نوع مقاله: پژوهشی

یادگیری چندنمایی برای تشخیص حرکت مبتنی بر داده‌های اسکلتی با استفاده از شبکه دو جریانی مبتنی بر مکانیزم توجه

رضا روحانی سروستانی[۱]، استادیار ، محمد ملکی سینی[۲]، دانشجو

۱- گروه مهندسی کامپیوتر – دانشکده فنی و مهندسی - دانشگاه شهرکرد- شهرکرد - ایران- rrohani@sku.ac.ir
۲- گروه مهندسی کامپیوتر – دانشکده فنی و مهندسی - دانشگاه شهرکرد- شهرکرد - ایران- malekisinii@gmail.com

چکیده: با پیشرفت تکنولوژی و استفاده روز افزون از ماشین‌های هوشمند، سیستم‌های تشخیص حرکت انسانی به یک موضوع مهم در حوزه بینایی ماشین تبدیل شده‌اند. در سال‌های اخیر و به لطف ویژگی‌های مختصر و مفید داده‌های اسکلتی، روش‌های تشخیص حرکت مبتنی بر اسکلت با استفاده از شبکه‌های کانولوشن گرافی (GCN) عملکرد قابل توجهی را به دست آورده‌اند. در روش‌های قبلی، از کانولوشن‌های محلی یک بعدی برای استخراج روابط زمانی بین فریم‌های مجاور استفاده می‌شود و از همبستگی‌های بین فریم‌های نامجاور غافل می‌شوند. از طرفی حرکت‌های انسان، شامل تغییرات زمانی زیادی است که یک وابستگی زمانی قوی را بین حرکات مفاصل نشان می‌دهد. بنابراین تشخیص یک حرکت، مستلزم تحلیل جامعی از همبستگی‌های بین مفاصل در حوزه‌های مکانی (spatial) و زمانی (temporal) است. در این مقاله شبکه‌ای به نام MV AT-AR پیشنهاد شده است که با استفاده از مکانیزم توجه، همبستگی‌های بین مفاصل را در زمان‌های مختلف یاد می‌گیرد. معماری شبکه پیشنهادی از دو جریان ورودی استفاده می‌کند و با استفاده از گراف مکمل اسکلتی، خصوصیات مختلف اسکلت انسان را منعکس می‌کند، که باعث تشخیص حرکت با دقت بالا می‌شود. ارزیابی تجربی روی مجموعه داده NTU RGB+D نشان می‌دهد که شبکه پیشنهادی به دقت ۹۶/۷ درصدی دست یافته است.

واژه‌های کلیدی: تشخیص حرکت انسان، داده اسکلتی، شبکه کانولوشن گرافی، مکانیزم توجه، یادگیری عمیق.

Multi-View Learning for Skeleton based Action Recognition using Two-Stream Network based on Attention Mechanism

Mohammad Maleki Sini, Student[1], Reza Roohani Sarvestani, Assistant Professor[2]

1- Faculty of Computer Engineering, University of Shahrekord, Shahrekord, Iran, Email: malekisinii@gmail.com
2- Faculty of Computer Engineering, University of Shahrekord, Shahrekord, Iran, Email: rrohani@sku.ac.ir

Abstract With the advancement of technology and the increasing use of intelligent machines, Human Action Recognition systems have become an important topic in the field of machine vision. In recent years, thanks to the concise and useful features of skeletal data, skeleton-based Action Recognition methods using graph convolutional neural networks (GCNs) have achieved significant performance. In previous methods, one-dimensional local convolutions used to investigate temporal relationships between adjacent frames and neglect correlations between non-adjacent frames. On the other hand, human movements include many changes that show a strong dependency between joint movements. Therefore, the recognition of an action requires a comprehensive analysis of correlations between joints in the spatial and temporal domain. In this paper, we propose MV AT-AR networks that learn the correlations between joints at different times by using the attention mechanism. The proposed network architecture uses two input streams and reflects the different characteristics of the human skeleton by using the skeletal complement graph, which enables Action Recognition with high accuracy. Evaluation on the NTU RGB+D dataset shows that the proposed network achieves an accuracy of 96.7%.

Keywords: Human Action Recognition, Skeletal Data, Convolutional Neural Network, Attention Mechanism, Deep Learning.

تاریخ ارسال مقاله: ۱۴۰۱/۰۷/۱۱
تاریخ اصلاح مقاله: ۱۴۰۲/۰۹/۱۶
تاریخ پذیرش مقاله: ۱۴۰۳/۰۶/۱۰
نام نویسنده مسئول: رضا روحانی سروستانی
نشانی نویسنده مسئول: ایران - شهرکرد – دانشگاه شهرکرد – دانشکده فنی و مهندسی– گروه آموزشی مهندسی کامپیوتر.

۱- مقدمه

در سال‌های اخیر، تشخیص حرکت انسان به یک موضوع تحقیقاتی مهم در حوزه بینایی ماشین تبدیل شده است. سیستم‌های تشخیص حرکت انسان در برنامه‌های زیادی از جمله نظارت تصویری، تعامل انسان با کامپیوتر، واقعیت مجازی، مراقبت‌های بهداشتی و بازیابی ویدیو کاربرد دارند [۱، ۲].

به طور کلی، تشخیص حرکت انسان را می‌توان با داده‌های مختلفی مانند تصاویر رنگی، داده‌های عمق، جریان نوری و داده‌های اسکلتی انجام داد [۳، ۴]. داده‌های اسکلتی که موقعیت سه بعدی مفاصل بدن انسان را نشان می‌دهند، در مقایسه با سایر داده‌ها از مقاومت بیشتری در برابر تغییرات نمای دوربین‌ها، تغییرات ظاهری انسان و تغییرات نور پس‌زمینه برخوردار هستند [۵، ۶].

داده‌های اسکلتی معمولاً با مکان‌یابی مختصاتِ مکانی مفاصل، با حسگرهای عمق مانند کینکت [۷] به دست می‌آیند و یا با استفاده از الگوریتم های تخمین‌پوز بر اساس ویدیو ها استخراج می‌شوند. الگوریتم های تخمین‌پوز مانند Openpose [۸] یا Densepose [۹] مختصات مفاصل اسکلتی را از روی تصاویر دوبعدی تخمین می‌زنند و دستگاه‌هایی مانند کینکت نیز داده‌های اسکلتی را با استفاده از سنسورهای مختلف به‌صورت نقاط سه‌بعدی ثبت و ذخیره می‌کنند. طبق تحقیقات صورت گرفته در این زمینه، ثابت‌شده است که داده‌های اسکلتی که موقعیت‌های سه‌بعدی مفاصل بدن انسان را نشان می‌دهند، در نمایش حرکت، محاسبات و ذخیره‌سازی بسیار کارآمد هستند [۱۰]. ازاین‌رو در سال‌های اخیر، محققان زیادی به تشخیص حرکت با استفاده از داده‌های اسکلتی روی آورده‌اند.

رویکردهای اولیه تشخیص حرکت مبتنی بر اسکلت، هر مفصل بدن انسان را به‌عنوان یک واحد مستقل حامل ویژگی در نظر می‌گیرند و با روش‌های دست‌ساز (Handcraft)، همبستگی‌ها را مدل می‌کنند، پس از آن ویژگی‌های اسکلتی سطح بالای تولید شده، به طبقه‌بندهای سنتی مانند k همسایه نزدیک [۱۱] و ماشین بردار پشتیبان [۱۲] فرستاده می‌شوند تا دسته‌بندی حرکات را انجام دهند. در این روش‌ها، ویژگی‌ها بر اساس زوایای بین مفاصل [۱۳]، ویژگی‌های سینماتیکی [۱۴] و یا مسیرهای بین مفاصل [۱۵] طراحی می‌شوند.

