Fusion de caractéristiques inertielles pour la reconnaissance de gestes

SAMUEL BERLEMONT, GREGOIRE LEFEBVRE

Orange Labs R&D
28 Chemin du Vieux Chêne, Meylan, France

prénom.nom@orange.com

Résumé - Nous nous intéressons ici à la problématique de la reconnaissance de gestes symboliques dans l'espace 3D, instrumentés sur terminal mobile. Nous proposons le couplage des données inertielles hétérogènes accélérométriques et gyrométriques pour augmenter la précision des méthodes classiques de reconnaissance de gestes symboliques. L'utilisation des données couplées de l'accéléromètre et du gyromètre améliore le taux de reconnaissance moyen pour des écart-types réduits, ce qui permet d'asseoir l'intérêt de ces modèles pour le cas « utilisateur unique » (99,87 ± 0,20 %), et proposer des pistes d'étude pour le cas « multi-utilisateur » (79,73 ± 1,72 %).

Abstract – In this paper, we focus on 3D symbolic gesture, performed with mobile devices. We suggest coupling heterogeneous accelerometric and gyrometric inertial data to improve the precision of classic symbolic gesture recognition methods. The coupled accelerometer and gyrometer data allow for a significant increase in mean recognition rates, with reduced standard deviations, which confirms the high-performance of these models for the one-user case (99.87 ± 0.20 %), and suggest some leads for tackling the multi-user case (79.73 ± 1.72 %).

1 Introduction

De nos jours, la multiplication des capteurs dans les terminaux mobiles offre de nouvelles fonctionnalités. L'interaction homme machine se développe notamment autour de la gestuelle naturelle instrumentée, dont la détection et la reconnaissance peuvent être assurées par des capteurs inertiels. La détection de gestes consiste à segmenter temporellement les signaux utiles pour la phase de reconnaissance, tandis que cette dernière permet de différencier et identifier des gestes statiques (ex: terminal retourné), des gestes dynamiques (ex : secouer le terminal) ou des gestes symboliques (ex: alphabet, chiffres, formes sémantiques). Ces derniers offrent à l'utilisateur de nouvelles modalités, par exemple pour s'authentifier sur terminal mobile [6], ou télécommander un périphérique distant [3].

Cet article est organisé comme suit : la section 2 présente un état de l'art de la reconnaissance de gestes symboliques instrumentés. Notre approche est ensuite exposée à la section 3. Dans la section 4, les résultats expérimentaux illustrent l'intérêt du couplage d'informations. Finalement, nous dressons quelques conclusions et perspectives.

2 Etat de l'art

2.1 Prétraitements

Classiquement, les signaux inertiels se présentent sous la forme d'un ensemble de relevés des accélérations linéaires selon les axes d'un repère fixe par rapport au terminal. Les accéléromètres possèdent des caractéristiques propres, et diffèrent en précision, en sensibilité, et en fréquence d'échantillonnage. Ces relevés sont soumis à des perturbations, telles que le bruit ou les interférences, qu'il convient de corriger à l'aide de prétraitements. Cette phase définit quatre étapes : la calibration, la normalisation, le filtrage et la mise à l'échelle temporelle.

Les relevés accélérométriques subissent d'abord une phase de calibration pour supprimer l'influence de la gravité. Les capteurs *MEMS* (*MicroElectroMechanical System*s), calculent en effet des accélérations linéaires au moyen de systèmes masse-ressort capacitifs. Ainsi, lorsque le terminal est à la verticale, la force de gravité entraîne un déplacement d'une masse, interprété comme le résultat d'une accélération vers le haut. Afin de pouvoir comparer deux gestes identiques, mais réalisés avec des inclinaisons différentes, il faut parvenir à isoler cette composante au cours du geste. Pylvänäinen [4] part de l'hypothèse que l'inclinaison de l'instrument est constante et très faible au cours du geste. Deux étapes sont donc nécessaires : le calcul d'un vecteur de référence, par exemple une estimation de l'influence de la gravité ; puis la transformation géométrique par rotation des données dans un repère commun à l'aide du vecteur calculé.

La phase de normalisation géométrique permet ensuite de déterminer une échelle commune pour tous les gestes, limitant ainsi l'influence des changements de rapidité d'exécution et d'amplitude entre utilisateurs. Kallio *et al.* [2] proposent une normalisation en amplitude avec une moyenne nulle et une variance unitaire pour toutes les composantes de l'accéléromètre, tandis que Kela *et al.* [3] opèrent une mise à l'échelle linéaire entre le minimum et le maximum des relevés.

