IDIAP RESEARCH REPORT

Reconnaissance de gestes 3D bi-manuels

Agnès Just [a] Sébastien Marcel [a]
Olivier Bernier [b] Jean Emmanuel Viallet [b]

IDIAP–RR 03-79

janvier 2004

à paraître dans
Atelier "Acquisition du geste par vision artificielle et applications", RFIA2004

Institut Dalle Molle d'Intelligence Artificielle Perceptive • CP 592 • Martigny • Valais • Suisse

téléphone +41 − 27 − 721 77 11
télécopieur +41 − 27 − 721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

a IDIAP, Martigny, CH-1920
b France Télécom Recherche & Développement, FR-22300 Lannion

Rapport de recherche de l'IDIAP 03-79

Reconnaissance de gestes 3D bi-manuels

Agnès Just   Sébastien Marcel   Olivier Bernier   Jean Emmanuel Viallet

janvier 2004

à paraître dans
Atelier "Acquisition du geste par vision artificielle et applications", RFIA2004

Résumé. Cet article présente une base de seize gestes dynamiques obtenus par suivi de différentes parties colorées du corps, suivi réalisé en temps réel par l'algorithme EM. Ces gestes réalisés avec une ou deux mains sont appris et reconnus avec des HMM. Les prétraitements effectués sur cette base de gestes ainsi que les résultats de classification sont présentés.

1 Introduction

Les interfaces homme machine (IHM) les plus couramment utilisées sont le clavier, la souris et pour les jeux le joystick. L'utilisation de gestes de la main permet d'envisager de communiquer avec un ordinateur de manière plus intuitive. L'observation des gestes à travers des caméras permet à l'utilisateur de s'affranchir du port de capteurs encombrants tels des gants numériques. L'utilisation de marqueurs visuels est une contrainte qui peut être envisagée notamment lors d'une phase préalable d'apprentissage ; en revanche, il est souhaitable de lever cette contrainte lors de l'utilisation réelle d'un système basé sur la reconnaissance de gestes. La reconnaissance visuelle de gestes est plus naturelle mais également plus difficile. Deux aspects de la reconnaissance des gestes doivent être distingués. Le premier, l'aspect statique, caractérise la posture ou la forme spécifique de l'image de la main que l'on cherche à identifier. Le second, l'aspect dynamique, concerne soit la trajectoire de la main soit la succession des postures de la main que l'on cherche à reconnaître. Dans cet article nous traitons de la reconnaissance dynamique de la trajectoire de gestes effectués avec une ou deux mains. Les gestes choisis n'appartiennent ni à un langage tel celui de la langue des signes ni à celui d'un jargon professionnel (alphabet sémaphore des marins ou code des plongeurs). Le vocabulaire a été choisi parmi des signes couramment effectués et reconnus par tous (au moins dans la culture occidentale) et ne nécessitant donc pas ou peu d'apprentissage et de mémorisation de la part d'une nouvelle personne amenée à réaliser ces gestes, si ce n'est de se souvenir des commandes associées aux gestes.

2 Travaux antérieurs

La reconnaissance de gestes dynamiques de la main est un problème de traitement de séquences qui peut être résolu de différentes manières.

Darrell et Pentland [1] utilisent une approche basée vision afin de modéliser les objets et le comportement. Les objets sont représentés par des ensembles de vues. Cette approche autorise l'apprentissage du modèle en utilisant les observations. L'inconvénient de cette approche est que les objets articulés complexes possèdent une gamme d'apparences très étendue. C'est la raison pour laquelle ces auteurs s'appuient sur une interpolation de l'apparence à partir d'un nombre réduit de vues. La classification des gestes est effectuée par appariement dynamique de motifs stéréoscopiques spatiaux-temporels (c'est à dire les gestes) avec des motifs de gestes pré-enregistrés.

Starner et Pentland [2] ont utilisé un vecteur à huit composantes, correspondant aux positions x et y de chaque main, à l'angle de l'axe de plus faible inertie et à l'excentricité de l'ellipse englobante. Des réseaux de HMM leurs ont permis de reconnaître une séquence de gestes issus de l'American Sign Language. L'apprentissage fut effectué en étiquetant avec le signe approprié la partie correspondante du flux vidéo. Une modélisation du langage permet de segmenter l'enchaînement des signes. L'algorithme de décodage de Viterbi fut employé avec et sans la grammaire correspondant à la structure connue des phrases.