با پیشرفت شبکه‌های یادگیری عمیق، شبکه‌های عصبی کانولوشن (CNN) و شبکه‌های عصبی بازگشتی (RNN) برای تشخیص حرکت مبتنی بر اسکلت به‌کار گرفته شدند. در روش‌های مبتنی بر شبکه‌های کانولوشن، داده‌های اسکلتی به‌عنوان شبه‌تصویر مدل‌سازی می‌شوند و هر مفصل به‌عنوان یک پیکسل تصویر در نظر گرفته می‌شود [۱۸-۱۶]. روش‌های مبتنی بر شبکه‌های عصبی بازگشتی عمدتاً از شبکه‌های حافظه کوتاه‌مدت (LSTM) برای مدل‌سازی دینامیک زمانی دنباله‌ای از اسکلت‌ها استفاده می‌کنند. این روش‌ها، هر اسکلت در یک توالی را به‌عنوان یک بردار نشان می‌دهند که توسط مختصات مفاصل به‌هم‌پیوسته تشکیل می‌شود [۲۱-۱۹]. از آنجایی که داده‌های اسکلتی در قالب گراف پیاده‌سازی می‌شوند، لذا روش‌های مبتنی بر CNN و RNN نمی‌توانند به‌طور کامل از ساختار غیراقلیدسی داده‌های اسکلتی بهره ببرند. پس از ایجاد شبکه‌های کانولوشن‌گرافی (GCN) [۲۲]، محققان زیادی از این شبکه برای تشخیص حرکت مبتنی بر اسکلت استفاده کردند.

GCN ها تعمیمی از شبکه‌های کانولوشن هستند که به جای داده‌های ساختاریافته، از داده‌های گرافی به عنوان ورودی استفاده می‌کنند. اخیراً این شبکه‌ها نتایج قابل توجهی را برای تشخیص حرکت مبتنی بر اسکلت به دست آورده‌اند. برای اولین بار شبکه‌ای بر اساس کانولوشن گرافی، به نام ST-GCN [۲۳] طراحی شد، که شامل شبکه کانولوشن‌گرافی و کانولوشن‌زمانی یک‌بعدی است. اولین جزء که کانولوشن‌گرافی است، برای استخراج ویژگی‌های مکانی اسکلتی استفاده می‌شود و دومین جزء، یعنی کانولوشن یک‌بعدی نیز برای یادگیری روابط زمانی بین مفاصل مربوطه در فریم‌های متوالی اعمال می‌شود. در سال‌های گذشته، کارهای زیادی بر اساس ساختار این شبکه طراحی شد.

تانگ و همکاران [۲۴] یک روش یادگیری تقویتی عمیق مبتنی بر GCN ارائه کردند که فریم‌های اسکلتی مهم را انتخاب می‌کند و فریم‌های مبهم را دور می‌اندازد. وِن و همکاران [۲۵] از کانولوشن‌گرافی برای رمزگذاری ساختار فضایی سلسله‌مراتبی اسکلت و یک بلوک متراکم زمانی متغیر برای بهره‌برداری از اطلاعات زمانی محلی در محدوده‌های مختلف دنباله‌های اسکلت انسانی استفاده کردند. شی و همکاران [۲۶] شبکه کانولوشن‌گرافی‌تطبیقی دوجریانی (2-AGCN) را برای یادگیری ارتباط بین مفاصل غیرمحلی برای وظایف مختلف پیشنهاد دادند. علاوه بر این 2-AGCN کارایی مدل را از طریق مدل‌سازی هم‌زمان داده‌های مفصلی و استخوان‌ها افزایش می‌دهد.

روش‌های مبتنی بر کانولوشن‌گرافی معرفی‌شده از دو محدودیت رنج می‌برند. این محدودیت‌ها باعث ایجاد مدلی غیرانعطاف‌پذیر برای تشخیص حرکت مبتنی بر اسکلت می‌شود.

اولین محدودیت این روش‌ها استفاده از وزن‌های انعطاف‌ناپذیر اشتراکی (Weight Sharing) است. از آنجایی که شبکه‌های کانولوشن گرافی تعمیمی از CNN ها هستند، لذا از یک فیلتر یکسان در یک لایه استفاده می‌کنند. استفاده از یک فیلتر یکسان، به دلیل اینکه مفاصل بدن از اهمیت نابرابری در حرکات مختلف برخوردارند، منجر به ایجاد شبکه‌ای غیر انعطاف‌پذیر می‌شود. برای برطرف کردن این مشکل، وانگ و همکاران [۲۷] نوعی کانولوشن‌گرافی به نام کانولوشن‌گرافی تفکیک‌پذیر مبتنی بر پارامتر (SPG-Conv) ارائه دادند که از وزن‌های اختصاصی بین دو مفصل به جای وزن اشتراکی استفاده می‌کند.

دومین محدودیت روش‌های ذکر شده نیز این است که تنها روابط زمانی بین فریم‌های مجاور را در نظر می‌گیرند و از روابط زمانی بین فریم‌های نامجاور غافل می‌شوند. به‌طورکلی یک حرکت معمولی از مراحل زمانی مختلف و مجزا تشکیل شده است. همبستگی بین چند مرحله‌ی زمانی کلیدی بافاصله‌ی زمانی مجاور یا نامجاور، به تشخیص

حرکت کمک می‌کند. برای مثال، حرکت "پریدن" شامل مراحل زمانی مختلفی مانند "شروع"، "دویدن"، "فرود آمدن" و "برخاستن" است و یادگیری همبستگی‌های بین این مراحل باعث بهبود در دقت تشخیص حرکت می‌شود. بنابراین در مدل‌سازی‌زمانی، همبستگی‌های بین مراحل زمانی کلیدی، به‌ویژه در فاصله زمانی غیرمجاور، باید به‌صراحت در یک روش انعطاف‌پذیر مورد بهره‌برداری قرار گیرد. یکی از راه‌حل‌های موجود برای حل این مشکل و تمرکز بر روابط بین فریم‌های نامجاور، استفاده از مکانیسم‌های توجه است.

در این مقاله، با استفاده از مکانیسم توجه و لایه‌های کانولوشن SPG-Conv [۲۷] مدلی برای تشخیص حرکت مبتنی بر اسکلت ارائه می‌شود که بر روابط بین فریم‌های نامجاور در یک حرکت متمرکز است و از وزن خصوصی بین دو مفصل برای تشخیص حرکت استفاده می‌کند.

۲- پیش‌زمینه

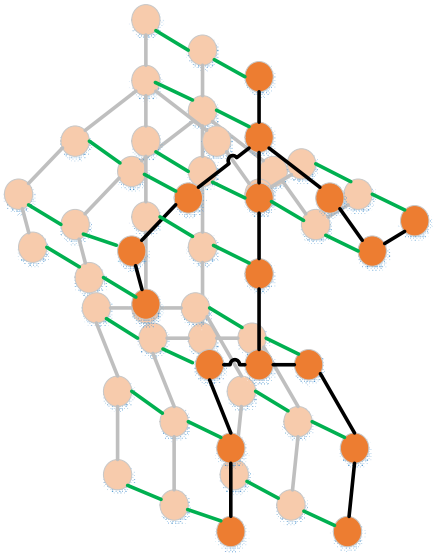
این بخش به بررسی ساختارهای مورد نیاز برای ایجاد مدل پیشنهادی می‌پردازد و این ساختارها را از جنبه‌های مختلف بررسی می‌کند.