La troisième étape permet un lissage par filtrage des données perturbées par des micro-mouvements. Classiquement, sont utilisés les filtres passe-bas et de Butterworth. Akl *et al.* [7] proposent également une compression temporelle, en utilisant des vecteurs moyens sur une fenêtre temporelle glissante.

Finalement, une phase de mise à l'échelle temporelle est réalisée, ce qui permet de donner plus d'importance

à l'information saillante du geste. Schlömer *et al.* [5] proposent une stratégie de seuillage pour supprimer l'information redondante, tandis que Kallio *et al.* [2] et Kela *et al.* [3] uniformisent le nombre de relevés pour chaque geste par interpolation.

2.2 Modélisation et classification

Diverses stratégies peuvent être appliquées pour reconnaitre des gestes instrumentés à partir des données inertielles prétraitées. Trois approches classiques sont mises en œuvre : l'approche géométrique, statistique ou par apprentissage automatique.

L'approche géométrique propose une mise en correspondance entre un signal test et des instances de référence qui composent un modèle de geste (cf. *Dynamic Time Warping (DTW)* [6,7,10]). La stratégie de classification majoritaire est l'algorithme des k plus proches voisins. Alors que Liu *et al.* [6] utilisent deux répétitions par classe, Choe *et al.* [7] opèrent une phase de *clustering* afin de modéliser des exemples représentatifs depuis la base d'apprentissage.

La seconde approche synthétise les signaux d'apprentissage sous la forme de modèles *HMM* (*Hidden Markov Model (HMM)*) [2-5]. Les signaux de test sont ensuite classés en sélectionnant le modèle fournissant le score maximum de vraisemblance. On distingue deux modalités : les *HMM* discrets *(dHMM),* dont les émissions suivent une variable aléatoire discrète, et les *HMM* continus *(cHMM)*, dont les émissions suivent une variable aléatoire continue. Ainsi, les *dHMM* requièrent une phase de quantification vectorielle pour discrétiser les relevés des signaux. Deux stratégies s'affrontent : la discrétisation de l'espace caractéristique des accélérations, avec par exemple, une distribution uniforme sur la sphère unité pour Schlömer *et al.* [5], et la détermination d'un *codebook* spécifique à chaque modèle par *clustering kmeans* [2,3].

Une troisième approche propose de calculer des caractéristiques (ex: moyenne, énergie, entropie, etc.) par segment temporel des gestes à apprendre, afin de construire un classifieur spécifique par apprentissage automatique (par ex: *Support Vector Machine, SVM*) [9]. Le classifeur détermine alors la classe d'appartenance d'un geste inconnu selon son vecteur caractéristique.

3 Notre approche

La littérature exploite une source unique de données inertielles : l'accéléromètre. Notre contribution est d'utiliser conjointement des données inertielles hétérogènes synchronisées : les accélérations linéaires et les vitesses angulaires, provenant respectivement de l'accéléromètre et du gyromètre du terminal. En effet, un gyromètre *MEMS* est composé de quatre pièces capacitives oscillantes. Lors d'une rotation, la force de Coriolis résultante provoque un changement de mode oscillatoire, entraînant une variation de capacité.

Ce couplage d'informations inertielles nécessite de même une étape de prétraitement avant la modélisation des gestes et leur classification.

3.1 Prétraitements

Notre approche s'appuie sur une phase de prétraitement en trois étapes : normalisation, filtrage et mise à l'échelle temporelle. La phase préliminaire de calibration est exclue car nous supposons que les gestes partagent une même orientation initiale, permettant ainsi d'isoler la composante due à la gravité.

L'étape de normalisation, dont la stratégie résulte de tests préliminaires, se définit par une échelle linéaire spécifique à chaque enregistrement. Pour chaque geste $G: \{G_{t_1}; \dots; G_{t_N}\}$, , les composantes vectorielles sont divisées par la norme maximale définie par :

$$\max_t(||G_{t_1}||, \dots, ||G_{t_n}||) \qquad (1)$$

L'étape de filtrage est ensuite réalisée à l'aide d'un filtre passe-bas discret, de paramètre β, donnant une importance décroissante aux données passées. Soit G le geste non filtré, le geste filtré GF est défini par :

$$GF_{t+1} = (1-\beta).G_{t+1} + \beta.GF_t \qquad (2)$$

Enfin, la mise à l'échelle temporelle résulte d'un seuillage, en supprimant les échantillons dont la distance euclidienne avec leur prédécesseur est inférieure à un seuil δ. Ainsi, le geste seuillé GS est constitué des relevés du geste G satisfaisant la condition :

$$\forall t, \|GS_t - GS_{t-1}\| > \delta \qquad (3)$$

La Figure 1 illustre la composante accélérométrique suivant l'axe $\boldsymbol{x}$ de dix gestes rapides vers la droite avant et après les prétraitements. On observe que la normalisation permet d'uniformiser l'amplitude sur l'ensemble des gestes ; que le filtrage effectue un lissage des variations rapides de faible amplitude ; et que le seuillage uniformise les débuts et fins de geste. Ces traitements contribuent ainsi à faire émerger la partie saillante du geste.

