Ciencias Aplicadas

MDPI

Artículo

Una nueva red neuronal simétrica fina-gruesa para humanos 3D

Reconocimiento de acciones basado en secuencias de nubes de puntos

Chang Li [1], Qian Huang [1,*], Yingchi Mao [1], weiwen qian [1] y Xing Li [2]

[1] Facultad de Ciencias de la Computación e Ingeniería de Software, Universidad Hohai, Nanjing 211100, China; lichang@hhu.edu.cn (CL); yingchimao@hhu.edu.cn (YM); qianweiwen@hhu.edu.cn (WQ)

[2] Facultad de Ciencias y Tecnología de la Información y Facultad de Inteligencia Artificial, Nanjing Forestry Universidad, Nanjing 210037, China; lixing@njfu.edu.cn

* Correspondencia: huangqian@hhu.edu.cn

Resumen: El reconocimiento de la acción humana ha facilitado el desarrollo de dispositivos de inteligencia artificial centrados en actividades y servicios humanos. Esta tecnología ha progresado introduciendo Nubes de puntos 3D derivadas de cámaras de profundidad o radares. Sin embargo, el comportamiento humano es complejo y Las nubes de puntos involucradas son vastas, desordenadas y complicadas, lo que plantea desafíos para la acción 3D. reconocimiento. Para resolver estos problemas, proponemos una red neuronal simétrica fina-gruesa (SFCNet) que analiza simultáneamente la apariencia y los detalles de las acciones humanas. En primer lugar, las secuencias de nubes de puntos se transforman y voxelizan en conjuntos de vóxeles 3D estructurados. Estos conjuntos luego se aumentan con un descriptor de intervalo-frecuencia para generar características 6D que capturan la dinámica espaciotemporal información. Al evaluar la ocupación del espacio de vóxeles mediante umbrales, podemos identificar eficazmente el partes esenciales. Después de eso, todos los vóxeles con la característica 6D se dirigen al flujo grueso global, mientras que los vóxeles dentro de las partes clave se enrutan al flujo fino local. Estas dos corrientes extraen características de apariencia global y partes críticas del cuerpo mediante el uso de PointNet++ simétrico. Después, La fusión de características de atención se emplea para capturar patrones de movimiento más discriminativos de forma adaptativa. Se validan los experimentos realizados con los conjuntos de datos de referencia públicos NTU RGB+D 60 y NTU RGB+D 120 La eficacia y superioridad de SFCNet para el reconocimiento de acciones 3D.

Palabras clave: análisis de nubes de puntos; reconocimiento de acciones 3D; reconocimiento de patrones; aprendizaje profundo

Cita: Li, C.; Huang, Q.; Mao, Y.; Qian, W.; Li, X. Una novela simétrica Red neuronal fina y gruesa para 3D Reconocimiento de la acción humana basado en Secuencias de nubes de puntos. Aplica. Ciencia. 2024, 14, 6335. https://doi.org/10.3390/aplicación14146335

Editor académico: Atsushi Mase

Recibido: 11 de junio de 2024
Revisado: 8 de julio de 2024
Aceptado: 18 de julio de 2024
Publicado: 20 de julio de 2024

1. Introducción

El reconocimiento de la acción humana tiene como objetivo ayudar a las computadoras a comprender la semántica del comportamiento humano a partir de diversos datos registrados por los dispositivos de adquisición. En particular, acción 3D. El reconocimiento está dedicado a extraer patrones de acción a partir de datos 3D que involucran movimientos humanos . Ha atraído cada vez más atención debido a sus aplicaciones generalizadas, como Monitoreo de seguridad pública, evaluación del desempeño, reconocimiento militar e inteligencia. transporte [1].

Los métodos actuales de reconocimiento de acciones 3D convencionales se pueden clasificar en métodos basados en profundidad (incluidos mapas de profundidad y secuencias de nubes de puntos) [2–4] y métodos basados en esqueletos [5,6], según el tipo de datos empleado. Limitado por la precisión Algoritmo de estimación de pose: la inevitable tarea anterior: métodos basados en esqueletos. enfrentan desafíos de robustez y consumo computacional. Por el contrario, basado en profundidad Los métodos son más independientes de la tarea y han atraído una atención generalizada. Existente Los enfoques de reconocimiento de acciones 3D basados en profundidad se dividen principalmente en dos categorías principales. El El primero es codificar movimientos 3D en una o más imágenes [2,3,7,8] y utilizar CNN [9]. para el reconocimiento de la acción. Sin embargo, el plano de la imagen 2D no puede caracterizar completamente el 3D. dinámica porque las acciones humanas son simultáneamente espaciotemporales y se llevan a cabo en el Espacio 3D. La otra es transformar el vídeo de profundidad en una secuencia de nube de puntos [10], que registra las coordenadas 3D de puntos en el espacio en múltiples instancias de tiempo. Así, comparado Con imágenes, las secuencias de nubes de puntos tienen la ventaja de conservar la apariencia 3D y

dinámica de la geometría a lo largo del tiempo, lo que permite un análisis avanzado y una comprensión de las acciones humanas. Además, las nubes de puntos se pueden obtener utilizando diversos dispositivos como escáneres láser, radares, sensores de profundidad y cámaras RGB+D, que se pueden montar en drones, farolas, vehículos y aviones de vigilancia, ampliando el alcance de aplicación del reconocimiento de acciones. . Sin embargo, debido a la estructura compleja y al enorme volumen de la nube de puntos, los métodos de reconocimiento de acciones 3D existentes basados en ella presentan los siguientes desafíos.

En primer lugar, las secuencias de nubes de puntos siempre tienen puntos masivos proporcionales a la dimensión temporal y el esquema de procesamiento de datos requiere mucho tiempo. Por lo tanto, desarrollar un modelo de secuencia de nube de puntos eficiente y liviano es fundamental para el reconocimiento de acciones en 3D. En segundo lugar, los puntos de las secuencias son irregulares y muestran información espacial intracuadro desordenada y detalles temporales intercuadro ordenados, lo que dificulta el análisis de los patrones de movimiento subyacentes. Sin embargo, los métodos de procesamiento de nubes de puntos existentes generalmente realizan una reducción de resolución indiferenciada de las nubes de puntos en general, lo que resulta en una pérdida uniforme de información esencial y sutil. Además, los esquemas de análisis de nubes de puntos existentes ignoran las partes críticas del cuerpo que contribuyen a las acciones, lo que resulta en una falta de matices de las características de acción extraídas, lo que finalmente limita el rendimiento del reconocimiento

Para resolver estos problemas, proponemos un marco de aprendizaje profundo llamado Red neuronal simétrica fina y gruesa (SFCNet) que combina simétricamente el análisis de características de movimiento desde perspectivas locales y globales, como se muestra en la Figura 1. En primer lugar, para ahorrar costos computacionales. , reducimos los puntos mediante muestreo de cuadros y muestreo de puntos más lejanos. A continuación, las nubes de puntos muestreadas se transforman en vóxeles 3D para crear una representación de nube de puntos compacta. Luego, las posiciones 3D originales se adjuntan con un descriptor de intervalo-frecuencia para representar la configuración espacial general y facilitar la identificación de partes esenciales del cuerpo, lo que nos permite dividir las secuencias de nubes de puntos en espacio fino local y espacio grueso global. Tratamos los vóxeles involucrados en estos dos espacios como puntos y empleamos PointNet++ [11] para extraer características de un extremo a otro. Finalmente, nuestro módulo de fusión de características combina la apariencia global y los detalles locales para obtener características discriminativas para el reconocimiento de acciones 3D. Los extensos experimentos en los conjuntos de datos a gran escala NTU RGB+D 60 y NTU RGB+D 120 demuestran la efectividad y preponderancia de SFCNet, mediante el cual la intención humana puede ser juzgada y asistida en aplicaciones de de

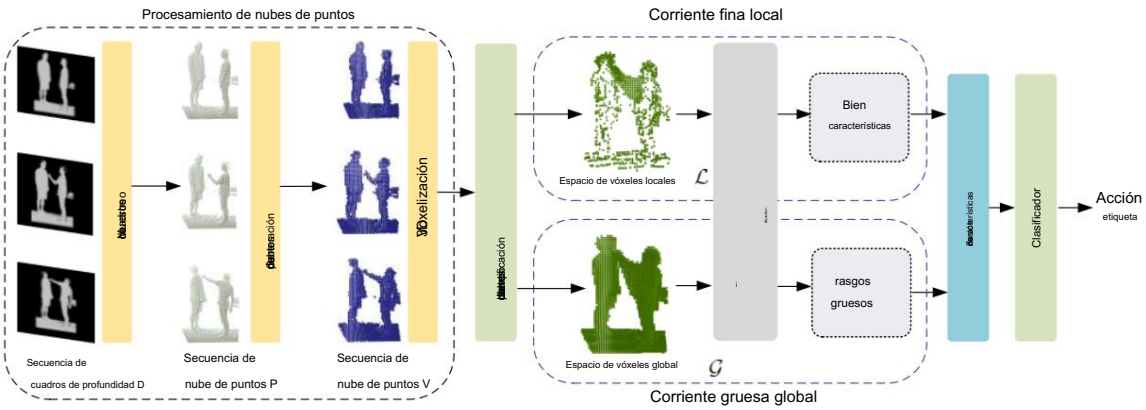


Figura 1. La canalización de SFCNet. Convierte fotogramas de profundidad en una nube de puntos y aplica operaciones de voxelización. Una estructura simétrica codifica vóxeles 3D, con partes cruciales e información dinámica global procesada por separado. El descriptor de frecuencia de intervalo adjunto caracteriza inicialmente la información de movimiento y luego es procesado por PointNet++ [11] para obtener características más profundas. Finalmente, el clasificador reconoce acciones 3D utilizando la característica agregada.

