Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
DOI: 10.22034/jscit.2025.222128
Volume 14, Number 1, Spring 2025, pp. 1-10
Received: 07/10/2024, Revised: 01/15/2025, Accepted: 03/18/2025

Using Graph Neural Networks and Attention Mechanism for Skeleton-Based Action Recognition

Mohammad Maleki Sini[1], Reza Roohani Sarvestani[2*]
1- Faculty of Computer Engineering, University of Shahrekord, Shahrekord, Iran.
2*- Faculty of Computer Engineering, University of Shahrekord, Shahrekord, Iran.
[1] malekisinii@gmail.com, [2*]rrohani@sku.ac.ir

Corresponding author's address: Reza Roohani Sarvestani, Faculty of Computer Engineering, Shahrekord University, Shahrekord, Iran.

Abstract- In recent years, graph convolutional networks (GCNs) have achieved significant performance in the field of skeleton-based action recognition. Existing GCN-based methods usually apply fixed graph topologies and a temporal convolutional filter to extract the spatial and temporal features of an action. Since an action is coordinated through different parts of the body in the time domain and shows different characteristics in the time domain, this work causes the loss of a lot of information about an action.

To address this issue, in this paper, we present an attention-based graph neural (AT-AR) network to discover distinct features from both spatial and temporal aspects. The proposed model uses an SPG Net convolution to learn Spatio-temporal features. In addition, the attention mechanism creates an attention score using temporal features, which can enhance the temporal correlations of an action. After establishing the two-stream structure, the AT-AR network has reached 96% and 97% accuracy under X-View and X-Sub on the NTU RGB+D dataset.

Keywords- Human Action Recognition, Skeletal Data, Attention Mechanism, Graph Neural Network, Deep Learning.

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
شناسه دیجیتال: 10.22034/jscit.2025.222128
جلد ۱۴، شماره ۱، بهار ۱۴۰۴ ، صفحه ۱-۱۰
دریافت: ۱۴۰۳/۰۴/۲۰، بازنگری: ۱۴۰۳/۱۰/۲۶، پذیرش: ۱۴۰۳/۱۲/۲۸

استفاده از شبکه‌های عصبی گرافی و مکانیزم توجه برای تشخیص حرکت انسان مبتنی بر داده‌های اسکلتی

محمد ملکی سینی[۱]، رضا روحانی سروستانی*[۲]
۱- دانشکده فنی و مهندسی، دانشگاه شهرکرد، شهرکرد، ایران.
*۲- دانشکده فنی و مهندسی، دانشگاه شهرکرد، شهرکرد، ایران.
[1] malekisinii@gmail.com, [2*]rrohani@sku.ac.ir

* نشانی نویسنده مسئول: رضا روحانی سروستانی، شهرکرد، بلوار رهبر، دانشگاه شهرکرد، دانشکده فنی و مهندسی.

چکیده- در سال‌های اخیر شبکه‌های کانولوشن گرافی (GCN) عملکرد قابل توجهی در زمینه‌ی تشخیص حرکت مبتنی بر اسکلت به دست آورده‌اند. روش‌های مبتنی بر GCN موجود، معمولاً توپولوژی‌های گرافی ثابت و یک فیلتر کانولوشنی زمانی ثابت را برای استخراج ویژگی‌های مکانی و زمانی یک حرکت اعمال می‌کنند. از آنجایی که یک حرکت انجام شده توسط انسان، از طریق بخش‌های مختلف بدن در حوزه‌های زمانی هماهنگ می‌شوند و ویژگی‌های مختلفی را در حوزه‌ی زمانی نشان می‌دهند، این کار باعث از دست رفتن اطلاعات زیادی برای یک حرکت می‌شود. برای پرداختن به این موضوع، در این مقاله یک شبکه‌ی عصبی گرافی مبتنی بر توجه (AT-AR) برای کشف ویژگی‌های متمایز از جنبه‌های مکانی و زمانی ارائه می‌کنیم. مدل پیشنهادی از یک کانولوشن SPG Net برای یادگیری ویژگی‌های مکانی – زمانی استفاده می‌کند.ما همچنین از دو ماژول توجه استفاده می‌کنیم. مکانیسم توجه STA، یک امتیاز توجه را با استفاده از ویژگی‌های زمانی ایجاد می‌کند، که می‌تواند همبستگی‌های زمانی یک حرکت را افزایش دهد و مکانیزم خودتوجه نیز مفاصلی را انتخاب می‌کند که برای تشخیص حرکت مهم‌تر هستند. این دو ماژول توجه در یک شبکه دوجریانی با هم ترکیب شده‌اند و با استفاده از ورودی یکسان در دیتاست NTU RGB+D 60 کار تشخیص حرکت مبتنی بر اسکلت را انجام می‌دهند.

واژه‌های کلیدی: تشخیص حرکت انسان، داده اسکلتی، مکانیسم توجه، شبکه عصبی گرافی، یادگیری عمیق.

۱- مقدمه

تشخیص حرکت انسان[۱] یک کار چالش‌برانگیز در زمینه بینایی ماشین است که به ماشین‌ها کمک می‌کند تا رفتارهای انسان را شناسایی کند. سیستم‌های تشخیص حرکت انسان کاربردهای زیادی در نظارت تصویری، نظارت هوشمند، سرگرمی، روباتیک نظارت بر سلامتی، تعامل بین انسان و کامپیوتر و واقعیت مجازی دارند [۵-۱]. اساساً تشخیص حرکت انسان را می‌توان با استفاده از داده‌های مختلفی مانند تصاویر رنگی [۶]، داده‌های عمق [۷]، توالی‌های مادون قرمز [۸] و داده‌های اسکلتی [۹] انجام داد. با توسعه حسگرهای حرکتی مانند مایکروسافت کینکت [۱۰] و Asus Xtion Pro LIVE [۱۱] و همچنین الگوریتم‌های تخمین‌پوز انسان[۲] مانند Open Pose [۱۲] و Alpha Pose [۱۳]، روش‌های تشخیص حرکت مبتنی بر داده‌های اسکلتی به طور گسترده‌ای در سالیان اخیر مورد مطالعه قرار گرفته‌اند. حسگرهای حرکتی، داده‌های اسکلتی را با استفاده از سنسورهای مختلف به صورت نقاط سه بعدی ثبت و ذخیره می‌کنند و الگوریتم‌های تخمین پوز نیز مختصات مفاصل اسکلتی را از روی تصاویر رنگی استخراج می‌کنند. به طور خاص داده‌های اسکلتی شامل یک توالی زمانی از مختصات دوبعدی یا سه بعدی از مفاصل بدن انسان است که

روابط بین مفاصل و استخوان‌ها را نشان می‌دهد.

