le corps et *extérieur* désigne la direction opposée. Finalement, l'orientation est définie par O_r pour le roulis et O_t pour le tangage ; chacun pouvant prendre quatre valeurs.

— Pour le mouvement, les trois dimensions de l'espace sont disponibles, et pour chacune, trois états sémantiques sont considérés : (0)une accélération négligeable, (1)une accélération positive et (2)une accélération négative. Ainsi, le mouvement est défini par le triplet A_x, A_y et A_z ; chaque élément pouvant prendre trois valeurs.

Finalement, en concaténant ces trois éléments, on obtient un symbole atomique dont la structure est la suivante : $< D_l, D_r, D_m, D_i, O_r, O_t, A_z, A_y, A_x >$. L'alphabet ainsi construit contient alors $2^4 * 4^2 * 3^3 = 6912$ symboles différents dont voici deux exemples :

— Poing fermé, statique, paume vers l'intérieur et direction des doigts vers l'avant $< 0, 0, 0, 0, 2, 0, 0, 0, 0 >= (512)_{10} = (200)_{16}$

— Main ouverte, statique, paume vers le haut et direction des doigts vers l'avant $< 1, 1, 1, 1, 0, 0, 0, 0, 0 >= (15360)_{10} = (3C00)_{16}$

5.1.3.4 Description du vocabulaire gestuel

Une fois l'alphabet défini, on peut écrire l'expression régulière pour chacun des gestes du vocabulaire à reconnaître ; et pour ce faire, nous partons du modèle anatomique de

Kita (Kita *et al.* 1997) présenté au Chapitre 2. Selon celui-ci, un geste comprend donc cinq grandes phases : la préparation, une pause pré-stroke, un stroke, une pause post-stroke et la rétraction.

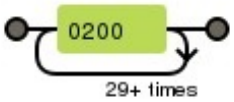
Pour l'écriture des expression régulières, nous ne tenons pas compte des phases de préparation et de rétraction qui constituent le départ et le retour de la main à sa position de repos. En effet, d'une part, ces deux phases ne participent pas directement à la sémantique du geste, et d'autre part, elles dépendent de la position de repos qui est variable lors d'un usage naturel et en rend donc difficile, voir impossible, la description formelle.

En se restreignant aux trois phases centrales, trois types de gestes se distinguent alors : les gestes statiques, les gestes dynamiques simples et les gestes dynamiques complexes.

Description des gestes statiques

Les gestes statiques ont la particularité de ne présenter aucun mouvement, changement d'orientation ni de configuration pendant la phase de stroke. Ainsi, pour ces gestes, les trois phases se confondent pour n'en former qu'une seule. Cela implique une expression régulière simple de la forme *\xFFFF{N,}* avec $FFFF$ un code hexadécimal et N un nombre entier. Le code *\xFFFF* correspond à la configuration statique de la main avec les trois composantes d'accélérations nulles. Le quantificateur $\{N,\}$, exprime la durée minimale de la pause pour être valide et se traduit par : *le symbole qui précède dois être présent consécutivement au moins N fois, sinon plus*. N indique donc une durée et dépend de la fréquence des données sélectionnées. Avec une fréquence de 60Hz et une pause d'au moins une demi-seconde : $N = 30$. Dans le dictionnaire défini, trois gestes sont statiques :

— **Valider** : la main est fermée, la direction des doigts vers l'avant et la paume vers l'intérieur ; ce qui se traduit par l'expression régulière **\x0200{30,}**. L'automate qui correspond à cette expression, et généré par l'outil en ligne *Regexper* est le suivant :



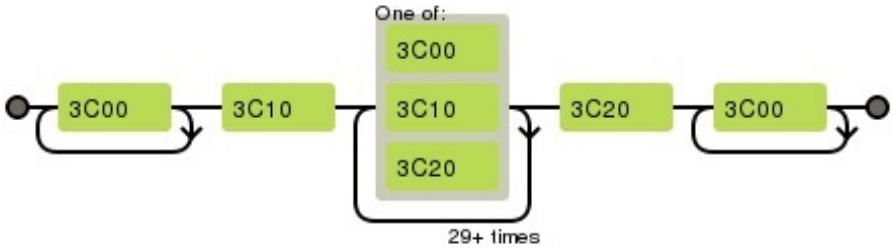
— **Annuler** : la main est fermée, la direction des doigts vers l'avant et la paume de la main vers l'extérieur : ce qui se traduit par l'expression régulière **\x0300{30,}**.

— **Stop** : la main est fermée, la direction des doigts vers le haut et la paume de la main vers l'intérieur ; ce qui se traduit par l'expression régulière **\x0280{30,}**.

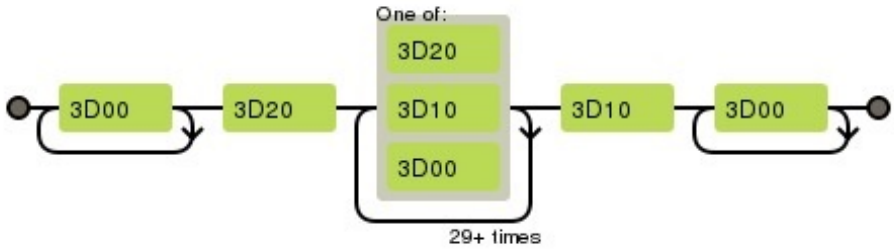
Description des gestes dynamiques simples

Les gestes dynamiques simples sont ceux pour lesquels le stroke comporte un mouvement uniforme qui peut être décrit simplement. Dans notre vocabulaire gestuel, les gestes de décollage, d'atterrissage et de retour à la base présentent seulement des translations verticales qui peuvent alors s'écrire :

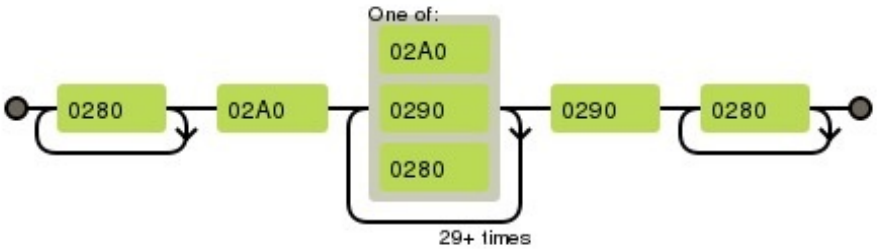
— **Décoller** : la main est ouverte, la direction des doigts vers l'avant et la paume de la main vers le haut. Le stroke est un mouvement linéaire vers le haut. L'expression régulière est alors **\x3C00+\x3C10[\x3C00\x3C10\x3C20]{30,}\x3C20\x3C00+**.



— **Atterrir** : la main est ouverte, la direction des doigts vers l'avant et la paume de la main vers le bas. Le stroke est un mouvement linéaire vers le bas. L'expression régulière est alors **\x3D00+\x3D20 [\x3D20\x3D10\x3D00]{30,} \x3D10\x3D00+**.



— **Retour à la base** : la main fermée, la direction des doigts vers le haut et la paume de la main vers l'intérieur. Le stroke est un mouvement linéraire vers le bas. L'expression régulière est alors **\x0280+\x02A0 [\x02A0\x0290\x0280]{30,} \x0290\x0280+**.



(a) (b)

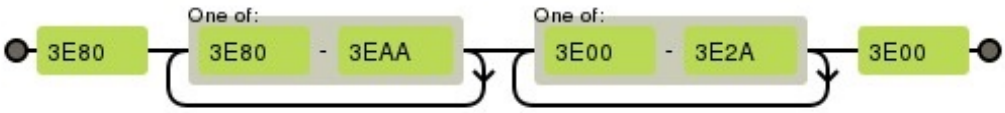
FIGURE 5.4 – Deux exemples d'exécutions valides pour le geste *aller au point suivant*. Il apparaît que le mouvement naturel lors d'un tel geste est complexe. Ce qui est sémantiquement important, c'est la rotation de la main dans le plan sagittal, mais cela peut s'accompager de mouvements de translation. Ainsi, une infinité de mouvements sont possibles ; parmi les quels : (a) un quart de cercle au niveau du coude, et (b) une rotation et une translation simultanées de la main assurées conjointement par les articulations du coude et de l'épaule.

Description des gestes dynamiques complexes

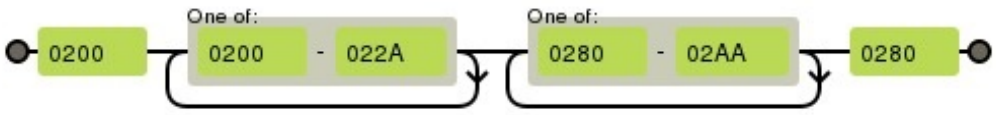
Finalement, pour les deux gestes restants, le mouvement du stroke est plus complexe. Bien que sémantiquement, ce qui importe, c'est un changement de 90 de l'orientation de la main dans le plan sagittal, comme l'illustre la Figure 5.4, ce changement peut être réalisé par une rotation pure au niveau du coude mais peut également être accompagné de translations permises conjointement par les articulations du coude et de l'épaule.

Le motif du stroke peut ici être exprimé par une succession de deux états : chacun reflétant une orientation différente mais avec n'importe quelle accélération. Cependant, en expression régulière, les symboles étant insécables, on ne peut donc pas exprimer directement *quelque soit l'accélération*. Il faut alors énumérer tous les cas ; ce qui représente $3^3 = 27$ symboles pour chaque état. Heureusement, dans notre cas, on peut utiliser la notation *[a-b]* qui exprime *tout les symboles de 'a' à 'b' (compris)*. Ainsi, de manière élégante, les expressions régulières pour les deux derniers gestes sont :

— **Suivant** : Au début du geste, la main est ouverte, la direction des doigts vers le haut et la paume vers l'intérieur. Puis, avec un mouvement en arc de cercle, la main atteint la position d'arrivée : main toujours ouverte, la direction des doigts vers l'avant et la paume toujours vers l'intérieur. L'expression régulière est alors **\x3E80 [\x3E80-\x3EAA]+ [\x3E00-\x3E2A]+ \x3E00**.



— **Précédent** : Au début du geste, la main est fermée (pouce tendu), la direction des doigts est vers l'avant et la paume vers l'intérieur. Puis, avec un mouvement en arc de cercle, la main atteint la position d'arrivée : main toujours fermée, la direction des doigts vers le haut et la paume toujours vers l'intérieur. L'expression régulière est alors **\x200 [\x200-\x22A]+ [\x280-\x2AA]+ \x280**.



Ainsi, après avoir redéfini un alphabet spécifique à notre dictionnaire de gestes, une expression régulière a été construite pour décrire formellement chacun des gestes à reconnaître. Une fois compilées, ces expression régulières, dont un résumé est présenté en Table 5.2, sont utilisables à la fois avec la fonction de validation et la fonction de recherche ; nous avons cependant fait le choix d'utiliser la fonction recherche associée à un mode de sélection des données de type fenêtre glissante.

Ce mode de sélection présente l'avantage d'être générique et de ne nécessiter aucune connaissance à priori des gestes. Ainsi les connaissances expertes sur le vocabulaire gestuel sont uniquement contenues dans les expressions régulières ; et de cette manière, ajouter de nouveaux gestes, ou changer ceux déjà définis implique seulement l'édition des expressions régulières et aucunement la modification de la méthode de sélection.

Finalement, la définition de l'alphabet utilisé ici impose de manière implicite la représentation des données (le descripteur) ; et pour compléter notre de reconnaissance gestuelle, il ne reste maintenant plus qu'a choisir un capteur et les traitements associés.

Valider	\x0200{30,}
Annuler	\x0300{30,}
Stop	\x0280{30,}
Décoller	\x3C00+\x3C10[\x3C00\x3C10\x3C20]{30,}\x3C20\x3C00+
Atterrir	\x3D00+\x3D20 [\x3D20\x3D10\x3D00]{30,} \x3D10\x3D00+
Retour à la base	\x0280+\x02A0 [\x02A0\x0290\x0280]{30,} \x0290\x0280+
Suivant	\x3E80 [\x3E80-\x3EAA]+ [\x3E00-\x3E2A]+ \x3E00
Précédent	\x200 [\x200-\x22A]+ [\x280-\x2AA]+ \x280

TABLE 5.2 – Synthèse des expressions régulières construites pour décrire notre dictionnaire gestuel.

5.1.4 Choix du capteur

5.1.4.1 Capteurs existants

Pour capter les mouvements d'un corps humain, que ce soit dans son ensemble, ou seulement d'une partie, différentes technologies existent. Une première distinction peut être faite entre deux grandes approches (Yang *et al.* 2013) : celles basées vision, c'est à dire lorsqu'une ou plusieurs caméras sont mises en oeuvre, et les approches non basées vision qui imposent le port de composants actifs.

Approches basées vision

Les approches basées vision sont certainement les plus représentées dans la littérature. Généralement peu intrusives, elle présentent l'avantage de ne pas gêner les mouvements. Cependant, l'usage d'une ou de plusieurs caméras, fixe un champs de vision et impose une zone de travail qui convient assez peu à un contexte de grande mobilité. De plus, suivant les technologies, la variation de l'environnement peut perturber, ou rendre complexe, la captation des mouvements (environnement visuellement complexe, bruité ou luminosité changeante). Les approches basées vision souffrent également de la problématique de masquage : les zones qui sont cachées, même temporairement ne peuvent naturellement pas être captées.

Deux paramètres permettent de distinguer les approches basées vision : d'une part les systèmes 2D peuvent être opposés aux systèmes 3D, et d'autre part,l'utilisation ou non de marqueurs.

Les systèmes 2D, sont les moins coûteux mais sont également plus complexes en terme de traitement d'image : seuls des critères de forme, de couleur et de déplacement sont utilisables pour localiser et suivre les gestes. Les systèmes 3D fournissent quant à eux, une information de profondeur robuste aux conditions d'éclairage (visible). Cette information de profondeur permet notamment de simplifier la détection et la localisation des objets en les séparant naturellement de l'evironnement qui est d'avantage éloigné. Différentes technologies 3D existent (Han *et al.* 2016) :

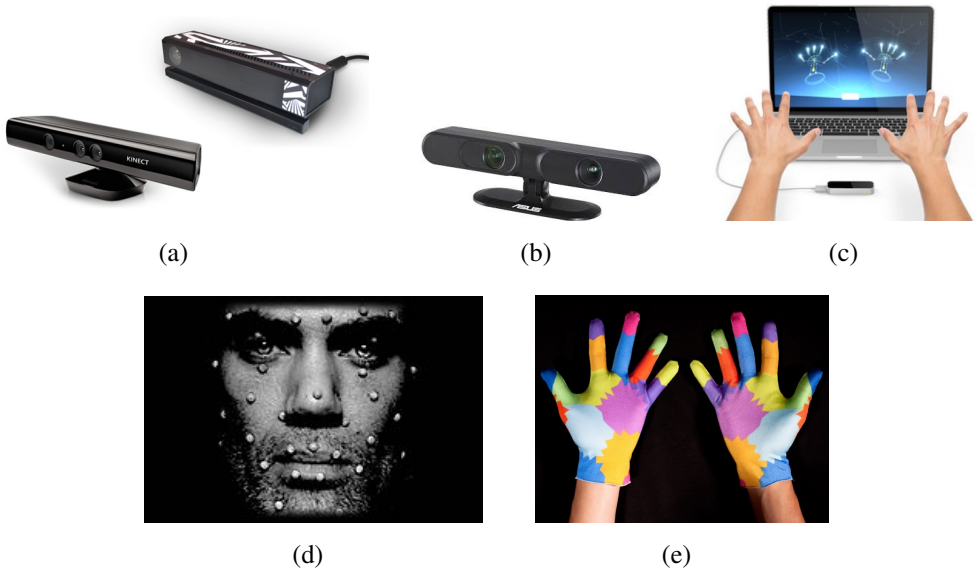
— Les **caméras stéréoscopiques** consistent à utiliser simultanément plusieurs caméras 2D dont les positions relatives sont connues. Les décalages des objets dans les flux vidéos permettent alors d'estimer leurs informations de profondeur. Les caméras stéréoscopiques peuvent nécessiter une phase de calibration avant utilisation.

— Les **caméras à lumière structurée** consistent à projeter dans la scène un pattern lumineux invisible à l'oeil (généralement de l'infra-rouge) et qui se déforme en fonction des surfaces. Ce pattern est ensuite capté par une caméra spécifiquement réceptive à la lumière projetée et l'analyse de la déformation permet d'estimer la carte de profondeur. La *Kinect v1* produite par Microsoft et le *Xtion PRO LIVE* de Asus sont des exemples de capteurs qui utilise cette technologie.

— Les **caméras à temps de vol (ToF)** consistent à illuminer la scène de manière temporaire et à mesurer le temps mis par la lumière pour revenir au capteur. A la manière d'un radar, le temps de vol permet alors d'estimer la distance des éléments de l'espace. La *Kinect v2* de Microsoft ou le capteur *Leap Motion* produit par la société du même nom sont des exemples de capteur low-cost utilisant cette technologie.

Certains systèmes utilisent des marqueurs : de petits éléments réflectifs, pour aider la localisation et le suivi de points particuliers dans les flux vidéos. L'objectif est de simplifier les traitements d'image et de les rendre plus précis. Ces éléments peuvent être utilisés sur n'importe quel objet, comme un drone, ou les différentes partie d'un corps humain pour en capter les mouvements. Cette technologie à laquelle il est souvent fait référence avec l'appellation *Mo-Cap* (motion capture) est particulièrement précise mais très coûteuse et contraignante dans sa mise en oeuvre.

Sur le même principe que les marqueurs, des vêtements présentant des motifs colorés



(a) (b) (c)

(d) (e)

FIGURE 5.5 – Différentes technologies basées vision pour la captation de gestes : (a) Les Kinect V1 (à gauche) et V2 (à droite) de Microsoft ; (b) la caméra de profondeur Xtion PRO LIVE de ASUS ; (c) le capteur Leap de la société du même nom ; (d) un exemple d'utilisation de la motion capture pour le visage ; et (e) un exemple de *color glove* .

particuliers peuvent également être utilisés. Pour les mains, on parle de *color gloves* (Wang & Popović 2009; Lamberti & Camastra 2011).

Approches non basées vision

L'autre grande approche pour capter les mouvements du corps est le positionnement de capteurs actifs. Contrairement aux approches basées vision, il n'y a pas de problématique de champ de vision ni de masquage. Cependant, cette approche est d'avantage invasive du fait du poids des capteurs (principalement des sources énergétiques) et de la connectique qui peuvent gêner les mouvements. A nouveau, différentes technologies existent :

— Les **centrales inertielles** permettent de mesurer l'orientation et l'accélération. Les modèles les plus complets de ces composants permettent la captation de 9 degrés de libertés (DOF) en utilisant 3 accéléromètres, 3 gyroscopes et 3 magnétomètres. Ce type de capteur peut être relativement léger mais souffre cependant de perturbations magnétiques, notamment lors de la proximité avec des éléments métalliques. De plus, les estimation de position et de vitesse peuvent être relativement bruités. Les centrales inertielles peuvent être placées de part et d'autre d'une articulation pour estimer l'état de celle-ci (angle flexion-extension). C'est ce principe qui est utilisé dans le gant *IGS-Cobra Glove* de la société Synertial : 12 centrales inertielles permettent l'estimation de la configuration de la main et de son déplacement.

— Les **capteurs électromagnétiques (et acoustiques)** permettent la captation de leur position et orientation 3D (6 DOF), dans un champs magnétique (ou acoustique) généré par un équipement placé dans l'environnement. Cette solution peut être

particulièrement précise et permet la mobilité mais est spécifique à un environnement contrôlé (du fait de la génération du champs). Les capteurs magnétiques de la société Polhemus sont un exemple souvent utilisé de ce type de capteur.

— Les **capteurs de flexion** permettent la mesure de l'angle de petites articulations (main) et plus généralement de la déformation d'un objet. Deux technologies principales existent :
  — les capteurs optiques consistent à placer une source lumineuse et un capteur de part et d'autre d'un tube transparent (une fibre optique pour les modèles les plus récents). Suivant le degré de flexion, le tube est plus ou moins écrasé ce qui conduit à un changement du niveau de lumière capté en sortie. Cette technologie a été proposée dans les années 1970 pour la mise en oeuvre d'un gant instrumenté : le *MIT-LED glove* (Sturman & Zeltzer 1994). Depuis, d'autres modèles de gants utilisent cette technologie comme le *5DT Glove* commercialisé par Fifth Dimension Technologies.
  — les capteurs résistifs reposent sur l'usage d'une bande flexible faite dans un matériau dont la résistance électrique varie en fonction de la flexion. Ce type de capteur est plus simple à mettre en oeuvre que les modèles optiques et peut même être directement intégré à un circuit imprimé. Les gants *CyberGlove II et III*, considérés parmi les meilleurs du marché (Kessler *et al.* 1995), utilisent cette technologie.

— Finalement, les capteurs **électromyographiques de surface** (sEMG) permettent la captation et l'analyse des signaux électriques qui sont envoyés aux muscles. Les signaux sont captés par des électrodes placées sur la peau, soit en nombre réduit mais de manière précise afin de s'intéresser à un muscle ou à un groupe musculaire en particulier, soit en nombre plus important avec des matrices d'électrodes qui renseignent sur l'ensemble des groupes musculaires d'une zone. Dans le premier cas, adapté aux études bio-mécaniques, le positionnement des électrodes est crucial ; plus complexe à mettre en oeuvre, plus invasif, mais les données captées sont simples à exploiter. Dans le second cas, d'avantage adapté aux IHM, la mise en oeuvre est rapide, simple et le matériel est moins intrusif ; l'exploitation des données est cependant plus complexe et reste encore mal maîtrisée à ce jour. L'électromyographie de surface est donc une technologie intéressante mais qui nécessite encore quelques années avant d'être utilisable.

5.1.4.2 Choix du capteur : un gant accélérométrique

Le cas d'usage et d'application considéré dans ces travaux est la commande d'un robot mobile par un soldat sur le champs de bataille. Ce contexte conduit assez naturellement à écarter les technologies de captation qui imposent le contrôle d'un environnement statique : les caméras fixes et les capteurs magnétiques.

L'usage de tout capteur émettant des infra-rouges est également proscrit puisque cela trahi immédiatement la position en cas d'utilisation d'équipements de vision thermique ; ce qui est raisonnablement le cas sur un champs de bataille. Les capteurs à lumière structurée

(a) (b) (c) (d)

FIGURE 5.6 – Différents modèles de gants instrumentés utilisant différentes technologies : (a) Le modèle *5DT-Glove* commercialisé par Fifth Dimension Technologies utilise des capteurs de flexions optiques ; (b) et (c) les modèles *Cyberglove II et III* utilisent la technologie des capteurs de flexion résistifs ; et (d) le modèle *IGS-Cobra Glove* de la société Synertial utilise la technologie des centrales inertielles .

et à temps de vol sont donc également exclus.

