Reconnaissance d'actions humaines 3D par l'analyse de forme des trajectoires de mouvement.

Maxime Devanne[1,2,3], Hazem Wannous[1], Stefano Berretti[3], Pietro Pala[3], Mohamed Daoudi[1,2] et Alberto Del Bimbo[3]

[1]Université de Lille 1 - LIFL (UMR Lille1/CNRS 8022)
[2]Institut Mines-Telecom
[3]Université de Florence, Italie

Résumé

La reconnaissance d'actions humaines dans des séquences vidéo 3D est un problème important, actuellement au cœur de nombreux domaines de recherche comme la vidéo surveillance, les interfaces Homme-Machine et la rééducation. Le développement d'algorithmes de reconnaissance d'actions précis et efficaces est une tâche difficile à cause des fortes variabilités des formes humaines, des vêtements et du mouvement. Dans ce papier, nous proposons un nouvel outil permettant de représenter de manière compacte, de comparer et de reconnaître des actions humaines capturées à partir de caméras de profondeur. Dans un premier temps, les coordonnées 3D de chaque articulation du squelette humain sont considérées comme une chaîne de mouvement. L'évolution spatiale et temporelle de ce vecteur caractéristique est ensuite représentée par une trajectoire dans l'espace des actions. Grâce à cette représentation basée sur les articulations 3D, nous sommes capable de capturer simultanément aussi bien l'apparence géométrique du corps humain que sa dynamique au cours du temps. Le problème de reconnaissance d'actions est ensuite formulé comme un problème de recherche de similarités entre la forme des trajectoires dans une variété riemannienne. La classification par l'algorithme des k-plus-proches-voisins est ensuite effectuée sur la variété pour bénéficier de la géométrie riemannienne dans l'espace des formes. Notre méthode est évaluée sur deux bases de données publiques. En comparaison avec les méthodes existantes dans l'état de l'art, les résultats obtenus montrent l'efficacité de l'approche proposée avec un taux supérieur à 91% *sur les deux bases de données.*

Abstract

Recognizing human actions in a 3D video sequence is an important open problem, which is currently at the heart of many research domains including surveillance, Human-Machine interfaces and rehabilitation.Developing algorithms for action recognition that are both accurate and efficient is challenging due to the variability of the human shape, clothing and motion. In this paper, we propose a new framework which allows compact representation, quick comparison and accurate recognition of human actions in video sequences from depth sensors. Initially, the 3D coordinates of the joints of a human skeleton are considered as one motion channel and the spatial and temporal evolution of this feature vector is represented as a trajectory in an action space. Thanks to such a 3D joint-based framework, we are able to capture both the geometric appearance and the dynamics of the human body simultaneously. The action and activity recognition problem is then formulated as the problem of computing the similarity between the shape of trajectories in a Riemannian manifold. Classification using kNN is finally performed on this manifold taking benefit from Riemannian geometry in the open curve shape space. Experiments on two action datasets, namely MSR Action 3D and UTKinect, are performed. Compared to state-of-the-art methods, results show high performance, above 91%*, on the two challenging datasets.*

Mots clé : Reconnaissance d'actions 3D, modélisation temporelle, espace des formes, variété riemannienne.

1. Introduction

Les technologies de l'imagerie ont récemment montré un avancement conséquent avec l'apparition de nouvelles caméras de profondeur comme la Kinect de Microsoft [Mic13] ou l'Asus Xtion PRO LIVE [ASU13]. Ces nouveaux périphériques d'acquisition ont stimulé le développement de divers applications prometteuses comme l'estimation et la reconstruction de la posture [SFC*11], l'estimation du flot de scène [HB11], la reconnaissance de gestes [RYZ11] et la super-résolution de visages [BDP12]. Une étude récente sur

les applications basées sur les caméras de profondeur peut être trouvée dans [HSXS13]. Les résultats encourageant montrés dans ces travaux peuvent être en partie expliqués par les avantages apportés par de telles caméras de profondeur, comme la segmentation premier-plan/arrière-plan et la robustesse aux changements de conditions d'éclairage. Par conséquent, plusieurs bibliothèques permettant la détection et le suivi du corps humain en temps réel ont vu le jour. Alors que ces méthodes d'extraction et de représentation du corps humain par des silhouettes ou des squelettes ont évolué rapidement, les techniques interprétant la dynamique de ces données, afin de comprendre les actions observées, sont assez limitées. Cette tâche est notamment compliquée par la nécessité d'être invariant aux transformations géométriques ainsi qu'à la vitesse d'exécution de l'action. De plus, d'autres défis importants comme les données bruitées et la variété de poses au sein d'un même type d'action rendent la reconnaissance d'actions d'autant plus difficile.

La reconnaissance d'actions humaines se basant sur l'analyse de données fournies par des caméras de profondeur ont attiré de nombreux groupes de recherche dans les dernières années. Les approches décrites par la suite peuvent être groupées en trois principales catégories selon la méthode d'utilisation de l'information de profondeur : les méthodes basées squelette, les méthodes basées carte de profondeur et les méthodes hybrides qui combinent et exploitent les deux types d'information. Suivant cette catégorisation, les méthodes existantes pour la reconnaissance d'action à partir de caméra de profondeur sont analysées par la suite.