برخلاف سایر داده‌ها، داده‌های اسکلتی در برابر تداخلات پس‌زمینه‌ای، تغییرات ظاهری در پس زمینه، تغییرات نماها و تغییرات روشنایی مقاوم هستند [۱۶-۱۴]. به طور کلی هدف تشخیص حرکت مبتنی بر داده‌های اسکلتی، پیش‌بینی حرکات از طریق ویژگی‌های اسکلتی بدن انسان است. روش‌های مبتنی بر یادگیری عمیق اولیه معمولاً از شبکه‌های عصبی کانولوشن[۳] و شبکه‌های عصبی بازگشتی[۴] برای تشخیص حرکت مبتنی بر اسکلت استفاده می‌کنند. روش‌های مبتنی بر CNN[۵] داده‌های اسکلتی را به عنوان یک شبه‌تصویر مدلسازی می‌کنند و از یک کانولوشن برای طبقه‌بندی تصاویر به دسته‌های حرکت استفاده می‌کنند [۲۱-۱۷].

روش‌های مبتنی بر RNN[۶] نیز معمولاً از شبکه‌های LSTM[۷] برای مدلسازی دینامیک زمانی دنباله‌ای از اسکلت‌ها استفاده می‌کنند. این روش‌ها هر اسکلت در یک دنباله اسکلتی را به عنوان یک بردار درنظر می‌گیرند که توسط مختصات به هم پیوسته تشکیل می‌شود [۲۶-۲۲].

روش‌های مبتنی بر CNN و RNN برای داده‌های ساختاریافته و مشبکه در فضای اقلیدسی تعبیه شده‌اند. از طرفی اسکلت بدن انسان به عنوان گراف در فضای غیراقلیدسی است که مفاصل به عنوان رئوس گراف و اتصالات طبیعی بین آن‌ها (استخوان‌ها) به عنوان یال‌های گراف می‌باشند. بنابراین روش‌های مبتنی بر CNN و RNN نمی‌توانند بستر مناسبی را برای ثبت و ذخیره ویژگی‌های حرکتی داده‌های اسکلتی فراهم کنند. پس از ایجاد شبکه‌های کانولوشن گرافی برای داده‌های گرافی، محققان زیادی به استفاده از این شبکه برای تشخیص حرکت مبتنی بر اسکلت روی آوردند و به نتایج قابل قبولی نیز دست یافتند.

اولین تلاش‌ها در این زمینه منجر به پیدایش شبکه ST-GCN [۲۷] شد که از دو جزء اصلی GCN[۸] و کانولوشن زمانی یک بعدی تشکیل شده است. در این شبکه، کانولوشن گرافی ویژگی‌های مکانی مرتبط با اسکلت را استخراج می‌کند و کانولوشن زمانی یک بعدی نیز بر روی فریم‌های متوالی اعمال می‌شود تا ویژگی‌های زمانی مرتبط با اسکلت بدن را استخراج کند.

روش‌های مبتنی بر ST-GCN به دنبال ثبت و ذخیره ویژگی‌های مکانی-زمانی یک دنباله حرکتی از طریق گراف هستند. با این حال اکثر این روش‌ها از دو مشکل کلی رنج می‌برند: ۱) به طور کلی یک حرکت معمولی از مراحل زمانی مختلفی تشکیل شده است. ویژگی‌های یک مفصل در یک حرکت به مفاصل درون فریمی و برون فریمی مرتبط هستند و همبستگی بین چند مرحله زمانی کلیدی با فاصله زمانی مجاور یا نامجاور به تشخیص حرکت کمک به سزایی می‌کند، اما روش‌های مبتنی بر ST-GCN فقط وابستگی‌های زمانی بین فریم‌های متوالی را در نظر می‌گیرند و از وابستگی‌های زمانی بین فریم‌های نامجاور غافل می‌شوند. ۲) در بسیاری از مجموعه داده‌های اسکلتی، برخی از مفاصل دارای اطلاعات قوی‌تری نسبت به سایر مفاصل می‌باشند. برای مثال در حرکت "دویدن" مفاصل موجود در "پاها" نقش بیشتری را نسبت به سایر مفاصل بدن ایفا می‌کنند و در مقابل برای حرکت "نوشتن" مفاصل موجود در "دست‌ها" اطلاعات بهتری را ارائه می‌دهند. بنابراین چگونگی بهره‌برداری کامل از وابستگی‌های مکانی-زمانی مفاصل بدن یکی از مشکلات مهم در روش‌های تشخیص حرکت با استفاده از داده‌های اسکلتی است.

با الهام از مشاهدات فوق، ما یک شبکه کانولوشن گرافی مبتنی بر توجه به نام AT-AR را پیاده‌سازی کردیم که بخش‌های اصلی آن در زیر خلاصه می‌شود:

• کار ما بر اساس لایه‌های SPG-Net [۲۸] پایه‌ریزی شده است که از گراف پارامتری برای یادگیری روابط وزن‌ها استفاده می‌کند و هزینه محاسباتی را نیز کاهش می‌دهد.

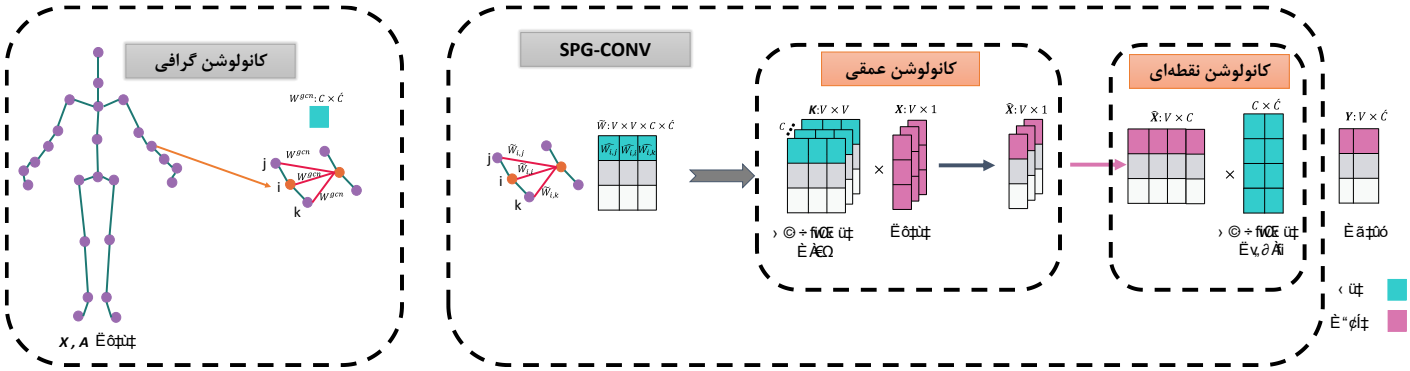
• شبکه AT-AR از معماری دو جریانی و سلسله‌مراتبی برای ذخیره ویژگی‌های اسکلتی استفاده می‌کند.

• ما از یک مکانیزم خودتوجه[۹] برای انتخاب رئوس مهم در گراف اسکلتی استفاده می‌کنیم و از یک مکانیزم توجه مکانی-زمانی[۱۰] استفاده می‌کنیم که امتیاز توجه را با استفاده از ویژگی‌های ایجاد می‌کند و همبستگی‌های درون فریمی و برون فریمی را به دست می‌آورد.