En general, las principales contribuciones de nuestro trabajo son las siguientes: • Proponemos un descriptor de intervalo-frecuencia para caracterizar los vóxeles 3D durante la ejecución de la acción, que preserva completamente los detalles del movimiento y proporciona pistas críticas para la percepción de partes clave del cuerpo. Hasta donde sabemos, nuestro trabajo es el primero en manejar secuencias de nubes de puntos de esta manera.

• Construimos un marco de aprendizaje profundo llamado SFCNet, que primero emplea una estructura simétrica para procesar secuencias de nubes de puntos. Codifica la dinámica local de las partes cruciales del cuerpo a través de una corriente fina y luego complementa estos detalles intrincados para la apariencia global capturada por una corriente gruesa. SFCNet puede enfatizar partes esenciales del cuerpo y capturar patrones de movimiento más discriminativos, abordando el problema de representación de acciones efectivas basado en nubes de puntos. • El SFCNet presentado ha demostrado su precisión superior en dos conjuntos de datos disponibles públicamente, NTU RGB+D 60 y NTU RGB+D 120, lo que demuestra que nuestro método tiene un potencial considerable para reconocer varios tipos de acciones, como acciones diarias, médicas- acciones relacionadas y acciones de interacción de dos personas.

2. Trabajos relacionados

2.1. Reconocimiento de acciones 3D basado en esqueletos

Los métodos de reconocimiento de acciones 3D existentes se pueden clasificar en métodos basados en esqueletos [12-17] y métodos basados en profundidad [3,7,12,18,19]. Generalmente existen cuatro enfoques principales para el reconocimiento de acciones basado en esqueletos. El primero es utilizar CNN [12] para aprender los patrones espacio-temporales de pseudoimágenes [13,14]. Caetano et al. [20] introdujo la imagen de articulaciones de referencia de estructura de árbol (TSRJI) para representar secuencias de esqueleto. El segundo es considerar las secuencias de esqueleto como series de tiempo [15-17] y utilizar columnas vertebrales como RNN [21] para la extracción de características. El tercero es ver los datos esqueléticos como gráficos [6,22] con las articulaciones como vértices y los huesos como aristas y recurrir a GCN [16] para la representación de la acción.

Por ejemplo, ST-GCN [23] representó eficazmente la información dinámica temporal de secuencias de esqueletos utilizando estrategias de partición y convolución de gráficos espacio-temporales. SkeleMotion [5] capturó información dinámica temporal calculando el tamaño y la orientación de las articulaciones del esqueleto en diferentes escalas de tiempo. El cuarto es codificar el esqueleto como tokens a través de Transformer. Plizzari et al. [24] emplearon atención espacio-temporal en Transformer y capturaron una relación dinámica entre cuadros de articulaciones. Sin embargo, dado que todavía existen desafíos importantes en la estimación precisa de la pose humana en 3D [25,26], los métodos de reconocimiento de acciones basados en esqueletos sufren de una cascada de rendimiento debido a esta inevitable tarea ascendente.

2.2. Reconocimiento de acciones 3D basado en profundidad

Para el reconocimiento de acciones 3D basado en profundidad, los primeros enfoques representan principalmente videos en profundidad mediante descriptores manuales [19]. Yang et al. [7] construyeron mapas de movimiento de profundidad (DMM) apilando las diferencias entre cuadros de los cuadros de profundidad proyectados. Luego, calcularon el histograma de gradientes orientados (HOG) para representar las acciones. Estos métodos tienen un poder expresivo limitado y, por lo tanto, suelen necesitar ayuda para capturar información espacio-temporal. En los últimos años, los métodos de aprendizaje profundo se han generalizado con el desarrollo de las redes neuronales. La mayoría de los investigadores intentaron comprimir vídeo profundo en imágenes y analizaron patrones de movimiento utilizando CNN [12]. Kamel et al. [27] ingresan imágenes de movimiento de profundidad (DMI) y descriptores de articulaciones móviles (MJD) a CNN para el reconocimiento de acciones. Para codificar información espacio-temporal de secuencias de profundidad, Adrián et al. [28] propusieron 3D-CNN para extraer características de movimiento. Además, propusieron ConvLSTM [29] para acumular patrones de movimiento discriminativos de unidades de largo plazo. Xiao y col. [3] rotaron la cámara virtual dentro del espacio 3D para proyectar densamente un video de profundidad sin procesar desde diferentes puntos de vista de imágenes virtuales y así construyeron imágenes dinámicas de múltiples vistas. Para invariantes de vista en perspectiva, Kumar et al. [30] propusieron una ActionNet basada en CNN y entrenada con un conjunto de datos de vistas múltiples recopilados utilizando cinco cámaras de profundidad. Ghosh et al. [31] calcularon la imagen del historial de movimiento detectado en el borde del descriptor de profundidad de vista múltiple (ED-MHI) como entrada de un modelo C
Wang y cols. [2] utilizaron secuencias de vídeo de profundidad segmentadas para generar tres tipos de imágenes de profundidad dinámicas. Sin embargo, el mapa de profundidad 2D todavía tiene dificultades para explotar plenamente los patrones de movimiento 3D debido a su estructura espacial compacta [10].

Recientemente, la conversión de mapas de profundidad en nubes de puntos para su procesamiento ha logrado mejores resultados tanto en el campo de reconocimiento como en el de segmentación. Numerosos estudios han demostrado

que las nubes de puntos tienen ventajas significativas en la representación de información espacial 3D debido a sus características, como el desorden y la invariancia de rotación. El aprendizaje profundo para nubes de puntos no solo se ha utilizado ampliamente en tareas de clasificación y segmentación, sino que también ha demostrado fuerza muscular en la reconstrucción de escenas [32] y la detección de objetivos [33]. Sin embargo, los métodos anteriores se centran únicamente en entidades dentro de nubes de puntos estáticas. Cuando se utilizan nubes de puntos para el reconocimiento de acciones 3D, es necesario extraer características dinámicas de acuerdo con los intervalos de tiempo y las características de apariencia de todo el proceso de acción. La clave para un procesamiento eficiente de secuencias de nubes de puntos reside en seleccionar un método de análisis de nubes de puntos adecuado. Tomás y col. [34] desarrollaron un método inspirado en la convolución basada en imágenes y emplearon un conjunto de puntos del núcleo para distribuir el peso de cada núcleo. Como herramienta eficaz para analizar y procesar conjuntos de puntos, PointNet++ [11] se aplica ampliamente para el reconocimiento de acciones 3D basado en secuencias de nubes de puntos. El primer método es 3DV [10], que ejecuta voxelización 3D hacia las secuencias de nubes de puntos y describe la apariencia 3D por ocupación espacial, y la agrupación de rangos temporales se utiliza para la extracción 3DV. Este método se centra principalmente en el movimiento general y los cambios de apariencia de una acción. Sin embargo, ignora los detalles de la acción, como un movimiento sutil de la mano, lo que limita su capacidad para representar el comportamiento con precisión. Por lo tanto, nuestro objetivo es capturar las partes cruciales de las acciones y su delicada información para reconocerlas como acciones humanas más sólidas.

3. Metodología 3.1.

Tubería La

tubería del SFCNet propuesto se muestra en la Figura 1. Primero, cada marco de profundidad se transforma en una nube de puntos para preservar mejor las características dinámicas y de apariencia en el espacio 3D. Para facilitar el análisis del uso espacial y delimitar el espacio local, realizamos operaciones de voxelización en las nubes de puntos. A continuación, construimos un marco simétrico para codificar vóxeles 3D, donde las partes clave y la información dinámica global se procesan por separado en el flujo fino local y el flujo grueso global. Luego, adjuntamos el descriptor de intervalo-frecuencia para complementar la información de movimiento. Empleamos PointNet++ [11] para capturar patrones de movimiento y enviar la característica agregada al clasificador para el reconocimiento de acciones 3D.

3.2. Generación de vóxeles tridimensionales

El vídeo de profundidad tiene la ventaja de resistir interferencias externas, como el fondo y la luz, en comparación con el modo RGB porque contiene la información de profundidad del sujeto de la acción. Esencialmente, el vídeo de profundidad es un tipo de datos de series temporales compuestos por mapas de profundidad dispuestos en orden cronológico. Matemáticamente, un vídeo de profundidad con t fotogramas se puede definir como D = {d1, d2 ,. . . , dt}, donde di es un mapa de profundidad del cuadro t en el que cada píxel representa una coordenada 3D (x, y, z) y z es la distancia desde la cámara de profundidad. Dado que es imposible clasificar la importancia de la acción en la dimensión temporal mediante un único criterio, el muestreo uniforme puede ayudarnos a comprender mejor el proceso de movimiento general en comparación con el muestreo aleatorio [10]. Por lo tanto, primero tomamos muestras del video en profundidad de manera uniforme para aliviar la carga computacional y al mismo tiempo mantener la integridad de la acción. La secuencia de profundidad después del muestreo se denota como Dˆ = {d1, d2, . . . , dT}, donde T es el número de fotogramas y el

El valor predeterminado es 64.