۲-۱- ساخت گراف اسکلتی

هر اسکلت به‌عنوان یک گراف نشان داده می‌شود که نمایش سلسله مراتبی توالی‌های اسکلت را تشکیل می‌دهد [۲۸]. اسکلت مرتبط با بدن انسان در یک حرکت را می‌توان با گراف $G = (V, E)$ نشان داد، که V مجموعه‌ی مفاصل بدن و E مجموعه‌ی اتصالات بین مفاصل (استخوان‌های بدن) است. یک توالی از فریم‌های اسکلتی را می‌توان به‌عنوان یک تنسور ویژگی $X \in \mathbb{R}^{C\times T\times V}$ نشان داد، که در آن C بیانگر بعد داده‌ها (دوبعدی یا سه‌بعدی)، T نشان‌دهنده‌ی تعداد فریم‌های اسکلتی در کل توالی حرکت و V نیز نشان‌دهنده‌ی تعداد مفاصل بدن است. مجموعه مفاصل را می‌توان به‌صورت زیر نشان داد [۲۸]:

$$\{v_{ti} | t, i \in z, 1 \le t \le T, 1 \le i \le V\}$$

ساختار گراف اسکلتی را می‌توان با استفاده از ماتریس مجاورت $A \in \mathbb{R}^{V\times V}$ نشان داد، بطوریکه $A_{i,j} \in \{0,1\}$ بیانگر وجود یا عدم وجود یال بین مفصل i − ام و j − ام است. به طور کلی در این مقاله از گراف فیزیکی-حرکتی (Spatio-Temporal) استفاده می‌شود که دو مرحله ساخته می‌شود؛ ابتدا مفاصل درون فریم با استفاده از یال‌هایی منطبق بر فیزیک بدن انسان به هم متصل می‌شوند، سپس هر مفصل به خودش در فریم‌های متوالی متصل است. شکل ۱ گراف فیزیکی-حرکتی را نشان می‌دهد. این نوع گراف برای اولین بار در شبکه ST-GCN [۲۳] به کار گرفته شد.در این روش دو نوع یال وجود دارد: ۱- یال‌های فیزیکی که در آن مفاصل بر طبق استخوان‌های طبیعی بدن انسان به هم متصل می‌شوند. ۲- یال‌های زمانی یا حرکتی که در آن هر مفصل در مراحل زمانی متوالی دقیقاً به خودش متصل است. در این روش تنها اطلاعات زمانی هر فریم با فریم‌های مجاورش سنجیده می‌شود و از در نظر گرفتن اطلاعات بین فریم‌های نامجاور صرف نظر می‌شود. برای برطرف کردن این مشکل در این مقاله از مکانیزم توجه استفاده شده است که در بخش‌های آتی به آن پرداخته می‌شود.



شکل ۱: گراف فیزیکی-حرکتی که در آن نقاط نارنجی رنگ مفاصل بدن انسان را نشان می‌دهد و یال‌های گراف اسکلتی نیز نشان‌دهنده استخوان‌های طبیعی بدن انسان می‌باشند. هر مفصل نیز به خودش در فریم‌های بعدی متصل است و گراف حرکتی را می‌سازد. در این گراف تنها اطلاعات زمانی بین فریم‌های مجاور در نظر گرفته می‌شود و از درنظر گرفتن اطلاعات زمانی بین فریم‌های نامجاور صرف نظر می‌شود.

۲-۲- کانولوشن تفکیک‌پذیر

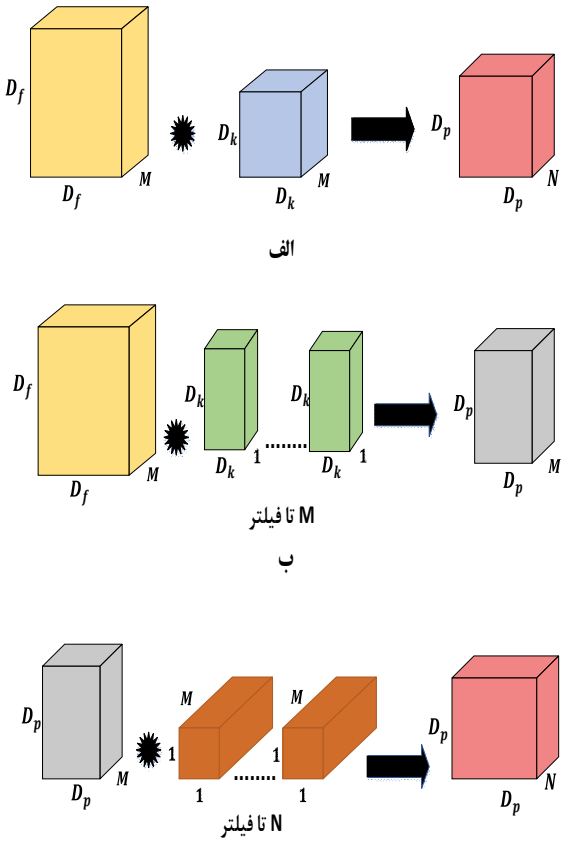
کانولوشن تفکیک پذیر برای اولین بار توسط گوگل در معماری Xception [۲۹] و MobileNet [۳۰] به کار گرفته شد. به‌طورکلی در یک کانولوشن تفکیک‌پذیر از دو عملیات کانولوشن با نام‌های کانولوشن عمقی (Depth-Wise) و کانولوشن‌نقطه‌ای (Point-Wise) برای کاهش تعداد پارامترهای شبکه کانولوشن استفاده می شود. کانولوشن عمقی، یک فیلتر را برای هر کانال اعمال می‌کند، درحالی‌که کانولوشن نقطه‌ای از یک لایه کانولوشن 1×1 برای ترکیب خروجی کانولوشن عمقی و تنظیم تعداد کانال‌های خروجی استفاده می‌کند. شکل ۲، به مقایسه کانولوشن استاندارد و کانولوشن تفکیک‌پذیر می‌پردازد.

با فرض اینکه اندازه ویژگی ورودی $D_f \times D_f \times M$ باشد و N عدد فیلتر با اندازه‌ی $D_k \times D_k \times M$ بر روی نقشه ویژگی ورودی اِعمال شود، خروجی کانولوشن استاندارد به‌صورت $D_p \times D_p \times N$ به‌دست می‌آید، که به تعداد $N \times D_p^2 \times D_k^2 \times M$ ضرب نیاز دارد.

در یک کانولوشن‌تفکیک‌پذیر، ابتدا M عدد فیلتر با اندازه $D_k \times D_k \times 1$ روی ورودی اعمال می‌شود و خروجی آن به‌صورت $D_p \times D_p \times M$ به‌دست می‌آید. از آنجایی که فیلتر به اندازه $D_p \times D_p$ بار در تمام

کانال‌های M جابه‌جا می‌شود، لذا خروجی کانولوشن عمقی به تعداد $D_p^2 \times D_k^2 \times M$ ضرب نیاز دارد. در کانولوشن نقطه‌ای نیز، تعداد N عدد فیلتر با اندازه $1 \times 1 \times M$ روی خروجی کانولوشن عمقی اعمال می‌شود، که به تعداد $N \times D_p^2 \times M$ ضرب نیاز است.

با جمع تعداد ضرب‌های انجام شده در کانولوشن‌عمقی و کانولوشن نقطه‌ای، پیچیدگی محاسباتی کانولوشن‌تفکیک‌پذیر به‌صورت $M \times D_k^2 \times D_p^2 + M \times D_p^2 \times N$ به‌دست می‌آید. همان‌طور که مشخص است خروجی کانولوشن‌استاندارد و تفکیک‌پذیر یکسان است، اما هزینه ی محاسباتی کانولوشن تفکیک‌پذیر بسیار کمتر است. جدول ۱ تعداد عملیات ضربی مورد استفاده در کانولوشن استاندارد و کانولوشن تفکیک‌پذیر را نشان می‌دهد. استفاده از کانولوشن‌تفکیک‌پذیر به کاهش اندازه‌ی پارامترهای مدل، کمک ویژه‌ای خواهد کرد.



شکل ۲: مقایسه کانولوشن استاندارد و کانولوشن تفکیک‌پذیر. شکل الف نمای کلی یک کانولوشن استاندارد را نشان می‌دهد. شکل ب و ج نیز دو مرحله کلی کانولوشن تفکیک پذیر را نشان می‌دهد که در شکل ب یک کانولوشن عمقی با تعداد M فیلتر و در شکل ج یک کانولوشن نقطه ای با N فیلتر نشان داده شده است. در نهایت مشاهده می‌شود که خروجی کانولوشن استاندارد و نقطه‌ای برابر است.