۲- پیش‌زمینه

۲-۱- ساخت گراف

برای تشخیص حرکت مبتنی بر داده‌های اسکلتی، هر اسکلت بدن انسان به عنوان یک گراف نشان داده می‌شود که نمایش سلسله‌مراتبی دنباله‌های اسکلتی را تشکیل می‌دهد. به طور کلی، برای یک اسکلت با N مفصل، گراف اسکلتی به صورت G = (V,E) تعریف می‌شود که $V = \{v_1, \dots, v_n\}$ مجموعه‌ی مفاصل و E مجموعه یال‌ها است که اتصالات بین مفاصل بدن (استخوان‌ها) می‌باشند. گراف اسکلتی را می‌توان با ماتریس مجاورت $A \in R^{N \times N}$ نشان داد که مقادیر موجود در آن اعداد ۰ و ۱ هستند که



شکل ۱: کانولوشن گرافی مبتنی بر پارامتر تفکیک پذیر و GCN برای مدلسازی داده‌های اسکلتی [۲۸]. به طور کلی تفاوت عمده بین SPG-Conv و کانولوشن گرافی معمولی این است که در کانولوشن گرافی معمولی هنگام مدلسازی فعل و انفعالات مختلف از وزن اشتراکی استفاده می‌شود اما در کانولوشن SPG-Conv از وزن‌های اختصاصی استفاده می شود. در شکل فوق تفاوت این دو گراف نشان داده شده است. اگر یک گراف محلی با سه راس را در نظر بگیرید، کانولوشن گرافی معمولی وزن مشترکی را برای همه یال‌ها اعمال می‌کند که در شکل فوق با W^{gcn} قابل مشاهده است، در حالیکه در SPG-Conv برای هر یال یک وزن خصوصی (برای مثال $\widehat{W}_{i,i}$ و $\widehat{W}_{i,k}$) تعیین می‌شود.

نشان‌دهنده اتصال یا عدم اتصال، مفاصل v_i و v_j به یکدیگر است. با توجه به توالی اسکلتی، مجموعه‌ای از ویژگی‌های $X=\{x_{t,n} \mid 1 \le t \le T\ ,\ 1 \le n \le N\ ;\ n, t \in Z\}$ می‌تواند به عنوان یک تنسور ویژگی $X \in \mathbb{R}^{C \times T \times V}$ نشان داده شود، که در آن C بیانگر بعد داده‌ها (۲-بعدی یا ۳-بعدی)، T نشان‌دهنده‌ی تعداد فریم‌های اسکلتی در کل توالی حرکت و V نیز نشان‌دهنده‌ی تعداد مفاصل اسکلت بدن است.

۲-۲- کانولوشن گرافی

برای پیاده‌سازی کانولوشن گرافی بر روی اسکلت انسان، مفاصل درون یک فریم به عنوان ماتریس مجاورت A نشان داده می‌شوند. عملیات پس انتشار در کانولوشن گرافی از لایه‌ی $H^{(l)}$ به لایه‌ی $H^{(l+1)}$ به صورت زیر تعریف می‌شود [۲۹]:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-.5}\,\tilde{A}\,\tilde{D}^{-.5}\,H^{(l)}\,W^{(l)}\right) \qquad (۱)$$

که در آن σ یک تابع فعالساز غیرخطی مانند ReLU است. $H^{(l)}$ ماتریس ویژگی لایه‌ی قبل و $W^{(l)}$ ماتریس وزن مرتبط با لایه‌ی قبل است. $\tilde{A}$، ماتریس مجاورتی است که با استفاده از ماتریس همانی I به صورت زیر بدست می‌آید:

$$\tilde{A} = A + I \qquad (۲)$$

$\tilde{D}$ نیز ماتریس درجه‌ی متناسب با ماتریس مجاورت $\tilde{A}$ است و به صورت زیر بیان می‌شود:

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{i,j} \qquad (۳)$$

که در آن $\tilde{A}_{i,j}$ درایه متناظر با راس i و j ماتریس $\tilde{A}$ می‌باشد.

یک تقریب مرتبه‌ی اول از کانولوشن گرافی به صورت زیر ارائه می‌شود [۲۹]:

$$Y = \tilde{A}\, X\, W^{GCN} \qquad (۴)$$

که در آن $X \in R^{V \times C}$ و $Y \in R^{V \times \acute{C}}$، به ترتیب نشان‌دهنده‌ی ورودی و خروجی نقشه‌های ویژگی کانولوشن گرافی هستند و W^{GCN} وزن‌های اشتراکی کانولوشن گرافی برای تمام مفاصل اسکلتی موجود در گراف اسکلتی بدن انسان است. $\tilde{A} \in R^{V \times V}$ نیز ماتریس مجاورت مربوط به گراف اسکلتی به صورت نرمال است.

یان و همکاران [۲۷] نیز یک کانولوشن گرافی برای داده‌های اسکلتی ارائه داده است، که به صورت زیر می‌باشد:

$$Y = \sum_{s=1}^{S} (\tilde{A}_s \odot M_s)\, X\, W_s^{ST\text{-}GCN} \qquad (۵)$$

که در آن S، تعداد زیرمجموعه‌های ایجاد شده با استفاده از استراتژی پارتیشن‌بندی موجود در ST-GCN است. M_s، برابر با s-امین زیرمجموعه‌ی قابل آموزش برای یادگیری ویژگی‌های یال‌های گراف اسکلتی است، W_s وزن اشتراکی زیر مجموعه s است، X بردار ویژگی ورودی و ⊙، نیز یک ضرب عنصری است.

از آنجایی تعداد مفاصل موجود در هر فریم ثابت است، می‌توان به جای وزن اشتراکی موجود در معادله (۵) از وزن خصوصی برای

مقابله با فعل‌وانفعالات مختلف مفاصل استفاده کرد که باعث ایجاد کانولوشن گرافی تفکیک‌پذیر مبتنی بر پارامتر[11] می‌شود [۲۸]. برای یک کانولوشن گرافیِ پارامتردار (PG-Conv)، وزن اشتراکی $W^{GCN}\in R^{C\times \acute{C}}$ موجود در معادله (۴) با وزن خصوصی $W_{i,j}$ جایگزین می‌شود. این کار باعث ذخیره‌ی فعل و انفعالات بین مفصل i - ام و j - ام می‌شود. در واقع معادله (۴) به صورت زیر تغییر می‌کند [۲۸]:

$$Y_i = \sum_{j=1}^{V} \tilde{A}_{i,j}\, X_j\, W_{i,j}$$
$$= \sum_{j=1}^{V} X_j\, [\tilde{A}_{i,j}\ \ W_{i,j}] = \sum_{j=1}^{V} X_j\, \widetilde{W}_{i,j} = X * \widetilde{W}_i \qquad (۶)$$

