نوع مقاله: پژوهشی

تشخیص حرکت انسانی با استفاده از روش‌های یادگیری عمیق به همراه تلفیق‌های با ناظر

مهسا مظفری وانانی[۱]، کارشناسی ارشد، رضا روحانی سروستانی[۲]، استادیار

۱- گروه مهندسی کامپیوتر نرم‌افزار - دانشکده فنی و مهندسی - دانشگاه شهرکرد - شهرکرد - ایران - mozafari.96ma@gmail.com
۲- گروه مهندسی کامپیوتر نرم‌افزار - دانشکده فنی و مهندسی - دانشگاه شهرکرد - شهرکرد - ایران - rrohani@sku.ac.ir

چکیده: با پیشرفت علم و گسترش روزافزون هوش مصنوعی، شناخت حرکات انسانی توسط رایانه‌ها به یک زمینه‌ی تحقیقاتی مهم تبدیل شده است. تشخیص حرکت به دلیل پیچیدگی حرکت انجام شده، شرایط آب و هوایی ناپایدار، شدت نور و تغییر زاویه دید دوربین با چالش‌های فراوانی روبه‌رو می‌باشد. برای نمایش حرکت، از دو نوع‌داده کارا و مفید ساختار اسکلتی بدن و ویدیو رنگی حاصل از دوربین کینکت استفاده می‌شود که در این مقاله مختصات اسکلت بدن علاوه بر دوربین کینکت با استفاده از روش‌های نوین یادگیری عمیق با دقت بالا تولید شده است. با استخراج ویژگی‌های تأثیرگذار از هر نوع‌داده به کمک روش‌های به‌روز ریاضیاتی و شبکه‌های عصبی، تشخیص حرکت با هر کدام از آن‌ها انجام می شود. هر یک از این نوع‌داده‌ها، به‌تنهایی قادر به ارائه‌ی اطلاعات خاص و متفاوت از حرکت انسان می‌باشند. در مرحله دوم این پژوهش، به هدف بهره‌گیری از ویژگی‌های هر دو نوع‌داده در کنار هم، از دو روش کارآمد تلفیق با نظارت MFA و BGLPCCA برای شناسایی حرکت انسانی استفاده می‌گردد. روش پیشنهادی پژوهش حاضر، چارچوبی برای تشخیص حرکت با یک نوع‌داده و چندین نوع‌داده است و دلیل برتری آن کسب مقدار صحت بالا در زمینه تشخیص حرکات انسانی می‌باشد.

واژه‌های کلیدی: تشخیص حرکات انسانی، ساختار سه‌بعدی اسکلت بدن، ویدیو رنگی، استخراج ویژگی، تلفیق ویژگی، یادگیری عمیق.

Human Action Recognition using Deep Learning and Supervised Data Fusion Approach

Mahsa Mozafari Vanani, Master of Science[1], Reza Rohani Sarvestani, Assistant[2]

1- Faculty of Computer Software Engineering, University of Sharekord, Sharekord, Iran, Email: mozafari.96ma@gmail.com
2- Faculty of Computer Software Engineering, University of Shahrekord, Shahrekord, Iran, Email: rrohani@sku.ac.ir

Abstract: With the advancement of science and the increasing expansion of artificial intelligence, the recognition of human actions by computers has become an important research field. Action recognition faces numerous challenges due to the complexity of the actions, unstable weather conditions, varying light intensity, and changes in camera viewing angles. To display action, two types of useful data are used: skeletal body structure and color video obtained from Kinect cameras. In this article, body skeletal coordinates are generated with high accuracy using advanced deep learning methods in addition to Kinect cameras. By extracting influential features from each type of data using modern mathematical methods and neural networks, action recognition is performed with each of them. Each type of data alone is capable of providing specific and different information about human action. In the second stage of this research, efficient combination methods with MFA and BGLPCCA supervision are used to utilize the features of both types of data together for human action identification. The proposed method in this research provides a framework for detecting action with single and multiple types of data, and its superiority lies in achieving high accuracy in human action recognition.

Keywords: Human Action Recognition, Skeleton Body Structure, RGB-Videos, Feature Extraction, Feature Fusion, Deep Learning.

تاریخ ارسال مقاله: ۱۴۰۱/۰۶/۱۹
تاریخ اصلاح مقاله: ۱۴۰۳/۰۷/۰۱
تاریخ پذیرش مقاله: ۱۴۰۳/۰۷/۲۱
نام نویسنده مسئول: رضا روحانی سروستانی
نشانی نویسنده مسئول: ایران - شهرکرد – دانشگاه شهرکرد – فنی و مهندسی– گروه آموزشی مهندسی کامپیوتر نرم‌افزار.