جدول۱: مقایسه کانولوشن استاندارد و کانولوشن‌تفکیک‌پذیر

نوع کانولوشن	تعداد عملیات ضربی مورد استفاده
کانولوشن استاندارد	$N \times D_p^2 \times D_k^2 \times M$
کانولوشن تفکیک‌پذیر	$M \times D_p^2 \times (D_k^2 + N)$

بر طبق جدول ۱ نسبت عملیات ضربی مورد استفاده در کانولوشن تفکیک‌پذیر به عملیات ضربی مورد استفاده در کانولوشن استاندارد به صورت زیر به دست می‌آید:

$$Ratio = \frac{M \times D_p^2 \times (D_k^2 + N)}{N \times D_p^2 \times D_k^2 \times M} = \frac{1}{N} + \frac{1}{D_k^2} \qquad (۱)$$

۲-۳- کانولوشن گرافی

برای پیاده‌سازی کانولوشن گرافی بر روی اسکلت انسان، مفاصل درون یک فریم به‌عنوان ماتریس مجاورت A نشان داده می‌شوند. عملیات پس انتشار در کانولوشن‌گرافی از لایه‌ی $H^{(l)}$ به لایه‌ی $H^{(l+1)}$ به‌صورت زیر تعریف می‌شود [۲۲]:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-.5}\,\tilde{A}\,\tilde{D}^{-.5}\,H^{(l)}\,W^{(l)}\right) \qquad (۲)$$

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 به‌صورت زیر به‌دست می‌آید:

$$\tilde{A} = A + I \qquad (۳)$$

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

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{i,j} \qquad (۴)$$

یک تقریب مرتبه‌ی اول از کانولوشن گرافی به‌صورت زیر ارائه می‌شود [۲۲]:

$$Y = \tilde{A}\,X\,W^{GCN} \qquad (۵)$$

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 به‌صورت نرمال است.

یان و همکاران [۲۳] نیز یک کانولوشن گرافی برای داده‌های اسکلتی ارائه داده است، که به‌صورت زیر می‌باشد:

$$Y = \sum_{s=1}^{S} (\tilde{A}_s \odot M_s)\,X\,W_s^{ST-GCN} \qquad (۶)$$

که در آن S، تعداد زیرمجموعه‌های ایجادشده با استفاده از استراتژی پارتیشن‌بندی موجود در ST-GCN [۲۳] است. M_s، برابر با s-امین زیرمجموعه‌ی قابل‌آموزش برای یادگیری ویژگی‌های یال‌های گراف اسکلتی است و ⊙، نیز یک ضرب عنصری است.

از آنجایی تعداد مفاصل موجود در هر فریم ثابت است، می‌توان به‌جای وزن اشتراکی موجود در معادله ۵ از وزن خصوصی برای مقابله با فعل‌وانفعالات مختلف مفاصل استفاده کرد که باعث ایجاد کانولوشن گرافی مبتنی بر پارامتر تفکیک‌پذیر می‌شود [۲۷]. برای یک کانولوشن گرافی پارامتر دار (PG-Conv)، وزن اشتراکی $W^{GCN} \in R^{C\times\acute{C}}$ موجود در معادله‌ی ۵ با وزن خصوصی $W_{i,j}$ جایگزین می‌شود. این کار باعث ذخیره‌ی فعل‌وانفعالات بین مفصل i - ام و j - ام می‌شود. درواقع معادله ۵ به‌صورت زیر تغییر می‌کند [۲۷]:

$$Y_i = \sum_{V_j} \frac{1}{Z_{i,j}} X_j\, W_{i,j} = \sum_{j=1}^{V} \tilde{A}_{i,j}\, X_j\, W_{i,j} = \sum_{j=1}^{V} X_j\, [\tilde{A}_{i,j}\, W_{i,j}] = \sum_{j=1}^{V} X_j\, \widetilde{W}_{i,j} = X * \widetilde{W}_i \quad (۷)$$

که در آن *، عملگر کانولوشن گرافی‌پارامتردار است و $\widetilde{W}_i \in R^{1\times V\times C\times\acute{C}}$، فعل‌وانفعالات بین مفصل i − ام و j − ام را به دست می‌آورد. بنابراین، خروجی نقشه‌ی ویژگی نهایی کانولوشن گرافی پارامتردار به‌صورت زیر به دست می‌آید:

$$Y = X * \widetilde{W} = X * [\tilde{A} \odot W] \quad (۸)$$

کانولوشن گرافی پارامتردار را می‌توان به یک کانولوشن تفکیک‌پذیر تبدیل کرد. بنابراین این کانولوشن، به دو کانولوشن عمقی و نقطه‌ای تقسیم می‌شود. برای هر کانال ورودی C، فیلتری با اندازه $V \times V$ در کانولوشن عمقی اعمال می‌شود. در این صورت خروجی نقشه‌ی ویژگی به‌صورت زیر به دست می‌آید:

$$\check{X}_C = (\tilde{A} \odot K_C)\, X_C \quad (۹)$$

در این معادله، $X_C \in R^{V\times 1}$ و $\check{X}_C \in R^{V\times 1}$ کانال‌های ورودی و خروجی ویژگی‌ها هستند و $K_C \in R^{V\times V}$، C- امین فیلتر می‌باشد. سپس یک کانولوشن (نقطه‌ای) با اندازه‌ی ۱×۱ روی خروجی کانولوشن عمقی (معادله ۸) اعمال می‌شود، تا خروجی آن به بعد $\acute{C}$ گسترش یابد. بدین‌صورت یک کانولوشن SPG-Conv به وجود می‌آید که خروجی آن با خروجی کانولوشن پارامتردار برابر است.

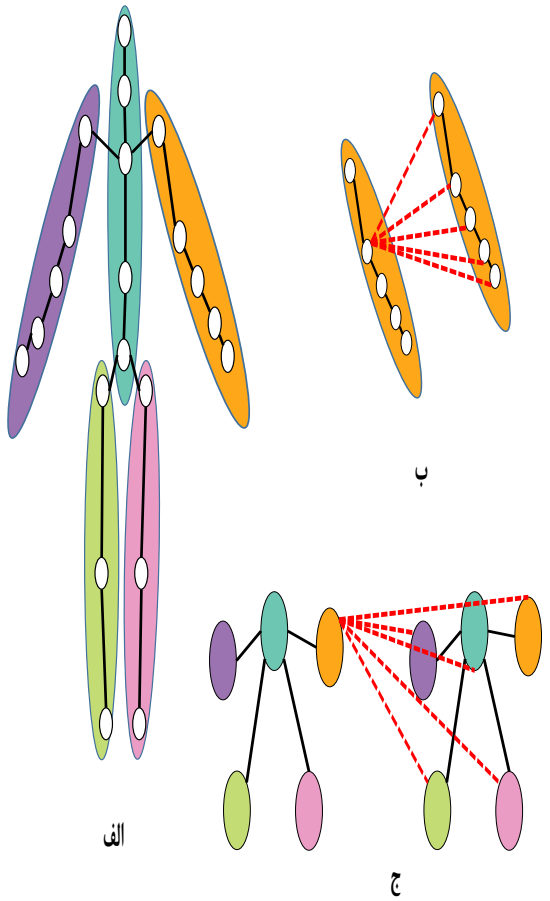
۳- مدل پیشنهادی

در این قسمت به بررسی مدل پیشنهادی و قسمت‌های مختلف آن می‌پردازیم.

