Sous échantillonnage et machine à noyaux élastiques pour la classification de données de mouvement capturé

Pierre-François Marteau, Sylvie Gibet, Clément Reverdy

UMR 6074 IRISA, Université de Bretagne Sud,
Campus de Tohannic, 56000 Vannes, France
prenom.nom AT univ-ubs DOT fr

Résumé. Dans le domaine de la reconnaissance de gestes isolés, bon nombre de travaux se sont intéressés à la réduction de dimension sur l'axe spatial pour réduire à la fois la complexité algorithmique et la variabilité des réalisations gestuelles. Il est assez étonnant de constater que peu de ces méthodes se sont explicitement penchées sur la réduction de dimension sur l'axe temporel. En matière de complexité, la réduction de dimension sur cet axe est un enjeu majeur quant à l'utilisabilité de distances élastiques en complexité quadratique. Par ailleurs, la prise en compte de la variabilité sur cet axe demeure une source avérée de gain de performance. Pour tenter d'apporter un éclairage en matière de réduction de dimension sur l'axe temporel, nous présentons dans cet article une approche basée sur un sous échantillonnage temporel associé à l'exploitation d'un apprentissage automatique à base de noyaux élastiques. Nous montrons expérimentalement, sur deux jeux de données très référencés dans la communauté et très opposés en matière de qualité de capture de mouvement, qu'il est possible de réduire sensiblement le nombre de postures sur les trajectoires temporelles tout en conservant, grâce à des noyaux élastiques, des performances de reconnaissance au niveau de l'état de l'art du domaine. Le gain de complexité obtenu rend une telle approche éligible pour des applications *temps-réel*.

1 Introduction

La reconnaissance de gestes est un domaine de recherche très actif depuis plusieurs décennies qui évolue et s'adapte en fonction des dispositifs de capture de mouvement et de l'état de l'art des méthodes de reconnaissance principalement basées apprentissage à partir d'exemples. Récemment, la mise sur le marché de technologies grand public, souvent associées à des consoles de jeux, a permis de démocratiser l'usage de ces capteurs non seulement dans le contexte des jeux interactifs mais également dans le cadre d'applications exploitant une interaction gestuelle. Ainsi des bases de données de très bonne qualité construites à partir de dispositifs coûteux dont l'exploitation nécessite une expertise spécifique côtoient aujourd'hui des bases de données bon marché, mais en contre-partie plus bruitées, produites plus rapidement grâce à ces nouveaux capteurs accessibles et largement diffusés qui ne nécessitent pas d'expertise particulière en matière d'usage. La communauté scientifique voit ainsi mises à sa

disposition des bases de données de mouvement capturé hétérogènes de qualités très diverses, constituant un véritable défi pour les algorithmes de reconnaissance, en particulier de gestes de commande. Ainsi il devient possible d'évaluer la robustesse et la capacité de généralisation des algorithmes de reconnaissance sur des paradigmes de capture de mouvement et des cadres d'utilisation très diversifiés. Outre la qualité de reconnaissance, la complexité des algorithmes et leur temps de réponse est également un enjeu majeur, notamment dans un contexte d'interaction temps réel.

Nous présentons dans cet article un algorithme de reconnaissance de gestes isolés, robuste et efficace, basé sur l'apprentissage de données mouvement capturées et représentées sous la forme de postures squelettiques évoluant au cours du temps. Dans une première partie nous détaillons la nature des données et les prétraitements principaux que nous considérons. La deuxième partie présente l'état de l'art des techniques récentes exploitées pour la reconnaissance de gestes isolés. La troisième partie expose les spécificités de l'algorithme proposé en le positionnant dans le contexte général de la classification de données multivariées séquentielles. Nous présentons dans la quatrième partie une évaluation de cet algorithme sur deux jeux de données de qualités très distinctes en comparant les performances obtenues avec ceux présentés dans l'état de l'art du domaine de la reconnaissance de gestes isolés. Une discussion finale est proposée ainsi que quelques perspectives.

2 Données mouvement capturées et modélisation à base de postures squelettiques

Nous nous intéressons dans cet article à la classification de données de mouvement humain acquises par l'intermédiaire de capteurs divers (caméras avec ou sans marqueurs passifs ou actifs, capteurs mécaniques de type combinaisons ou exo-squelettes), mais pré-traitées de manière homogène. Plus précisément les données capturées sont en général reconstruites sous la forme de trajectoires dans l'espace 3D des points d'articulation d'un squelette sous-jacent plus ou moins proche du squelette réel de l'acteur ayant produit le mouvement. Cette identification d'un modèle de squelette à partir de données capturées est réalisé par le biais d'algorithmes de mise en correspondance par optimisation (de Aguiar et al., 2006), (O'Brien et al., 2000), (Shotton et al., 2011). Les applications exploitant un modèle de squelette convertissent ainsi les données 3D des capteurs en coordonnées cartésiennes ou angulaires permettant de définir plus ou moins précisément l'état des articulations prises en compte.