Les approches basées squelette sont devenues populaires suite au travail de Shotton et al. [SFC*11] où une méthode de prédiction précise en temps réel des positions des articulations 3D du corps humain à partir de cartes de profondeur est proposée. S'appuyant sur la position de ces articulations, [XCA12] propose une approche qui calcule des histogrammes des positions de 12 articulations comme une représentation compacte de la posture. Ces histogrammes de posture sont ensuite regroupés en k mots visuels. L'évolution temporelle de ces mots visuels est modélisée par un model de Markov caché. Dans [YT12], la reconnaissance d'actions humaines est obtenue par l'extraction de trois caractéristiques pour chaque articulation basées sur la différence deux à deux des positions des articulations : dans la trame actuelle, entre la trame actuelle et la précédente et entre la trame actuelle et la première de la séquence représentant la posture neutre. L'analyse des composantes principales (PCA) est utilisée pour réduire les redondances et le bruit de ces caractéristiques, et ainsi obtenir une représentation compacte appelée *EigenJoints* pour chaque trame. Finalement, un classifieur bayésien naïf est utilisé pour la classification multi-classes.

Les méthodes basées sur les cartes de profondeur s'appuient sur l'extraction de descripteurs à partir de l'ensemble des points de l'image de profondeur. La modélisation de la dynamique de l'action est ainsi un défi important résolu de manière différente selon les approches. L'approche proposée dans [LZL10] emploie des silhouettes 3D pour décrire les postures et utilise un modèle graphique pour modéliser la dynamique de l'action. Dans [YZT12], la dynamique est décrite par l'intermédiaire de cartes de mouvement de profondeur qui mettent en avant les zones de la scène où un mouvement est effectué. D'autres méthodes proposent de travailler dans un espace 4D divisé en cellules spatiotemporelles pour y extraire des caractéristiques représentant l'apparence de profondeur comme *Spatio-Temporal Occupancy Pattern* [VNO*12], *Random Occupancy Pattern* [WLC*12] and *Depth Cuboid Similarity Feature* [XA13]. Enfin le travail dans [OL13] propose de quantifier l'espace 4D en utilisant les sommets d'un polychore puis en modélisant la distribution des vecteurs normaux pour chaque cellule.

Les solutions hybrides combinent les informations issues des deux flux (squelette et carte de profondeur) pour modéliser l'action. Wang et al. [WLWY12] propose de calculer un descripteur appelé *Local Occupancy Pattern* autour de chaque articulation. Dans [OBT13], les actions sont caractérisées par la combinaison des angles entre articulations calculés sur les squelettes et les histogrammes de gradients orientés calculés sur l'image de profondeur.

Ces approches, basées sur les données de profondeur, bénéficient des nombreux travaux réalisés sur la reconnaissance d'action à partir de vidéos couleur 2D [WRB11, TCSU08, BTR12, Pop10]. Outre les méthodes euclidiennes [SZTL14], des récentes techniques reformulent de manière intéressante le problème de reconnaissance d'actions à travers des espaces non-euclidiens comme les variétés riemanniennes [VRCC05, HSWL12, AAASC11, Lui12], que nous souhaitons exploiter dans ce papier.

2. Vue d'ensemble de l'approche

Une action humaine est naturellement caractérisée par l'évolution du corps humain au cours du temps. Les données de squelettes contenant la position 3D des différentes parties du corps fournissent une représentation précise de la posture du corps humain. Ces caractéristiques peuvent facilement être extraites et suivies à partir des cartes de profondeur. De plus elles fournissent une information locale sur le corps humain. Cependant, même si les positions 3D précises des différentes articulations sont disponibles, la tâche de reconnaissance d'actions reste difficile à cause de variations temporelles et spatiales dans la manière d'effectuer une action.

Ces considérations nous motivent à aborder le problème de la reconnaissance d'action en proposant une approche basée sur l'analyse de l'évolution des articulations du squelette au cours de la séquence vidéo. Pour cela, nous modélisons un squelette par un vecteur multidimensionnel obtenu en concaténant les coordonnées 3D de ses articulations. Ensuite nous considérons la trajectoire que ce vecteur décrit dans l'espace euclidien multidimensionnel modélisant la dynamique de l'ensemble des articulations. Ces trajectoires sont ensuite interprétées dans une variété riemannienne afin de comparer leur forme par l'intermédiaire d'un recalage temporelle dans *l'espace des formes*. De ce fait, nous

