Réseaux de neurones profonds géométriques basés sur des transformations rigides et non rigides pour la reconnaissance de l'action humaine

Rasha Friji[1,2] Hassen Drira[3,4] Faten Chaieb[1,5] Hamza Kchok [6]

[1] Labo CRISTAL - ENSI, Université Manouba Campus, Manouba, Tunisie
[2] Talan Tunisia, 10 Rue de l'énergie solaire Impasse N°1 Charguia 1, Tunis 2035, Tunisie
[3] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France
[4] Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189 CRIStAL, F-59000 Lille, France
[5] Efrei Paris Research Lab
[6] Université Paris-Est Créteil (UPEC), France

Résumé

Bien que les architectures d'apprentissage profond soient efficaces dans la plupart des tâches de vision par ordinateur, elles ont été conçues pour des données avec une structure euclidienne. Cette condition n'est généralement pas remplie puisque les données prétraitées peuvent se situer sur un espace non linéaire. Dans cet article, nous proposons une approche d'apprentissage profond sensible à la géométrie utilisant l'optimisation des transformations rigides et non rigides pour la reconnaissance d'action basée sur le squelette. Les séquences de squelette sont d'abord modélisées comme des trajectoires sur l'espace de formes de Kendall, puis mappées sur l'espace tangent linéaire. Les données structurées résultantes sont ensuite transmises à une architecture d'apprentissage profond, qui comprend une couche qui optimise les transformations rigides et non rigides des squelettes 3D, suivie d'un réseau CNN-LSTM. L'évaluation sur NTU-RGB+D et NTU-RGB+D 120, a prouvé que l'approche proposée surpasse les méthodes existantes d'apprentissage géométrique profond et dépasse les approches récemment publiées pour la majorité des configurations.

Abstract

Although deep Learning architectures have been successful in various computer vision tasks, they were designed for data with an underlying Euclidean structure, which is not usually fulfilled since pre-processed data may lie on a non-linear space. In this paper, we propose a geometry-aware deep learning architecture using rigid and non rigid transformation optimization for skeleton-based action recognition. Skeleton sequences are first modeled as trajectories on Kendall's shape space and then mapped to the linear tangent space. The mapped data are then fed to a deep learning architecture, which includes a layer that optimizes over rigid and non rigid transformations of the 3D skeletons, followed by a CNN-LSTM network. The assessment on NTU-RGB+D and NTU-RGB+D 120, has proven that the proposed approach outperforms existing geometric deep learning methods and exceeds recent approaches with respect to the majority of configurations.

1. Introduction

L'analyse du comportement humain via divers types de données est devenue un problème de recherche actif en vision par ordinateur. À notre connaissance, les principales approches précédentes d'apprentissage profond géométrique ont été conçues pour des espaces de caractéristiques [7, 6] ou pour une variété en 3D du corps humain [1, 10]. Très peu de travaux ont étudié ce problème sur les espaces de formes [4, 5]. Dans ce travail, nous proposons une nouvelle approche d'apprentissage profond géométrique sur l'espace de formes de Kendall, appelée KShapeNet, pour la reconnaissance d'action à partir de données squelettes. Les séquences de squelettes sont d'abord modélisées comme des tra-

jectoires sur l'espace de formes de Kendall en filtrant les transformations rigides et le facteur d'échelle. Les séquences sont, par la suite, projetées dans un espace tangent linéaire et les données structurées résultantes sont introduites dans une architecture d'apprentissage profond. Cette dernière comprend une nouvelle couche qui apprend la meilleure transformation rigide ou non rigide à appliquer aux squelettes 3D pour reconnaître les actions.

Contributions : Les principales contributions sont : 1-Nous présentons une nouvelle architecture profonde sur l'espace de formes de Kendall qui apprend les transformations des squelettes pour la reconnaissance d'action.

2-Le réseau profond proposé comprend une nouvelle couche de transformation qui optimise les transformations rigides et non rigides des squelettes afin d'augmenter la précision de la reconnaissance des actions.

2. Modélisation des trajectoires dans l'espace de formes

Nous utilisons la représentation du squelette humain basée sur les formes des points de repère (Landmarks), et les outils géométriques de l'analyse de formes de Kendall [3] pour modéliser les formes du squelette et leur évolution temporelle.