Nous considérons dans cet article des données squelettiques reconstruites à partir de données acquises via le système Kinect de Microsoft et via le dispositif Vicon-MX utilisé par le Max Planck Institute pour produire la base de donnée HDM05.

Ces données sont intrinsèquement bruitées du fait principalement de deux sources d'erreur :
- les bruits liés aux capteurs et à l'acquisition des données (dérives diverses, phénomène d'imprécision et d'occultation, etc.),
- les bruits liés à la reconstruction du squelette à partir des données de capteur.

Par ailleurs, du fait même de la nature des dispositifs utilisés (du nombre de capteurs exploités), la nature des squelettes reconstruits est variable pour deux raisons principales :

– la morphologie des acteurs (longueur des segments) est la principale source de variabilité pour un dispositif fixé,
– le nombre d'articulations (de degrés de liberté) varie en fonction du dispositif exploité.

D'une manière très générale, nous considérons ainsi une donnée mouvement sous la forme d'un vecteur multivarié décrivant une trajectoire au cours du temps, autrement dit une série temporelle : $Y_T = [y(t) \in \mathbb{R}^k, t = 1 \cdots T] = [y(1), ..., y(T)]$, où $k = 3 \cdot N$ (N étant le nombre d'articulations) varie typiquement entre 20 et 150 en fonction des dispositifs utilisés et de la tâche considérée.

Ce vecteur n'est évidemment pas constitué de caractéristiques scalaires indépendantes. Les redondances spatio-temporelles qu'il encode ouvrent ainsi des perspectives en matière de réduction dimensionnelle et de réduction de bruit, particulièrement intéressantes en reconnaissance de mouvement puisque l'on peut espérer en tirer un bénéfice tant en matière de temps de calcul que de performance.

3 État de l'art de la reconnaissance de gestes isolés

Le domaine de l'analyse et de la reconnaissance de gestes est très large et récemment particulièrement actif compte tenu de la démocratisation des systèmes de capture de mouvement par caméra. Il recouvre des aspects traitement du signal, traitement d'images, modélisation dynamique, approches statistiques et apprentissage automatique. Nous faisons ici un bref panorama non exhaustif des méthodes principales de l'état de l'art actuel.

Les méthodes de reconnaissance de geste se focalisent en premier lieu sur la modélisation de caractéristiques susceptibles de bien représenter la cinématique ou la dynamique qui caractérisent les déformations soit du squelette dans son ensemble, soit des portions de celui-ci. Parmi ces méthodes, on retrouve des approches classiques et novatrices que les auteurs ont adaptées aux données mouvement. On peut citer en particulier les méthodes suivantes :

– Les modèles à base de modèles dynamiques linéaires (Veeraraghavan et al., 2004) exploitent par exemple les modèles autoregressifs (AR) et autoregressifs à moyenne ajustée (ARMA) pour caractériser la cinématique des mouvements.
– Les modèles dynamiques non-linéaires (Bissacco et al., 2007) mettent en ?uvre des principes d'analyse et de reconnaissance de mouvement basés sur des modèles dynamiques contrôlés par des processus Gaussiens.
– Les modèles de Markov cachés (Mitra et Acharya, 2007) ont été exploités avec un certain succès à la reconnaissance de geste.
– Les champs aléatoires conditionnels, (Wang et al., 2006) ont été développé et exploité pour modéliser les dépendances entre les articulations et augmenter ainsi la discrimination des modèles type HMM.
– L'utilisation de réseaux de neurones récurrents (Martens et Sutskever, 2011), ou de machines de Boltzman conditionnelles et restreintes (Larochelle et al., 2012) ont récemment été proposés dans un contexte de modélisation dynamique de séries temporelles et de classification.

Certaines méthodes s'attaquent plus spécifiquement à la problématique de réduction de dimension dans un objectif explicite ou implicite de réduction de la variabilité en recherchant en général un gain d'efficacité. On retrouve parmi ces approches :

