angewandte Wissenschaften

MDPI

Artikel

# Ein neuartiges symmetrisches Fein-Grob-Neuralnetzwerk für 3D-Menschen

## Aktionserkennung auf Basis von Punktwolkensequenzen

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

[1] Hochschule für Informatik und Softwaretechnik, Hohai-Universität, Nanjing 211100, China; lichang@hhu.edu.cn (CL); yingchimao@hhu.edu.cn (YM); qianweiwen@hhu.edu.cn (WQ)

[2] Hochschule für Informationswissenschaft und Technologie und Hochschule für künstliche Intelligenz, Nanjing Forestry Universität, Nanjing 210037, China; lixing@njfu.edu.cn

***** Korrespondenz: huangqian@hhu.edu.cn

**Zusammenfassung:** Die Erkennung menschlicher Handlungen hat die Entwicklung von Geräten mit künstlicher Intelligenz erleichtert, die sich auf menschliche Aktivitäten und Dienste konzentrieren. Diese Technologie hat sich weiterentwickelt durch die Einführung 3D-Punktwolken, die von Tiefenkameras oder Radargeräten abgeleitet werden. Das menschliche Verhalten ist jedoch komplex und Die beteiligten Punktwolken sind riesig, ungeordnet und kompliziert und stellen eine Herausforderung für 3D-Aktionen dar. Erkennung. Um diese Probleme zu lösen, schlagen wir ein symmetrisches fein-grobes neuronales Netzwerk (SFCNet) vor. das gleichzeitig das Aussehen und die Details menschlicher Handlungen analysiert. Zunächst werden die Punktwolkensequenzen transformiert und in strukturierte 3D-Voxelsätze voxelisiert. Diese Sätze werden dann erweitert mit einem Intervall-Frequenz-Deskriptor zur Generierung von 6D-Features zur Erfassung der räumlich-zeitlichen Dynamik Informationen. Durch die Auswertung der Voxelraumbelegung mittels Schwellenwertbildung können wir effektiv die wesentliche Teile. Danach werden alle Voxel mit dem 6D-Merkmal in den globalen Grobstrom geleitet, während die Voxel innerhalb der Schlüsselteile zum lokalen Feinstrom geleitet werden. Diese beiden Ströme extrahieren globale Erscheinungsmerkmale und kritische Körperteile durch die Verwendung von symmetrischem PointNet++. Anschließend Die Aufmerksamkeitsmerkmalfusion wird eingesetzt, um besser diskriminierbare Bewegungsmuster adaptiv zu erfassen. Experimente mit den öffentlichen Benchmark-Datensätzen NTU RGB+D 60 und NTU RGB+D 120 bestätigen Die Effektivität und Überlegenheit von SFCNet bei der 3D-Aktionserkennung.

**Schlüsselwörter:** Punktwolkenanalyse; 3D-Aktionserkennung; Mustererkennung; Deep Learning



**Zitat:** Li, C.; Huang, Q.; Mao, Y.; Qian, W.; Li, X. Eine neuartige symmetrische Fein-Grobes neuronales Netzwerk für 3D Erkennung menschlicher Handlungen basierend auf Punktwolkensequenzen. Anwendungswissenschaft. **2024,** 14, 6335. https://doi.org/10.3390/app14146335

Wissenschaftlicher Herausgeber: Atsushi Mase

Empfangen: 11. Juni 2024

Überarbeitet: 8. Juli 2024

Akzeptiert: 18. Juli 2024

Veröffentlicht: 20. Juli 2024



### 1. Einleitung

Die Erkennung menschlicher Handlungen soll Computern helfen, die Semantik menschlichen Verhaltens anhand verschiedener Daten zu verstehen, die von den Erfassungsgeräten aufgezeichnet werden. Insbesondere 3D-Aktionen Anerkennung widmet sich der Gewinnung von Aktionsmustern aus 3D-Daten, die menschliche Bewegungen beinhalten . Es hat aufgrund seiner weit verbreiteten Anwendungen zunehmend Aufmerksamkeit erregt, wie z. B. Überwachung der öffentlichen Sicherheit, Leistungsbeurteilung, militärische Aufklärung und intelligente Transport [1].

Die derzeit gängigen 3D-Aktionserkennungsmethoden können je nach verwendetem Datentyp in tiefenbasierte Methoden (einschließlich Tiefenkarten und Punktwolkensequenzen) [2–4] und skelettbasierte Methoden [5,6] eingeteilt werden. Begrenzt durch die Genauigkeit Pose-Estimation-Algorithmus – die unvermeidliche vorgelagerte Aufgabe – skelettbasierte Methoden stehen vor Herausforderungen in puncto Rechenleistung und Robustheit. Im Gegensatz dazu Methoden sind aufgabenunabhängiger und haben breite Aufmerksamkeit erregt. Bestehende tiefenbasierte 3D-Aktionserkennungsansätze lassen sich hauptsächlich in zwei Hauptkategorien einteilen. Die erste besteht darin, 3D-Bewegungen in ein oder mehrere Bilder zu kodieren [2,3,7,8] und CNNs zu nutzen [9]. zur Aktionserkennung. Die 2D-Bildebene kann jedoch die 3D-Bildebene nicht vollständig charakterisieren. Dynamik, weil menschliche Handlungen gleichzeitig raumzeitlich und in der 3D-Raum. Die andere Möglichkeit besteht darin, das Tiefenvideo in eine Punktwolkensequenz umzuwandeln [10], die zeichnet die 3D-Koordinaten von Punkten im Raum zu mehreren Zeitpunkten auf. Im Vergleich Bei Bildern haben Punktwolkensequenzen den Vorteil, dass sie das 3D-Erscheinungsbild beibehalten und

Geometriedynamik im Zeitverlauf, was eine erweiterte Analyse und ein besseres Verständnis menschlicher Handlungen ermöglicht. Darüber hinaus können Punktwolken mithilfe verschiedener Geräte wie Laserscannern , Radaren, Tiefensensoren und RGB+D-Kameras gewonnen werden, die an Drohnen, Straßenlaternen, Fahrzeugen und Überwachungsflugzeugen montiert werden können, wodurch der Anwendungsbereich der Aktionserkennung erweitert wird . Aufgrund der komplexen Struktur und des enormen Volumens der Punktwolke sind die darauf basierenden bestehenden 3D-Aktionserkennungsmethoden jedoch mit den folgenden Herausforderungen verbunden.

Erstens haben Punktwolkensequenzen immer massive Punkte, die proportional zur Zeitdimension sind , und das Datenverarbeitungsschema ist zeitaufwändig. Daher ist die Entwicklung eines effizienten und leichten Punktwolkensequenzmodells für die 3D-Aktionserkennung von entscheidender Bedeutung.
Zweitens sind die Punkte in den Sequenzen unregelmäßig und weisen ungeordnete räumliche Informationen innerhalb des Bildes und geordnete zeitliche Details zwischen den Bildern auf, was die Analyse der zugrunde liegenden Bewegungsmuster erschwert. Bestehende Methoden zur Verarbeitung von Punktwolken führen jedoch normalerweise ein undifferenziertes Downsampling der gesamten Punktwolken durch, was zu einem gleichmäßigen Verlust wesentlicher und subtiler Informationen führt. Darüber hinaus ignorieren bestehende Punktwolkenanalyseschemata die kritischen Körperteile, die zu den Aktionen beitragen, was zu einem Mangel an Nuancen der extrahierten Aktionsmerkmale führt, was letztendlich die Leistung der Aktionserkennung einschränkt.

Zur Lösung dieser Probleme schlagen wir ein Deep-Learning-Framework namens Symmetric Fine-coarse Neural Network (SFCNet) vor, das die Analyse von Bewegungsmerkmalen aus lokaler und globaler Perspektive symmetrisch kombiniert, wie in Abbildung 1 dargestellt. Um Rechenkosten zu sparen , reduzieren wir zunächst die Punkte durch Frame-Sampling und Farthest-Point-Sampling. Als nächstes werden die abgetasteten Punktwolken in 3D-Voxel umgewandelt, um eine kompakte Punktwolkendarstellung zu erstellen . Die ursprünglichen 3D-Positionen werden dann mit einem Intervall-Frequenz- Deskriptor versehen, um die gesamte räumliche Konfiguration darzustellen und die Identifizierung wesentlicher Körperteile zu erleichtern, sodass wir die Punktwolkensequenzen in einen lokalen Feinraum und einen globalen Grobraum aufteilen können. Wir behandeln die in diesen beiden Räumen beteiligten Voxel als Punkte und verwenden PointNet++ [11], um Merkmale durchgängig zu extrahieren. Schließlich kombiniert unser Feature-Fusion-Modul das globale Erscheinungsbild und lokale Details, um Unterscheidungsmerkmale für die 3D-Aktionserkennung zu erhalten. Die umfangreichen Experimente mit den groß angelegten Datensätzen NTU RGB+D 60 und NTU RGB+D 120 demonstrieren die Wirksamkeit und Leistungsfähigkeit von SFCNet, mit dem die menschliche Absicht bei Fernerkundungsanwendungen beurteilt und unterstützt werden k

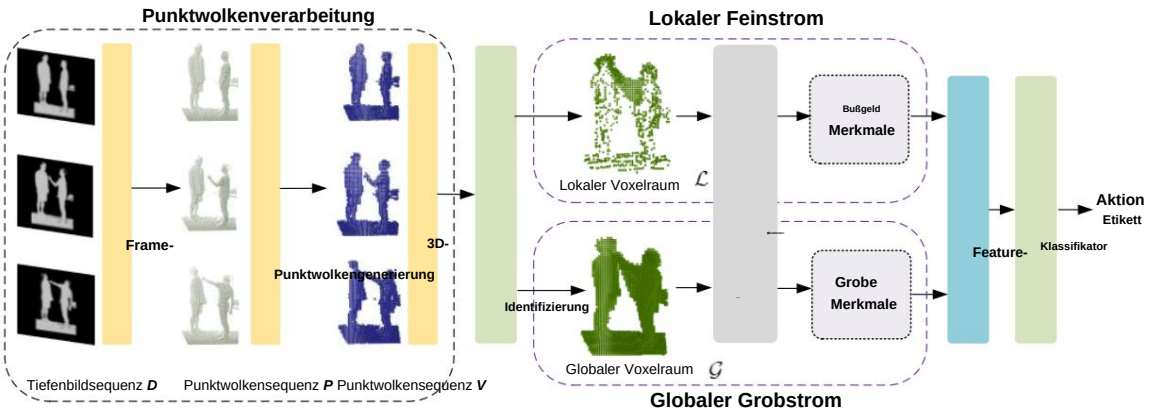



**Abbildung 1.** Die Pipeline von SFCNet. Sie konvertiert Tiefenbilder in eine Punktwolke und wendet Voxelisierungsoperationen an . Eine symmetrische Struktur kodiert 3D-Voxel, wobei wichtige Teile und globale dynamische Informationen separat verarbeitet werden. Der angehängte Intervall-Frequenz-Deskriptor charakterisiert zunächst die Bewegungsinformationen und wird dann von PointNet++ [11] für tiefere Merkmale verarbeitet. Schließlich erkennt der Klassifikator 3D-Aktionen mithilfe des aggregierten Merkmals.

