sciences appliquées

MDPI

Article

Un nouveau réseau neuronal symétrique fin-grossier pour l'humain 3D

Reconnaissance d'actions basée sur des séquences de nuages de points

Chang Li [1], Qian Huang [1,*], Yingchi Mao [1], Weiwen Qian [1] et Xing Li [2]

[1] Collège d'informatique et de génie logiciel, Université Hohai, Nanjing 211100, Chine ; lichang@hhu.edu.cn (CL); yingchimao@hhu.edu.cn (YM); qianweiwen@hhu.edu.cn (WQ)
[2] Collège des sciences et technologies de l'information et Collège d'intelligence artificielle, Nanjing Forestry Université, Nanjing 210037, Chine ; lixing@njfu.edu.cn
* Correspondance : huangqian@hhu.edu.cn

Résumé : La reconnaissance de l'action humaine a facilité le développement de dispositifs d'intelligence artificielle axés sur les activités et les services humains. Cette technologie a progressé en introduisant Nuages de points 3D issus de caméras de profondeur ou de radars. Cependant, le comportement humain est complexe et les nuages de points impliqués sont vastes, désordonnés et compliqués, ce qui pose des défis à l'action 3D reconnaissance. Pour résoudre ces problèmes, nous proposons un réseau neuronal symétrique fin-coarse (SFCNet) qui analyse simultanément l'apparence et les détails des actions humaines. Premièrement, les séquences de nuages de points sont transformées et voxélisées en ensembles de voxels 3D structurés. Ces ensembles sont ensuite augmentés avec un descripteur de fréquence d'intervalle pour générer des caractéristiques 6D capturant la dynamique spatio-temporelle information. En évaluant l'occupation de l'espace voxel à l'aide du seuillage, nous pouvons identifier efficacement le pièces essentielles. Après cela, tous les voxels dotés de la fonctionnalité 6D sont dirigés vers le flux grossier global, tandis que les voxels dans les parties clés sont acheminés vers le flux fin local. Ces deux flux extraient caractéristiques d'apparence globale et parties du corps critiques en utilisant PointNet++ symétrique. Ensuite, la fusion des caractéristiques d'attention est utilisée pour capturer de manière adaptative des modèles de mouvement plus discriminants. Les expériences menées sur les ensembles de données de référence publics NTU RGB+D 60 et NTU RGB+D 120 valident L'efficacité et la supériorité de SFCNet pour la reconnaissance d'actions 3D.

Mots-clés : analyse de nuages de points ; Reconnaissance d'actions 3D ; la reconnaissance de formes; l'apprentissage en profondeur

Citation : Li, C. ; Huang, Q. ; Mao, Y. ; Qian, W. ; Li, X. Un roman symétrique Réseau neuronal fin-grossier pour la 3D Reconnaissance de l'action humaine basée sur Séquences de nuages de points. Appl. Sci. 2024, 14, 6335. https://doi.org/10.3390/app14146335

Rédacteur académique : Atsushi Mase

Reçu : 11 juin 2024
Révisé : 8 juillet 2024
Accepté : 18 juillet 2024
Publié : 20 juillet 2024

1. Introduction

La reconnaissance de l'action humaine vise à aider les ordinateurs à comprendre la sémantique du comportement humain à partir de diverses données enregistrées par les dispositifs d'acquisition. En particulier, l'action 3D La reconnaissance est dédiée à l'extraction de schémas d'action à partir de données 3D impliquant des mouvements humains . Il a attiré une attention croissante en raison de ses applications répandues, telles que surveillance de la sécurité publique, évaluation des performances, reconnaissance militaire et intelligence transport [1].

Les méthodes actuelles de reconnaissance d'actions 3D peuvent être classées en méthodes basées sur la profondeur (y compris les cartes de profondeur et les séquences de nuages de points) [2-4] et en méthodes basées sur le squelette [5,6] en fonction du type de données utilisé. Limité par la précision algorithme d'estimation de pose - la tâche inévitable en amont - méthodes basées sur le squelette sont confrontés à des défis de consommation informatique et de robustesse. En revanche, basé sur la profondeur les méthodes sont plus indépendantes des tâches et ont attiré une large attention. Existant Les approches de reconnaissance d'actions 3D basées sur la profondeur se répartissent principalement en deux catégories principales. Le la première consiste à encoder des mouvements 3D en une ou plusieurs images [2,3,7,8] et à utiliser des CNN [9] pour la reconnaissance des actions. Cependant, le plan image 2D ne peut pas caractériser complètement le plan image 3D. dynamique car les actions humaines sont à la fois spatio-temporelles et menées dans le Espace 3D. L'autre consiste à transformer la vidéo de profondeur en une séquence de nuages de points [10], qui enregistre les coordonnées 3D de points dans l'espace à plusieurs instances temporelles. Ainsi, comparé avec les images, les séquences de nuages de points ont l'avantage de conserver l'apparence 3D et

dynamique de la géométrie au fil du temps, permettant une analyse et une compréhension avancées des actions humaines. De plus, les nuages de points peuvent être obtenus à l'aide de divers appareils tels que des scanners laser, des radars, des capteurs de profondeur et des caméras RVB+D, qui peuvent être montés sur des drones, des lampadaires, des véhicules et des avions de surveillance, élargissant ainsi le champ d'application de la reconnaissance d'action. . Cependant, en raison de la structure complexe et du volume massif du nuage de points, les méthodes existantes de reconnaissance d'actions 3D qui en découlent présentent les défis suivan

Premièrement, les séquences de nuages de points comportent toujours des points massifs proportionnels à la dimension temporelle, et le schéma de traitement des données prend du temps. Par conséquent, le développement d'un modèle de séquence de nuages de points efficace et léger est essentiel pour la reconnaissance d'actions 3D. Deuxièmement, les points dans les séquences sont irréguliers, présentant des informations spatiales intra-trame non ordonnées et des détails temporels inter-trame ordonnés, ce qui rend difficile l'analyse des schémas de mouvement sous-jacents. Cependant, les méthodes de traitement des nuages de points existantes effectuent généralement un sous-échantillonnage indifférencié de l'ensemble des nuages de points, ce qui entraîne une perte uniforme d'informations essentielles et subtiles. De plus, les schémas d'analyse de nuages de points existants ignorent les parties critiques du corps contribuant aux actions, ce qui entraîne un manque de nuances dans les caractéristiques d'action extraites, ce qui limite finalement les performances de reconnaissance des actions.

Pour résoudre ces problèmes, nous proposons un cadre d'apprentissage profond appelé Réseau neuronal symétrique fin-coarse (SFCNet) qui combine symétriquement l'analyse des caractéristiques de mouvement d'un point de vue local et global, comme le montre la figure 1. Premièrement, pour réduire les coûts de calcul. , nous réduisons les points par échantillonnage de base et par échantillonnage du point le plus éloigné. Ensuite, les nuages de points échantillonnés sont transformés en voxels 3D pour créer une représentation compacte du nuage de points. Les positions 3D originales sont ensuite associées à un descripteur de fréquence d'intervalle pour décrire la configuration spatiale globale et faciliter l'identification des parties essentielles du corps, nous permettant de diviser les séquences de nuages de points en espace fin local et espace grossier global. Nous traitons les voxels impliqués dans ces deux espaces comme des points et utilisons PointNet++ [11] pour extraire les caractéristiques de bout en bout. Enfin, notre module de fusion de fonctionnalités combine l'apparence globale et les détails locaux pour obtenir des fonctionnalités discriminantes pour la reconnaissance d'actions 3D. Les expériences approfondies sur les ensembles de données à grande échelle NTU RGB+D 60 et NTU RGB+D 120 démontrent l'efficacité et la prépondérance de SFCNet, grâce auquel l'intention humaine peut être jugée et assistée dans les applications d

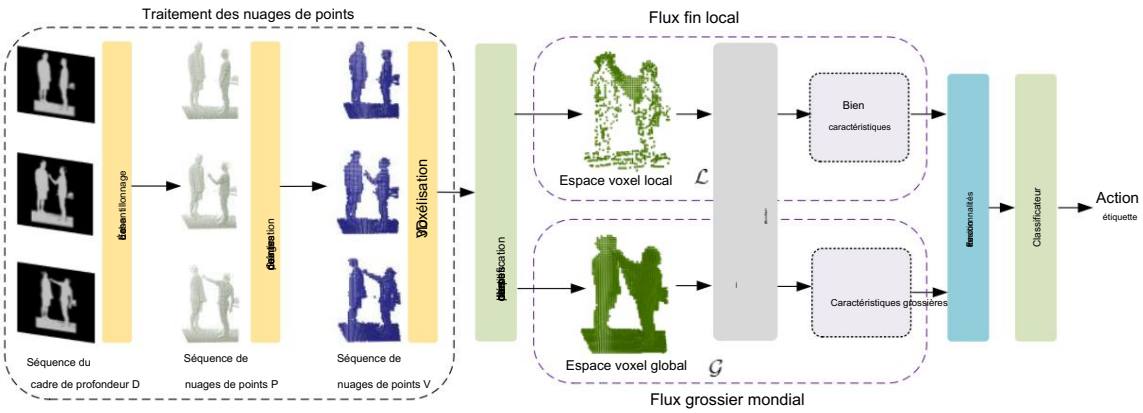


Figure 1. Le pipeline de SFCNet. Il convertit les images de profondeur en un nuage de points et applique des opérations de voxelisation. Une structure symétrique code les voxels 3D, avec des parties cruciales et des informations dynamiques globales traitées séparément. Le descripteur de fréquence d'intervalle ci-joint caractérise initialement les informations de mouvement, puis est traité par PointNet++ [11] pour des fonctionnalités plus approfondies. Enfin, le classificateur reconnaît les actions 3D à l'aide de la fonctionnalité agrégée.