- L'Analyse en Composante Principale a été très largement exploitée en analyse et reconnaissance de geste dans un objectif de réduction temporelle (Masoud et Papanikolopoulos, 2003).
- D'autres méthodes linéaires comme les projections préservant les voisinages locaux (Locality Preserving Projection) (He et Niyogi, 2003) ou leurs analogues non linéaires tel qu'ISOMAP (Tenenbaum et al., 2000) ont été mises en oeuvre pour plonger les postures dans des espaces à plus faible dimension au sein desquels un alignement temporel (DTW, voir section 4.2) plus efficace associé à la distance de Hausdorff peux être exploité pour classer des mouvements.
- Des méthodes plus adhoc de réduction de dimension s'avèrent également efficaces : on peut citer par exemple les travaux récents d'(Yu et Aggarwal, 2009) qui proposent de ne s'intéresser qu'aux trajectoires des cinq extrémités du squelette (les 2 pieds, les 2 mains et la tête).
- Les modèles à base de processus Gaussiens à variables latentes ont été également largement mis à contribution, en particulier une version hiérarchique exploitée en reconnaissance de geste d'action. (Han et al., 2010)

D'autres méthodes enfin ciblent plutôt l'identification de variables significatives maximisant la discrimination des catégories de mouvement considérées.

- Les forêts d'arbre de décision, Fothergill et al. (2012), Zhao et al. (2012) ont récemment fait l'objet d'expérimentation sur des données capturées par Kinect.
- (Ofli et al., 2013) ont récemment proposé de sélectionner automatiquement quelques articulations du squelette réputées les plus informatives pour expliquer l'action en cours.
- Dans la même lignée, (Hussein et al., 2013) utilisent des matrices de covariance évaluées sur les points d'articulation du squelette se déformant au cours du temps comme des descripteurs discriminants pour caractériser une séquence de mouvement. L'exploitation de fenêtres glissantes peut s'apparenter ici à une réduction de dimension sur l'axe temporel.
- Dans (Wang et al., 2012) des *actionlets* sont définis à partir de coefficients de Fourier pour caractériser les articulations les plus discriminantes.
- Etc.

Concernant l'exploitation de distances élastiques dans un processus de reconnaissance on peut citer parmi les nombreux travaux existants les applications décrites dans (Sempena et al., 2011), et les accélérations matérielles proposées dans (Hussain et Rashid, 2012). Cependant, aucun travail exploitant ce type de distance n'a étudié à notre connaissance la question de la réduction de données sur l'axe temporel.

4 Sous échantillonnage et machines à noyaux élastiques

4.1 Réduction de dimension sur l'axe temporel et distances élastiques

Lorsque l'on s'intéresse aux distances ou noyaux élastiques pour profiter de leur capacité à gérer certaines formes de variabilité temporelle, on est vite confronté à des difficultés liées à leur coût algorithmique, en général quadratique avec la longueur des séries temporelles traitées et linéaire avec le nombre de dimensions *spatiales* que celles-ci comportent. Cette complexité importante est une limitation certaine quant à leur utilisation notamment lorsque de grands volumes de données doivent être traités ou lorsque des contraintes dites *temps-réels* s'imposent. Il est donc a priori particulièrement pertinent de considérer une réduction de la complexité algorithmique sur l'axe temporel au même titre que sur l'axe spatial. Dans la littérature très riche portant sur les questions de reconnaissance de geste il est notable de constater que si certaines études ont permis de réduire avec un certain succès la dimensionnalité des données sur l'axe spatial, très peu à notre connaissance ont considéré une réduction de dimensionalité sur l'axe temporel. On peut citer les travaux de (Keogh et Pazzani, 2000) qui ont explicitement mentionné un sous-échantillonnage temporel associé à une comparaison dynamique (DTW) dans un contexte de fouille de séquences temporelles. Dans le contexte du traitement des données mouvement, il semble donc particulièrement utile de déterminer si une redondance spatio-temporelle des trajectoires de mouvement peut-être exploitée, notamment dans la perspective d'une utilisation des distances élastiques.

Une première approche que nous développons ici consiste à sous-échantillonner les données mouvement de manière à ramener toutes les trajectoires de mouvement sous la forme de séquences de même longueur, i.e. contenant le même nombre réduit de postures, puis d'effectuer une tâche de classification ou de reconnaissance à l'aide de machine à vecteurs supports (SVM) à base de noyaux élastiques afin d'évaluer les taux de performance en fonction du degré de sous-échantillonnage considéré.

4.2 Noyaux élastiques et leur régularisation

Déformation temporelle dynamique
La mesure DTW (Dynamic Time Warping) (Velichko et Zagoruyko, 1970), (Sakoe et Chiba, 1971), de loin la mesure élastique la plus exploitée, est définie de la manière suivante :

$$\begin{aligned} d_{dtw}(X_p, Y_q) &= d_E^2(x(p), y(q)) \\ &+ \text{Min} \begin{cases} d_{dtw}(X_{p-1}, Y_q) & suppression \\ d_{dtw}(X_{p-1}, Y_{q-1}) & substitution \\ d_{dtw}(X_p, Y_{q-1}) & insertion \end{cases} \end{aligned} \quad (1)$$

où $d_E(x(p), y(q)$ est la distance euclidienne (éventuellement élevée au carré) définie sur $\mathbb{R}^k$ entre les deux postures des séquences X et Y prises aux instants p et q respectivement. Outre le fait que cette mesure ne respecte pas l'inégalité triangulaire, elle ne permet pas de définir un noyau défini positif. Lorsqu'une telle mesure est exploitée par une machine à vecteurs supports (SVM), le problème d'optimisation inhérent à la phase d'apprentissage de ce type d'algorithme n'est plus quadratique. La convergence vers l'*optimorum* n'est donc plus

garantie ce qui, en fonction de la complexité de la tâche considérée peut être pénalisant.