Chaque squelette X dans une séquence d'actions est représenté comme un ensemble de n points de repère dans $\mathbb{R}^3$, c'est-à-dire que $X \in \mathbb{R}^{n\times 3}$. Dans le cadre de travail, nous modélisons les séquences de formes squelettiques et nous utilisons la représentation de formes de Kendall pour obtenir les invariances requises en matière de translation, d'échelle et de rotation. Les variabilités de translation et d'échelle peuvent être supprimées de l'espace de représentation par normalisation, comme suit. Soit H la sous-matrice $(n-1)\times n$ d'une matrice de Helmert, comme détaillé dans [3], où la première ligne est supprimée. Afin de centrer un squelette X, nous le prémultiplions par H, $HX \in \mathbb{R}^{(n-1)\times 3}$; alors, HX contient les coordonnées euclidiennes centrées de X. Soit $C_0 = \{HX \in \mathbb{R}^{(n-1)\times 3}.|X \in \mathbb{R}^{n\times 3}\}$, qui est un espace vectoriel de dimension $3(n-1)$, qui peut être identifié à $\mathbb{R}^{3(n-1)}$. En utilisant le produit interne (norme) euclidien standard sur C_0, nous mettons à l'échelle tous les squelettes centrés pour qu'ils aient une norme unitaire. Par conséquent, nous définissons l'espace de préforme comme étant $C = \{HX \in C_0 | \|HX\|^2 = (HX)^T(HX) = 1\}$; en raison de la contrainte de la norme unitaire, C est une sphère unitaire à $(3n-4)$ dimensions dans $\mathbb{R}^{3(n-1)}$. Dans le reste de l'article, nous appellerons un élément de C un $\tilde{X}$, c'est-à-dire un squelette centré et de norme unitaire. L'espace tangent à toute préforme $\tilde{X}$ est donné par $T_{\tilde{X}}(C) = \{V \in \mathbb{R}^{3n-4} | \langle V, X\rangle = V^T X = 0\}$.

Dans les analyses ultérieures, notre représentation des séquences de squelettes passe à l'espace tangent. On définit alors deux outils géométriques riemanniens qui permettent de faire correspondre des points 1) de l'espace pré-forme à un espace tangent, et 2) vice-versa La tâche 1) peut être réalisée par la carte logarithmique (log map), $log_{\tilde{X}} : C \to T_{\tilde{X}}(C)$, définie par (pour $\tilde{X}$, $\tilde{Y} \in C$) :

$$log_{\tilde{X}}(\tilde{Y}) = \frac{\theta}{\sin(\theta)}(\tilde{Y} - \cos(\theta)\tilde{X}), \qquad (1)$$

où $\theta = \cos^{-1}\left(\langle \tilde{X}, \tilde{Y}\rangle\right)$ est la distance en arc de cercle entre $\tilde{X}$ and $\tilde{Y}$ on C. La tâche 2) s'effectue via la carte exponentielle (Exponential map), $exp_{\tilde{X}} : T_{\tilde{X}}(C) \to C$, définie comme (for $\tilde{X} \in C$ and $V \in T_{\tilde{X}}(C)$) :

$$\tilde{Y} = \cos(\|V\|)\tilde{X} + \sin(\|V\|)\frac{V}{\|V\|}, \qquad (2)$$

où $\|V\| = \sqrt{V^T V}$.

La variabilité de rotation est éliminée par paire (ou par rapport à un modèle donné), en alignant de manière optimale deux configurations $\tilde{X}$ et $\tilde{Y}$ via une analyse de Procrustes [3] On peut ainsi utiliser les mêmes outils géométriques riemanniens que sur l'espace des préformes C.

3. Architecture profonde dans l'espace de formes

L'architecture proposée, KShapeNet, est illustrée dans la Fig. 1. Les séquences d'entrée sont d'abord modélisées comme des trajectoires sur C, après chaque squelette $\tilde{X}$ est mis en correspondance avec un espace tangent commun $T_{\tilde{X}_0}(C)$ au niveau d'une forme de référence $\tilde{X}_0$ (pose neutre). Ensuite, une couche de transformation est construite dans cet espace tangent pour augmenter les dissimilarités globales ou locales entre les classes des actions, suivie par un bloc CONV et un réseau LSTM à une seule couche. En sortie, un bloc entièrement connecté donne la classe d'action correspondante. Le nombre de paramètres entrainables de notre modèle est 886504 paramètres.