Un système de caméra 2D ou 3D sans émission sont envisageables, cependant la nature visuelle complexe de l'environnement conduit à des traitements d'image difficiles. De plus, le positionnement dynamique de la caméra est également problématique. Il est envisageable d'utiliser directement la caméra du robot ; mais cela implique qu'une commande n'est possible qu'à courte distance et sans obstacle. Or, un objectif des robots mobiles est d'acquérir de l'information à distance et en dehors du champs de vision. L'usage des approches basées vision sont donc écartées.

Ainsi, des différentes technologies disponibles, seuls les capteurs inertiels et de flexion semblent utilisables. Finalement, parmi les principaux modèles de gants commercialement disponibles et présentés Figure 5.6, nous avons fait le choix d'un gant *IGS-Cobra Glove* de la société Synertial. Ce modèle repose sur la technologie inertielle et se distingue particulièrement par son intégration de bonne qualité.

5.1.4.3 Caractéristiques techniques et traitements

Le modèle de gant retenu comporte 12 centrales inertielles (Figure 5.7(a)) maintenues à l'intérieur d'un gant textile double couche (Figure 5.7(b)). Le gant est relié par un câble à une batterie et à un transmetteur Wifi. Tous deux pouvant être positionnés de différentes manières, nous avons fait le choix de les placer à la ceinture afin qu'ils limitent le moins possible les mouvements.

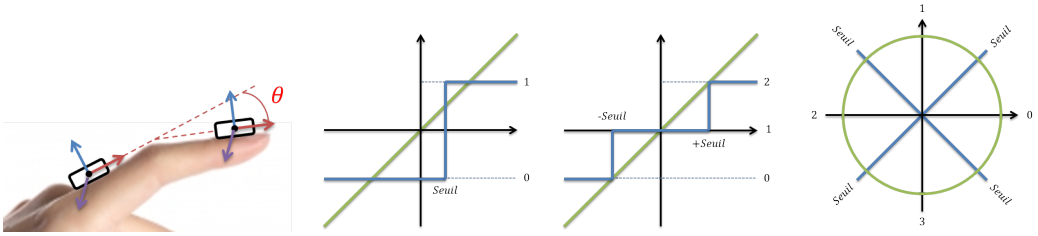
Les capteurs inertiels sont positionnés sur la main conformément à la Figure 5.7(c) : 1 capteur sur le dos de la main et 2 capteurs par doigt sauf pour le pouce qui en a 3. Ces capteurs acquièrent 9 informations à une fréquence de 60Hz : 3 accélérations linaires, 3 accélérations angulaires et 3 orientations absolues dans le référentiel terrestre.

Les informations fournies par les différents capteurs sont finalement traitées pour être mises en forme conformément à l'alphabet précédemment défini.

— **Configuration** : On estime pour chaque doigt si il est plié ou tendu. Pour cela, et

(a) (b) (c)

FIGURE 5.7 – Nous avons fait le choix d'un gant *IGS-Cobra Glove*. Il comporte 12 (a) centrales inertielles de petite taille, intégrées dans (b) un gant textile double couche et positionnées comme en (c) .



(a) (b) (c) (d)

FIGURE 5.8 – (a) angle de courbure d'un doigt et les trois types de seuillage utilisés : (b) à un niveau, (c) à deux niveaux et, (d) cyclique à quatre niveaux .

comme illustré en Figure 5.8(a), on calcule l'angle de rotation pour le couple de central positionné sur chaque doigt. Puis, on applique un seuillage à un niveau comme illustré en Figure 5.8(b).

— **Orientations** : En utilisant l'orientation estimée par le capteur positionné sur le dos de la main, on conserve les informations de roulis et de tangage. Pour chacun de ces deux angles, on applique un seuillage cyclique à quatre niveaux comme illustré en Figure 5.8(d).

— **Accélérations** : En utilisant le triplet d'accélération estimé par le capteur positionné sur le dos de la main, on annule la composante additive due à la gravité. Puis, pour chaque dimension, on applique un seuillage à deux niveaux comme illustré en Figure 5.8(c).

Finalement, les trois types d'informations labélisées sont concaténées pour construire les symboles atomiques de notre alphabet gestuel. Ils sont ensuite stoqués temporairement pour être fournis, selon le modèle de fenêtrage glissant, aux différents automates qui en font l'analyse.

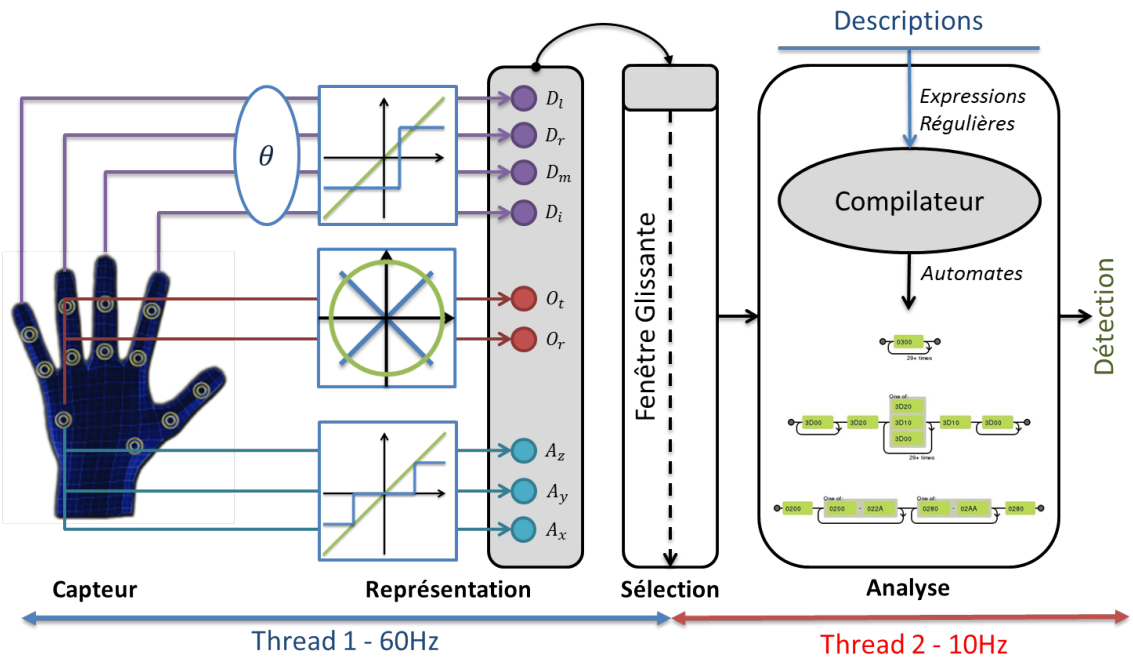


FIGURE 5.9 – Synthèse de la brique de reconnaissance spécifiée et implémentée. Elle a de particulier de reposer sur l'usage d'automates construits automatiquement par compilation de descriptions formelles spécifiées explicitement par un expert sous la forme d'expressions régulières.

5.1.5 Synthèse de l'architecture

En résumé, un système complet de reconnaissance de gestes a été spécifié et implémenté. Ce système reprend l'architecture classique de tout système de reconnaissance gestuel : un capteur, une représentation des données, un mode sélection et une analyse.

L'architecture de ce système de reconnaissance, résumée en Figure 5.9, comprend donc :

— un gant instrumenté utilisant un ensemble de centrales inertielles ;
— les données fournies sont seuillées et concaténées pour être représentées sous forme de symboles atomiques d'un alphabet gestuel spécifique.
— Ces symboles sont ensuite stockés temporairement et fournis selon un mode de sélection de type fenêtrage glissant à un banc d'automates.
— Ces automates sont construits automatiquement en compilant les expressions régulières écrites par un opérateur humain. Chaque automate réalise alors, en parallèle, et à fréquence régulière, une recherche dans les données fournies. Lorsqu'un automate trouve une séquence de données valide au regard de la grammaire régulière qu'il représente, il génère alors un évènement de détection.

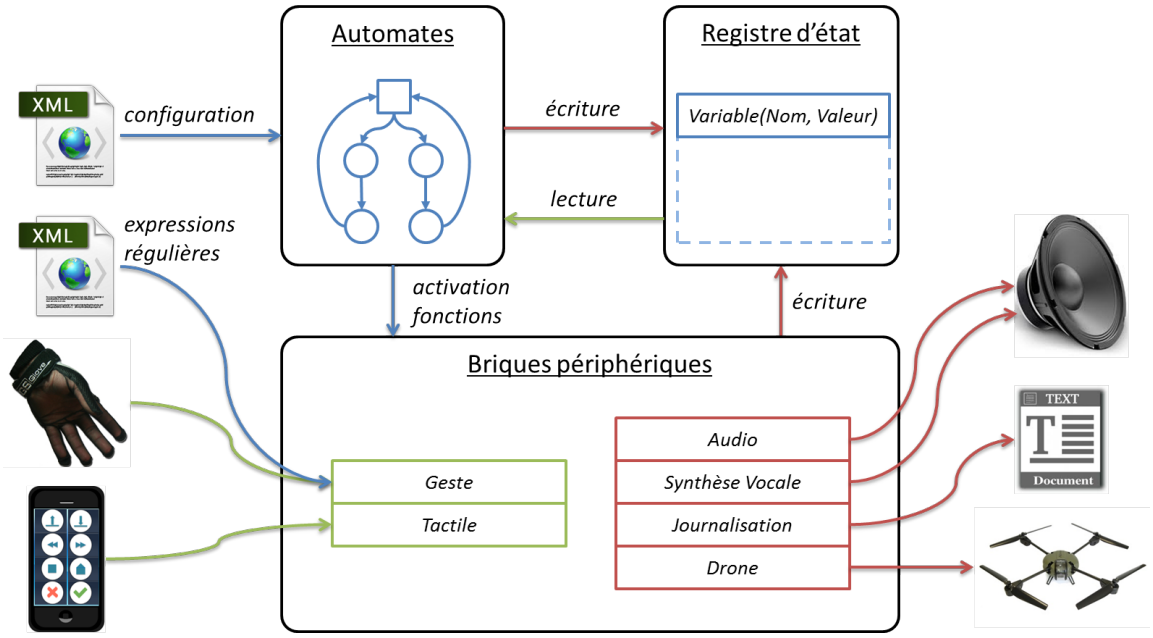


FIGURE 5.10 – Architecture de la plate-forme interactive implémentée.

5.2 Plate-forme interactive

Nous disposons donc d'une brique de reconnaissance gestuelle qui génère des évènements lors de la détection de la réalisation de gestes sémaphoriques spécifiques. Néanmoins, le développement ne s'arrête pas là. En effet, il nous semble que cette brique ne se suffit pas et qu'il est nécessaire de l'intégrer au sein d'une architecture plus large. Comme proposé dans le modèle d'interaction, cette architecture doit permettre la génération de feedbacks, la sécurisation des commandes et l'activation et le suivi des fonctions d'une plate-forme robotique.

5.2.1 Architecture

Comme illustré en Figure 5.10, nous avons fait le choix d'une architecture comportant trois éléments : un *registre d'état*, des *briques périphériques* et un ensemble d'*automates déterministes*.

Le registre d'état est un ensemble de variables nommées, accessibles à la fois par les briques périphériques et les automates. De nouvelles variables d'état peuvent être crées, modifiées et détruites dynamiquement en fonction du besoin.

Les briques périphériques sont différents composants logiciels qui permettent de traiter les éléments de l'interaction. Chaque brique assure un ensemble de fonctions activables ou automatiques et signale l'évolution de son état dans le registre. On peut distinguer deux types de briques périphériques : les briques d'entrées et les briques de sortie. Parmi les briques périphérique d'entrée, on trouve la brique de reconnaissance gestuelle et une brique *interface tactile*. Pour les briques périphériques de sortie, on trouve deux briques de feedback

sonore : une pour la lecture de fichiers audios et une pour la synthèse vocale ; une brique *drone* qui réalise l'interface avec différents types de drones et une brique *journalisation* qui permet d'inscrire dans un fichier différents évènements horodatés.

Les automates sont un ensemble de programmes qui déclenchent les actions des briques périphériques de manière contextuelle en fonction de l'état du registre. Ces automates sont simplement composés d'états et de transitions logiques chargées depuis des fichiers. Par leur intermédiaire, il est possible de définir différents comportements et ainsi de spécifier le modèle d'interaction à mettre en oeuvre.

5.2.2 Briques périphériques

Brique reconnaissance gestuelle

La brique de reconnaissance gestuelle consiste à détecter l'exécution de gestes décrits par des expression régulières. Cette brique comporte les différents traitements présentés précédemment. Les expressions régulières sont éditées ou chargées depuis un fichier et à chacune est associé un nom. A la création ou au chargement d'une nouvelle expression régulière, une variable booléenne et portant le même nom est crée dans la base d'état. Lorsque la réalisation d'un geste est détectée, la variable qui lui correspond est mise à jour. Enfin, une variable supplémentaire indique l'état actuel de la brique : (0) le gant n'est pas détecté, (1) le gant et détecté mais non connecté et (2) le gant est connecté et prêt pour la reconnaissance.

Brique interface tactile

La brique interface tactile consiste à connecter, via le réseau Tcp-Ip, l'interface d'un smartphone. Celle si, implémentée en C# avec le moteur Unity3D, présente sur l'écran différents boutons ; et lors de l'activation de l'un d'eux, un message est envoyé à la brique périphérique qui met à jour une variable du registre. A nouveau, une variable spécifique indique si l'interface tactile est (0) non détectée, (1) détectée mais non connectée et (2) connectée.

L'objectif de cette brique est de mettre en oeuvre une modalité d'entrée complémentaire et compatible de la modalité gestuelle afin de pouvoir les comparer. Ainsi, l'interface tactile est composée de huit boutons (un par geste) ; chacun identifiables par une icône.

Brique fichiers audio

La brique périphérique de lecture audio permet de lancer et d'interrompre la lecture d'un fichier MP3 désigné par son nom. Une variable spécifique indique alors l'état de cette brique : (0) brique disponible ou (1) lecture en cours. L'objectif de cette brique est de mettre en oeuvre des feedbacks audios iconiques simples : par exemple des "bip" dont la tonalité révèle un caractère positif ou négatif.

Brique synthèse vocale

La brique de synthèse vocale assure la verbalisation de phrases écrites en langage naturel fournies sous forme de chaînes de caractères. Cette fonctions est réalisée simplement en utilisant l'API native de Windows (Microsoft Speech API : SAPI). Une variable spécifique indique l'état de cette brique : (0) brique disponible ou (1) synthèse en cours.

Brique journalisation

La brique journalisation permet l'ouverture, la fermeture et l'écriture de données horodatées dans un fichier. Cette brique n'a pas directement pour objectif l'interaction mais la sauvegarde des évènements afin de pouvoir en réaliser une analyse, au besoin et à posteriori.

Brique drone

Finalement, la dernière brique périphérique est la brique drone. Celle-ci permet de se connecter, de commander et de suivre la progression d'un drone sur un parcour prédéfini.

Six fonctions ont donc été adressées : décoller, atterrir, avancer sur le parcour, reculer, aller à la base et interrompre un déplacement. Une variable renseigne sur l'état du drone : (0) non détecté, (1) détecté mais non connecté, (2) au sol, (3) en cours de décollage, (4) d'atterrissage, (5) de déplacement ou (6) à l'arrêt en vol (maintien de la position).

Dans le cadre de ces travaux, deux types de drones ont été adressés :

— Un simulateur de comportement : la réalisation des fonctions est simulée seulement par des temporisations. Les fonctions de décollage et d'atterrissage ont une durée fixe et les fonctions de déplacement on une durée calculée en fonction de la distance théorique à parcourir et d'une vitesse moyenne fixée.

— Un simulateur 3D : le drone et différents environnements de vol sont simulés avec le moteur Unity3D. Les déplacements sont simulés par de simples translations et non de manière physico-réaliste. La Figure 5.11 présente deux illustration de ce simulateur.

La connexion à un drone réel a également été entreprise mais n'a pu être achevée. En effet, l'usage d'un drone physique impose aujourd'hui la prise en compte d'un certain nombre de contraintes. L'obligation de la présence d'un pilote de sécurité disposant d'une commande prioritaire ajoute notamment de la complexité et des problématiques de non-déterminisme dans le déroulement des commandes.

5.2.3 Configuration pour le modèle d'interaction défini

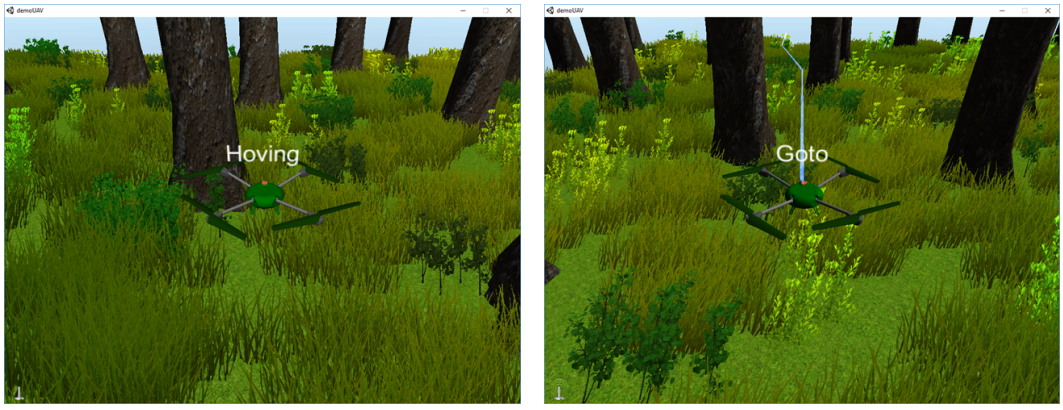
Finalement, différents automates ont été configurés pour mettre en oeuvre le comportement prévu par le modèle d'interaction défini et présenté au Chapitre 2.

Ainsi, pour chaque commande, le protocole suivant a été implémenté :

— Le système attend l'exécution d'un geste de commande.

— Si cette commande est incompatible avec le contexte actuel (par exemple atterrissage alors que le drone est au sol), un bip sonore spécifique indique que la commande a bien été prise en compte mais qu'elle ne peut pas être exécutée.

— Si cette commande est compatible avec le contexte actuel, un message vocal explicite la commande détectée et demande une confirmation pour l'activation de celle-ci.



(a) (b)

FIGURE 5.11 – Deux vues du simulateur 3D de drone réalisé avec le moteur Unity3D ; le drone (a) à l'arrêt en vol et (b) en cours de déplacement vers un point de passage .

Par exemple : "décollage demandé, veuillez confirmer".

— Si le geste de validation est détecté dans les secondes qui suivent, la commande est activée, un bip sonore spécifique confirme la prise en compte de cette consigne et le système attend une nouvelle commande.
— Si le geste d'annulation est détecté ou qu'une durée de 5 seconde s'est écoulée sans que la geste de validation ne soit détecté, un bip sonore spécifique signale que la commande n'a pas été prise en compte et le système attend une nouvelle commande.

Parallèlement à ce protocole, différents messages sont générés pour signaler à l'utilisateur la fin de la réalisation d'une commande : par exemple "décollage terminé".

Des messages d'alerte en langage naturel ont également été ajoutés pour signaler d'éventuels problèmes relatifs à la connexion du gant, de l'interface tactile et du drone.

5.3 Synthèse et discussion

Dans ce chapitre, nous avons donc présenté une brique de reconnaissance gestuelle basée sur le principe de la description formelle des gestes à reconnaître. Cette brique est ensuite intégrée au sein d'un système qui permet la mise en oeuvre d'un modèle d'interaction qui ne se limite pas à la simple activation d'une commande en réaction à un geste mais, avec un ensemble de feedback sonores, sécurise les commandes et renseigne sur l'état du drone.

5.3.1 Description formelle et expression régulières

Nous avons donc fait un choix différents des méthodes habituelles pour la modélisation des gestes. Ainsi, plutôt que de s'appuyer sur des méthodes d'apprentissage automatique d'exemples, nous exploitons la connaissance a-priori des gestes pour expliciter à la machine ce qu'elle doit reconnaître.

Cette méthode semble particulièrement pertinente mais requiert cependant une forme d'expertise sur la manière de décrire un geste et d'exprimer cette description en expression régulière. Ce qui semble alors fondamental, c'est d'une part, la connaissance du modèle anatomique générique des gestes ainsi que la distinction entre gestes statiques et dynamiques ; et d'autre part, la recherche de la limite sémantique de chaque geste.

Il est en effet particulièrement important de se demander ce que *n'est pas un geste* : à partir de quel changement dans les paramètres a-t-on un autre geste ? Par exemple, la configuration de la main est-elle importante dans les gestes d'atterrissage et de décollage : utiliser un poing fermé au lieu de la main à plat représente-il toujours la même commande ? Un autre exemple est la nature du mouvement dans le geste pour faire progresser le drone au point suivant : est-ce qu'une translation couplée à une rotation est équivalente sémantiquement à une rotation pure ?

Il est donc fondamental pour chaque élément d'un geste (configuration, orientation, mouvement), de se demander ce qui est sémantiquement acceptable, et sur cette base, d'écrire l'expression régulière qui correspond. Dans notre cas, ces questions de limite des gestes avaient été posées aux participants qui les ont proposés lors de la phase d'élicitation des gestes (étape 1) du protocole de construction du vocabulaire gestuel. Cela avait alors pour objectif de permettre le regroupement des propositions de gestes identiques en gestes candidats. La question était alors : à partir de quant deux propositions de gestes peuvent-ils être regroupés ?

Il est donc intéressant de noter que dans le cas présent, ce sont les informations renseignées par les *créateurs* des gestes eux-mêmes qui ont servit à l'écriture de leur description formelle en expression régulière. Le protocole de choix des gestes, tel que nous l'avons présenté et appliqué semble donc d'autant plus pertinent lorsque l'on souhaite réaliser une description des gestes et non un apprentissage automatique.

Finalement, la syntaxe des expressions régulière est intéressante car elle permet à la fois l'énumération : ce qui correspond à la logique utilisée par la DTW ; et la modélisation : ce qui correspond à logique utilisée par les HMM. Néanmoins, utilisée directement, cette syntaxe spécifique nécessite une certaine habitude.