Dans [3], Marcel et al. ont utilisé des GIOHMM [1] pour différencier les gestes déictiques des gestes symboliques. Le taux de confusion obtenu en considérant un ensemble constitué de deux gestes déictiques et deux gestes symboliques était inférieur à 16 % . Hong et al. [4] utilisent les positions 2D du centre de la tête et des mains pour reconnaître en temps réel quatre gestes effectués par une seule main. Lee et Kim [5] utilisent un vocabulaire de 10 gestes réalisés avec une seule main dans le plan de l'image pour piloter une application de présentation nommée PowerGesture. Les trajectoires des gestes sont choisies de façon à être distinctes les unes des autres mais doivent être apprises par un utilisateur. Le taux de reconnaissance moyen est proche de 100%, à partir de 196 exemples d'apprentissage et 54 exemples de test de chaque geste (les 250 exemples sont obtenus auprès de huit personnes différentes). Chateau et al. [7] après élimination du fond et à partir d'une seule caméra, reconnaissent la pose d'une personne, revêtue de manchons colorés, en utilisant un filtre à particules et identifient l'un des seize gestes 3D effectués à partir de la trajectoire 3D (correspondant à la pose 3D) projetée en 2D et

[1] qui sont une extension du modèle IOHMM

par comparaison de cette trajectoire 2D en employant une distance de Haussdorf modifiée. Bahranie et Vannobel [8] identifient des gestes issus de la langue des signes correspondant à des mouvements rectilignes ou circulaires, simples ou composés et obtiennent, pour un vocabulaire de 18 signes (20 occurrences du signe sont apprises et 5 sont testées) un taux de reconnaissance proche de 100%, chaque HMM possédant 5 états cachés et 4 gaussiennes.

3 L'approche HMM

Les chaînes de Markov cachées (HMM) sont aujourd'hui la technique la plus couramment utilisée pour la reconnaissance de gestes dynamiques, notamment en raison de leur succès dans les problèmes de reconnaissance de la parole et de l'écriture manuscrite. Les HMM sont une méthode statistique qui modélisent des séquences d'états, nommés états cachés car non observables. Le modèle inclut des probabilités de transition entre ces états et des probabilités d'émission à partir de ces états afin de modéliser les observations. Soit q_t l'état du système et y_t la sortie à l'instant t. L'émission $P(y_t|q_t)$ de chaque état est probabiliste et ne dépend que de l'état courant q_t. La transition entre deux états $P(q_t|q_{t-1})$ est également probabiliste et ne dépend que de l'état précédent. Cette relation causale est indiquée sur la figure 1.

Les chaînes de Markov cachées sont en fait basées sur des chaînes de Markov homogènes étant donnée que la dynamique du système est déterminée uniquement par les probabilités de transition qui sont indépendantes du temps.

Afin de tirer efficacement parti des HMM, il est nécessaire d'imposer une topologie au graphe des états. L'objet de cette topologie est d'obtenir un meilleur contrôle sur le nombre de paramètres libres et de pouvoir injecter une connaissance *a priori* sur la nature des données. Nous utilisons ici la topologie gauche-droite. Une topologie se caractérise par la présence ou l'absence de transition entre états. Une fois la topologie choisie, l'apprentissage des HMM peut s'effectuer en utilisant l'algorithme *Expectation-Maximization* (EM) [6].

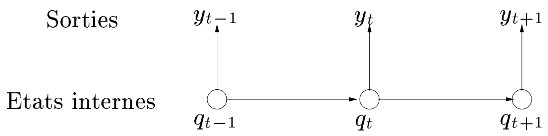

FIG. 1 – Dépendances entre la sortie y et l'état caché q du modèle.

Les HMM sont des modèles génératifs. Cela implique que pour une tâche de classification, un HMM distinct doit être entraîné pour chacune des classes considérées. De façon à séparer les gestes, un classifieur Bayes naïf est utilisé, avec des probabilités a priori égales pour chaque geste [2].

