Une approche géométrique pour analyser l'intention sociale à partir du mouvement de marqueurs 3D

Paul Audain Desrosiers, Mohamed Daoudi, Maria-Francesca Gigliotti, Yann Coello

Une approche géométrique pour analyser l'intention sociale à partir du mouvement de marqueurs 3D

Paul Audain Desrosiers [1] , Mohamed Daoudi [2,3], Maria-Francesca Gigliotti[1], Yann Coello [1]
Université de Lille, Sciences Cognitives et Sciences Affectives (SCALab) - UMR 9193

[2] IMT Lille Douai, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France;
[3] Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189 CRIStAL, F-59000 Lille, France.

Résumé : *Dans ce papier, nous proposons un cadre géométrique capable de prédire en temps réel une intention sociale vs une intention personnelle. Le participant doit réaliser un ensemble de gestes comportant une intention sociale ou personnelle, en portant un gant contenant 4 marqueurs passifs. L'utilisation d'un système de capture de mouvement permet d'obtenir la trajectoire de la main du participant contenant les différents marqueurs 3D. Les données 3D obtenues sont définies dans un espace de forme de courbes ouvertes, puis analysées dans une variété Riemannienne. Nous avons obtenu un taux de reconnaissance moyen pour les deux gestes (intention sociale, personnelle) de 68%, ce qui est comparable au score moyen produit par l'évaluation humaine. Les résultats expérimentaux montrent également que le taux de classification pourrait être utilisé pour améliorer la communication sociale entre les agents humains et virtuels. A notre connaissance, il s'agit de la première étude en temps réel qui utilise des techniques de vision par ordinateur pour analyser l'effet de l'intention sociale sur l'action motrice afin d'améliorer la communication sociale entre un humain et un agent virtuel.*

Mots-clés : Intention sociale et personnelle, analyse des trajectoires, géométrie Riemannienne, motion capture (Mocap).

1 Introduction

La reconnaissance des actions et des comportements d'individus est l'un des domaines les plus actifs en vision par ordinateur. Cependant, la motricité volontaire est organisée à partir des intentions motrices et sociales [4]. Des recherches récentes en psychologie cognitive ont montré que lorsque nous réalisons une action avec une intention sociale au lieu d'une intention personnelle, nous amplifions les paramètres spatiaux et temporels de l'action motrice [7]. De plus, un observateur est capable de percevoir ces changements cinématiques et anticiper l'intention sociale dans les actions motrices effectuées par d'autres, afin d'agir de manière complémentaire [5]. Toutefois, les liens entre intention sociale et motricité sont encore mal connus, notamment dans le domaine de la vision par ordinateur. C'est dans ce contexte que Zunino et al. [9] proposent une approche de prédiction d'intention à partir de mouvements représentés par des matrices de covariance. L'utilisation des matrices de covariance est étendue au cas des séquences temporelles d'articulations 3D, en proposant une approche de la reconnaissance de l'action humaine à partir de séquences squelettes 3D extraites de données de profondeur. Contrairement au travail de Zunino et al [9], nous représentons les mouvements de la main par un ensemble de trajectoires [1].

2 Méthodologie

2.1 Représentation du mouvement dans $\mathbb{R}^3$

Nous représentons par P_t l'état du point à un instant t, $P_t = [x_1(t), y_1(t), z_1(t) \dots x_k(t), y_k(t), z_k(t)]^T$. Un mouvement est une séquence de poses et peut être vue comme le résultat d'une trajectoire continue dans l'espace des mouvements. La trajectoire est définie par le mouvement au cours du temps des points caractéristiques encodant les coordonnées 3D des articulations de la main. Soit une trajectoire dans l'espace des actions représentée par une fonction paramétrée $\alpha(t) : I = [0, 1] \rightarrow \mathbb{R}^3$.