Im Allgemeinen sind die wichtigsten Beiträge unserer Arbeit wie folgt: •

Wir schlagen einen Intervall-Frequenz-Deskriptor vor, um die 3D-Voxel während der Aktionsausführung zu charakterisieren, der die Bewegungsdetails vollständig bewahrt und wichtige Hinweise für die Wahrnehmung wichtiger Körperteile liefert. Nach bestem Wissen ist unsere Arbeit die erste, die Punktwolkensequenzen auf diese Weise verarbeitet.

- Wir konstruieren ein Deep-Learning-Framework namens SFCNet, das zunächst eine symmetrische Struktur zur Verarbeitung von Punktwolkensequenzen verwendet. Es kodiert die Dynamik der lokalen, wichtigen Körperteile über einen feinen Strom und ergänzt diese komplizierten Details dann zum globalen Erscheinungsbild, das von einem groben Strom erfasst wird. Das SFCNet kann wichtige Körperteile hervorheben und diskriminierendere Bewegungsmuster erfassen und löst so das Problem der effektiven Aktionsdarstellung auf der Grundlage von Punktwolken. • Das vorgestellte SFCNet hat seine überlegene Genauigkeit anhand von zwei öffentlich verfügbaren Datensätzen, NTU RGB+D 60 und NTU RGB+D 120, unter Beweis gestellt, was beweist, dass unsere Methode erhebliches Potenzial bei der Erkennung verschiedener Arten von Aktionen hat, wie z. B. alltägliche Aktionen, medizinisch bedingte Aktionen und Aktionen der Interaktion zwischen zwei Personen.

## 2. Verwandte Werke

### 2.1. Skelettbasierte 3D-Aktionserkennung

Bestehende Methoden zur 3D-Aktionserkennung können in skelettbasierte Methoden [12–17] und tiefenbasierte Methoden [3,7,12,18,19] eingeteilt werden . Es gibt im Allgemeinen vier gängige Ansätze für die skelettbasierte Aktionserkennung. Der erste Ansatz besteht darin, CNN [12] zu verwenden, um die räumlich-zeitlichen Muster aus Pseudobildern zu lernen [13,14]. Caetano et al. [20] führten das Tree Structure Reference Joints Image (TSRJI) ein, um Skelettsequenzen darzustellen. Der zweite Ansatz besteht darin, die Skelettsequenzen als Zeitreihen zu betrachten [15–17] und Backbones wie RNN [21] zur Merkmalsextraktion zu verwenden. Der dritte Ansatz besteht darin, die Skelettdaten als Graphen [6,22] mit Gelenken als Eckpunkten und Knochen als Kanten anzuzeigen und sich GCN [16] zur Aktionsdarstellung zuzuwenden. Beispielsweise konnte ST-GCN [23] die zeitlichen dynamischen Informationen von Skelettsequenzen effektiv darstellen, indem es räumlich-zeitliche Graphenfaltungs- und Partitionierungsstrategien verwendete. SkeleMotion [5] erfasste zeitlich dynamische Informationen, indem es die Größe und Ausrichtung von Skelettgelenken in verschiedenen Zeitskalen berechnete. Die vierte Möglichkeit besteht darin, das Skelett über Transformer als Token zu kodieren. Plizzari et al. [24] verwendeten räumlich-zeitliche Aufmerksamkeit in Transformer und erfassten eine dynamische Inter-Frame-Beziehung von Gelenken. Da es jedoch immer noch erhebliche Herausforderungen bei der genauen dreidimensionalen Schätzung der menschlichen Pose gibt [25,26], leiden skelettbasierte Aktionserkennungsmethoden aufgrund dieser unvermeidlichen vorgelagerten Aufgabe unter Leistungseinbußen.

### 2.2. Tiefenbasierte 3D-Aktionserkennung

Bei der tiefenbasierten 3D-Aktionserkennung stellen frühe Ansätze Tiefenvideos hauptsächlich durch manuelle Deskriptoren dar [19]. Yang et al. [7] konstruierten Tiefenbewegungskarten (DMMs), indem sie die Unterschiede zwischen den Bildern der projizierten Tiefenbilder stapelten. Dann berechneten sie das Histogramm der orientierten Gradienten (HOGs), um die Aktionen darzustellen. Solche Methoden haben eine begrenzte Ausdruckskraft und benötigen daher normalerweise Hilfe bei der Erfassung räumlich-zeitlicher Informationen. In den letzten Jahren sind Deep-Learning-Methoden mit der Entwicklung neuronaler Netzwerke zum Mainstream geworden. Die meisten Forscher versuchten, Tiefenvideos in Bilder zu komprimieren und Bewegungsmuster mithilfe von CNNs zu analysieren [12]. Kamel et al. [27] gaben Tiefenbewegungsbilder (DMIs) und bewegliche Gelenkdeskriptoren (MJDs) in CNNs zur Aktionserkennung ein. Um räumlich-zeitliche Informationen von Tiefensequenzen zu kodieren, schlugen Adrián et al. [28] 3D-CNN vor, um Bewegungsmerkmale zu extrahieren . Darüber hinaus schlugen sie ConvLSTM [29] vor , um diskriminierende Bewegungsmuster aus langen Kurzzeiteinheiten zu akkumulieren. Xiao et al. [3] rotierten die virtuelle Kamera im 3D-Raum, um ein Rohtiefenvideo aus verschiedenen virtuellen Bildperspektiven dicht zu projizieren und so dynamische Multiview-Bilder zu erstellen. Für perspektivische Ansichtsinvarianten schlugen Kumar et al. [30] ein ActionNet auf CNN-Basis vor und trainierten es mit einem Multiview-Datensatz, der mit fünf Tiefenkameras erfasst wurde. Ghosh et al. [31] berechneten ein Multiview-Tiefendeskriptor- Edge-Detected-Motion-History-Bild (ED-MHI) als Eing Wang et al. [2] verwendeten segmentierte Tiefenvideosequenzen, um drei Arten dynamischer Tiefenbilder zu erzeugen. Aufgrund ihrer kompakten räumlichen Struktur hat die 2D-Tiefenkarte jedoch immer noch Schwierigkeiten, 3D-Bewegungsmuster vollständig auszunutzen [10].

In jüngster Zeit wurden durch die Umwandlung von Tiefenkarten in Punktwolken zur Verarbeitung sowohl im Bereich der Erkennung als auch der Segmentierung bessere Ergebnisse erzielt. Zahlreiche Studien haben gezeigt,

dass Punktwolken aufgrund ihrer Eigenschaften wie Unordnung und Rotationsinvarianz erhebliche Vorteile bei der Darstellung dreidimensionaler räumlicher Informationen haben . Deep Learning für Punktwolken wird nicht nur häufig bei Klassifizierungs- und Segmentierungsaufgaben eingesetzt, sondern hat sich auch bei der Szenenrekonstruktion [32] und der Zielerkennung [33] als äußerst nützlich erwies Die oben genannten Methoden konzentrieren sich jedoch nur auf Merkmale innerhalb statischer Punktwolken. Bei der Verwendung von Punktwolken zur 3D-Aktionserkennung ist es notwendig, dynamische Merkmale entsprechend den Zeitintervallen und den Erscheinungsmerkmalen des gesamten Aktionsprozesses zu extrahieren. Der Schlüssel zu einer effizienten Verarbeitung von Punktwolkensequenzen liegt in der Auswahl einer geeigneten Punktwolkenanalysemethode . Thomas et al. [34] entwickelten eine Methode, die von der bildbasierten Faltung inspiriert ist, und verwendeten eine Reihe von Kernelpunkten, um jedes Kernelgewicht zu verteilen. Als effizientes Werkzeug zur Analyse und Verarbeitung von Punktsätzen wird PointNet++ [11] häufig für die 3D-Aktionserkennung basierend auf Punktwolkensequenzen eingesetzt. Die erste Methode ist 3DV [10], die eine 3D-Voxelisierung der Punktwolkensequenzen durchführt und das 3D-Erscheinungsbild durch räumliche Besetzung beschreibt, und für die 3DV-Extraktion wird zeitliches Rangpooling verwendet. Diese Methode konzentriert sich in erster Linie auf die allgemeine Bewegung und die Erscheinungsänderungen einer Aktion. Sie ignoriert jedoch die Details der Aktion, wie z. B. eine subtile Handbewegung, was ihre Fähigkeit, das Verhalten genau darzustellen, einschränkt . Unser Ziel ist es daher, die entscheidenden Teile von Aktionen und ihre sensiblen Informationen zu erfassen, um sie als robustere menschliche Aktionen zu erkennen.

## **3. Methodik** 3.1.

Pipeline Die

Pipeline des vorgeschlagenen SFCNet ist in Abbildung 1 dargestellt. Zunächst wird jeder Tiefenrahmen in eine Punktwolke umgewandelt, um die dynamischen und Erscheinungsbildmerkmale im 3D-Raum besser zu erhalten. Um die Analyse der Raumnutzung zu erleichtern und den lokalen Raum abzugrenzen, führen wir Voxelisierungsoperationen an den Punktwolken durch. Als Nächstes erstellen wir ein symmetrisches Framework zur Kodierung von 3D-Voxeln, wobei Schlüsselteile und globale dynamische Informationen separat im lokalen Feinstrom und im globalen Grobstrom verarbeitet werden. Dann fügen wir den Intervall-Frequenz-Deskriptor hinzu, um Bewegungsinformationen zu ergänzen. Wir verwenden PointNet++ [11], um Bewegungsmuster zu erfassen und das aggregierte Merkmal zur 3D-Aktionserkennung an den Klassifikator zu senden.