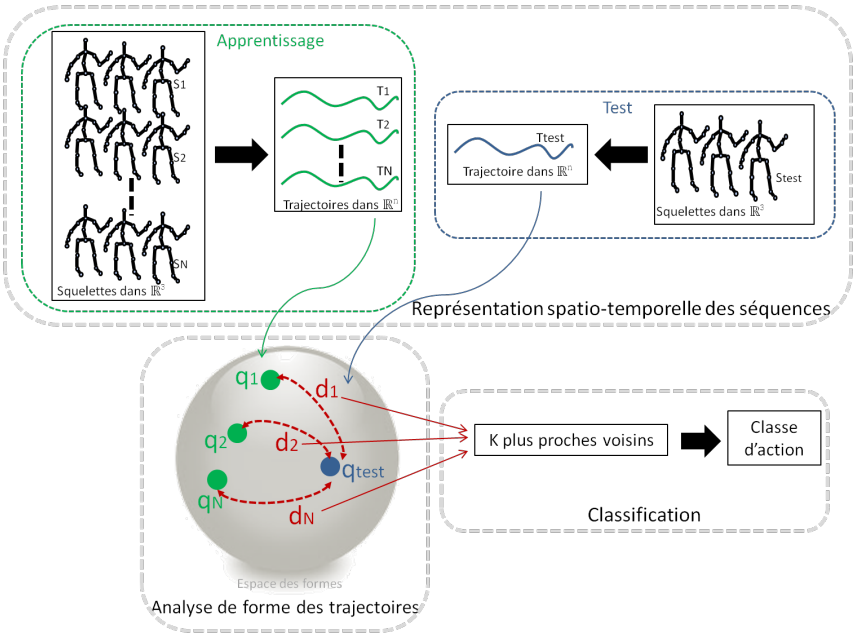

Figure 1: Vue d'ensemble de notre approche. Tout d'abord les séquences de squelettes sont modélisées par des trajectoires dans *l'espace des actions*. Ces trajectoires sont ensuite représentées dans *l'espace des formes*. La reconnaissance est finalement conduite grâce à l'algorithme des *k*-plus-proches-voisins sur cette variété.

reformulons le problème de reconnaissance d'action par une analyse statistique dans *l'espace des formes*. Une métrique élastique est utilisée sur cette variété pour comparer la forme des trajectoires. Cette distance permet à l'approche d'être invariante à l'élasticité des trajectoires. En d'autres termes, elle nous permet de faire face à un défi important de la reconnaissance d'actions : l'invariance à la vitesse d'exécution de l'action. La figure 1 illustre l'approche proposée.

Le reste du papier est organisé comme suit : La section 3 décrit la représentation spatiotemporelle d'une action par une trajectoire. La section 4 introduit l'outil riemannien utilisé pour l'analyse et la comparaison de la forme des trajectoires. Dans la section 5, nous présentons un outil statistique sur la variété riemannienne ainsi que l'algorithme utilisé pour la classification. La section 6 présente les différents résultats obtenus sur deux bases de données publiques. Enfin, la section 7 conclut le papier et discute de futures directions de recherche.

3. Représentation dans l'espace des actions

Grâce aux caméras de profondeur, un squelette humanoïde 3D peut être efficacement extrait à partir des cartes de profondeur depuis l'apparition du travail de Shotton et al. [SFC*11]. Ces squelettes contiennent la position 3D d'un certain nombre d'articulations représentant différentes parties du corps humain. Le nombre d'articulations estimées dépend de l'outil utilisé en combinaison avec le périphérique. Les squelettes extraits grâce au SDK de Microsoft contiennent 20 articulations alors que ceux extraits à partir du SDK de PrimeSense NiTE n'en contiennent que 15. Pour chaque trame t d'une séquence, la position 3D de chaque articulation i est représentée par trois coordonnées exprimées dans le système de référence de la caméra $p_i(t) = (x_i(t), y_i(t), z_i(t))$. Afin de garantir une invariance aux transformations géométriques (rotation et translation), nous alignons l'ensemble des squelettes par rapport à un squelette de référence en calculant la matrice de transformation optimale entre les squelettes. Soit N_j le nombre d'articulations contenues dans un squelette, le vecteur caractéristique à la trame t est défini comme :

$$v(t) = [x_1(t)\ y_1(t)\ z_1(t)\ \ldots\ x_{N_j}(t)\ y_{N_j}(t)\ z_{N_j}(t)]^T . \quad (1)$$

La taille d'un tel vecteur caractéristique est $3N_j$. Pour une séquence de N_f trames, un nombre correspondant de vecteurs colonne est défini et concaténé pour obtenir une matrice caractéristique décrivant la séquence entière :

$$M = \begin{pmatrix} v(1) & v(2) & \ldots & v(N_f) \end{pmatrix} . \quad (2)$$

Cette matrice caractéristique représente l'évolution de la posture au cours du temps, où chaque vecteur colonne v est vu comme un échantillon d'une trajectoire continue dans R^{3N_j}, représentant l'action dans un espace de $3N_j$ dimensions appelé *espace des actions*.

4. Analyse dans l'espace des formes

Une action est une séquence de poses et peut être vue comme le résultat d'un échantillonnage d'une trajectoire continue dans *l'espace des actions*. La trajectoire est définie comme le mouvement au cours du temps des points caractéristiques encodant les coordonnées 3D des articulations du squelette. Soit une trajectoire dans *l'espace des actions* représentée comme une fonction $\beta : I \rightarrow \mathbb{R}^n$, pour $I = [0,1]$. Pour analyser la forme de β, nous représentons la trajectoire

par la *square-root velocity function* (SRVF) $q : I \to \mathbb{R}^n$, définie comme :

$$q(t) \doteq \frac{\dot{\beta}(t)}{\sqrt{\|\beta(t)\|}} , \tag{3}$$

$q(t)$ est une fonction particulière introduite dans [JKSJ07] qui capture la forme de β tout en offrant des facilités de calcul. Comme montré dans [JKSJ07], la norme $\mathbb{L}^2$ représente la métrique pour comparer la forme de deux trajectoires. L'ensemble de toutes les trajectoires, noté $\mathcal{C}$, est ainsi défini comme :

$$\mathcal{C} = \{q : I \to \mathbb{R}^n | \ \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^n) , \tag{4}$$