Algunos métodos actuales de reconocimiento de acciones [35,36] eligen mapear marcos de profundidad a espacios 2D para su procesamiento directo. Aunque estos enfoques a veces pueden lograr un buen rendimiento , no pueden superar el problema de la representación inadecuada de la información 3D Por lo tanto, para representar mejor el movimiento humano en el espacio 3D, transformamos cada cuadro di en una nube de puntos P = {p1, p2 ,. . . , pn}, donde n es el número de puntos, generando así una secuencia de nube de puntos S = {P1, P2, . . . , PT}. Al generar nubes de puntos, se requieren parámetros intrínsecos de la cámara porque definen el modelo de imagen de la cámara, incluidos

distancia focal y coordenadas del punto principal (cx, cy). Para cada píxel (x, y, z) en la imagen de profundidad, su nube de puntos correspondiente p(x , y , z) se puede obtener mediante la siguiente fórmula:

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

donde fx y fy representan la distancia focal de la cámara de profundidad en la dirección horizontal y vertical, que se puede obtener de los parámetros del dispositivo. fz está configurado en 1 de forma predeterminada.

A diferencia de las imágenes tradicionales (datos estructurados normales), los puntos de la nube de puntos están desordenados, por lo que su procesamiento resulta complicado. Muchos algoritmos existentes están diseñados para datos de cuadrícula regulares. Sin embargo, la nube de puntos desordenada es un grupo de puntos distribuidos aleatoriamente en el espacio 3D, por lo que su estructura es compleja de procesar y analizar directamente. Para resolver este problema, transformamos la nube de puntos en una cuadrícula 3D regular (espacio vóxel) mediante voxelización para regularizar la representación de la nube de puntos. Primero, definimos el tamaño de la cuadrícula de vóxeles Vgrid = (Vx, Vy, Vz) en coordenadas tridimensionales, lo que determina la resolución del proceso de voxelización. Cada celda de esta cuadrícula es un vóxel potencial y el tamaño de cada celda se indica como Vvoxel(dx, dy, dz). Dado un punto p(x, y, z) en la nube de puntos, se asigna a la cuadrícula encontrando el índice de vóxel correspondiente Vindex(x, y, z) de acuerdo con la siguiente ecuación:

$$Vindex(x, y, z) = \left(\frac{x - xmín}{dx}, \frac{y - ymín}{dy}, \frac{z - zmín}{dz}\right) \quad (2)$$

donde xmin, ymin y zmin son las coordenadas mínimas de todas las nubes de puntos. dx, dy y dz se calculan como el tamaño total dividido por el número de celdas en cada dimensión (Vx, Vy, Vz). La función de suelo . redondea hacia abajo al punto más cercano. Definimos que un vóxel está ocupado si contiene una nube de puntos. Luego, la información de apariencia 3D se puede describir observando si los vóxeles han sido ocupados o no, sin tener en cuenta el punto excluido, como se muestra en la Ecuación (3):

$$V^t_{vóxel}(x, y, z) = \begin{cases} 1, & \text{si } V^t_{voxel}(x, y, z) \text{ está ocupado} \\ 0, & \text{de lo contrario} \end{cases} \quad (3)$$

donde $V^t_{voxel}(x, y, z)$ indica un determinado vóxel en el cuadro t . (x, y, z) es el índice de posición 3D normal , es decir, Vindex en la ecuación (2). Esta estrategia tiene dos beneficios principales. En primer lugar, los conjuntos de vóxeles 3D binarios generados son regulares, como se muestra en la Figura 2. Por lo tanto, se reduce la complejidad del procesamiento de la nube de puntos . Además, la voxelización puede comprimir eficazmente las nubes de puntos porque los vóxeles vecinos pueden tener características similares. Esta compresión no sólo reduce la cantidad de puntos sino que también ayuda a reducir la sobrecarga de almacenamiento y cál

3.3. Identificación y representación de partes clave La

cuestión vital en las tareas de reconocimiento de acciones 3D es capturar y representar de manera eficiente características dinámicas dentro de secuencias de nubes de puntos. Por ahora, los métodos de estimación basados en el flujo de escenas [37,38] pueden ayudar a comprender el movimiento 3D, pero requieren mucho tiempo. Algunos estudios utilizan la agrupación de rangos temporales [3,39] para preservar los procesos de movimiento en el espacio 3D dividiendo segmentos de tiempo. Estos métodos pueden capturar más información temporal, pero a menudo solo dividen una pequeña cantidad de intervalos, lo que da como resultado características dinámicas de grano grueso. Proponemos un módulo de codificación e identificación de piezas cruciales para centrarse mejor en la dinámica crítica durante el movimiento. Puede extraer las partes principales del espacio vóxel 3D global de acuerdo con los tiempos de ocupación del espacio y codificar los detalles de la a Específicamente, primero analizamos la ocupación del espacio construyendo un espacio U 3D con los límites exactos como secuencias de nubes de puntos para cada ubicación espacial, y los valores iniciales.

de U se establecen en 0. Procesamos los m grupos uniformes de la secuencia Dˆ en orden. Entonces, el 3D El uso del espacio se puede calcular según la ecuación (4):

$$Uvoxel(x, y, z) = \begin{cases} Uvoxel(x, y, z) + 1, vi = 0 \\ Uvoxel(x, y, z), \text{en caso contrario} \end{cases} \quad (4)$$

La ocupación total del espacio u para cada posición se puede obtener después de contar todos los m conjuntos de puntos. Además, dado que los conjuntos de puntos están naturalmente ordenados en el tiempo, podemos fácilmente Registre la primera y la última vez que se tomó, f y l, respectivamente, para cada ubicación espacial. Entonces, definimos el umbral θ para dividir el espacio local prominente. Los lugares ocupados menores que θ se tratan como ruido incidental, y aquellos que registran más y menos que m constituyen las partes críticas del movimiento S como la Ecuación (5):

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

Si el valor de Svoxel(x, y, z) es igual a 0, esto denota que el vóxel pertenece al grupo global. espacio G; de lo contrario, pertenece al espacio local L. Comparado con la nube de puntos métodos de procesamiento [10,40], que comúnmente adoptan operaciones de reducción de resolución uniformes, El método propuesto es más eficaz, especialmente para acciones que implican sólo un pequeño número. de las partes de las extremidades, porque dividir el espacio local no sólo puede superar el fondo Efectos hasta cierto punto, pero también mejoran eficazmente el contenido de oro del punto de muestra. datos. Como se muestra en la Figura 2, el espacio local L conserva completamente la información detallada del partes principales del cuerpo, lo que proporciona señales críticas para el reconocimiento de acciones en 3D y, al mismo tiempo, reduciendo la redundancia.

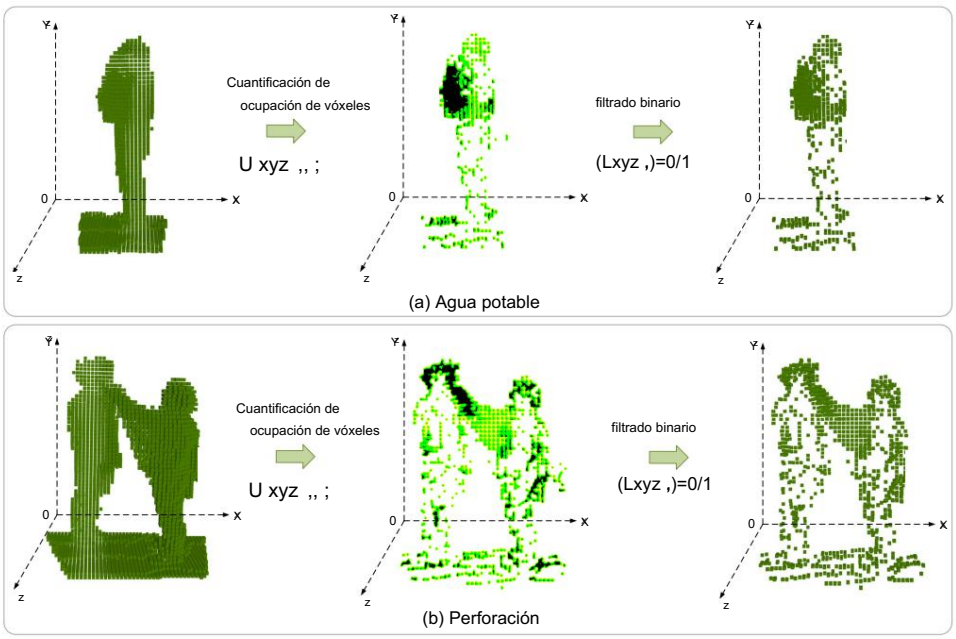


Figura 2. El proceso de división del espacio local. Cuantificamos la contribución de un vóxel a una acción mediante su recuento de ocupación de espacio. Al establecer un umbral, las partes críticas se pueden dividir como un comprimido espacio local, que elimina información redundante y reduce la carga computacional.