### 3.2. Dreidimensionale Voxelerzeugung

Tiefenvideo hat im Vergleich zu RGB-Modal den Vorteil, dass es externen Störungen wie Hintergrund und Licht widersteht, da es die Tiefeninformationen des Aktionsobjekts enthält. Im Wesentlichen ist Tiefenvideo eine Art Zeitreihendaten, die aus in chronologischer Reihenfolge angeordneten Tiefenkarten bestehen. Mathematisch kann ein Tiefenvideo mit t Bildern als D = {d1, d2, . . . , dt} definiert werden , wobei di eine Tiefenkarte von t Bildern ist, in der jedes Pixel eine 3D-Koordinate (x, y, z) darstellt und z die Entfernung von der Tiefenkamera ist. Da es unmöglich ist, die Wichtigkeit einer Aktion in der Zeitdimension anhand eines einzigen Kriteriums zu klassifizieren, kann uns eine gleichmäßige Abtastung helfen, den gesamten Bewegungsvorgang besser zu verstehen als eine zufällige Abtastung [10]. Daher tasten wir das Tiefenvideo zunächst gleichmäßig ab, um den Rechenaufwand zu verringern und gleichzeitig die Integrität der Aktion aufrechtzuerhalten. Die Tiefensequenz nach der Abtastung wird als Dˆ = {d1, d2, . . . , dT} bezeichnet , wobei T die Anzahl der Bilder und die

Der Standardwert ist 64.

Einige aktuelle Methoden zur Aktionserkennung [35,36] verwenden die Abbildung von Tiefenbildern in 2D- Räume zur direkten Verarbeitung. Obwohl diese Ansätze manchmal gute Ergebnisse erzielen , können sie das Problem der unzureichenden Darstellung von 3D-Informationen nicht lö Um die menschliche Bewegung im 3D-Raum besser darzustellen, transformieren wir daher jedes Bild di in eine Punktwolke P = {p1, p2, . . . , pn}, wobei n die Anzahl der Punkte ist, und erzeugen so eine Punktwolkensequenz S = { P1, P2, . . . , PT}. Beim Erzeugen von Punktwolken sind intrinsische Parameter der Kamera erforderlich, da sie das Bildmodell der Kamera definieren, einschließlich

Brennweite und Hauptpunktkoordinaten (cx, cy). Für jedes Pixel (x, y, z) im Tiefenbild kann die entsprechende Punktwolke p(x ) mit der folgenden Formel berechnet werden:

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

wobei fx und fy die Brennweite der Tiefenkamera in horizontaler und vertikaler Richtung darstellen, die aus den Geräteparametern abgerufen werden kann. fz ist standardmäßig auf 1 eingestellt .

Im Gegensatz zu herkömmlichen Bildern (regulär strukturierte Daten) sind die Punkte in der Punktwolke ungeordnet, sodass ihre Verarbeitung schwierig ist. Viele vorhandene Algorithmen sind für reguläre Gitterdaten ausgelegt. Die ungeordnete Punktwolke ist jedoch eine Gruppe zufällig verteilter Punkte im 3D-Raum, sodass ihre Struktur komplex zu verarbeiten und direkt zu analysieren ist. Um dieses Problem zu lösen, transformieren wir die Punktwolke durch Voxelisierung in ein reguläres 3D-Gitter (Voxelraum), um die Punktwolkendarstellung zu regulieren. Zuerst definieren wir die Größe des Voxelgitters Vgrid = (Vx, Vy, Vz) in dreidimensionalen Koordinaten, die die Auflösung des Voxelisierungsprozesses bestimmen. Jede Zelle in diesem Gitter ist ein potenzielles Voxel und die Größe jeder Zelle wird als Vvoxel(dx, dy, dz) bezeichnet. Gegeben sei ein Punkt p(x, y, z) in der Punktwolke, der auf das Gitter abgebildet wird, indem der entsprechende Voxelindex Vindex(x, y, z) gemäß der folgenden Gleichung ermittelt wird:

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

wobei xmin, ymin und zmin die Mindestkoordinaten aller Punktwolken sind. dx, dy und dz werden als Gesamtgröße geteilt durch die Anzahl der Zellen in jeder Dimension (Vx, Vy, Vz) berechnet. Die Bodenfunktion ⌊·⌋ rundet auf den nächsten Punkt ab. Wir definieren, dass ein Voxel besetzt ist, wenn es eine Punktwolke enthält. Dann können die 3D-Erscheinungsinformationen beschrieben werden, indem beobachtet wird, ob die Voxel besetzt sind oder nicht, wobei der ausgeschlossene Punkt außer Acht gelassen wird, wie in Gleichung (3) dargestellt :

$$V_{voxel}(x, y, z) = \begin{cases} 1, & \text{wenn } V_{voxel}(x, y, z) \text{ besetzt ist} \\ 0, & \text{ansonsten} \end{cases}, \quad (3)$$

wobei $V_{voxel}(x, y, z)$ bezeichnet einen bestimmten Voxel im t-ten Frame. (x, y, z) ist der reguläre 3D-Positionsindex , also Vindex in Gleichung (2). Diese Strategie hat zwei Hauptvorteile. Erstens sind die erzeugten binären 3D-Voxelsätze regulär, wie in Abbildung 2 dargestellt. Dadurch wird die Komplexität der Punktwolkenverarbeitung reduziert. Darüber hinaus kann die Voxelisierung Punktwolken effektiv komprimieren, da benachbarte Voxel ähnliche Eigenschaften haben können. Diese Komprimierung reduziert nicht nur die Anzahl der Punkte, sondern trägt auch dazu bei, den Speicher- und Berechnungsaufwand zu reduzieren

### 3.3. Identifizierung und Darstellung wichtiger Teile Das

entscheidende Problem bei 3D-Aktionserkennungsaufgaben ist die effiziente Erfassung und Darstellung dynamischer Merkmale in Punktwolkensequenzen. Derzeit können auf Szenenfluss basierende Schätzmethoden [37,38] dabei helfen, 3D-Bewegungen zu verstehen, aber dies ist sehr zeitaufwändig. Einige Studien verwenden zeitliches Rangpooling [3,39], um Bewegungsprozesse im 3D-Raum durch Aufteilung von Zeitsegmenten zu bewahren. Diese Methoden können mehr zeitliche Informationen erfassen, unterteilen jedoch häufig nur eine kleine Anzahl von Intervallen, was zu grobkörnigen dynamischen Merkmalen führt. Wir schlagen ein Modul zur Identifizierung und Kodierung wichtiger Teile vor, um sich besser auf die kritische Dynamik während der Bewegung zu konzentrieren . Es kann die Hauptteile aus dem globalen 3D-Voxelraum entsprechend den Raumbelegungszeiten extrahieren und die Aktionsdet
Konkret analysieren wir zunächst die Raumbelegung, indem wir einen 3D-Raum U mit den genauen Grenzen als Punktwolkensequenzen für jeden räumlichen Standort und den Anfangswerten konstruieren

von U werden auf 0 gesetzt. Wir verarbeiten die m gleichmäßigen Gruppen der Folge Dˆ der Reihe nach. Dann ist die 3D Der Platzbedarf kann nach Gleichung (4) berechnet werden:

$$Uvoxel(x, y, z) = \begin{matrix} Uvoxel(x, y, z) + 1, \text{ vi } \ddot{y} = 0 \\ Uvoxel(x, y, z), \text{ andernfalls} \end{matrix} \quad (4)$$

Der gesamte Raumbedarf u für jede Position ergibt sich aus der Zählung aller m Punktmengen. Da die Punktmengen außerdem natürlicherweise zeitlich geordnet sind, können wir leicht Notieren Sie die erste und letzte Zeit, f und l, für jeden räumlichen Standort. Dann Wir definieren den Schwellenwert *ÿ,* um den prominenten lokalen Raum aufzuteilen. Die besetzten Orte weniger als *ÿ* werden als zufälliges Rauschen behandelt, und diejenigen, die mehr als und weniger als m aufzeichnen bilden die kritischen Bewegungsanteile S gemäß Gleichung (5):

$$Svoxel(x, y, z) = \begin{matrix} 0, & Uvoxel(x, y, z) < \ddot{y} \\ 1, \ddot{y} \ddot{y} & Uvoxel(x, y, z) \ddot{y} m \end{matrix} \quad (5)$$

Wenn der Wert von Svoxel(x, y, z) gleich 0 ist, bedeutet dies, dass der Voxel zum globalen Raum G, andernfalls gehört es zum lokalen Raum L. Verglichen mit der Punktwolke Verarbeitungsmethoden [10,40], die üblicherweise einheitliche Downsampling-Operationen anwenden, Die vorgeschlagene Methode ist effektiver, insbesondere bei Aktionen, an denen nur eine kleine Anzahl beteiligt ist von Gliedmaßenteilen, denn die Teilung des lokalen Raumes kann nicht nur den Hintergrund überwinden Effekte bis zu einem gewissen Grad, sondern auch effektiv erhöhen den Goldgehalt der beprobten Punkt Daten. Wie in Abbildung 2 dargestellt, bewahrt der lokale Raum L die detaillierten Informationen der Hauptkörperteile, die wichtige Hinweise für die 3D-Aktionserkennung liefern und gleichzeitig wesentlich Reduzierung der Redundanz.

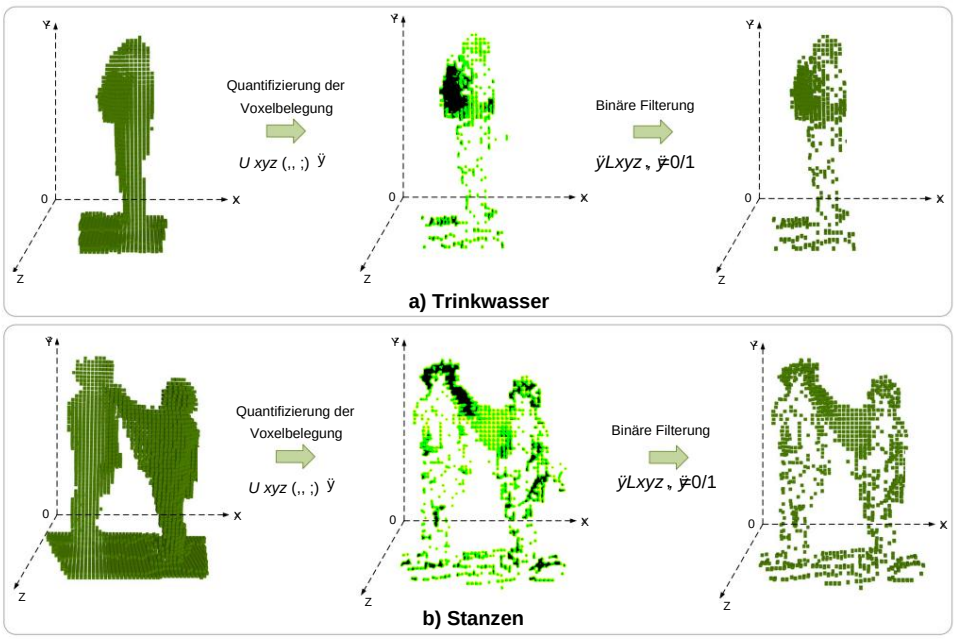



**Abbildung 2.** Der Prozess der lokalen Raumaufteilung. Wir quantifizieren den Beitrag eines Voxels zu einer Aktion durch seine Raumbelegungszahl. Durch Festlegen eines Schwellenwerts können die kritischen Teile als komprimierte lokaler Speicherplatz, der redundante Informationen entfernt und die Rechenlast reduziert.