Avec la norme $\mathbb{L}^2$ sur son plan tangent, cette variété $\mathcal{C}$ devient alors une variété riemannienne. Chaque élément de $\mathcal{C}$ représente une trajectoire. On définit la distance entre deux éléments q_1 et q_2 par la longueur du chemin géodésique entre q_1 et q_2 sur la variété $\mathcal{C}$. Comme ces éléments ont une norme $\mathbb{L}^2$ unitaire, $\mathcal{C}$ peut être vu comme une hypersphère de l'espace de Hilbert. Ainsi la distance géodésique entre q_1 et q_2 est définie comme :

$$d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle) . \tag{5}$$

Afin de garantir l'invariance à la vitesse d'exécution de l'action, nous devons comparer la forme des trajectoires indépendamment de leur élasticité. Cela nécessite une invariance à la re-paramétrisation des trajectoires. Nous définissons la classe d'équivalence de la forme q par :

$$[q] = \{\sqrt{\dot{\gamma}(t)}(q \circ \gamma)(t) | \ \gamma \in \Gamma\} . \tag{6}$$

où Γ est le groupe de re-paramétrisation. Tous les éléments de cette classe d'équivalence, auxquels est associée une certaine re-paramétrisation γ, sont équivalents à la forme de q. L'ensemble de telles classes d'équivalences est appelé *espace des formes* et noté $\mathcal{S}$. La comparaison entre deux éléments q_1 et q_2 requiert une comparaison entre leur classe d'équivalence $[q_1]$ et $[q_2]$. La distance géodésique dans $\mathcal{S}$ devient ainsi :

$$d_s([q_1], [q_2]) = d_c(q_1, q_2^*) . \tag{7}$$

où q_2^* est l'élément q_2 re-paramétré par rapport à q_1. En pratique, la programmation dynamique est utilisée pour trouver la re-paramétrisation optimal entre deux éléments.

5. Reconnaissance d'action dans l'espace des formes

La méthode proposée pour la reconnaissance d'actions est basée sur l'algorithme des k-plus-proches-voisins appliqué dans *l'espace des formes* sur l'ensemble des données d'apprentissage ou sur des séquences représentatives calculées à l'aide de la moyenne de Karcher [Kar77].

5.1. Calcul de trajectoires moyennes

Un des avantages de l'utilisation d'une approche riemannienne pour la reconnaissance d'action est que cela nous permet d'exploiter des outils statistiques sur les éléments de la variété. Par exemple, nous pouvons utiliser la notion de moyenne de Karcher [Kar77] pour calculer des trajectoires moyennes à partir d'un ensemble de trajectoires. Ainsi, une trajectoire moyenne peut être d'une part calculée à partir d'un ensemble de trajectoires différentes pour représenter une trajectoire intermédiaire. D'autre part, elle peut être calculée à partir d'un ensemble de trajectoires similaires afin d'obtenir un modèle moyen qui peut être vue comme une trajectoire représentative de l'ensemble. De plus, pour classifier une séquence, la distance géodésique doit être calculée avec toutes les séquences d'apprentissage. Pour un grand nombre de séquences d'apprentissage, cela implique un temps de calcul élevé. Utiliser des trajectoires moyennes représentatives peut ainsi diminuer le nombre de séquences d'apprentissage et donc le temps de calcul. Pour un ensemble de trajectoires d'apprentissage représentées dans *l'espace des formes* $q_1, \ldots, q_n$, leur moyenne de Karcher peut être définie comme :

$$\mu = \arg\min \sum_{i=1}^{n} d_s([q], [q_i])^2 . \tag{8}$$

La figure 2 présente un exemple de calcul de moyenne de Karcher pour cinq trajectoires ($q_1 \ldots q_5$). Dans l'étape initiale, q_1 est sélectionné comme la trajectoire moyenne. De manière itérative, la moyenne est mise à jour grâce à la métrique élastique calculée entre toutes les trajectoires q_i. Après convergence, la trajectoire moyenne est donné par q_m.

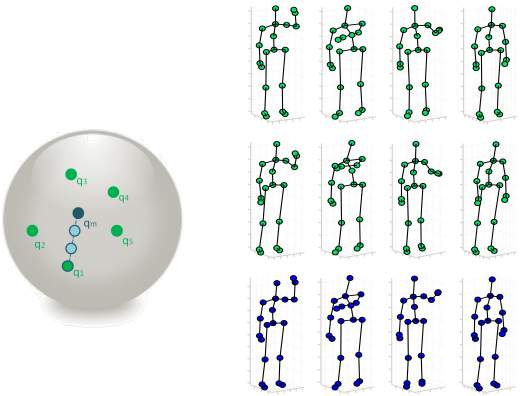

Figure 2: Calcul de la moyenne de Karcher entre 5 trajectoires représentées dans *l'espace des formes* (à gauche) et dans *l'espace des actions* (à droite). Les deux premières séquences correspondent aux trajectoires q_1 et q_2 tandis que la séquence du bas correspond à la trajectoire moyenne q_m.