3.4. Extracción de características simétricas

Para los vóxeles 3D procesados, la forma más intuitiva es emplear 3DCNN [41,42], pero está limitado por el tamaño del vóxel y requiere mucho tiempo. Optamos por utilizar PointNet++ [11] como extractor de características en este trabajo, como se muestra en la Figura 3. Está diseñado explícitamente para aprendizaje de características en conjuntos de puntos desordenados en el espacio métrico, lo que le permite capturar local patrones de grano fino en nubes de puntos. Para lograr esto, PointNet++ divide el punto

nube en regiones locales superpuestas según una métrica de distancia en el espacio subyacente. Para obtener señales visuales 3D detalladas, PointNet++ emplea recursivamente PointNet [43] para extraer características locales, que luego se fusionan para el análisis de apariencia global. PointNet++ es una excelente alternativa a 3DCNN, ya que funciona bien en la captura de patrones 3D locales esenciales para el reconocimiento de acciones. Además, aplicarlo es relativamente sencillo y sólo requiere transformar vóxeles 3D en conjuntos de puntos.

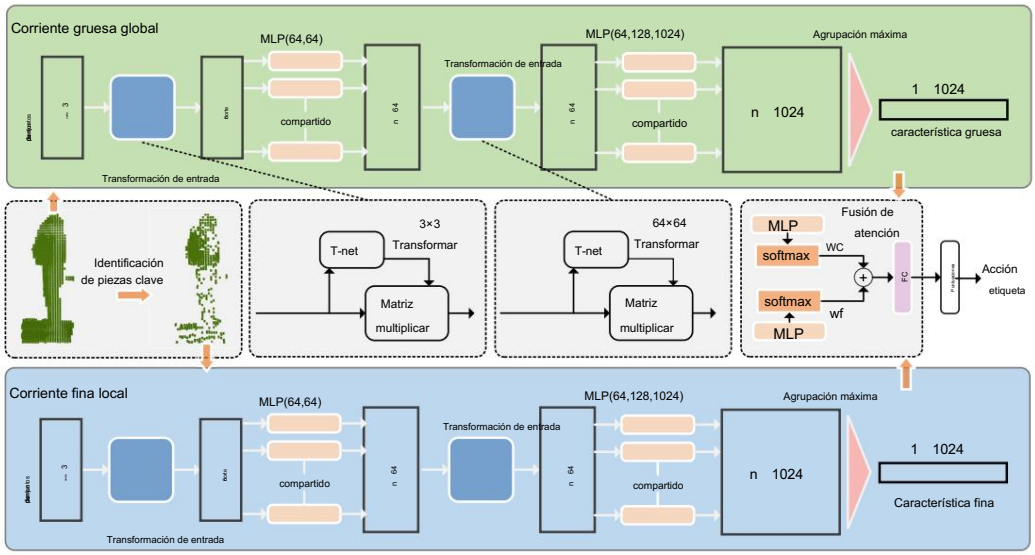


Figura 3. La estructura de red de SFCNet. Es una estructura de red simétrica que comprende el flujo grueso global y un flujo fino local. El flujo grueso global toma la información de apariencia global como entrada, mientras que el flujo fino local solo adopta la información de la parte clave comprimida. PointNet++ extrae las características de cada flujo y las fusiona con los pesos que se pueden aprender como características de acción finales, que se envían al clasificador para el reconocimiento de acciones 3D.

Para ajustar PointNet++, cada punto p(x, y, z) se abstrae como el vóxel Vvoxel(x, y, z) con la característica descriptiva de (x, y, z, I), donde I es la frecuencia del intervalo. descriptor que incluye tres variables (o, f , I) que denotan marcas de tiempo de inicio, marcas de tiempo de finalización y la frecuencia de ocupación general de los vóxeles, respectivamente. Adjuntamos I a las posiciones 3D originales de los vóxeles para obtener los puntos 6D p′ (x, y, z, o, f , I), como se muestra en la Figura 4. Matemáticamente, para un vóxel Vvoxel, podemos determinar el tiempo de inicio y finalización (o y f) de su ocupación espacial voxel(x, y, z) = 1 usando la Ecuación (3), que indica el índice de tiempo cuando se cumplen las condiciones V t y V t voxel(x, y, z) = 0 se satisfacen primero. Además, la ocupación total del espacio u se puede calcular mediante la ecuación (4). En consecuencia, la frecuencia de ocupación I se puede calcular como I = u/(f − o). El descriptor de frecuencia de intervalo cubre el intervalo de tiempo y la frecuencia de ocupación general de las ubicaciones espaciales no solo puede ayudarnos a distinguir las acciones inversas que cubren el mismo espacio, sino que también ayuda a resaltar las diferencias entre acciones al retener información detallada dentro del proceso de acción. Finalmente, los conjuntos de puntos obtenidos se utilizan como entrada de PointNet++ para extraer características de acción.

Además, diseñamos una red de dos flujos para procesar la nube de puntos en el espacio global y local, respectivamente (como se muestra en la Figura 3). El espacio global contiene todos los puntos voxelizados y la corriente gruesa global captura los patrones de movimiento generales. Teniendo en cuenta que las partes vitales del cuerpo pueden proporcionar información dinámica más específica y discriminativa para el reconocimiento de acciones, dividimos los puntos voxelizados de las partes esenciales del cuerpo en espacios locales e ingresamos el flujo fino para extraer características. Después de eso, el módulo de fusión de características se configura para la representación de acciones finas y generales. Considerando las características de las diferentes acciones, su dependencia de las características globales y locales es diversa. Para acciones que solo involucran movimientos de las extremidades, como saludar y patear, el modelo debe enfatizar las características locales detalladas de la corriente. Por el contrario, para grandes movimientos de todo el cuerpo, como caer y saltar, el modelo debería centrarse en las características de la corriente g

Para esta percepción de acción específica, empleamos el módulo de fusión de características con el mecanismo Primero, proyectamos las características Xf R y los de atención n×1024 y Xc R n×1024 extraídos del finura. flujos gruesos, respectivamente, en el espacio de características inferior para reducir la carga computacional y obtener X y Xc . Luego, las características intermedias se extraen mediante un perceptrón f multicapa (MLP).

Después de eso, los pesos aprendibles Wf y Wc de la corriente gruesa global y la corriente fina local se obtienen a través de la función de activación SoftMax, respectivamente.
Finalmente, las características globales y locales se fusionan como se muestra en la Ecuación (6) para obtener la característica de movimiento Xˆ:

$$\begin{aligned} Wf &= \text{SoftMax MLP } X_f' \\ Wc &= \text{SoftMax MLP } X_c' \\ \hat{X} &= \varphi \left(\sum_{yo=1}^{k} {}_{\text{Wisconsin}}X_{ff}(i) + Wi\, cX_c'(i) \right) \end{aligned} \tag{6}$$

donde φes la capa lineal y K es la longitud de la salida de las entidades por MLP. Finalmente, la capa completamente conectada restauró dimensionalmente la característica fusionada Xˆ y el clasificador SoftMax obtuvo las puntuaciones de predicción finales. A diferencia de los métodos existentes que analizan directamente el estado de movimiento de todas las nubes de puntos, nuestro trabajo se centra en la parte crítica del cuerpo al realizar acciones, lo que ayuda a superar la influencia de datos redundantes, como los antecedentes, en el reconocimiento de acciones 3D. Además, los detalles de las partes cruciales y la apariencia global del cuerpo humano se complementan entre sí, lo que estimula la extracción de características discriminativas para el reconocimiento de acciones en 3D.

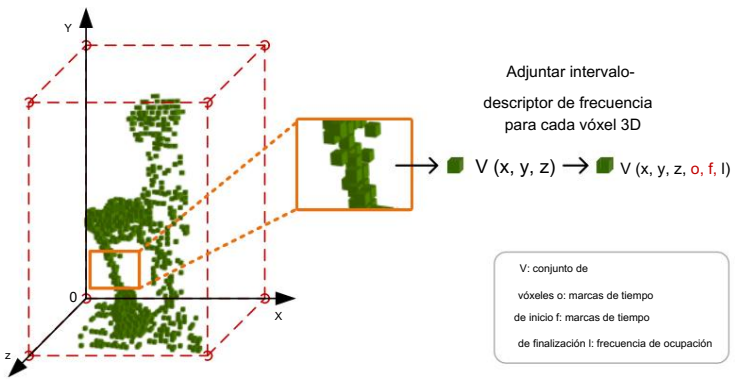


Figura 4. Ilustración del descriptor adicional de intervalo-frecuencia. Contiene marcas de tiempo de inicio, marcas de tiempo de finalización y la frecuencia de ocupación general de los vóxeles y se denota como (o, f , l) en la Sección 3.4; así, la información espacial de profundidad 3D se transforma en características de seis canales.

4. Experimentos 4.1.
Conjuntos de datos

Conjunto de datos NTU RGB+D 60. El NTU RGB+D 60 [44] es un conjunto de datos de reconocimiento de acciones 3D a gran escala que contiene alrededor de 56,880 muestras de acciones RGB+D. Utiliza Microsoft Kinect V2 para capturar 60 categorías de acciones realizadas por 40 sujetos. El conjunto de datos sigue dos principios de evaluación. En el caso de la vista cruzada, las muestras capturadas por la cámara 1 se utilizaron como conjunto de prueba y las cámaras 2 y 3 se consideraron como conjunto de entrenamiento. Es decir, el número de muestras de prueba es 18.960 y se utilizan 37.920 muestras para el entrenamiento. En el caso de temas cruzados, los datos se dividen en un conjunto de prueba de 16.560 muestras y un conjunto de entrenamiento de 40.320 muestras según la identificación del sujeto.