2.2 Représentation du mouvement dans l'hypersphère $\mathcal{C}$

Dans le but de modéliser et d'étudier nos courbes, nous adoptons la fonction appelée square-root velocity function (SRVF) [8]. Elle a été exploitée avec succès pour la reconnaissance d'actions humaines [2], la reconnaissance de visages en 3D [3] et plus récemment dans la génération d'expressions faciales [6]. Plus précisément, pour une courbes donnée $\alpha(t) : I \rightarrow \mathbb{R}^3$, la fonction SRVF $q(t) : I \rightarrow \mathbb{R}^3$ est définie par :

$$q(t) = \frac{\dot{\alpha}(t)}{\sqrt{\|\dot{\alpha}(t)\|}} , \tag{1}$$

où, $\| \cdot \|$ est la norme L_2 dans $\mathbb{R}^3$.

Afin d'éliminer la variabilité d'échelle des courbes, nous les mettons à l'échelle pour qu'elles soient de longueur 1. Par conséquent, les SRVF correspondant à ces courbes sont des éléments d'une hypersphère unitaire dans le l'espace de Hilbert $\mathbb{L}^2(I, \mathbb{R}^3)$ comme expliqué dans [8]. Nous appellerons cette hypersphère $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3 | \ \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$. Chaque élément de $\mathcal{C}$ représente une courbe dans $\mathbb{R}^3$ associée à un mouvement humain. $\mathcal{C}$ est une hypersphère, la longueur ou la distance géodésique entre deux éléments q_1 et q_2 est définie comme suit :

$$d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle) . \tag{2}$$

Nous définissons les opérations $\log_\mu(.)$ et $\exp_\mu(.)$, les logarithme et exponentielle sur la sphère, utilisées pour projecter les données SRVF dans les deux sens vers l'espace tangent $T_\mu(C)$ à un point de référence μ. Elles sont données par :

$$\begin{aligned} \log_\mu(q) &= \frac{d_{\mathcal{C}}(q, \mu)}{sin(d_{\mathcal{C}}(q, \mu))}(q - cos(d_{\mathcal{C}}(q, \mu))\mu) , \\ \exp_\mu(s) &= cos(\|s\|)\mu + sin(\|s\|)\frac{s}{\|s\|}, \end{aligned} \tag{3}$$

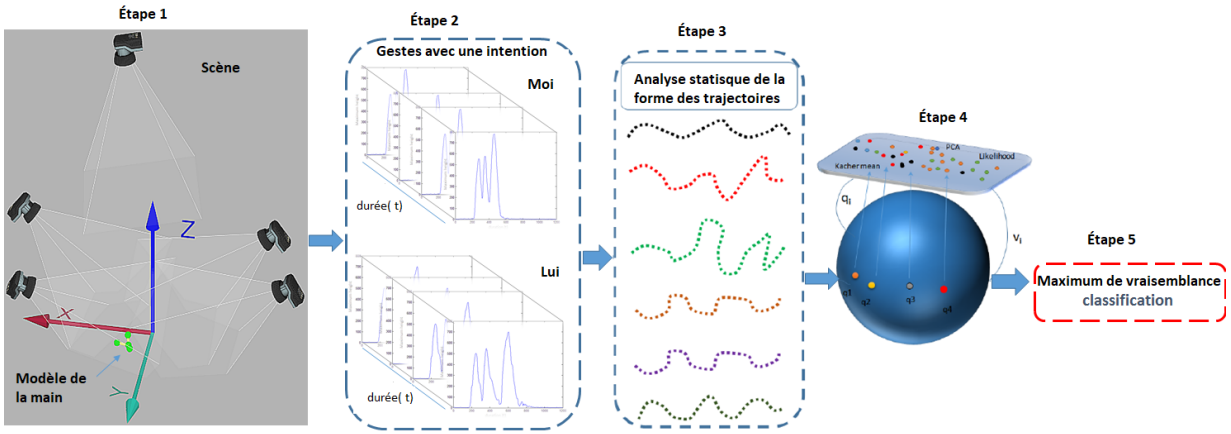

FIGURE 1 – Vue générale de la méthode proposée.