En calculant une telle trajectoire moyenne pour chaque classe, nous supposons qu'il existe uniquement un seul moyen d'effectuer chaque action, ce qui n'est pas forcément vrai dans tous les cas. Par exemple, une personne gauchère et une personne droitière effectueront différemment une même action. Dans ce cas, calculer une unique trajectoire moyenne pour cette action peut donner une trajectoire non représentative de l'action. Pour cette raison, nous calculons des trajectoires moyennes au sein d'une même classe pour chaque sujet séparément. Au lieu d'avoir une seule trajectoire modèle par classe, nous avons maintenant un modèle par action et par sujet. Ainsi, les différents moyens d'effectuer une même

action sont maintenant séparés et les trajectoires moyennes résultantes sont des trajectoires représentatives des actions.

5.2. k-plus-proches-voisins avec la distance élastique

Soit $\{(X_i,y_i)\}$, $i = 1,\ldots,N$, un ensemble d'apprentissage où X_i appartient à *l'espace des formes* $\mathcal{S}$, et y_i est le label de classe prenant une valeur de $\{1,\ldots,N_c\}$, avec N_c le nombre de classes. L'objectif est de trouver une function $F(X) : \mathcal{S} \longmapsto \{1,\ldots,N_c\}$ pour regrouper les données représentées dans *l'espace des formes* à partir des données labélisées de l'ensemble d'apprentissage. Pour cela, nous proposons d'utiliser le classifieur des k-plus-proches-voisins sur la variété riemannienne appris sur des séquences représentées dans *l'espace des formes*. Cette méthode d'apprentissage nous permet d'exploiter les propriétés géométriques de *l'espace des formes*, et plus particulièrement sa métrique élastique. La classification repose sur le calcul des distances géodésiques entre une trajectoire de test et l'ensemble des trajectoires d'apprentissage. Plus précisément, un ensemble de trajectoires d'apprentissage $X_i : i = 1,\ldots,N$ est représenté comme un ensemble d'éléments $q_i : i = 1,\ldots,N$ dans *l'espace des formes*. Ensuite, pour une séquence de test représentée dans *l'espace des formes*, les distances géodésiques avec l'ensemble des séquences d'apprentissage sont calculées pour trouver les k séquences les plus similaires. Le label de classe associé à la séquence de test est celui le plus représenté parmis les k plus proches séquences d'apprentissage.

6. Résultats expérimentaux

La performance de notre approche est évaluée et comparée avec les méthodes existantes de l'état de l'art sur deux bases de données publiques : MSR Action 3D [LZL10] et UTKinect [XCA12].

6.1. MSR Action 3D

Cette base de donnée publique a été collectée par Microsoft Research [LZL10] et peut être vue comme une base référence pour la reconnaissance d'actions, utilisée dans de nombreux travaux. Elle inclut 20 actions effectuées par 10 personnes 2 ou 3 fois. Au total, 567 séquences sont disponibles. Les différentes actions orientées jeux vidéo sont choisies pour couvrir différentes variations du mouvement des bras, des jambes, du torse et de leur combinaison. Chaque sujet est positionné au centre de la scène face à la caméra. Il était demandé aux personnes d'effectuer les actions avec le bras droit ou la jambe droite lorsque l'action nécessite un seul membre. De plus, toutes les actions sont effectuées sans interactions avec des objets. Les deux principaux défis de cette base de données sont la forte similarité entre certaines actions et la variation de vitesse d'exécution de l'action. Pour chaque séquence les informations de couleur, de profondeur et de squelette sont fournies. Dans notre cas, seules les données de squelette sont utilisées. Comme reporté dans [WLWY12], 10 actions ne sont pas utilisées dans les tests car les squelettes sont soit manquants, soit trop bruités. Pour nos expérimentations, nous utilisons donc 557 séquences.
Nous testons notre approche avec ses différentes méthodes mentionnées dans la section 5.1. Les résultats sont reportés dans la table 1.

Table 1: MSR Action 3D. Nous testons notre approche avec ses différentes méthodes de classification ($k$pp, $k$pp et moyenne de Karcher par action, $k$pp et moyenne de Karcher par action et par sujet.).

Methode	Taux (%)
kpp	88.3
kpp & moyenne de Karcher par action	89.0
kpp & moyenne de Karcher par action/subject	**92.1**

En analysant les résultats, nous pouvons remarquer que le meilleur taux de reconnaissance est obtenu en utilisant la notion de moyenne de Karcher par action et par sujet. En comparaison avec l'utilisation de la moyenne de Karcher par action uniquement, on peut voir que séparer les différentes façons d'effectuer une même action permet d'augmenter le taux de reconnaissance. De plus, les résultats montrent que l'utilisation de trajectoires moyennes est plus efficace que d'utiliser l'ensemble des séquences d'apprentissage. Cela peut s'expliquer pour le cas d'actions similaires. Dans ce cas, une séquence appartenant à une première classe peut être très proche de séquences appartenant à une seconde classe, et ainsi sélectionnée comme un faux positif lors de la classification. Le calcul de trajectoires moyennes peut ainsi augmenter la distance inter-classes et donc améliorer le taux de classification. Par exemple, les deux premières actions de la base (*high arm wave* and *horizontal high arm wave*) sont très proches. Utiliser de telles trajectoires moyennes permet de réduire la confusion entre ces deux actions. Ceci peut être visualiser dans la figure 3 représentant la matrice de confusion pour les deux méthodes de classification différentes.