که در آن *، عملگر کانولوشن گرافی پارامتردار است و $\widetilde{W}_i\in R^{1\times V\times C\times \acute{C}}$، فعل و انفعالات بین مفصل i – ام و j – ام را بدست می‌آورد. بنابراین، خروجی نقشه‌ی ویژگی نهایی کانولوشن گرافیِ پارامتردار به صورت زیر بدست می‌آید:

$$Y = X * \widetilde{W} = X * [\,\widetilde{A} \odot W\,] \qquad (۷)$$

کانولوشن گرافی پارامتردار را می‌توان به یک کانولوشن تفکیک‌پذیر تبدیل کرد. بنابراین این کانولوشن، به دو کانولوشن عمقی[12] و نقطه‌ای[13] تقسیم می‌شود. برای هر کانال ورودیِ C، فیلتری با اندازه V×V در کانولوشن عمقی اعمال می‌شود. در این صورت خروجی نقشه‌ی ویژگی به صورت زیر بدست می‌آید:

$$\breve{X}_C = (\widetilde{A} \odot K_C)\, X_C \qquad (۸)$$

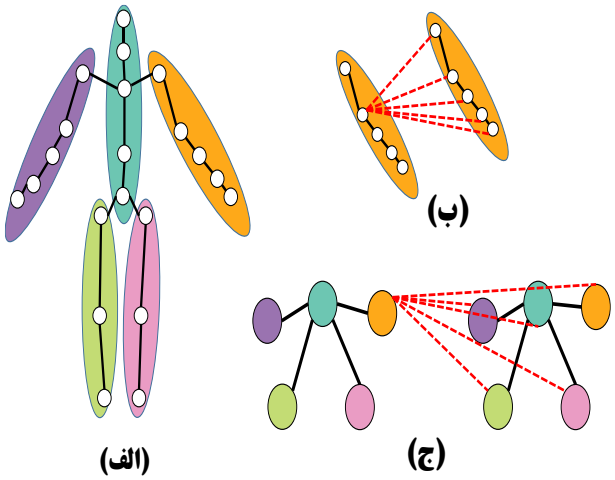
در این معادله، $X_C\in R^{V\times 1}$ و $\breve{X}_C\in R^{V\times 1}$ کانال‌های ورودی و خروجی ویژگی‌ها هستند و $K_C\in R^{V\times V}$، C-امین فیلتر می‌باشد. سپس یک کانولوشن (نقطه‌ای) با اندازه‌ی 1×1 روی خروجی کانولوشن عمقی (معادله ۷) اعمال می‌شود، تا خروجی آن به بعد $\acute{C}$ گسترش یابد. بدین صورت یک کانولوشن SPG-Conv بوجود می‌آید که خروجی آن با خروجی کانولوشن پارامتردار برابر است.

شکل (۱) تفاوت یک کانولوشن گرافی تفکیک‌پذیر مبتنی بر پارامتر و یک کانولوشن گرافی را نشان می‌دهد.

۲-۳- گروه‌بندی مفاصل اسکلتی

برای استخراج ویژگی‌های معنایی – حرکتیِ سطح بالا، ابتدا تمامی مفاصل موجود در گراف اسکلتی بدن را به ۵ گروه مختلف تقسیم می‌کنیم، که عبارت‌اند از: "سر و تنه"، "بازوی چپ"، "بازوی راست"، "پای چپ" و "پای راست". در مرحله‌ی بعد ویژگی‌ها و اطلاعات میان این گروه‌ها توسط مدل یاد گرفته می‌شود تا تشخیص حرکت با دقت بالاتری انجام پذیرد. شکل (۲) عملیات تقسیم مفاصل و ارتباط بین گروه‌های مفصلی را نشان می‌دهد.

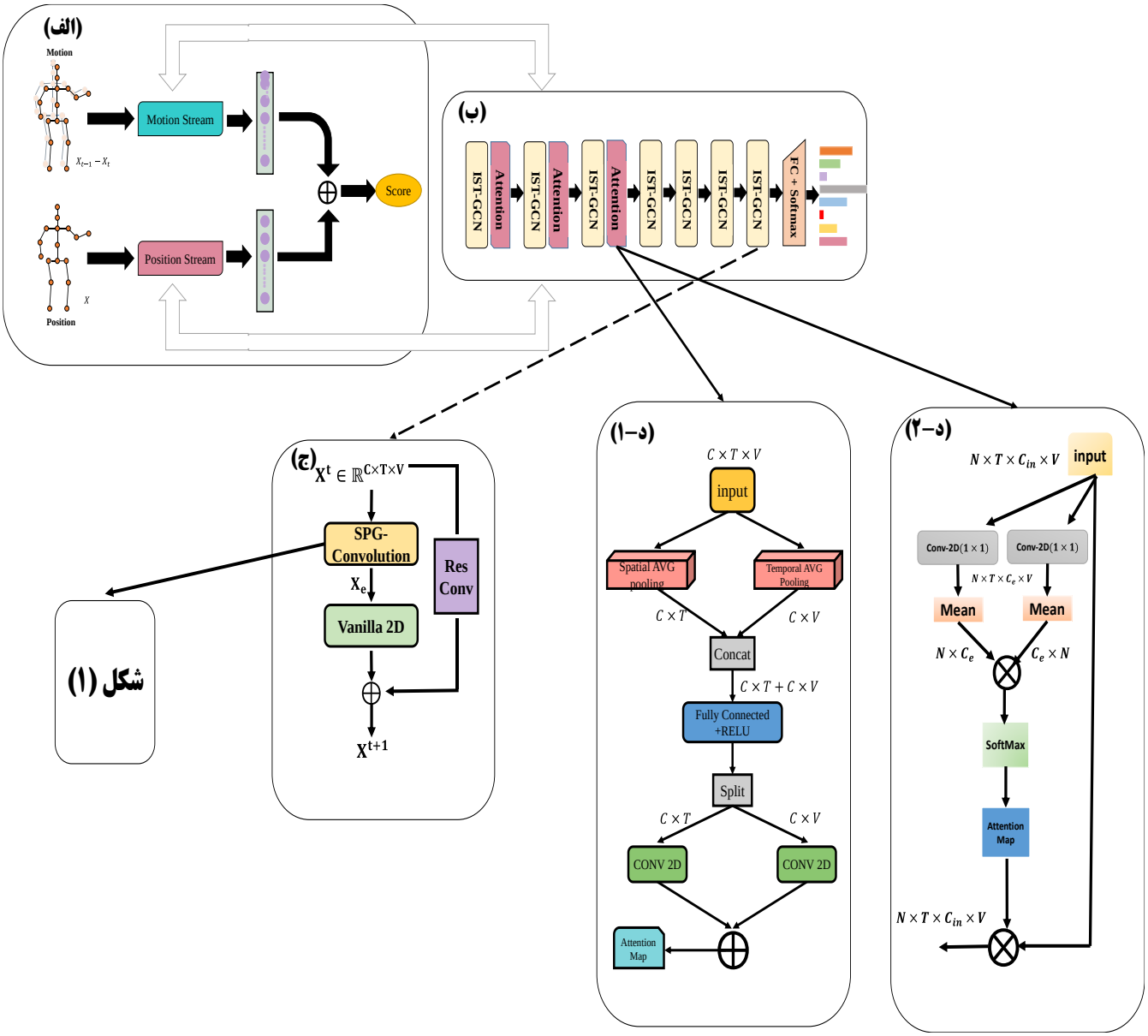
ابتدا از یک ماتریس مجاورت A^T برای تقسیم‌بندی مفاصل استفاده می‌کنیم، بطوریکه درایه‌های این ماتریس احتمال تعلق داشتن یک مفصل به یکی از پنج گروه مفاصل اسکلتی را نشان می‌دهد. اگر تعداد مفاصل گراف اسکلتی برابر v باشد، آنگاه $A^T \in \mathbb{R}^{5\times v}$. سپس اطلاعات گروه‌ها و ماتریس مجاورت A^T مستقیماً وارد SPG-Conv و معادله‌ی (۸) می‌شوند تا اطلاعات بین گروه‌ها یاد گرفته شوند.