Noyau DTW régularisé

Des travaux récents (Cuturi et al., 2007), (Marteau et Gibet, 2013) ont permis de proposer de nouvelles orientations pour régulariser les noyaux construits à partir des mesures élastiques telles que la DTW. Une solution générale proposée par (Marteau et Gibet, 2013), qui découle du théorème de (Haussler, 1999) sur les noyaux de convolution définis sur structures discrètes, définit le noyau DTW régularisé $\mathcal{K}_{dtwr}$ sous la forme suivante qui exploite deux termes récursifs, K^{xy}_{dtwr} et K^{xx}_{dtwr} :

$$\begin{aligned}
&\mathcal{K}_{dtwr}(X_p, Y_q) = K^{xy}_{dtwr}(X_p, Y_q) + K^{xx}_{dtwr}(X_p, Y_q) \\
&K^{xy}_{dtwr}(X_p, Y_q) = \tfrac{1}{3} e^{-\nu d^2_E(x(p),y(q))} \sum \begin{cases} h(p-1,q) K^{xy}_{dtwr}(X_{p-1}, Y_q) \\ h(p-1,q-1) K^{xy}_{dtwr}(X_{p-1}, Y_{q-1}) \\ h(p,q-1) K^{xy}_{dtwr}(X_p, Y_{q-1}) \end{cases} \\
&K^{xx}_{dtwr}(X_p, Y_q) = \tfrac{1}{3} \sum \begin{cases} \frac{1}{2} h(p-1,q) K^{xx}_{dtwr}(X_{p-1}, Y_q) \\ \quad \left(e^{-\nu d^2_E(x(p),y(p))} + e^{-\nu d^2_E(x(q),y(q))} \right) \\ \Delta_{p,q} h(p-1,q-1) K^{xx}_{dtwr}(X_{p-1}, Y_{q-1}) e^{-\nu d^2_E(x(p),y(q))} \\ \frac{1}{2} h(p,q-1) K^{xx}_{dtwr}(X_p, Y_{q-1}) \\ \quad \left(e^{-\nu d^2_E(x(q),y(q))} + e^{-\nu d^2_E(x(p),y(p))} \right) \end{cases}
\end{aligned} \tag{2}$$

où $h(.,.)$ est une fonction indicatrice définissant un corridor symétrique, $d_E(.,.)$ est la distance euclidienne définie sur $\mathbb{R}^k$, $\nu \in \mathbb{R}^+$ et $\Delta_{p,q}$ est le symbole de Kronecker.

La principale idée derrière la régularisation de noyaux construits à partir d'une distance élastique est de remplacer les opérateurs $\min$ ou $\max$ (qui empêchent la symétrisation des noyaux) par un opérateur de sommation ($\sum$). Ceci amène à considérer, non plus uniquement le meilleur alignement possible, mais tous les meilleurs voire également les *bons* chemins en sommant leurs coût globaux. Le paramètre ν permet de contrôler ce que l'on entend par *bon* alignement en pénalisant plus ou moins les alignements trop éloignés des alignements optimaux. Le paramètre ν peut-être aisément optimisé, en adoptant, par exemple, une procédure de validation croisée sur des données d'apprentissage.

Le noyau $\mathcal{K}_{dtwr}$ défini par l'équation 2 est défini positif. Les deux termes K^{xy}_{dtwr} et K^{xx}_{dtwr} le rendent équivalent à une somme de noyaux de convolution (Marteau et Gibet, 2013).

Noyaux élastiques

Nous considérons uniquement les noyaux exponentiels (de type Gaussien ou RBF) construits à partir des deux mesures précédentes d_{dtw} et $\mathcal{K}_{dtwr}$, ainsi que le noyau obtenu à partir d'une distance Euclidienne [1] à titre de référence, i.e. :

$$K_{dtw}(.,.) = e^{-d_{dtw}(.,.)/\sigma} \quad K_{dtwr}(.,.) = e^{\mathcal{K}_{dtwr}(.,.)/\sigma} \quad K_{ed}(.,.) = e^{-d_{ed}(.,.)/\sigma} \tag{3}$$

1. La distance Euclidienne est exploitable uniquement parce qu'un nombre fixe de postures est considéré pour caractériser chacun des mouvements, ceci quelque soit leur longueur initiale.