4 La base des gestes

La base de gestes est constituée des trajectoires du centre de la tête, des deux mains et du torse des différents gestes. Chaque trajectoire est obtenue par la méthode de suivi décrite dans [9]. L'acquisition se fait avec deux caméras (distance inter-caméra de l'ordre de 1,20 m) orientées vers une personne située à près de deux mètres des caméras. La calibration des caméras est réalisée avec un objet de taille connue et par une technique similaire à celle utilisée pour le suivi. Le fond est d'abord éliminé puis on effectue un filtrage de couleur adapté à chacune des parties considérées c'est-à-dire de la tête, des deux mains et du torse. L'extraction des régions de la tête, des mains et du torse est simplifiée en équipant la personne qui effectue les gestes d'un vêtement et de deux gants ayant chacun une couleur différente du visage et différente du fond. Lors d'une phase d'apprentissage de gestes, le port

[2] en supposant que chaque geste à la même probabilité d'occurrence.

de vêtements colorés n'est pas une contrainte rédhibitoire. En revanche, en utilisation réelle, une telle contrainte nuit à l'ergonomie de l'IHM gestuelle et il est souhaitable que la trajectoire des mains nues soit obtenue comme cela a été fait en [9]. La figure 2 montre le résultat d'un tel suivi d'une personne sans gants colorés sur les mains : les ellipsoïdes sont correctement positionnés sur la tête et sur les mains et les trajectoires des mains peuvent ainsi être connues au cours d'un geste. Lors de l'apprentissage des gestes, le modèle statistique qui permet le suivi est constitué de quatre ellipsoïdes : un pour la tête et un pour le torse et un pour chacune des deux mains. Chacun des ellipsoïdes est projeté sous la forme d'une ellipse dans le plan de chacune des deux caméras. Une densité de probabilité gaussienne, de même centre et de mêmes dimensions que l'ellipse considérée, est associée à chacune des ellipses. Les paramètres du modèle (positions, tailles et orientations des ellipsoïdes) sont adaptés aux observations des pixels de couleur filtrés lors de l'étape de prétraitement. L'adaptation prend en compte simultanément les pixels détectés issus des deux caméras et est basée sur la recherche du maximum de vraisemblance obtenu par l'algorithme EM.

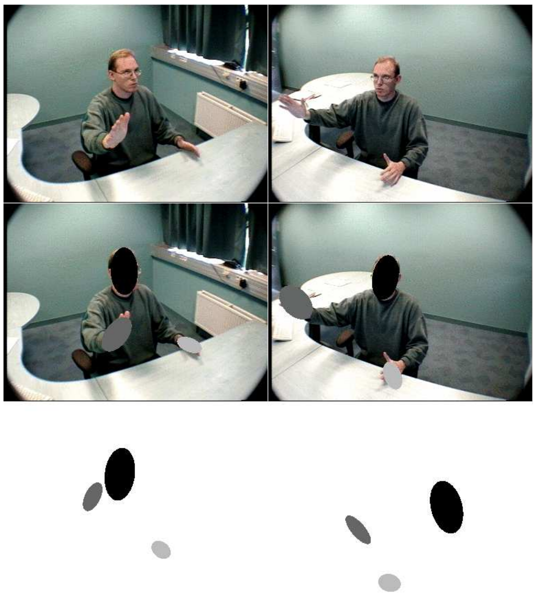

FIG. 2 – Haut : Images d'origine droite et gauche d'une personne sans gant coloré sur les mains. Milieu : Images d'origine avec projection des ellipsoïdes. Bas Gauche : Projection des ellipsoïdes dans le plan xy d'une caméra virtuelle située entre les deux caméras. Bas Droit : Projection des ellipsoïdes dans le plan yz d'une caméra virtuelle située entre les deux caméras.

FIG. 3 – Images droite et gauche d'une personne présentant une extension maximale des bras. Les deux caméras avec un champ de 70ř sont à 1,60m du sol et distantes de 1,20 m ; la personne est située à 1,90 m des deux caméras.