### 3.4. Symmetrische Merkmalsextraktion

Für die verarbeiteten 3D-Voxel ist der intuitivste Weg die Verwendung von 3DCNN [ 41,42]. es ist durch die Voxelgröße begrenzt und zeitaufwendig. Wir entscheiden uns für PointNet++ [11] als Feature-Extraktor in dieser Arbeit, wie in Abbildung 3 dargestellt. Er ist explizit für hierarchische Feature-Learning auf ungeordneten Punktmengen im metrischen Raum, wodurch es lokale feinkörnige Muster in Punktwolken. Um dies zu erreichen, partitioniert PointNet++ die Punktwolke

Cloud in überlappende lokale Regionen basierend auf einer Distanzmetrik im zugrunde liegenden Raum. Um detaillierte 3D-visuelle Hinweise zu erhalten, verwendet PointNet++ rekursiv PointNet [43], um lokale Merkmale zu extrahieren, die dann für eine globale Erscheinungsanalyse zusammengeführt werden. PointNet++ ist eine hervorragende Alternative zu 3DCNNs, da es lokale 3D-Muster, die für die Aktionserkennung wichtig sind, gut erfasst. Darüber hinaus ist die Anwendung relativ unkompliziert und erfordert lediglich die Umwandlung von 3D-Voxeln in Punktmengen.

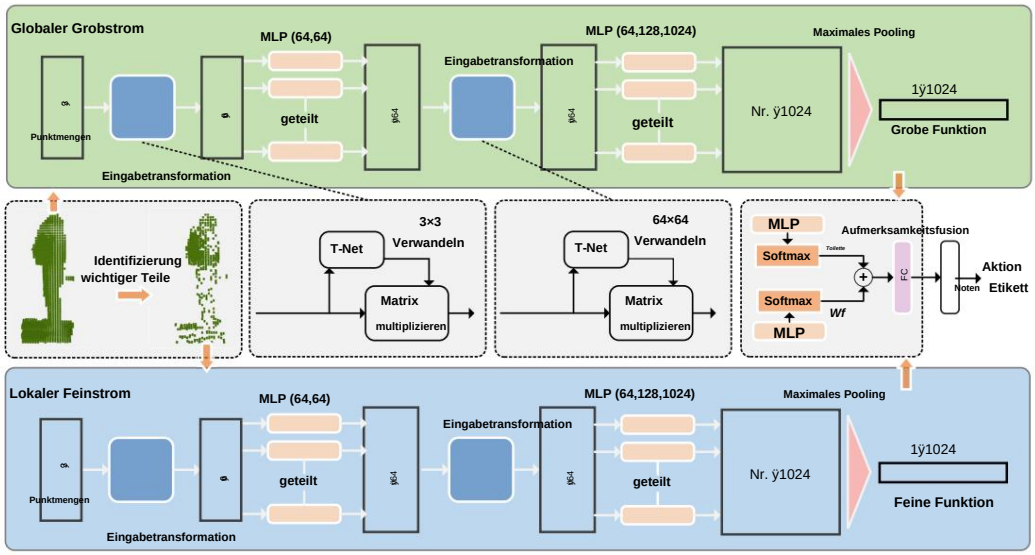



**Abbildung 3.** Die Netzwerkstruktur von SFCNet. Es handelt sich um eine symmetrische Netzwerkstruktur, die aus einem globalen Grobstrom und einem lokalen Feinstrom besteht. Der globale Grobstrom verwendet die globalen Erscheinungsinformationen als Eingabe, während der lokale Feinstrom nur die komprimierten Schlüsselteilinformationen übernimm Die Merkmale jedes Streams werden von PointNet++ extrahiert und mit den lernbaren Gewichten zu den endgültigen Aktionsmerkmalen zusammengeführt, die zur 3D-Aktionserkennung an den Klassifikator gesendet werden.

Um PointNet++ anzupassen, wird jeder Punkt p(x, y, z) als Voxel Vvoxel(x, y, z) mit der beschreibenden Eigenschaft (x, y, z, I) abstrahiert, wobei I der Intervall-Frequenz-Deskriptor ist, der drei Variablen (o, f , I) enthält , die Startzeitstempel, Endzeitstempel bzw. die Gesamtbelegungsfrequenz der Voxel bezeichnen. Wir hängen I an die ursprünglichen 3D-Positionen der Voxel an, um die 6D-Punkte p zu erhalten

ŷ
(x, y, z, o, f , I), wie in Abbildung 4 dargestellt. Mathematisch können wir für einen Voxel Vvoxel die Start- und Endzeit (o und f) seiner Raumbelegung voxel(x, y, z) = 1 mithilfe von Gleichung (3) bestimmen , die den Zeitindex angibt, wenn die Bedingungen V t und V

T

voxel(x, y, z) = 0 sind zuerst erfüllt. Zusätzlich kann die gesamte Raumbelegung u mit Gleichung (4) berechnet werden. Folglich kann die Belegungshäufigkeit I als I = u/(f ÿ o) berechnet werden . Der Intervall-Häufigkeits-Deskriptor deckt Zeitintervalle ab, und die Gesamtbelegungshäufigkeit der räumlichen Standorte kann uns nicht nur helfen, die umgekehrten Aktionen zu unterscheiden, die denselben Raum abdecken, sondern auch, die Unterschiede zwischen Aktionen hervorzuheben, indem detaillierte Informationen innerhalb des Aktionsprozesses beibehalten werden. Schließlich werden die erhaltenen Punktsätze als Eingabe von PointNet++ verwendet, um Aktionsmerkmale zu extrahieren.

Darüber hinaus entwerfen wir ein Zwei-Stream-Netzwerk, um die Punktwolke im globalen bzw. lokalen Raum zu verarbeiten (siehe Abbildung 3). Der globale Raum enthält alle voxelisierten Punkte, und der globale grobe Stream erfasst die gesamten Bewegungsmuster. In Anbetracht der Tatsache , dass lebenswichtige Körperteile gezieltere und differenziertere dynamische Informationen für die Aktionserkennung liefern können, teilen wir die voxelisierten Punkte wichtiger Körperteile in lokale Räume auf und geben den feinen Stream ein, um Merkmale zu extrahieren. Danach wird das Merkmalsfusionsmodul für die fein-grobe Aktionsdarstellung eingerichtet. In Anbetracht der Merkmale verschiedener Aktionen ist ihre Abhängigkeit von globalen und lokalen Merkmalen unterschiedlich. Bei Aktionen, die nur Gliedmaßenbewegungen beinhalten, wie Winken und Treten, sollte das Modell lokale feinkörnige Stream-Merkmale betonen. Im Gegensatz dazu sollte sich das Modell bei großen Ganzkörperbewegungen wie Fallen und Springen auf die Merkmale des globalen Streams konzentrieren. Um di

Für diese aktionsspezifische Wahrnehmung verwenden wir das Merkmalsfusionsmodul mit Aufmerksamkeitsmechanismus, der aus dem Feinsinn extrahiert wird. Zuerst projizieren wir die Merkmale Xf ÿ R bzw. grobe Ströme in den unteren Merkmalsraum, um den Rechenaufwand zu reduzieren und X und X c zu erhalten. Dann werden die Zwischenmerkmale durch ein mehrschichtiges f- Perzeptron (MLP) extrahiert. Danach werden die lernbaren Gewichte Wf und Wc des globalen groben Stroms bzw. des lokalen feinen Stroms durch die Aktivierungsfunktion SoftMax erhalten .

Abschließend werden die globalen und lokalen Merkmale gemäß Gleichung (6) fusioniert , um das Bewegungsmerkmal Xˆ zu erhalten:

$$\begin{aligned} Wf &= \text{SoftMax MLP } X_f \\ Wc &= \text{SoftMax MLP } X_c \\ \hat{X} &= \ddot{y} \sum_{i=1}^{K} W_{LAN}^{(i)} X^{(i)} + \text{Mit c} X_c^{(i)} \end{aligned} \tag{6}$$

wobei *ÿ* die lineare Schicht und K die Länge der von MLP ausgegebenen Merkmale ist. Schließlich stellte die vollständig verbundene Schicht das fusionierte Merkmal Xˆ dimensional wieder her und der SoftMax -Klassifikator erhielt die endgültigen Vorhersagewerte. Anders als bestehende Methoden, die den Bewegungszustand der gesamten Punktwolken direkt analysieren, konzentriert sich unsere Arbeit auf den kritischen Körperteil bei der Ausführung von Aktionen, was dazu beiträgt, den Einfluss redundanter Daten wie des Hintergrunds auf die 3D-Aktionserkennung zu überwinden. Darüber hinaus ergänzen sich die Details der entscheidenden Teile und das allgemeine Erscheinungsbild des menschlichen Körpers, was die diskriminative Merkmalsextraktion für die 3D-Aktionserkennung fördert.

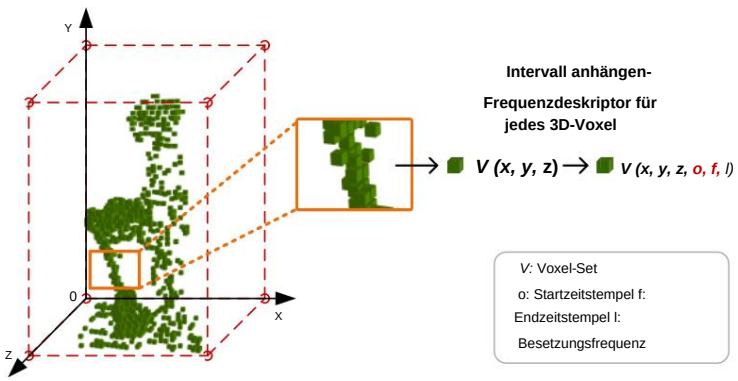



**Abbildung 4.** Darstellung des zusätzlichen Intervall-Frequenz-Deskriptors. Er enthält Startzeitstempel, Endzeitstempel und die Gesamtbelegungsfrequenz der Voxel und wird in Abschnitt 3.4 als (o, f , l) bezeichnet; somit werden die 3D-Tiefenrauminformationen in Merkmale von sechs Kanälen umgewandelt.