5 Expérimentations

Nous montrons dans les expérimentations qui suivent, que, comparativement à l'exploitation de noyaux élastiques non définis positifs tels que K_{dtw}, l'exploitation des noyaux élastiques régularisés améliorent les performances des classifieurs à base de machine à vecteurs supports en rendant convexe le problème d'optimisation sous-jacent à l'apprentissage de ces machines.

5.1 Bases de données mouvement et tâches considérées

Pour estimer la robustesse de l'approche proposée, nous l'évaluons sur deux bases de données mouvement de qualité opposée, l'une développée au Max Planck Institute, l'autre dans les laboratoires de recherche de Microsoft.

Base de données HDM05 : Cette base de données (Müller et al., 2007) du Max Planck Institute est constituée de données capturées par un système Vicon-MX à base de marqueurs optiques réflectifs suivis par 6 caméras haute définition et configurées pour des enregistrements à 120Hz. Les actions gestuelles segmentées sont transformées en séquences de postures qui correspondent à un squelette constitué de $N = 31$ articulations. A chacune des articulations est ainsi associée une position 3D (x, y, z) qui décrit une trajectoire au cours de l'action. En pratique la position de l'articulation centrale du squelette et son orientation (localisée au niveau du plexus) servant de référentiel, seules les positions dans ce référentiel des 30 articulations restantes sont exploitées, ce qui conduit à représenter chaque posture par un vecteur $Y_t \in \mathbb{R}^k$, avec $k = 90$. Les deux tâches considérées, HDM05-1 et HDM05-2 sont celles proposées respectivement par (Ofli et al., 2012) (reprises dans les travaux de (Hussein et al., 2013)) et (Ofli et al., 2013). Pour les deux tâches, 3 sujets participent à l'apprentissage et 2 sujets distincts participent aux tests. Pour la tâche HDM05-1, 11 actions gestuelles sont considérées : *{deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat, and throw basketball}*. Cela constitue 249 séquences de mouvement. Pour la tâche HDM05-2, les sujets sont les mêmes, mais 5 actions supplémentaires sont considérées en plus des 11 précédentes : *{jump, jumping jacks, throw, sit down, and stand up}*. Le jeu de données considéré comporte 393 séquences de mouvement au total. Pour ces deux tests, la longueur des séquences d'action est comprise entre 56 et 901 postures (0,5-7,5s).

Base de données MSR-Action3D : Cette base de données (Li et al., 2010) a récemment été développée pour proposer un *benchmark* constitué de séquences d'images 3D (*depth map*) capturées par le capteur Kinect de Microsoft. La base de données est constituée de 20 actions gestuelles typiques d'une interaction avec une console de jeu et étiquetées comme suit : [*high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw*]. Chaque action a été réalisée par 10 sujets faisant face à la caméra 2 ou 3 fois. Le jeu de données comporte 567 séquences de mouvement dont les longueurs varient de 14 à 76 postures. Les images 3D de taille 640x480 ont été acquises à la fréquence de 15 images par seconde. De chaque image 3D est extraite une posture squelettique comportant $N = 20$ articulations caractérisées par 3 coordonnées. Nous caractérisons les postures en référence à l'articulation centrale du squelette, ce qui conduit

à représenter chaque posture par un vecteur $Y_t \in \mathbb{R}^k$, avec $k = 57$. La tâche consiste à fournir une validation croisée sur les sujets, i.e. 5 sujets participant à l'apprentissage et 5 sujets participant au test, avec toutes les configurations possibles, soit 252 au total.

5.2 Résultats et analyse

Nous présentons, pour les tâches considérées, les résultats obtenus en exploitant un classifieur SVM construit à partir de la bibliothèque LIBSVM (Chang et Lin, 2001) et des noyaux élastiques K_{dtw} et K_{dtwr}. Le sous-échantillonnage considéré est homogène ; dans toutes les tâches considérées, nous conservons la première et la dernière posture, et nous sélectionnons un nombre fixe de postures à intervalle fixe. A titre de référence nous indiquons également les résultats obtenus sur la base d'un noyau construit à partir de la distance euclidienne, K_{ed}. Nous fournissons les taux de bonne classification sur les données d'apprentissage obtenus par validation croisée (80% apprentissage, 20% test) et sur les données tests.