شکل۲: تقسیم‌بندی مفاصل اسکلتی بدن به ۵ گروه. گروه‌بندی مفاصل بدن انسان. برای گروه‌بندی، ۲۵ مفصل بدن انسان به ۵ گروه تقسیم می‌شوند. در شکل۲-الف این ۵ گروه با رنگ‌های مختلف قابل مشاهده است. در این مقاله از دو نوع ارتباط درون گروهی و برون گروهی استفاده می‌شود. شکل۲-ب ارتباطات درون گروهی متعلق به گروه بازوی چپ را نشان می‌دهد که در آن ارتباط یک مفصل با سایر مفاصل به دست آورده می‌شود. شکل ۲-ج نیز ارتباط برون گروهی را نشان می‌دهد که در آن ارتباط گروه بازوی چپ با سایر گروه‌ها به دست می‌آید.

۲-۴- بلوک کانولوشن مکانی-زمانی بهبود یافته (IST-GCN)

در بلوک اصلی مدل پیشنهادی ابتدا از یک لایه SPG-Conv بر روی ویژگی اسکلتی $X \in R^{C\times T\times V}$ استفاده می‌کنیم. سپس یک کانولوشن زمانی دوبعدی با اندازه فیلتر ۹×۱ به دنبال آن بکار می‌گیریم تا روابط زمانی بین فریم‌های مجاور در نظر گرفته شود. در مرحله آخر نیز از یک بلوک اتصال ثابت برای اعمال ویژگی‌های ورودی ازدست‌رفته استفاده می‌کنیم. شکل (۳-ج) بلوک کانولوشن مکانی-زمانی و اجزاء آن را نشان می‌دهد.



شکل (۱)

شکل۳: نمای کلی شبکه پیشنهادی را نشان می‌دهد. در قسمت الف، شبکه AT-AR را با دو جریان ورودی نشان می‌دهد که شکل (ب) هر جریان ورودی، شامل ۷ بلوک IST-GCN و سه بلوک توجه را نشان می‌دهد. در قسمت (ج) نیز بلاک کلی IST-GCN و جزئیات آن را نشان می‌دهد که شامل یک کانولوشن Vanilla دو بعدی و یک کانولوشن پارامتریک تفکیک پذیر(SPG-Conv) است که جزئیات SPG-Conv نیز به طور کامل در شکل (۱) نشان داده شده است. همچنین بلوک های توجه به کار رفته در این معماری نیز در قسمت د-۱ و د-۲ قابل مشاهده است. در واقع مدل پیشنهادی یکبار با مکانیزم توجه مکانی-زمانی (د-۱) و یکبار با مکانیزم خودتوجهی (د-۲) پیاده سازی شده است.

۳- مدل پیشنهادی

۳-۱- مکانیزم خودتوجهی

ورودی یک مکانیزم توجه شامل مجموعه‌ای از بردارهای query (Q)، Key (K) با بعد C و Value (V) است و خروجی آن به صورت زیر حاصل می‌شود [۳۰]:

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{C}}\right) \quad (۹)$$

در این معادله هنگامی که مقادیر ورودی یکسان باشند، به آن مکانیزم خود توجهی گفته می‌شود.

برای یک توالی اسکلتی با ورودی $C \times T \times V$ ماژول توجهی به صورت زیر پیاده سازی کردیم:

$$x_{out(t,n)} = \sum_{t+1}^{T}\sum_{n+1}^{N} \text{SoftMax}\left(\frac{W_{tn,t+1n+1}}{\sqrt{C_e}}\right) x_{in(t+1,n+1)} \quad (۱۰)$$

که در این معادله $W_{tn,t+1n+1}$، وزن توجه مرتبط با مفصل $x_{(t+1,n+1)}$ و $x_{(t,n)}$ است که به صورت زیر محاسبه می‌شود:

$$W_{tn,t+1n+1} = f_1(x_{(t,n)}) f_2(x_{(t+1,n+1)})^T \quad (۱۱)$$

که f_1 و f_2 دو تابع تعبیه هستند، به طوری‌که $f : C_{in} \rightarrow C_e$ و C_{in} و C_e به ترتیب کانال‌های ورودی و تعبیه هستند. در شکل (۳:د-۲) نمای کلی این ماژول توجه، قابل مشاهده است.

۳-۲- توجه مکانی-زمانی

بلوک‌های IST-GCN[۱۴] موجود در مدل پیشنهادی، تنها از یک فیلتر زمانی ثابت ۹×۱ استفاده می‌کند که در مقابله با حرکات مختلف ناکارآمد است. از طرفی بلوک IST-GCN تنها همبستگی‌های زمانی بین فریم‌های متوالی را درنظر می‌گیرد و از همبستگی‌های بین فریم‌های نامجاور غافل می‌شود. برای رفع این محدودیت، از یک ماژول توجه مکانی – زمانی استفاده می‌کنیم تا اطلاعات بین تمامی فریم‌های حرکتی را در نظر گیرد. شکل (۳:د-۱) نمای کلی این مکانیسم توجه را نشان می‌دهد.

در مکانیسم توجه پیشنهادی، ابتدا توالی اسکلتی ورودی با اندازه‌ی $C \times T \times V$ دریافت می‌شود و سپس دو لایه‌ی کاهش اندازه میانگین[۱۵] در سطح فریم و مفصل روی داده‌ی ورودی اعمال می‌شود. سپس این دو بردار ویژگی کاهش‌یافته با هم ترکیب می‌شوند و از یک لایه‌ی کاملاًمتصل عبور داده می‌شوند. در مرحله بعد، از دولایه‌ی کاملاً متصل[۱۶] مجزا استفاده می‌شود تا دو مجموعه‌ی امتیاز توجه[۱۷] برای بعد مفاصل و بعد فریم به‌دست آید. سپس خروجی این دو لایه در هم ضرب می‌شوند و خروجی کلی ماژول امتیاز توجه، برای کل توالی حرکتی می‌باشد. درواقع این ماژول توجه را می‌توان به‌صورت زیر نمایش داد:

$$f_{out} = sigmoid(f_e.W_1) \otimes sigmoid(f_e.W_2) \quad (۱۲)$$

و f_e به‌صورت زیر به دست می‌آید:

$$f_e = ReLU(GAP_{frame}(f_{in}) + GAP_{joint}(f_{in})).W \quad (۱۳)$$

در این معادلات f_{in} و f_{out} ورودی و خروجی نقشه‌های ویژگی هستند و GAP_{frame} و GAP_{joint} به ترتیب لایه کاهش اندازه برحسب میانگین بر روی بعد فریم و مفصل می‌باشند. $W, W_1, W_2 \in R^{C \times C}$، پارامترهای یادگیری هستند و $\otimes$ بیانگر ضرب خارجی می‌باشد.