Digital Object Identifier 10.22034/JASP.

۱- مقدمه

در زندگی روزمره، انسان‌ها از طریق دیدار و شنیدار قادر به برقراری ارتباط با محیط پیرامون خود هستند و با انجام حرکاتی در پاسخ به اتفاقات اطراف خود واکنش نشان می‌دهند. حرکات صورت گرفته توسط انسان بیانگر احساس، توانایی، افکار و وضعیت فیزیکی او می‌باشند که منجر به فهم بهتر انسان می‌شوند. درک این حرکات توسط رایانه، یک زمینه‌ی پژوهشی در هوش مصنوعی به نام شناخت حرکت انسانی[۱] به وجود آورده است که در دهه‌های اخیر مورد توجه بسیاری از محققان قرار گرفته است. شناخت حرکات انسانی به دلیل وجود بسیاری از مشکلات مانند تغییر شرایط نور، اغتشاشات[۲] حسگرها، تغییر سرعت در انجام حرکت‌ها و سایر شرایط محیطی و دستگاهی، به یک موضوع تحقیقاتی جذاب تبدیل شده و علاقه‌ی بسیاری را برای تحقیق و رفع چالش‌های موجود در این زمینه برانگیخته است [۱].

در حوزه‌ی تشخیص حرکت انسانی، بینایی ماشین کاربرد بسیاری دارد. از ابزارهای بینایی ماشین می‌توان انواع حسگرها و دوربین‌هایی نظیر دوربین معمولی ضبط تصاویر رنگی[۳]، کینکت[۴] و غیره را نام برد که از هر یک بسته به نوع اطلاعات جمع‌آوری شده، مطابق با الزامات کار استفاده می‌شود [۲]. حرکت‌های انسانی توسط این ابزار ضبط می‌گردد و به عنوان ورودی برای الگوریتم‌های یادگیری ماشین جهت انجام عملیات مورد نظر فرستاده می‌شوند. دوربین کینکت معروف‌ترین ابزار در زمینه‌ی شناخت حرکات انسان است که در اکثر پژوهش‌ها مورد استفاده قرار می‌گیرد. کینکت دارای دو عدد لنز و یک درگاه مادون قرمز است که یکی از دوربین‌ها به همراه درگاه مادون‌قرمز برای تشخیص عمق تصاویر و دوربین دیگر، برای تشخیص تصاویر رنگی به کار گرفته می‌شود. دوربین کینکت به خوبی قادر به ضبط ویدیوهای رنگی از حرکت انسان‌ها می‌باشد.

طی دهه‌های گذشته، پژوهشگران عمدتاً برای تجزیه و تحلیل حرکت‌های انسانی، از تصاویر و ویدیوهای رنگی استفاده می‌کردند. بهره‌گیری از دوربین‌های ضبط حرکت، معمولاً بر اساس روش‌هایی مبتنی بر شباهت‌های انسانی، مدل‌های مکانی-زمانی و توصیف‌کننده‌های محلی می‌باشد؛ اما ویدیوهای رنگی بر خلاف نمایش واضح حرکت با مشکلاتی در شناخت عملکرد انسان، مانند حساسیت به نورهای مختلف و شامل بودن اطلاعات بی‌فایده از فضای پشت صحنه حرکت شخص روبرو هستند؛ همچنین، حجم اطلاعات فیلم‌های رنگی بسیار زیاد و پردازش آن‌ها سنگین است. علاوه بر تصاویر و ویدیوهای رنگی، دوربین کینکت قادر به استخراج نوع‌داده‌ی دیگری به اسم مختصات مفاصل اسکلت[۵] بدن انسان نیز می‌باشد. در ساختارهای اسکلتی، بدن انسان به تعدادی مفصل تقسیم می‌شود. این امر منجر به کاهش بسیار پیچیدگی به هنگام تشخیص حرکت می‌شود [۳]، لذا استفاده از مختصات نقاط اسکلتی بدن در پژوهش‌های امروزی بسیار چشم‌گیر شده است. علاوه بر این، مختصات اسکلتی بدن انسان به دلیل مختصر بودن[۶]، عدم حساسیت[۷]، نمایش مستقل از زاویه دید[۸] و حجم بسیار کمتر به دلیل حذف پس‌زمینه ویدیو، یکی از بهترین و پرکاربردترین نوع‌داده‌ها جهت شناسایی حرکات انسان می‌باشد [۴]؛ اما باز هم تشخیص حرکت به کمک نقاط اسکلتی نیز دارای ضعف‌هایی می‌باشد که گاهی اوقات در حرکاتی خاص، مشکلاتی ایجاد می‌کند.