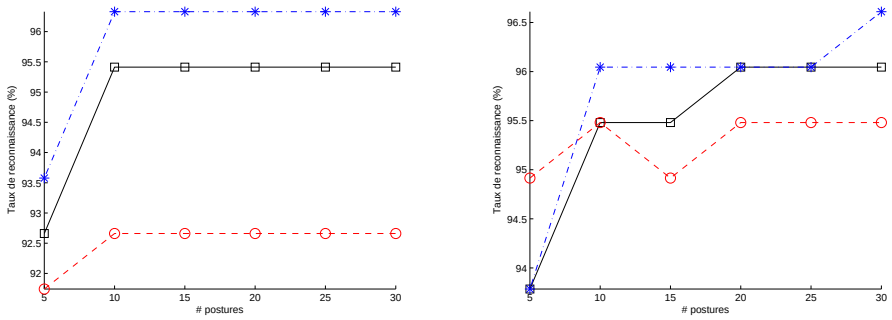

FIG. 1 – *Taux de bonne classification obtenus sur la base HDM05 pour la tâche HDM05-1 définie dans (Ofli et al., 2012) à gauche, et pour la tâche HDM05-2 définie dans (Ofli et al., 2013) à droite pour un nombre de postures variant dans* $\{5, 10, 15, 20, 25, 30\}$ *:* K_{ed} *courbes (rouge, rond, trait),* K_{dtw} *courbes (noire, carré, plein),* K_{dtwr} *courbes (bleu, étoile, point-trait).*

La figure 1 présente les taux de bonne classification lorsque le nombre de postures retenus varie de 5 à 30 sur les tâches HDM05-1 (à gauche) et HDM05-2 (à droite). Nous observons sur cette figure que le sous-échantillonnage ne dégrade pas les résultats de manière catastrophique. Des taux de sous échantillonnage importants (10 à 15 postures par mouvement, ce qui représente un taux de compression moyen de 97% sur la base HDM05 et de 70% sur la base MSRAction3D) conduisent à des résultats très satisfaisants (96 à 98% pour les 2 tâches). Le classifieur SVM construit sur la base du noyau K_{dtwr} régularisé produit les meilleurs taux de reconnaissance. Le même type de résultat est obtenu sur la base MSRAction3D, avec toutefois des performances bien moindres pour les SVM construits sur la base de la distance Euclidienne. Par ailleurs si l'on obtient des très bon taux de classification (96%) sur les données d'apprentissage, du fait de la nature bruitée des données Kinect et de la variabilité inter sujets, les taux de reconnaissance sur les données de test descendent autour de 82%.

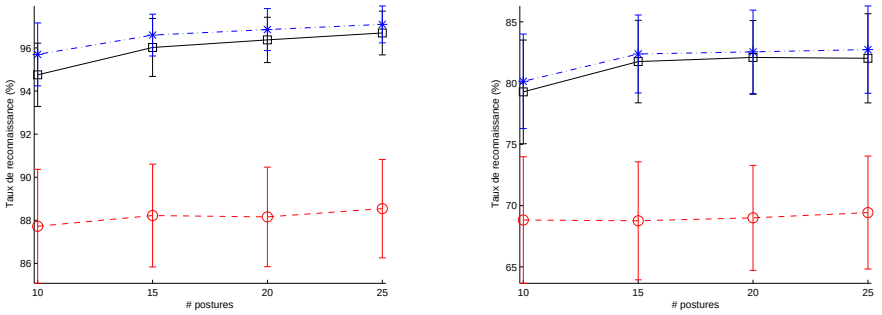

FIG. 2 – *Taux de bonne classification obtenus sur la base MSRAction3D, sur les données d'apprentissage (à gauche) et de test (à droite) pour un nombre de postures par mouvement variant dans* $\{10, 15, 20, 25\}$ *: K_{ed} courbes (rouge, rond, trait), K_{dtw} courbes (noire, carré, plein), K_{dtwr} courbes (bleu, étoile, point-trait).*

	K_{ed}A	K_{ed}T	K_{dtw}A	K_{dtw}T	K_{dtwr}A	K_{dtwr}T
Moyenne	87,71	69,73	96,04	81,41	96,65	82,50
Ecart type	2,34	5,73	1,36	5,04	1,13	3,22

TAB. 1 – *Moyennes et écarts types des taux de bonne classification obtenus sur la base de données MSRAction3D par validation croisée sur les sujets (252 tests) ; A : sur les données d'apprentissage, T : sur les données de test pour un nombre de postures égal à 15.*

La table 1 précise sur les données MSRAction3D et pour les SVM à base de K_{ed}, K_{dtw} et K_{dtwr}, les moyennes et écarts-types obtenus pour les données d'apprentissage (A) et de test (T) des taux de reconnaissance en validation croisée sur les sujets (252 configurations) lorsque les mouvements sont représentés sous la forme de séquences de 15 postures. A titre de comparaison, nous donnons dans la table 2 les résultats obtenus par différentes méthodes de l'état de l'art et mettons en vis-vis les résultats obtenus par nos SVM construits à partir de la DTW régularisée en association avec un sous-échantillonnage à 15 postures. Cette analyse comparée montre que les SVM construits à partir de noyaux DTW régularisés obtiennent des résultats légèrement au dessus de l'état actuel sur les jeux de données considérés.