۳-۳- معماری شبکه

روش پیشنهادی ما از یک معماری دوجریانه تبعیت می‌کند که در آن هر جریان ورودی شامل هفت لایه کانولوشن گرافی IST-GCN و بلوک‌های توجه معرفی شده در بخش‌های قبلی هستند. هر جریان ورودی نیز ویژگی‌های اسکلتی را به صورت سلسله‌مراتبی می‌آموزد.

جریان موقعیت[۱۸]، اطلاعات مکانی مفاصل را دریافت می‌کند، در حالیکه جریان حرکت[۱۹] برای مدلسازی تغییرات زمانی برای فریم‌های متوالی طراحی شده است. جریان موقعیت، مختصات توالی X را وارد می‌کند و جریان حرکت، تغییر مکان مفاصل اسکلتی است که می‌تواند توسط دو فریم متوالی یعنی، X_{t+1}-X_t، محاسبه شود. در نهایت امتیازات خروجی دو جریان از طریق لایه تماماً متصل و سافت‌مکس عبورداده می‌شود، تا حرکت تشخیص داده شود. شکل (۳-ب) نمای کلی یک جریان ورودی شبکه را نشان می‌دهد.

کانال‌های خروجی به ترتیب دارای اندازه‌های ۶۴، ۶۴، ۶۴، ۱۲۸، ۱۲۸، ۲۵۶ و ۲۵۶ می‌باشند. به طور کلی، سه بلوک اولیه لایه‌های IST-GCN با ماتریس مجاورت A است، که اطلاعات مرتبط با گراف اسکلتی را استخراج می‌کند و پس از هر کدام، یک ماژول توجه مکانی-زمانی استفاده شده است. بلوک چهارم تعداد مفاصل را از بعد V=۲۵ (تعداد مفاصل بدن) به U=۵ (تعداد گروه‌های مفاصل) کاهش می‌دهد. بلوک پنجم رابطه‌های درون گروهی

مفاصل را می‌آموزد و دو بلوک نهایی برای یادگیری رابطه بین گروه‌ها استفاده می‌شود.

با استفاده از ترکیب دو جریان ورودی، مدلی به نام AT-AR ایجاد می‌شود که فرایند تشخیص حرکت را با استفاده از گراف اسکلتی انجام می‌دهد. شکل (۳-الف) طرح کلی شبکه پیشنهادی AT-AR را نشان می‌دهد، که از خروجی دو جریان موقعیت و حرکت بدست می‌آید.

۴- آزمایش‌ها

۴-۱- مجموعه‌داده

در این تحقیق از مجموعه داده NTU RGB+D [۳۵] برای ارزیابی شبکه پیشنهادی استفاده می‌کنیم. NTU-RGB+D، یک مجموعه داده در مقیاس بزرگ برای تشخیص حرکت است، که از ۵۶۸۸۰ نمونه تشکیل شده است. به طور کلی، این مجموعه داده از چهار نوع داده‌ی مختلف تشکیل شده است و ما فقط از داده‌های اسکلتی موجود در آن که شامل موقعیت سه بعدی ۲۵ مفصل بدن انسان است، استفاده می‌کنیم.

مجموعه داده NTU-RGB+D در مجموع دارای ۶۰ کلاس حرکتی است که به سه دسته‌ی کلی تقسیم می‌شوند: ۴۰ حرکت معمولی روزانه (نوشتن، خواندن و ...)، ۹ حرکت مرتبط با سلامت بدن (عطسه کردن، درد گردن و...) و ۱۱ حرکت واکنش‌گرا (مشت زدن، هل دادن و...).

مجموعه داده‌ی NTU RGB+D دارای دو معیار ارزیابی هنگام تقسیم داده‌ها به مجموعه‌های آموزش و آزمایش می‌باشد، که عبارتند از: Cross Subject (CS) و Cross View (CV) .

در ارزیابی با استفاده از CS، ۴۰ فردی که برای انجام حرکات انتخاب شدند، به دو گروه آموزش و آزمایش تقسیم می‌شوند و هر گروه از ۲۰ نفر تشکیل می‌شود. در ارزیابی به روش CV، تمامی نمونه‌های ضبط شده از دوربین‌های ۲ و ۳ به عنوان نمونه‌های آموزش انتخاب می‌شوند و نمونه‌های ضبط شده از دوربین ۱ به عنوان نمونه‌های آزمایش انتخاب می‌شوند.

۴-۲- ارزیابی مدل

در مدل پیشنهادی، اندازه‌ی توالی اسکلتی ورودی را برابر ۲۰۰ در نظر می‌گیریم. در روند آموزش شبکه، از الگوریتم گرادیان کاهشی تصادفی در ۱۰۰ دوره[۲۰] استفاده می‌کنیم تا آموزش شبکه بهینه شود. مقدار نرخ یادگیری اولیه[۲۱] را برابر ۰/۲، نرخ حذف تصادفی[۲۲] برابر ۰/۴ و اندازه دسته[۲۳] را برابر ۱۲۸ در نظر می‌گیریم. برای ارزیابی مدل، از مکانیزم‌های توجه موجود در بخش‌های ۳-۱ و ۳-۲ به‌صورت جداگانه استفاده می‌کنیم.

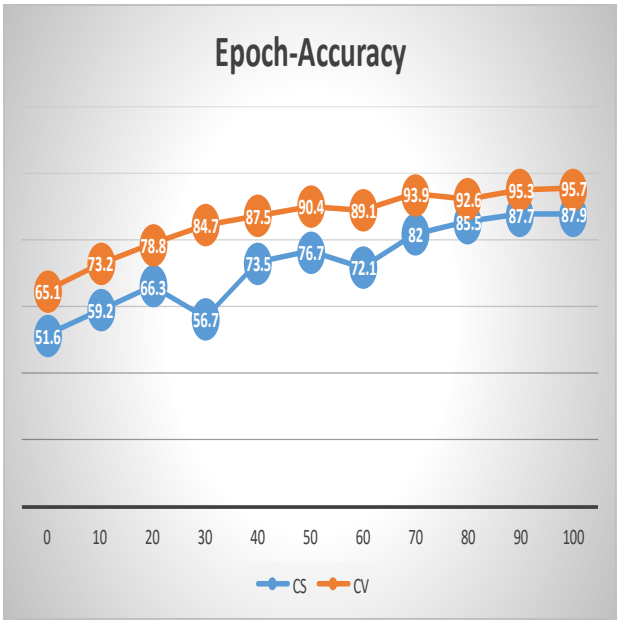
برای بررسی شبکه AT-AR تعاریف زیر را دنبال می‌کنیم:

۱- SPG-Net : شبکه پایه مورد استفاده در این مقاله است که از وزن اختصاصی بین دو مفصل به جای وزن اشتراکی استفاده می‌کند.