## 4. Experimente

### 4.1. Datensätze

**NTU RGB+D 60-Datensatz.** NTU RGB+D 60 [44] ist ein umfangreicher 3D-Aktionserkennungsdatensatz , der rund 56.880 RGB+D-Aktionsbeispiele enthält. Er verwendet Microsoft Kinect V2, um 60 Aktionskategorien zu erfassen, die von 40 Probanden ausgeführt werden. Der Datensatz folgt zwei Bewertungsprinzipien. Im Fall der Queransicht wurden die von Kamera 1 erfassten Beispiele als Testsatz verwendet, und Kameras 2 und 3 gelten als Trainingssatz. Das heißt, die Anzahl der Testbeispiele beträgt 18.960 und 37.920 Beispiele werden für das Training verwendet. Im Fall der Queransicht werden die Daten basierend auf der ID des Probanden in einen Testsatz mit 16.560 Beispielen und einen Trainingssatz mit 40.320 Beispielen aufgeteilt.

**NTU RGB+D 120-Datensatz.** NTU RGB+D 120 [45] ist ein umfangreicher Datensatz zur 3D- Aktionserkennung, der aus 114.480 Beispielen und 120 Aktionskategorien besteht, die von 106 Probanden ausgefüllt wurden. Dieser Datensatz enthält alltägliche Aktionen, medizinisch bedingte Aktionen und Aktionen der Interaktion zwischen zwei Personen . Die Beispiele werden an verschiedenen Orten und mit verschiedenen Hintergründen gesammelt, die als 32 Setups bezeichnet werden. Zusätzlich zu den allgemeinen probandenübergreifenden Einstellungen enthält das Setup-übergreifend

Auswertung eingeführt wird, wobei der Trainingssatz aus Proben mit ungeraden Setup-IDs stammt, und aus dem Rest ergibt sich der Testsatz.

### 4.2. Schulungsdetails

Standardmäßig werden das SFCNet und seine Varianten mit dem Adaptive Moment trainiert. Schätzungsoptimierer (Adam) für 60 Epochen unter dem PyTorch Deep Learning Framework, sofern nicht anders angegeben. Wir verwenden den standardmäßigen Cross-Entropy-Loss und wenden Datenerweiterungstechniken wie zufällige Rotation, Dithering und Dropout auf die Trainingsdaten an. Die Lernrate beginnt bei 0,001 und sinkt alle zehn Epochen um 0,5. Um Fairness zu gewährleisten , folgen wir strikt dem Mustersegmentierungsschema der beiden Datensätze gemäß Benchmarks.

### 4.3. Parameteranalyse

**Die Größe von 3D-Voxeln.** Punktwolken bestehen typischerweise aus einer großen Anzahl ungeordneter Punktemengen. Aufgrund der Komplexität dieser Daten kann die Speicherung sehr zeitaufwändig sein. und verarbeiten. Um dieses Problem zu lösen, betten wir die ungeordneten Punkte im 3D-Raum in ein regelmäßige Gitterstruktur durch Rasterung. Dadurch wird die Punktwolke in 3D-Voxel umgewandelt basierend auf der Raumbelegung, wobei der kontinuierliche 3D-Raum in ein regelmäßiges Raster diskretisiert wird. Dies Der Prozess bietet eine regelmäßige und gut verstandene Struktur, die den Rechenaufwand reduziert Komplexität und komprimiert die Punktwolkendaten, was die Rechenleistung deutlich reduziert Belastung. Es ist wichtig, die Punktwolke angemessen zu voxelisieren, da die Größe der 3D-Voxel bestimmt die Stärke der Punktwolkenkompression und die Granularität der die Punktwolkendarstellung. Um den Einfluss der Voxelisierung auf die Ergebnisse zu untersuchen, evaluierte die Leistung von SFCNet auf dem NTU RGB+D 60 Datensatz für verschiedene Voxel. Die Ergebnisse sind in Tabelle 1 dargestellt und zeigen, dass das Modell am besten funktioniert für eine Würfelgröße 35 mm. Wenn Sie die Größe zu groß oder zu klein einstellen, kann dies zu einer Verringerung der Genauigkeit führen.

**Tabelle 1.** Leistung im NTU RGB+D 60-Datensatz mit Voxelisierung unterschiedlicher Größe.

| Voxelgröße (mm) | Themenübergreifend | Queransicht |
|---|---|---|
| 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 % |

**Die Festlegung der Schwelle** ÿ. Menschliches Verhalten umfasst in der Regel nur die Bewegung von bestimmte Körperteile, wie Armbewegungen, Beinbewegungen, Kopfdrehungen usw. Diese Lokalität bedeutet dass die Verhaltensanalyse sich mehr auf wesentliche Körperteile konzentrieren sollte als auf die ganzen Körper. Mit Hilfe der Belegungshäufigkeitsvariable l im Intervall-Frequenz Deskriptor können wir das Engagement jedes Voxels beschreiben, das positiv verwandt ist mit der Beitrag des Körperteils im Aktionsausführungsprozess. Da wir die Tiefenwirkungsfolge in Gruppen gleicher Dauer, die Belegungshäufigkeit l positiv korreliert mit der Anzahl der Belegung u in Abschnitt 3.3. Dann wird ein Schwellenwert *ÿ* verwendet, um Bewerten Sie die Aufmerksamkeit, die dem Körperteil geschenkt wird. Um den Einfluss der Schwelle zu untersuchen, Vergleichen Sie die Leistung des SFCNet im NTU RGB+D 60-Datensatz mit verschiedenen Werten. Die Ergebnisse sind in Tabelle 2 dargestellt. Das optimale Ergebnis kann erreicht werden, wenn *ÿ* gleich 30. Unsere Untersuchungen ergaben, dass geringfügige Änderungen von ÿ um nicht mehr als 5 dazu führten, in Schwankungen in der Genauigkeit, was die Bedeutung der Untersuchung von Schwellenwerten unterstreicht und Abbau wichtiger Körperteile.

**Tabelle 2.** Leistung im NTU RGB+D 60-Datensatz mit verschiedenen ÿ-Werten.

| Der Wert des Schwellenwertes *ÿ* | Themenübergreifend | Queransicht |
|---|---|---|
| 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. Ablationsstudie

**Wirksamkeit des Intervall-Frequenz-Deskriptors.** Nur die ursprünglichen 3D-Punktwolkendaten enthalten die Standortinformationen der Punkte im 3D-Raum. Selbst wenn die Zeitdimension in die Punktwolkensequenzen eingeführt wird, ist es immer noch eine Herausforderung, die Gesamtheit zu beschreiben räumlich-zeitliche Dynamik des Punktes, indem wir uns nur auf diese Hinweise verlassen. Wir haben ein Intervall-Frequenz-Deskriptor, der den Beginn und die Anzahl der Voxel erfasst Belegung. Diese zusätzlichen Informationen helfen uns, den menschlichen Verhalten durch Erfassung zusätzlicher Bewegungsmerkmale. Wir führten Ablationsstudien an den NTU RGB+D 60-Datensatz, bei dem wir Bewegungsmerkmalinformationen in zwei Streams entfernt haben, und Die in SFCNet eingegebenen Punktmengen hatten nur 3D-Koordinaten (x, y, z). Die Vergleichsergebnisse sind in Tabelle 3 dargestellt. Wir haben festgestellt, dass ohne zusätzliche dreidimensionale Merkmale Es kam zu einer erheblichen Leistungsverschlechterung von SFCNet um mehr als 10 %.

**Tabelle 3.** Wirksamkeit des Intervall-Frequenz-Deskriptors auf NTU RGB+D 60.

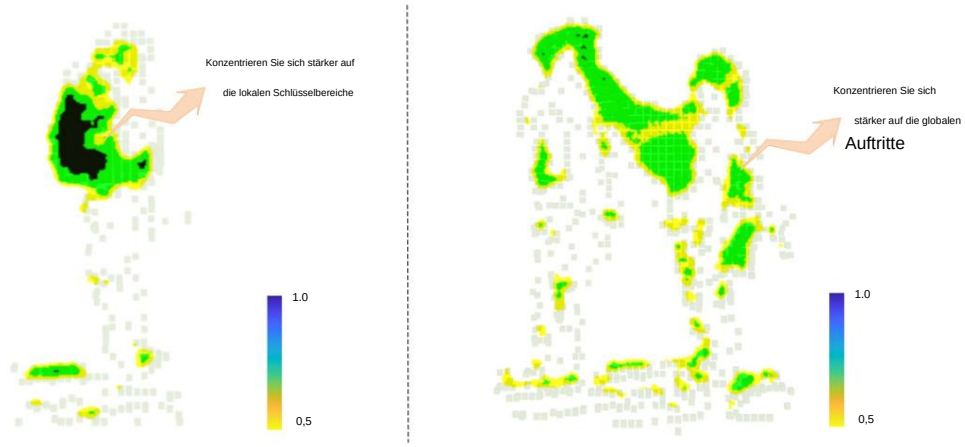
| Punkt-Feature | Themenübergreifend | Queransicht |
|---|---|---|
| (x, y, z) | 78,0 % | 82,3 % |
| (x, y, z, o, f, l) | **89,9 %** | **96,7 %** |

Dies deutet darauf hin, dass der Intervall-Frequenz-Deskriptor effektiv die dynamische Funktionen innerhalb des gesamten Aktionsprozesses, die eine entscheidende Rolle bei der 3D-Aktionserkennung spielen.