به دلیل کاربرد زیاد و مؤثر مختصات نقاط اسکلتی، نحوه‌ی تشخیص این نقاط مسئله مهمی می‌باشد. مختصات مفاصل انسان از راه‌های مختلفی به دست می‌آید؛ همان‌طور که اشاره شد راحت‌ترین راه دستیابی به این نقاط استفاده از خود دوربین کینکت است. اما به هنگام استفاده از این دوربین، شرایط محیطی مانند: میزان نور محیط، فاصله دوربین از شخص، شلوغی محیط و غیره بر تصویربرداری بسیار تاثیرگذار هستند و اگر موارد ذکر شده در حالت مطلوب نباشند، عملکرد دوربین کینکت با مشکل مواجه می شود [۵]. با این وجود، راه دیگری که برای دستیابی به مختصات نقاط مفصلی بدن انسان حائز اهمیت می‌شود؛ استفاده از یادگیری عمیق است. روش‌های یادگیری عمیق به دلیل آموزش دادن شبکه عصبی با مختصات واقعی نقاط، می‌توانند نتایج مطلوبی را برای تشخیص مفاصل اسکلت بدن به دست آورند.

هر یک از نوع‌داده‌هایی که برای تشخیص حرکت انسانی وجود دارد؛ هر کدام به نوبه خود از مزایا و معایبی برخوردار هستند. از نظر عملی، هر روش نوع خاصی از اطلاعات را ضبط می‌کند که ممکن است دیگری بسته به شرایط قادر به دریافت آن اطلاعات نباشد یا آن‌ها را از دست بدهد؛ ولی این نوع‌داده‌ها می‌توانند به صورت تکمیل‌کننده‌ی هم عمل کنند. در واقع، اطلاعات مختصات اسکلتی بدن انسان مکمل اطلاعات تصاویر رنگی است؛ لذا ادغام این اطلاعات با یکدیگر می‌تواند باعث بهبود کیفیت و افزایش دقت تشخیص رفتار و حرکت انسان شود. برای ادغام نوع‌داده های مختلف، بعد از استخراج ویژگی از آن ها، می‌توان از روش‌های گوناگون تلفیق استفاده کرد و ویژگی‌های به دست آمده از اطلاعات نوع‌داده‌ها را با همدیگر ترکیب نمود.

مسئله شناخت حرکات انسانی به دلیل وابستگی به هوش مصنوعی، در محیط‌های مختلف نقش اساسی در تعامل انسان و ربات/رایانه دارد [۶]، همچنین دارای کاربردهای بسیاری در حوزه بازی‌های کامپیوتری، ساخت وسایل ورزشی، مسائل پزشکی و مراقبت‌های بهداشتی، سرگرمی و بازی، امنیت و دوربین‌های مدار بسته و نظارتی می‌باشد [۷]؛ لذا ساخت ماشینی که بتواند اقدامات و حرکات انسان را به طور دقیق شناسایی کند، از اهمیت کارهای تحقیقاتی هوش مصنوعی می‌باشد. در دهه‌های اخیر به دلیل اهمیت و کارایی چشم‌گیر تشخیص حرکت انسانی در علم و صنعت، تحقیقات بسیاری در این زمینه صورت گرفته است که به مرور برخی از آن‌ها پرداخته می‌شود. در ابتدا مرجع [۸] در زمینه‌ی تشخیص حرکت، تنها با نوع‌داده تصاویر رنگی یک نمایش ویدیویی مبتنی بر مسیرهای متراکم و توصیف‌گر حرکت مرزی[۹] ارائه داد و به کمک مسیرهای متراکم که پوشش خوبی از حرکت پیش‌زمینه فراهم می‌کرد، اطلاعات محلی هر

فیلم را ضبط کرد و به عنوان توصیف‌کننده، ویژگی‌های مطابق با مسیر مختصات نقاط و مدل حرکتی هیستوگرام‌های جریان‌های نوری را استخراج کرد و به تشخیص حرکت پرداخت. در مرجع [۹] با تعمیم تکنیک های موجود در تشخیص شئ و همچنین با استفاده از ویژگی‌های استخراج شده از فیلم، مدل‌های عملیاتی جدیدی برای تشخیص حرکت ساختند و به کمک شبکه‌های عصبی پیچشی، ویژگی‌های مکانی-زمانی را برای ساخت طبقه‌بندهای قوی استخراج کردند. علاوه بر این، روش پیشنهادی ارائه شده را به ردیابی حرکت در زمان ارتباط دادند و آن را خط لوله حرکت[۱۰] نامیدند.