En général, les principales contributions de nos travaux sont les suivantes :

- Nous proposons un descripteur de fréquence d'intervalle pour caractériser les voxels 3D lors de l'exécution d'une action, qui préserve pleinement les détails du mouvement et fournit des indices critiques pour la perception des parties clés du corps. À notre connaissance, notre travail est le premier à traiter ainsi des séquences de nuages de points.

• Nous construisons un framework d'apprentissage profond nommé SFCNet, qui utilise d'abord une structure symétrique pour traiter les séquences de nuages de points. Il code la dynamique des parties cruciales locales du corps via un flux fin, puis complète ces détails complexes à l'apparence globale capturée par un flux grossier. Le SFCNet peut mettre l'accent sur les parties essentielles du corps et capturer des modèles de mouvement plus discriminants, résolvant ainsi le problème de représentation efficace des actions basé sur des nuages de points. • Le SFCNet présenté a démontré sa précision supérieure sur deux ensembles de données accessibles au public, NTU RGB+D 60 et NTU RGB+D 120, ce qui prouve que notre méthode a un potentiel considérable pour reconnaître divers types d'actions telles que les actions quotidiennes, les actions associées et actions d'interaction à deux.

2. Travaux connexes

2.1. Reconnaissance d'action 3D basée sur le squelette

Les méthodes existantes de reconnaissance d'actions 3D peuvent être classées en méthodes basées sur le squelette [12-17] et en méthodes basées sur la profondeur [3,7,12,18,19]. Il existe généralement quatre approches principales pour la reconnaissance d'actions basée sur un squelette. La première consiste à utiliser CNN [12] pour apprendre les modèles spatio-temporels à partir de pseudo-images [13,14]. Caetano et coll. [20] ont introduit l'image des joints de référence de structure arborescente (TSRJI) pour représenter les séquences squelettes. La seconde consiste à considérer les séquences squelettes comme des séries chronologiques [15-17] et à utiliser des squelettes tels que RNN [21] pour l'extraction de caractéristiques. La troisième consiste à visualiser les données squelettiques sous forme de graphiques [6,22] avec les articulations comme sommets et les os comme arêtes et à se tourner vers GCN [16] pour la représentation des actions.

Par exemple, ST-GCN [23] a représenté efficacement les informations dynamiques temporelles des séquences squelettes en utilisant des stratégies de convolution et de partitionnement de graphes spatio-temporels. SkeleMotion [5] a capturé des informations dynamiques temporelles en calculant la taille et l'orientation des articulations du squelette à différentes échelles de temps. La quatrième consiste à encoder le squelette sous forme de jetons via Transformer. Plizzari et coll. [24] ont utilisé l'attention spatio-temporelle dans Transformer et ont capturé une relation dynamique entre les images entre les images. Cependant, comme il existe encore des défis importants dans l'estimation précise de la pose humaine en 3D [25,26], les méthodes de reconnaissance d'actions basées sur le squelette souffrent d'une cascade de performances en raison de cette tâche inévitable en amont.

2.2. Reconnaissance d'action 3D basée sur la profondeur

Pour la reconnaissance d'actions 3D basée sur la profondeur, les premières approches représentent principalement les vidéos de profondeur par des descripteurs manuels [19]. Yang et coll. [7] ont construit des cartes de mouvement en profondeur (DMM) en empilant les différences inter-images des images de profondeur projetées. Ensuite, ils ont calculé l' histogramme des gradients orientés (HOG) pour représenter les actions. De telles méthodes ont un pouvoir d'expression limité et nécessitent donc généralement de l'aide pour capturer des informations spatio-temporelles. Ces dernières années, les méthodes d'apprentissage profond se sont généralisées avec le développement des réseaux de neurones. La plupart des chercheurs ont tenté de compresser des vidéos profondes en images et ont analysé les modèles de mouvement à l'aide de CNN [12]. Kamel et coll. [27] saisissez des images de mouvement en profondeur (DMI) et des descripteurs d'articulations mobiles (MJD) dans les CNN pour la reconnaissance des actions. Pour coder les informations spatio-temporelles des séquences de profondeur, Adrián et al. [28] ont proposé 3D-CNN pour extraire les caractéristiques de mouvement. De plus, ils ont proposé ConvLSTM [29] pour accumuler des modèles de mouvement discriminants à partir d'unités à long terme. Xiao et coll. [3] ont fait pivoter la caméra virtuelle dans l'espace 3D pour projeter de manière dense une vidéo brute en profondeur à partir de différents points de vue d'imagerie virtuelle et ainsi construire des images dynamiques multi-vues. Pour les invariants de vue en perspective, Kumar et al. [30] ont proposé un ActionNet basé sur des CNN et formé avec un ensemble de données multi-vues collectées à l'aide de cinq caméras de profondeur. Ghosh et coll. [31] ont calculé une image d'historique de mouvement détecté par les bords (ED-MHI) du descripteur de profondeur multi-vues en tant qu'entré Wang et coll. [2] ont utilisé des séquences vidéo de profondeur segmentées pour générer trois types d' images de profondeur dynamiques. Cependant, la carte de profondeur 2D a encore du mal à exploiter pleinement les modèles de mouvement 3D en raison de sa structure spatiale compacte [10].

Récemment, la conversion de cartes de profondeur en nuages de points pour le traitement a donné de meilleurs résultats dans les domaines de la reconnaissance et de la segmentation. De nombreuses études ont montré

que les nuages de points présentent des avantages significatifs dans la représentation des informations spatiales 3D en raison de leurs caractéristiques, telles que le désordre et l'invariance de rotation. L'apprentissage profond des nuages de points a non seulement été largement utilisé dans les tâches de classification et de segmentation, mais a également démontré la force musculaire dans la reconstruction de scènes [32] et la détection de cibles [33]. Cependant, les méthodes ci-dessus se concentrent uniquement sur les entités contenues dans des nuages de points statiques. Lors de l'utilisation de nuages de points pour la reconnaissance d'actions 3D, il est nécessaire d'extraire des caractéristiques dynamiques en fonction des intervalles de temps et des caractéristiques d'apparence de l'ensemble du processus d'action. La clé d'un traitement efficace des séquences de nuages de points réside dans la sélection d'une méthode d'analyse de nuages de points appropriée. Thomas et coll. [34] ont développé une méthode inspirée de la convolution basée sur des images et ont utilisé un ensemble de points de noyau pour distribuer le poids de chaque noyau. En tant qu'outil efficace pour analyser et traiter des ensembles de points, PointNet++ [11] est largement appliqué pour la reconnaissance d'actions 3D basée sur des séquences de nuages de points. La première méthode est la 3DV [10], qui exécute une voxélisation 3D vers les séquences de nuages de points et décrit l'apparence 3D par occupation spatiale, et le regroupement de rangs temporels est utilisé pour l'extraction 3DV. Cette méthode se concentre principalement sur le mouvement général d'une action et sur les changements d'apparence. Cependant, il ignore les détails de l'action, comme un mouvement subtil de la main, ce qui limite sa capacité à représenter le comportement avec précision. Par conséquent, nous visons à capturer les parties cruciales des actions et leurs informations délicates pour les reconnaître comme des actions humaines plus robustes.

3. Méthodologie 3.1.

Pipeline Le

pipeline du SFCNet proposé est illustré à la figure 1. Tout d'abord, chaque image de profondeur est transformée en un nuage de points pour mieux préserver les caractéristiques dynamiques et d'apparence dans l' espace 3D. Pour faciliter l'analyse des usages spatiaux et délimiter l'espace local, nous effectuons des opérations de voxélisation sur les nuages de points. Ensuite, nous construisons un cadre symétrique pour coder les voxels 3D, dans lequel les éléments clés et les informations dynamiques globales sont traités séparément dans le flux fin local et le flux grossier global. Ensuite, nous attachons le descripteur de fréquence d'intervalle pour compléter les informations sur le mouvement. Nous utilisons PointNet++ [11] pour capturer des modèles de mouvement et envoyer la fonctionnalité agrégée au classificateur pour la reconnaissance d'actions 3D.

3.2. Génération de voxels tridimensionnels

La vidéo de profondeur a l'avantage de résister aux interférences externes, telles que l'arrière-plan et la lumière, par rapport au modal RVB, car elle contient les informations de profondeur du sujet de l'action. Essentiellement, la vidéo de profondeur est une sorte de série de données chronologiques composée de cartes de profondeur classées par ordre chronologique. Mathématiquement, une vidéo de profondeur avec t images peut être définie comme D = {d1, d2, . . . , dt}, où di est une carte de profondeur de l'image t dans laquelle chaque pixel représente une coordonnée 3D (x, y, z) et z est la distance de la caméra de profondeur. Puisqu'il est impossible de classer l'importance de l'action dans la dimension temporelle selon un seul critère, l'échantillonnage uniforme peut nous aider à mieux comprendre le processus de mouvement global par rapport à l'échantillonnage aléatoire [10]. Par conséquent, nous échantillonnons d'abord la vidéo en profondeur de manière uniforme pour alléger la charge de calcul tout en préservant l'intégrité de l'action. La séquence de profondeur après échantillonnage est notée Dˆ = {d1, d2, . . . , dT}, où T est le nombre de trames et le

la valeur par défaut est 64.

Certaines méthodes actuelles de reconnaissance d'actions [35,36] choisissent de mapper les images de profondeur sur des espaces 2D pour un traitement direct. Bien que ces approches puissent parfois atteindre de bonnes performances , elles ne peuvent pas résoudre le problème de la représentation inadéquate des informations 3D. Par conséquent, pour mieux représenter le mouvement humain dans l'espace 3D, nous transformons chaque image di en un nuage de points P = {p1, p2, . . . , pn}, où n est le nombre de points, générant ainsi une séquence de nuages de points S = {P1, P2, . . . , PT}. Lors de la génération de nuages de points, les paramètres intrinsèques de la caméra sont requis car ils définissent le modèle d'imagerie de la caméra, notamment

distance focale et coordonnées du point principal (cx, cy). Pour chaque pixel (x, y, z) de l' image de profondeur, son nuage de points correspondant p(x) peut être obtenu par la formule suivante :

$$p(x', y', z') = \left(\frac{(x - cx) \times z}{fx}, \frac{(y - cy) \times z}{fy}, \frac{z}{fz}\right) \quad (1)$$

où fx et fy représentent la distance focale de la caméra de profondeur dans la direction horizontale et dans la direction verticale, qui peut être obtenue à partir des paramètres de l'appareil. fz est défini sur 1 par défaut.