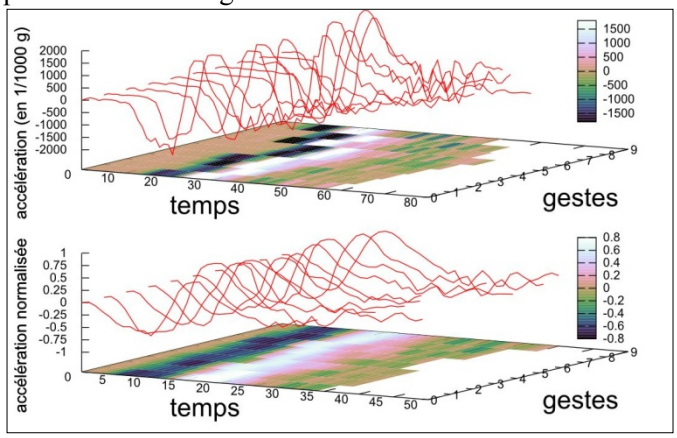

Figure 1: Composante accélérométrique brute (haut) et prétraitée (bas) de plusieurs gestes rapides vers la droite.

3.2 Modélisation et classification

Après prétraitements, nous établissons une comparaison des résultats de classification pour les modèles géométriques et les modèles statistiques. Nous privilégions ces modélisations, basées sur le traitement

des signaux inertiels temporels, et excluons de l'étude l'apprentissage automatique basé sur un vecteur fixe.

L'approche géométrique étudiée se fonde sur [10] et consiste en un modèle basé sur le *DTW*, avec une classification de gestes inconnus à l'aide de l'algorithme des k plus proches voisins. Le *DTW* permet d'offrir une mesure élastique entre le geste inconnu et les représentants de chaque classe de geste dans la base d'apprentissage. Une telle modélisation approxime alors la règle de classification bayésienne.

Pour l'approche statistique, notre choix se porte sur les *HMM* discrets et continus. Une architecture « left-to-right » [1] est adoptée, pour laquelle les états ne peuvent opérer de transition vers des états antérieurs, afin de refléter le caractère temporel du geste. Nous limitons le nombre d'états successifs vers lesquels un état peut opérer une transition. La constitution du *codebook* des *dHMM*, dont la taille est à déterminer expérimentalement, est réalisée par l'algorithme *kmeans*. Ce dernier est initialisé de manière à assurer une représentation uniforme du geste au cours du temps, en prenant les moyennes des vecteurs tous les $\frac{1}{k-1}.T_{max}$ pour chaque geste utilisé en apprentissage, avec T_{max} le temps du geste. Les lois d'émissions des c*HMM* sont définies par des lois normales, dont les matrices de variance-covariance $\sum_i$ sont initialisées en découpant chaque geste en apprentissage en autant de séquences qu'il y a d'états s_i, $\sum_i$ correspondant alors à une estimation non-biaisée sur la séquence correspondante.

4 Expérimentations

4.1 Protocole expérimental

La littérature ne présentant pas de base de référence pour les gestes symboliques, nous proposons une phase d'acquisition et d'évaluation pour évaluer l'apport du couplage d'informations inertielles.

Les performances des méthodes de reconnaissance *DTW, dHMM* et *cHMM* sont évaluées sur deux bases de données, constituées de 14 gestes symboliques : BD_0, effectuée par un utilisateur unique, formée de 40 répétitions pour chaque geste, représentant ainsi 560 enregistrements et BD_1, effectuée avec 22 utilisateurs et formée de 5 répétitions par geste et par utilisateur, pour un ensemble total de 1540 enregistrements. Des gestes de complexités différentes sont représentés dans notre étude, avec des mouvements courts et rectilignes tels que *flickS, flickN, flickE, flickW* (translations dans le plan horizontal), *up* et *down* (translations selon l'axe vertical), ou des gestes à plus forte valeur sémantique et présentant des variations de courbure avec *heart*, *alpha*, *N, Z, clockwise, counterclockwise* et *throw*.