۲- Self AT-AR : از معماری دو جریانی و ۷ بلوک IST-GCN استفاده می‌کند و با استفاده از مکانیزم Self-Attention اطلاعات مفاصل مهم حرکتی را به دست می‌آورد.

۳- ST AT-AR : از معماری دو جریانی و ۷ بلوک IST-GCN استفاده می‌کند و با استفاده از مکانیزم توجه مکانی-زمانی روابط بین فریم‌های نامجاور را به دست می‌آورد.

برای ارزیابی مدل از دو معیار ارزیابی CV و CS استفاده می‌کنیم. در شکل (۴) همگرایی شبکه ST AT-AR در طول ۱۰۰ دوره آموزشی و با استفاده از معیارهای ارزیابی CV و CS نشان داده شده است و همانطور که مشاهده می‌شود، مدل از نودمین دوره به بعد همگرا شده است.



شکل ۴: دقت مدل ST AT-AR در ۱۰۰ دوره.

در جدول (۱) به مقایسه مدل پیشنهادی و مدل پایه‌ای (SPG-Net) پرداختیم و مشاهده می‌کنیم که مدل ما نسبت به مدل پایه‌ای از دقت بیشتری برخوردار است، به‌طوری که در حالت دوجریانی، شبکه ST AT-AR به ترتیب ۱/۴ و ۱/۳ درصد تحت معیار ارزیابی CV و CS از شبکه پایه‌ای بهتر عمل کرده است.

شبکه Self AT-AR نیز در حالت دو جریانی به ترتیب به میزان ۰/۸ و ۰/۶ تحت ارزیابی‌های CV و CS از شبکه پایه‌ای بهتر عمل کرده است و از دقت بیشتری برخوردار است.

همانطور که در جدول (۱) مشخص است، دقت شبکه در حالت استفاده از جریان موقعیت بهتر از حالت استفاده از حرکتی موقعیت است و استفاده همزمان از این دو جریان باعث بهبود دقت کلی می‌شود.

جدول ۱: دقت مدل پایه‌ای و مدل پیشنهادی و تعداد پارامترهای آن‌ها.

Model	Configuration	CS	CV	#Params
SPG-Net	Two-Stream	۸۶/۶	۹۴/۳	۰/۸۴۸
	Position	۸۴/۱	۹۲/۴	۰/۴۲۴
	Motion	۸۳/۸	۹۱/۸	۰/۴۲۴
Self AT-AR	Two-Stream	۹۲/۸	۹۶/۵	۰/۸۹۰
	Position	۹۰/۸	۹۴/۲	۰/۴۴۵
	Motion	۹۱/۵	۹۴/۴	۰/۴۴۵
ST AT-AR	Two-Stream	۹۳/۷	۹۷/۴	۰/۹
	Position	۹۲/۱	۹۵/۸	۰/۴۵
	Motion	۹۱/۹	۹۵/۱	۰/۴۵

جدول (۲) به مقایسه مدل AT-AR و مدل‌های قبلی کانولوشن گرافی برای تشخیص حرکت مبتنی بر داده‌های اسکلتی می‌پردازد. همانطورکه مشخص است مدل AT-AR به دلیل استفاده از کانولوشن تفکیک‌پذیر از تعداد پارامترهای کمتری استفاده می‌کند و لذا سرعت آموزش این مدل از مدل‌های قبلی به مراتب بهتر است.

جدول۲: دقت مدل پیشنهادی و مدل‌های پیشین.

model	CV (%)	CS (%)	#Parameter
ST-GCN [۲۷]	۸۸/۳	۸۱/۶	۳/۱
2s-AGCN [۳۱]	۹۵/۱	۸۸/۵	۶/۹۴
RA-GCN [۳۲]	۹۳/۵	۸۵/۵	۱۰/۱
AGC-LSTM [۳۳]	۹۵	۸۹/۲	۲۲/۸۹
MS-G3D [۳۴]	۹۴/۷	۸۷/۲	۶/۴
STC-NET [۳۶]	۹۶/۷	۹۲/۵	-
HD-GCN[۳۷]	۹۷/۲	۹۳/۴	-
Fusing + AGE [۳۸]	۹۴/۵	۸۸/۷	-
TD-GCN [۳۹]	۹۶/۸	۹۲/۸	-
Hypersa [۴۰]	۹۶/۵	۹۲/۹	-
Self AT-AR (ours)	۹۶/۵	۹۲/۸	۰/۸۹۰
ST AT-AR (ours)	۹۷/۴	۹۳/۷	۰/۹

جدول (۳) نیز زمان مورد نیاز برای آموزش مدل در هر دوره را با استفاده از یک گرافیک RTX 2080 Ti نشان می‌دهد. در این جدول زمان بر حسب تعداد نمونه‌های آموزشی محاسبه شده است و می‌توان فهمید که هر چه تعداد نمونه‌های داده‌های اسکلتی بیشتر باشد شبکه به زمان بیشتری برای آموزش داده‌ها نیاز دارد.

جدول ۳: میزان زمان آموزش شبکه بر حسب نمونه‌های ورودی.

Number of sample	Model	Time (Hour)
۱۵۰۰۰	ST AT-AR	۰/۱۳
	Self AT-AR	۰/۱
۴۰۰۰۰	ST AT-AR	۰/۳۵
	Self AT-AR	۰/۳
۵۶۸۸۰	ST AT-AR	۰/۴۲
	Self AT-AR	۰/۳۸

۵- نتیجه‌گیری

از آنجایی که تشخیص حرکت یک شاخه مهم در تحقیقات هوش مصنوعی می‌باشد و روش‌های پیشین مبتنی بر داده‌های اسکلتی از اطلاعات بین فریم‌های نامجاور غافل می‌شدند لذا در این مقاله یک مدل تشخیص حرکت مبتنی بر داده‌های اسکلتی با استفاده از مکانیسم توجه و شبکه‌های کانولوشن گرافی ارائه دادیم که اطلاعات بیشتری را نسبت به روش‌های پیشین به دست می‌آورد. شبکه پیشنهادی ما از یک مدل دوجریانی پیروی می‌کند، که در هر جریان از چندین بلوک کانولوشن‌گرافی و ماژول توجه استفاده می‌کند؛ تا همبستگی‌های مکانی بین مفاصل و همبستگی‌های زمانی بین فریم‌های مجاور و نامجاور برای یک حرکت یاد گرفته شوند. استفاده از مکانیزم توجه منجر به بهبود دقت شبکه پیشنهادی شده است. از آنجایی که شبکه پیشنهادی ما از کانولوشن تفکیک‌پذیر استفاده می‌کند و شبکه‌های کانولوشنی تفکیک‌پذیر به نسبت شبکه‌های کانولوشنی معمولی از پیچیدگی محاسباتی کمتری برخوردارند، لذا مدل پیشنهادی نسبت به سایر مدل‌های پیشین از تعداد پارامترهای کمتری استفاده می‌کند و این امر منجر به سریع‌تر شدن اجرای کدهای شبکه می‌شود. همچنین شبکه ما به دلیل استفاده از دو ماژول توجه، نسبت به مدل پایه‌ای SPG-Net از دقت بیشتری برخوردار است.