La base de gestes est constituée de seize gestes tous effectués par seize personnes différentes. Pour chaque personne et pour chaque geste, cinq sessions espacées dans le temps ont été effectuées. Lors de chaque session, chaque geste a été répété dix fois à la suite. Les personnes commencent et terminent

FIG. 4 – Une occurrence du geste "nager". Le geste, répété ici deux fois, entre les deux attitudes de repos (première et dernière images), dure 50 images. Une image sur deux est représentée (toujours la vue de droite).

chaque geste dans une même attitude de repos : avec les bras le long du corps et les mains sur les cuisses. A l'issue de chaque session, le repérage du début et de la fin de la session est effectué manuellement. Pour chaque geste de la session, la trajectoire est extraite. On dispose ainsi de 800 trajectoires par geste et d'un nombre total de 12800 trajectoires.

Les gestes sont effectués avec une ou deux mains. Les gestes mono-manuels sont réalisés avec la main dominante (de la main gauche pour les gauchers et de la main droite pour les droitiers). Les gauchers étant moins nombreux, les trajectoires miroir sont extraites comme si le geste avait été réalisé par un droitier.

L'amplitude des gestes dépend de la longueur des bras de chaque personne. Pour chaque personne, un couple d'images est enregistré correspondant à l'extension maximale des bras de la personne (figure 3). Cette extension maximale permet de normaliser les trajectoires des gestes effectués par les différentes personnes.

Nom	Description	**S/R**	**M/B**
Oui/Stop	Main levée au niveau du visage paume ouverte	S	M
Non/Effacer	Idem avec mouvements de gauche à droite	R	M
Lever main	Main levée au dessus de la tête	S	M
Bonjour	Idem avec mouvements de gauche à droite	R	M
Gauche	Main vers la taille, mouvements à gauche	R	M
Droite	Idem avec mouvements à droite	R	M
Haut	Idem avec mouvements vers le haut	R	M
Bas	Idem avec mouvements vers le bas	R	M
Avant	Idem avec mouvements vers l'avant	R	M
Arrière	Idem avec mouvements vers l'arrière	R	M
Nager	Geste mimant la brasse	R	B
Voler	Geste mimant le vol d'oiseau	R	B
Applaudir	Main à hauteur du torse	R	B
Pointer à gauche	Main à hauteur du torse	S	M
Pointer au centre	Main à hauteur du torse	S	M
Pointer à droite	Main à hauteur du torse	S	M

TAB. 1 – Caractéristiques des 16 gestes. **S**imple : le geste est effectué une seule fois entre les positions de repos. **R**épété : le geste est effectué plusieurs fois entre les positions de repos. **M**ono : geste effectué avec une seule main. **B**i-manuel : geste effectué avec les deux mains.

Les seize gestes différents sont décrits dans la Table 1. Certains gestes sont effectués avec deux mains (tels que voler, nager et applaudir). Certains gestes sont effectués une seule fois entre les deux attitudes de repos. D'autres gestes, tels que "nager" sont effectués plusieurs fois entre les positions de repos (figure 4). Le nombre de répétitions est libre et varie selon la personne et la session. L'utilisation de deux gants de couleurs distinctes évitent les problèmes d'occultations et de séparation des deux mains.

La base de données est partitionnée en trois sous-ensembles : le premier pour l'apprentissage, le second pour la validation et le dernier pour l'évaluation (Tableau 2). Les trajectoires effectuées par une même personne n'appartiennent qu'à un seul des trois sous-ensembles, ce qui permet d'évaluer les capacités de généralisation inter-personne de l'apprentissage. Il s'agit donc de reconnaissance multi-signeur où l'on cherche à reconnaître un geste appris effectué par une autre personne. Les gestes étant effectués par des personnes différentes et avec un nombre libre de répétition du noyau ont des durées variables, exprimée en nombre d'images, (tableau 2).