Ainsi, le reproche principal qui peut être adressé à la description formelle sur la base des expressions régulières, semble être la complexité de mise en oeuvre imposant un degré d'expertise. Il pourrait néanmoins être intéressant, pour répondre à cette problématique, de proposer une procédure guidée utilisant des formulaires à choix multiples et en langage naturel. Les réponses fournies pourraient alors être utilisées pour générer automatiquement l'expression régulière qui correspond. Cette question n'a cependant pas été traitée au cours des travaux de la thèse et reste ouverte pour des développements futurs.

5.3.2 Alphabet gestuel utilisé

Il a également été fait le choix, de construire un alphabet de symboles atomiques en concaténant des informations simples de types accélération, orientation et configuration. Ce choix est ici important car il permet de conserver un langage de description relativement représentatif des données d'origine ; ce qui permet, de manière plus immédiate, la réflexion

et l'écriture des expressions régulières

A posteriori, il semble qu'il serait plus avantageux de travailler avec des vitesses orientées au lieux d'accélérations. Cependant, cela n'est pas aisément faisable avec les capteurs choisis : en effet, cela implique d'intégrer les accélérations pour obtenir des vitesses ; ce qui conduit à des données bruitées difficilement exploitables.

Ainsi, nous avons conservé l'utilisation des accélérations, mais dans un contexte différent : avec des capteurs de position et d'orientation absolues, le choix de travailler avec des vitesses semblerait plus pertinent.

5.3.3 Performances

L'architecture proposée a donc été implémentée. Cela a été réalisé en *C++* avec la plate-forme *Qt* et, pour les expressions régulières, la classe générique *QRegExp* a été utilisée.

Une fois l'implantation terminée, la question des performances a alors été posée : le système est-il fiable ? De manière usuelle, pour évaluer les performances d'un système de reconnaissance, on lui fourni un ensemble d'exemples dont on connaît la nature, et on comptabilise les réponses suivant le type : exemple correctement détecté et classifié, exemple non détecté et exemple détecté mais mal classifié.

Sur ce principe, nous avons donc réalisé un test : nous avons demandé à dix participants de réaliser 5 fois chacun des gestes. Pour chaque participant, les gestes à réaliser étaient imposés dans un ordre aléatoire par le système via une interface simple : pour chaque nouvelle acquisition, les participants avaient 5 secondes pour réaliser le geste désigné par son nom et une image.

Au final, en retirant les exemples qui ne correspondent pas exactement la description théorique de leur geste, le système mis en oeuvre présente un taux de reconnaissance et de détection de 100%. En effet, tous les gestes réalisés et respectant la description sémantique ont correctement été reconnus.

La question se pose alors, pour les exécutions qui ont été exclues : cette exclusion est-elle légitime ? Habituellement, pour les systèmes reposant sur un apprentissage automatique, on considère que l'utilisateur ne se trompe pas et, si une erreur de détection survient, c'est alors que la machine a mal appris. En effet, avec une telle approche, on ne dispose pas d'une description formelle de ce qu'est précisément le geste et donc de ce qu'est *se tromper*. Avec notre approche en revanche, il est possible de décider si un geste correspond ou non aux règles qui le définissent. Un exemple qui a été rencontré, est la réalisation du geste atterrir avec une orientation de la main incorrecte au début ou à la fin : main en biais et non parallèle au sol. Deux cas semblent alors envisageables :

1. La personne n'a pas compris le geste ou a mal articulé et l'a donc réalisé de manière impropre. Ici, la machine ne semble aucunement impliquée et la question se pose d'avantage en termes de comment le geste a été présenté et de la nécessité d'une formation plus complète de l'utilisateur.
2. La réalisation sort du cadre des règles définies, mais la personne estime que cela

correspond tout de même au geste demandé. Il peut alors être pertinent de s'interroger sur la définition du geste ce qui conduit à modifier l'expression régulière qui lui correspond.

Pour l'ensemble des cas qui ont été rencontrés, comme celui du geste d'atterrissage précédemment décrit, il a été possible d'expliquer en quoi la réalisation sortait du cadre des règles fixées ; et pour chacun, il a été raisonnablement décidé qu'il s'agissait d'une erreur d'exécution et non d'une règle impropre ; et c'est au final, avec ce raisonnement, que nous estimons que le système mis en oeuvre présente des performances optimales pour le vocabulaire considéré.

CHAPITRE 6

Évaluation de l'utilisabilité du système en situation écologique

Sommaire

6.1 Hypothèse, utilisabilité et test utilisateur

6.1.1 Hypothèse

Nous avons donc proposé un modèle d'interaction (Chapitre 2), construit un dictionnaire gestuel (Chapitre 3) et montré la pertinence du geste comme modalité de commande (Chapitre 4). Un système technique a également été développé (Chapitre 5). Celui-ci assure la reconnaissance de gestes spécifiques en temps réel, l'émission de feedbacks adaptés et gère le comportement d'un drone virtuel.

Ces différents travaux ont été conduits avec pour objectif de proposer le geste comme alternative aux interfaces classiques : tactile ou joystick, pour commander un système en situation de mobilité et en environnement complexe, voir hostile. Notre hypothèse majeure est que dans un tel contexte, le geste libère l'utilisateur d'un certain nombre de contraintes : visuelles, physiques voire cognitives, le rendant alors globalement plus disponible pour traiter son environnement.

Dans ce chapitre, nous étudions donc la validité de cette hypothèse en se plaçant dans le contexte de la commande d'un drone militaire de contact et avec le modèle d'interaction proposé, le vocabulaire construit et le système technique développé.

6.1.2 Utilisabilité

Pour évaluer la validité de notre hypothèse, nous avons cherché à évaluer l'*utilisabilité* de la commande au geste : L'utilisabilité est un concept qui cherche à décrire la qualité d'un système vis-à-vis de son utilisateur. En remplaçant le concept plus ancien de *user-friendly* qui énonçait qu'un système est bon lorsqu'il ne cause aucune frustration à son utilisateur, l'utilisabilité propose une définition plus précise qui permet d'objectiver sa compréhension et son évaluation.

Pour définir l'utilisabilité, il convient de se référer aux deux définitions les plus connues :

> usability is about learnability, efficiency, memorability, errors, and satisfaction
>
> (Nielsen 1994)

> the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
>
> (ISO 2000)

Ainsi, l'utilisabilité est le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec *efficacité*, *efficience* et *satisfaction*, dans un contexte d'utilisation précis.

L'efficacité est alors définie comme étant la la capacité d'un produit à permettre à son utilisateur d'atteindre le résultat prévu, l'efficience est la minimisation de l'effort à fournir par l'utilisateur pour atteindre son but le plus rapidement possible, et la satisfaction est l'évaluation subjective du confort et de l'interaction pour l'utilisateur.

Finalement, au regard de cette définition, l'utilisabilité est la caractéristique d'un système à être adapté à l'environnement et surtout à l'utilisateur lui-même ; et c'est particulièrement ce qui nous intéresse dans le cadre de notre étude du geste. Mais à quoi correspond l'utilisabilité dans le contexte militaire de la commande d'un drone de contact ?

Tout d'abord, l'efficacité est ici la capacité de l'opérateur à commander le drone ; est-il en mesure d'activer une commande lorsqu'il en a besoin ? La mémorisation des gestes, et la robustesse du système technique sont ici les deux éléments les plus impactants. En effet, si l'utilisateur oublie quel geste utiliser, ou si le système échoue à reconnaître un geste fait, il devient alors impossible de commander le drone.

Ensuite, l'efficience est le temps et la quantité de ressources (physiques ou cognitives) que l'opérateur utilise pour passer une commande. Dans le cadre de cette étude, il s'agit du critère le plus important. En effet, les ressources de l'opérateur sont partagées entre la commande du drone et l'accomplissement de la mission dont la progression, le maintien de l'arme et le contrôle de l'environnement sont les actes élémentaires. Or, l'objectif est de doter le fantassin d'un outil : un drone pour l'aider dans l'accomplissement d'une mission et non de dédier le fantassin à la commande du drone. Ainsi, il s'agit d'une situation de double tâche où la priorité sera toujours donnée à la mission et c'est pour cela qu'il est fondamental de minimiser les ressources nécessaires à l'usage du drone ; donc, de maximiser son efficience. Le modèle d'interaction impacte l'utilisateur d'un point de vue cognitif : sa compréhension de la situation et son choix des actions à réaliser. La qualité du vocabulaire gestuel impacte quant à lui l'opérateur de manière globale : la complexité des gestes, que ce soit leur durée, difficulté de réalisation ou de mémorisation ont un impact à la fois physique et cognitif. Enfin, la robustesse du système et sa latence vont également impacter l'interaction : un geste mal ou trop tardivement reconnu impose de répéter son exécution en y portant une attention supplémentaire.

Finalement, la satisfaction prend également en compte le modèle d'interaction, les gestes et la robustesse du système mais de manière indirecte, surtout par comparaison avec les moyens existants ou idéalisés. D'autres éléments interviennent également comme l'acceptation sociale et la résistance au changement : Comment un militaire se voit-il utiliser le geste sur le terrain, vis-à-vis de ses équipiers et de sa hiérarchie ? En occident, le geste ayant souvent une connotation péjorative (gesticuler) en s'opposant au calme et à la maîtrise de soit, l'acceptabilité de son usage n'est pas évidente.

6.1.3 Test utilisateur

On souhaite donc évaluer la commande au geste avec notre système technique. Pour ce faire deux stratégies existent (Dumas & Redish 1999) : l'*évaluation heuristique* qui, comme utilisée au chapitre 2, consiste à contrôler le respect d'un ensemble de règles pratiques, et le *test utilisateur* qui consiste à confronter un groupe de participants au système dans un contexte écologique pour en observer le comportement et déterminer les éléments du systèmes qui ne sont pas adaptés.

L'évaluation heuristique est une solution basée sur l'application de modèles ou de règles théoriques. Elle est rapide et relativement peut coûteuse. Un test utilisateur, quant à lui, repose sur des mesures objectives et subjectives. Cette seconde méthode, bien que plus coûteuse et complexe à mettre en oeuvre du fait de l'implication de participants et de matériel, peut néanmoins conduire à une évaluation plus précise et plus complète.

Puisqu'à ce jour, il n'existe à notre connaissance, aucun retour d'expérience sur l'utilisabilité de la commande au geste en situation de mobilité,aussi nous avons choisi de tester nos hypothèses en réalisant un test utilisateur.

6.2 Protocole

6.2.1 Plan du protocole

On cherche donc à évaluer l'utilisabilité globale de notre système. Pour ce faire, des participants militaires commandent un drone virtuel au cours de différents scénarios imposés tout en cherchant à se déplacer, à conserver les deux mains sur une arme qu'ils portent et à rester attentifs à l'environnement. Le système est testé dans quatre conditions et des mesures objectives et subjectives sont alors recueillies.

6.2.1.1 Conditions étudiées

D'une part, le système est utilisé soit avec la reconnaissance gestuelle, soit avec une interface tactile. Les deux configurations sont en tout point identiques hormis la modalité de commande. L'objectif, en utilisant également la modalité tactile, est de pouvoir comparer les résultats et tirer un bilan objectif quant à la commande gestuelle.

D'autre part, le système est testé dans deux environnements que nous appelons *réel* et *virtuel*. Dans l'environnement réel, les participants se déplacent sur un chemin balisé en environnement urbain extérieur. Dans l'environnement virtuel, les participants sont immergés dans une scène 3D projetée sur un écran et progressent via une métaphore de *marche sur place* (Slater *et al.* 1995). Celle-ci converti le piétinement des participants sur un tapis de détection en déplacement dans la scène virtuelle.

6.2.1.2 Mesures

Pour chaque condition, trois mesures sont réalisées :

La première mesure est la durée cumulée de l'interaction. Il s'agit de mesurer le temps durant lequel le participant n'a pas les deux mains sur son arme et n'est donc pas en mesure d'en faire usage immédiatement. Ainsi, nous avons choisi cette mesure qui nous semble représentative de la disponibilité des mains de l'utilisateur.

La seconde mesure est le nombre de pas réalisés au cours d'un scénario. Il s'agit de mesurer la capacité à se déplacer : à progresser sur un itinéraire. Nous avons choisi cette mesure qui nous semble représentative de l'impact de la commande sur la capacité à se déplacer.

Enfin, la troisième mesure est un score de satisfaction globale du participant au regard d'une modalité et d'une session. Cette mesure est subjective et réalisée à l'aide d'un questionnaire papier.

6.2.1.3 Sessions

Chaque participant teste donc le système dans quatre conditions qui sont réparties en deux sessions :

- Session *Réel* :
  - Commande gestuelle en environnement réel.
  - Commande tactile en environnement réel.
- Session *Virtuel* :
  - Commande gestuelle en environnement virtuel.
  - Commande tactile en environnement virtuel.

Pour chaque participant, l'ordre des sessions et des modalités est choisi aléatoirement afin de compenser tout effet d'ordre. Ainsi, au cours d'une session, chaque participant utilise les deux modalités, geste et tactile, dans un même environnement. Une session comporte six phases, soit trois par modalité :

1. Préparation du participant : explication de la modalité, équipement, entraînement.
2. Quatre scénarios de commande du drone virtuel en situation de déplacement.
3. Renseignement d'un questionnaire papier.

La figure 6.1 présente un résumé des variables de cette étude. Les participants, le matériel et les trois phases pour une modalité d'une session vont maintenant être détaillés.

6.2.2 Participants

Dix participants ont pris part à cette étude. Tous étaient membres de la *Section Tactique de l'Armée de Terre* française (STAT), et plus précisément du groupe *Drone* du groupement *Renseignement* (RENS). Ce groupement a pour mission de vérifier l'adéquation au besoin militaire des nouveaux systèmes destinés aux forces aéro-terrestres et possède ainsi une parfaite connaissance à la fois des drones et du besoin opérationnel.

6.2.3 Matériel

Pour cette étude, deux systèmes techniques ont été mis en oeuvre :

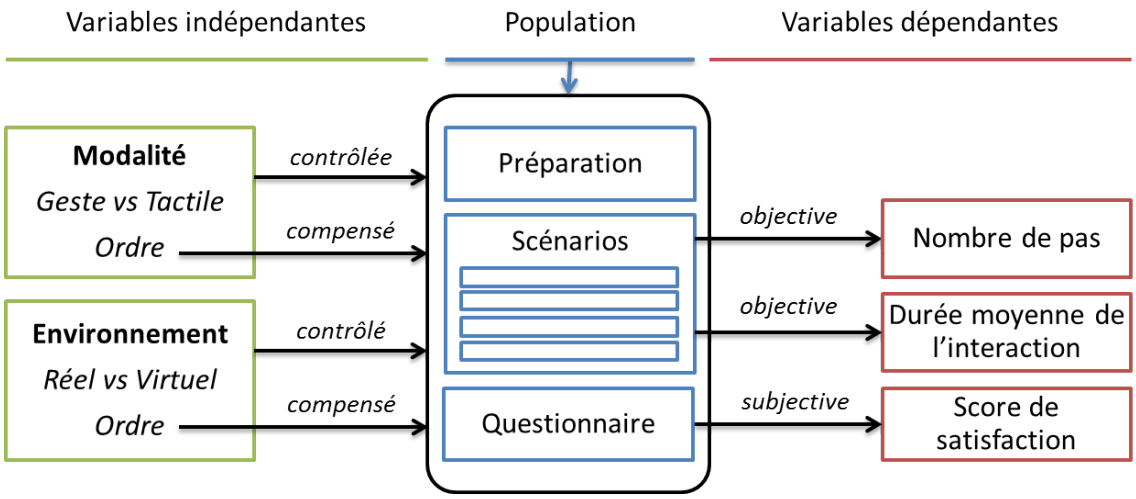


FIGURE 6.1 – Synthèse du protocole mis en oeuvre pour l'évaluation de l'utilisabilité de la commande au geste.

— Le premier, commun aux deux environnements, est le système interactif qui assure la détection d'évènements gestuels ou tactiles, simule le comportement d'un drone et génère les feedback associés.
— Le second, spécifique à l'environnement virtuel, permet l'affichage d'une scène 3D et la progression dans celle-ci à l'aide d'une métaphore de marche.

En plus de ces deux systèmes qui vont être détaillés, une réplique d'arme a été utilisée : le participant est équipé d'une arme de type Famas fictif.

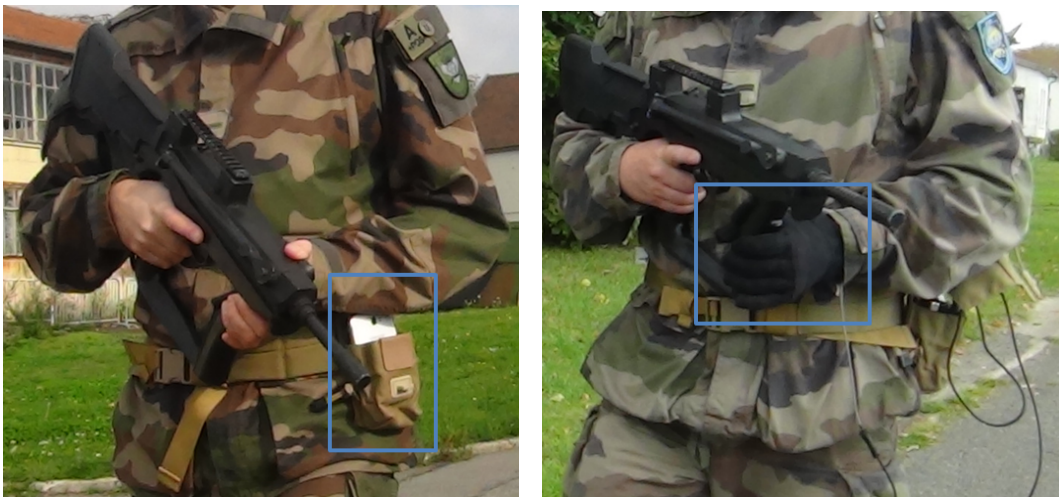
6.2.3.1 Matériel commun aux deux environnements

Le premier équipement, commun aux deux environnements est donc le système interactif. Sont rôle est d'assurer la captation du vocabulaire gestuel ou la détection d'évènements tactiles sur un smart-phone, de simuler le comportement d'un drone avec un plan de mission et d'émettre les feedbacks adéquats. Ce système comporte les cinq éléments matériels présentés en Figure 6.2 :

En modalité gestuelle, le participant est équipé d'un gant instrumenté de type *CyberGlovve IGS*, connecté à un émetteur-récepteur wifi et une batterie. Ce gant comporte douze centrales inertielles dont les valeurs sont lues et émises en temps réel. Le gant est porté à la main gauche du participant et l'émetteur et la batterie sont positionnés à sa ceinture dans une poche prévue à cet effet. Une illustration est proposée en Figure 6.3(b).

En modalité tactile, le participant est équipé d'un smart-phone. Le modèle utilisé est un Samsung Galaxy Note 3. Il dispose d'un écran de 5.7 pouces et pèse 168 grammes. Une interface logicielle a été développée pour ce téléphone. Celle-ci présente 8 boutons, chacun correspondant à une des commandes activables, et les évènements de ces boutons sont émis par wifi. Ce téléphone est placé dans une poche à la ceinture du participant, d'où il peut le sortir pour utilisation. Une illustration est proposée en Figure 6.3(a).

FIGURE 6.2 – Configuration matérielle commune aux deux environnements de l'étude.



(a) (b)

FIGURE 6.3 – (a) équipement de la modalité tactile et (b) équipement de la modalité gestuelle .

Quelle que soit la modalité, l'expérimentateur est équipé d'un ordinateur portable connecté à un routeur wifi et un haut-parleur portable. Le routeur assure la création d'un réseau local permettant la communication entre l'ordinateur, le gant et le téléphone. Le haut-parleur permet quant à lui d'amplifier les feedbacks sonores afin que ceux-ci soient clairement audibles à la fois du participant et de l'expérimentateur.

Sur l'ordinateur, la plate-forme interactive présentée au chapitre 5.2 assure différentes fonctions. Elle réalise la reconnaissance gestuelle et la réception des évènements tactiles.

Elle simule également le comportement d'un drone avec un plan de mission. Différents itinéraires composés de points de passages peuvent être édités et les six fonctions élémentaires du drones sont simulés : le décollage, l'atterrissage, la progression au point de passage suivant sur l'itinéraire actif, le retour au point de passage précédent ou à la base et l'arrêt en vol du déplacement du drone. La simulation est ici purement temporelle : l'activation d'une commande génère un évènement au bout d'une durée configurée et aucun feedback visuel n'est proposé. Ce type de simulation a été choisi car la situation écologique visée ne requiert pas que l'opérateur voit le drone.

Enfin, la plate-forme interactive comporte un automate configuré pour la gestion contextuelle des différents feedbacks sonores. En plus des feedbacks prévus par le modèle d'interaction, des messages ont été ajoutés pour indiquer au participant les commandes à activer pour chaque scénario. La configuration est la suivante :

Un message sonore indique au participant la commande à activer. Puis, lorsque le participant active cette commande (geste ou tactile), le nom de la commande est explicitée et une confirmation est demandée. Si le participant valide la commande (geste ou tactile), une confirmation est émise par un bip sonore et l'exécution de la commande est simulée. Lorsque la simulation est complétée (temporisation terminée), un message indique le résultat : décollage, atterrissage ou déplacement terminé. Enfin, un message contextuel est émis avant de passer à la commande suivante du scénario. Ce message contextuel a pour objectif de justifier la transition entre de commandes et ainsi d'augmenter le réalisme de la situation. Les messages possibles sont :

- **Début de la mission, faites décoller le drone** : Au début d'un scénario.
- **Décollage terminé, faites avancer le drone**.
- **Déplacement terminé. Aucune menace détectée, faites avancer le drone**.
- **Déplacement terminé. Menace suspectée, faites reculer le drone** : On demande au participant de mettre le drone en sécurité le temps qu'un analyste vidéo (opérateur militaire imaginaire) d'analyser les données que le drone a recueillies.
- **Déplacement terminé. Menace non confirmée, faites avancer le drone**.
- **Déplacement terminé. Menace confirmée, faites rentrer le drone à la base** : La menace ayant été confirmée le drone doit être renvoyé à son point de ralliement (programmé dans l'itinéraire) pour être récupéré.
- **Déplacement terminé. Dernier point atteint, faites atterrir le drone**.
- **Déplacement à la base terminé, faites atterrir le drone**.
- **Atterrissage terminé, fin de la mission** : Indique la fin du scénario.