Contrairement aux images traditionnelles (données structurées classiques), les points du nuage de points ne sont pas ordonnés, ce qui rend leur traitement difficile. De nombreux algorithmes existants sont conçus pour des données de grille régulières. Cependant, le nuage de points non ordonné est un groupe de points distribués de manière aléatoire dans l'espace 3D, leur structure est donc complexe à traiter et à analyser directement. Pour résoudre ce problème, nous transformons le nuage de points en une grille 3D régulière (espace voxel) par voxélisation pour régulariser la représentation du nuage de points. Tout d'abord, nous définissons la taille de la grille de voxels Vgrid = (Vx, Vy, Vz) en coordonnées tridimensionnelles, qui détermine la résolution du processus de voxélisation. Chaque cellule de cette grille est un voxel potentiel et la taille de chaque cellule est notée Vvoxel (dx, dy, dz). Étant donné un point p(x, y, z) dans le nuage de points, il est mappé sur la grille en trouvant l'indice de voxel correspondant Vindex(x, y, z) selon l'équation suivante :

$$Vindex(x, y, z) = \left(\frac{x - xmin}{dx}, \frac{y - ymin}{dy}, \frac{z - zmin}{dz}\right) \quad (2)$$

où xmin, ymin et zmin sont les coordonnées minimales de tous les nuages de points. dx, dy et dz sont calculés comme la taille totale divisée par le nombre de cellules dans chaque dimension (Vx, Vy, Vz). La fonction plancher . arrondit au point le plus proche. On définit qu'un voxel est occupé s'il contient un nuage de points. Ensuite, les informations d'apparence 3D peuvent être décrites en observant si les voxels ont été occupés ou non, sans tenir compte du point exclu, comme le montre l'équation (3) :

$$V^t_{voxel}(x, y, z) = \begin{cases} 1, & \text{si } V^t_{voxel}(x, y, z) \text{ est occupé} \\ 0, & \text{sinon} \end{cases} \quad (3)$$

où $V^t_{voxel}(x, y, z)$ indique un certain voxel à la ième image. (x, y, z) est l' indice de position 3D régulier , c'est-à-dire Vindex dans l'équation (2). Cette stratégie présente deux bénéfices principaux. Premièrement, les ensembles de voxels 3D binaires obtenus sont réguliers, comme le montre la figure 2. Ainsi, la complexité du traitement des nuages de points est réduite. De plus, la voxélisation peut compresser efficacement les nuages de points car les voxels voisins peuvent avoir des caractéristiques similaires. Cette compression réduit non seulement le nombre de points, mais contribue également à réduire les frais de stockage et de calcul.

3.3. Identification et représentation des éléments clés Le

problème essentiel dans les tâches de reconnaissance d'actions 3D est de capturer et de représenter efficacement les caractéristiques dynamiques au sein de séquences de nuages de points. Pour l'instant, les méthodes d'estimation basées sur le flux de scènes [37,38] peuvent aider à comprendre le mouvement 3D, mais cela prend beaucoup de temps. Certaines études utilisent le regroupement de rangs temporels [3,39] pour préserver les processus de mouvement dans l'espace 3D en divisant les segments temporels. Ces méthodes peuvent capturer davantage d'informations temporelles, mais ne divisent souvent qu'un petit nombre d'intervalles, ce qui donne lieu à des caractéristiques dynamiques à granularité grossière. Nous proposons un module crucial d'identification et d'encodage des pièces pour mieux se concentrer sur les dynamiques critiques lors du mouvement. Il peut extraire les principales parties de l'espace voxel 3D global en fonction des temps d'occupation de l'espace et coder les détails de l'actio Plus précisément, nous analysons d'abord l'occupation de l'espace en construisant un espace 3D U avec les limites exactes sous forme de séquences de nuages de points pour chaque emplacement spatial, et les valeurs initiales

de U sont fixés à 0. Nous traitons les m groupes uniformes de séquence Dˆ dans l'ordre. Ensuite, la 3D l'utilisation de l'espace peut être calculée selon l'équation (4) :

$$Uvoxel(x, y, z) = \begin{cases} Uvoxel(x, y, z) + 1, vi & = 0 \\ Uvoxel(x, y, z), sinon & \end{cases} \quad (4)$$

L'occupation totale de l'espace u pour chaque position peut être obtenue après avoir compté tous les m ensembles de points. De plus, puisque les ensembles de points sont naturellement ordonnés dans le temps, nous pouvons facilement enregistrez la première et la dernière fois prises, f et l, respectivement, pour chaque emplacement spatial. Alors, nous définissons le seuil θ pour partitionner l'espace local important. Les emplacements occupés inférieur à θ sont traités comme bruit incident, et ceux qui enregistrent plus et moins de m constituent les parties critiques du mouvement S comme l'équation (5) :

$$Svoxel(x, y, z) = \begin{cases} Uvoxel(x, y, z) < \theta \\ 0, 1, \theta \le Uvoxel(x, y, z) \le m \end{cases} \quad (5)$$

Si la valeur de Svoxel(x, y, z) est égale à 0, cela signifie que le voxel appartient au global espace G ; sinon, il appartient à l'espace local L. Par rapport au nuage de points les méthodes de traitement [10,40], qui adoptent généralement des opérations de sous-échantillonnage uniformes, la méthode proposée est plus efficace, notamment pour les actions impliquant seulement un petit nombre des parties des membres, car diviser l'espace local peut non seulement surmonter l'arrière-plan effets dans une certaine mesure, mais améliorent également efficacement la teneur en or du point échantillonné données. Comme le montre la figure 2, l'espace local L préserve entièrement les informations détaillées du parties du corps principal, qui fournissent des indices critiques pour la reconnaissance d'action 3D tout en étant substantiellement réduisant la redondance.

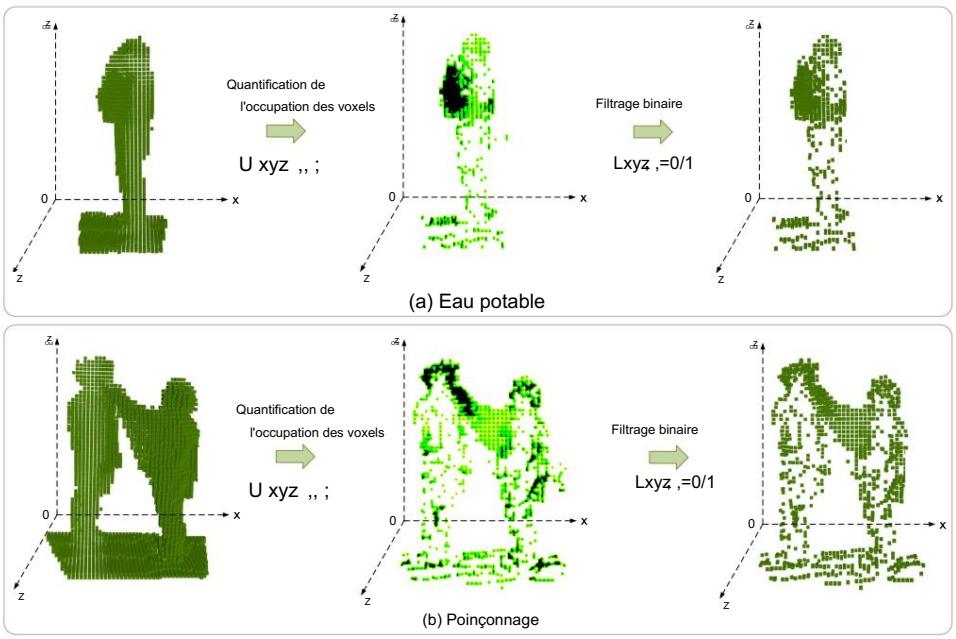


Figure 2. Le processus de division de l'espace local. Nous quantifions la contribution d'un voxel à une action par son nombre d'occupation d'espace. En définissant un seuil, les parties critiques peuvent être divisées sous forme compressée. espace local, qui supprime les informations redondantes et réduit la charge de calcul.

3.4. Extraction de fonctionnalités symétriques

Pour les voxels 3D traités, la manière la plus intuitive consiste à utiliser 3DCNN [41,42], mais elle est limitée par la taille du voxel et prend du temps. Nous choisissons d'utiliser PointNet++ [11] comme extracteur de fonctionnalités dans ce travail, comme le montre la figure 3. Il est conçu explicitement pour les fonctionnalité d'apprentissage sur des ensembles de points non ordonnés dans l'espace métrique, ce qui lui permet de capturer des motifs à grain fin dans des nuages de points. Pour y parvenir, PointNet++ partitionne le point

nuage en régions locales qui se chevauchent en fonction d'une métrique de distance dans l'espace sous-jacent. Pour obtenir des repères visuels 3D détaillés, PointNet++ utilise de manière récursive PointNet [43] pour extraire les caractéristiques locales, qui sont ensuite fusionnées pour une analyse d'apparence globale. PointNet++ est une excellente alternative aux 3DCNN car il permet de capturer des modèles 3D locaux essentiels à la reconnaissance des actions. De plus, son application est relativement simple et ne nécessite que la transformation des voxels 3D en ensembles de points.