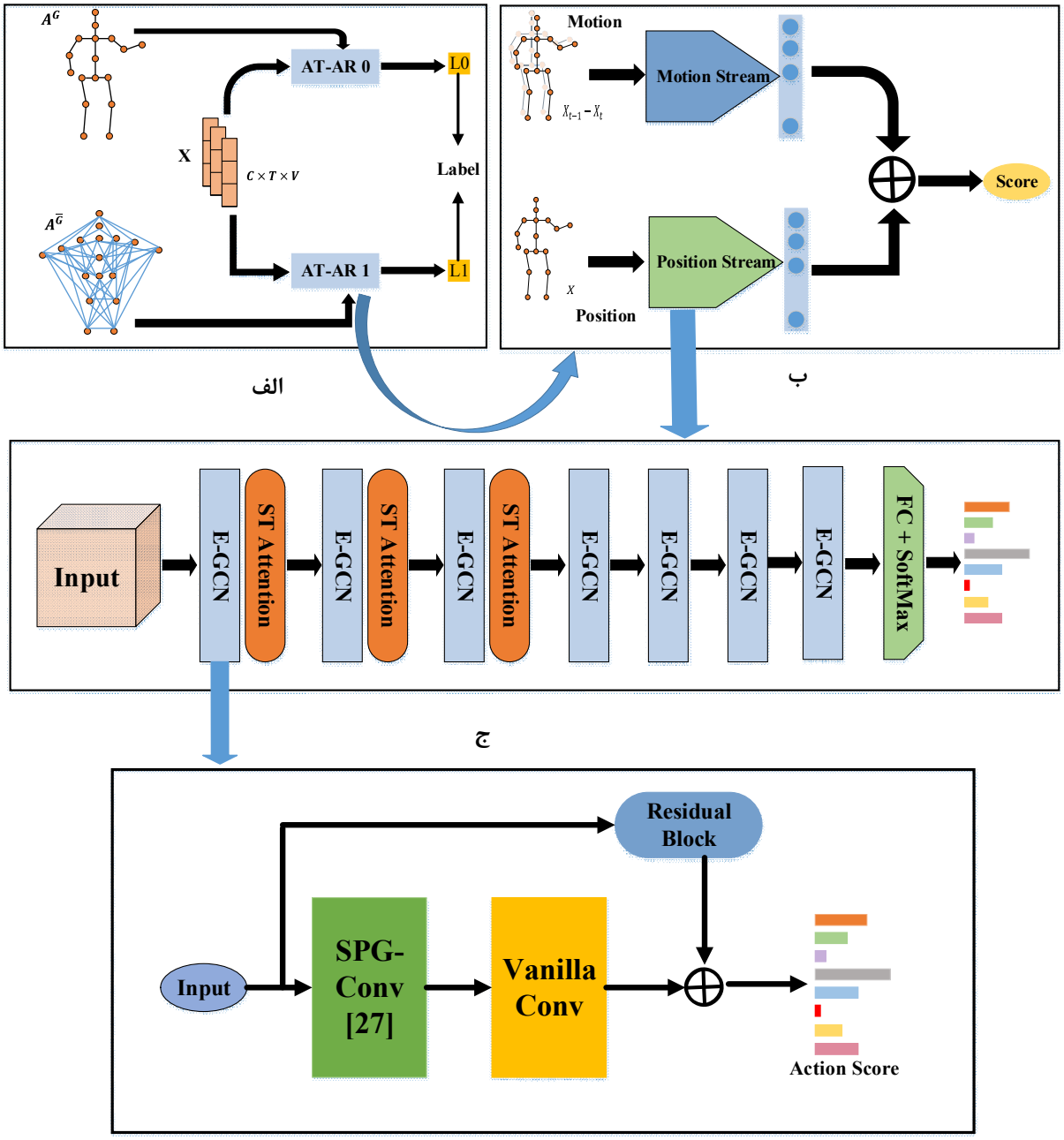
۳-۱- گروه‌بندی مفاصل

برای استخراج ویژگی‌های معنایی – حرکتی سطح بالا، ابتدا تمامی مفاصل موجود در گراف اسکلتی بدن را به ۵ گروه مختلف تقسیم می‌شود، 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 ۳ عملیات گروه‌بندی مفاصل و ارتباط بین گروه‌های مفصلی را نشان می‌دهد.

ابتدا از یک ماتریس مجاورت A^T برای تقسیم‌بندی مفاصل استفاده می‌شود، بطوریکه درایه‌های این ماتریس احتمال تعلق داشتن یک مفصل به یک گروه 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 ۹ می‌شوند تا اطلاعات بین گروه‌ها یاد گرفته شوند.



شکل۳:گروه‌بندی مفاصل بدن انسان. برای گروه‌بندی، ۲۵ مفصل بدن انسان به ۵ گروه تقسیم می‌شوند. در شکل۳-الف این ۵ گروه با رنگ‌های مختلف قابل مشاهده است. در این مقاله از دو نوع ارتباط درون گروهی و برون گروهی استفاده می‌شود. شکل ۳-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 ۳-ج نیز ارتباط برون گروهی را نشان می‌دهد که در آن ارتباط گروه بازوی چپ با سایر گروه‌ها به دست می‌آید.



شکل ۴: نمای کلی مدل پیشنهادی: شکل (الف) نمای شبکه MV AT-AR را نشان می‌دهد؛ شبکه MV AT-AR از دو شبکه AT-AR تشکیل شده است که به صورت مستقل از هم آموزش می‌بینند. ورودی شبکه AT-AR0 گراف اصلی اسکلتی و ورودی شبکه AT-AR1 گراف مکمل اسکلتی است. هر کدام از مدل‌های AT-AR شامل دو جریان ورودی position و motion هستند که در شکل (ب) وجود دارد. هر کدام از این جریان‌های موجود در شکل (ب) شامل ۷ بلوک E-GCN و سه بلوک توجه می‌باشد که در شکل (ج) وجود دارد. بلوک‌های E-GCN از دو بلوک SPG-[۲۷] Conv و Vanilla-Conv تشکیل شده است که در شکل (د) وجود دارد و بلوک ST Attention مورد استفاده در شبکه نیز در شکل (۵) قابل مشاهده است.

۳-۲- بلوک اصلی مدل (E-GCN)

برای بلوک اصلی مدل پیشنهادی ابتدا از یک لایه SPG-Conv بر روی ویژگی اسکلتی $X \in R^{C\times T\times V}$ استفاده می‌شود. سپس یک کانولوشن زمانی دوبعدی با اندازه فیلتر ۹×۱ به دنبال آن بکار گرفته می‌شود تا روابط زمانی بین فریم‌های مجاور در نظر گرفته شود. درنهایت از یک بلوک اتصال ثابت (Residual Connection) نیز برای اعمال ویژگی‌های ورودی ازدست‌رفته استفاده خواهد شد. شکل (۴- د) بلوک اصلی مدل پیشنهادی را نشان می‌دهد.

۳-۳- مکانیزم توجه زمانی – مکانی

بلوک‌های E-GCN موجود در مدل پیشنهادی، تنها از یک فیلتر زمانی ثابت ۹×۱ استفاده می‌کند که در مقابله با حرکات مختلف ناکارآمد می‌باشد. از طرفی بلوک E-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 (Spatio-Temporal Attention) استفاده می‌شود تا اطلاعات بین تمامی فریم‌های حرکتی را در نظر گیرد. شکل ۵ نمای کلی این مکانیزم توجه را نشان می‌دهد.

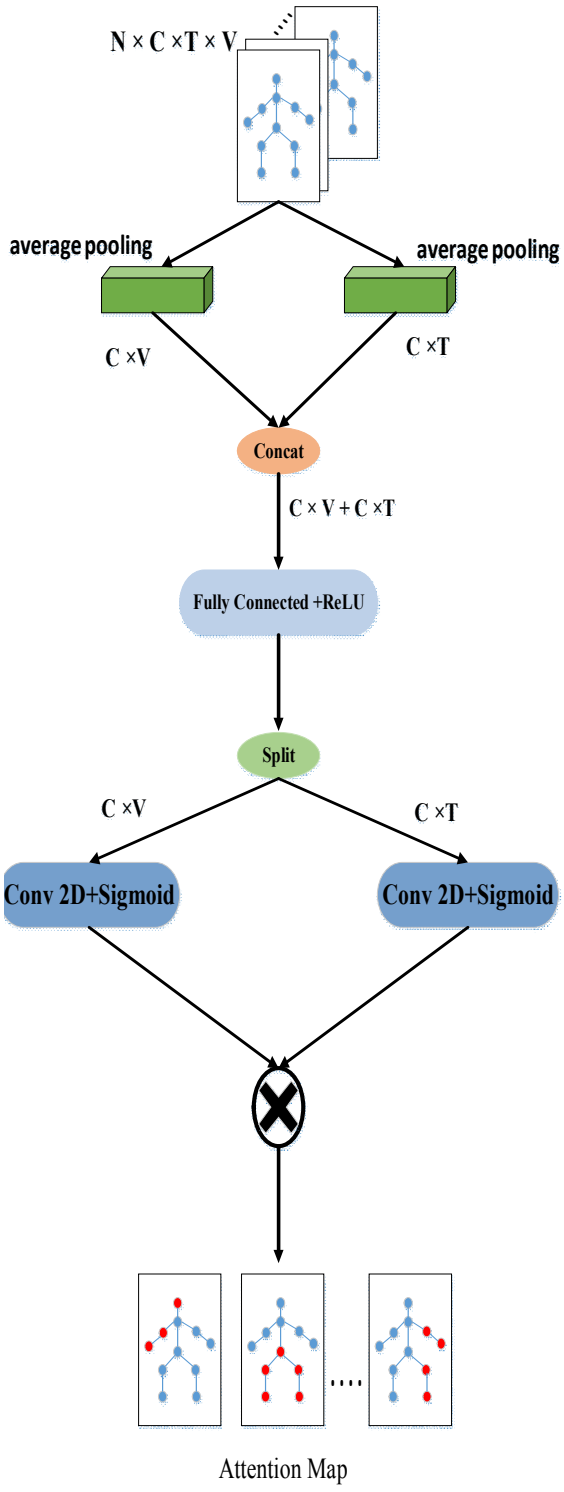
در مکانیزم توجه پیشنهادی، ابتدا توالی اسکلتی ورودی با اندازه‌ی $C \times T \times V$ دریافت می‌شود و سپس دو لایه‌ی کاهش اندازه میانگین (average pooling) در سطح فریم و مفصل روی داده‌ی ورودی اعمال می‌شود. سپس این دو بردار ویژگی کاهش‌یافته با هم ترکیب می‌شوند و از یک لایه‌ی کاملاًمتصل عبور داده می‌شوند. در مرحله بعد، از دولایه‌ی کاملاً متصل مجزا استفاده می‌شود تا دو مجموعه‌ی امتیاز توجه (attention score) 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:

$$f_{out} = sigmoid\,(f_e.W_1) \otimes sigmoid\,(f_e.W_2) \qquad (۱۰)$$

و f_e به‌صورت زیر به دست می‌آید:

$$f_e = \sigma(GAP_{frame}(f_{in}) + GAP_{joint}(f_{in})).W \qquad (۱۱)$$

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. σ تابع فعال‌ساز ReLU است. $W, W_1, W_2 \in R^{C\times C}$، پارامترهای یادگیری هستند و $\otimes$ بیانگر ضرب می‌باشد.



شکل ۵: بلوک توجه مکانی - زمانی

۳-۴- معماری دو جریانی

در این تحقیق از یک معماری دو جریانی استفاده می‌شود، که هر جریان از هفت بلوک کانولوشنی E-GCN و یک بلوک توجه مکانی – زمانی

تشکیل شده است. هر جریان ویژگی‌های اسکلتی را به‌صورت سلسله مراتبی یاد می‌گیرد.

جریان موقعیت (position)، اطلاعات مکانی مفاصل را دریافت می‌کند، درحالی‌که جریان حرکت (motion) 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 لایه‌های تماماً متصل و سافت‌مکس عبور داده می‌شود، تا حرکت تشخیص داده شود. شکل (۴-ج) نمای کلی یک جریان ورودی شبکه را نشان می‌دهد.

کانال‌های خروجی به ترتیب دارای اندازه‌های ۶۴، ۶۴، ۶۴، ۱۲۸، ۱۲۸، ۲۵۶ و ۲۵۶ می‌باشند. به‌طورکلی، سه بلوک اولیه لایه‌های E-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 که کانال‌های خروجی با اندازه ۶۴ دارند. بلوک چهارم تعداد مفاصل را از بعد ۲۵ (تعداد مفاصل بدن) به بعد 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 (۴-ب) طرح کلی شبکه پیشنهادی AT-AR را نشان می‌دهد، که از خروجی دو جریان موقعیت و حرکت به دست می‌آید.

۳-۵- یادگیری چند نمایی داده‌های اسکلتی

یادگیری چند نمایی یک چارچوب یادگیری ماشین است که در آن داده‌ها توسط چندین گروه ویژگی مجزا نشان داده می‌شوند و هر گروه ویژگی به عنوان یک نمای خاص شناخته می‌شود. یک ویژگی کلی برای گراف‌ها بدین‌صورت است که مکمل یک گراف حاوی اطلاعات تکمیلی درباره‌ی آن گراف است [۳۱]. مشابه ایزومرها در شیمی، دو گراف اسکلتی که مفاصل یکسان ولی توپولوژی‌های مختلفی دارند، می‌توانند خصوصیات مختلف حرکت انسان را منعکس کنند [۲۷]. بنابراین یک راه حل برای تشخیص حرکت چند نمایی با داده‌های اسکلتی، استفاده از توپولوژی‌های مختلف گرافی برای ایجاد ارتباطات تکمیلی نماهای مختلف است.

در این مقاله برای تشخیص حرکت از یک گراف اسکلتی و مکمل آن استفاده می‌شود تا ویژگی‌های مکمل گراف اسکلتی توسط مدل یاد گرفته شوند. استفاده از گراف اسکلتی و مکمل آن منجر به ایجاد شبکه‌ای به نام MV AT-AR می‌شود که در شکل (۴-الف) قابل مشاهده است.

اگر ماتریس مجاورت مرتبط با گراف اسکلتی بدن برابر A^G در نظر گرفته شود، آنگاه ماتریس مجاورت مرتبط با گراف اسکلتی مکمل به صورت زیر به دست می‌آید [۲۷]:

$$A^{\bar{G}} = 1 - A^G \qquad (۱۲)$$

که در آن $A^{\bar{G}}$ ماتریس مجاورت گراف مکمل است و دارای شرط زیر می‌باشد:

$$\forall A^{\bar{G}}_{i,j} \in \{0,1\} \qquad A^{\bar{G}}_{i,j} + A^{G}_{i,j} = 1 \qquad (۱۳)$$

هدف اصلی MV AT-AR استفاده از گراف‌های A^G و $A^{\bar{G}}$، برای استخراج ویژگی‌های مکمل از گراف‌های مختلف است. MV AT-AR یک مدل کلی متشکل از دو مدل AT-AR است که دو گراف فوق را به‌صورت مجزا دریافت می‌کنند و به‌صورت مستقل از هم نیز آموزش می‌بینند. هر مدل، داده‌های یکسان را از نظر موقعیت و حرکت به عنوان ورودی می‌گیرند و برای دسته‌بندی حرکات، برچسب‌های یکسان دریافت می‌کند. برای به دست آوردن دقت نهایی مدل در داده‌های آزمایش به سادگی از امتیازات پیش‌بینی شده دو مدل میانگین گرفته می‌شود.

۴- ارزیابی مدل

در این بخش به بررسی مجموعه داده مورد استفاده و نتایج ارزیابی مدل بر روی این مجموعه داده پرداخته می‌شود.

۴-۱- مجموعه داده

این مقاله از مجموعه داده NTU RGB+D [۳۲] برای ارزیابی شبکه پیشنهادی استفاده می‌کند. 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 فقط از داده‌های اسکلتی موجود در آن، که شامل موقعیت سه‌بعدی ۲۵ مفصل بدن انسان است، استفاده می‌شود.

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 عبارت‌اند 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 ضبط‌شده از دوربین‌های ۲ و ۳ به عنوان نمونه‌های آموزش انتخاب می‌شوند و نمونه‌های ضبط‌شده از دوربین ۱ به‌عنوان نمونه‌های آزمایش انتخاب می‌شوند.

۴-۲- آزمایش‌ها

در مدل پیشنهادی، اندازه‌ی توالی اسکلتی ورودی را برابر ۲۰۰ در نظر گرفته می‌شود. در روند آموزش شبکه از الگوریتم گرادیان کاهشی تصادفی در ۱۰۰ دوره (epoch) استفاده می‌شود تا آموزش شبکه بهینه شود. مقدار نرخ یادگیری اولیه برابر ۰/۰۲، Dropout برابر ۰/۴ و Batch-size برابر ۱۲۸ در نظر گرفته می‌شود. طبق شکل ۶ مشاهده می‌شود میزان دقت شبکه تحت ارزیابی Cross View بعد از دوره معینی تمایل به پایداری پیدا می‌کند.