Conjunto de datos NTU RGB+D 120. NTU RGB+D 120 [45] es un extenso conjunto de datos para el reconocimiento de acciones 3D, que consta de 114.480 muestras y 120 categorías de acciones completadas por 106 sujetos. Este conjunto de datos contiene acciones diarias, acciones relacionadas con la medicina y acciones de interacción de dos personas . Las muestras se recolectan en varios lugares y fondos, que se indican como 32 configuraciones. Además de las configuraciones generales entre temas, la configuración cruzada

Se introduce la evaluación, donde el conjunto de entrenamiento proviene de muestras con ID de configuración impares, y el conjunto de pruebas proviene del resto.

4.2. Detalles del entrenamiento

De forma predeterminada, SFCNet y sus variaciones se entrenan utilizando el momento adaptativo. Optimizador de estimación (Adam) para 60 épocas bajo el marco de aprendizaje profundo PyTorch, a menos que se diga lo contrario. Usamos la pérdida de entropía cruzada estándar y aplicamos técnicas de aumento de datos como rotación aleatoria, tramado y abandono a los datos de entrenamiento. La tasa de aprendizaje comienza en 0,001 y decae en 0,5 cada diez épocas. Para garantizar la equidad , seguimos estrictamente el esquema de segmentación de la muestra de los dos conjuntos de datos de acuerdo con puntos de referencia.

4.3. Análisis de parámetros

El tamaño de los vóxeles 3D. Las nubes de puntos suelen estar formadas por un gran número de puntos desordenados. conjuntos de puntos. Debido a la complejidad de estos datos, almacenarlos puede llevar mucho tiempo. y proceso. Para aliviar este problema, incorporamos los puntos desordenados en el espacio 3D en un Estructura de cuadrícula regular mediante rasterización. Esto convierte la nube de puntos en vóxeles 3D. basado en la ocupación del espacio, discretizando el espacio 3D continuo en una cuadrícula regular. Este El proceso proporciona una estructura regular y bien entendida, lo que reduce la complejidad computacional. complejidad y comprime los datos de la nube de puntos, lo que reduce significativamente carga. Es esencial voxelizar la nube de puntos adecuadamente porque el tamaño de la El vóxel 3D determina la fuerza de la compresión de la nube de puntos y la granularidad de la representación de la nube de puntos. Para examinar el impacto de la voxelización en los resultados, evaluó el rendimiento de SFCNet en el conjunto de datos NTU RGB+D 60 para diferentes tamaños vóxeles. Los resultados se presentan en la Tabla 1, lo que indica que el modelo funciona mejor para un Tamaño del cubo de 35 mm. Configurar el tamaño demasiado grande o demasiado pequeño puede provocar una disminución de la precisión.

Tabla 1. Rendimiento en el conjunto de datos NTU RGB+D 60 con voxelización de diferentes tamaños.

Tamaño del vóxel (mm)	Tema cruzado	Vista cruzada
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%

El establecimiento del umbral θ . El comportamiento humano normalmente implica sólo el movimiento de partes específicas del cuerpo, como agitar los brazos, caminar con las piernas, girar la cabeza, etc. Esta localidad significa que el análisis del comportamiento debería centrarse más en partes esenciales del cuerpo que en las todo el cuerpo. Con la ayuda de la variable de frecuencia de ocupación l en el intervalo-frecuencia descriptor, podemos describir el compromiso de cada vóxel, que está relacionado positivamente con la contribución de la parte del cuerpo en el proceso de ejecución de la acción. Ya que tomamos muestras de secuencia de acción en profundidad en grupos de igual duración, la frecuencia de ocupación l positivamente se correlaciona con el número de ocupación u en la Sección 3.3. Entonces, se emplea un umbral θ para Evaluar la atención prestada a la parte del cuerpo. Para investigar la influencia del umbral, Compare el rendimiento de SFCNet en el conjunto de datos NTU RGB+D 60 con varios valores. Los hallazgos se presentan en la Tabla 2. El resultado óptimo se puede obtener cuando θ es igual 30. Nuestra investigación descubrió que ligeras modificaciones en θ , de no más de 5, resultaban en fluctuaciones en la precisión, lo que subraya la importancia de investigar los umbrales y descomponer partes importantes del cuerpo.

Tabla 2. Rendimiento en el conjunto de datos NTU RGB+D 60 con varios valores de θ .

El valor del umbral θ	Tema cruzado	Vista cruzada
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. Estudio de ablación

Efectividad del descriptor intervalo-frecuencia. Solo los datos originales de la nube de puntos 3D. Contienen la información de ubicación de los puntos en el espacio 3D. Incluso si la dimensión del tiempo se introduce en las secuencias de nubes de puntos, todavía es un desafío describir el conjunto dinámica espaciotemporal del punto basándose únicamente en estas pistas. hemos diseñado un descriptor de intervalo-frecuencia que captura el tiempo de inicio y el número de vóxeles ocupación. Esta información adicional nos ayuda a describir de manera integral las características humanas. comportamiento capturando características de movimiento adicionales. Realizamos estudios de ablación en el Conjunto de datos NTU RGB+D 60 donde eliminamos información de características de movimiento en dos flujos, y los conjuntos de puntos ingresados a SFCNet solo tenían coordenadas 3D (x, y, z). Los resultados de la comparación se presentan en la Tabla 3. Observamos que sin características tridimensionales adicionales, Hubo una degradación significativa del rendimiento de SFCNet en más del 10%.

Tabla 3. Efectividad del descriptor intervalo-frecuencia en la NTU RGB+D 60.

Característica de punto	Tema cruzado	Vista cruzada
(x, y, z)	78,0%	82,3%
(x, y, z, o, f, l)	89,9%	96,7%

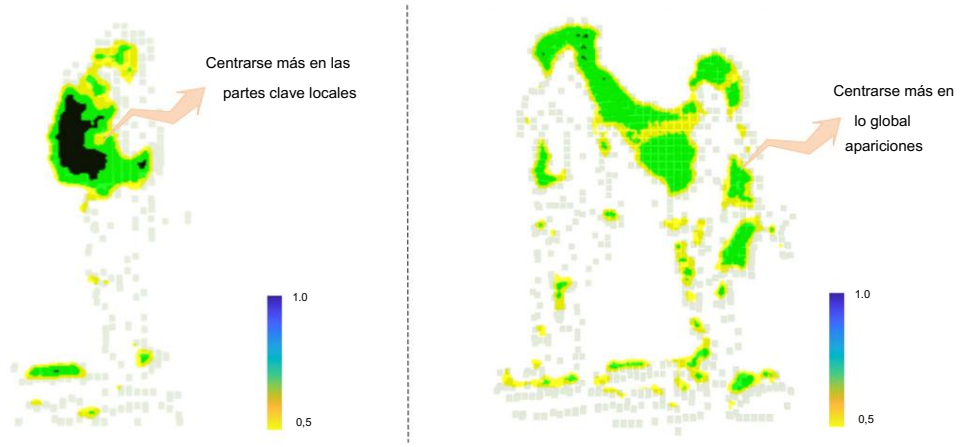
Esto indica que el descriptor de intervalo-frecuencia representa efectivamente la dinámica características dentro de todo el proceso de acción, que juega un papel vital en el reconocimiento de acciones 3D.

Efectividad de la fusión de características de dos flujos. Diferentes acciones humanas contienen diferentes Dinámica global y local, describimos el patrón de movimiento general de todo el ser humano. cuerpo durante la acción a través de la corriente gruesa global. Por el contrario, la fina corriente local Describe la dinámica de partes cruciales del cuerpo, que presta más atención a los detalles y características locales de las acciones y ayuda a capturar los cambios sutiles y los patrones de acción complejos. Los resultados en la Tabla 4 muestran que nuestro SFCNet propuesto fusiona las características finas y gruesas del Dos corrientes pueden comprender y reconocer las acciones humanas de manera más integral. Primero nosotros Analice el rendimiento del reconocimiento en el estado de flujo único. Se puede observar que el local La secuencia tiene una mayor capacidad para representar la acción que la secuencia global, lo que indica que Es esencial prestar atención a las partes principales involucradas para eliminar la redundancia. Además, Comparamos tres estrategias diferentes de fusión de características para demostrar la superioridad de fusión de características basada en la atención propuesta en SFCNet (ver Ecuación (6)). Como se muestra en la tabla 4. SFCNet (fusión) tiene ventajas aparentes sobre la cascada nativa o la fusión aditiva estrategias. La razón principal es que la fusión de características basada en la atención puede asignar de forma adaptativa la atención del modelo a las características de la corriente gruesa global y la corriente fina local. Como se muestra en la Figura 5, beber agua sólo implica la interacción entre manos y cabeza, y el La amplitud del movimiento es pequeña, por lo que el modelo enfatiza las características del local. flujo para capturar patrones de movimiento detallados. Por otro lado, el puñetazo implica la interacción de dos personas, y el movimiento de golpe es grande y poderoso, por lo que Se presta más atención a la apariencia global del cuerpo mientras se enfatizan algunos detalles de las manos y la cabeza. Este mecanismo de extracción de características específicas de la acción mejora la capacidad de generalización de SFCNet y precisión para el reconocimiento de acciones 3D.

$$U_{xyz}; \quad L_{xyz}, =0/1$$

Tabla 4. Efectividad de la fusión de características de dos flujos en el conjunto de datos NTU RGB+D 60.