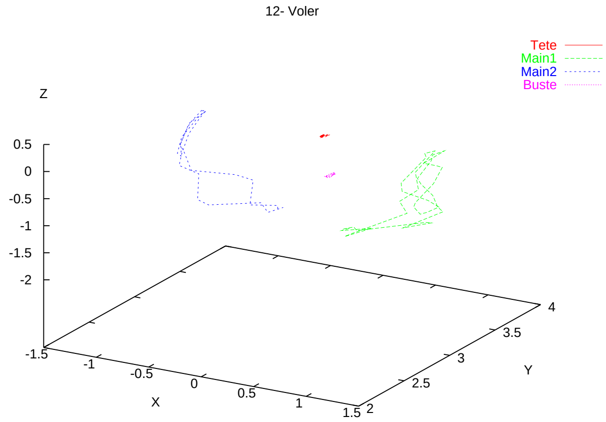

FIG. 5 – Exemple de trajectoire 3D de la tête et du torse (stable) et des deux mains battants des ailes d'une occurrence du geste "voler".

	Apprentissage	Validation	Evaluation
nombre de personnes	4	4	8
nombre de répétitions d'un geste	50	50	50
durée minimale d'un geste	12	6	10
durée moyenne d'un geste	25	24	28
durée maximale d'un geste	64	71	89

TAB. 2 – Partition de la base des trajectoires et durée minimale, moyenne et maximale d'un geste (exprimée en nombre d'images) des trois sous-ensembles.

5 Résultats

5.1 Prétraitement des trajectoires

Pour chaque personne, les trajectoires sont normalisées à partir de l'extension maximale des bras de cette personne. La trajectoire se produit dans un cube d'arête unité correspondant à l'extension maximale des bras. Ainsi les coordonnées x, y et z des mains appartiennent à l'intervalle compris entre -0.5 et 0.5. La figure 5 présente les trajectoires de la tête, du buste et des deux mains d'une occurrence du geste "voler" dans un espace à trois dimensions. L'axe z est l'axe vertical naturel de la personne. Les coordonnées 3D de la tête et du torse changent peu en général au cours d'un geste. C'est pourquoi seules les trajectoires normalisées des deux mains sont prises en compte pour reconnaître un geste.

5.2 Reconnaissance des gestes

Afin de déterminer les méta-paramètres optimaux du modèle (nombre d'états cachés, nombre de gaussiennes par état caché), nous avons suivi le protocole expérimental suivant. De nombreux modèles ont été entraînés sur l'ensemble d'apprentissage. Puis les performances de ces modèles sont déterminées sur l'ensemble de validation. Le meilleur résultat a été obtenu avec 16 états cachés et 20 gaussiennes par état. Ce meilleur modèle a été ré-entraîné sur l'union des ensembles d'apprentissage et de validation et ses performances sont alors estimées sur l'ensemble d'évaluation. Le taux d'erreur moyen, évalué sur l'ensemble des 16 classes, est de 36.70%. Ce taux, supérieur à celui obtenu par [3], peut s'expliquer par le plus grand nombre de gestes distincts, par le fait qu'un certain nombre de trajectoires de gestes différents occupent la même partition de l'espace 3D et par la reconnaissance du type multi-signeur.

Les tableaux 3 et 4 détaillent les résultats du meilleur HMM sur l'ensemble d'évaluation. La matrice de confusion a été séparée en deux pour des raisons de mise en page et la diagonale est représentée en gras. On constate que les gestes bi-manuel sont bien reconnus. Quelques confusions se produisent entre les gestes "nager" et "applaudir" (pour ces deux gestes les mains se rejoignent régulièrement). Le geste de pointage le mieux reconnu est "pointer à gauche" et correspond à une position de la main droite peu présente lors des autres gestes. De manière générale, on constate que les gestes ayant été effectués avec une grande variabilité peuvent conduire à des trajectoires similaires.

6 Conclusions

On constate que les gestes utilisant les deux mains sont plutôt bien classés. En ce qui concerne les gestes gauche, droite, haut, bas, avant et arrière, on s'aperçoit, au vu de la matrice de confusion, que ces gestes sont plutôt mal classés. Cela est dû aux trajectoires en 3 dimensions qui ne sont pas assez discriminantes. Il pourrait être utile d'ajouter de nouvelles caractéristiques, telles les données angulaires ou bien les dérivées premières suivant les 3 axes.