L'acquisition des données est basée sur le Smartphone Android *Samsung Nexus S*. Les relevés inertiels sont composés des mesures simultanées de l'accéléromètre (*KR3DM 3-axis accelerometer)* et du gyromètre *(KR3G gyroscope sensor)*, à une fréquence moyenne de 40 Hz.

Les relevés sont segmentés temporellement de manière automatique sur un critère de puissance instantanée calculé à partir des données accélérométriques. Ce critère, nul lorsque le terminal est inactif, augmente jusqu'à atteindre un niveau plateau lorsque l'utilisateur débute le mouvement. Une technique de seuillage permet ainsi d'identifier le début et la fin du geste utile à la reconnaissance. La segmentation est alors uniformisée sur l'ensemble des enregistrements, à l'inverse d'une segmentation manuelle par appui de bouton, qui induit des différences inter- et intra- utilisateurs.

4.2 Configurations testées

Trois configurations de test ont été définies, afin de cerner au mieux les limites de chaque méthode. Chaque configuration est étudiée dans trois situations : en utilisant uniquement les relevés accélérométriques (acc); uniquement les relevés gyrométriques (gyr); en combinant ces deux informations (acc+gyr). Les paramètres de prétraitement et des modélisations sont ajustés pour chaque configuration en fonction des bases d'apprentissage. La configuration C_1, basée sur BD_0, permet de tester les performances dans le cas d'un utilisateur unique. Pour chaque modèle et pour chaque geste, respectivement 5 et 16 échantillons sont sélectionnés aléatoirement pour l'apprentissage et le test. La configuration C_2, basée sur BD_1, consiste à sélectionner aléatoirement 2 échantillons par geste et par utilisateur pour l'apprentissage, et les 3 échantillons restants pour la classification, afin de couvrir les variabilités qui existent entre les différents utilisateurs durant l'apprentissage. Les configurations C_1 et C_2 sont répétées 10 fois, afin d'obtenir un taux de reconnaissance moyen, et de limiter l'influence du choix des échantillons d'apprentissage et de test. La configuration C_3, basée sur BD_1, permet de tester le potentiel de généralisation de chaque modèle. Tous les enregistrements pour un utilisateur sont utilisés en apprentissage, tandis que ceux des autres utilisateurs sont classés. On réalise ici une validation croisée pour 3 utilisateurs.

4.3 Résultats et comparaisons des performances

Selon les résultats présentés à la Table 1, lors de l'utilisation d'un capteur unique, l'accéléromètre se montre dans tous les cas plus performant que le gyromètre. Cependant, il est remarquable d'observer que l'utilisation jointe des informations fournies par ces deux capteurs permet une augmentation significative du taux de classification moyen. Globalement, les résultats de l'approche DTW sont supérieurs, notamment grâce à une recherche exhaustive parmi les exemples de référence. Les résultats des *dHMM* sont moindres, dus au caractère temporel du geste. En effet, les émissions composant le geste ne sont pas statistiquement indépendantes. Ainsi, la loi de probabilité discrète n'est pas suffisante pour gérer les variations du temps passé dans chaque état au cours de chaque répétition d'un

geste, à l'inverse de celle des *cHMM*, qui obtient de très bons résultats.

Pour chacune de ces configurations, le couplage des capteurs permet de garder un écart-type faible pour les meilleurs résultats (0,20% pour C_1, 0,15% pour C_2, 1,72% pour C_3), et se montre donc très compétitive par rapport aux résultats utilisant seulement l'accéléromètre.

Pour la configuration C_1, qui correspond au cas le plus favorable pour ces méthodes de reconnaissance, le couplage des capteurs aboutit à un résultat amélioré de 0,3% pour le *DTW* et 0,85% pour les *cHMM*, et ce malgré des scores déjà élevés de 99,40% et 99,02% respectivement pour l'accéléromètre simple.

Pour la configuration C_2, le challenge réside dans le passage à l'échelle avec 22 utilisateurs en apprentissage et en test. La méthode *DTW* offre le meilleur résultat, avec un score moyen de 94,05%, car elle attribue intrinsèquement plus d'importance aux enregistrements de référence appartenant à l'utilisateur à reconnaître. Néanmoins, le temps de classification moyen par geste est de 34,57 ms pour l'approche *DTW* (i.e. recherche exhaustive des k plus proches voisins) et de 12,82 ms pour l'approche *cHMM* (i.e. chargements des modèles et calcul du maximum de vraisemblance). Ceci peut limiter l'usage du DTW pour de grandes bases de gestes.