جدول ۲ میزان دقت شبکه با استفاده از گراف اسکلتی و ترکیب آن با گراف مکمل را نشان می‌دهد. همچنین جدول ۲ و شکل ۷به مقایسه ی درصد دقت مدل پیشنهادی و مدل‌های قبلی می‌پردازد و نشان می‌دهد که مدل پیشنهادی از مدل‌های پیشین بهتر عمل کرده است. شبکه پیشنهادی در حالت استفاده از یک گراف، به دلیل در نظر گرفتن روابط زمانی بین فریم‌های نامجاور، تحت ارزیابی CV و CS به میزان ۰/۸ و ۰/۵ درصد بهتر از شبکه‌ی HPGNet [۲۷] پایه‌ای و هنگام استفاده از دو گراف مکمل و اصلی تحت ارزیابی CV و CS به میزان ۰/۹ و ۱/۳ درصد بهتر از شبکه‌ی MV HPGNet [۲۷] پایه‌ای عمل کرده است.

شبکه‌ی پیشنهادی به دلیل استفاده از لایه‌های کانولوشن تفکیک‌پذیر از تعداد پارامترهای کمتری استفاده می‌کند، لذا به محاسبات و منابع کمتری نیاز دارد. ستون آخر جدول ۲ نیز تعداد پارامترهای مدل پیشنهادی و سایر مدل‌ها را برحسب میلیون نشان می‌دهد. با یک مقایسه‌ی ساده بین مدل پیشنهادی در این مقاله و سایر مدل‌ها می‌توان فهمید که مدل AT-AR یک شبکه‌ی بهینه برای تشخیص حرکت با استفاده از داده‌های اسکلتی است، که با استفاده از مکانیسم توجه مکانی- زمانی روابط بین فریم‌های نامجاور را درنظر می‌گیرد و به دقت مطلوبی نیز دست می‌یابد. همچنین این مدل را از دو جنبه مختلف می‌توان بررسی کرد. زمانی که از گراف اسکلتی اصلی برای تشخیص حرکت استفاده شده شبکه به دقت ۹۵/۵ درصدی دست یافته که به میزان ۱/۲ درصد، دقت کمتری از حالت چندنمایی (MV AT-AR) دارد و این امر نشان می‌دهد که گراف اسکلتی مکمل، حاوی اطلاعات مفیدی است که می‌تواند به بهبود سیستم تشخیص حرکت کمک کند.

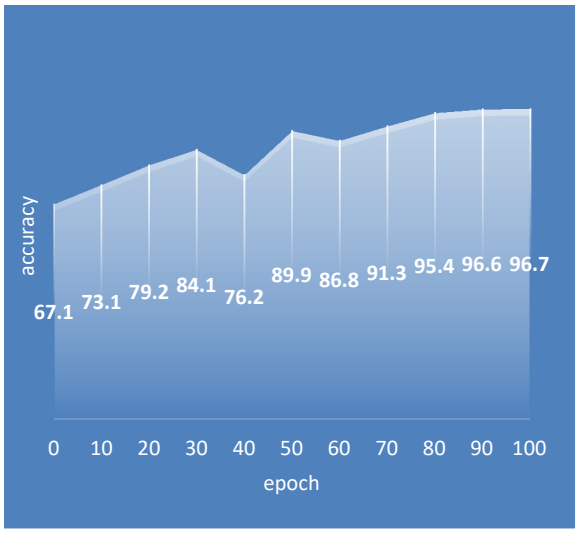
جدول ۲: مقایسه مدل پیشنهادی با سایر روش‌ها بر روی دیتاست NTU60.

model	CV (%)	CS (%)	#Parameter(M)
ST-GCN [23]	۸۸/۳	۸۱/۶	۳/۱
2s-AGCN [26]	۹۵/۱	۸۸/۵	۶/۹۴
RA-GCN [33]	۹۳/۵	۸۵/۵	۱۰/۱
AGC-LSTM [34]	۹۵	۸۹/۲	۲۲/۸۹
HCN [35]	۹۱/۱	۸۶/۵	۲/۶۴
MS-G3D [36]	۹۴/۷	۸۷/۲	۶/۴
SR-TSL [37]	۹۲/۴	۸۴/۸	۱۹/۰۷
HPGNet [27]	۹۴/۷	۸۷/۲	۰/۷۹۳
MV HPGNet [27]	۹۵/۸	۸۸/۶	۱/۶۵۰
AT-AR(ours)	۹۵/۵	۸۷/۷	۰/۸۴۱
MV AT-AR (ours)	۹۶/۷	۸۹/۹	۱/۷۸

جدول ۳ زمان مورد نیاز برای آموزش مدل در هر دوره را نشان می‌دهد، که با استفاده از آن می‌توان فهمید هر چه تعداد نمونه‌های اسکلتی مورد استفاده بیشتر باشند، شبکه به زمان بیشتری برای آموزش نیاز دارد.

همانطور که قبلاً ذکر شد، مدل پیشنهادی از دو جریان ورودی استفاده می‌کند. جریان position، مختصات سه‌بعدی مفاصل و جریان motion، تغییر مکان مفاصل اسکلتی را به عنوان ورودی دریافت می‌کند. جدول ۴ دقت شبکه AT-AR با استفاده از دو جریان ورودی را به صورت جداگانه نشان می‌دهد. همانطور که در این جدول مشخص است، دقت شبکه در حالت استفاده از جریان position بهتر از حالت استفاده از جریان motion است و استفاده همزمان از این دو جریان باعث بهبود دقت کلی می‌شود.

شبکه پیشنهادی با زبان پایتون و چارچوب پایتورچ پیاده سازی شده است و سیستم مورد استفاده برای اجرای آن در جدول ۵ قابل مشاهده است.



شکل ۶: دقت شبکه تحت ارزیابی cross view بر روی دیتاست NTU60.

جدول ۳: زمان مورد نیاز برای آموزش هر دوره، با توجه به تعداد نمونه‌های ورودی.

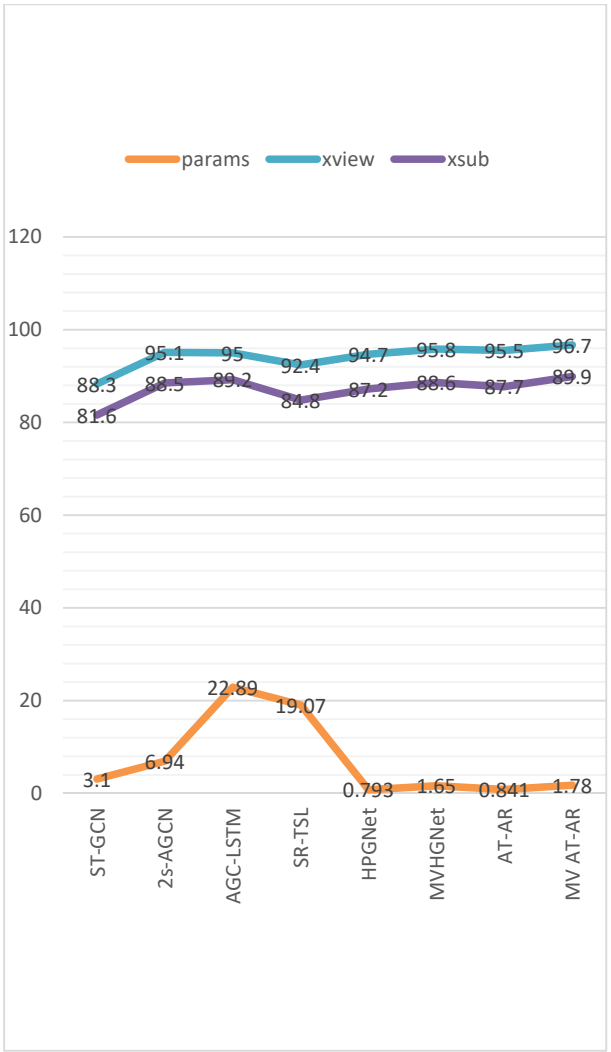
Model	Number of sample	Time (Hour)
MV AT-AR	۱۵۰۰۰	۰/۱۳
MV AT-AR	۴۰۰۰۰	۰/۳
MV AT-AR	۵۶۸۸۰	۰/۴۱

جدول ۴: دقت شبکه AT-AR بر روی دیتاست NTU60 با توجه به تعداد جریان‌های ورودی.