2.3 Analyse statistique des trajectoires

L'objectif principal de notre étude est de classer l'intention de l'utilisateur parmi deux classes c_k qu'on dénote {personnelle, sociale}. Pour cela, nous proposons d'apprendre des distributions représentatives des trajectoires pour chaque classe. La variété $\mathcal{C}$ n'est pas un espace vectoriel. Les structures euclidiennes telles que la norme et le produit scalaire, les algorithmes d'apprentissage automatique, y compris l'analyse en composantes principales (ACP) et l'algorithme de classification par maximum de vraisemblance, ne peuvent pas être appliqués dans leur forme originale sur le variété $\mathcal{C}$. Une approche commune utilisée pour faire face à cette non-linéarité est d'exploiter les propriétés Euclidiennes de l'espace tangent en un point particulier de la variété, par exemple, la moyenne de Karcher des données, μ. Un tel espace tangent est un espace vectoriel linéaire qui est plus pratique pour calculer des statistiques. Par conséquent, afin d'apprendre la distribution des vecteurs dans l'espace tangent, nous pouvons d'abord effectuer une ACP pour apprendre un sous-espace principal appelé $\mathcal{B}$. La matrice de covariance dans cette base est définie par $\Sigma = \sum_{i=1}^{N} v_i v_i^T$, où v_i sont les vecteurs tangents de la projection dans la base $\mathcal{B}$.

Enfin, la distribution normale multivariée de la trajectoire c_k, $p(v|c_k; |\Sigma|)$ est apprise en utilisant la matrice de covariance Σ calculée à partir de l'ensemble des v_i où $|\Sigma|$ est le déterminant de la matrice de covariance Σ, voir l'équation 4.

$$p\left(v \mid c_k; \Sigma\right) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2} v^T \Sigma^{-1} v} \qquad (4)$$

3 Résultats Expérimentaux

3.1 Protocole expérimental

L' expérimentation comprend 2 parties : a) l'acquisition de données et une étape d'apprentissage ; b) la classification et l'analyse du mouvement de la main des sujets pour interagir avec un agent virtuel. Pour une évaluation efficace de notre méthode, nous avons collecté un ensemble de données sur 15 sujets dont l'âge varie de 20 à 50 ans. Toute la scène est couverte par 5 caméras (Infra-rouge) de capture de mouvement (Mocap) qui fonctionnent à une fréquence de 200 Hz chacune. Les sujets ont été invités dans la scène, et à s'asseoir devant une table. Le sujet porte un gant noir qui contient 4 marqueurs passifs. Ces marqueurs sont placés dans une position spécifique sur le gant : l'index (pointe), le pouce (pointe), la main et le poignet. Une tasse est située à une position particulière sur la table. Lorsque le mot («Moi»,«Lui») est diffusé, le sujet doit déplacer la tasse du point A au point B, ces 2 positions étant visibles sur la table, voir figure 2. La distance totale entre le point initial (PI) et le point B est de 48 cm, et la distance entre le point A et le point B est de 24 cm. Les distances ont été choisies en fonction d'une vraie table barman.

FIGURE 2 – Dans cette figure, il est possible d'observer les différentes positions (A, B, position initiale) et mouvements de la main du participant sur la table. PI : position initiale (hand laying), A : déplacement de la main ; B : saisie de la tasse.

1) Avant la réalisation d'un geste, on vérifie bien que la main du participant se trouve dans la position initiale. Lorsque le mot « Moi » est diffusé, le sujet doit déplacer la tasse sur la table du point A vers le point B avec une intention personnelle (déplacer la tasse sans vouloir inclure une autre personne dans l'action).