6.2.3.2 Matériel spécifique à l'environnement virtuel

Le second système, spécifique à l'environnement virtuel, permet l'affichage d'une scène 3D et la progression dans celle-ci à l'aide d'une métaphore de marche. Ce système est composé d'un écran monoscopique rétro-projeté et d'un tapis détectant la pression des pieds en neufs zones distinctes ; tous deux connectés à un ordinateur. Comme présenté en Figure 6.4, ces équipements sont installés dans une grande salle sans fenêtre permettant ainsi de contrôler la luminosité ambiante.

FIGURE 6.4 – Équipement spécifique de d'environnement virtuel : (1) écran vidéo-projeté affichant une scène virtuelle 3D, (2) caméra réalisant un *plan en pied* du participant et (3) un tapis comportant des zones de pression pour détecter le piétinement du participant.

Au cours des scénarios d'une séance en environnement virtuel, le participant est placé debout sur le tapis, à un mètre face à l'écran.

Une scène virtuelle représentant un environnement extérieur a été implémentée. Le participant observe cette scène en vue à la première personne et progresse grâce à une métaphore de marche sur place (Slater *et al.* 1995). Cette métaphore repose sur le principe suivant : à chaque piétinement du participant, la vue avance d'une distance correspondant à un pas dans l'environnement virtuel. L'itinéraire suivi et les orientations de la caméra sont pré-programmés ; le participant ne contrôle donc que la progression.

Pour implémenter cette métaphore, nous avons choisi d'utiliser un avatar virtuel doté d'une animation réaliste de la marche issue de la motion capture. Une illustration de l'avatar utilisé est proposée en Figure 6.5(b). Ainsi, sans qu'il le perçoive, le participant est représenté dans l'environnement virtuel par un avatar et perçoit la scène à travers ses yeux. Lorsque l'un des pieds du participant quitte le tapis, l'animation de marche de l'avatar est déclenchée et fait avancer celui-ci. Puisqu'il est placé dans un environnement à la même échelle que lui et qu'il est doté d'une animation de marche réaliste, l'usage de cet avatar permet de suivre automatiquement le relief et de ne pas avoir à calibrer le déplacement de la caméra.

La trajectoire de l'avatar et l'orientation de son regard suivent donc un itinéraire pré-programmé. Avec cette métaphore, le participant ne contrôle donc que la progression ce qui est relativement limité mais permet néanmoins un bon compromis entre simplicité de mise en oeuvre et réalisme.

Enfin, l'environnement comptabilise automatiquement le nombre de pas effectués par le participant au cours d'un scénario.

(a) (b)

FIGURE 6.5 – (a) Exemple de ce que voit le participant et (b) illustration de l'avatar du participant dans la scène virtuelle. Cet avatar avance d'un pas sur un itinéraire pré-programmé à chaque piétinement du participant sur le tapis.

6.2.4 Phase 1 : Préparation du participant

Pour chaque modalité d'une session, la première étape est d'équiper le participant du matériel de commande : soit le système reconnaissance de geste, soit la tablette tactile. L'utilisation d'un plan de mission ainsi que les six actions réalisées par le drone sont expliquées.

Ensuite, pour chaque action, la commande et les feedbacks associés sont présentés : pour l'interface tactile, le bouton est indiqué et pour l'interface gestuelle, le geste est démontré par l'expérimentateur. La démonstration est répétée autant de fois que le participant le souhaite.

Une fois toutes les commandes vues, le participant est invité à les pratiquer librement jusqu'à ce qu'il se sente capable de les utiliser de manière fluide. Finalement, il est demandé au participant d'activer chacune des 6 commandes, dans un ordre aléatoire, pour s'assurer de la bonne maîtrise de l'interface avant de commencer les scénarios.

6.2.5 Phase 2 : Scénarios

Une fois la préparation terminée, le participant complète quatre scénarios d'une durée de trois minutes environ. Les quatre scénarios sont réalisés à la suite mais dans un ordre différent pour chaque participant.

Un scénario consiste à activer des commandes suggérées par le système via des messages sonores tout en cherchant à avancer, conserver les deux mains sur l'arme et rester attentif à l'environnement (qu'il soit réel ou virtuel).

Un scénario comporte toujours sept commandes espacées du temps d'exécution par le drone virtuel : environ 20 secondes. Ce déroulement temporel a été choisi afin de rendre le participant relativement actif pendant la durée du scénario tout en lui laissant néanmoins le temps de progresser entre chaque action.

Un scénario commence et se termine toujours le drone au sol. Ainsi, les deux premières commandes sont toujours respectivement le décollage et la progression au point de passage

suivant ; et la dernière commande est toujours l'atterrissage. Les quatre commandes intermédiaires sont différentes pour chaque scénario et ne sont pas connues du participant. Pour ces quatre commandes, il peut être demandé au participant de faire avancer le drone, de le faire reculer ou de le faire rentrer à la base.

Parmi les différentes combinaisons respectant ces conditions, les quatre scénarios retenus sont les suivants :

1. Décollage → Suivant → Précédent → Suivant → Suivant → Suivant → Atterrissage
2. Décollage → Suivant → Suivant → Précédent → Suivant → Suivant → Atterrissage
3. Décollage → Suivant → Suivant → Suivant → Précédent → Suivant → Atterrissage
4. Décollage → Suivant → Précédent → Suivant → Précédent → Base → Atterrissage

Au cours de chaque scénario, le nombre de pas effectués par le participant est comptabilité : soit à vue par l'expérimentateur en environnement réel, soit par le tapis en environnement virtuel.

Le système technique enregistre chaque évènement dans un fichier et le participant est filmé pendant toute la durée des scénarios. En environnement virtuel un *plan pied* est réalisé par une caméra fixe placée devant lui. En environnement réel, un second expérimentateur précède le participant de quelques mètres pour le filmer.

Ces données sont ensuite analysées pour en extraire la durée cumulée pendant laquelle les deux mains ne sont pas sur l'arme. Pour ce faire, on utilise un outil d'annotation de vidéo.

Parmi différents outils proposés par la communauté scientifique (Dasiopoulou *et al.* 2011), nous avons fait le choix du logiciel *ELAN* (Lausberg & Sloetjes 2009) (Eudico Linguistic ANnotator) pour sa simplicité et la non divulgation des données traitées. Une vue de l'outil ELAN avec une vidéo traitée et les différentes pistes d'annotations est présentée en Figure 6.6.

Avec cet outil, pour chacune des 160 vidéos (10 participants, 4 conditions et 4 scénarios), on procède de la manière suivante : Dans un nouveau projet, on importe une vidéo et les évènements enregistrés par le système (demande de commande, commandes et feedbacks). Puis, on synchronise manuellement ces deux éléments. Ensuite, on crée une nouvelle piste d'annotation. Enfin, on annote la vidéo en renseignant chaque évènement de lâcher ou de reprise de l'arme par le participant. La durée cumulée de lâcher de l'arme est finalement obtenu avec l'outil de statistique d'annotation d'ELAN.

6.2.6 Phase 3 : Questionnaire

Enfin, après avoir terminé les quatre scénarios pour une même modalité et un même environnement, le participant est déséquipé. Puis, il est invité à renseigner un questionnaire papier dont l'objectif est de recueillir son niveau de satisfaction global vis-à-vis du système qui vient d'être utilisé.

Le questionnaire utilisé lors de cette étude est le SUS (Brooke *et al.* 1996) (*System Usability Scale*). Contrairement au questionnaire utilisé lors de l'étude portant sur l'attention

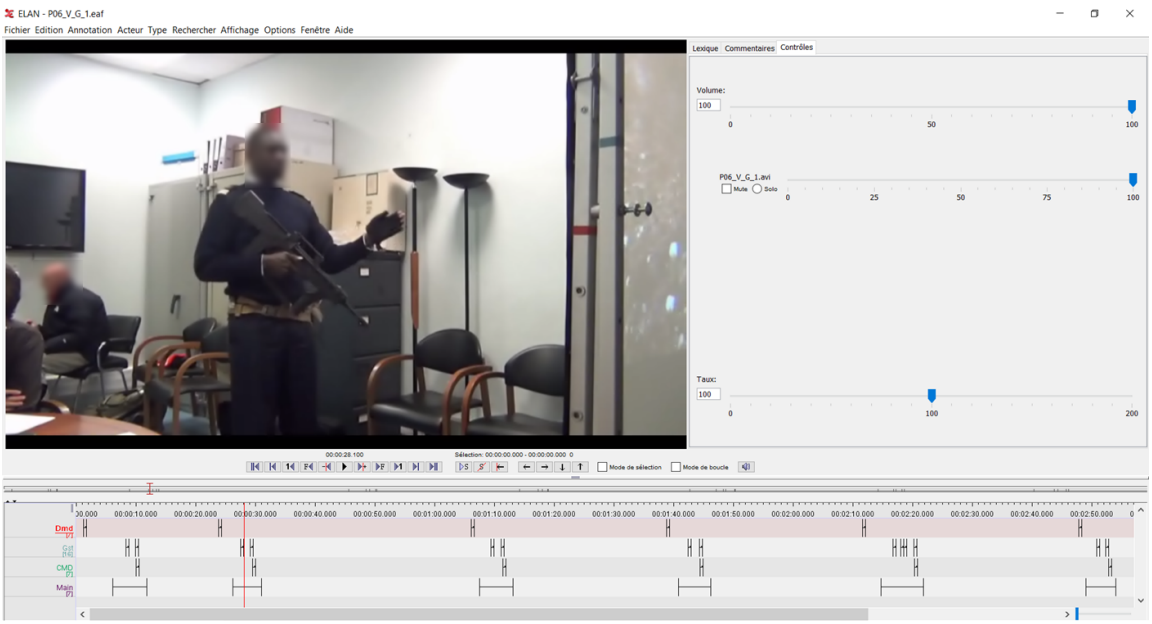


FIGURE 6.6 – Vue de l'outil d'annotation de vidéo ELAN présentant une vidéo traitée ainsi que les pistes d'annotation crées et éditées.

visuelle, nous avons ici fait le choix d'un questionnaire standard de la littérature. Parmi d'autre questionnaires de la littérature comme le *After-Scenario Questionnaire* (ASQ), le *Post-Study System Usability Questionnaire* (PSSUQ) ou encore le *Computer System Usability Questionnaire* (CSUQ) (Lewis 1995) , nous avons sélectionné le SUS pour sa rapidité, sa simplicité et l'applicabilité de l'ensemble de ses questions à notre système. Ce questionnaire comporte les dix affirmations suivantes :

1. J'aimerais utiliser ce système fréquemment
2. Je trouve ce système inutilement complexe
3. Je pense que ce système est facile à utiliser
4. J'aurais besoin d'un support technique pour pouvoir utiliser ce système
5. Les différentes fonctionnalités de ce système sont bien intégrées
6. Ce système est truffé d'incohérences
7. Le grand public peut apprendre à utiliser ce système très rapidement
8. Ce système est lourd à utiliser
9. J'ai confiance en ce système
10. J'ai du apprendre beaucoup choses avant de pouvoir utiliser le système

Pour chacune de ces affirmations, il est demandé au participant d'indiquer son niveau d'accord ou de désaccord. Pour cela, le participant attribue une note allant de 1 à 5 où 1 correspond à *pas du tout d'accord* et 5 correspond à *tout à fait d'accord*.

Une fois le questionnaire rempli, un score de satisfaction est calculé à partir de la formule proposée par les auteurs du questionnaire (Brooke *et al.* 1996) :

$$S = 2.5 * (\sum_{q=[1,3,5,7,9]} (Q_q - 1) + \sum_{q=[2,4,6,8,10]} (5 - Q_q)) \tag{6.1}$$

Dans cette formule, Qq est le score de la question q pouvant prendre une valeur de 1 à 5. Le score S estime donc la satisfaction globale du participant sur une plage de 0 à 100 : 0 étant *aucunement satisfait* et 100 étant *totalement satisfait*.

Ce score est finalement analysé pour comparer les résultats entre les deux modalités et les deux environnements.

6.3 Résultats

6.3.1 Déplacement

Le nombre de pas effectués par chaque participant a donc été recueilli pour chaque scénario et une ANOVA à deux facteurs (modalité et environnement) et à mesures répétées a été réalisée. Les résultats, sont présentés en Table 6.2 et Figure 6.7.

Pas	Réel	Virtuel	Moyenne
Tactile	230	164	197
Gestuel	244	158	201
Moyenne	237	161	199

ANOVA	F	p
Modalité	0.20	0.66
Environnement	18.24	0.002

TABLE 6.1 – Moyennes du nombre de pas effectués par les participants au cours d'un scénario, en fonction de la modalité et de l'environnement ; et résultat de l'analyse ANOVA à deux facteurs et mesures répétées.

6.3.2 Disponibilité des mains

L'analyse des 160 vidéos a donc permis de mesurer précisément les durées pendant lesquelles les participants quittent l'arme d'une main pour activer une commande au cours des scénarios. A partir de ces mesures, une ANOVA à deux facteurs (modalité et environnement) et mesures répétées a été réalisée. Les résultats, sont présentés en Table 6.2.

6.3.3 Satisfaction

Enfin, le dépouillement des questionnaires a permis d'évaluer subjectivement la satisfaction des participants pour chaque modalité et environnement. Les résultats, sont présentés en Table 6.3 et Figure 6.8.

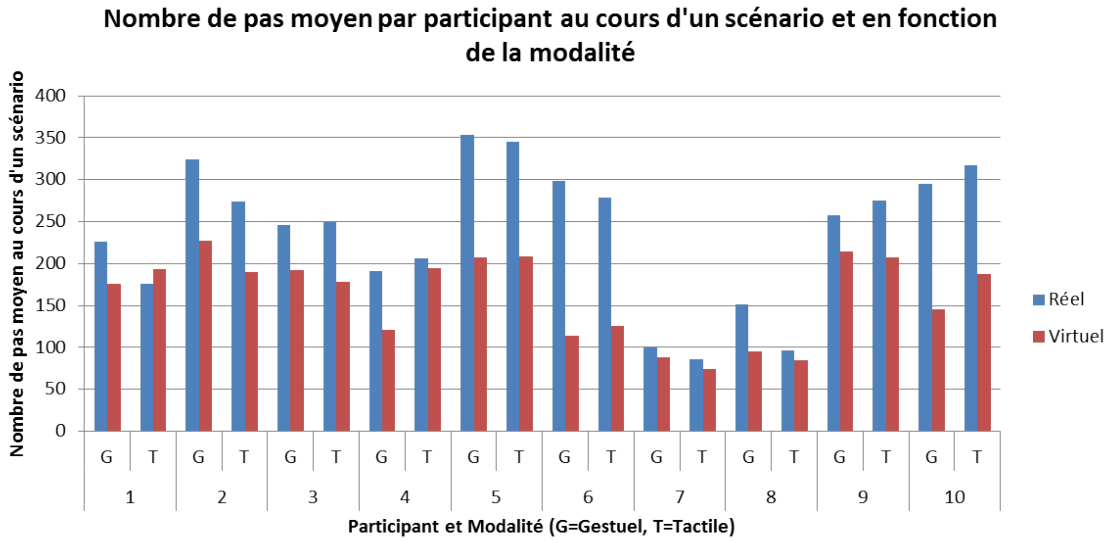


FIGURE 6.7 – Nombre de pas moyen effectué par les participants en fonction de la modalité et de l'environnement.

Durées	Réel	Virtuel	Moyenne
Tactile	101	104	103
Gestuel	66	39	52
Moyenne	84	71	78

ANOVA	F	p
Modalité	22.45	0.009
Environnement	1.07	0.35

TABLE 6.2 – Durée moyenne, en seconde, pendant laquelle les participants lâchent l'arme d'une main pour activer une commande au cours d'un scénario et résultat de l'analyse ANOVA à deux facteurs et mesures répétées.

SUS	Réel	Virtuel	Moyenne
Tactile	81.5	77.7	79.6
Gestuel	91.6	89.7	90.6
Moyenne	86.5	83.7	85.1

ANOVA	F	p
Modalité	5.18	0.049
Environnement	1.23	0.296

TABLE 6.3 – Moyennes des score de satisfaction des participants, recueillis sur une échelle de 0 à 100 avec le questionnaire SUS, en fonction de la modalité et de l'environnement ; et résultat de l'analyse ANOVA à deux facteurs et mesures répétées.

6.4 Discussion

6.4.1 Une utilisabilité globalement bonne

Dans cette étude, un modèle d'interaction, un vocabulaire gestuel et un système technique ont donc été évalués. Les différents résultats obtenus constituent, à notre connaissance, un premier retour d'expérience quant à l'utilisabilité de l'association de ces trois éléments.

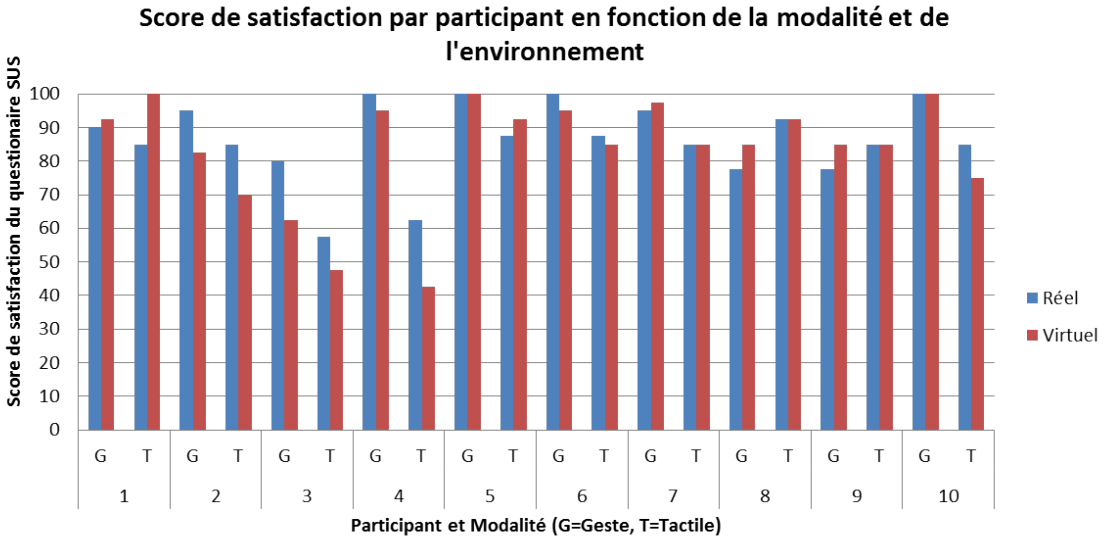


FIGURE 6.8 – Scores de satisfaction pour chaque participant en fonction de la modalité et de l'environnement.

6.4.1.1 Le geste n'est pas moins efficace que le tactile

Le premier enjeu était donc de contrôler la capacité du geste permette la commande effective d'un drone. Or, l'ensemble des participants ont pu activer les sept commandes de chaque scénario quel que soit l'environnement et la modalité.

Ce constat montre d'une part que, sans considération de modalité, le modèle d'interaction est adapté. En effet, la supervision reposant sur un plan de mission a été décrite comme un *bon compromis* permettant de bénéficier des avantages de la planification, et plus généralement de l'autonomie croissante des robots, tout en conservant un certain contrôle sur la machine.

D'autre part, ce constat montre que le vocabulaire gestuel a été appris par les participants que la brique de reconnaissance a correctement permis la détection de ces gestes. Ce second point nous semble particulièrement important puisque la maîtrise des gestes, à la fois par les participants et la machine, impacte directement la capacité à commander le drone et donc, l'efficacité du système.

6.4.1.2 Une meilleur efficience en partie confirmée

Le second enjeu de cette étude, et certainement le plus important, était de montrer l'efficience de la modalité gestuelle, c'est à dire sa capacité à permettre une interaction globalement moins coûteuse (physiquement et cognitivement). Dans ce but, la disponibilité des mains et la capacité à se déplacer au cours d'un scénario ont été mesurées. Ces résultats confirment la capacité du geste à libérer les mains. L'étude de l'impact sur le déplacement n'a quant-à lui pas été concluant ce qui, cependant, nous semble être dû au protocole mis en oeuvre plus qu'a la modalité.

La modalité gestuelle libère les mains

La commande avec l'interface tactile est significativement plus lente que la commande au geste (F=22.45, p=0.009) : au cours d'un scénario les participants ont passé en moyenne 103 secondes à commander le système avec l'interface tactile contre seulement 52 secondes avec l'interface gestuelle. Puisque chaque scénario comporte exactement 7 commandes, une interaction tactile dure donc en moyenne 15 secondes contre 7 secondes pour une interaction gestuelle. Ce résultat nous semble particulièrement important et deux facteurs sont à considérer :

Le premier facteur est que pour l'interaction tactile, le participant devait sortir et ranger la tablette à chaque commande, alors qu'en gestuelle, il lui suffisait de quitter l'arme pour réaliser le geste immédiatement.

Le second facteur est que chaque interaction est composée de 3 étapes : (a) activation d'une commande, (b) un message audio d'explicitation de la commande et et de demande de confirmation, et (c) confirmation. Or, en gestuel, certains participants replaçaient systématiquement la main sur l'arme pendant le message intermédiaire ; alors que cela ne c'est jamais produit avec l'interface tactile. En effet, le temps de sortir et ranger l'interface tactile étant plus long que la durée du message, les participants ont choisi de conserver l'interface en main.

Cette mesure montre donc bien que la modalité gestuelle sémaphorique permet d'activer une commande rapidement tout en conservant une bonne mobilité de la main. En outre, la capacité à reprendre rapidement l'arme en main, même au cours d'une interaction avec le système est un avantage important en environnement hostile. La commande au geste semble donc particulièrement adaptée à notre cas d'étude : la commande d'un drone militaire de contact.

Pas d'impact du geste sur le déplacement

Une de nos hypothèses quant à l'usage du geste est qu'il permet une plus grande mobilité du participant en impactant moins sa vision et son équilibre qu'avec une interface tactile. Afin de contrôler cette hypothèse, le nombre de pas réalisés par les participants au cours des différents scénarios ont donc été comptabilités et il était attendu des valeurs plus importantes en gestuel qu'en tactile. Cependant, l'effet de la modalité attendu n'a pas été constaté (F=.20, p=0.66) et ne permet donc pas de confirmer notre hypothèse.

L'élément majeur qui nous semble expliquer ce résultat est un biais possible dans le protocole mis en oeuvre : l'interprétation de la consigne de déplacement par les participants.

Il avait été explicitement demandé aux participants de chercher à progresser *tant que cela leur semblait possible*. Il était alors attendu que le seul élément pouvant empêcher le déplacement était l'activation d'une commande : perturbant la vision, voir l'équilibre. La durée plus importante de l'interaction avec la modalité tactile devait également renforcer cet effet.