7 Remerciements

Les auteurs tiennent à exprimer leurs remerciements à Bernard Rolland qui a réalisé le logiciel de capture d'images et de trajectoires, à Joël Guérin qui a conduit les sessions d'acquisition de gestes et d'extraction des trajectoires ainsi qu'aux différentes personnes qui ont prêté leur gestuelle.

	oui	non	lever	bonjour	gauche	droite	haut	bas
oui	**230**	127	0	0	8	4	11	9
non	49	**133**	0	0	8	6	23	7
lever	7	5	**314**	253	0	1	0	0
bonjour	1	2	66	**116**	0	0	0	0
gauche	14	0	0	0	**212**	34	60	58
droite	1	10	0	0	4	**236**	9	35
haut	27	43	0	0	48	67	**134**	78
bas	8	7	0	0	35	18	45	**89**
avant	6	20	0	0	23	15	24	46
arrière	53	53	0	0	8	11	50	32
nager	0	0	0	0	0	0	0	0
voler	0	0	0	0	0	0	0	0
appl(audir)	0	0	0	0	0	0	0	0
pointer G	0	0	0	0	24	0	0	0
pointer A	4	0	0	0	2	0	0	14
pointer D	0	0	20	34	28	8	44	32

Tab. 3 – Matrice de confusion sur l'ensemble d'évaluation, gestes 1 à 8 (colonne : geste réalisé, ligne : geste obtenu). Chaque geste étant réalisé 400 fois, on indique le nombre de fois où un geste est obtenu en réponse.

	avant	arrière	nager	voler	appl	point G	point A	point D
oui	12	54	1	0	2	12	7	16
non	13	9	0	0	0	0	0	43
lever	2	0	0	0	0	0	0	1
bonjour	0	0	0	0	0	0	0	4
gauche	22	18	0	0	0	0	0	0
droite	19	0	0	0	0	0	0	15
haut	21	38	0	0	0	0	0	0
bas	5	5	0	0	0	0	0	0
avant	**218**	49	0	0	0	0	18	3
arrière	18	**221**	0	0	0	1	0	3
nager	0	0	**399**	0	22	0	0	0
voler	0	0	0	**400**	0	0	0	
appl(audir)	0	0	0	0	**376**	0	0	0
pointer G	1	0	0	0	0	**387**	46	0
pointer A	7	2	0	0	0	0	**300**	26
pointer D	62	4	0	0	0	0	29	**289**

Tab. 4 – Matrice de confusion sur l'ensemble d'évaluation, gestes 9 à 16 (colonne : geste réalisé, ligne : geste obtenu).

Références

[1] T. Darrell et A. Pentland, "Space-time gestures", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 335-340, New York, 1993.

[2] T. E. Starner et A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models", *Int. Symposium on Computer Vision*,pp 265-270, Coral Gables, 1995.

[3] S. Marcel, O. Bernier, J.E. Viallet et D. Collobert, "Hand Gesture Recognition using Input-Ouput Hidden Markov Models", *Proc. of the FG'2000 Conference on Automatic Face and Gesture Recognition*, pp 456-651, Grenoble, 2000.

[4] P. Hong, M. Turk et T.S. Huang, "Gesture Modeling and Recognition Using Finite State Machines", *Proc. of the FG'2000 Conference on Automatic Face and Gesture Recognition*, pp 410-415, Grenoble, 2000.

[5] H-K Lee et J. H. Kim, "An HMM-Based Threshold model Approach for Gesture Recognition", *IEEE PAMI*, Vol 21, Nř 10, pp 961-973, 1999.

[6] A.P. Dempster, N.M. Laird et D.B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", *Journal of Royal Statistical Society*, Vol. 39, pp. 1-38, 1977.

[7] T. Chateau, F. Jurie, M. Dhome et R. Marc "Analyse de mouvements humains : formation de stagiaires chefs de manoeuvre sur ponts roulants" *Journées du GDR ISIS et AS 70 Perception, Modélisation et Interprétation du Geste Humain"*, Grenoble, Mars 2003.

[8] S. Bahrami et J-M Vannobel "Reconnaissance de signes par la modélisation de primitives de mouvements" *Journées du GDR ISIS et AS 70 Perception, Modélisation et Interprétation du Geste Humain"*, Grenoble, Mars 2003.

[9] O. Bernier et D. Collobert, "Head and Hands 3D Tracking in Real Time by the EM algorithm" *Proceeding of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)*, pp 75-81, Vancouver, 2001.