6 Conclusion et perspectives

Dans le contexte de la reconnaissance de gestes isolés, où peu de travaux considèrent explicitement la réduction de dimension sur l'axe temporel, nous avons présenté une approche basée sur un sous-échantillonnage des séquences de mouvement associée à l'exploitation de machines à noyaux élastiques. Sur les tâches considérées, nous avons pu montrer qu'un sous-échantillonnage, même relativement important, ne dégrade que très modérément les résultats de reconnaissance. La redondance temporelle est donc élevée et peu utile du point de vue de la discrimination de ce type de mouvement. Le bénéfice en terme de complexité algorithmique est

HDM05-1	Taux de bonne classification (%)
SMIJ (Ofli et al., 2012)	84.40
Cov3DJ, L = 3 (Hussein et al., 2013)	95.41
$SVMK_{dtwr}$, 15 postures	**96.33**
HDM05-2	Taux de bonne classification (%)
SMIJ (Ofli et al., 2013), 1-NN	91.53
SMIJ (Ofli et al., 2013), SVM	89.27
$SVMK_{dtwr}$, 15 postures	**96.05**
MSR-Action3D	Taux de bonne classification (%)
HON4D (Oreifej et Liu, 2013)	82.15 ± 4.18
$SVMK_{dtwr}$, 15 postures	**82.50 ± 3.22**

TAB. 2 – *Résultats comparés sur les bases HDM05 et MSRAction3D.*

quadratique avec la réduction du nombre de postures sur l'axe temporel. L'élasticité des noyaux apporte un gain de performance important (comparativement à des noyaux à base de distance Euclidienne) lorsque les données sont caractérisées par une grande variabilité. Nos résultats montrent qu'un SVM à base de DTW régularisée est très compétitif vis-à-vis de l'état de l'art pour les 2 jeux de données testés, même lorsque la réduction de dimension sur l'axe temporel est importante. Cette étude ouvre donc des perspectives notamment vers l'exploitation d'autres noyaux élastiques plus sophistiqués (Marteau, 2009) et des techniques d'échantillonnage adaptatif (Marteau et Gibet, 2005) (Marteau et Ménier, 2009) susceptibles d'extraire les postures les plus significatives d'une séquence de mouvement. La réduction de dimension conjointe sur les axes spatial et temporel est aussi un enjeu important.

Références

Bissacco, A., A. Chiuso, et S. Soatto (2007). Classification and recognition of dynamical models : The role of phase, independent components, kernels and optimal transport. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*(11), 1958–1972.

Chang, C.-C. et C.-J. Lin (2001). *LIBSVM : a library for support vector machines*.

Cuturi, M., J.-P. Vert, O. Birkenes, et T. Matsui (2007). A Kernel for Time Series Based on Global Alignments. In *Proceedings of ICASSP'07*, Honolulu, HI, pp. II–413 – II–416. IEEE.

de Aguiar, E., C. Theobalt, et H.-P. Seidel (2006). Automatic learning of articulated skeletons from 3d marker trajectories. In B. et al. (Ed.), *ISVC*, Volume 4291 of *Lecture Notes in Computer Science*, pp. 485–494. Springer.

Fothergill, S., H. Mentis, P. Kohli, et S. Nowozin (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, pp. 1737–1746. ACM.

Han, L., X. Wu, W. Liang, G. Hou, et Y. Jia (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Comput. 28*(5), 836–849.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, University of California, Santa Cruz. Technical Report.

He, X. et P. Niyogi (2003). Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press.

Hussain, S. et A. Rashid (2012). User independent hand gesture recognition by accelerated dtw. In *Int. Conf. on Informatics, Electronics Vision (ICIEV)*, pp. 1033–1037.

Hussein, M. E., M. Torki, M. A. Gowayyed, et M. El-Saban (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*.

Keogh, E. J. et M. J. Pazzani (2000). Scaling up dynamic time warping for datamining applications. In *Proc. of the Sixth ACM SIGKDD*, KDD '00, New York, NY, USA, pp. 285–289.

Larochelle, H., M. Mandel, R. Pascanu, et Y. Bengio (2012). Learning algorithms for the classification restricted boltzmann machine. *J. of Machine Learning Research 13*, 643–669.

Li, W., Z. Zhang, et Z. Liu (2010). Action recognition based on a bag of 3d points. In I. C. Press (Ed.), *Proc. IEEE Int'l Workshop on CVPR for Hum. Comm. Behav. Analysis*, pp. 9–14.