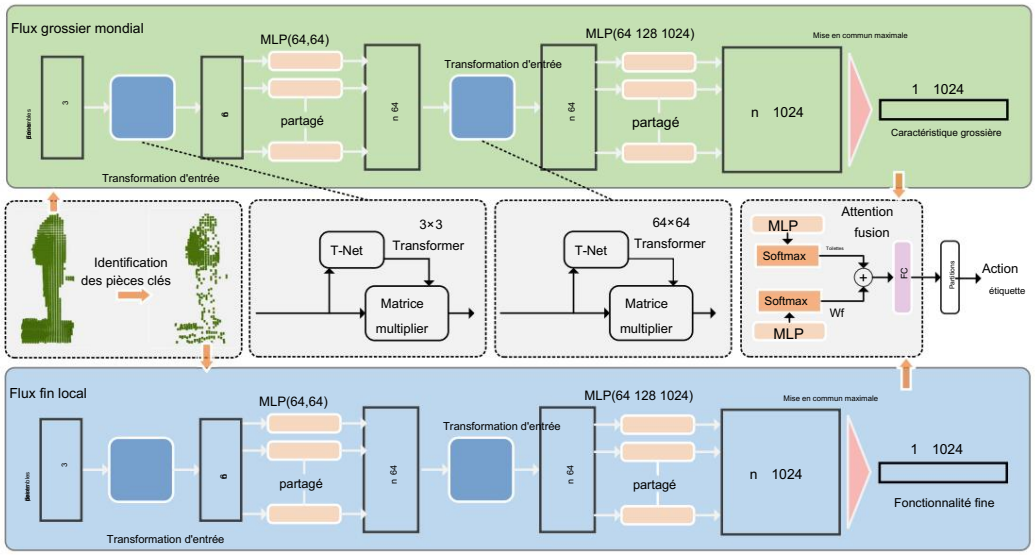


Figure 3. La structure du réseau de SFCNet. Il s'agit d'une structure de réseau symétrique comprenant le flux grossier global et un flux fin local. Le flux grossier global prend en entrée les informations d'apparence globale, tandis que le flux fin local n'adopte que les informations de partie clé compressées. Les caractéristiques de chaque flux sont extraites par PointNet++ et fusionnées par les poids apprenables en tant que caractéristiques d'action finale, qui sont envoyées au classificateur pour la reconnaissance d'action 3D.

Pour s'adapter à PointNet++, chaque point p(x, y, z) est ensuite résumé sous la forme du voxel Vvoxel(x, y, z) avec la caractéristique descriptive de (x, y, z, I), où I est la fréquence d'intervalle descripteur comprenant trois variables (o, f , I) qui désignent respectivement les horodatages de début, les horodatages de fin et la fréquence d'occupation globale des voxels. Nous attachons I aux positions 3D originales des voxels pour obtenir les points 6D p

I
(x, y, z, o, f , I), comme le montre la figure 4. Mathématiquement, pour un voxel Vvoxel, nous pouvons déterminer l'heure de début et de fin (o et f) de son occupation spatiale voxel(x, y, z) = 1 en utilisant l'équation (3), qui indique l'indice de temps lorsque les conditions V t

t
et V

voxel(x, y, z) = 0 sont d'abord satisfaits. De plus, l'occupation totale de l'espace u peut être calculée par l'équation (4). Par conséquent, la fréquence d'occupation I peut être calculée comme I = u/(f − o). Le descripteur d'intervalle-fréquence couvre l'intervalle de temps et la fréquence d'occupation globale des emplacements spatiaux peut non seulement nous aider à distinguer les actions inverses qui couvrent le même espace, mais également aider à mettre en évidence les différences entre les actions en conservant des informations détaillées dans le processus d'action. Enfin, les ensembles de points obtenus sont utilisés comme entrée de PointNet++ pour extraire les caractéristiques d'action.

De plus, nous concevons un réseau à deux flux pour traiter le nuage de points dans l' espace global et local, respectivement (comme le montre la figure 3). L'espace global contient tous les points voxélisés et le flux grossier global capture les modèles de mouvement globaux. Considérant que les parties vitales du corps peuvent fournir des informations dynamiques plus ciblées et discriminantes pour la reconnaissance des actions, nous divisons les points voxélisés des parties essentielles du corps en espaces locaux et saisissons le flux fin pour en extraire les caractéristiques. Après cela, le module de fusion de fonctionnalités est configuré pour une représentation fine-grossière des actions. Compte tenu des caractéristiques des différentes actions, leur dépendance vis-à-vis des caractéristiques globales et locales est variable. Pour les actions qui impliquent uniquement des mouvements des membres, comme agiter et donner des coups de pied, le modèle doit mettre l'accent sur les caractéristiques locales du flux à granularité fine. En revanche, pour les mouvements importants du corps entier, tels que les chutes et les sauts, le modèle doit se concentrer sur les caractéristiques du flux global. Atteir

Pour cette perception spécifique à l'action, nous utilisons le module de fusion de caractéristiques avec un mécanisme d'attention d'abord, nous projetons respectivement les n×1024 et Xc R n×1024 extrait du finisme. Tout caractéristiques Xf R et les flux grossiers dans l'espace de caractéristiques inférieur pour réduire la charge de calcul et obtenir X et X c . Ensuite, les caractéristiques intermédiaires sont extraites par perceptron f multicouche (MLP). Après cela, les poids apprenables Wf et Wc du flux grossier global et du flux fin local sont obtenus respectivement via la fonction d'activation SoftMax .

Enfin, les caractéristiques globales et locales sont fusionnées comme le montre l'équation (6) pour obtenir la caractéristique de mouvement Xˆ :

$$\begin{aligned} Wf &= \text{SoftMax MLP } Xf \\ Wc &= \text{SoftMax MLP } X_c \\ \hat{X} &= \varphi \sum_{je=1}^{K} Wi_{Xff}^{(je)} + Wi\text{-}CX_c^{(je)} \end{aligned} \tag{6}$$

où φ est la couche linéaire et K est la longueur des entités générées par MLP. Enfin, la couche entièrement connectée a restauré dimensionnellement la caractéristique fusionnée Xˆ et le classificateur SoftMax a obtenu les scores de prédiction finaux. Contrairement aux méthodes existantes qui analysent directement l'état de mouvement de l'ensemble des nuages de points, notre travail se concentre sur la partie critique du corps lors de l'exécution d'actions, ce qui permet de surmonter l'influence des données redondantes telles que l'arrière-plan sur la reconnaissance d'actions 3D. De plus, les détails des parties cruciales et l'apparence globale du corps humain se complètent, ce qui stimule l'extraction de caractéristiques discriminantes pour la reconnaissance d'actions 3D.

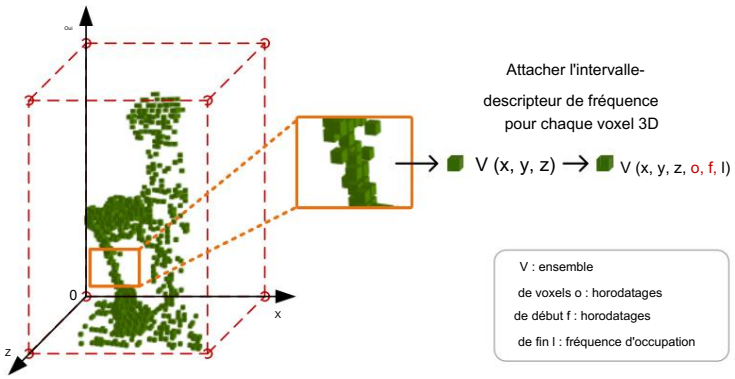


Figure 4. Illustration du descripteur intervalle-fréquence supplémentaire. Il contient les horodatages de début, les horodatages de fin et la fréquence d'occupation globale des voxels et est noté (o, f , l) dans la section 3.4 ; ainsi, les informations spatiales de profondeur 3D sont transformées en caractéristiques de six canaux.

4. Expériences 4.1.

Ensembles de données

Ensemble de données NTU RVB+D 60. Le NTU RGB+D 60 [44] est un ensemble de données de reconnaissance d'actions 3D à grande échelle qui contient environ 56 880 échantillons d'actions RVB+D. Il utilise Microsoft Kinect V2 pour capturer 60 catégories d'actions réalisées par 40 sujets. L'ensemble de données suit deux principes d'évaluation. Dans le cas d'une vue croisée, les échantillons capturés par la caméra 1 ont été utilisés comme ensemble de test, et les caméras 2 et 3 sont considérées comme l'ensemble d'apprentissage. Autrement dit, le nombre d'échantillons de test est de 18 960 et 37 920 échantillons sont utilisés pour la formation. Dans le cas d'un sujet croisé, les données sont divisées en un ensemble de tests de 16 560 échantillons et un ensemble d'apprentissage de 40 320 échantillons en fonction de l'identité du sujet.

Ensemble de données NTU RVB+D 120. NTU RGB+D 120 [45] est un ensemble de données complet pour la reconnaissance d'actions 3D, composé de 114 480 échantillons et 120 catégories d'actions complétées par 106 sujets. Cet ensemble de données contient des actions quotidiennes, des actions médicales et des actions d'interaction à deux. Les échantillons sont collectés dans divers endroits et arrière-plans, désignés par 32 configurations. En plus des paramètres généraux inter-sujets, la configuration inter

l'évaluation est introduite, où l'ensemble d'apprentissage provient d'échantillons avec des ID de configuration impairs, et l'ensemble de tests vient du reste.

4.2. Détails de la formation

Par défaut, le SFCNet et ses variantes sont formés à l'aide du Adaptive Moment Optimiseur d'estimation (Adam) pour 60 époques sous le framework d'apprentissage profond PyTorch, sauf indication contraire. Nous utilisons la perte d'entropie croisée standard et appliquons des techniques d'augmentation des données telles que la rotation aléatoire, le tramage et l'abandon aux données d'entraînement. Le taux d'apprentissage commence à 0,001 et diminue à 0,5 toutes les dix époques. Pour garantir l'équité , nous suivons strictement le schéma de segmentation des échantillons des deux ensembles de données selon des repères.