**Wirksamkeit der Zweistrom-Merkmalsfusion.** Unterschiedliche menschliche Handlungen enthalten unterschiedliche globale und lokale Dynamik, beschreiben wir das Gesamtbewegungsmuster des gesamten Menschen Körper während der Aktion durch den globalen groben Strom. Im Gegensatz dazu der lokale feine Strom beschreibt die Dynamik wichtiger Körperteile, wobei den Details mehr Aufmerksamkeit gewidmet wird und lokale Besonderheiten von Aktionen und hilft dabei, subtile Änderungen und komplexe Aktionsmuster zu erfassen. Die Ergebnisse in Tabelle 4 zeigen, dass unser vorgeschlagenes SFCNet, das die feinen und groben Merkmale des Zwei Strömungen können menschliche Handlungen umfassender verstehen und erkennen. Erstens diskutieren wir die Erkennungsleistung im Single-Stream-Zustand. Es ist ersichtlich, dass die lokale Stream hat eine höhere Fähigkeit, die Aktion darzustellen als der globale Stream, was darauf hinweist, dass er ist es wichtig, auf die Hauptteile zu achten, um Redundanz zu vermeiden. Darüber hinaus Wir vergleichen drei verschiedene Strategien zur Merkmalsfusion, um die Überlegenheit von aufmerksamkeitsbasierte Merkmalsfusion, die in SFCNet vorgeschlagen wird (siehe Gleichung (6)). Wie in Tabelle 4. SFCNet (Fusion) hat offensichtliche Vorteile gegenüber der nativen Kaskaden- oder additiven Fusion Strategien. Der Hauptgrund ist, dass die aufmerksamkeitsbasierte Merkmalsfusion adaptiv zuordnen kann die Aufmerksamkeit des Modells auf die Merkmale des globalen Grobstroms und des lokalen Feinstroms. Wie in Abbildung 5 dargestellt, erfolgt das Trinken von Wasser nur durch die Interaktion von Hand und Kopf. Die Bewegungsamplitude ist klein, daher betont das Modell die Eigenschaften der lokalen Stream, um feinkörnige Bewegungsmuster zu erfassen. Beim Schlagen hingegen handelt es sich um die Interaktion zweier Menschen, und die Schlagbewegung ist groß und kraftvoll, so mehr Aufmerksamkeit wird dem globalen Erscheinungsbild des Körpers gewidmet, während einige Details betont werden von Hände und Kopf. Dieser aktionsspezifische Merkmalsextraktionsmechanismus verbessert die Generalisierungsfähigkeit von SFCNet und Genauigkeit für die 3D-Aktionserkennung.

**Tabelle 4.** Wirksamkeit der Zwei-Stream-Feature-Fusion im NTU RGB+D 60-Datensatz.

| Eingabestrom | Themenübergreifend | Queransicht |
|---|---|---|
| 1s-SFCNet (L) 1s-SFCNet (G) | 85,0 %<br>81,8 % | 94,6 %<br>86,6 % |
| SFCNet (verkettet) | 88,9 % | 94,8 % |
| SFCNet (hinzufügen) | 86,7 % | 93,9 % |
| SFCNet (Fusion) | 89,9 % | 96,7 % |




**a) Trinkwasser** **b) Stanzen**

**Abbildung 5.** Visualisierung der Feature-Aufmerksamkeit. Wir visualisieren die Heatmap von Trinkwasser und Stanzen.

### 4.5. Vergleich mit bestehenden Methoden

Um die Leistung des vorgeschlagenen SFCNet zu bewerten, vergleichen wir es mit bestehenden Methoden auf zwei großen Benchmark-Datensätzen, wie in den Tabellen 5 und 6 dargestellt. Wir teilen bestehende 3D-Aktion Erkennungsmethoden in skelettbasierte und tiefenbasierte Methoden. Bei den skelettbasierten Methoden vergleichen wir verschiedene Backbone-basierte Methoden, darunter CNN [5], LSTM [15,46] und GCN [6,23,47]. Für tiefenbasierte Methoden vergleichen wir 2D-bildbasierte Methoden [19,35,48], 3D CNN-basierte Methoden [28] und 3D Voxel-basierte Methoden [10]. Die Vergleichsergebnisse sind in den Tabellen 5 und 6 angegeben . Für den NTU RGB+D 60-Datensatz erreicht das vorgeschlagene SFCNet 89,9 % und 96,7 % Genauigkeit bei Cross-Subject- und Cross-View-Einstellungen. Darüber hinaus haben wir Vergleichen Sie SFCNet mit zwei Methoden, die auf multimodalen Daten basieren. Im Vergleich zu ED-MHI [31] welche Tiefen- und Skelettdaten kombiniert, verbessert unsere Methode die Genauigkeit um 4,3% in der TS-CNN-LSTM [49] fusionierte Daten aus drei Modalitäten, nämlich RGB, Tiefe, und Skelett, aber es ist 2,6 % und 4,9 % niedriger als SFCNet in den Cross-Subject-Einstellungen bzw. Cross- View-Einstellungen. Für den NTU RGB+D 120-Datensatz erreicht SFCNet auch wettbewerbsfähige Ergebnisse, die Genauigkeiten von 83,6 % und 93,8 % unter Cross-Subject- und Cross-View-Einstellungen erreichten, Im Allgemeinen ist SFCNet effektiv und ausgezeichnet und übertrifft traditionelle Methoden mit manueller Merkmalsextraktion [19,35,50] und Deep-Learning-Methoden, die die Tiefe komprimieren Video in Bilder zur Weiterverarbeitung [2,3,36] oder Punktwolkensequenzen [10]. Die experimentellen Ergebnisse beweisen, dass das SFCNet für die Erfassung diskriminierender menschlicher Verhaltensmuster überlegen ist und Dies ist für die 3D-Aktionserkennung von Vorteil.

**Tabelle 5.** Vergleich verschiedener Methoden zur Genauigkeit der Aktionserkennung (%) auf dem NTU RGB+D 60 Datensätze.

| Methode | Queransicht | Jahr |
|---|---|---|
| **Themenübergreifender Input: 3D Skelett** | | |
| GCA-LSTM [15] | 74,4 | 82,8 | 2017 |
| Zwei-Stream-Aufmerksamkeit LSTM [46] | 77,1 | 85.1 | 2018 |
| ST-GCN [23] | 81,5 | 88,3 | 2018 |
| Skelettbewegung [5] | 69,6 | 80.1 | 2019 |
| [ 0042 ] 2s- | 86,8 | 94,2 | 2019 |
| AGCN [0043] | 88,5 | 95,1 | 2019 |
| ST-TR (neu) [24] | 89,9 | 96,1 | 2021 |
| DSwarm-Net (neu) [51] | 85,5 | 90,0 | 2022 |
| Aktionsnetz [30] | 73,2 | 76.1 | 2023 |
| SGMSN (neu) [52] | 90,1 | 95,8 | 2023 |
| **Eingabe: Tiefenkarten** | | |
| 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-Punktnetz++ [10] | 88,8 | 96,3 | 2020 |
| DOGV (neu) [53] | 90,6 | 94,7 | 2021 |
| 3DFCNN [28] | 78.1 | 80,4 | 2022 |
| 3D-Beschneiden [54] | 83,6 | 92,4 | 2022 |
| ConvLSTM (neu) [29] | 80,4 | 79,9 | 2022 |
| CBBMC (neu) [48] | 83,3 | 87,7 | 2023 |
| PointMapNet (neu) [55] | 89,4 | 96,7 | 2023 |
| SFCNet (unseres) | 89,9 | 96,7 | - |
| **Input: Multimodalitäten** | | |
| ED-MHI [31] | 85,6 | - | 2022 |
| TS-CNN-LSTM [49] | 87,3 | 91,8 | 2023 |

**Tabelle 6.** Vergleich verschiedener Methoden zur Genauigkeit der Aktionserkennung (%) auf dem NTU RGB+D 120 Datensätze.

| Methode | Themenübergreifend | Kreuz-Set | Jahr |
|---|---|---|---|
| **Eingabe: 3D-Skelett** | | |
| GCA-LSTM [15] | 58,3 | 59,3 | 2017 |
| Karte der Körperhaltungsentwicklung [56] | 64,6 | 66,9 | 2018 |
| Zwei-Stream-Aufmerksamkeit LSTM [46] | 61,2 | 63,3 | 2018 |
| ST-GCN [23] | 70,7 | 73,2 | 2018 |
| NTU RGB+D 120 Basislinie [45] | 55,7 | 57,9 | 2019 |
| FSNet [57] | 59,9 | 62,4 | 2019 |
| Skelettbewegung [5] | 67,7 | 66,9 | 2019 |
| TSRJI [20] | 67,9 | 62,8 | 2019 |
| [ 0042 ] 2s- | 77,9 | 78,5 | 2019 |
| AGCN [0043] | 82.9 | 84,9 | 2019 |
| ST-TR (neu) [24] | 82.7 | 84,7 | 2021 |
| SGMSN (neu) [52] | 84.8 | 85,9 | 2023 |
| **Eingabe: Tiefenkarten** | | |
| APSR [45] | 48,7 | 40,1 | 2019 |
| 3DV-Punktnetz++ [10] | 82,4 | 93,5 | 2020 |
| DOGV (neu) [53] | 82,2 | 85,0 | 2021 |
| 3D-Beschneiden [54] | 76,6 | 88,8 | 2022 |
| SFCNet (unseres) | 83,6 | 93,8 | - |

**5. Diskussion**

Um die Vor- und Nachteile der vorgeschlagenen Methode zu analysieren, haben wir die Erkennungsgenauigkeit von SFCNet im NTU RGB+60-Datensatz für die übergreifenden Einstellungen für jede Kategorie dargestellt. Die Ergebnisse werden in Form einer Verwirrungsmatrix in Abbildung 6 (links) angezeigt. Wir haben einige verwirrende Aktionen für eine klarere Darstellung ausgewählt und sie lokal vergrößert, wie in Abbildung 6 (rechts) gezeigt. Die Ergebnisse zeigen, dass SFCNet eine robuste Fähigkeit zur Analyse menschlicher Aktionen mit einer Erkennungsgenauigkeit von über 90 % in den meisten Kategorien hat. Beispielsweise hat es eine Genauigkeit von 100 % beim Hüpfen und 99 % beim Hochspringen erreicht . SFCNet ist jedoch bei der Erkennung einiger ähnlicher Aktionen verwirrt. Beispielsweise sind Lesen und Schreiben sowie das Tragen und Ausziehen von Schuhen die verwirrendsten Beispielpaare. Darüber hinaus wurden 25 % derjenigen, die mit ihren Telefonen spielten, fälschlicherweise als Lesen (9 %), Schreiben (8 %) und Tippen auf der Tastatur (8 %) klassifiziert. Die Genauigkeit beim Tippen auf der Tastatur beträgt nur 66 %, und 12 % der Beispiele werden fälschlicherweise als Schreiben klassifiziert. Aus der Analyse haben wir herausgefunden, dass diese Aktionen nur geringfügige Unterschiede aufweisen und die Bewegungsamplitude gering ist. Dies ist der Hauptgrund