Dans un deuxième temps, nous comparons notre approche aux autres méthodes de l'état de l'art. Les résultats sont reportés dans la table 2. Pour une comparaison équitable, nous utilisons le même protocole expérimental que les travaux évalués sur la base MSR Action 3D. Les cinq premiers sujets sont choisis pour l'apprentissage, les cinq autres pour le test. En analysant les résultats, nous pouvons voir que notre méthode dépasse les méthodes de l'état de l'art exceptée celle proposée dans [OBT13]. Cependant, cette aproche utilise une combinaison des informations de squelette et de profondeur. Ils reportent qu'en utilisant uniquement les données de squelette, leur méthode donne un taux de reconnaissance de 83.5% plus bas que le taux obtenu avec notre approche.

Pas la suite, nous conduisons les mêmes expérimentations avec toutes les combinaisons possibles de choisir la moitié des sujets comme apprentissage et l'autre moitié comme test. Pour chacune des 252 combinaisons, nous utilisons la moyenne de Karcher par action et par sujet pour l'ensemble d'apprentissage. Nous obtenons un taux de reconnaissance moyen de $87.28 \pm 2.41\%$ (moyenne $\pm$ écart-type). Parmi les 252 combinaisons, le plus bas taux de reconnaissance obtenu est de 81.31% alors que le plus élevé est de 93.04%. En comparaison avec le travail présenté dans [OL13], où le taux moyen est aussi calculé pour toutes les combinaisons possibles, nous dépassons leur résultat de $82.15 \pm 4.18\%$.

Table 2: MSR Action 3D. Comparaison de notre méthode avec les méthodes les plus pertinentes de l'état de l'art.

Methode	Taux (%)
EigenJoints [YT12]	82.3
STOP [VNO*12]	84.8
DMM & HOG [YZT12]	85.5
Random Occupancy Pattern [WLC*12]	86.5
Actionlet [WLWY12]	88.2
DCSF [XA13]	89.3
JAS & HOG^2 [OBT13]	**94.8**
HON4D [OL13]	88.9
Ours	**92.1**

De plus, la faible valeur de l'écart-type dans nos expérimentations montre que notre méthode est très peu dépendante des données choisies pour l'apprentissage. Afin de présenter le taux de reconnaissance obtenu par notre méthode pour chaque action séparément, les matrices de confusion sont calculées pour chacun des cas et présentées dans la figure 3.

Nous pouvons remarquer qu'un très faible taux de reconnaissance est obtenu pour les actions *hammer* et *hand catch*. Cela peut être expliqué par le fait que ces actions sont similaires à d'autres actions. De plus, la façon d'exécuter ces actions varie considérablement selon les sujets. Par exemple, pour l'action *hammer*, les sujets de l'ensemble d'apprentissage ne donnent qu'un seul coup de marteau alors que certains sujets de l'ensemble de test en donnent plusieurs. Dans ce cas, les formes des trajectoires sont très différentes et les séquences correspondantes ne sont pas détectées comme similaires. La figure 4 illustre cet exemple. Comme il est difficile de visualiser les trajectoires dans un espace de grande dimension, ces dernières sont ici représentées en trois dimensions correspondant à une seule articulation (main droite). Les quatre trajectoires correspondent à des échantillons différents de l'action *hammer* où un seul coup de marteau est donné pour les deux premiers cas alors que plusieurs coups sont donnés pour les deux derniers cas. Nous pouvons remarquer que la forme des trajectoires est ainsi différente.

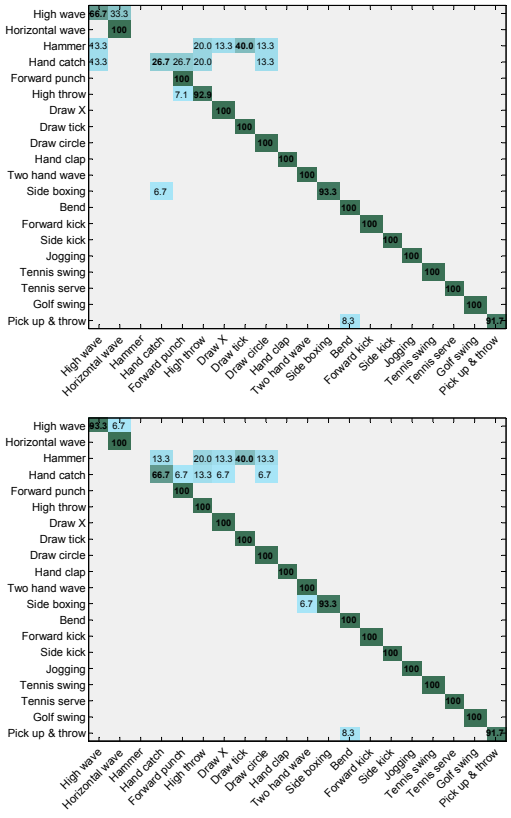

Figure 3: MSR Action 3D. Matrices de confusion obtenues avec notre approche. La classification est effectuée à l'aide le l'algorithme des k-plus-proches-voisins apris sur l'ensemble des séquences d'apprentissage (en haut) et sur des trajectoires moyennes par action et par sujet calculées grâce à la moyenen de Karcher (en bas).