Or, il a été constaté chez la majorité des participants, un impact des messages audio sur les déplacements : les participants s'immobilisaient lorsqu'une menace était signalée et ne reprenaient le déplacement que lorsque la menace était invalidée. Si la menace était

confirmée, les participants restaient alors statiques jusqu'à la fin du scénario. Ainsi, cet effet inattendu et constaté quelle que soit l'environnement (réel ou virtuel) a fortement impacté les résultats.

Il a par ailleurs été constaté, au regard des vidéos, que lors d'une commande en situation de déplacement, les participants pouvaient aussi bien s'immobiliser momentanément ou maintenir le déplacement quelle que soit la modalité. Il peut cependant sembler que l'immobilisation est plus fréquente avec la modalité tactile, mais le nombre trop réduit de données ne permet pas une analyse statistique.

Ainsi, avec cette étude, il ne semble pas possible de confirmer ni d'infirmer notre hypothèse de la capacité du geste à permettre une meilleur mobilité en comparaison d'une interface tactile standard.

6.4.1.3 Un modèle d'interaction et des gestes satisfaisants

La satisfaction des participants a donc été mesurée à l'aide d'un questionnaire standard de la littérature. D'une part, les résultats révèlent l'adéquation du modèle d'interaction avec le besoin utilisateur, et d'autre part, ils montrent une nette préférence pour la modalité gestuelle.

Modèle d'interaction et fonctions adressées

Ainsi, quelle que soit la modalité, le score de satisfaction global est relativement élevé (85.1) et peut être caractérisé d' *excellent* (Bangor *et al.* 2009). Ce résultat montre qu'au delà de la modalité, le modèle d'interaction dans son ensemble est relativement simple et adapté au contexte.

La plupart participants ont indiqué que le choix de la supervision plutôt que de la télé-opération est un premier élément fondamental pour libérer l'opérateur et lui permettre d'accomplir sa mission. En effet, l'objectif ici n'est pas la précision du vol mais l'acquisition d'informations élémentaires. Les fonctions du drone adressées ainsi que les feedbacks sonores ont été décrits comme *simples et efficaces*. La nécessité de la sécurisation des commandes par une étape de validation a également été considérée comme nécessaire ; et ce, quelle que soit la modalité.

Si l'automatisation des commandes de vol du drone ont été considérées comme *raisonnables et nécessaires* par les participants, le traitement des informations captées a bien plus été discuté. En effet, les participants ont indiqué qu'analyser une vidéo pour y détecter des menaces est une tâche complexe qui requiert une expertise perçue comme encore inaccessible à une intelligence artificielle.

Une réelle préférence pour les gestes

Le résultat du questionnaire révèlent que la commande au geste a été perçue comme significativement plus satisfaisante que la commande tactile ($F=5.18$, $p=0.049$) avec un score moyen de 90.6 pour la modalité gestuelle contre 79.6 pour la modalité tactile.

Ce résultat s'explique par le fait que les participants ont indiqué se sentir plus libres avec l'interface gestuelle. La manipulation et la visualisation de l'interface tactile, bien que

considérées comme normales, ont été ressenties comme une réelle limitation en comparaison de la modalité gestuelle. Des évènements ponctuels comme le reflet du soleil sur l'écran en environnement réel ou des activations involontaires de boutons lors de la manipulation de l'interface ont par ailleurs amplifié cette différence de contrainte imposée par les deux modalités.

En début d'étude, quelques participants ont émis des réserves quant à l'usage des gestes sur le terrain. Ils ont alors formulé des inquiétudes relevant effectivement de l'acceptabilité sociale et organisationnelle. Les participants ont indiqué craindre des gestes inadaptés et peut discrets. La capacité d'un système technique à détecter des gestes de manière robuste a également été questionnée.

Cependant, en fin d'étude, les appréhensions de ces participants semblent avoir disparu et l'élément qui semble avoir permis de lever ces craintes est le vocabulaire gestuel lui-même. En effet, les gestes ont été perçus comme très concrets, intuitifs et simples à réaliser. Ainsi, la faible amplitude des gestes et leur durée très courte semblent avoir impacté l'acceptation de leur usage et la perception de la capacité pour une machine de les reconnaître. Ce constat montre ici tout l'impact et la nécessité de faire intervenir des utilisateurs pour la construction du vocabulaire gestuel.

6.4.2 Des pistes pour une nouvelle étude

6.4.2.1 Conserver la marche sur place mais modifier sa mise en oeuvre

Dans cette étude, un environnement virtuel a donc été mis en oeuvre ; d'une part par sécurité vis à vis des conditions extérieures, et d'autre part, pour sa répétabilité. Cet environnement était composé d'une scène virtuelle dans laquelle les participants progressaient via une métaphore de *marche sur place*. Disposant ici de mesures à la fois en environnement réel et en environnement virtuel, nous nous sommes interrogés sur la cohérence des résultats entre ces deux conditions.

L'analyse des mesures révèle que l'usage du virtuel n'a pas impacté significativement l'activation des commandes et la manipulation de l'arme ($F=1.07$, $p=0.35$), ni la satisfaction relative au système testé ($F=1.23$, $p=0.296$). Ceci conforte l'usage de la RV comme outil pour l'organisation de testes utilisateurs.

En revanche l'environnement a eu un impact significatif sur le déplacement des participants ($F=18.24$, $p=0.002$). Il semble donc pertinent de questionner la métaphore de marche choisie ainsi qu'a sa mise en oeuvre technique.

La métaphore de marche sur place

Alors que les participants progressaient de manière fluide en environnement réel, il leur a été demandé de piétiner en environnement virtuel. Or, quatre participants ont indiqué que cette métaphore leur semblait intéressante d'un point de vue immersion mais toutefois moins naturelle que de simplement marcher. Ce constat est par ailleurs cohérent avec le résultat d'une étude qui a en effet montré que la marche sur place est perçue comme plus complexe et bien moins naturelle (Usoh *et al.* 1999).

Cependant, dans le cadre de l'étude présentée ici, la complexité ajoutée par la métaphore de marche ne semble pas nécessairement incompatible avec notre objectif qui est de mesurer l'éventuelle interaction entre l'action de commander un drone via une modalité gestuelle et celle de se déplacer en environnement complexe. Nous émettons l'hypothèse que réduire les automatismes de marche chez les participants rendrait plus évident l'interaction entre déplacement et commande du drone permettant ainsi de le mesurer ; si, effectivement, un tel effet existe bien.

Mise en oeuvre de la métaphore

Ainsi, la métaphore de marche sur place semble rester pertinente. Toutefois, sa mise en oeuvre technique semble devoir être modifiée. En effet, un phénomène particulier a été constaté chez tous les participants en environnement virtuel : en piétinant, les participants avançaient progressivement jusqu'à quitter le tapis de détection. C'est en constatant qu'ils ne progressaient plus dans la scène virtuelle que les participants prenaient conscience qu'ils avaient quitté les zones de détection. Ils se replaçaient alors immédiatement, avant de poursuivre.

Ce phénomène de décalage des participants semble dû à l'usage d'un tapis possédant des zones de détections qui, par ailleurs, sont relativement limitées. Pour une nouvelle étude, ce phénomène pourrait être évité en choisissant d'instrumenter le participant plutôt que l'environnent, comme nous l'avons fait avec le tapis. Pour ce faire, la reconnaissance d'actions sur place (marche, course et saut), à partir de données inertielles (Pfeiffer *et al.* 2016) semble être une solution de remplacement avantageuse.

6.4.2.2 Commande gestuelle en situation complexe

La commande gestuelle a donc été testée par des utilisateurs militaires ; et cette évaluation a montré la pertinence de cette modalité en situation de mobilité. Cependant, les deux environnements de cette étude : que ce soit l'environnement réel ou l'environnement virtuel, ne présentaient de toute évidence aucune menace et n'imposaient donc aucune contrainte forte aux participants.

Pour aller plus loin dans l'évaluation de de la commande au geste, il serait intéressant d'imposer des conditions supplémentaires afin, éventuellement, de faire apparaître des limitations pratiques ou techniques.

Il nous semble alors pertinent de considérer des contraintes à la fois cognitives et physiques pouvant être imposées, soit par l'environnement, soit par l'activité elle même. Quatre éléments nous semblent alors relativement écologiques tout en restant simples à mettre en oeuvre :

— **Pression temporelle :** Imposer aux participants de chercher à minimiser le temps requis pour activer chaque commande.
— **Pression environnementale :** Imposer aux participants d'utiliser la commande gestuelle dans un environnement stressant : par exemple un environnement particulièrement bruyant ou visuellement complexe.
— **Forte mobilité :** Imposer aux participants de chercher à utiliser la commande

gestuelle, non seulement en situation de marche, mais également de course ou de franchissement d'obstacle.

— **Encombrement physique :** Imposer aux participants un équipement plus complet pouvant limiter leurs mouvements comme le port d'un gilet pare balle.

6.4.2.3 Automaticité de la commande gestuelle

Au cours de l'étude, plusieurs participants ont indiqué que plus ils utilisaient les gestes, plus ils gagnaient en confiance et en fluidité. Quelques participants ont également exprimé le sentiment qu'avec une utilisation prolongée, la réalisation des gestes pourrait devenir automatique. Il a ainsi été suggéré que le vocabulaire gestuel puisse être retenu à long terme et utilisé de manière presque inconsciente, même en situation de stress.

De manière plus précise, sur un plan cognitif, ces remarques faites par des participants conduisent à adresser la question de la création d'automatismes. Formellement, un automatisme est défini comme étant un *processus dépourvu d'attention consciente* (Singer 2002). Il possède deux propriétés fondamentales (Perruchet 1988) : d'une part l'absence de coût cognitif (ou charge mentale), et d'autre part l'absence de contrôle attentionnel.

L'absence de coût cognitif est un élément particulièrement intéressant dans notre contexte. En effet, il renforcerait la capacité d'un utilisateur à activer une commande gestuelle de manière instantanée tout en poursuivant l'accomplissement d'une autre tâche : ici le déplacement, le contrôle de l'environnement et la manipulation de l'arme. Alors que lors de l'étude de l'impact de la modalité de commande sur la capacité à percevoir l'environnement (Chapitre 4), nous avons justement montré le moindre coût de la commande gestuelle par rapport à la commande tactile ; la possibilité d'une automatisation de la commande gestuelle suggère que cette différence pourrait être rendue encore plus significative.

L'absence de contrôle attentionnel correspond quant à lui à un aspect généralement considéré comme plus négatif des automatismes. Il s'agit du fait qu'une action soit réalisée sans prise de décision consciente, donc pouvant ignorer des informations importantes. Ce qui devient alors mentalement coûteux, c'est le déploiement d'activités inhibitrices (Perruchet 1988) pour modifier ou interrompre un comportement automatique. Ainsi, les automatismes peuvent devenir problématiques lorsqu'ils produisent des réactions inadaptées en situation critiques, comme par exemple en aviation (Leplat 2005) ; cependant, dans le contexte de la commande gestuelle d'un drone de contact, cet élément ne semble pas particulièrement critique.

Ainsi, la possibilité de l'automatisation de la commande au geste nous semble tout à fait bénéfique et renforcerait la pertinence de cette modalité. Cependant, bien que notre vocabulaire gestuel soit de taille très restreinte et ait été construit de manière à être simple à comprendre et à apprendre, il n'existe à notre connaissance aucune certitude quant à l'automaticité de la commande gestuelle ni sur son impact significatif sur l'utilisabilité.

Finalement, les questions de l'automaticité de la commande gestuelle et de son impact restent ouvertes et nous semblent particulièrement pertinentes à adresser.

6.5 Conclusion

Dans cette étude qui constitue le dernier travail de cette thèse, un modèle d'interaction, le vocabulaire gestuel méthodiquement constitué et un système technique complet ont donc été expérimentés dans leur globalité. Nous avons alors adressé la question de l'utilisabilité de la commande gestuelle, autant sur son aspect pratique que sur son aspect technique.

Pour ce faire, des tests utilisateurs ont été conduits : dix participants militaires ont expérimenté le système dans différentes conditions et des mesures objectives et subjectives ont été recueillies. Ces mesures ont montré que pour l'activation d'un nombre réduit de commandes, l'efficacité des modalités gestuelles et tactiles sont comparables ; validant ainsi le modèle d'interaction et le dictionnaire de gestes.

Concernant l'efficience, ces mesures ont également confirmé notre hypothèse d'une modalité moins contraignante. En effet, il a été montré que la commande au geste est instantanée et plus rapide à activer. En revanche, aucun élément de réponse n'a peu être apporté quant à l'impact de la commande gestuelle sur la mobilité.

Enfin, un questionnaire standard de la littérature ainsi que des remarques des participants au cours des différentes sessions ont révélé un fort degré de satisfaction vis-à-vis du modèle d'interaction proposé pour la commande de drone, et un intérêt certain pour la modalité gestuelle malgré des craintes formulées en début d'étude.

Finalement, ces tests utilisateurs constituent un premier retour d'expérience encourageant pour la commande gestuelle. Ils ont permis de montrer la réelle valeur opérationnelle de cette modalité et d'ouvrir de nouvelles perspectives comme son usage en milieu plus complexe et l'étude de caractéristiques fondamentales telles que l'automaticité.

CHAPITRE 7

Conclusion et perspectives

Sommaire

7.1 Conclusion sur la commande gestuelle sémaphorique

7.1.1 Pourquoi les gestes sémaphoriques ?

Le geste est une modalité intéressante en ce qu'elle permet une interaction directe, sans support matériel. Ce sont alors les mouvements du corps qui permettent directement de modifier l'environnement ou d'exprimer une intention. Pour cette raison, cette modalité intéresse particulièrement les chercheurs de différents domaines dont celui des interfaces homme-machine.

Cependant, la notion de geste est large. De ce fait, il est difficile d'objectiver son bon usage et sa mise en oeuvre technique.

En nous basant sur la littérature en psycho-linguistique nous proposons qu'un geste est avant tout une signification donnée au mouvement d'un corps humain, par un observateur et au regard d'une norme. Ainsi, ce qui est signifiant dans un mouvement dépend de la perspective de cet observateur et pour cette raison, il est légitime de trouver différentes classifications et modélisations adressant différents niveaux de détail.

Adressant la question de l'interaction homme-robot en environnement spécifique : commande d'un drone en mobilité et en environnement hostile, il nous importe de catégoriser les gestes suivant leur emploi afin de déterminer le ou lesquels seraient les plus pertinents.

En considérant la fonction des mouvements, on peut distinguer les gestes manipulatifs qui servent à modifier l'environnement et les gestes communicatifs qui servent à transmettre

des informations. Puis, pour les gestes communicatifs, en fonction de leur contexte d'usage, on peut considérer les trois catégories majeures du continuum de Kendon : les gestes co-verbaux : symboles improvisés qui accompagnent la parole, les gestes sémaphoriques : symboles standardisés et utilisés seuls, et les langues des signes : ensembles de symboles standardisés et structurés grammaticalement.

Ainsi, nous avons choisi de considérer quatre grandes catégories de gestes : (1) manipulatifs, (2) co-verbaux, (3) sémaphoriques et (4) langues des signes. Chacun de ces types de geste possède des propriétés pratiques et techniques le rendant plus ou moins pertinent en fonction du contexte d'utilisation adressé.

Dans le cadre de la commande d'un drone militaire de contact, nous avons choisi en premier lieu, d'écarter l'usage des gestes manipulatifs et des gestes co-verbaux. En effet, les gestes manipulatifs ont pour propriété de permettre un contrôle continu d'une ou plusieurs variables alors que dans le contexte adressé ici, l'objectif est d'activer des fonctions autonomes du drone pour libérer l'utilisateur d'une partie de la charge de travail. Les gestes co-verbaux ont quant-à eux été écartés en raison de l'usage conjointe de la parole qui ne satisfait pas la contrainte de discrétion impérative en milieu hostile.

Ainsi, les gestes sémaphoriques et les langues des signes sont les seuls types de geste à priori pertinents puisque ce sont des ensembles de symboles standardisés qui permettent d'exprimer discrètement une intention et donc d'activer les commandes d'un robot. Les langues des signes étant structurées grammaticalement, elles présentent une expressivité plus importante que les gestes co-verbaux. Des intentions plus complexes peuvent donc être exprimées ce qui autoriserait une interaction plus complète avec un robot. En plus d'indiquer quelle commande activer, l'utilisateur pourrait alors en préciser certains paramètres comme le lieu le moment ou la durée.

Cependant, la question de la faisabilité technique c'est alors posée. En effet, à notre connaissance, la reconnaissance robuste de ces deux types de gestes n'était pas évidente. La reconnaissance des gestes sémaphoriques étant plus simple que celle des langues des signes en raison de l'absence de grammaire et de phénomènes tels que la co-articulation, nous avons choisi d'adresser l'usage des gestes sémaphoriques dans un premier temps.

Notre choix était alors en contradiction avec des considérations pratiques formulées par certains auteurs. Wexelblat par exemple, remet en question la pertinence de leur utilisation puisque, à juste titre, les gestes sémaphoriques sont artificiels par natures (standardisés). Cependant, nous avons formulé l'hypothèse que, bien que les gestes sémaphoriques soient artificiels en ce qu'ils ont besoin d'être définis et appris, ils permettent d'activer des commandes autonomes en impactant faiblement les capacités de leur utilisateur à percevoir l'environnement, se déplacer et manipuler des objets.

Ainsi, nous avons fait le choix d'adresser l'usage des gestes sémaphoriques pour commander un robot en situation de mobilité et en environnement hostile. L'enjeu de nos travaux était donc de montrer l'utilisabilité de cette modalité autant sur le plan pratique que sur le plan technique.

7.1.2 Comment utiliser les gestes sémaphoriques ?

Les gestes sémaphoriques sont donc des gestes communicatifs standardisés qui expriment des concepts. Le principe de base de leur emploi est simple : il s'agit d'associer à chaque fonction autonome d'un système, la réalisation d'un geste spécifique. Cependant, sous cette apparente simplicité se cache deux problématiques : d'une part, le geste est uniquement une modalité d'entrée qui doit donc être complétée par un modèle d'interaction respectant un ensemble d'heuristiques d'ergonomie. D'autre part, pour une application donnée, quels gestes utiliser et comment les choisir ?

Modèle d'interaction :

Ainsi, nous nous sommes interrogés lors du Chapitre 2 sur ce qu'il faut à la modalité gestuelle sémaphorique pour être complète. Une évaluation heuristique nous a permis de mettre en évidence la nécessité de feedbacks et d'un mécanisme de sécurisation des commandes.

Pour assurer la visibilité constante de l'état du robot mobile, y compris lorsqu'il est en dehors du champ de vision de l'opérateur, des messages de prise en compte d'une commande et de fin d'accomplissement d'une commande ont été ajoutés.

Par ailleurs, alors qu'une interface matérielle peut être activée ou désactivée, l'usage du corps sur lequel repose la modalité gestuelle est quant à lui permanent. Ainsi, il est possible qu'un geste soit réalisé involontaire ou détecté par erreur par la machine ; conduisant alors à l'activation non voulue d'une fonction du robot mobile. Ce comportement de l'interface constitue donc une source d'erreur qu'il convient d'empêcher ; et particulièrement dans un contexte militaire ou toute action involontaire peut directement nuire à la sécurité des soldats.

Nous avons donc proposé d'ajouter un mécanisme de sécurisation reposant sur un principe de validation volontaire de la part de l'utilisateur. Ainsi, pour activer une commande du robot mobile, il est nécessaire de compléter une coute phase de dialogue comportant quatre étapes : (1) demande d'activation d'une fonction spécifique par la réalisation du geste sémaphorique correspondant, (2) explicitation par le système de la commande invoquée et demande de confirmation, (3) confirmation de l'activation par la réalisation d'un gestes spécifique, et (4) activation par le système de l'exécution de la fonction autonome et confirmation de sa réalisation avec un message sonore.

Choix des gestes sémaphoriques :

Les gestes sémaphoriques, comme les langues des signes, sont des symboles standardisés qui doivent être connus des différents interlocuteurs de la situation de communication. Dans le cadre de l'interaction gestuelle proposée ici, les gestes doivent donc être partagés par l'opérateur et la machine.

Il est donc nécessaire que le vocabulaire de gestes soit défini et appris par les utilisateurs. Différentes stratégies existent alors pour constituer un vocabulaire de gestes :

Les gestes peuvent être définis par les concepteurs du système afin d'assurer des performances optimales. Cette solution n'est cependant pas acceptable car elle peut conduire

à un vocabulaire de gestes inadaptés : difficiles à réaliser et à apprendre.

Les gestes peuvent également être définis par les utilisateurs : soit de manière individuelle permettant à chacun de posséder son propre vocabulaire, soit de manière collective pour constituer un unique vocabulaire commun à tous. Bien que ces deux dernière propositions semblent pertinentes, nous avons choisi d'explorer la voie de la définition d'un vocabulaire unique.

Il a donc été nécessaire d'adresser la problématique de la méthodologie à appliquer pour construire de manière collective un vocabulaire de gestes. L'objectif est alors de construire un dictionnaire adapté aux utilisateurs. Les gestes doivent être faciles à mémoriser et pour cela, il nous semble qu'ils doivent être logiques sémantiquement (la forme du geste représente la fonction à activer) et ne doivent pas être ambigus (un même geste ne peut pas représenter plusieurs fonctions et inversement).

Nous avons proposé de combiner deux méthodologies issues de la littérature pour en former une nouvelle composée de quatre étapes. Dans un premier temps, des participants élicitent individuellement un geste par fonction à adresser. Puis, les propositions de gestes recueillies sont analysées par des examinateurs pour déterminer les groupes de propositions identiques : les gestes candidats. L'ensemble des gestes candidats forment alors un catalogue qui est ensuite présenté à de nouveaux participants. Lors de cette troisième phase, il est alors demandé aux participant d'élire, dans le catalogue, un geste par fonction du système. Pour chaque fonction, le geste ayant reçu le plus de suffrages est conservé pour constituer le vocabulaire gestuel définitif. Finalement, il est demandé à un troisième groupe de participants de reconstituer les associations geste-fonction du vocabulaire élu alors que celles-ci sont présentées dans un ordre aléatoire. La capacité à reconstituer le vocabulaire par des participants n'ayant pas pris part à sa construction est alors un indice de la qualité de ce vocabulaire : les gestes sont compréhensibles et non ambigus.