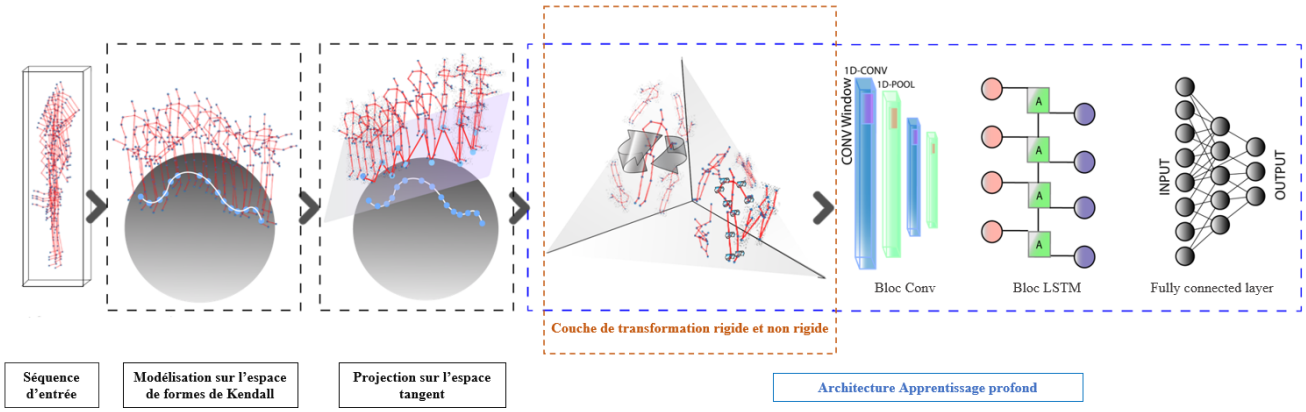

FIGURE 1. Illustration de l'architecture complète de KShapeNet et de ses différents blocs : 1- Modélisation de la séquence d'entrée comme des trajectoires sur l'espace de formes de Kendall. 2- Projection sur l'espace tangent 3-Architecture d'apprentissage profond intégrant la couche de transformation rigide et non-rigide..

3.1. Optimisation sur les transformations rigides

Pour optimiser les transformations rigides, des rotations 3D sont appliquées aux squelettes individuels des séquences (Fig. 2), et sont mises à jour pendant l'étape d'apprentissage.

Soit $\tilde{Y}_i$ le $i^{ième}$ squelette centré de norme unitaire dans une séquence S, et $\hat{Y}_i$ son représentant dans l'espace tangent (remodelé d'un vecteur $3(n-3)$ en une matrice $3 \times (n-1)$ représentée dans les coordonnées ambiantes). La couche de transformation est exécutée sur chaque séquence, résultant en une sortie cachée h, donnée par :

$$h_i = O_i \hat{Y}_i \tag{3}$$

où $O_i \in SO(3)$. Dans la phase de rétropropagation, la descente de gradient adapte directement les noyaux O_i de sorte qu'ils peuvent ne pas se situer dans $SO(3)$. Pour s'assurer que les noyaux mis à jour se trouvent dans $SO(3)$, nous proposons une deuxième variante de cette couche, appelée "basée sur l'angle", où l'optimisation est effectuée sur les angles de rotation. Les matrices de rotation sont ensuite générées dans la passe de feed-forward.

3.2. Optimisation sur les transformations non rigides

L'optimisation des transformations locales est réalisée en trouvant les meilleures rotations des articulations du squelette 3D, par rapport aux axes x, y et z, qui améliorent les performances de la tâche de reconnaissance d'action (3).

Soit $\tilde{Y}_i$ la même représentation du squelette que la section précédente, et $q_i^j \in \mathbb{R}^3$ le $j^{ième}$ joint de $\hat{Y}_i$. La couche de transformation est effectuée sur chaque séquence, résultant en une sortie cachée h, donnée par :

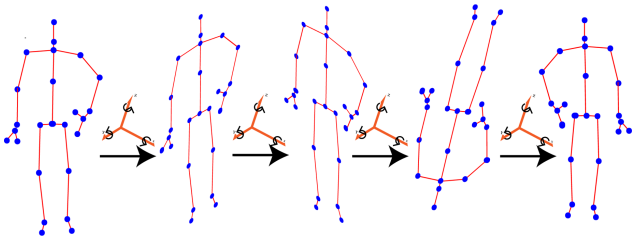

FIGURE 2. Optimisation sur la transformation rigide : Des rotations 3D du squelette entier sont appliquées pendant l'étape d'apprentissage.

$$h_i = \{O_{i,j} q_i^j\}_{j=1}^{n}, \tag{4}$$

où $O_{i,j} \in SO(3)$. Comme dans le cas des transformations rigides, une variante d'optimisation basée sur les angles est proposée pour garantir que chaque $O_{i,j}$ soit une matrice de rotation.