2) Lorsque le mot « Lui » est diffusé, le sujet doit déplacer la tasse du point A vers le point B avec une intention sociale (avec la volonté d'impliquer une autre personne dans l'action). Les sujets réalisent 50 gestes "Lui" et 50 gestes "Moi" et peuvent commencer aléatoirement par la condition "Lui" ou par la condition "Moi". Pour le traitement des données nous avons éliminé les gestes de retour à la position initiale de la main, ainsi que tous les gestes qui sont incorrects c'est-à-dire : tremblements, hésitations de la main, signaux inexploitables ou données manquantes. Un filtre médian 3D a été utilisé pour éliminer le bruit. Il a été vérifié que la courbe de vitesse de chaque geste contient 4 minimum et 3 maximum ce qui permet de garantir que le mouvement a été réalisé de façon correcte sur le

plan cinématique. La position initiale (PI) de la main du participant détermine le premier minimum, le deuxième minimum correspond au moment où le participant prend la tasse sur la table (position A). Le troisième minimum correspond au moment où le participant dépose la tasse sur la table (position B). Le quatrième minimum correspond au retour à la position initiale (RPI), voir figure 3. Les 3 maximums correspondent aux pics de vitesses des 3 mouvements réalisés (saisie de la tasse, déplacement de la tasse, retour en position initiale). Pour rappel, la fréquence d'échantillonnage du Mocap est de 200 Hz, et les participants sont invités à réaliser le geste dans un intervalle de 0 à 4s. L'axe des abscisses représente la durée (t) du mouvement, et l'axe des ordonnées la vitesse, voir figure 3. Pour analyser l'effet de l'intention sociale sur la cinématique du mouvement, nous représentons le mouvement de la main par une série temporelle des points en 3D. Ensuite nous sélectionnons les moments pertinents du geste (saisie de la tasse, déplacement de la tasse). La question qui se pose est comment classer le mouvement de la main en deux classes (intention personnelle et sociale). Notre approche consiste à représenter cette série temporelle de points en 3D par une trajectoire ou une courbe en 3D. Dans la figure 4, les courbes en rouge correspondent à des trajectoires des gestes «Lui», et les courbes en bleu les trajectoires des gestes «Moi». Nous effectuons par la suite une analyse statistique de la forme de ces trajectoires.

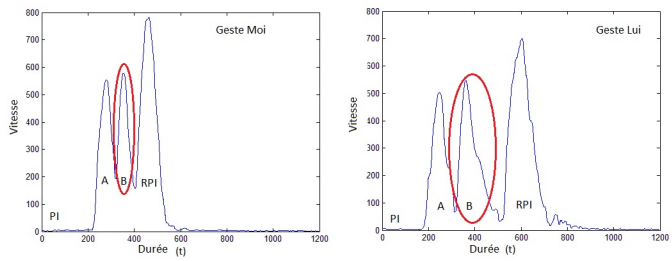

FIGURE 3 – PI : Position Initiale, A : déplacement de la main pour saisie la tasse ; B : saisie de la tasse ; RPI : retour position initiale de la main.

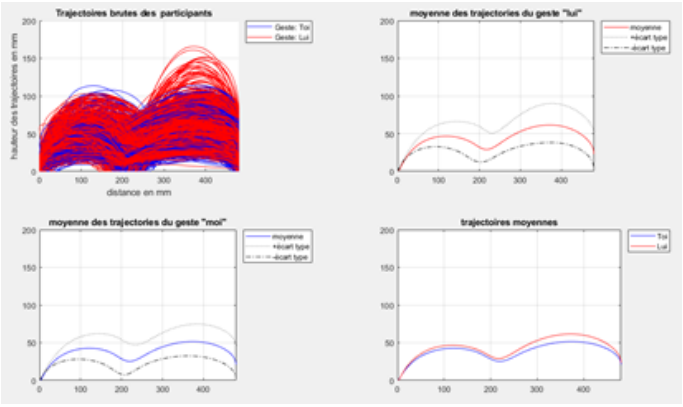

FIGURE 4 – Dans cette figure, il est possible d'observer l'ensemble des gestes (trajectoires) réalisés par les participants avec une intention sociale (les courbes rouges) ou personnelle (les courbes en bleu).