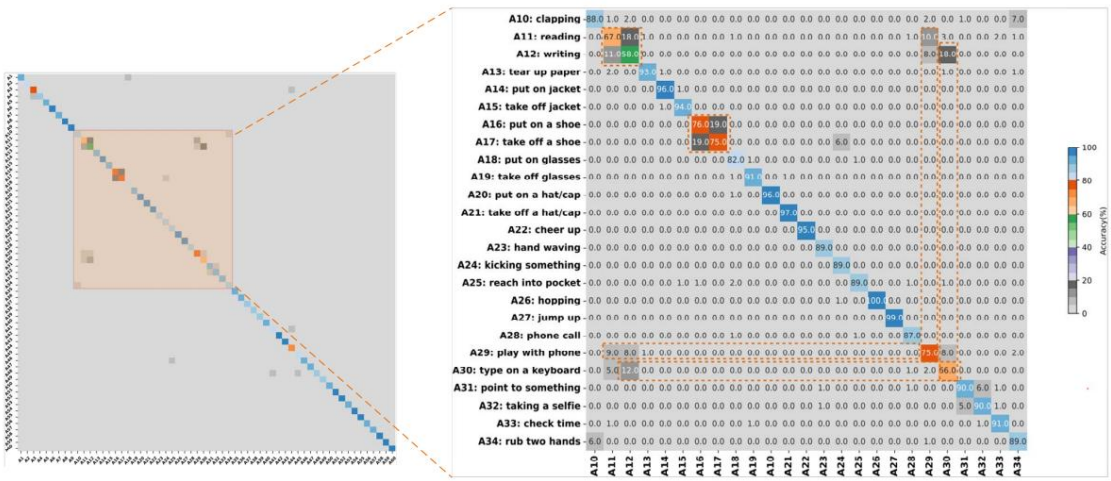



**Abbildung 6.** Konfusionsmatrix für klassenspezifische Erkennungsgenauigkeit. Das Bild **(links)** enthält alle Aktionskategorien. Um einige der Verschleierungsaktionen hervorzuheben, wird (rechts) eine lokale Vergrößerung angezeigt .

**6. Schlussfolgerungen**

In diesem Artikel schlagen wir ein symmetrisches neuronales Netzwerk, SFCNet, vor, um 3D-Aktionen aus Punktwolkensequenzen zu erkennen. Es enthält einen globalen groben Stream und einen lokalen feinen Stream, der PointNet++ als Merkmalsextraktor verwendet. Die Punktwolkensequenzen werden als strukturierte Voxelsätze reguliert, an die der vorgeschlagene Intervall-Frequenz-Deskriptor angehängt wird, um 6D-Merkmale zu generieren, die räumlich-zeitliche dynamische Informationen erfassen. Der globale grobe Stream erfasst die grobkörnigen Aktionsmuster anhand des Aussehens des menschlichen Körpers, und der lokale feine Stream extrahiert aktionsspezifische feinkörnige Merkmale aus kritischen Teilen. Nach der Merkmalsfusion kann SFCNet diskriminierende Bewegungsmuster ermitteln, die allgemeine räumliche Änderungen beinhalten und entscheidende Details durchgängig hervorheben. Laut den experimentellen Ergebnissen an zwei großen Benchmark-Datensätzen, NTU RGB+D 60 und NTU RGB+D 120, ist SFCNet für die 3D-Aktionserkennung effektiv und hat das Potenzial für Ferner

Das vorgeschlagene SFCNet weist jedoch noch Einschränkungen bei der Unterscheidung ähnlicher Aktionen auf. Unsere zukünftige Arbeit wird sich auf die Erkennung ähnlicher Aktionen und die Erfassung subtiler Muster konzentrieren, um die Genauigkeit zu verbessern.

**Beiträge der Autoren:** Konzeptualisierung, CL und QH; Methodik, CL, WQ und XL; Software, QH und YM; Validierung, CL, WQ und XL; formale Analyse, QH und YM; Untersuchung, CL und WQ; Ressourcen, QH und YM; Datenkuratierung, WQ; Schreiben – Vorbereitung des Originalentwurfs, CL und WQ; Schreiben – Überprüfung und Bearbeitung, CL, YM, WQ, QH und XL; Visualisierung, WQ; Überwachung,

QH und YM; Projektverwaltung, QH und YM; Mittelbeschaffung, QH, YM und CL. Alle Autoren haben die veröffentlichte Version des Manuskripts gelesen und stimmen ihr zu.

**Finanzierung:** Diese Forschung wurde finanziert durch das Postgraduate Research & Practice Innovation Program der Provinz Jiangsu (Zuschussnummer KYCX23_0753), den Fundamental Research Funds for the Central Universities (Zuschussnummer B230205027), das Key Research and Development Program of China (Zuschussnummer 2022YFC3005401), das Key Research and Development Program of China, Provinz Yunnan (Zuschussnummer 202203AA080009), den 14. Fünfjahresplan für Erziehungswissenschaften der Provinz Jiangsu (Zuschussnummer D/2021/01/39) und das Jiangsu Higher Education Reform Research Project (Zuschussnummer 2021JSJG143); und das APC wurde finanziert durch den Fundamental Research Funds for the Central Universities.

**Erklärung des Institutional Review Board:** Nicht zutreffend.

**Einverständniserklärung:** Nicht zutreffend.

**Datenverfügbarkeitserklärung:** Die in diesem Dokument verwendeten Datensätze NTU RGB+D 60 und NTU RGB+D 120 sind öffentlich, kostenlos und verfügbar unter: https://rose1.ntu.edu.sg/dataset/actionRecognition/ (abgerufen am 21. Dezember 2020).

**Interessenkonflikte:** Die Autoren erklären, dass keine Interessenkonflikte bestehen.

## Verweise

1. Riaz, W.; Gao, C.; Azeem, A.; Saifullah; Bux, JA; Ullah, A. Verkehrsanomalie-Vorhersagesystem mithilfe eines prädiktiven Netzwerks. Remote Sens. **2022,** 14, 1–19.
2. Wang, P.; Li, W.; Gao, Z.; Tang, C.; Ogunbona, PO Tiefenpooling-basierte groß angelegte 3D-Aktionserkennung mit Convolutional Neural Networks. IEEE Trans. Multimed. **2018,** 20, 1051–1061.
3. Xiao, Y.; Chen, J.; Wang, Y.; Cao, Z.; Zhou, JT; Bai, X. Aktionserkennung für Tiefenvideo mit dynamischen Multi-View-Bildern. Inf. 2019 , 480, 287–304.
4. Li, C.; Huang, Q.; Li, X.; Wu, Q. Erkennung menschlicher Handlungen basierend auf mehrskaligen Merkmalskarten aus Tiefenvideosequenzen. Tools Appl. **2021,** 80, 32111–32130.
5. Caetano, C.; Sena, J.; Brémond, F.; Dos Santos, JA; Schwartz, WR Skelemotion: Eine neue Darstellung von Skelettgelenksequenzen basierend auf Bewegungsinformationen zur 3D-Aktionserkennung. In Proceedings der IEEE International Conference on Advanced Video and Signal Based Surveillance, Taipeh, Taiwan, 18.–21. September 2019; IEEE: Piscataway, NJ, USA, 2019; S. 1–8.
6. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Aktional-strukturelle Graph-Faltungsnetzwerke für skelettbasierte Aktionserkennung. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Long Beach, CA, USA, 15.–20. Juni 2019; S. 3595–3603.
7. Yang, X.; Zhang, C.; Tian, Y. Erkennen von Aktionen mithilfe von auf Tiefenbewegungskarten basierenden Histogrammen orientierter Gradienten. In Proceedings of the ACM International Conference on Multimedia, Nara, Japan, 29. Oktober–2. November 2012; S. 1057–1060.
8. Elmadany, NED; He, Y.; Guan, L. Informationsfusion zur Erkennung menschlicher Handlungen über Biset/Multiset-Globalitätslokalität Erhaltung der kanonischen Korrelationsanalyse. IEEE Trans. Image Process. **2018,** 27, 5275–5287.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning für die Bilderkennung. In Proceedings der IEEE-Konferenz zu Computer Vision und Mustererkennung, Las Vegas, NV, USA, 27.–30. Juni 2016; S. 770–778.
10. Wang, Y.; Xiao, Y.; Xiong, F.; Jiang, W.; Cao, Z.; Zhou, JT; Yuan, J. 3DV: 3D-Dynamisches Voxel zur Aktionserkennung in Tiefenvideo. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Seattle, WA, USA, 14.–19. Juni 2020; S. 508–517.
11. Qi, CR; Yi, L.; Su, H.; Guibas, LJ Pointnet++: Tiefes hierarchisches Merkmalslernen auf Punktmengen in einem metrischen Raum. Adv. Neural Inf. Prozess. Syst. **2017,** 30, 5105–5114.
12. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, PO Aktionserkennung aus Tiefenkarten mittels Deep Convolutional neuronale Netzwerke. IEEE Trans. Hum. -Mach. Syst. **2015,** 46, 498–509.
13. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Eine neue Darstellung von Skelettsequenzen für die 3D-Aktionserkennung. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Honolulu, HI, USA, 21.–26. Juli 2017; S. 3288–3297.
14. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Co-Occurrence-Feature-Learning aus Skelettdaten zur Aktionserkennung und -erfassung mit hierarchischer Aggregation. arXiv **2018,** arXiv:1804.06055v1 .
15. Liu, J.; Gang, W.; Ping, H.; Duan, LY; Kot, AC Globale kontextbewusste Aufmerksamkeits-LSTM-Netzwerke für 3D-Aktionserkennung. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Honolulu, HI, USA, 21.–26. Juli 2017; S. 3671–3680.
16. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skelettbasierte Aktionserkennung mit gerichteten Graph-Neuralnetzen. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Long Beach, CA, USA, 15.–20. Juni 2019; S. 7912–7921.

17. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. Ein aufmerksamkeitsverstärktes Graph-Convolutional-LSTM-Netzwerk für skelettbasierte Aktionserkennung . In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Long Beach, CA, USA, 15.–20. Juni 2019; S. 1227–1236.

18. Yu, Z.; Wenbin, C.; Guodong, G. Auswertung räumlich-zeitlicher Merkmale interessanter Punkte für tiefenbasierte Aktionserkennung. Bildvisualisierung. 2014 , 32, 453–464.
19. Oreifej, O.; Liu, Z. Hon4d: Histogramm orientierter 4D-Normalen zur Aktivitätserkennung aus Tiefensequenzen. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Portland, OR, USA, 23.–28. Juni 2013; S. 716–723.
20. Caetano, C.; Brémond, F.; Schwartz, WR Skelettbilddarstellung für 3D-Aktionserkennung basierend auf Baumstruktur und Referenzverbindungen. In Proceedings der 32. SIBGRAPI-Konferenz zu Grafiken, Mustern und Bildern, Rio de Janeiro, Brasilien, 28.–31. Oktober 2019; S. 16–23.

21. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Räumlich-zeitliches LSTM mit Vertrauensgates zur 3D-Erkennung menschlicher Aktionen. In Proceedings der European Conference on Computer Vision, Amsterdam, Niederlande, 11.–14. Oktober 2022; Springer: Berlin/Heidelberg, Deutschland, 2016; S. 816–833.

22. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. Ansicht adaptiver neuronaler Netzwerke für hochleistungsfähige skelettbasierte Erkennung menschlicher Handlungen. IEEE Trans. Pattern Anal. Mach. Intell. **2019,** 41, 1963–1978.
23. Yan, S.; Xiong, Y.; Lin, D. Räumlich-zeitliche Graph-Faltungsnetzwerke für skelettbasierte Aktionserkennung. In Proceedings der 32. AAAI-Konferenz über künstliche Intelligenz, New Orleans, LA, USA, 2.–7. Februar 2018; S. 7444–7452.
24. Plizzari, C.; Cannici, M.; Matteucci, M. Skelettbasierte Aktionserkennung über räumliche und zeitliche Transformatornetzwerke. Comput. Vis. Bildverstand. **2021,** 208-209, 103219.
25. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Echtzeiterkennung menschlicher Posen in Teilen aus einzelnen Tiefenbildern. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Colorado Springs, CO, USA, 20.–25. Juni 2011; IEEE: Piscataway, NJ, USA, 2011; S. 1297–1304.
26. Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, JT; Yuan, J. A2j: Anchor-to-Joint-Regressionsnetzwerk zur dreidimensionalen artikulierten Posenschätzung aus einem einzelnen Tiefenbild. In Proceedings der IEEE International Conference on Computer Vision, Seoul, Republik Korea, 27. Oktober–2. November 2019; S. 793–802.
27. Kamel, A.; Sheng, B.; Yang, P.; Li, P.; Shen, R.; Feng, DD Tiefe Convolutional Neural Networks zur Erkennung menschlicher Handlungen Verwenden von Tiefenkarten und Körperhaltungen. IEEE Trans. Syst. Man Cybern. Syst. **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: Aktionserkennung in Echtzeit unter Verwendung tiefer neuronaler 3D-Netze mit rohen Tiefeninformationen. Multimed. Werkzeuge Appl. **2022,** 81, 24119–24143.
29. Sánchez-Caballero, A.; Fuentes-Jiménez, D.; Losada-Gutiérrez, C. Echtzeiterkennung menschlicher Handlungen mithilfe von Rohtiefenvideos. basierte rekurrierende neuronale Netzwerke. Multimed. Tools Appl. **2022,** 82, 16213–16235.
30. Kumar, DA; Kishore, PVV; Murthy, G.; Chaitanya, TR; Subhani, S. Ansichtsinvariante Erkennung menschlicher Aktionen mithilfe von Oberflächenkarten über Faltungsnetzwerke. In Proceedings der International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering, Chennai, Indien, 1.–2. November 2023; S. 1–5.
31. Ghosh, SK; M, R.; Mohan, BR; Guddeti, RMR Deep Learning-basierte Multi-View-3D-Erkennung menschlicher Aktionen unter Verwendung von Skelett- und Tiefendaten. Multimed. Tools Appl. **2022,** 82, 19829–19851.
32. Li, R.; Li, X.; Fu, CW; Cohen-Or, D.; Heng, PA Pu-gan: ein Point-Cloud-Upsampling-Kontrastnetzwerk. In Proceedings der IEEE/CVF International Conference on Computer Vision, Seoul, Republik Korea, 27. Oktober–2. November 2019; S. 7203–7212.
33. Qi, CR; Litany, O.; He, K.; Guibas, LJ Deep Hough Voting für 3D-Objekterkennung in Punktwolken. In Proceedings der IEEE/CVF International Conference on Computer Vision, Seoul, Republik Korea, 27. Oktober–2. November 2019; S. 9277–9286.
34. Thomas, H.; Qi, CR; Deschaud, JE; Marcotegui, B.; Goulette, F.; Guibas, L. KPConv: Flexible und verformbare Faltung für Punktwolken. In Proceedings der IEEE/CVF International Conference on Computer Vision, Seoul, Republik Korea, 27. Oktober–2. November 2019; S. 6410–6419.

35. Ohn-Bar, E.; Trivedi, MM Gemeinsame Winkelähnlichkeiten und HOG2 für Aktionserkennung. In Proceedings der IEEE/CVF- Konferenz über Computer Vision und Mustererkennungs-Workshops, Portland, OR, USA, 23.–28. Juni 2013; S. 465–470.
36. Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, MS Unüberwachtes Lernen von ansichtsinvarianten Aktionsdarstellungen. Adv. Neural Inf. Prozess. Syst. **2018,** 31, 1262–1272.
37. Liu, X.; Qi, CR; Guibas, LJ Flownet3d: Lernen von Szenenfluss in 3D-Punktwolken. In Proceedings der IEEE/CVF-Konferenz über Computer Vision und Mustererkennung, Long Beach, CA, USA, 15.–20. Juni 2019; S. 529–537.
38. Zhai, M.; Xiang, X.; Lv, N.; Kong, X. Optischer Fluss und Szenenflussschätzung: Eine Untersuchung. Mustererkennung. **2021,** 114, 107861.
39. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank Pooling zur Aktionserkennung. IEEE Trans. Pattern Anal. Mach. Intel. **2016,** 39, 773–787.
40. Liu, J.; Xu, D. GeometryMotion-Net: Eine starke Zwei-Stream-Baseline für die 3D-Aktionserkennung. IEEE Trans. Circuits Syst. Video Technol. **2021,** 31, 4711–4721.
41. Dou, W.; Chin, WH; Kubota, N. Wachsendes Speichernetzwerk mit zufälligem Gewicht 3DCNN für die kontinuierliche Erkennung menschlicher Handlungen. In Proceedings der IEEE International Conference on Fuzzy Systems, Incheon, Republik Korea, 13.–17. August 2023; S. 1–6.

42. Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; Kankanhalli, M. PSTNet: Punkt-räumlich-zeitliche Faltung auf Punktwolkensequenzen. In Proceedings der International Conference on Learning Representations, Addis Abeba, Äthiopien, 26.–30. April 2020; S. 1–6.
43. Qi, CR; Su, H.; Mo, K.; Guibas, LJ Pointnet: Deep Learning auf Punktmengen für 3D-Klassifizierung und -Segmentierung. In Proceedings der IEEE-Konferenz zu Computer Vision und Mustererkennung, Honolulu, HI, USA, 21.–26. Juli 2017; S. 652–660.
44. Shahroudy, A.; Liu, J.; Ng, TT; Wang, G. NTU RGB+D: Ein groß angelegter Datensatz für die 3D-Analyse menschlicher Aktivitäten. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Las Vegas, NV, USA, 27.–30. Juni 2016; S. 1010–1019.
45. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, LY; Kot, AC Ntu rgb+ d 120: Ein groß angelegter Benchmark für dreidimensionale menschliche Aktivität Verständnis. IEEE Trans. Pattern Anal. Mach. Intell. **2019,** 42, 2684–2701.
46. Liu, J.; Wang, G.; Duan, LY; Abdiyeva, K.; Kot, AC Skelettbasierte Erkennung menschlicher Aktionen mit globaler Kontextwahrnehmung Achtung LSTM-Netzwerke. IEEE Trans. Image Process. **2018,** 27, 1586–1599.
47. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Zwei-Stream-adaptive Graph-Faltungsnetzwerke für skelettbasierte Aktionserkennung. In Proceedings der IEEE/CVF-Konferenz zu Computer Vision und Mustererkennung, Long Beach, CA, USA, 15.–20. Juni 2019; S. 12026–12035.

48. Li, X.; Huang, Q.; Wang, Z. Räumliche und zeitliche Informationsfusion zur Erkennung menschlicher Handlungen mittels Center Boundary Balancing Multimodaler Klassifikator. J. Vis. Commun. Bilddarstellung. **2023,** 90, 103716.
49. Zan, H.; Zhao, G. Forschung zur Erkennung menschlicher Handlungen basierend auf Fusion TS-CNN und LSTM-Netzwerken. Arab. J. Sci. Eng. **2023,** 48, 2331–2345.
50. Yang, X.; Tian, Y. Supernormaler Vektor zur Aktivitätserkennung mithilfe von Tiefensequenzen. In Proceedings der IEEE/CVF-Konferenz über Computer Vision und Mustererkennung, Columbus, OH, USA, 23.–28. Juni 2014; S. 804–811.
51. Basak, H.; Kundu, R.; Singh, PK; Ijaz, MF; Wo´zniak, M.; Sarkar, R. Eine Kombination aus Deep Learning und schwarmbasierter Optimierung zur 3D-Erkennung menschlicher Handlungen. Sci. Rep. **2022,** 12, 1–17.
52. Qi, Y.; Hu, J.; Zhuang, L.; Pei, X. Semantisch gesteuerte mehrskalige Aktionserkennung des menschlichen Skeletts. Appl. Intell. Int. J. Artif. Intell. Neuronale Netze. Komplexe Problemlösungstechnologie. **2023,** 53, 9763–9778.
53. Ji, X.; Zhao, Q.; Cheng, J.; Ma, C. Nutzung der räumlich-zeitlichen Darstellung zur 3D-Erkennung menschlicher Aktionen anhand einer Tiefenkarte Sequenzen. Knowl. -Based Syst. **2021,** 227, 107040.
54. Guo, J.; Liu, J.; Xu, D. 3D-Pruning: Ein Modellkomprimierungsframework für effiziente 3D-Aktionserkennung. IEEE Trans. Circuits Syst. Video Technol. **2022,** 32, 8717–8729.
55. Li, X.; Huang, Q.; Zhang, Y.; Yang, T.; Wang, Z. PointMapNet: Point Cloud Feature Map Network zur 3D- Erkennung menschlicher Aktionen. Symmetry **2023,** 15, 1–17.
56. Liu, M.; Yuan, J. Erkennen menschlicher Handlungen als Entwicklung von Pose-Estimation-Karten. In Proceedings der IEEE/CVF- Konferenz zu Computer Vision und Mustererkennung, Salt Lake City, UT, USA, 18.–23. Juni 2018; S. 1159–1168.
57. Liu, J.; Shahroudy, A.; Wang, G.; Duan, LY; Kot, AC Skelettbasierte Online-Aktionsvorhersage unter Verwendung eines Skalenauswahlnetzwerks. IEEE Trans. Pattern Anal. Mach. Intell. **2019,** 42, 1453–1467.