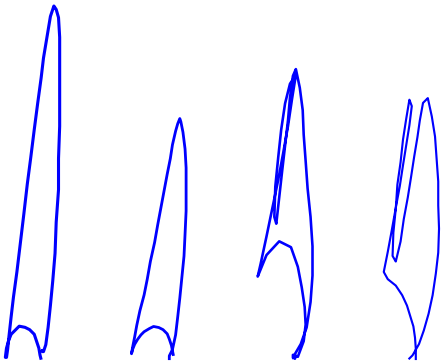

Figure 4: Visualisation d'un cas d'échec pour l'action *hammer*. Quatre trajectoires de la main droite sont représentées. Un seul coup de marteau est donnée pour les deux premières trajectoires tandis que deux coups sont donnés pour les deux trajectoires de droite.

6.2. UTKinect

Afin de confirmer l'efficacité de notre approche, nous proposons également de l'évaluer sur une seconde base de données appelée UTKinect [XCA12] qui présente d'autres défis. Dans cette base, 10 sujets effectuent 10 actions différentes deux fois pour un total de 200 séquences. La base présente trois défis principaux : tout d'abord, les actions sont capturées de différents points de vue ; ensuite, certaines actions nécessitent une interaction avec des objets ; enfin, une autre difficulté est ajoutée avec la présence d'occultations causées par les objets de la scène ou par le champ de vision restreint.

Pour être comparable avec le travail dans [XCA12], nous suivons le même protocole expérimental (leave-one-out-cross-validation). A chaque itération, une séquence est utilisée comme test et toutes les autres pour l'apprentissage. L'opération est répétée afin que chaque séquence soit utilisée une fois comme test. Nous obtenons un taux de reconnaissance de 91.5%, plus élevé que le taux reporté dans [XCA12] s'élevant à 90.9% . Cela montre que notre méthode est robuste aux changements de points de vue et aux occultations de certaines parties du corps. Cependant, en

analysant la matrice de confusion présentée dans la figure 5, nous pouvons remarquer que les plus faibles taux sont obtenus pour les actions utilisant des objets comme *carry* et *throw*. Ces actions sont confondues avec des actions similaires mais sans objet comme *walk* et *push*, respectivement. Cette limite est due au fait que notre approche ne prend en compte que les données de squelette. Ainsi, aucune information à propos de l'objet tenu par le sujet n'est disponible.

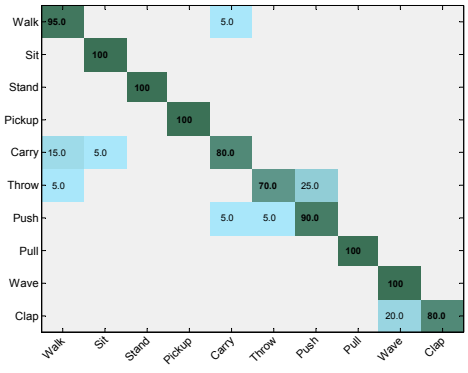

Figure 5: UTKinect. Matrice de confusion obtenue avec notre approche

7. Conclusion

Une approche efficace de reconnaissance d'actions humaines est proposée en utilisant une modélisation spatiotemporelle de trajectoires de mouvement dans une variété riemannienne. La position 3D de chaque articulation du squelette à chaque trame de la séquence est concaténée pour ainsi représenter l'action comme une trajectoire de mouvement dans *l'espace des actions*. Chaque trajectoire est ensuite exprimée comme un élément dans une variété riemannienne appelée *espace des formes*. Grâce à la géométrie riemannienne de la variété, la classification de l'action est résolue grâce l'algorithme des *k*-plus-proches-voisins en utilisant la distance élastique entre les formes des trajectoires. Les résultats expérimentaux sur deux bases de données MSR Action 3D et UTKinect démontrent que notre méthode dépasse les méthodes existantes de l'état de l'art dans la plupart des cas. Comme perspectives, nous envisageons tout d'abord d'intégrer dans notre approche d'autres descripteurs basés sur l'image de profondeur afin de gérer les cas d'interaction avec des objets. De plus, nous souhaitons approfondir les cas d'échec comme les répétitions de gestes pour fournir une approche plus robuste à ces variations. Enfin, nous réfléchissons à de possibles applications de notre travail notamment dans le domaine de la thérapie physique et de la rééducation assistée.

Références

[AAASC11] ABDELKADER M. F., ABD-ALMAGEED W., SRIVASTAVA A., CHELLAPPA R. : Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding. Vol. 115*, Num. 3 (2011), 439–455.

[ASU13] ASUS XTION PRO LIVE, : http ://www.asus.com/multimedia/xtion_pro/, 2013.

[BDP12] BERRETTI S., DEL BIMBO A., PALA P. : Superfaces : A super-resolution model for 3D faces. In *Proc. Work. on Non-Rigid Shape Analysis and Deformable Image Alignment* (Florence, Italy, Oct. 2012), pp. 73–82.

[BTR12] BIAN W., TAO D., RUI Y. : Cross-domain human action recognition. *IEEE Trans. on Systems, Man, and Cybernetics, Part B : Cybernetics. Vol. 42*, Num. 2 (avril 2012), 298–307.