Model	Configuration	CV (%)	CS (%)
AT-AR	Two-Stream	۹۵/۵	۸۷/۷
AT-AR	Position Stream	۹۳/۸	۸۵/۲
AT-AR	Motion Stream	۹۲/۶	۸۴/۷

جدول۵ : مشخصات سخت افزار مورد استفاده برای اجرای شبکه

	Characteristics
Memory (Ram)	128 GB
Processor (CPU)	AMD Ryzen 3970X
Graphic (GPU)	ASUS GeForce RTX 2080
Operation System (OS)	Windows 10
Frame Work	Python (Torch)



شکل ۷: کارایی مدل پیشنهادی و مدل‌های پیشین بر روی دیتاست NTU60.

۵- نتیجه‌گیری

در این مقاله یک مدل تشخیص حرکت مبتنی بر داده‌های اسکلتی با استفاده از مکانیسم توجه ارائه شده است. شبکه پیشنهادی از یک مدل دوجریانی پیروی می‌کند، که در هر جریان از چندین بلوک کانولوشن گرافی و ماژول توجه استفاده می‌کند؛ تا همبستگی‌های مکانی بین مفاصل و همبستگی‌های زمانی بین فریم‌های مجاور و نامجاور برای یک حرکت یاد گرفته شوند. در نهایت با استفاده از یک گراف اسکلتی مکمل عملیات تشخیص حرکت چندنمایی را برای داده‌های اسکلتی انجام شده است. ارزیابی مدل تحت ۱۰۰ دوره نشان می‌دهد که مدل پیشنهادی بر روی مجموعه داده NTU RGB+D به ترتیب به دقت ۹۵/۵ و ۸۷/۷ درصد تحت ارزیابی‌های CV و CS هنگام استفاده از گراف اسکلتی اصلی رسیده است و هنگام استفاده از هر دو گراف اسکلتی و مکمل به دقت ۹۶/۷ و ۸۹/۹ درصدی تحت ارزیابی‌های CV و CS دست می‌یابد. همچنین تعداد پارامترهای شبکه پیشنهادی نسبت به مدل‌های قبلی به دلیل استفاده از شبکه‌های کانولوشن‌تفکیک‌پذیر بسیار کمتر شده است.

مراجع

[1] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, *Actional-structural graph convolutional networks for skeleton-based action recognition*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[2] J. K. Aggarwal and M. S. Ryoo, *Human activity analysis: A review*, ACM Comput. Survey, vol. 43, no. 3, p. 16, 2011.

[3] Y. Du, W. Wang, and L. Wang, *Hierarchical recurrent neural network for skeleton based action recognition*, in IEEE conference on computer vision and pattern recognition, pp. 1110–1118, 2015.

[4] J. Liu, A. Shahroudy, D. Xu, and G. Wang, *Spatio-temporal lstm with trust gates for 3d human action recognition*, in European conference on computer vision, pp. 816–833, Springer, 2016.

[5] F. Han, B. Reily, W. Hoff, and H. Zhang, *Space-time representation of people based on 3d skeletal data: A review*, Computer Vision and Image Understanding, vol. 158, pp. 85–105, 2017.

[6] C. Xu, Z. Guan, W. Zhao, H. Wu, B. Ling, *Adversarial incomplete multi-view clustering*, in: Twenty-Eighth International Joint Conference On Artificial Intelligence, IJCAI-19, 2019.

[7] Z. Zhang, *Microsoft kinect sensor and its effect*, IEEE MultiMedia, vol. 19, no. 2, pp. 4–10, 2012.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, *Realtime multiperson 2d pose estimation using part affinity fields*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7291–7299, July 2017.

[9] R. A. Güler, N. Neverova, and I. Kokkinos, *Densepose: Dense human pose estimation in the wild*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[10] C. Li, Q. Zhong, D. Xie, and S. Pu, *Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation*. In IJCAI, pages 786–792, July 2018.

[11] T.W. Chua, K. Leman, N.T. Pham, *Human action recognition via sum-rule fusion of fuzzy K-nearest neighbor classifiers*, in: IEEE International Conference On Fuzzy Systems, pp. 484–489, 2011.

[12] Z. Zhang, D. Tao, *Slow feature analysis for human action recognition*, arXiv: 1907.06670, 2019.

[13] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. *Sequence of the most informative joints (smij): A new representation for human skeletal action recognition*. Journal of Visual Communication and Image Representation, pp.24–38, 2014.

[14] M. Zanfir, M. Leordeanu, and C. Sminchisescu. *The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection*. In Proceedings of the IEEE international conference on computer vision, pages 2752–2759, 2013.

[15] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. *Graph based skeleton motion representation and similarity measurement for action recognition*. In European conference on computer vision, pages 370–385. Springer, 2016.

[16] Y. Du, Y. Fu, and L. Wang, *Skeleton based action recognition with convolutional neural network*, in Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR), pp. 579–583, 2015.

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

[18] 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), pp. 601–604, Jul. 2017.

[19] J. Liu, A. Shahroudy, D. Xu, and G. Wang, ''*Spatio-temporal LSTM with trust gates for 3D human action recognition*,'' in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland, pp. 816–833, Springer, 2016.

[20] 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 Int. Conf. Comput. Vis. (ICCV), pp. 1012–1020, Oct. 2017.

[21] 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), pp. 2117–2126, Oct. 2017.

[22] T. N. Kipf and M. Welling, "*Semi-supervised classification with graph convolutional networks*," in Proc. of ICLR, 2017.

[23] S. Yan, Y. Xiong, and D. Lin, "*Spatial temporal graph convolutional networks for skeleton-based action recognition*," in Proc. of AAAI, 2018.

[24] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, *Deep progressive reinforcement learning for skeleton-based action recognition*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5323– 5332, 2018.

[25] Y. Wen, L. Gao, H. Fu, F. Zhang, S. Xia, *Graph cnns with motif and variable temporal block for skeleton-based action recognition*, in: Proc. The 33rd AAAI Conference on Artificial Intelligence, 2019.

[26] 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), pp. 12026–12035. Jun. 2019.

[27] M. Wang, B. Ni, X. Yang, *Learning multi-view interactional skeleton graph for action recognition*. IEEE Trans. Pattern Anal. Mach. Intell. 2020.

[28] N. Heidari and A. Iosifidis, "*Temporal Attention Augmented Graph Convolutional Network for Efficient Skeleton-Based Human Action Recognition*," in International Conference on Pattern Recognition, 2020.

[29] F. Chollet. *Xception: Deep learning with depthwise separable convolutions*, arXiv preprint arXiv: 1610.02357v2, 2016.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. *MobileNets: Efficient convolutional neural networks for mobile vision applications*, arXiv preprint arXiv: 1704.04861, 2017.

[31] F. Barioli, W. Barrett, S. Fallat, H.T. Hall, L. Hogben, and H. van der Holst. *On the graph complement conjecture associated with minimum rank*. Lin. Alg. Appl., 436: 4373– 4391, 2012.

[32] A. Shahroudy, J. Liu, T. Ng, and G. Wang. *Ntu rgb+d: A large scale dataset for 3d human activity analysis*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1010–1019, 2016.

[33] Y.-F. Song, Z. Zhang, L. Wang, *Richly activated graph convolutional network for action recognition with incomplete skeletons*, in: 2019 IEEE International Conference on Image Processing, ICIP, pp. 1–5, 2019.

[34] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, *An attention enhanced graph convolutional LSTM network for skeleton-based action recognition*, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1227–1236, 2019.

[35] Li, C.; Zhong, Q.; Xie, D.; Pu, S. *Co-Occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation*. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18, Stockholm, Sweden, 13–19 July 2018.

[36] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, *Disentangling and unifying graph convolutions for skeleton-based action recognition*, in: 2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition, CVPR, pp. 140–149, , 2020.

[37] Chenyang Si, Ya Jing,WeiWang, LiangWang, and Tieniu Tan, "*Skeleton-based action recognition with spatial reasoning and temporal stack learning*," in ECCV, 2018.