4.3. Analyse des paramètres

La taille des voxels 3D. Les nuages de points sont généralement constitués d'un grand nombre de points non ordonnés. ensembles de points. En raison de la complexité de ces données, leur stockage peut prendre beaucoup de temps. et le processus. Pour atténuer ce problème, nous intégrons les points non ordonnés de l'espace 3D dans un structure de grille régulière grâce à la rastérisation. Cela convertit le nuage de points en voxels 3D basé sur l'occupation de l'espace, discrétisant l'espace 3D continu en une grille régulière. Ce Le processus fournit une structure régulière et bien comprise, ce qui réduit les calculs complexité et compresse les données du nuage de points, réduisant considérablement les calculs fardeau. Il est essentiel de voxéliser le nuage de points de manière appropriée car la taille du Le voxel 3D détermine la force de compression du nuage de points et la granularité de la représentation du nuage de points. Pour examiner l'impact de la voxélisation sur les résultats, nous évalué les performances de SFCNet sur l'ensemble de données NTU RGB+D 60 pour des voxels. Les résultats sont présentés dans le tableau 1, indiquant que le modèle fonctionne le mieux pour un Taille des cubes de 35 mm. Définir une taille trop grande ou trop petite peut entraîner une diminution de la précision.

Tableau 1. Performances sur l'ensemble de données NTU RGB+D 60 avec voxélisation de différentes tailles.

Taille du voxel (mm)	Sujets transversaux	Vue croisée
25 × 25 × 25 35	87,1%	94,9%
× 35 × 35 45 ×	89,9%	96,7%
45 × 45 55 × 55	88,1%	95,5%
× 55	86,5%	93,6%

Le réglage du seuil θ . Le comportement humain implique généralement uniquement le mouvement des parties spécifiques du corps, telles que l'agitation des bras, la marche des jambes, la rotation de la tête, etc. Cette localité signifie que l'analyse comportementale devrait être davantage axée sur les parties essentielles du corps plutôt que sur les tout le corps. A l'aide de la variable de fréquence d'occupation l dans l'intervalle de fréquence descripteur, nous pouvons décrire l'engagement de chaque voxel, qui est positivement lié à la contribution de la partie du corps dans le processus d'exécution de l'action. Puisque nous échantillonnons le séquence d'action en profondeur en groupes de durée égale, la fréquence d'occupation l positivement est en corrélation avec le nombre d'occupants u de la section 3.3. Ensuite, un seuil θ est utilisé pour évaluer l'attention portée à la partie du corps. Pour étudier l'influence du seuil, nous comparez les performances du SFCNet sur l'ensemble de données NTU RGB+D 60 avec différentes valeurs. Les résultats sont présentés dans le tableau 2. Le résultat optimal peut être obtenu lorsque θ est égal à 30. Nos recherches ont découvert que de légères modifications de θ , de pas plus de 5, entraînaient dans les fluctuations de précision, soulignant l'importance de l'étude des seuils et briser des parties importantes du corps.

Tableau 2. Performances sur l'ensemble de données NTU RGB+D 60 avec différentes valeurs de θ .

La valeur du seuil θ	Sujets transversaux	Vue croisée
15	79,5%	85,1%
20	83,5%	93,2%
25	86,5%	94,4%
30	89,9%	96,7%
35	87,3%	94,9%
40	86,9%	93,7%

4.4. Étude sur l'ablation

Efficacité du descripteur de fréquence d'intervalle. Les données originales du nuage de points 3D uniquement contiennent les informations de localisation des points dans l'espace 3D. Même si la dimension temporelle est introduit dans les séquences de nuages de points, il est encore difficile de décrire l'ensemble dynamique spatio-temporelle du point en s'appuyant uniquement sur ces indices. Nous avons conçu un descripteur de fréquence d'intervalle qui capture l'heure d'apparition et le nombre de voxels occupation. Ces informations supplémentaires nous aident à décrire de manière exhaustive l'humain comportement en capturant des caractéristiques de mouvement supplémentaires. Nous avons mené des études d'ablation sur Ensemble de données NTU RGB+D 60 dans lequel nous avons supprimé les informations sur les caractéristiques de mouvement dans deux flux, et les ensembles de points entrés dans SFCNet n'avaient que des coordonnées 3D (x, y, z). Les résultats de la comparaison sont présentés dans le tableau 3. Nous avons observé que sans caractéristiques tridimensionnelles supplémentaires, il y a eu une dégradation significative des performances de SFCNet de plus de 10 %.

Tableau 3. Efficacité du descripteur de fréquence d'intervalle sur le NTU RGB+D 60.

Caractéristique ponctuelle	Sujets transversaux	Vue croisée
(x, y, z)	78,0%	82,3%
(x, y, z, o, f, l)	89,9%	96,7%

Cela indique que le descripteur de fréquence d'intervalle représente effectivement la dynamique fonctionnalités dans l'ensemble du processus d'action, qui joue un rôle essentiel dans la reconnaissance des actions 3D.

Efficacité de la fusion de fonctionnalités à deux flux. Différentes actions humaines contiennent différentes dynamique globale et locale, nous décrivons le modèle de mouvement global de l'ensemble de l'être humain corps pendant l'action à travers le flux grossier global. En revanche, le flux local décrit la dynamique des parties cruciales du corps, qui accorde plus d'attention aux détails et caractéristiques locales des actions et aide à capturer les changements subtils et les modèles d'action complexes. Les résultats du tableau 4 montrent que notre proposition SFCNet fusionnant les caractéristiques fines-grossières du deux courants peuvent comprendre et reconnaître les actions humaines de manière plus complète. Premièrement, nous discuter des performances de reconnaissance dans l'état à flux unique. On voit que le local Le flux a une plus grande capacité à représenter l'action que le flux global, ce qui indique qu'il Il est essentiel de prêter attention aux principales parties impliquées pour supprimer la redondance. En outre, nous comparons trois stratégies de fusion de fonctionnalités différentes pour démontrer la supériorité de fusion de fonctionnalités basée sur l'attention proposée dans SFCNet (voir l'équation (6)). Comme le montre le tableau 4, SFCNet (fusion) présente des avantages apparents par rapport à la cascade native ou à la fusion additive stratégies. La raison principale est que la fusion de fonctionnalités basée sur l'attention peut attribuer de manière adaptative l'attention du modèle sur les caractéristiques du flux grossier global et du flux fin local.

Comme le montre la figure 5, l'eau potable implique uniquement une interaction entre les mains et la tête, et le l'amplitude du mouvement est faible, le modèle met donc l'accent sur les caractéristiques de la zone locale. flux pour capturer des modèles de mouvement à granularité fine. En revanche, le coup de poing implique l'interaction de deux personnes et le mouvement de frappe est large et puissant, donc une plus grande attention est accordée à l'apparence globale du corps tout en mettant l'accent sur certains détails de les mains et la tête. Ce mécanisme d'extraction de fonctionnalités spécifiques à une action améliore la capacité de généralisation de SFCNet et précision pour la reconnaissance d'actions 3D.

(a) Eau potable

Quantification de l'occupation des voxels

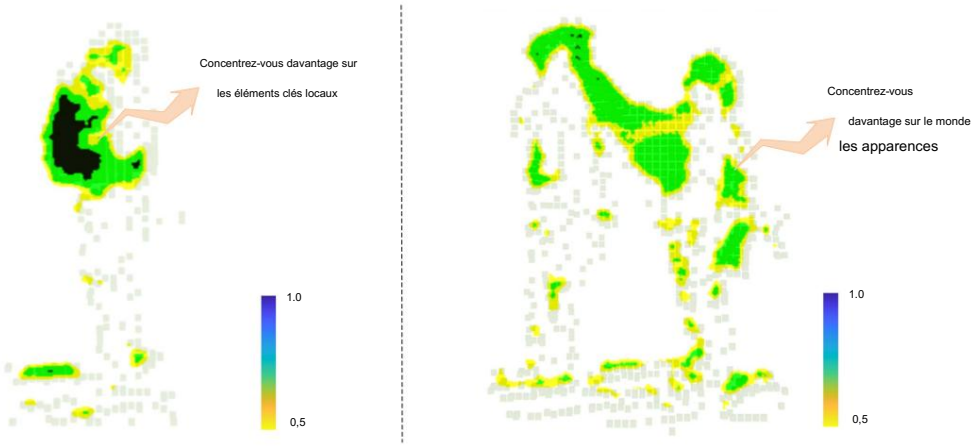
Filtrage binaire

(b) Poinçonnage

$$U_{xyz} = \begin{cases} L_{xyz}, & \text{if } L_{xyz} = 0/1 \\ 0, & \text{if } L_{xyz} \neq 0/1 \end{cases}$$

Tableau 4. Efficacité de la fusion de fonctionnalités à deux flux sur l'ensemble de données NTU RGB+D 60.

Flux d'entrée	Sujets transversaux	Vue croisée
1s-SFCNet (L) 1s-SFCNet (G)	85,0 % 81,8 %	94,6% 86,6%
SFCNet (concaténé)	88,9%	94,8%
SFCNet (ajouter)	86,7%	93,9%
SFCNet (fusion)	89,9%	96,7%



(a) Eau potable (b) Poinçonnage

Figure 5. Visualisation de l'attention des fonctionnalités. Nous visualisons la carte thermique de l'eau potable et du poinçonnage.