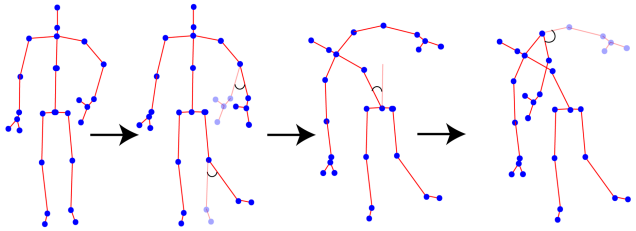

FIGURE 3. Optimisation sur la transformation non rigide : Des rotations 3D sont appliquées aux articulations pendant l'étape d'apprentissage.

4. Résultats expérimentaux

Dans cette section, nous présentons les résultats expérimentaux de KShapenet sur NTU-RGB+D[11] en utilisant les protocoles cross subject (CS) et cross view (CV), et NTU-RGB+D120 [8] en utilisant les protocoles cross subject (CS) et cross setup (CSet).Le premier bloc du tableau 1 présente les résultats de reconnaissance de KShapeNet pour les quatre variantes de la couche de transformation. Chaque ligne fait référence à l'une de ces variantes : 1) optimisation des transformations rigides avec la variante basée sur la matrice de rotation, 2) optimisation des transformations rigides avec la variante basée sur l'angle, 3) optimisation des transformations non rigides avec la variante basée sur la matrice de rotation et 4) optimisation des transformations non rigides avec la variante basée sur l'angle. Le second bloc du tableau 1compare quelques résultats de l'état de l'art aux résultats de KShapeNet.

Pour la configuration finale de KShapeNet, nous avons choisi d'optimiser les transformations non rigides en utilisant la variante basée sur les angles, ce qui permet une modélisation flexible des transformations inter-articulaires ; les résultats de reconnaissance correspondants sont mis en gras dans le tableau 1.

Dataset	NTU		NTU120	
Protocol	CS	CV	CS	Cset
Rigide basée sur la matrice	97.0	97.1	90.2	85.9
Rigide basée sur l'angle	96.9	96.3	89.1	84.9
Non rigide basée sur la matrice	96.8	96.9	90.6	84.3
Non rigide basée sur l'angle	**97.0**	**98.5**	**90.6**	**86.7**
Comparaison avec l'état de l'art				
AGC-LSTM[12]	89.2	95.0	-	-
Intrinsic SCDL [13]	73.8	82.9	-	-
Deep learning on $SO(3)^n$ [6]	61.3	66.9	-	-
MS-G3D Net[9]	-	-	86.9	88.4
SkeleMotion[2]	-	-	67.7	66.9

TABLE 1. Résultats des différentes variantes de la couche de transformation et comparaison avec état de l'art (accuracy en %).

5. Conclusion

Dans cet article, nous avons proposé une architecture géométrique profonde, KShapeNet, pour la reconnaissance d'actions basée sur la modélisation des actions humaines sur l'espace de formes de Kendall. Dans notre cadre, nous avons introduit une nouvelle couche de transformation pour augmenter les dissimilarités globales ou locales entre différents types d'actions. Des expériences, menées sur deux grands ensembles de données de référence, démontrent que, KShapeNet, dépasse les taux de reconnaissance des approches de pointe pour la majorité des configurations.

Références

[1] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning : Going beyond euclidean data. *IEEE Signal Processing Magazine*, 2017. 1

[2] C. Caetano, J. S. de Souza, F. Brémond, J. A. dos Santos, and W. R. Schwartz. Skelemotion : A new representation of skeleton joint sequences based on motion information for 3d action recognition. 4

[3] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. Wiley, 1998. 2

[4] R. Friji, H. Drira, and F. Chaieb. Geometric deep learning on skeleton sequences for 2d/3d action recognition. 1

[5] N. Hosni and B. B. Amor. A geometric convnet on 3d shape manifold for gait recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops*. 1

[6] Z. Huang, C. Wan, T. Probst, and L. V. Gool. Deep learning on lie groups for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 1, 4

[7] Z. Huang, J. Wu, and L. V. Gool. Building deep networks on grassmann manifolds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18) 2018*. 1

[8] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Yu Duan, and A. C. Kot. NTU RGB+D 120 : A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 4

[9] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2020. 4

[10] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst. Shapenet : Convolutional neural networks on non-euclidean manifolds. Technical report. 1

[11] A. Shahroudy, J. Liu, T. Ng, and G. Wang. NTU RGB+D : A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 4

[12] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4

[13] A. B. Tanfous, H. Drira, and B. . Amor. Sparse coding of shape trajectories for facial expression and action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 4