Flujo de entrada	Tema cruzado	Vista cruzada
1s-SFCNet (L)	85,0%	94,6%
1s-SFCNet (G)	81,8%	86,6%
SFCNet (concat)	88,9%	94,8%
SFCNet (agregar)	86,7%	93,9%
SFCNet (fusión)	89,9%	96,7%



(a) Agua potable (b) Perforación

Figura 5. Visualización de atención de funciones. Visualizamos el mapa de calor del agua potable y del ponche.

4.5. Comparación con métodos existentes

Para evaluar el rendimiento del SFCNet propuesto, lo comparamos con métodos existentes. en dos grandes conjuntos de datos de referencia, como se muestra en las Tablas 5 y 6. Dividimos la acción 3D existente métodos de reconocimiento en métodos basados en esqueleto y métodos basados en profundidad. En los métodos basados en esqueleto , comparamos diferentes métodos basados en backbone, incluidos CNN [5], LSTM [15,46] y GCN [6,23,47]. Para los métodos basados en profundidad, comparamos los métodos basados en imágenes 2D [19,35,48], Métodos basados en CNN 3D [28] y métodos basados en vóxeles 3D [10]. Los resultados de la comparación son reportado en las Tablas 5 y 6. Para el conjunto de datos NTU RGB+D 60, el SFCNet propuesto logra Precisión del 89,9% y del 96,7% en el caso de configuraciones entre sujetos y vistas cruzadas. Además, nosotros Compare SFCNet con dos métodos basados en datos multimodales. En comparación con ED-MHI [31], que combina datos de profundidad y esqueleto, nuestro método mejora la precisión en un 4,3% en el configuración transversal. TS-CNN-LSTM [49] fusionó datos de tres modalidades, a saber, RGB, profundidad, y esqueleto, pero es un 2,6% y un 4,9% más bajo que SFCNet en la configuración de temas cruzados y de vistas cruzadas, respectivamente. Para el conjunto de datos NTU RGB+D 120, SFCNet también logra resultados competitivos resultados, logrando precisiones del 83,6% y 93,8% en configuraciones de sujetos cruzados y vistas cruzadas, respectivamente. En general, SFCNet es efectivo y excelente, lo que supera a los métodos tradicionales que utilizan extracción manual de características [19,35,50] y métodos de aprendizaje profundo que comprimen la profundidad. vídeo en imágenes para su procesamiento [2,3,36] o secuencias de nubes de puntos [10]. Los resultados experimentales demostrar que SFCNet es superior para capturar patrones de comportamiento humano discriminativos y por lo tanto, es beneficioso para el reconocimiento de acciones en 3D.

Tabla 5. Comparación de diferentes métodos para la precisión del reconocimiento de acciones (%) en la NTU RGB+D 60 conjuntos de datos.

Método	Entrada	Vista cruzada	Año
	transversal: esqueleto 3D		
GCA-LSTM [15]	74,4	82,8	2017
Atención de dos flujos LSTM [46]	77.1	85.1	2018
ST-GCN [23]	81,5	88.3	2018
Movimiento esquelético [5]	69,6	80.1	2019
AS-GCN [6] 2s-	86,8	94.2	2019
AGCN [47]	88,5	95.1	2019
ST-TR (nuevo) [24]	89,9	96.1	2021
DSwarm-Net (nuevo) [51]	85,5	90.0	2022
Red de acción [30]	73.2	76.1	2023
SGMSN (nuevo) [52]	90.1	95,8	2023
	Entrada: mapas de profundidad		
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 (nuevo) [53]	90,6	94,7	2021
3DFCNN [28]	78.1	80,4	2022
Poda 3D [54]	83,6	92,4	2022
ConvLSTM (nuevo) [29]	80,4	79,9	2022
CBBMC (nuevo) [48]	83,3	87,7	2023
PointMapNet (nuevo) [55]	89,4	96,7	2023
SFCNet (nuestro)	89,9	96,7	-
	Entrada: Multimodalidades		
ED-MHI [31]	85,6	-	2022
TS-CNN-LSTM [49]	87,3	91,8	2023

Tabla 6. Comparación de diferentes métodos para la precisión del reconocimiento de acciones (%) en la NTU RGB+D 120 conjuntos de datos.

Método	Tema cruzado	Conjunto cruzado	Año
	Entrada: esqueleto 3D		
GCA-LSTM [15]	58.3	59.3	2017
Mapa de evolución de la postura corporal [56]	64,6	66,9	2018
Atención de dos flujos LSTM [46]	61.2	63.3	2018
ST-GCN [23]	70,7	73.2	2018
Línea base NTU RGB+D 120 [45]	55,7	57,9	2019
FSNet [57]	59,9	62,4	2019
Movimiento esquelético [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 (nuevo) [24]	82,7	84,7	2021
SGMSN (nuevo) [52]	84,8	85,9	2023
	Entrada: mapas de profundidad		
APSR [45]	48,7	40,1	2019
3DV-PointNet++ [10]	82,4	93,5	2020
DOGV (nuevo) [53]	82,2	85,0	2021
Poda 3D [54]	76,6	88,8	2022
SFCNet (nuestro)	83,6	93,8	-

5. Discusión

Para analizar las ventajas y desventajas del método propuesto, presentamos la precisión de reconocimiento de SFCNet en el conjunto de datos NTU RGB+60 en las configuraciones entre sujetos para cada categoría. Los resultados se muestran en forma de matriz de confusión en la Figura 6 (izquierda). Hemos seleccionado algunas acciones confusas para una visualización más clara y las hemos ampliado localmente, como se muestra en la Figura 6 (derecha). Los resultados muestran que SFCNet tiene una sólida capacidad de análisis de la acción humana, con una precisión de reconocimiento superior al 90% en la mayoría de las categorías. Por ejemplo, ha logrado un 100% de precisión al saltar y un 99% al saltar . Sin embargo, SFCNet está confundida acerca del reconocimiento de algunas acciones similares. Por ejemplo, leer y escribir, usar zapatos y quitarse los zapatos son los pares de muestra más confusos. Además, el 25% de los que jugaban con sus teléfonos fueron clasificados erróneamente como lectura (9%), escritura (8%) y escritura en el teclado (8%). La precisión al escribir en el teclado es solo del 66% y el 12% de las muestras se clasifican erróneamente como escritura. A partir del análisis, hemos descubierto que estas acciones solo tienen diferencias sutiles y la amplitud del movimiento es pequeña. Ésta es la razón principal por la que tales acciones son difíciles de distinguir.

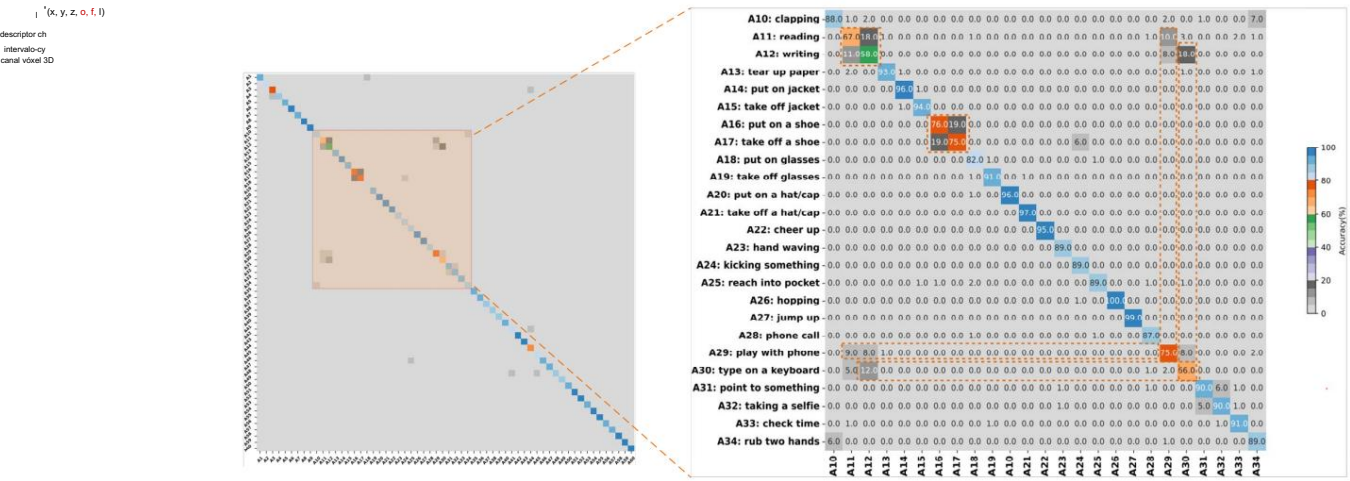


Figura 6. Matriz de confusión para la precisión del reconocimiento específico de una clase. La imagen de la (izquierda) contiene todas las categorías de acciones. Para enfatizar algunas de las acciones de ofuscación, se muestra un acercamiento local a la derecha.