4.5. Comparaison avec les méthodes existantes

Pour évaluer les performances du SFCNet proposé, nous le comparons aux méthodes existantes sur deux grands ensembles de données de référence, comme indiqué dans les tableaux 5 et 6. Nous divisons l'action 3D existante méthodes de reconnaissance en méthodes basées sur le squelette et basées sur la profondeur. Dans les méthodes basées sur le squelette , nous comparons différentes méthodes basées sur le squelette, notamment CNN [5], LSTM [15,46] et GCN [6,23,47]. Pour les méthodes basées sur la profondeur, nous comparons les méthodes basées sur des images 2D [19,35,48], Méthodes basées sur CNN 3D [28] et méthodes basées sur des voxels 3D [10]. Les résultats de la comparaison sont indiqué dans les tableaux 5 et 6. Pour l'ensemble de données NTU RGB+D 60, le SFCNet proposé atteint Précision de 89,9 % et 96,7 % dans le cas de paramètres multi-sujets et multi-vues. De plus, nous comparer SFCNet avec deux méthodes basées sur des données multimodales. Par rapport à ED-MHI [31], combinant des données de profondeur et de squelette, notre méthode améliore la précision de 4,3 % dans le cadre transversal. TS-CNN-LSTM [49] a fusionné les données de trois modalités, à savoir RVB, profondeur, et squelette, mais il est respectivement 2,6 % et 4,9 % inférieur à SFCNet dans les paramètres inter-sujets et multi-vues. Pour l'ensemble de données NTU RGB+D 120, SFCNet réalise également des résultats compétitifs résultats, atteignant des précisions de 83,6 % et 93,8 % dans des paramètres multi-sujets et multi-vues, respectivement. En général, SFCNet est efficace et excellent, ce qui surpasse les méthodes traditionnelles utilisant l'extraction manuelle de caractéristiques [19,35,50] et les méthodes d'apprentissage en profondeur qui compressent la profondeur. vidéo en images pour le traitement [2,3,36] ou en séquences de nuages de points [10]. Les résultats expérimentaux prouver que le SFCNet est supérieur pour capturer les modèles de comportement humain discriminants et est donc bénéfique pour la reconnaissance d'actions 3D.

Tableau 5. Comparaison des différentes méthodes de précision de reconnaissance d'action (%) sur le NTU RGB+D 60 ensembles de données.

Méthode	Entrée multi-	Vue croisée	Année
	sujets : squelette 3D		
GCA-LSTM [15]	74,4	82,8	2017
Attention à deux flux LSTM [46]	77.1	85.1	2018
ST-GCN [23]	81,5	88,3	2018
SkeleMotion [5]	69,6	80,1	2019
AS-GCN [6] 2s-	86,8	94,2	2019
AGCN [47]	88,5	95.1	2019
ST-TR (nouveau) [24]	89,9	96.1	2021
DSwarm-Net (nouveau) [51]	85,5	90,0	2022
ActionNet [30]	73.2	76.1	2023
SGMSN (nouveau) [52]	90,1	95,8	2023
	Entrée : cartes de profondeur		
HON4D[19]	30,6	7.3	2013
HOG2 [35]	32.2	22.3	2013
SNV [50]	31,8	13.6	2014
Li. [36]	68.1	83,4	2018
Wang. [2]	87.1	84,2	2018
MVDI [3]	84,6	87,3	2019
3DV-PointNet++ [10]	88,8	96,3	2020
DOGV (nouveau) [53]	90,6	94,7	2021
3DFCNN [28]	78.1	80,4	2022
Taille 3D [54]	83,6	92,4	2022
ConvLSTM (nouveau) [29]	80,4	79,9	2022
CBBMC (nouveau) [48]	83,3	87,7	2023
PointMapNet (nouveau) [55]	89,4	96,7	2023
SFCNet (le nôtre)	89,9	96,7	-
	Entrée : Multimodalités		
ED-MHI [31]	85,6	-	2022
TS-CNN-LSTM [49]	87,3	91,8	2023

Tableau 6. Comparaison des différentes méthodes de précision de reconnaissance d'action (%) sur le NTU RGB+D 120 ensembles de données.

Méthode	Sujets transversaux	Ensemble croisé	Année
	Entrée : squelette 3D		
GCA-LSTM [15]	58.3	59.3	2017
Carte d'évolution de la pose du corps [56]	64,6	66,9	2018
Attention à deux flux LSTM [46]	61.2	63,3	2018
ST-GCN [23]	70,7	73.2	2018
Ligne de base NTU RVB+D 120 [45]	55,7	57,9	2019
FSNet [57]	59,9	62,4	2019
SkeleMotion [5]	67,7	66,9	2019
TSRJI [20]	67,9	62,8	2019
AS-GCN [6] 2s-	77,9	78,5	2019
AGCN [47]	82,9	84,9	2019
ST-TR (nouveau) [24]	82,7	84,7	2021
SGMSN (nouveau) [52]	84,8	85,9	2023
	Entrée : Cartes de profondeur		
APSR [45]	48,7	40,1	2019
3DV-PointNet++ [10]	82,4	93,5	2020
DOGV (nouveau) [53]	82,2	85,0	2021
Taille 3D [54]	76,6	88,8	2022
SFCNet (le nôtre)	83,6	93,8	-

5. Discussion

Afin d'analyser les avantages et les inconvénients de la méthode proposée, nous avons présenté la précision de reconnaissance de SFCNet sur l'ensemble de données NTU RGB+60 sur les paramètres inter-sujets pour chaque catégorie. Les résultats sont présentés sous la forme d'une matrice de confusion dans la figure 6 (à gauche). Nous avons sélectionné quelques actions déroutantes pour un affichage plus clair et les avons agrandies localement, comme le montre la figure 6 (à droite). Les résultats montrent que SFCNet possède une solide capacité d'analyse de l'action humaine, avec une précision de reconnaissance supérieure à 90 % dans la plupart des catégories. Par exemple, il a atteint une précision de 100 % pour sauter et de 99 % pour sauter . Cependant, SFCNet est confus quant à la reconnaissance de certaines actions similaires. Par exemple, lire et écrire, porter des chaussures et enlever ses chaussures sont les paires d'échantillons les plus déroutantes. De plus, 25 % de ceux qui jouent avec leur téléphone ont été classés à tort comme lisant (9 %), écrivant (8 %) et tapant sur le clavier (8 %). La précision de la frappe sur le clavier n'est que de 66 % et 12 % des échantillons sont classés à tort comme écrits. À partir de l' analyse, nous avons constaté que ces actions ne présentent que des différences subtiles et que l' amplitude du mouvement est faible. C'est la principale raison pour laquelle de telles actions sont difficiles à distinguer.

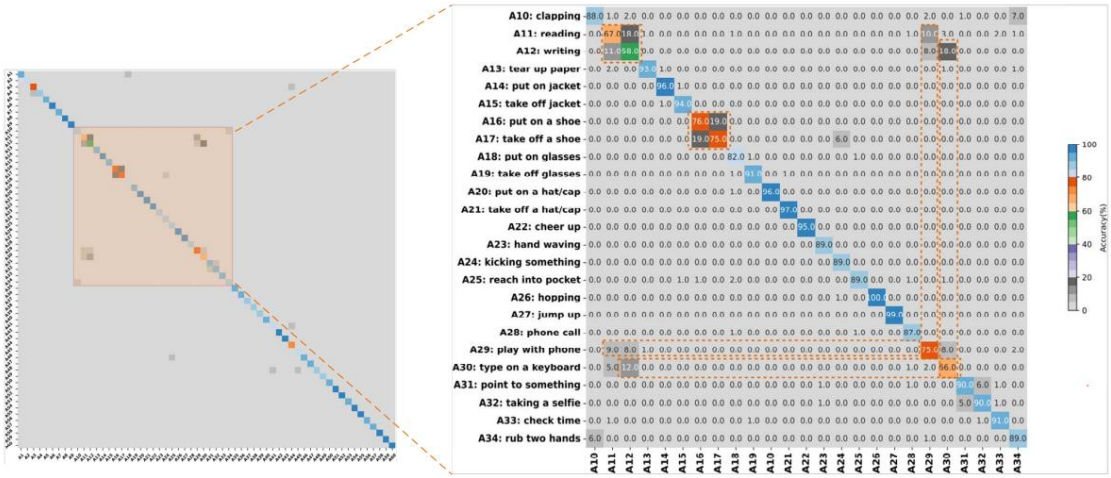


Figure 6. Matrice de confusion pour la précision de la reconnaissance spécifique à une classe. L'image de (à gauche) contient toutes les catégories d'actions. Pour souligner certaines actions d'obscurcissement, un zoom local est affiché à droite .

6. Conclusions

Dans cet article, nous proposons un réseau neuronal symétrique, SFCNet, pour reconnaître les actions 3D à partir de séquences de nuages de points. Il contient un flux grossier global et un flux fin local qui utilise PointNet++ comme extracteur de fonctionnalités. Les séquences de nuages de points sont régularisées sous forme d' ensembles de voxels structurés ajoutés au descripteur de fréquence d'intervalle proposé pour générer des caractéristiques 6D qui capturent des informations dynamiques spatio-temporelles. Le flux grossier global capture les modèles d'action à gros grain en fonction de l'apparence du corps humain, et le flux local délicat extrait les caractéristiques à grain fin spécifiques à l'action des parties critiques. Après la fusion des fonctionnalités, SFCNet peut exploiter des modèles de mouvement discriminants qui impliquent des changements spatiaux globaux et mettent l'accent sur les détails cruciaux de bout en bout. Selon les résultats expérimentaux sur deux grands ensembles de données de référence, NTU RGB+D 60 et NTU RGB+D 120, le SFCNet est efficace pour la reconnaissance d'actions 3D et présente un potentiel pour des applications Cependant, le SFCNet proposé présente encore des limites pour distinguer des actions similaires. Nos futurs travaux se concentreront sur la reconnaissance d'actions similaires et la capture de modèles subtils pour améliorer la précision.