Cette méthodologie a été appliquée pour construire un dictionnaire de gestes permettant d'adresser six fonctions d'un drone (décollage, atterrissage, suivant, précèdent, stop et base) ainsi qu'un geste de confirmation et un geste d'annulation requis pour le mécanisme de sécurisation des commandes du modèle d'interaction.

Nous avons montré que cette méthodologie permet effectivement de construire un dictionnaire consensuel et non ambigu particulièrement adapté aux utilisateurs. De plus, il a été constaté que l'implication d'utilisateurs finaux dès cette phase amont de construction du vocabulaire (alors qu'aucun système n'est encore utilisable) permet aux participants de se projeter dans son usage et rend également d'avantage légitimes les gestes recueillis vis-à-vis des futurs utilisateurs, participants à son acceptation.

7.1.3 Quel système pour reconnaître des gestes sémaphoriques ?

Une fois un dictionnaire de gestes construit, il convient de permettre à un système de les détecter et reconnaître de manière robuste et temps réel (sans latence perceptible).

Pour cela, il est nécessaire dans un premier temps, de déterminer les différentes fonctions fondamentales d'un système gestuel pour ensuite en déduire une architecture logicielle et ses différents composants. Au regard de la littérature, les fonctions fondamentales d'un

tel système sont (1) la captation et la représentation des mouvements, (2) la sélection temporelle, (3) la détection et la classification, et (4) dans certains cas, le suivi de la progression temporelle.

Dans le contexte de la commande d'un robot mobile avec des gestes sémaphoriques, la fonction de suivi des gestes ne semble à priori pas pertinente. En effet, les gestes sont ici très courts et l'information de progression n'intervient pas dans le modèle d'interaction.

La captation et la représentation des mouvements consiste à déterminer les caractéristiques géométriques et temporelles signifiantes dans les gestes du vocabulaire pour choisir et mettre en oeuvre des capteurs adaptés et transformer les données fournies par ceux-ci pour les exprimer dans une représentation pertinente. Pour le dictionnaire de gestes considéré, les informations physiques signifiantes sont la configuration (état plié-tendu des doigts), l'orientation et le mouvement de la main qui réalise les gestes. Au regard des informations requises et du contexte applicatif, un ensemble de capteurs ont été choisis. La situation de mobilité en environnement non contrôlé interdit l'usage de la majorité des système de captation basés vision (caméras) et conduit naturellement à la mise en oeuvre de capteurs portés. Parmi les différentes technologies disponibles, nous avons fait le choix d'un gant instrumenté comportant un groupe de centrales inertielles de petite taille positionnées de part et d'autre des articulations principales de la main. Le gant utilisé est un modèle standard parfaitement intégré et utilisant 12 centrales inertielles. L'ensemble des données captées sont traitées en temps réel pour ne conserver que les quelques informations nécessaires à la reconnaissance des gestes du dictionnaire.

La fonction de sélection temporelle consiste à extraire du flot continue de données fournies par les capteurs, des séquences de taille finie pour être analysées par les fonctions de détection et de classification. Parmi les stratégies existantes, nous avons fait le choix d'un simple fenêtrage glissant, dont les paramètres sont la taille de la fenêtre et son déplacement. Au regard des gestes du dictionnaire et de la réactivité souhaitée, nous avons choisi une fenêtre large de 2 secondes et un pas de déplacement de 100 millisecondes.

Finalement, les fonctions de détection et de classification ont pour objectif d'analyser les séquences fournies par la fonction de sélection temporelle pour déterminer si elles correspondent à un geste standard et, dans l'affirmative, duquel il s'agit. Pour ce faire, deux approches sont généralement considérées dans la littérature : les modèles de Markov cachés (MMC) et l'alignement temporel dynamique (DTW). Ces deux approches ont en commun d'estimer la similitude entre une séquence analysée et des modèles appris à partir d'exemples. Si la similitude estimée est suffisamment forte (dépasse un seuil statique ou dynamique), il y a alors détection et le modèle vainqueur désigne quel geste a été détecté.

Cependant, nous avons choisi de considérer une autre approche ne reposant pas sur un principe d'apprentissage à partir d'exemples, mais de spécification explicite par un expert, sous forme textuelle dans un langage spécifique.

En effet, puisque nous avons de par la méthodologie de construction du vocabulaire, sa connaissance théorique (les éléments signifiants de chaque geste), il nous semble plus pertinent de les indiquer directement à la machine. Il n'est donc pas nécessaire de passer par une phase de construction de base d'exemples et d'apprentissage automatique dont le risque est de perdre de l'information en plus d'être particulièrement coûteux en temps (implication

de participants).

Ainsi, nous avons proposé de spécifier directement les gestes à la machine. Il a alors fallu répondre à deux problématiques : d'une part, selon quel formalisme spécifier des gestes, et d'autre part, comment utiliser une spécification pour analyser une séquence et déterminer si elle correspond effectivement à la réalisation d'un geste.

Ces deux problématiques ont finalement été adressées à l'aide d'un couple d'outils particulièrement classiques en informatique : les expressions régulières et les automates déterministes. Les expressions régulières sont des chaînes de caractères écrites avec un alphabet et une syntaxe spécifiques. Elles sont ensuite compilées pour construire automatiquement des automates déterministes composés d'états et de transitions logiques. Ce sont les automates qui finalement permettent d'analyser (parser) une séquence de données pour déterminer si elle respecte le modèle décrit. Le point particulier pour pouvoir appliquer ce couple d'outils à la reconnaissance gestuelle est de choisir un alphabet fini d'éléments atomiques permettant d'exprimer un geste sous forme d'expression régulière. Pour ce faire, nous avons assez simplement combiné les éléments physiques signifiants de notre vocabulaire gestuel (configuration, orientation, mouvement).

Finalement, un système technique complet permettant la reconnaissance d'un dictionnaire de gestes a été mis en oeuvre. Un ensemble de centrales inertielles fournissent un flot continu de données dont les éléments sémantiques sont fournis sous forme de séquences temporelles à un ensemble d'automates déterministes, construits à partir d'expressions régulières, et qui assurent les fonctions de détection et de reconnaissance.

Nous avons montré par des tests fonctionnels qu'un tel système est relativement robuste et qu'il ne semble pas présenter de latence perceptible. Ce résultat est cohérent du fait du nombre restreint de gestes adressés, de leur faible durée (de l'ordre de la seconde) et de leur grande simplicité.

L'utilisation des expressions régulières rend par ailleurs l'implémentation du système extrêmement simple puisque cet outil est géré nativement par la majorité des langages de programmation (C++, Java, Python, ...).

La difficulté majeure de cette approche réside dans l'écriture de l'expression régulière qui nécessite d'une part la connaissance théorique des éléments signifiants d'un geste et de leur anatomie, et d'autre part, une certaine forme d'expertise sur la manière de transcrire ces éléments signifiants.

7.1.4 Opérationnellement, quel est le bilan ?

Nous avons donc également évalué l'utilisabilité de la commande gestuelle sémaphorique. Pour cela, deux études ont été conduites : l'une en laboratoire et magicien d'Oz, et l'autre, en situation écologique et avec un système technique réel.

Attention visuelle :

La première étude, présentée au Chapitre 4, a eu pour objectif de confirmer notre hypothèse que le geste sémaphorique libère les yeux de l'utilisateur.

En effet, si il parait évident que le fait d'avoir à regarder une interface physique telle qu'une tablette tactile, détourne le regard de l'environnement et empêche sa bonne perception, il n'est pas moins évident que l'utilisation de gestes n'empêche pas la compréhension de l'environnement en imposant une charge cognitive par la mémorisation et le rappel des gestes à utiliser.

Pour cela un protocole de double tâche a été mis en oeuvre. Il a été demandé à des participants de signaler le plus rapidement possible l'apparition de stimuli visuels connus dans un environnement virtuel tout en activant les commandes d'un drone virtuel soit en utilisant un vocabulaire de gestes sémaphoriques préalablement appris, soit une interface tactile simple. Dans ce protocole, la mesure principale est le temps moyen mis par chaque participant pour signaler l'apparition des stimuli visuels. En fonction de la modalité utilisée, plus ce temps de réaction est long, plus la commande a un impact négatif fort.

Les résultats obtenus ont révélé une différence significative entre la modalité gestuelle et une interface tactile simple. Les participants mettaient en moyenne moins de temps pour détecter un évènement visuel avec la commande gestuelle, confirmant notre hypothèse et donc la pertinence de la commande gestuelle.

Manipulation et déplacement :

La seconde étude, présentée au Chapitre 6, a eu pour objectif d'évaluer notre hypothèse d'une libération de la mobilité des utilisateurs ; en particulier les capacités à utiliser les mains et à se déplacer.

En effet, la dématérialisation de l'interface physique ainsi que l'absence de support visuel sont des éléments qui semblent favoriser la grande liberté d'action de l'utilisateur.

Un second protocole de double tâche a été mis en oeuvre. Des participants militaires ont utilisé la commande gestuelle, ainsi qu'une interface tactile, pour commander un drone virtuel tout en cherchant à progresser sur un itinéraire en environnement extérieur et en conservant les deux mains sur une arme.

Dans ce protocole, la durée des évènements de lâchée de l'arme, imposés par la commande du drone virtuel a été mesurée afin d'évaluer l'impact de la modalité sur la disponibilité des mains. La distance moyenne parcourue au cours d'un scénario a également été mesurée pour étudier la relation entre la modalité et la capacité de déplacement.

Les résultats obtenus ont révélé une différence significative entre la modalité gestuelle et une interface tactile simple pour la disponibilité des mains mais n'ont pas permis de conclure quant à un éventuel effet sur la capacité à se déplacer.

En effet, la modalité gestuelle est significativement plus rapide que la modalité tactile. Elle contraint moins longtemps l'usage des mains et impose donc de lâcher l'arme pour des périodes plus courtes. Il a également été montré qu'au delà de la durée de l'interaction, une commande gestuelle peut être interrompue instantanément ; ce qui constitue un apport significatif dans le domaine militaire où la capacité à réagir face à es menaces immédiates est fondamentale.

Le protocole n'a cependant pas permis de conclure quant-à un apport de la modalité gestuelle pour favoriser le déplacement de l'utilisateur. En effet, au cours des scénarios, le déplacement des participants ayant été impacté d'avantage par des considérations opération-

nelles que par la modalité elle-même, les résultats ne sont pas exploitables. L'hypothèse d'un impact positif du geste sur la capacité à se déplacer reste donc à adresser.

Apprentissage et satisfaction :

Lors des deux études conduites, des participants à la fois civils et militaires ont appris et utilisé le modèle d'interaction ainsi que le vocabulaire de gestes respectivement présentés aux Chapitre 2 et 3.

Il a été confirmé que l'apprentissage et l'usage des gestes était possible avec un court entraînement : de l'ordre de quelques minutes seulement. Le vocabulaire gestuel, présenté soit par des images accompagnées d'une courte description textuelle, par des vidéos ou par des démonstrations, a parût logique et simple à retenir et à réaliser. Cependant, il a également été montré que cette phase d'apprentissage est impérative pour la modalité gestuelle alors qu'une modalité tactile simple est quant-à elle immédiatement utilisable.

Par ailleurs, il a été constaté que l'usage prolongé de l'interface gestuelle pouvait conduire à une plus grande maîtrise des gestes renforçant son aspect rapide et instantané. La question de la création d'automatismes pour la réalisation des commandes gestuelles a été soulevée et reste encore ouverte.

Que ce soit en magicien d'Oz ou avec un système réel, la commande gestuelle a été perçue comme étant particulièrement plus satisfaisante que la modalité tactile dans un contexte de double tâche. En effet, les différentes contraintes imposées par l'interaction tactile ont constitué des sources de frustration évidentes. Tourner la tête, manipuler et maintenir une interface physique, activer avec précision des éléments de taille réduite sont autant d'éléments qui impactent l'activité.

Le type de geste choisi, la qualité sémantique du vocabulaire et l'implication d'utilisateurs experts dans la construction du dictionnaire semblent grandement participer à l'appréciation subjective globale positive de la commande gestuelle et à sa bonne acceptation.

7.2 Perspectives - Aller plus loin avec les gestes sémaphoriques

Ainsi, les travaux réalisés ont permis de montrer la pertinence à la fois technique et pratique de la commande gestuelle sémaphorique pour la commande d'un robot mobile. De nombreuses questions restent cependant encore ouvertes et ouvrent des perspectives pour de futurs travaux.

7.2.1 Intégration d'un système opérationnel

Le système technique développé et présenté au Chapitre 5 comprend donc différents éléments dont une brique de reconnaissance gestuelle, une plate-forme interactive et des composants de simulation de drone.

Le test utilisateur présenté au Chapitre 6 a par ailleurs montré la bonne robustesse de ce système. Cependant, il ne s'agit actuellement que d'un prototype qui permet de réaliser de premières études mais qui de toute évidence ne correspond pas à une réalité opérationnelle.

Aussi, premier travail d'ingénierie nécessaire serait d'industrialiser un système technique complètement intégré. Les points fondamentaux à adresser sont alors (1) la spécification d'un gant instrumenté ne comportant que les capteurs requis et compatibles des contraintes du domaine militaire, (2) le portage des traitements logiciels permettant la reconnaissance des gestes sur une plate-forme de faible encombrement (calculateur de type smart-phone), (3) la connexion d'un périphérique de restitution sonore à conduction osseuse, et (4) la connexion à un drone réel.

Les trois premiers points semblent ne pas présenter d'obstacles majeurs. En effet, les technologies de captations des gestes sont déjà relativement anciennes et bien maîtrisées, l'usage des expressions régulières pour la description et la reconnaissance des gestes est supportée nativement par tous les langages de programmation actuels et ne nécessite donc pas à priori de développement spécifique, et les systèmes de restitution sonore discrets sont usuels dans le domaine militaire et tendent à se démocratiser dans le secteur civil.

Il nous semble qu'à ce jour, seule la connexion à un drone réel constitue un enjeu important. En effet, les problématiques de sécurité des biens et des personnes dont la législation relative évolue rapidement, imposent des contraintes fortes qui nécessitent une attention particulière.

7.2.2 Poursuivre l'étude des propriétés des gestes sémaphoriques

Dans ces travaux, nous avons montré la pertinence du geste d'un point de vue pratique. Certaines hypothèses restent cependant encore ouvertes et de nouveaux questionnements, plus précis sont apparus.

L'impact de la commande gestuelle sémaphorique sur la capacité à se déplacer reste un élément majeur que nos travaux ont échoué à adresser. Cette hypothèse nous semble cependant particulièrement importante dans le contexte militaire. Comme pour l'attention visuelle, il nous semble fondamental pour répondre à cette problématique, de distinguer l'impact physique et l'impact cognitif de la modalité. Si une interaction existe entre la modalité gestuelle et l'action de se déplacer, repose-t-elle sur des conflits mécaniques au sein du corps humain ou sur une problématique plus générale d'équilibre et de perception de l'environnement ?

Trois éléments de questionnement plus précis sont également apparus au cours de nos travaux : (1) comment représenter des gestes, (2) comment et à quel point peut-on les apprendre, et (3) des automatismes gestuels peuvent ils être créés et quels seraient leur impact opérationnel.

La représentation des gestes nous semble particulièrement importante puisqu'elle conditionne l'apprentissage des gestes et donc la possibilité d'utiliser effectivement cette modalité. Lors des différentes études conduites, les participants ont toujours été accompagnés par une personne experte du dictionnaire de gestes ; et le remplacement de cette présence humaine par un "manuel" de formation, quel que soit son format est un point qui semble fondamental.

La question de l'apprentissage des gestes semble également importante. En effet, dans nos travaux, le nombre de gestes était relativement restreint : inférieur à la dizaine. Or, l'un des avantages significatifs de la modalité gestuelle, en comparaison du tactile, est son expressivité très grande voir illimitée. Ainsi, il semble possible d'adresser un très grand nombre de fonctions là ou les dimensions d'une interface tactile sont une limitation forte. Ainsi, même si il ne semble pas y avoir de limite théorique à la quantité de gestes qui peuvent être appris, il nous semble important de pouvoir caractériser la difficulté d'apprentissage en fonction du nombre de gestes et de leur qualité sémantique (si ils sont logiques pour les utilisateurs). Ainsi, il serait possible de déterminer objectivement et à priori l'acceptabilité de l'usage des gestes pour une application donnée.

Finalement, la question des automatismes gestuels complète les questions de la représentation et de l'apprentissage des gestes en étendant l'étude sur une échelle de temps plus long. Cette question nous semble particulièrement intéressante car elle pourrait également en constituer un outil de mesure plus précis. En effet, la capacité à constituer des automatismes gestuels et donc d'amplifier les avantages de cette modalité est un critère dont l'apparition et le temps avant apparition peut être indicatif du processus global de l'usage des gestes (représentation, apprentissage, usage). Ainsi, si utiliser quelques gestes, quels qu'ils soient, semble relativement simple à court terme, des variations inhérentes à une forme de qualité et de complexité des gestes pourrait devenir mesurable avec un usage prolongé.

Beaucoup de questions restent ainsi ouvertes quant à ces phénomènes et sur la manière de les mesurer objectivement.

7.2.3 Adresser de nouveaux cas d'application

Finalement, nous avons adressé l'usage de la modalité gestuelle sémaphorique d'un point de vue pratique et technique. Cela a permis de mettre en évidence des avantages de cette modalité qui est ainsi particulièrement pertinente pour le contexte opérationnel de la commande de robots mobiles en environnement complexe ; en particulier la commande des drones militaires de contact en environnement hostile.

Si l'usage des geste a été perçu comme une proposition particulièrement pertinente, de nombreuses questions restent cependant encore ouvertes quant à la manière d'utiliser les drones. Leur autonomie énergétique et décisionnelle constituent encore des freins pratiques qui, au cours de nos travaux ont parfois atténué l'intérêt de la modalité de commande elle même.

En effet, alors que nous avons montré que le geste permet d'être globalement plus disponible, les contraintes techniques actuelles des drones de faible taille imposent à elles seules des contraintes opérationnelles fortes qui limitent l'apport de la modalité gestuelle. Cependant, les drones ont fait l'objet de progrès techniques très importants au cours des dernières décennies et de nouvelles évolutions sont encore raisonnablement certaines. Ainsi, la commande gestuelle sémaphorique pourrait devenir d'autant plus pertinente dans les années à venir.

En dehors de la problématique de la commande des robots mobiles, l'aspect d'un encombrement moindre de la modalité gestuelle sémaphorique semble également pertinent, et

dès aujourd'hui, pour les environnements virtuels. En effet, l'absence de support physique à regarder et manipuler nous semblent être compatibles de certains enjeux tels que l'immersion, l'interaction et le sentiment de présence. De plus, l'expressivité à priori illimitée de la modalité gestuelle semble un avantage fort pour des environnements par définition illimités tant dans leurs dimensions que par les objets qu'ils permettent de représenter.

La modalité gestuelle sémaphorique favorise donc la disponibilité et la mobilité qui sont des enjeux importants pour les domaines tels que les objets connectés, la domotique, les véhicules autonomes et les usines du futur.

Bibliographie

[Amft *et al.* 2005] Oliver Amft, Holger Junker et Gerhard Tröster. *Detection of eating and drinking arm gestures using inertial body-worn sensors*. In Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on, pages 160–163. IEEE, 2005. *pas de citation*

[Archer 1997] Dane Archer. *Unspoken diversity : Cultural differences in gestures*. Qualitative Sociology, vol. 20, no. 1, pages 79–105, 1997. *pas de citation*

[Bangor *et al.* 2009] Aaron Bangor, Philip Kortum et James Miller. *Determining what individual SUS scores mean : Adding an adjective rating scale*. Journal of usability studies, vol. 4, no. 3, pages 114–123, 2009. *pas de citation*

[Bastien & Scapin 1993] JM Christian Bastien et Dominique L Scapin. *Ergonomic criteria for the evaluation of human-computer interfaces*. 1993. *pas de citation*

[Bauer *et al.* 2009] Andrea Bauer, Klaas Klasing, Georgios Lidoris, Quirin Mühlbauer, Florian Rohrmüller, Stefan Sosnowski, Tingting Xu, Kolja Kühnlenz, Dirk Wollherr et Martin Buss. *The autonomous city explorer : Towards natural human-robot interaction in urban environments*. International Journal of Social Robotics, vol. 1, no. 2, pages 127–140, 2009. *pas de citation*

[Baum & Sell 1968] Leonard E Baum et GR Sell. *Growth functions for transformations on manifolds*. . ¡. Math, vol. 27, no. 2, pages 211–227, 1968. *pas de citation*

[Baum *et al.* 1967] Leonard E Baum, John Alonzo Eagon*et al. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology*. Bull. Amer. Math. Soc, vol. 73, no. 3, pages 360–363, 1967. *pas de citation*

[Bellugi 1979] Ursula Bellugi. The signs of language. Harvard University Press, 1979. *pas de citation*

[Bevilacqua *et al.* 2009] Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy et Nicolas Rasamimanana. *Continuous realtime gesture following and recognition*. In Gesture in embodied communication and human-computer interaction, pages 73–84. Springer, 2009. *pas de citation*

[Bevilacqua *et al.* 2011] Frédéric Bevilacqua, Norbert Schnell, Nicolas Rasamimanana, Bruno Zamborlin et Fabrice Guédy. *Online gesture analysis and control of audio processing*. In Musical Robots and Interactive Multimodal Systems, pages 127–142. Springer, 2011. *pas de citation*

[Boehme *et al.* 1998] Hans-Joachim Boehme, Anja Brakensiek, Ulf-Dietrich Braumann, Markus Krabbes et Horst-Michael Gross. *Neural networks for gesture-based remote control of a mobile robot*. In Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, volume 1, pages 372–377. IEEE, 1998. *pas de citation*

[Bolt 1980] Richard A Bolt. "put-that-there" : Voice and gesture at the graphics interface, volume 14. ACM, 1980. *pas de citation*

[Bressem & Ladewig 2011] Jana Bressem et Silva H Ladewig. *Rethinking gesture phases : Articulatory features of gestural movement ?* Semiotica, vol. 2011, no. 184, pages 53–91, 2011. *pas de citation*

[Brooke *et al.* 1996] John Brooke*et al. SUS-A quick and dirty usability scale*. Usability evaluation in industry, vol. 189, no. 194, pages 4–7, 1996. *pas de citation*

[Calbris 1990] Geneviève Calbris. The semiotics of french gestures, volume 1900. Indiana Univ Pr, 1990. *pas de citation*

[Cameron *et al.* 1987] Jonathan M Cameron, Brian K Cooper, Robert A Salo et Brian H Wilcox. *Fusing global navigation with computer-aided remote driving of robotic vehicles*. In Cambridge Symposium_Intelligent Robotics Systems, pages 85–89. International Society for Optics and Photonics, 1987. *pas de citation*