6. Conclusiones

En este artículo, proponemos una red neuronal simétrica, SFCNet, para reconocer acciones 3D a partir de secuencias de nubes de puntos. Contiene un flujo grueso global y un flujo fino local que emplea PointNet++ como extractor de características. Las secuencias de nubes de puntos se regularizan como conjuntos de vóxeles estructurados añadidos por el descriptor de frecuencia de intervalo propuesto para generar características 6D que capturan información dinámica espaciotemporal. La corriente gruesa global captura los patrones de acción de grano grueso de la apariencia del cuerpo humano, y la corriente delicada local extrae características de grano fino específicas de la acción de partes críticas. Después de la fusión de características, SFCNet puede extraer patrones de movimiento discriminativos que involucran cambios espaciales generales y enfatizar detalles cruciales de un extremo a otro. Según los resultados experimentales en dos grandes conjuntos de datos de referencia, NTU RGB+D 60 y NTU RGB+D 120, SFCNet es eficaz para el reconocimiento de acciones 3D y tiene potencial para aplicaciones de detección remota. Sin embargo, la SFCNet propuesta todavía tiene limitaciones a la hora de distinguir acciones similares. Nuestro trabajo futuro se centrará en reconocer acciones similares y capturar patrones sutiles para mejorar la precisión.

Contribuciones de los autores: Conceptualización, CL y QH; metodología, CL, WQ y XL; software, QH y YM; validación, CL, WQ y XL; análisis formal, QH y YM; investigación, CL y WQ; recursos, QH y YM; curación de datos, WQ; redacción: preparación del borrador original, CL y WQ; redacción: revisión y edición, CL, YM, WQ, QH y XL; visualización, WQ; supervisión,

QH y YM; administración de proyectos, QH y YM; adquisición de financiación, QH, YM y CL Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación fue financiada por el Programa de Innovación en Práctica e Investigación de Postgrado de la provincia de Jiangsu (número de subvención KYCX23_0753), los Fondos de Investigación Fundamental para las Universidades Centrales (número de subvención B230205027), el Programa Clave de Investigación y Desarrollo de China (número de subvención 2022YFC3005401) , el Programa Clave de Investigación y Desarrollo de China, Provincia de Yunnan (número de subvención 202203AA080009), el 14.º Plan Quinquenal para las Ciencias de la Educación de la Provincia de Jiangsu (número de subvención D/2021/01/39) y el Proyecto de Investigación de Reforma de la Educación Superior de Jiangsu (número de subvención 2021JSJG143); y el APC fue financiado por los Fondos de Investigación Fundamental de las Universidades Centrales.

Declaración de la Junta de Revisión Institucional: No aplicable.

Declaración de Consentimiento Informado: No aplicable.

Declaración de disponibilidad de datos: los conjuntos de datos NTU RGB+D 60 y NTU RGB+D 120 utilizados en este documento son públicos, gratuitos y están disponibles en: https://rose1.ntu.edu.sg/dataset/actionRecognition/ (consultado el 21 de diciembre de 2020).

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Referencias

1. Riaz, W.; Gao, C.; Azeem, A.; Saifullah; Bux, JA; Ullah, A. Sistema de predicción de anomalías de tráfico mediante red predictiva. Sensores remotos 2022, 14, 1–19.
2. Wang, P.; Li, W.; Gao, Z.; Tang, C.; Ogunbona, PO Reconocimiento de acciones tridimensionales a gran escala basado en agrupación de profundidad con redes neuronales convolucionales. Traducción IEEE. Multimed. 2018, 20, 1051–1061.
3. Xiao, Y.; Chen, J.; Wang, Y.; Cao, Z.; Zhou, JT; Bai, X. Reconocimiento de acciones para videos en profundidad utilizando imágenes dinámicas de múltiples vistas. inf. Ciencia. 2019, 480, 287–304.
4. Li, C.; Huang, Q.; Li, X.; Wu, Q. Reconocimiento de acciones humanas basado en mapas de características de múltiples escalas a partir de secuencias de video en profundidad. Multimed. Herramientas Aplica. 2021, 80, 32111–32130.
5. Caetano, C.; Sena, J.; Bremond, F.; Dos Santos, JA; Schwartz, WR Skelemotion: una nueva representación de secuencias de articulaciones esqueléticas basadas en información de movimiento para el reconocimiento de acciones en 3D. En actas de la Conferencia internacional IEEE sobre vigilancia avanzada basada en señales y vídeo, Taipei (Taiwán), 18 a 21 de septiembre de 2019; IEEE: Piscataway, Nueva Jersey, EE. UU., 2019; págs. 1–8.
6. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Redes convolucionales de gráficos estructurales de acción para el reconocimiento de acciones basado en esqueletos. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Long Beach, CA, EE. UU., 15 a 20 de junio de 2019; págs. 3595–3603.
7. Yang, X.; Zhang, C.; Tian, Y. Reconocimiento de acciones utilizando histogramas de gradientes orientados basados en mapas de movimiento de profundidad. En Actas de la Conferencia Internacional ACM sobre Multimedia, Nara, Japón, 29 de octubre a 2 de noviembre de 2012; págs. 1057-1060.
8. Elmadany, NED; Ey.; Guan, L. Fusión de información para el reconocimiento de la acción humana a través de localidad de globalidad biset / multiset preservando el análisis de correlación canónica. Traducción IEEE. Proceso de imagen. 2018, 27, 5275–5287.
9. Él, K.; Zhang, X.; Ren, S.; Sun, J. Aprendizaje residual profundo para el reconocimiento de imágenes. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Las Vegas, NV, EE. UU., 27 a 30 de junio de 2016; págs. 770–778.
10. Wang, Y.; Xiao, Y.; Xiong, F.; Jiang, W.; Cao, Z.; Zhou, JT; Yuan, J. 3DV: Vóxel dinámico 3D para vídeo en profundidad de reconocimiento de acciones. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Seattle, WA, EE. UU., 14 a 19 de junio de 2020; págs. 508–517.
11. Qi, CR; Yi, L.; Su, H.; Guibas, LJ Pointnet++: Aprendizaje profundo de características jerárquicas en conjuntos de puntos en un espacio métrico. Adv. Inf. neuronal. Proceso. Sistema. 2017, 30, 5105–5114.
12. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, PO Reconocimiento de acciones a partir de mapas de profundidad utilizando convolución profunda Redes neuronales. Traducción IEEE. Tararear. -Mach. Sistema. 2015, 46, 498–509.
13. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Una nueva representación de secuencias de esqueletos para el reconocimiento de acciones en 3D. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Honolulu, HI, EE. UU., 21 a 26 de julio de 2017; págs. 3288–3297.
14. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Aprendizaje de funciones de coocurrencia a partir de datos esqueléticos para el reconocimiento y detección de acciones con agregación jerárquica. arXiv 2018, arXiv:1804.06055v1.
15. Liu, J.; Pandilla, W.; Ping, H.; Duan, LY; Kot, AC Redes LSTM de atención global consciente del contexto para el reconocimiento de acciones 3D. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Honolulu, HI, EE. UU., 21 a 26 de julio de 2017; págs. 3671–3680.
16. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Reconocimiento de acciones basado en esqueletos con redes neuronales de gráficos dirigidos. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Long Beach, CA, EE. UU., 15 a 20 de junio de 2019; págs. 7912–7921.

17. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. Una red lstm convolucional de gráficos de atención mejorada para el reconocimiento de acciones basado en esqueleto . En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Long Beach, CA, EE. UU., 15 a 20 de junio de 2019; págs. 1227-1236.

18. Yu, Z.; Wenbin, C.; Guodong, G. Evaluación de características de puntos de interés espaciotemporales para el reconocimiento de acciones basado en profundidad. Imagen Vis. Computadora. 2014, 32, 453–464.
19. Oreifej, O.; Liu, Z. Hon4d: Histograma de normales 4d orientadas para el reconocimiento de actividad a partir de secuencias de profundidad. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Portland, Oregón, EE. UU., 23 a 28 de junio de 2013; págs. 716–723.
20. Caetano, C.; Bremond, F.; Schwartz, WR Representación de imágenes de esqueleto para reconocimiento de acciones en 3D basado en la estructura del árbol y las juntas de referencia. En Actas de la Trigésima Segunda Conferencia SIBGRAPI sobre Gráficos, Patrones e Imágenes, Río de Janeiro, Brasil, 28 al 31 de octubre de 2019; págs. 16-23.

21. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Lstm espacio-temporal con puertas de confianza para el reconocimiento de acciones humanas en 3D. En Actas de la Conferencia europea sobre visión por computadora, Ámsterdam (Países Bajos), 11 a 14 de octubre de 2022; Springer: Berlín/Heidelberg, Alemania, 2016; págs. 816–833.

22. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. Ver redes neuronales adaptativas basadas en esqueletos de alto rendimiento reconocimiento de la acción humana. Traducción IEEE. Patrón Anal. Mach. Intel. 2019, 41, 1963–1978.
23. Yan, S.; Xiong, Y.; Lin, D. Redes convolucionales de gráficos temporales espaciales para el reconocimiento de acciones basado en esqueletos. En Actas de la Trigésima Segunda Conferencia AAAI sobre Inteligencia Artificial, Nueva Orleans, LA, EE. UU., 2 a 7 de febrero de 2018; págs. 7444–7452.
24. Plizzari, C.; Cannici, M.; Matteucci, M. Reconocimiento de acciones basado en esqueletos a través de redes transformadoras espaciales y temporales. Computadora. Vis. Comprensión de la imagen. 2021, 208-209, 103219.
25. Shotton, J.; Fitzgibbon, A.; Cocinero, M.; Afilado, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Reconocimiento de pose humana en tiempo real en partes a partir de imágenes de profundidad única. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Colorado Springs, CO, EE. UU., 20 a 25 de junio de 2011; IEEE: Piscataway, Nueva Jersey, EE. UU., 2011; págs. 1297-1304.
26. Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, JT; Yuan, J. A2j: Red de regresión de anclaje a articulación para estimación de pose articulada en 3D a partir de una única imagen de profundidad. En Actas de la Conferencia Internacional IEEE sobre Visión por Computadora, Seúl, República de Corea, 27 de octubre a 2 de noviembre de 2019; págs. 793–802.
27. Kamel, A.; Sheng, B.; Yang, P.; Li, P.; Shen, R.; Feng, DD Redes neuronales convolucionales profundas para el reconocimiento de la acción humana Uso de mapas de profundidad y posturas. Traducción IEEE. Sistema. Hombre Cibernético. Sistema. 2019, 49, 1806–1819.
28. Sánchez-Caballero, A.; de López-Diz, S.; Fuentes-Jiménez, D.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Casillas-Pérez, D.; Sarker, MI 3DFCNN: Reconocimiento de acciones en tiempo real utilizando redes neuronales profundas 3D con información de profundidad sin procesar. Multimed. Herramientas Aplica. 2022, 81, 24119–24143.