Contributions des auteurs : Conceptualisation, CL et QH ; méthodologie, CL, WQ et XL ; logiciels, QH et YM ; validation, CL, WQ et XL ; analyse formelle, QH et YM ; enquête, CL et WQ ; ressources, QH et YM ; conservation des données, WQ ; rédaction : préparation de l'ébauche originale, CL et WQ ; rédaction : révision et édition, CL, YM, WQ, QH et XL ; visualisation, WQ ; surveillanc

QH et YM ; administration de projet, QH et YM ; acquisition de financement, QH, YM et CL Tous les auteurs ont lu et accepté la version publiée du manuscrit.

Financement : Cette recherche a été financée par le programme d'innovation de recherche et de pratique postuniversitaire de la province du Jiangsu (numéro de subvention KYCX23_0753), les fonds de recherche fondamentale pour les universités centrales (numéro de subvention B230205027), le programme clé de recherche et de développement de Chine (numéro de subvention 2022YFC3005401). , le programme clé de recherche et de développement de la Chine, province du Yunnan (numéro de subvention 202203AA080009), le 14e plan quinquennal pour les sciences de l'éducation de la province du Jiangsu (numéro de subvention D/2021/01/39) et le projet de recherche sur la réforme de l'enseignement supérieur du Jiangsu. (numéro de subvention 2021JSJG143) ; et l'APC a été financé par les Fonds de recherche fondamentale des universités centrales.

Déclaration du comité d'examen institutionnel : sans objet.

Déclaration de consentement éclairé : sans objet.

Déclaration de disponibilité des données : les ensembles de données NTU RGB+D 60 et NTU RGB+D 120 utilisés dans cet article sont publics, gratuits et disponibles sur : https://rose1.ntu.edu.sg/dataset/actionRecognition/ (consulté le 21 décembre 2020).

Conflits d'intérêts : Les auteurs ne déclarent aucun conflit d'intérêts.

Les références

1. Riaz, W. ; Gao, C. ; Azim, A. ; Saifullah ; Bux, JA; Ullah, A. Système de prévision des anomalies de trafic utilisant un réseau prédictif. Télésens.2022 , 14 , 1–19.

2. Wang, P. ; Li, W. ; Gao, Z. ; Tang, C. ; Ogunbona, PO Reconnaissance d'actions 3D à grande échelle basée sur la mise en commun de profondeur avec des réseaux de neurones convolutifs. IEEETrans. Multimed. 2018, 20, 1051-1061.
3. Xiao, Y. ; Chen, J. ; Wang, Y. ; Cao, Z. ; Zhou, JT ; Bai, X. Reconnaissance d'actions pour la vidéo en profondeur à l'aide d'images dynamiques multi-vues. Inf. Sci. 2019, 480, 287-304.
4. Li, C. ; Huang, Q. ; Li, X. ; Wu, Q. Reconnaissance de l'action humaine basée sur des cartes de caractéristiques multi-échelles à partir de séquences vidéo en profondeur. Multimed. Outils Appl. 2021, 80, 32111-32130.
5. Caetano, C. ; Sena, J. ; Brémond, F. ; Dos Santos, JA; Schwartz, WR Skelemotion : une nouvelle représentation des séquences d'articulations du squelette basée sur les informations de mouvement pour la reconnaissance d'actions en 3D. Dans les actes de la conférence internationale de l'IEEE sur la vidéosurveillance avancée et basée sur les signaux, Taipei, Taiwan, 18-21 septembre 2019 ; IEEE : Piscataway, New Jersey, États-Unis, 2019 ; p. 1 à 8.
6. Li, M. ; Chen, S. ; Chen, X. ; Zhang, Y. ; Wang, Y. ; Tian, Q. Réseaux convolutifs de graphes actionnels-structurels pour la reconnaissance d'actions basée sur un squelette. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Long Beach, Californie, États-Unis, 15-20 juin 2019 ; pp. 3595-3603.
7. Yang, X. ; Zhang, C. ; Tian, Y. Reconnaître les actions à l'aide d'histogrammes de gradients orientés basés sur des cartes de mouvement en profondeur. Dans Actes de la Conférence internationale ACM sur le multimédia, Nara, Japon, 29 octobre-2 novembre 2012 ; pp. 1057-1060.
8. Elmadany, NED; Hé.; Guan, L. Fusion d'informations pour la reconnaissance de l'action humaine via la localité de globalité biset/multiset préserver l'analyse de corrélation canonique. IEEETrans. Processus d'images. 2018, 27, 5275-5287.
9. Lui, K. ; Zhang, X. ; Ren, S. ; Sun, J. Apprentissage résiduel profond pour la reconnaissance d'images. Dans Actes de la conférence IEEE sur la vision par ordinateur et la reconnaissance de formes, Las Vegas, NV, États-Unis, 27-30 juin 2016 ; pp. 770-778.
10. Wang, Y. ; Xiao, Y. ; Xiong, F. ; Jiang, W. ; Cao, Z. ; Zhou, JT ; Yuan, J. 3DV : Voxel dynamique 3D pour la reconnaissance d'action dans une vidéo en profondeur. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Seattle, WA, États-Unis, 14-19 juin 2020 ; pp. 508-517.
11. Qi, CR ; Yi, L. ; Su, H. ; Guibas, LJ Pointnet++ : Apprentissage hiérarchique approfondi de fonctionnalités sur des ensembles de points dans un espace métrique. Av. Informations neuronales. Processus. Système. 2017, 30, 5105-5114.
12. Wang, P. ; Li, W. ; Gao, Z. ; Zhang, J. ; Tang, C. ; Ogunbona, PO Reconnaissance d'actions à partir de cartes de profondeur utilisant la convolution profonde les réseaux de neurones. IEEETrans. Hum. -Mach. Système. 2015, 46, 498-509.
13. Ke, Q. ; Bennamoun, M. ; Un, S. ; Sohel, F. ; Boussaid, F. Une nouvelle représentation de séquences squelettes pour la reconnaissance d'actions 3D. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Honolulu, HI, États-Unis, 21-26 juillet 2017 ; pp. 3288-3297.
14. Li, C. ; Zhong, Q. ; Xie, D. ; Pu, S. Fonctionnalité de cooccurrence apprenant à partir de données squelettes pour la reconnaissance et la détection d'actions avec agrégation hiérarchique. arXiv 2018, arXiv:1804.06055v1 .
15. Liu, J. ; Gang, W. ; Ping, H. ; Duan, LY; Kot, AC Réseaux LSTM d'attention contextuelle mondiale pour la reconnaissance d'actions 3D. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Honolulu, HI, États-Unis, 21-26 juillet 2017 ; pages 3671 à 3680.
16. Shi, L. ; Zhang, Y. ; Cheng, J. ; Lu, H. Reconnaissance d'actions basée sur le squelette avec des réseaux neuronaux à graphes dirigés. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Long Beach, Californie, États-Unis, 15-20 juin 2019 ; pages 7912 à 7921.

17. Si, C. ; Chen, W. ; Wang, W. ; Wang, L. ; Tan, T. Un réseau lstm convolutif de graphes amélioré pour l'attention pour la reconnaissance d'actions basée sur un squelette . Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Long Beach, Californie, États-Unis, 15-20 juin 2019 ; pages 1227 à 1236.

18. Yu, Z. ; Wenbin, C. ; Guodong, G. Évaluation des caractéristiques des points d'intérêt spatio-temporels pour la reconnaissance d'actions basée sur la profondeur. Image Vis. Calculer. 2014, 32, 453-464.
19. Oreifej, O. ; Liu, Z. Hon4d : Histogramme de normales 4D orientées pour la reconnaissance d'activité à partir de séquences de profondeur. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Portland, OR, États-Unis, 23-28 juin 2013 ; pp. 716-723.
20. Caetano, C. ; Brémond, F. ; Schwartz, WR Représentation d'image squelette pour la reconnaissance d'actions 3D basée sur la structure arborescente et les articulations de référence. Dans Actes de la trente-deuxième conférence SIBGRAPI sur les graphiques, les motifs et les images, Rio de Janeiro, Brésil, 28-31 octobre 2019 ; p. 16-23.

21. Liu, J. ; Shahroudy, A. ; Xu, D. ; Wang, G. Lstm spatio-temporel avec portes de confiance pour la reconnaissance de l'action humaine en 3D. Dans Actes de la Conférence européenne sur la vision par ordinateur, Amsterdam, Pays-Bas, 11-14 octobre 2022 ; Springer : Berlin/Heidelberg, Allemagne, 2016 ; pp. 816-833.

22. Zhang, P. ; Lan, C. ; Xing, J. ; Zeng, W. ; Xue, J. ; Zheng, N. Afficher les réseaux de neurones adaptatifs pour des performances élevées basées sur un squelette reconnaissance de l'action humaine. IEEETrans. Modèle Anal. Mach. Intell. 2019, 41, 1963-1978.
23. Yan, S. ; Xiong, Y. ; Lin, D. Réseaux convolutifs de graphes temporels spatiaux pour la reconnaissance d'actions basée sur un squelette. Dans Actes de la trente-deuxième conférence AAAI sur l'intelligence artificielle, Nouvelle-Orléans, LA, États-Unis, 2-7 février 2018 ; pages 7444 à 7452.
24. Plizzari, C. ; Cannici, M. ; Matteucci, M. Reconnaissance d'actions basée sur le squelette via des réseaux de transformateurs spatiaux et temporels. Calculer. Vis. Image comprise. 2021, 208-209, 103219.
25. Shotton, J. ; Fitzgibbon, A. ; Cuisinier, M. ; Pointu, T. ; Finocchio, M. ; Moore, R. ; Kipman, A. ; Blake, A. Reconnaissance de pose humaine en temps réel dans certaines parties à partir d'images à profondeur unique. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Colorado Springs, CO, États-Unis, 20-25 juin 2011 ; IEEE : Piscataway, New Jersey, États-Unis, 2011 ; pp. 1297-1304.
26. Xiong, F. ; Zhang, B. ; Xiao, Y. ; Cao, Z. ; Yu, T. ; Zhou, JT ; Yuan, J. A2j : Réseau de régression ancre-à-joint pour l'estimation de pose articulée 3D à partir d'une image de profondeur unique. Dans les actes de la conférence internationale de l'IEEE sur la vision par ordinateur, Séoul, République de Corée, 27 octobre-2 novembre 2019 ; pp. 793-802.
27. Kamel, A. ; Sheng, B. ; Yang, P. ; Lèvre.; Shen, R. ; Feng, DD Réseaux de neurones à convolution profonde pour la reconnaissance de l'action humaine Utilisation des cartes de profondeur et des postures. IEEETrans. Système. Homme Cybern. Système. 2019, 49, 1806-1819.
28. Sánchez-Caballero, A. ; de López-Diz, S. ; Fuentes-Jimenez, D. ; Losada-Gutiérrez, C. ; Marrón-Romera, M. ; Casillas-Pérez, D. ; Sarker, MI 3DFCNN : Reconnaissance d'actions en temps réel à l'aide de réseaux neuronaux profonds 3D avec des informations brutes de profondeur. Multimed. Outils Appl. 2022, 81, 24119-24143.