Marteau, P. F. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Trans. Pattern Anal. Mach. Intell. 31*(2), 306–318.

Marteau, P. F. et S. Gibet (2005). Adaptive sampling of motion trajectories for discrete task-based analysis and synthesis of gesture. In *LNAI Proc. of Int. Gesture Workshop*, pp. 224–235. Springer.

Marteau, P.-F. et S. Gibet (2013). Constructing positive elastic kernels with application to time series classification. Technical report, UMR 6074 IRISA, http ://arxiv.org/abs/1005.5141.

Marteau, P.-F. et G. Ménier (2009). Speeding up simplification of polygonal curves using nested approximations. *Pattern Anal. Appl. 12*(4), 367–375.

Martens, J. et I. Sutskever (2011). Learning recurrent neural networks with hessian-free optimization. In *ICML*, pp. 1033–1040.

Masoud, O. et N. Papanikolopoulos (2003). A method for human action recognition. *Image Vision Comput. 21*(8), 729–743.

Mitra, S. et T. Acharya (2007). Gesture recognition : A survey. *Trans. Sys. Man Cyber Part C 37*(3), 311–324.

Müller, M., T. Röder, M. Clausen, B. Eberhardt, B. Krüger, et A. Weber (2007). Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn.

O'Brien, J. F., R. E. Bodenheimer, G. J. Brostow, et J. K. Hodgins (2000). Automatic joint parameter estimation from magnetic motion capture data. In *Proceedings of Graphics Interface 2000*, pp. 53–60.

Ofli, F., R. Chaudhry, G. Kurillo, R. Vidal, et R. Bajcsy (2012). Sequence of the most informative joints (smij) : A new representation for human skeletal action recognition. In *CVPR Workshops*, pp. 8–13. IEEE.

Ofli, F., R. Chaudhry, G. Kurillo, R. Vidal, et R. Bajcsy (2013). Sequence of the most informative joints (smij) : A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation 0*(0), 1–20.

Oreifej, O. et Z. Liu (2013). Hon4d : Histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE ICPR*.

Sakoe, H. et S. Chiba (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the 7th International Congress of Acoustic*, pp. 65–68.

Sempena, S., N. Maulidevi, et P. Aryan (2011). Human action recognition using dynamic time warping. In *Int. Conf. on Electrical Engineering and Informatics (ICEEI)*, pp. 1–5.

Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, et A. Blake (2011). Real-time human pose recognition in parts from single depth images. In *Conf. on Computer Vision and Pattern Recognition*, CVPR '11, pp. 1297–1304. IEEE.

Tenenbaum, J. B., V. de Silva, et J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319.

Veeraraghavan, A., A. K. R. Chowdhury, et R. Chellappa (2004). Role of shape and kinematics in human movement analysis. In *CVPR (1)*, pp. 730–737.

Velichko, V. M. et N. G. Zagoruyko (1970). Automatic recognition of 200 words. *International Journal of Man-Machine Studies 2*, 223–234.

Wang, J., Z. Liu, Y. Wu, et J. Yuan (2012). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE int. conf. CVPR*, pp. 1290–1297.

Wang, S. B., A. Quattoni, L. Morency, D. Demirdjian, et T. Darrell (2006). Hidden conditional random fields for gesture recognition. In *IEEE int. conf. CVPR*, Volume 2, pp. 1521–1527.

Yu, E. et J. Aggarwal (2009). Human action recognition with extremities as semantic posture representation. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 0*, 1–8.

Zhao, X., Z. Song, J. Guo, Y. Zhao, et F. Zheng (2012). Real-time hand gesture detection and recognition by random forest. In M. Zhao et J. Sha (Eds.), *Communications and Information Processing*, Volume 289, pp. 747–755. Springer Berlin Heidelberg.

Summary

In the field of human gestural action recognition, many studies have focused on dimensionality reduction along the spatial axis to reduce both the computational complexity and variability of gestural utterances. It is quite surprising that very few of these methods have explicitly addressed the dimension reduction along the time axis. In terms of complexity, dimensionality reduction on this axis is a major issue regarding the usability of elastic distances since they have a quadratic complexity. Moreover, taking into account the variability along the time axis has demonstrated to be very useful to improve recognition scores. In an attempt to highlight dimension reduction along the time axis, we present in this paper an approach based on temporal down-sampling associated to elastic kernel machine learning. We show experimentally, on two data sets widely referenced in the domain of human action recognition and very opposed in terms of motion capture quality, that it is possible to significantly reduce the number of frames characterizing the skeleton trajectories while maintaining the recognition performances at the the state-of-the-art level on these datasets. Gain obtained in computational complexity makes such an approach eligible for *real-time* applications.