[HB11] HADFIELD S., BOWDEN R. : Kinecting the dots : Particle based scene flow from depth sensors. In *Proc. Int. Conf. on Computer Vision* (Barcelona, Spain, Nov. 2011), pp. 2290–2295.

[HSWL12] HARANDI M. T., SANDERSON C., WILIEM A., LOVELL B. C. : Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In *Proc. IEEE Work. on the Applications of Computer Vision* (Washington, DC, USA, 2012), WACV'12, IEEE Computer Society, pp. 433–439.

[HSXS13] HAN J., SHAO L., XU D., SHOTTON J. : Enhanced computer vision with microsoft kinect sensor : A review. *IEEE Trans. on Cybernetics. Vol. 43*, Num. 5 (2013), 1318–1334.

[JKSJ07] JOSHI S. H., KLASSEN E., SRIVASTAVA A., JERMYN I. : A novel representation for riemannian analysis of elastic curves in R^n. In *Proc IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Minneapolis, MN, USA, June 2007), pp. 1–7.

[Kar77] KARCHER H. : Riemannian center of mass and mollifier smoothing. *Comm. on Pure and Applied Math.. Vol. 30* (1977), 509–541.

[Lui12] LUI Y. M. : Tangent bundles on special manifolds for action recognition. In *IEEE Trans. on Circuits and Systems for Video Technology* (2012), vol. 22, pp. 930–942.

[LZL10] LI W., ZHANG Z., LIU Z. : Action recognition based on a bag of 3D points. In *Proc. Work. on Human Communicative Behavior Analysis* (San Francisco, California, USA, June 2010), pp. 9–14.

[Mic13] MICROSOFT KINECT : http ://www.microsoft.com/en-us/kinectforwindows/, 2013.

[OBT13] OHN-BAR E., TRIVEDI M. M. : Joint angles similarities and HOG^2 for action recognition. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data* (Portland, Oregon, USA, June 2013), pp. 465–470.

[OL13] OREIFEJ O., LIU Z. : HON4D : Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proc. Int. Conf. on Computer Vision and*

Pattern Recognition (Portland, Oregon, USA, June 2013), pp. 716–723.

[Pop10] POPPE R. : A survey on vision-based human action recognition. *Image Vision Comput.. Vol. 28*, Num. 6 (juin 2010), 976–990.

[RYZ11] REN Z., YUAN J., ZHANG Z. : Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proc. ACM Int. Conf. on Multimedia* (Scottsdale, Arizona, USA, Nov. 2011), pp. 1093–1096.

[SFC*11] SHOTTON J., FITZGIBBON A., COOK M., SHARP T., FINOCCHIO M., MOORE R., KIPMAN A., BLAKE A. : Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Colorado Springs, Colorado, USA, June 2011), pp. 1–8.

[SZTL14] SHAO L., ZHEN X., TAO D., LI X. : Spatio-temporal laplacian pyramid coding for action recognition. *IEEE Trans. on Cybernetics. Vol. PP*, Num. 99 (2014), 1–1.

[TCSU08] TURAGA P., CHELLAPPA R., SUBRAHMANIAN V. S., UDREA O. : Machine recognition of human activities : A survey. *IEEE Trans. on Circuits and Systems for Video Technology. Vol. 18*, Num. 11 (novembre 2008), 1473–1488.

[VNO*12] VIEIRA A. W., NASCIMENTO E. R., OLIVEIRA G. L., LIU Z., CAMPOS M. F. : STOP : Space-time occupancy patterns for 3D action recognition from depth map sequences. In *Iberoamerican Congress on Pattern Recognition* (Buenos Airies, Argentina, Sept. 2012), pp. 252–259.

[VRCC05] VEERARAGHAVAN A., ROY-CHOWDHURY A., CHELLAPPA R. : Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 27*, Num. 12 (2005), 1896–1909.

[WLC*12] WANG J., LIU Z., CHOROWSKI J., CHEN Z., WU Y. : Robust 3D action recognition with random occupancy patterns. In *Proc. Europ. Conf. on Computer Vision* (Florence, Italy, Oct. 2012), pp. 1–8.

[WLWY12] WANG J., LIU Z., WU Y., YUAN J. : Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (Providence, Rhode Island, USA, June 2012), pp. 1–8.

[WRB11] WEINLAND D., RONFARD R., BOYER E. : A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding. Vol. 115*, Num. 2 (février 2011), 224–241.

[XA13] XIA L., AGGARWAL J. K. : Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data* (Portland, Oregon, USA, June 2013), pp. 2834–2841.

[XCA12] XIA L., CHEN C.-C., AGGARWAL J. K. : View invariant human action recognition using histograms of 3D joints. In *Proc. Work. on Human Activity Understanding from 3D Data* (Providence, Rhode Island, USA, June 2012), pp. 20–27.

[YT12] YANG X., TIAN Y. : Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proc. Work. on Human Activity Understanding from 3D Data* (Providence, Rhode Island, June 2012), pp. 14–19.

[YZT12] YANG X., ZHANG C., TIAN Y. : Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. ACM Int. Conf. on Multimedia* (Nara, Japan, Oct. 2012), pp. 1057–1060.