3.2 Résultats

Dans la phase de test, 15 nouveaux participants totalement naïfs ont été invités dans la scène expérimentale qui est identique à l'étape d'apprentissage. Un agent virtuel animé a été projeté sur un téléviseur de 165 cm. L'agent virtuel consiste en un barman dans son propre environnement de travail. Dans la figure 5, on peut observer 3 positions sur la table : 1) La position initiale (PI, Hand laying) ; 2) la position A ; 3) la position B. En effet, le déplacement de la main du participant en partant de la position initiale pour aller vers la position A pour prendre la tasse, et pour la déposer dans la position B, permet de définir la trajectoire parcourue par la main du participant pour réaliser le geste. Ainsi, le participant s'assied devant la table avec la télé. Lorsque le mot (« Lui », « Moi ») est diffusé comme dans l'étape d'apprentissage, le participant déplace la tasse de la position A à la position B avec l'intention sociale ou personnelle. Le participant doit réaliser la bonne intention sociale pour déclencher l'action appropriée du barman virtuel. Ainsi, nous obtenons un score de reconnaissance de 73%. Dans la figure de gauche, le participant réalise une intention personnelle («Moi»), tandis que dans la figure de droite le participant réalise un geste avec une intention sociale («Lui»). L'argent virtuel réagit selon l'intention du geste qui est détectée. L'avantage de la méthode proposée c'est qu'elle est invariante par rapport à la position et la rotation, autrement dit le participant pourrait déposer la tasse n'importe où sur la table sans aucun autre apprentissage supplémentaire. Pour bien mener notre étude, nous avons préféré que les participants fassent tous le même geste avec le même point de départ et même point d'arrivée. Pour rappel, dans cette étude nous avons utilisé le profile de vitesse des deux gestes comme un bon indicateur qui permet d'analyser l'effet de l'intention sur la cinématique, et la trajectoire nous permet d'observer comment le participant module les deux gestes. Dans la figure 5 on observe que les courbes en rouge qui correspondent à une intention sociale possèdent une amplitude plus grande que les courbes en bleu qui représentent une intention personnelle. A partir de ces résultats, il est possible de dire que dans une interaction, parfois on amplifie nos gestes lorsqu'on souhaite inclure une autre personne dans notre action.

FIGURE 5 – Dans la figure de gauche, le participant réalise un geste avec une intention personnelle, donc le barman virtuel le regard. Tandis que dans le figure de droite, le participant réalise un geste avec une intention sociale, et il a été servi par le barman virtuel.

4 Conclusion et perspectives

Dans ce papier, nous avons proposé une approche basée sur l'analyse cinématique et de la géométrie Riemannienne pour analyser le mouvement du bras humain, lorsque les individus réalisent des gestes avec une intention sociale ou personnelle. Les résultats obtenus sur l'ensemble des données nous permettent d'avoir un taux de reconnaissance pour les deux gestes de 73%. Les résultats montrent que la méthode proposée est comparable aux scores produits par [7].

Références

[1] Mohamed Daoudi, Yann Coello, Paul Audain Desrosiers, and Laurent Ott. A new computational approach to identify human social intention in action. In *13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 512–516, 2018.

[2] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo. 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. *IEEE TC*, 45(7) :1340–1352, 2014.

[3] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3D face recognition under expressions, occlusions, and pose variations. *PAMI*, 35(9) :2270–2283, 2013.

[4] Maria Francesca Gigliotti, Adriana Sampaio, Angela Bartolo, and Coello Yann. The combined effect of motor and social goals on the kinematics of object-directed motor action. volume 10, page 6369, 2020.

[5] Daniel Lewkowicz, Quesque Francois, Coello Yann, and N. Delevoye-Turrell Yvonne. Individual differences in reading social intentions from motor deviants. volume 6, 2015.

[6] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional Wasserstein generative adversarial nets. *PAMI*, pages 1–1, 2020.

[7] Francois Quesque, Daniel Lewkowicz, Yvonne N. Delevoye-Turrell, and Yann Coello. Effects of social intention on movement kinematics in cooperative actions. volume 7, 2013.

[8] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *PAMI*, 33(7) :1415–1428, 2011.

[9] Andrea Zunino, Jacopo Cavazza, Atesh Koul, Andrea Cavallo, Cristina Becchio, and Vittorio Murino. Intention from motion. *CoRR*, abs/1605.09526, 2016.