[Canan 1999] James W Canan. *Seeing more, and risking less, with UAVs*. Aerospace America, vol. 37, no. 10, pages 26–31, 1999. *pas de citation*

[Candan *et al.* 2012] K Selçuk Candan, Rosaria Rossini, Xiaolan Wang et Maria Luisa Sapino. *sDTW : computing DTW distances using locally relevant constraints based on salient feature alignments*. Proceedings of the VLDB Endowment, vol. 5, no. 11, pages 1519–1530, 2012. *pas de citation*

[Carmona & Climent 2012] Josep Maria Carmona et Joan Climent. *A performance evaluation of hmm and dtw for gesture recognition*. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pages 236–243. Springer, 2012. *pas de citation*

[Carreiro & Burke 2006] Louis G Carreiro et A Alan Burke. *Unmanned underwater vehicles*. In 7th Annual Solid State Energy Conversion Alliance (SECA) Workshop and Peer Review, Department of Energy, Philadelphia, PA, 2006. *pas de citation*

[Choe *et al.* 2010] BongWhan Choe, Jun-Ki Min et Sung-Bae Cho. *Online gesture recognition for user interface on accelerometer built-in mobile phones*. In Neural Information Processing. Models and Applications, pages 650–657. Springer, 2010. *pas de citation*

[Choi *et al.* 2012] Eunjung Choi, Sunghyuk Kwon, Donghun Lee, Hojin Lee et Min K Chung. *Can user-derived gesture be considered as the best gesture for a command ? : Focusing on the commands for smart home system*. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 56, pages 1253–1257. SAGE Publications, 2012. *pas de citation*

[Chomsky 1956] Noam Chomsky. *Three models for the description of language*. Information Theory, IRE Transactions on, vol. 2, no. 3, pages 113–124, 1956. *pas de citation*

[Claassen *et al.* 1990] Wim Claassen, Edwin Bos et Carla Huls. *The Pooh way in human-computer interaction : Towards multimodal interfaces*. SPIN/MMC Research Report, no. 5, 1990. *pas de citation*

[Clark 1996] Herbert H Clark. Using language. Cambridge university press, 1996. *pas de citation*

[Cockton *et al.* 2008] Gilbert Cockton, Alan Woolrych, Darryn Lavery, A Sears, J Jacko, I Tsuchiya et G Grandy. *Inspection based evaluations*. 2008. *pas de citation*

[Corradini & Gross 2000] Andrea Corradini et Horst-Michael Gross. *Camera-based gesture recognition for robot control*. In Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, volume 4, pages 133–138. IEEE, 2000. *pas de citation*

[Coupeté *et al.* 2016] Eva Coupeté, Fabien Moutarde, Sotiris Manitsaris et Olivier Hugues. *Recognition of Technical Gestures for Human-Robot Collaboration in Factories*. In The Ninth International Conference on Advances in Computer-Human Interactions, 2016. *pas de citation*

[Coutaz *et al.* 1996] Joëlle Coutaz, Laurence Nigay et Daniel Salber. *Agent-based architecture modelling for interactive systems*. In Critical issues in user interface systems engineering, pages 191–209. Springer, 1996. *pas de citation*

[Dasiopoulou *et al.* 2011] Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti et Yiannis Kompatsiaris. *A survey of semantic image and video annotation tools*. In Knowledge-driven multimedia information extraction and ontology evolution, pages 196–239. Springer, 2011. *pas de citation*

[Davis & Shah 1994] James Davis et Mubarak Shah. *Visual gesture recognition*. In Vision, Image and Signal Processing, IEE Proceedings-, volume 141, pages 101–106. IET, 1994. *pas de citation*

[Demazeau & Müller 1991] Yves Demazeau et J-P Müller. Decentralized ai, 2. Elsevier, 1991. *pas de citation*

[Dempster *et al.* 1977] Arthur P Dempster, Nan M Laird et Donald B Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977. *pas de citation*

[Dos Santos Souza 1999] Ivani Dos Santos Souza. *Quand les gestes deviennent une proto-langue*. Analyse globale descriptive du lexique et des échanges interactionnels d'un sourd brésilien. Mémoire de DEA en Sciences du Langage, Université Paris VIII, 1999. *pas de citation*

[Dumas & Redish 1999] Joseph S Dumas et Janice Redish. A practical guide to usability testing. Intellect Books, 1999. *pas de citation*

[Dupont & Marteau 2015] Marc Dupont et Pierre-François Marteau. *Coarse-DTW : Exploiting Sparsity in Gesture Time Series*. In Advanced Analytics and Learning on Temporal Data (AALTD), 2015. *pas de citation*

[Efron 1941] David Efron. *Gesture and environment.* 1941. *pas de citation*

[Fikkert 2010] Fredrik Willem Fikkert. Gesture interaction at a distance. University of Twente, Centre for Telematics and Information Technology, 2010. *pas de citation*

[Fong & Thorpe 2001] Terrence Fong et Charles Thorpe. *Vehicle teleoperation interfaces*. Autonomous robots, vol. 11, no. 1, pages 9–18, 2001. *pas de citation*

[Fong *et al.* 2001] Terrence W Fong, François Conti, Sébastien Grange et Charles Baur. *Novel interfaces for remote driving : gesture, haptic, and PDA*. In Intelligent Systems and Smart Manufacturing, pages 300–311. International Society for Optics and Photonics, 2001. *pas de citation*

[Fong *et al.* 2003] Terrence Fong, Charles Thorpe et Betty Glass. *Pdadriver : A handheld system for remote driving*. In IEEE international conference on advanced robotics, numéro LSRO2-CONF-2003-004, 2003. *pas de citation*

[Forney Jr 1973] G David Forney Jr. *The viterbi algorithm*. Proceedings of the IEEE, vol. 61, no. 3, pages 268–278, 1973. *pas de citation*

[Fothergill *et al.* 2012] Simon Fothergill, Helena Mentis, Pushmeet Kohli et Sebastian Nowozin. *Instructing people for training gestural interactive systems*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1737–1746. ACM, 2012. *pas de citation*

[Franklin & Graesser 1996] Stan Franklin et Art Graesser. *Is it an Agent, or just a Program ? : A Taxonomy for Autonomous Agents*. In Intelligent agents III agent theories, architectures, and languages, pages 21–35. Springer, 1996. *pas de citation*

[Friedman 1976] Lynn Friedman. *Phonology of a soundless language : phonological structure of the American Sign Language*. 1976. *pas de citation*

[Frohlich] DM Frohlich. *The Design Space of Interfaces : Multimedia Systems, Interaction and Applications*. In Proceedings 1st Eurographics Workshop, Stockholm Sweden. *pas de citation*

[Frohlich & Luff 1990] David Frohlich et Paul Luff. *Applying the technology of conversation to the technology for conversation*. Computers and conversation, pages 187–220, 1990. *pas de citation*

[Gage 1995] Douglas W Gage. *UGV history 101 : A brief history of Unmanned Ground Vehicle (UGV) development efforts*. Rapport technique, DTIC Document, 1995. *pas de citation*

[Gazette 1989] Marine Corps Gazette. *The Changing Face of War : Into the Fourth Generation William S. Lind, Colonel Keith Nightengale (USA), Captain John F. Schmitt (USMC), Colonel Joseph W. Sutton (USA), and Lieutenant Colonel Gary I. Wilson (USMCR)*. Marine Corps Gazette, pages 22–26, 1989. *pas de citation*

[Gerwing & Bavelas 2004] Jennifer Gerwing et Janet Bavelas. *Linguistic influences on gesture's form*. Gesture, vol. 4, no. 2, pages 157–195, 2004. *pas de citation*

[Godenschweger & Strothotte 1998] Frank Godenschweger et Thomas Strothotte. *Modeling and generating sign language as animated line drawings*. In Proceedings of the third international ACM conference on Assistive technologies, pages 78–84. ACM, 1998. *pas de citation*

[Goldin-Meadow 1993] Susan Goldin-Meadow. *When does gesture become language ? A study of gesture used as a primary communication system by deaf children of hearing parents*. Tools, language and cognition in human evolution, pages 63–85, 1993. *pas de citation*

[Gray 2012] Colin S Gray. Another bloody century : Future warfare. Hachette UK, 2012. *pas de citation*

[Grosjean & Lane 1977] François Grosjean et Harlan Lane. *Pauses and syntax in American sign language*. Cognition, vol. 5, no. 2, pages 101–117, 1977. *pas de citation*

[Guo & Sharlin 2008] Cheng Guo et Ehud Sharlin. *Exploring the use of tangible user interfaces for human-robot interaction : a comparative study*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 121–130. ACM, 2008. *pas de citation*

[Han *et al.* 2016] Fei Han, Brian Reily, William Hoff et Hao Zhang. *Space-Time Representation of People Based on 3D Skeletal Data : A Review*. arXiv preprint arXiv :1601.01006, 2016. *pas de citation*

[Holler & Wilkin 2011] Judith Holler et Katie Wilkin. *Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue*. Journal of Nonverbal Behavior, vol. 35, no. 2, pages 133–153, 2011. *pas de citation*

[Hummels *et al.* 1997] Caroline Hummels, Gerda Smets et Kees Overbeeke. *An intuitive two-handed gestural interface for computer supported product design*. In Gesture and Sign Language in Human-Computer Interaction, pages 197–208. Springer, 1997. *pas de citation*

[Hutchins *et al.* 1985] Edwin L Hutchins, James D Hollan et Donald A Norman. *Direct manipulation interfaces*. Human–Computer Interaction, vol. 1, no. 4, pages 311–338, 1985. *pas de citation*

[Hutchins 1987] Edwin Hutchins. *Metaphors for interface design*. Rapport technique, DTIC Document, 1987. *pas de citation*

[Iba *et al.* 1999] Soshi Iba, J Michael Vande Weghe, Christiaan JJ Paredis et Pradeep K Khosla. *An architecture for gesture-based control of mobile robots*. In Intelligent Robots and Systems, 1999. IROS'99. Proceedings. 1999 IEEE/RSJ International Conference on, volume 2, pages 851–857. IEEE, 1999. *pas de citation*

[ISO 2000] ISO. Software product quality - quality model, iso/iec 9126. International Organization for Standardization, 2000. *pas de citation*

[Itakura 1975] Fumitada Itakura. *Minimum prediction residual principle applied to speech recognition*. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 23, no. 1, pages 67–72, 1975. *pas de citation*

[Javed *et al.* 2011] Waqas Javed, Niklas Elmqvist, Ji Soo Yi*et al.* *Direct manipulation through surrogate objects*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 627–636. ACM, 2011. *pas de citation*

[Jégo *et al.* 2013] Jean-François Jégo, Alexis Paljic et Philippe Fuchs. *User-defined gestural interaction : A study on gesture memorization*. In 3D User Interfaces (3DUI), 2013 IEEE Symposium on, pages 7–10. IEEE, 2013. *pas de citation*

[Jones *et al.* 2010] Geraint Jones, Nadia Berthouze, Roman Bielski et Simon Julie. *Towards a situated, multimodal interface for multiple UAV control*. In Robotics and Automation (ICRA), 2010 IEEE International Conference on, pages 1739–1744. IEEE, 2010. *pas de citation*

[Junker *et al.* 2008] Holger Junker, Oliver Amft, Paul Lukowicz et Gerhard Tröster. *Gesture spotting with body-worn inertial sensors to detect user activities*. Pattern Recognition, vol. 41, no. 6, pages 2010–2024, 2008. *pas de citation*

[Kahneman 1973] Daniel Kahneman. Attention and effort. Citeseer, 1973. *pas de citation*

[Kahol *et al.* 2004] Kanav Kahol, Priyamvada Tripathi et Sethuraman Panchanathan. *Automated gesture segmentation from dance sequences*. In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pages 883–888. IEEE, 2004. *pas de citation*

[Karam *et al.* 2006] Maria Karam*et al. Investigating user tolerance for errors in vision-enabled gesture-based interactions*. In Proceedings of the working conference on Advanced visual interfaces, pages 225–232. ACM, 2006. *pas de citation*

[Karam 2006] Maria Karam. *PhD Thesis : A framework for research and design of gesture-based human-computer interactions*. PhD thesis, University of Southampton, 2006. *pas de citation*

[Kay 1995] Jennifer S Kay. *STRIPE : Remote driving using limited image data*. In Conference Companion on Human Factors in Computing Systems, pages 59–60. ACM, 1995. *pas de citation*

[Kendon 1980] Adam Kendon. *Gesticulation and speech : Two aspects of the process of utterance*. The relationship of verbal and nonverbal communication, vol. 25, pages 207–227, 1980. *pas de citation*

[Kendon 1986] Adam Kendon. *Current issues in the study of gesture*. The biological foundations of gestures : Motor and semiotic aspects, vol. 1, pages 23–47, 1986. *pas de citation*

[Kendon 1988a] Adam Kendon. *How gestures can become like words*. Cross-cultural perspectives in nonverbal communication, vol. 1, pages 131–141, 1988. *pas de citation*

[Kendon 1988b] Adam Kendon. Sign languages of aboriginal australia : Cultural, semiotic and communicative perspectives. Cambridge University Press, 1988. *pas de citation*

[Kendon 1996] Adam Kendon. *An agenda for gesture studies*. Semiotic review of books, vol. 7, no. 3, pages 8–12, 1996. *pas de citation*

[Kendon 2004] Adam Kendon. Gesture : Visible action as utterance. Cambridge University Press, 2004. *pas de citation*

[Kendon 2007] Adam Kendon. *On the origins of modern gesture studies*. Gesture and the dynamic dimension of language, pages 13–28, 2007. *pas de citation*

[Keogh & Pazzani 1999] Eamonn J Keogh et Michael J Pazzani. *Scaling up dynamic time warping to massive datasets*. In Principles of Data Mining and Knowledge Discovery, pages 1–11. Springer, 1999. *pas de citation*

[Kessler *et al.* 1995] G Drew Kessler, Larry F Hodges et Neff Walker. *Evaluation of the CyberGlove as a whole-hand input device*. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 2, no. 4, pages 263–283, 1995. *pas de citation*

[Kita *et al.* 1997] Sotaro Kita, Ingeborg Van Gijn et Harry Van der Hulst. *Movement phases in signs and co-speech gestures, and their transcription by human coders*. In Gesture and sign language in human-computer interaction, pages 23–35. Springer, 1997. *pas de citation*

[Kortenkamp *et al.* 1996] David Kortenkamp, Eric Huber, R Peter Bonasso*et al. Recognizing and interpreting gestures on a mobile robot*. In Proceedings of the National Conference on Artificial Intelligence, pages 915–921, 1996. *pas de citation*

[Kurtenbach & Hulteen 1990] Gordon Kurtenbach et Eric A Hulteen. *Gestures in human-computer communication*. The art of human-computer interface design, pages 309–317, 1990. *pas de citation*

[Lamberti & Camastra 2011] Luigi Lamberti et Francesco Camastra. *Real-time hand gesture recognition using a color glove*. In Image Analysis and Processing–ICIAP 2011, pages 365–373. Springer, 2011. *pas de citation*

[Laurel 1986] Brenda Laurel. *Interface as mimesis*. User centered system design : New perspectives on human-computer interaction, pages 67–85, 1986. *pas de citation*

[Lausberg & Sloetjes 2009] Hedda Lausberg et Han Sloetjes. *Coding gestural behavior with the NEUROGES-ELAN system*. Behavior research methods, vol. 41, no. 3, pages 841–849, 2009. *pas de citation*

[Lee & Kim 1999] Hyeon-Kyu Lee et Jin H Kim. *An HMM-based threshold model approach for gesture recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 21, no. 10, pages 961–973, 1999. *pas de citation*

[Lee & Xu 1996] Christopher Lee et Yangsheng Xu. *Online, interactive learning of gestures for human/robot interfaces*. In Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on, volume 4, pages 2982–2987. IEEE, 1996. *pas de citation*

[Leplat 2005] Jacques Leplat. *Les automatismes dans l'activité : pour une réhabilitation et un bon usage*. Activités, vol. 2, no. 2-2, 2005. *pas de citation*

[Leroi-Gourhan 2013] André Leroi-Gourhan. Le geste et la parole- : Technique et langage, volume 1. Albin Michel, 2013. *pas de citation*

[Lewis 1995] James R Lewis. *IBM computer usability satisfaction questionnaires : psychometric evaluation and instructions for use*. International Journal of Human-Computer Interaction, vol. 7, no. 1, pages 57–78, 1995. *pas de citation*

[Liddell 1977] Scott Kent Liddell. An investigation into the syntatic structure of american sign language. UMI, 1977. *pas de citation*

[Liu *et al.* 2009] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya et Venu Vasudevan. *uWave : Accelerometer-based personalized gesture recognition and its applications*. Pervasive and Mobile Computing, vol. 5, no. 6, pages 657–675, 2009. *pas de citation*

[Marec 2013] Jean-Pierre Marec. *Réflexions sur la robotique militaire*, 2013. *pas de citation*

[Martinic 2014] Gary Martinic. *Les technologies terrestres robotisées, foisonnantes et aussi variées qu'utiles*. Revue militaire canadienne, vol. 14, no. 4, 2014. *pas de citation*

[Mashood *et al.* 2015] Ahmed Mashood, Hassan Noura, Imad Jawhar et Nader Mohamed. *A gesture based kinect for quadrotor control*. In Information and Communication Technology Research (ICTRC), 2015 International Conference on, pages 298–301. IEEE, 2015. *pas de citation*

[McGovern 1991] Douglas E McGovern. *Experience and results in teleoperation of land vehicles*. In Pictorial communication in virtual and real environments, pages 182–195. Taylor & Francis, Inc., 1991. *pas de citation*

[McNeill 1985] David McNeill. *So you think gestures are nonverbal ?* Psychological review, vol. 92, no. 3, page 350, 1985. *pas de citation*

[McNeill 1992] David McNeill. Hand and mind : What gestures reveal about thought. University of Chicago press, 1992. *pas de citation*

[McNeill 2000] David McNeill. Language and gesture, volume 2. Cambridge University Press, 2000. *pas de citation*

[Monajjemi *et al.* 2013] Valiallah Mani Monajjemi, Jens Wawerla, Rodney Vaughan et Greg Mori. *Hri in the sky : Creating and commanding teams of uavs with a vision-mediated gestural interface*. In Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, pages 617–623. IEEE, 2013. *pas de citation*

[Morris & Ochs 1997] Desmond Morris et Edith Ochs. Le langage des gestes : un guide international. Calmann-Lévy, 1997. *pas de citation*

[Myers & Rabiner 1981] Cory S Myers et Lawrence R Rabiner. *A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition*. Bell System Technical Journal, vol. 60, no. 7, pages 1389–1409, 1981. *pas de citation*

[Nahapetyan & Khachumov 2015] VE Nahapetyan et VM Khachumov. *Gesture recognition in the problem of contactless control of an unmanned aerial vehicle*. Optoelectronics, Instrumentation and Data Processing, vol. 51, no. 2, pages 192–197, 2015. *pas de citation*

[Navon & Gopher 1979] David Navon et Daniel Gopher. *On the economy of the human-processing system.* Psychological review, vol. 86, no. 3, page 214, 1979. *pas de citation*

[Ng & Sharlin 2011] Wai Shan Ng et Ehud Sharlin. *Collocated interaction with flying robots*. In RO-MAN, 2011 IEEE, pages 143–149. IEEE, 2011. *pas de citation*

[Nielsen *et al.* 2003] Michael Nielsen, Moritz Störring, Thomas B Moeslund et Erik Granum. *A procedure for developing intuitive and ergonomic gesture interfaces for HCI*. In Gesture-Based Communication in Human-Computer Interaction, pages 409–420. Springer, 2003. *pas de citation*

[Nielsen 1994] Jakob Nielsen. Usability engineering. Elsevier, 1994. *pas de citation*

[Nielsen 1995] Jakob Nielsen. *10 usability heuristics for user interface design*. Fremont : Nielsen Norman Group.[Consult. 20 maio 2014]. Disponível na Internet, 1995. *pas de citation*

[Norman 1988] Donald A Norman. The psychology of everyday things. Basic books, 1988. *pas de citation*

[Pashler 1994] Harold Pashler. *Dual-task interference in simple tasks : data and theory.* Psychological bulletin, vol. 116, no. 2, page 220, 1994. *pas de citation*

[Perruchet 1988] Pierre Perruchet. Les automatismes cognitifs, volume 174. Editions Mardaga, 1988. *pas de citation*

[Perzanowski *et al.* 1998] Dennis Perzanowski, Alan C Schultz et William Adams. *Integrating natural language and gesture in a robotics domain*. In Intelligent Control (ISIC), 1998. Held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), Intelligent Systems and Semiotics (ISAS), Proceedings, pages 247–252. IEEE, 1998. *pas de citation*

[Perzanowski *et al.* 2001] Dennis Perzanowski, Alan C Schultz, William Adams, Elaine Marsh et Magda Bugajska. *Building a multimodal human-robot interface*. Intelligent Systems, IEEE, vol. 16, no. 1, pages 16–21, 2001. *pas de citation*

[Pfeiffer *et al.* 2016] Thies Pfeiffer, Aljoscha Schmidt et Patrick Renner. *Detecting Movement Patterns from Inertial Data of a Mobile Head-Mounted-Display for Navigation via Walking-in-Place*. IEEE Virtual Reality 2016, 2016. *pas de citation*

[Pfeil *et al.* 2013] Kevin Pfeil, Seng Lee Koh et Joseph LaViola. *Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles*. In Proceedings of the 2013 international conference on Intelligent user interfaces, pages 257–266. ACM, 2013. *pas de citation*

[Quek *et al.* 2002] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough et Rashid Ansari. *Multimodal human discourse : gesture and speech*. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 9, no. 3, pages 171–193, 2002. *pas de citation*

[Quek 1995] Francis KH Quek. *Eyes in the interface*. Image and vision computing, vol. 13, no. 6, pages 511–525, 1995. *pas de citation*

[Rabiner & Juang 1986] Lawrence R Rabiner et Biing-Hwang Juang. *An introduction to hidden Markov models*. ASSP Magazine, IEEE, vol. 3, no. 1, pages 4–16, 1986. *pas de citation*

[Rimé & Schiaratura 1991] Bernard Rimé et Loris Schiaratura. *Gesture and speech*. 1991. *pas de citation*

[Rogalla *et al.* 2002] O Rogalla, M Ehrenmann, R Zöllner, R Becher et R Dillmann. *Using gesture and speech control for commanding a robot assistant*. In Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on, pages 454–459. IEEE, 2002. *pas de citation*