لازم به ذکر است کدهای شبیه‌سازی شبکه AT_AR در پاورقی قابل مشاهده می باشد.[۲۴]

مراجع

[1] P. Yin, J. Ye, G. Lin, and Q. Wu, "Graph neural network for 6D object pose estimation," *Knowl.-Based Syst.*, vol. 218, Art. no. 106839, 2021, doi: 10.1016/j.knosys.2021.106839.

[2] D. Kong, Y. Bao, and W. Chen, "Collaborative learning based on centroid distance-vector for wearable devices," *Knowl.-Based*

[19] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton-based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Hong Kong, China, Jul. 2017, pp. 601–604, doi: 10.1109/ICMEW.2017.8026282.

[20] T. S. Kim and A. Reiter, "Interpretable 3-D human action analysis with temporal convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Snowbird, UT, USA, Jun. 2017, pp. 1623–1631, doi: 10.1109/CVPRW.2017.207.

[21] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view-invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017, doi: 10.1016/j.patcog.2017.03.029.

[22] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3-D human action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2016, pp. 816–833, doi: 10.1007/978-3-319-46487-9_50.

[23] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1012–1020, doi: 10.1109/ICCV.2017.115.

[24] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2117–2126, doi: 10.1109/ICCV.2017.229.

[25] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4800–4814, Sep. 2022, doi: 10.1109/TNNLS.2021.3061115.

[26] W. Li, L. Wen, M.-C. Chang, S.-N. Lim, and S. Lyu, "Adaptive RNN tree for large-scale human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1444–1452, doi: 10.1109/ICCV.2017.159.

[27] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, New Orleans, LA, USA, Feb. 2018, pp. 7444–7452.

[28] M. Wang, B. Ni, and X. Yang, "Learning multi-view interactional skeleton graph for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6940–6954, Jun. 2023, doi: 10.1109/TPAMI.2020.3032738.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017.

[30] S. Cho, M. Maqbool, F. Liu, and H. Foroosh, "Self-attention network for skeleton-based human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass Village, CO, USA, Jan. 2020, pp. 635–644, doi: 10.1109/WACV45572.2020.9093439.

[31] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12026–12035, doi: 10.1109/CVPR.2019.01230.

[32] Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 1–5, doi: 10.1109/ICIP.2019.8802917.

[33] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1227–1236, doi: 10.1109/CVPR.2019.00132.

Syst., vol. 194, Art. no. 105569, 2020, doi: 10.1016/j.knosys.2020.105569.

[3] M. Zhang, G. Tian, Y. Zhang, and P. Duan, "Service skill improvement for home robots: autonomous generation of action sequence based on reinforcement learning," *Knowl.-Based Syst.*, vol. 212, Art. no. 106605, 2021, doi: 10.1016/j.knosys.2020.106605.

[4] I. S. MacKenzie, *Human-Computer Interaction: An Empirical Research Perspective*, 2nd ed. Amsterdam, The Netherlands: Morgan Kaufmann, 2024, doi: 10.1016/C2020-0-02009-6.

[5] X. Guo, S. Guo, C. Wu, J. Li, C. Liu, and W. Chen, "Intelligent monitoring for safety-enhanced lithium-ion/sodium-ion batteries," *Adv. Energy Mater.*, vol. 13, no. 10, Art. no. 2300268, 2023, doi: 10.1002/aenm.202300268.

[6] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6202–6211, doi: 10.1109/ICCV.2019.00630.

[7] S. Baek, Z. Shi, M. Kawade, and T.-K. Kim, "Kinematic-layout-aware random forests for depth-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1468–1477, doi: 10.1109/CVPR.2016.160.

[8] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based CNNs for infrared action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 927–931, Jun. 2018, doi: 10.1109/LSP.2018.2823654.

[9] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2117–2126, doi: 10.1109/ICCV.2017.229.

[10] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Apr. 2012, doi: 10.1109/MMUL.2012.24.

[11] ASUS, "Xtion PRO LIVE," 2011. [Online]. Available: https://www.asus.com/3D-Sensor/Xtion_PRO/. [Accessed: 03-May-2025].

[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," arXiv preprint arXiv:1812.08008, Dec. 2018.

[13] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 7157–7173, Aug. 2022, doi: 10.1109/TPAMI.2022.3222784.

[14] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowl.-Based Syst.*, vol. 239, Art. no. 108146, Dec. 2022, doi: 10.1016/j.knosys.2022.108146.

[15] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Underst.*, vol. 158, pp. 85–105, Jan. 2017, doi: 10.1016/j.cviu.2017.01.011.

[16] R. Hou and Z. Wang, "Self-attention based anchor proposal for skeleton-based action recognition," arXiv preprint arXiv:2112.09413, Dec. 2021, doi: 10.48550/arXiv.2112.09413.

[17] Y. Du, Y. Fu, and L. Wang, "Skeleton-based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Kuala Lumpur, Malaysia, Nov. 2015, pp. 579–583, doi: 10.1109/ACPR.2015.7486569.

[18] H. Liu, J. Tu, and M. Liu, "Two-stream 3D convolutional neural network for skeleton-based action recognition," arXiv preprint arXiv:1705.08106, May 2017.

[34] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 140–149, doi: 10.1109/CVPR42600.2020.00023.

[35] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1010–1019, doi: 10.1109/CVPR.2016.114.

[36] J. Lee, M. Lee, S. Cho, S. Woo, S. Jang, and S. Lee, "Leveraging spatio-temporal dependency for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10255–10264, doi: 10.1109/ICCV51070.2023.00958.

[37] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10444–10453, doi: 10.1109/ICCV51070.2023.00958.

[38] Z. Qin, Y. Liu, P. Ji, D. Kim, L. Wang, R. I. McKay, S. Anwar, and T. Gedeon, "Fusing higher-order features in graph neural networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 4783–4797, Apr. 2024, doi: 10.1109/TNNLS.2022.3201518.

[39] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 811–823, May 2023, doi: 10.1109/TMM.2023.3271811.

[40] Y. Zhou, Z.-Q. Cheng, C. Li, Y. Fang, Y. Geng, X. Xie, and M. Keuper, "Hypergraph transformer for skeleton-based action recognition," arXiv preprint arXiv:2211.09590, Nov. 2022, doi: 10.48550/arXiv.2211.09590.

پاورقی‌ها:

[1] Human Action Recognition
[2] Human Pose Estimation Algorithms
[3,5] Convolutional Neural Networks
[4,6] Recurrent Neural Networks
[7]Long Short Term Memory
[8]Graph Convolution Neural Network
[9] Self Attention
[10] Spatio-Temporal Attention
[11] Separable Parametric Graph Convolution
[12] Depth-Wise Conv
[13] Point-Wise Conv
[14] Improved Spatio Temporal GCN
[15] Average Pooling
[16] Fully Connected
[17] Attention Score
[18] Position Stream
[19] Motion Stream
[20] Epoch
[21] Learning Rate
[22] Dropout
[23] Batch-Size
[24] https://github.com/malekisini/AT_AR-Network/tree/master