29. Sánchez-Caballero, A. ; Fuentes-Jiménez, D. ; Losada-Gutiérrez, C. Reconnaissance de l'action humaine en temps réel à l'aide d'une vidéo en profondeur brute. basés sur des réseaux de neurones récurrents. Multimed. Outils Appl. 2022, 82, 16213-16235.
30. Kumar, DA; Kishore, PVV ; Murthy, G. ; Chaitanya, TR ; Subhani, S. Visualisez la reconnaissance invariante de l'action humaine à l'aide de cartes de surface via des réseaux convolutifs. Dans Actes de la Conférence internationale sur les méthodologies de recherche en gestion des connaissances, en intelligence artificielle et en ingénierie des télécommunications, Chennai, Inde, 1er et 2 novembre 2023 ; p. 1 à 5.
31. Ghosh, Saskatchewan ; M.; Mohan, BR ; Guddeti, RMR Reconnaissance d'actions humaines 3D multi-vues basée sur l'apprentissage profond à l'aide de données de squelette et de profondeur. Multimed. Outils Appl. 2022, 82, 19829-19851.
32. Li, R. ; Li, X. ; Fu, CW; Cohen-Or, D. ; Heng, PA Pu-gan : un réseau contradictoire de suréchantillonnage de nuages de points. Dans Actes de la conférence internationale IEEE/CVF sur la vision par ordinateur, Séoul, République de Corée, 27 octobre-2 novembre 2019 ; pages 7203 à 7212.
33. Qi, CR ; Litanie, O. ; Lui, K. ; Guibas, LJ Deep a voté pour la détection d'objets 3D dans les nuages de points. Dans Actes de la conférence internationale IEEE/CVF sur la vision par ordinateur, Séoul, République de Corée, 27 octobre-2 novembre 2019 ; pages 9277 à 9286.
34. Thomas, H. ; Qi, CR ; Deschaud, JE; Marcotegui, B. ; Goulette, F. ; Guibas, L. KPConv : Convolution flexible et déformable pour les nuages de points. Dans Actes de la conférence internationale IEEE/CVF sur la vision par ordinateur, Séoul, République de Corée, 27 octobre-2 novembre 2019 ; pages 6410 à 6419.

35. Ohn-Bar, E. ; Trivedi, similitudes des angles articulaires MM et HOG2 pour la reconnaissance des actions. Dans les actes de la conférence IEEE/CVF sur les ateliers de vision par ordinateur et de reconnaissance de formes, Portland, Oregon, États-Unis, 23-28 juin 2013 ; pp. 465-470.
36. Li, J. ; Wong, Y. ; Zhao, Q. ; Kankanhalli, MS Apprentissage non supervisé des représentations d'actions invariantes de vue. Av. Informations neuronales. Processus. Système. 2018, 31, 1262-1272.
37. Liu, X. ; Qi, CR ; Guibas, LJ Flownet3d : Apprentissage du flux de scènes dans des nuages de points 3D. Dans les actes de la conférence IEEE/CVF sur Vision par ordinateur et reconnaissance de formes, Long Beach, Californie, États-Unis, 15-20 juin 2019 ; pp. 529-537.
38. Zhai, M. ; Xiang, X. ; Lv, N. ; Kong, X. Estimation du flux optique et du flux de scène : une enquête. Reconnaissance de modèles. 2021, 114, 107861.
39. Fernando, B. ; Gavves, E. ; Oramas, J. ; Ghodrati, A. ; Tuytelaars, T. Mise en commun des classements pour la reconnaissance des actions. IEEETrans. Modèle Anal. Mach. Intell. 2016, 39, 773-787.
40. Liu, J. ; Xu, D. GeometryMotion-Net : Une base de référence solide à deux flux pour la reconnaissance d'actions 3D. IEEETrans. Système de circuits. Technologie vidéo . 2021, 31, 4711-4721.
41. Dou, W. ; Menton, WH ; Kubota, N. Réseau de mémoire croissant avec 3DCNN à poids aléatoire pour la reconnaissance continue de l'action humaine. Dans les actes de la conférence internationale de l'IEEE sur les systèmes flous, Incheon, République de Corée, 13-17 août 2023 ; p. 1 à 6.

42. Fan, H. ; Yu, X. ; Ding, Y. ; Yang, Y. ; Kankanhalli, M. PSTNet : Convolution spatio-temporelle de points sur des séquences de nuages de points. Dans Actes de la Conférence internationale sur les représentations d'apprentissage, Addis-Abeba, Éthiopie, 26-30 avril 2020 ; p. 1 à 6.
43. Qi, CR ; Su, H. ; Mo, K. ; Guibas, LJ Pointnet : Deep learning sur des ensembles de points pour la classification et la segmentation 3D. Dans Actes de la conférence IEEE sur la vision par ordinateur et la reconnaissance de formes, Honolulu, HI, États-Unis, 21-26 juillet 2017 ; pp. 652-660.
44. Shahroudy, A. ; Liu, J. ; Ng, TT; Wang, G. NTU RGB+D : un ensemble de données à grande échelle pour l'analyse de l'activité humaine en 3D. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Las Vegas, NV, États-Unis, 27-30 juin 2016 ; pp. 1010-1019.
45. Liu, J. ; Shahroudy, A. ; Pérez, M. ; Wang, G. ; Duan, LY; Kot, AC Ntu rgb+ d 120 : une référence à grande échelle pour l'activité humaine en 3D compréhension. IEEETrans. Modèle Anal. Mach. Intell. 2019, 42, 2684-2701.
46. Liu, J. ; Wang, G. ; Duan, LY; Abdiyeva, K. ; Kot, AC Reconnaissance de l'action humaine basée sur le squelette et tenant compte du contexte mondial Attention aux réseaux LSTM. IEEETrans. Processus d'images. 2018, 27, 1586-1599.
47. Shi, L. ; Zhang, Y. ; Cheng, J. ; Lu, H. Réseaux convolutionnels de graphes adaptatifs à deux flux pour la reconnaissance d'actions basée sur un squelette. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Long Beach, Californie, États-Unis, 15-20 juin 2019 ; pages 12026 à 12035.
48. Li, X. ; Huang, Q. ; Wang, Z. Fusion d'informations spatiales et temporelles pour la reconnaissance de l'action humaine via Center Boundary Balancing Classificateur multimodal. J. Vis. Commun. L'image représente. 2023, 90, 103716.
49. Zan, H. ; Zhao, G. Recherche sur la reconnaissance de l'action humaine basée sur les réseaux Fusion TS-CNN et LSTM. Arabe. J. Sci. Ing. 2023, 48, 2331-2345.
50. Yang, X. ; Tian, Y. Vecteur super normal pour la reconnaissance d'activité à l'aide de séquences de profondeur. Dans les actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Columbus, OH, États-Unis, 23-28 juin 2014 ; pp. 804-811.
51. Basak, H. ; Kundu, R. ; Singh, PK ; Ijaz, MF; Wozniak, M. ; Sarkar, R. Une union d'apprentissage en profondeur et d'optimisation basée sur des essaims pour la reconnaissance de l'action humaine en 3D. Sci. Rep.2022 , 12, 1-17.
52. Qi, Y. ; Hu, J. ; Zhuang, L. ; Pei, X. Reconnaissance d'action du squelette humain multi-échelle guidée par la sémantique. Appl. Intell. Int. J.Artif. Intell. Réseau neuronal. Problème complexe. -Résoudre la technologie. 2023, 53, 9763-9778.
53. Ji, X. ; Zhao, Q. ; Cheng, J. ; Ma, C. Exploiter la représentation spatio-temporelle pour la reconnaissance 3D de l'action humaine à partir d'une carte de profondeur séquences. Connaître. -Système basé. 2021, 227, 107040.
54. Guo, J. ; Liu, J. ; Xu, D. 3D-Pruning : un cadre de compression de modèles pour une reconnaissance efficace des actions 3D. IEEETrans. Système de circuits . Technologie vidéo. 2022, 32, 8717-8729.
55. Li, X. ; Huang, Q. ; Zhang, Y. ; Yang, T. ; Wang, Z. PointMapNet : réseau de cartes de caractéristiques de nuages de points pour la reconnaissance de l'action humaine en 3D. Symétrie 2023, 15, 1–17.
56. Liu, M. ; Yuan, J. Reconnaître les actions humaines comme l'évolution des cartes d'estimation de pose. Dans Actes de la conférence IEEE/CVF sur la vision par ordinateur et la reconnaissance de formes, Salt Lake City, UT, États-Unis, 18-23 juin 2018 ; pp. 1159-1168.
57. Liu, J. ; Shahroudy, A. ; Wang, G. ; Duan, LY; Kot, AC Prédiction d'action en ligne basée sur Skeleton utilisant un réseau de sélection d'échelle. IEEETrans. Modèle Anal. Mach. Intell. 2019, 42, 1453-1467.