[Sakoe & Chiba 1978] Hiroaki Sakoe et Seibi Chiba. *Dynamic programming algorithm optimization for spoken word recognition*. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 26, no. 1, pages 43–49, 1978. *pas de citation*

[Sala *et al.* 1995] Sergio Della Sala, Alan Baddeley, Costanza Papagno et Hans Spinnler. *Dual-task paradigm : a means to examine the central executive*. Annals of the New York Academy of Sciences, vol. 769, no. 1, pages 161–172, 1995. *pas de citation*

[Sanna *et al.* 2013] Andrea Sanna, Fabrizio Lamberti, Gianluca Paravati et Federico Manuri. *A Kinect-based natural interface for quadrotor control*. Entertainment Computing, vol. 4, no. 3, pages 179–186, 2013. *pas de citation*

[Seyfeddinipur 2006] Mandana Seyfeddinipur. Disfluency : Interrupting speech and gesture. MPI-Series in Psycholinguistics, 2006. *pas de citation*

[Sheridan 1992] Thomas B Sheridan. Telerobotics, automation, and human supervisory control. MIT press, 1992. *pas de citation*

[Shneiderman & Maes 1997] Ben Shneiderman et Pattie Maes. *Direct manipulation vs. interface agents*. interactions, vol. 4, no. 6, pages 42–61, 1997. *pas de citation*

[Shneiderman & Plaisant 1987] Ben Shneiderman et Catherine Plaisant. *Designing the user interface : Strategies for effective human-computer interaction*, 1987. *pas de citation*

[Shneiderman 1981] Ben Shneiderman. *Direct manipulation : A step beyond programming languages*. In ACM SIGSOC Bulletin, volume 13, page 143. ACM, 1981. *pas de citation*

[Shneiderman 1982] Ben Shneiderman. *The future of interactive systems and the emergence of direct manipulation*†. Behaviour & Information Technology, vol. 1, no. 3, pages 237–256, 1982. *pas de citation*

[Shun-Chiu 1992] Yau Shun-Chiu. Creation gestuelle et debuts du langage : creation de langues gestuelles chez des sourds isoles. Langages croisés, 1992. *pas de citation*

[Singer 2002] Robert N Singer. *Sport Psychology*. Journal of sport & exercise psychology, vol. 24, pages 359–375, 2002. *pas de citation*

[Slater *et al.* 1995] Mel Slater, Martin Usoh et Anthony Steed. *Taking steps : the influence of a walking technique on presence in virtual reality*. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 2, no. 3, pages 201–219, 1995. *pas de citation*

[Sparrell 1993] Carlton James Sparrell. *Coverbal iconic gesture in human-computer interaction*. PhD thesis, Massachusetts Institute of Technology, 1993. *pas de citation*

[Stern *et al.* 2004] Helman I Stern, Juan P Wachs et Yael Edan. *Hand gesture vocabulary design : a multicriteria optimization*. In Systems, Man and Cybernetics, 2004 IEEE International Conference on, volume 1, pages 19–23. IEEE, 2004. *pas de citation*

[Stern *et al.* 2008a] Helman I Stern, Juan P Wachs et Yael Edan. *Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors*. International Journal of Semantic Computing, vol. 2, no. 01, pages 137–160, 2008. *pas de citation*

[Stern *et al.* 2008b] Helman I Stern, Juan P Wachs et Yael Edan. *Optimal consensus intuitive hand gesture vocabulary design*. In Semantic Computing, 2008 IEEE International Conference on, pages 96–103. IEEE, 2008. *pas de citation*

[Stiefelhagen *et al.* 2004] Rainer Stiefelhagen, Christian Fügen, Petra Gieselmann, Hartwig Holzapfel, Kai Nickel et Alex Waibel. *Natural human-robot interaction using speech, head pose and gestures*. In Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, volume 3, pages 2422–2427. IEEE, 2004. *pas de citation*

[Stokoe 1960] William C Stokoe. *Sign language structure (Studies in Linguistics*. Occasional paper, vol. 8, 1960. *pas de citation*

[Stokoe 2005] William C Stokoe. *Sign language structure : An outline of the visual communication systems of the American deaf*. Journal of deaf studies and deaf education, vol. 10, no. 1, pages 3–37, 2005. *pas de citation*

[Strayer & Johnston 2001] David L Strayer et William A Johnston. *Driven to distraction : Dual-task studies of simulated driving and conversing on a cellular telephone*. Psychological science, vol. 12, no. 6, pages 462–466, 2001. *pas de citation*

[Sturman & Zeltzer 1994] David J Sturman et David Zeltzer. *A survey of glove-based input*. Computer Graphics and Applications, IEEE, vol. 14, no. 1, pages 30–39, 1994. *pas de citation*

[Thinh *et al.* 2016] Nguyen Truong Thinh, Nguyen Tan Viet Tuyen et Dang Thai Son. *Gait of Quadruped Robot and Interaction Based on Gesture Recognition*. Journal of Automation and Control Engineering Vol, vol. 4, no. 1, 2016. *pas de citation*

[Tognazzini 1993] Bruce Tognazzini. *Principles, techniques, and ethics of stage magic and their application to human interface design*. In Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, pages 355–362. ACM, 1993. *pas de citation*

[Triesch & Von Der Malsburg 1998] Jochen Triesch et Christoph Von Der Malsburg. *A gesture interface for human-robot-interaction*. In fg, page 546. IEEE, 1998. *pas de citation*

[Urban *et al.* 2004] Martin Urban, Peter Bajcsy, Rob Kooper et Jean-Chistophe Lementec. *Recognition of arm gestures using multiple orientation sensors : Repeatability assessment*. In Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on, pages 553–558. IEEE, 2004. *pas de citation*

[Usoh *et al.* 1999] Martin Usoh, Kevin Arthur, Mary C Whitton, Rui Bastos, Anthony Steed, Mel Slater et Frederick P Brooks Jr. *Walking> walking-in-place> flying, in virtual environments*. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 359–364. ACM Press/Addison-Wesley Publishing Co., 1999. *pas de citation*

[Van den Bergh *et al.* 2011] Michael Van den Bergh, Daniel Carton, Roderick De Nijs, Nikos Mitsou, Christian Landsiedel, Kolja Kuehnlenz, Dirk Wollherr, Luc Van Gool et Martin Buss. *Real-time 3D hand gesture interaction with a robot for understanding directions from humans*. In RO-MAN, 2011 IEEE, pages 357–362. IEEE, 2011. *pas de citation*

[Vatavu & Wobbrock 2015] Radu-Daniel Vatavu et Jacob O Wobbrock. *Formalizing agreement analysis for elicitation studies : new measures, significance test, and toolkit*. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 1325–1334. ACM, 2015. *pas de citation*

[Vertut 2013] Jean Vertut. Teleoperation and robotics : applications and technology, volume 3. Springer Science & Business Media, 2013. *pas de citation*

[Viterbi 1967] Andrew J Viterbi. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. Information Theory, IEEE Transactions on, vol. 13, no. 2, pages 260–269, 1967. *pas de citation*

[Waldherr *et al.* 2000] Stefan Waldherr, Roseli Romero et Sebastian Thrun. *A gesture based interface for human-robot interaction*. Autonomous Robots, vol. 9, no. 2, pages 151–173, 2000. *pas de citation*

[Wang & Popović 2009] Robert Y Wang et Jovan Popović. *Real-time hand-tracking with a color glove*. ACM transactions on graphics (TOG), vol. 28, no. 3, page 63, 2009. *pas de citation*

[Wexelblat 1994] Alan Daniel Wexelblat. *A feature-based approach to continuous-gesture analysis*. PhD thesis, Massachusetts Institute of Technology, 1994. *pas de citation*

[Wexelblat 1997] Alan Wexelblat. *Research challenges in gesture : Open issues and unsolved problems*. In Gesture and sign language in human-computer interaction, pages 1–11. Springer, 1997. *pas de citation*

[Wickens 1980] Christopher D Wickens. *The structure of attentional resources*. Attention and performance VIII, vol. 8, 1980. *pas de citation*

[Widmer *et al.* 2014] Antoine Widmer, Roger Schaer, Dimitrios Markonis et Henning Müller. *Gesture Interaction for Content–based Medical Image Retrieval*. In Proceedings of International Conference on Multimedia Retrieval, page 503. ACM, 2014. *pas de citation*

[Wilcox & Bush 1992] Lynn D Wilcox et Marcia A Bush. *Training and search algorithms for an interactive wordspotting system*. In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, volume 2, pages 97–100. IEEE, 1992. *pas de citation*

[Wilson & Wilson 2004] Daniel Wilson et Andy Wilson. *Gesture recognition using the xwand*. 2004. *pas de citation*

[Wobbrock *et al.* 2005] Jacob O Wobbrock, Htet Htet Aung, Brandon Rothrock et Brad A Myers. *Maximizing the guessability of symbolic input*. In CHI'05 extended abstracts on Human Factors in Computing Systems, pages 1869–1872. ACM, 2005. *pas de citation*

[Wobbrock *et al.* 2009] Jacob O Wobbrock, Meredith Ringel Morris et Andrew D Wilson. *User-defined gestures for surface computing*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1083–1092. ACM, 2009. *pas de citation*

[Yang *et al.* 2013] Shuai Yang, Prashan Premaratne et Peter Vial. *Hand gesture recognition : An overview*. In Broadband Network & Multimedia Technology (IC-BNMT), 2013 5th IEEE International Conference on, pages 63–69. IEEE, 2013. *pas de citation*

[Zhang & Liu 2015] Jiali Zhang et Guixi Liu. *Dynamic gesture recognition and human-computer interaction*. 2015. *pas de citation*

[Zhu *et al.* 2002] Yuanxin Zhu, Guangyou Xu et David J Kriegman. *A real-time approach to the spotting, representation, and recognition of hand gestures for human–computer interaction*. Computer Vision and Image Understanding, vol. 85, no. 3, pages 189–208, 2002. *pas de citation*

ANNEXE A

Catalogue des gestes recueillis

Lors de la première phase du protocole permettant la construction d'un vocabulaire gestuel consensuel et non ambiguë (Chapitre **??**), 20 participants proposent chacun un geste pour chacune des 8 fonctions considérées (Décoller, Atterrir, Suivant, Précédent, Base, Stop, Valider et Annuler).

Il arrive alors que plusieurs participants proposent le même geste. Après analyse, les gestes identiques sont regroupés et forment alors un geste candidat.

Un catalogue est constitué à partir de l'ensemble des gestes candidats. u nombre de 46 (sur 160 gestes proposés initialement), l'ensemble de ces gestes et présenté dans les figurent qui suivent. Ils sont organisés par fonction : le décollage Figure A.1, l'atterrissage Figure A.2, aller au point suivant Figure A.3, revenir au point précédent Figure A.4, aller à la base Figure A.5, interrompre le déplacement A.6, valider une commande Figure A.7 et annuler une commande Figure A.8.

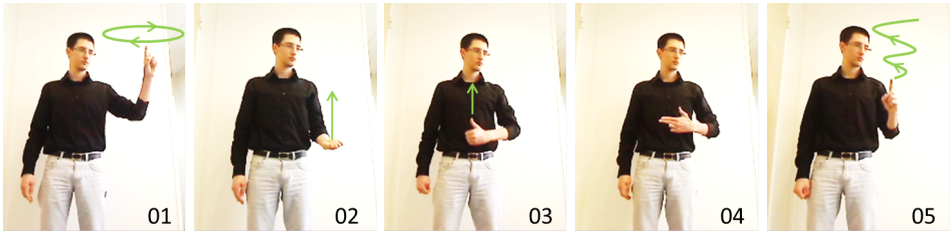


FIGURE A.1 – Gestes proposés pour la commande **Décoller**. Le geste élu est le geste portant le numéro 02.

FIGURE A.2 – Gestes proposés pour la commande **Atterrir**. Le geste élu est le geste portant le numéro 06.

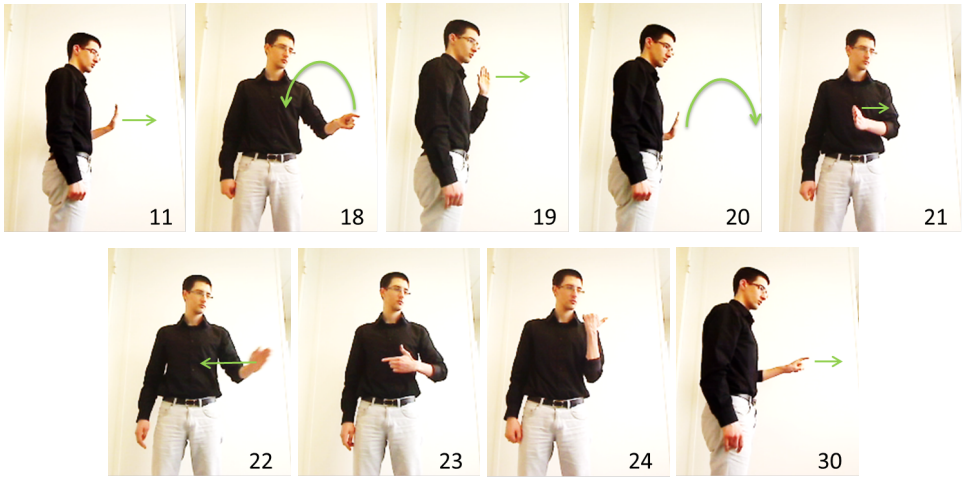


FIGURE A.3 – Gestes proposés pour la commande **Suivant**. Le geste élu est le geste portant le numéro 19.

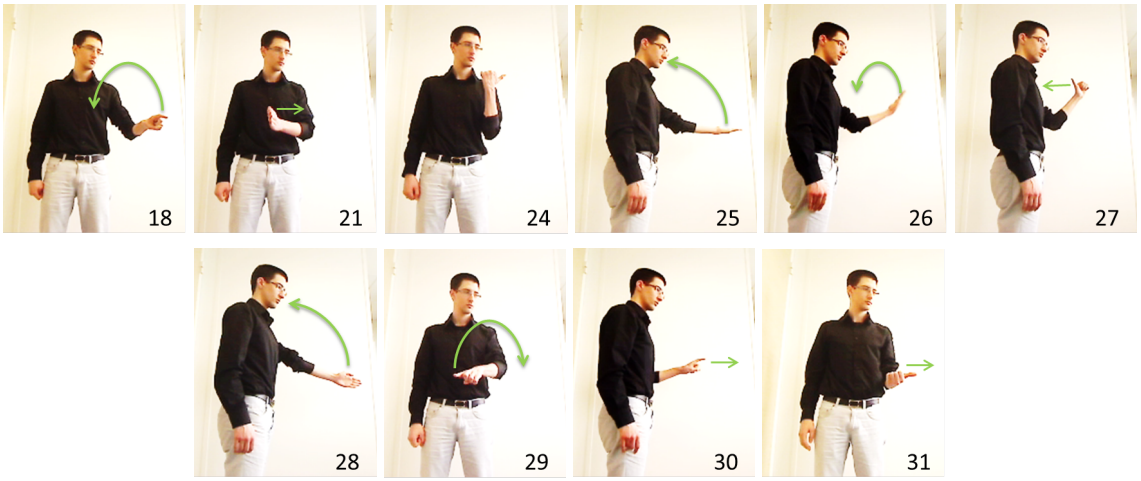


FIGURE A.4 – Gestes proposés pour la commande **Précédent**. Le geste élu est le geste portant le numéro 27.

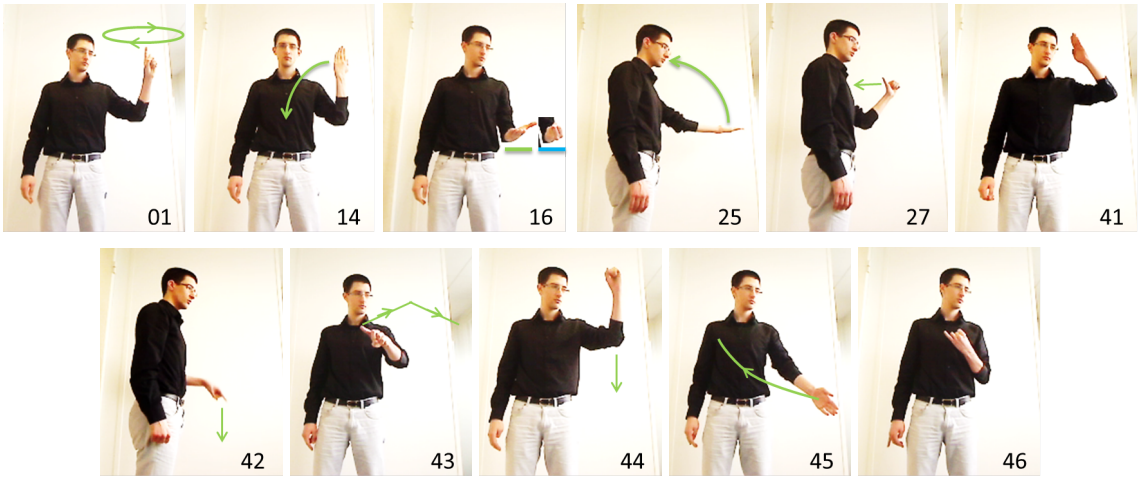


FIGURE A.5 – Gestes proposés pour la commande **Base**. Le geste élu est le geste portant le numéro 44.

FIGURE A.6 – Gestes proposés pour la commande **Stop**. Le geste élu est le geste portant le numéro 13.

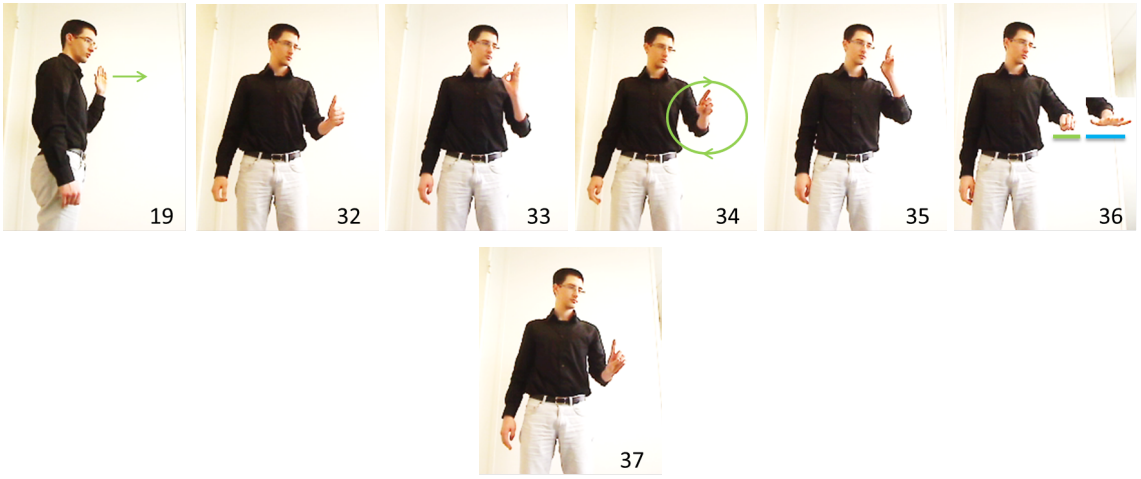


FIGURE A.7 – Gestes proposés pour la commande **Validation**. Le geste élu est le geste portant le numéro 32.

FIGURE A.8 – Gestes proposés pour la commande **Annulation**. Le geste élu est le geste portant le numéro 39.

Résumé

Utiliser une interface visuo-tactile peut être une gêne lorsqu'il est nécessaire de rester mobile et conscient de son environnement. Cela s'avère particulièrement problématique en milieu hostile comme pour la commande d'un drone militaire de contact.
Dans ces travaux nous faisons l'hypothèse que le geste est une modalité de commande moins contraignante puisqu'elle n'impose pas de visualiser ni de manipuler une interface physique.
Aussi, nous avons mis en place une démarche centrée utilisateur afin de confirmer d'une part, les avantages pratiques, et d'autre part, la faisabilité technique de la commande d'un robot mobile par geste.
Tout d'abord, l'étude théorique du geste a permis de construire un modèle d'interaction. Celui-ci consiste en l'activation de commandes automatiques par la réalisation de gestes sémaphoriques normalisés. Des messages sonores permettent de renseigner l'opérateur et un mécanisme de confirmation sécurise l'interaction.
Ensuite, le dictionnaire des gestes à utiliser a été constitué. Pour cela, une méthodologie a été proposée et appliquée : des utilisateurs élicitent puis élisent les gestes les plus pertinents.
Notre modèle d'interaction et le vocabulaire gestuel ont ensuite été évalués. Une étude en laboratoire nous a permis de montrer que l'interaction gestuelle telle que proposée est simple à apprendre et utiliser et qu'elle permet de conserver une bonne conscience de l'environnement.
Un système interactif complet a ensuite été développé. Son architecture a été déduite du modèle d'interaction et une brique de reconnaissance gestuelle a été mise en oeuvre. En marge des méthodes classiques, la brique proposée utilise un principe de description formelle des gestes avec une grammaire régulière.
Finalement, ce système a été évalué lors de tests utilisateurs. L'évaluation par des participants militaires a confirmé notre hypothèse de la pertinence du geste pour une interaction visuellement et physiquement moins contraignante.

Mots Clés

IHM,Geste,Robots,Drones

Abstract

Using a visuo-tactil interface may be restrictive when mobility and situation awareness are required. This is particularly problematic in hostile environment as for commanding a drone on a battlefield.
In the work presented here, we hypothesize that gesture is a less restrictive modaility since it doesn't require to manipulate nor to look at a device.
Thus we followed a user-centered approach to confirm practical advantages and technical feasibility of gestural interaction for drones.
First, the theoretical study of gestures allowed us to design an interaction model. It consists on activating commands by executing standardized semaphoric gestures. Sound messages inform the user and a confirmation mechanism secure the interaction.
Second, a gestural vocabulary has been built. To do so, a methodology has been proposed and used : end users elicited then elected the most appropriate gestures.
Then, the interaction model and the selected gestures have been evaluated. A laboratory study showed that they are both easy to learn and use and helps situation awareness.
An interactive system as then been developed. It's architecture has been deducted from our interaction model and a gesture recognizer as been buit. Different from usual methods, the recognizer we proposed is based on formal description of gestures using regular expressions.
Finaly, we conducted a user testing of the proposed system. The evaluation by end-users confirmed our hypothesis that gestures are a modality less distractive both visualy and physicaly.

Keywords

HMI,Gestures,Robots,Drones