La configuration C_3 présente le défi le plus important pour ces méthodes de reconnaissance. On teste ici leur adaptabilité et leur potentiel de généralisation, en ne gardant qu'un utilisateur pour l'apprentissage. Les *cHMM* et la méthode *DTW* fournissent des résultats comparables, avec des taux de classification moyens respectifs de 77,64% et 79,73%. Les résultats reflètent la similarité naturelle qui existe entre des gestes exécutés par différents utilisateurs, mais les nuances telles que la manière de tenir le terminal mobile ou la rapidité d'exécution ne sont pas saisies, et limitent les performances. On retrouve par exemple des confusions fréquentes entre *alpha* et *clockwise,* dont les dynamiques sont proches ; ou entre *N* et *Up*, lorsque les rapidités d'exécution diffèrent entre utilisateurs, donnant plus d'importance au début du geste.

5 Conclusions et perspectives

Cet article démontre donc l'intérêt d'utiliser de manière couplée les relevés accélérométriques et gyrométriques pour les problématiques de reconnaissance de gestes. De plus, il offre un récapitulatif des performances des méthodes de reconnaissance les plus utilisées actuellement dans les cas de l'utilisateur unique et de l'utilisateur multiple. Des perspectives sont envisagées dans la sélection et le couplage de données hétérogènes pour modéliser les variabilités intra- et inter- utilisateurs.

Remerciements

Les auteurs remercient Mr. Eric Petit et Mr. Sébastien Roux pour leur aide et pour l'implémentation de la méthode géométrique [10].

Tab 1 : Taux de reconnaissance moyen et écarts-types.

Méthodes		C_1		C_2		C_3	
		Moyenne	Ecart-type	Moyenne	Ecart-type	Moyenne	Ecart-type
DTW	gyr	95,39%	0,56%	80,63%	2,39%	73,35%	3,60%
	acc	99,40%	0,21%	92,59%	0,20%	78,23%	1,67%
	acc+gyr	99,70%	0,42%	**94,05%**	**0,15%**	**79,73%**	**1,72%**
cHMM	gyr	95,05%	2,62%	70,92%	0,74%	68,46%	2,60%
	acc	99,02%	0,81%	83,99%	1,09%	75,44%	1,57%
	acc+gyr	**99,87%**	**0,20%**	85,79%	0,67%	77,64%	2,89%
dHMM	gyr	57,50%	3,24%	43,13%	2,35%	28,14%	7,50%
	acc	77,14%	5,18%	64,09%	1,60%	39,14%	2,95%
	acc+gyr	81,02%	3,72%	69,46%	2,11%	43,31%	2,61%

Références

[1] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] S. Kallio, J. Kela, and J. Mäntyjärvi, Online gesture recognition system for mobile interaction, presented at the *IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2070 – 2076, 2003.

[3] J. Kela, P. Korpipää, J. Mäntyjärvi, S. Kallio, G. Savino, L. Jozzo, and D. Marca, Accelerometer-based gesture control for a design environment, *Personal Ubiquitous Computing*, pp. 285–299, 2006.

[4] T. Pylvänäinen, Accelerometer based gesture recognition using continuous HMMs, *Pattern Recognition and Image Analysis*, no. 8, pp. 639–646, 2005.

[5] T. Schlömer, B. Poppinga, N. Henze, and S. Boll, Gesture recognition with a Wii controller, *Proceedings of the 2nd international conference on Tangible and embedded interaction,* pp. 11–14, 2008.

[6] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, uWave: Accelerometer-based personalized gesture recognition and its applications, *Pervasive Mob. Comput.*, vol. 5, no. 6, pp. 657–675, 2009.

[7] B. Choe, J. Min and S. Cho, Online Gesture Recognition for User Interface on Accelerometer Built-in Mobile Phones, *Proc. in on Neural information processing: models and applications* (2), pp.650-657, 2010.

[8] A. Akl and S. Valaee, Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing, presented at the *International Conference on Acoustics Speech and Signal Processing*, pp. 2270 –2273, 2010.

[9] J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li, Gesture recognition with a 3-d accelerometer, *Ubiquitous Intelligence and Computing*, vol. 5585, pp. 25–38, 2009.

[10] E. Petit, GRASP: Moteur de reconnaissance de gestes. *Dépôt Logiciel, France Télécom*, IDDNFR.001.030023.000.S.P.2010.000.31500, 2010.