29. Sánchez-Caballero, A.; Fuentes-Jiménez, D.; Losada-Gutiérrez, C. Reconocimiento de acciones humanas en tiempo real mediante vídeo en profundidad sin procesar. Redes neuronales recurrentes basadas en Multimed. Herramientas Aplica. 2022, 82, 16213–16235.
30. Kumar, fiscal del distrito; Kishore, PVV; Murthy, G.; Chaitanya, TR; Subhani, S. Ver el reconocimiento invariante de la acción humana utilizando mapas de superficie a través de redes convolucionales. En Actas de la Conferencia Internacional sobre Metodologías de Investigación en Gestión del Conocimiento, Inteligencia Artificial e Ingeniería de Telecomunicaciones, Chennai, India, 1 y 2 de noviembre de 2023; págs. 1 a 5.
31. Ghosh, SK; SEÑOR.; Mohán, BR; Guddeti, RMR Reconocimiento de acción humana 3D multivista basado en aprendizaje profundo utilizando datos de esqueleto y profundidad. Multimed. Herramientas Aplica. 2022, 82, 19829–19851.
32. Li, R.; Li, X.; Fu, CW; Cohen-Or, D.; Heng, PA Pu-gan: una red adversaria de muestreo mejorado de nubes de puntos. En Actas de la Conferencia Internacional IEEE/CVF sobre Visión por Computadora, Seúl, República de Corea, 27 de octubre a 2 de noviembre de 2019; págs. 7203–7212.
33. Qi, CR; Letanía, O.; Él, K.; Guibas, LJ Deep hough votando a favor de la detección de objetos 3D en nubes de puntos. En Actas de la Conferencia Internacional IEEE/CVF sobre Visión por Computadora, Seúl, República de Corea, 27 de octubre a 2 de noviembre de 2019; págs. 9277–9286.
34. Thomas, H.; Qi, CR; Deschaud, JE; Marcotegui, B.; Goleta, F.; Guibas, L. KPConv: Convolución flexible y deformable para nubes de puntos. En Actas de la Conferencia Internacional IEEE/CVF sobre Visión por Computadora, Seúl, República de Corea, 27 de octubre a 2 de noviembre de 2019; págs. 6410–6419.

35. Ohn-Bar, E.; Trivedi, MM Similitudes de ángulos conjuntos y HOG2 para el reconocimiento de acciones. En Actas de la Conferencia IEEE/CVF sobre talleres de visión por computadora y reconocimiento de patrones, Portland, Oregón, EE. UU., 23 a 28 de junio de 2013; págs. 465–470.
36. Li, J.; Wang, Y.; Zhao, Q.; Kankanhalli, MS Aprendizaje no supervisado de representaciones de acciones invariantes de vista. Adv. Inf. neuronal. Proceso. Sistema. 2018, 31, 1262–1272.
37. Liu, X.; Qi, CR; Guibas, LJ Flownet3d: Aprendizaje del flujo de escenas en nubes de puntos 3D. En las actas de la Conferencia IEEE/CVF sobre Visión por computadora y reconocimiento de patrones, Long Beach, CA, EE. UU., 15 a 20 de junio de 2019; págs. 529–537.
38. Zhai, M.; Xiang, X.; Lv, N.; Kong, X. Estimación del flujo óptico y del flujo de la escena: una encuesta. Reconocimiento de patrones. 2021, 114, 107861.
39. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Agrupación de rangos para el reconocimiento de acciones. Traducción IEEE. Patrón Anal. Mach. Intel. 2016, 39, 773–787.
40. Liu, J.; Xu, D. GeometryMotion-Net: una sólida línea de base de dos flujos para el reconocimiento de acciones 3D. Traducción IEEE. Sistema de circuitos. Tecnología de vídeo . 2021, 31, 4711–4721.
41. Dou, W.; Chin, WH; Kubota, N. Red de memoria en crecimiento con peso aleatorio 3DCNN para el reconocimiento continuo de la acción humana. En Actas de la Conferencia Internacional IEEE sobre Sistemas Difusos, Incheon, República de Corea, 13-17 de agosto de 2023; págs. 1–6.

42. Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; Kankanhalli, M. PSTNet: Convolución espacio-temporal de puntos en secuencias de nubes de puntos. En Actas de la Conferencia Internacional sobre Representaciones del Aprendizaje, Addis Abeba, Etiopía, 26-30 de abril de 2020; págs. 1–6.
43. Qi, CR; Su, H.; Mo, K.; Guibas, LJ Pointnet: Aprendizaje profundo sobre conjuntos de puntos para clasificación y segmentación 3D. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Honolulu, Hawaii, EE. UU., 21 a 26 de julio de 2017; págs. 652–660.
44. Shahroudy, A.; Liu, J.; Ng, TT; Wang, G. NTU RGB+D: un conjunto de datos a gran escala para análisis de actividad humana en 3D. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Las Vegas, NV, EE. UU., 27 a 30 de junio de 2016; págs. 1010-1019.
45. Liu, J.; Shahroudy, A.; Pérez, M.; Wang, G.; Duan, LY; Kot, AC Ntu rgb+ d 120: Un punto de referencia a gran escala para la actividad humana en 3D comprensión. Traducción IEEE. Patrón Anal. Mach. Intel. 2019, 42, 2684–2701.
46. Liu, J.; Wang, G.; Duan, LY; Abdiyeva, K.; Kot, AC Reconocimiento de acción humana basado en esqueletos con conciencia del contexto global Atención Redes LSTM. Traducción IEEE. Proceso de imagen. 2018, 27, 1586–1599.
47. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Redes convolucionales de gráficos adaptativos de dos flujos para el reconocimiento de acciones basado en esqueletos. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Long Beach, CA, EE. UU., 15 a 20 de junio de 2019; págs. 12026-12035.
48. Li, X.; Huang, Q.; Wang, Z. Fusión de información espacial y temporal para el reconocimiento de la acción humana mediante el equilibrio de límites centrales Clasificador multimodal. J. Vis. Comunitario. Representación de imagen. 2023, 90, 103716.
49. Zan, H.; Zhao, G. Investigación sobre el reconocimiento de la acción humana basada en las redes Fusion TS-CNN y LSTM. Árabe. J. Ciencias. Ing. 2023, 48, 2331–2345.
50. Yang, X.; Tian, Y. Vector súper normal para el reconocimiento de actividades mediante secuencias de profundidad. En actas de la conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Columbus, OH, EE. UU., 23 a 28 de junio de 2014; págs. 804–811.
51. Basak, H.; Kundu, R.; Singh, PK; Ijaz, MF; Wo´zniak, M.; Sarkar, R. Una unión de aprendizaje profundo y optimización basada en enjambre para el reconocimiento de acciones humanas en 3D. Ciencia. Representante 2022, 12, 1-17.
52. Qi, Y.; Hu, J.; Zhuang, L.; Pei, X. Reconocimiento de acciones de esqueleto humano a múltiples escalas guiado semánticamente. Aplica. Intel. En t. J. Artif. Intel. Red neuronal. Problema complejo. -Tecnología de resolución. 2023, 53, 9763–9778.
53. Ji, X.; Zhao, Q.; Cheng, J.; Ma, C. Explotación de la representación espacio-temporal para el reconocimiento de acciones humanas en 3D a partir de un mapa de profundidad secuencias. Conocimiento. -Sistema basado. 2021, 227, 107040.
54. Guo, J.; Liu, J.; Xu, D. Poda 3D: un marco de compresión de modelos para un reconocimiento eficiente de acciones 3D. Traducción IEEE. Sistema de circuitos . Tecnología de vídeo. 2022, 32, 8717–8729.
55. Li, X.; Huang, Q.; Zhang, Y.; Yang, T.; Wang, Z. PointMapNet: Red de mapas de características de nubes de puntos para el reconocimiento de acciones humanas en 3D. Simetría 2023, 15, 1–17.
56. Liu, M.; Yuan, J. Reconocer las acciones humanas como la evolución de los mapas de estimación de pose. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Salt Lake City, UT, EE. UU., 18 a 23 de junio de 2018; págs. 1159-1168.
57. Liu, J.; Shahroudy, A.; Wang, G.; Duan, LY; Kot, AC Predicción de acciones en línea basada en esqueletos utilizando una red de selección de escala. Traducción IEEE. Patrón Anal. Mach. Intel. 2019, 42, 1453–1467.