در گذر زمان و پیشرفت علم جهت تشخیص سریع حرکت، نظر محققان به نوع‌داده‌ی ساختار اسکلتی جلب شد. در مرجع [۱۰] مختصات نقاط اسکلتی بدن انسان که توسط دوربین کینکت تخمین زده شده بود، مورد توجه قرار گرفت. این مرجع یک الگوریتم تکاملی پیشنهادی برای تعیین زیرمجموعه بهینه مفاصل و استخراج ویژگی ارائه کرد که منجر به بهبود میزان دقت نهایی تشخیص حرکت شد. یک روش مبتنی بر طبقه‌بندی تصویر برای شناسایی حرکت در مقیاس بزرگ از فیلم‌های اسکلت سه‌بعدی در مرجع [۱۱] ارائه شد. در این مرجع نگاشت فیلم‌های اسکلت سه‌بعدی برای تصاویر رنگی موجود به دست آمد و سپس یک شبکه عصبی پیچشی چندمقیاسی برای دسته‌بندی تصویر معرفی گردید. مرجع [۱۲] یک روش ساده و کارآمد برای رمزگذاری اطلاعات مکانی-زمانی توالی‌های اسکلت سه‌بعدی ارائه داد و آن را نقشه‌های خط سیر مفصل[۱۱] نامید و از شبکه‌های عصبی عمیق برای استخراج ویژگی در تشخیص حرکت انسان استفاده کرد.

مرجع [۱۳] در کار خود یک چارچوب یادگیری برای تشخیص حرکت پیشنهاد کرد که یک روش یادگیری رمزگذاری جدید معرفی می‌کند. در این مرجع ابتدا، یک روش یادگیری مبتنی بر گلوگاه اطلاعات طراحی کردند تا مدل را برای یادگیری یک نمایش پنهان آموزنده راهنمایی کنند. برای ارائه اطلاعات متمایز برای طبقه‌بندی کنش ها، نمودار مبتنی بر توجه را معرفی کردند. علاوه بر این، یک نمایش چندوجهی از اسکلت بدن انسان را با استفاده از موقعیت نسبی مفاصل ارائه کردند که برای ارائه اطلاعات سه‌بعدی مکمل مفاصل طراحی شده بود و در تشخیص حرکت انسان کارایی مناسبی داشت.

در نهایت محققان با توجه به کم و کاستی اطلاعات هر نوع‌داده در شرایط مختلف، برای بهره‌گیری از خاصیت تمام نوع‌داده‌ها به کارهایی در زمینه‌ی تلفیق پرداختند. در مرجع [۱۴] برای تشخیص عمل انسان، ویژگی‌های سه نوع‌داده را با یکدیگر ادغام کردند و نشان دادند تلفیق نوع‌های مختلف داده در مقایسه با شرایطی که از هر نوع‌داده جداگانه استفاده شود باعث بهبود میزان تشخیص حرکت می‌شود. در مرجع [۱۵] روشی مبتنی بر شبکه‌های عصبی پیچشی عمیق برای شناخت حرکت با ترکیب چندین نوع‌داده ارائه گردید. در مرجع [۱۶] مسئله شناسایی حرکات انسان را به‌گونه‌ای بررسی کردند که در آن هر عمل توسط حسگر چندگانه نظیر کینکت فیلم‌برداری شده است. آن‌ها برای تلفیق ویژگی‌های استخراج شده از انواع داده‌ها، دو روش نوین را با هدف ترکیب ویژگی‌های محلی و عمومی ارائه دادند. مرجع [۱۷] یک مدل عمیق D3D-LSTM مبتنی بر 3D-CNN و LSTM برای شناسایی عمل و تعامل انسان بر روی تصاویر رنگی و عمق ارائه داد. این مرجع در مدل خود ویژگی‌های هر دو نوع تصویر را با استفاده از روش‌های تلفیق مؤثر و کارا، ترکیب نمود و عملکرد خوبی از خود نشان داد.

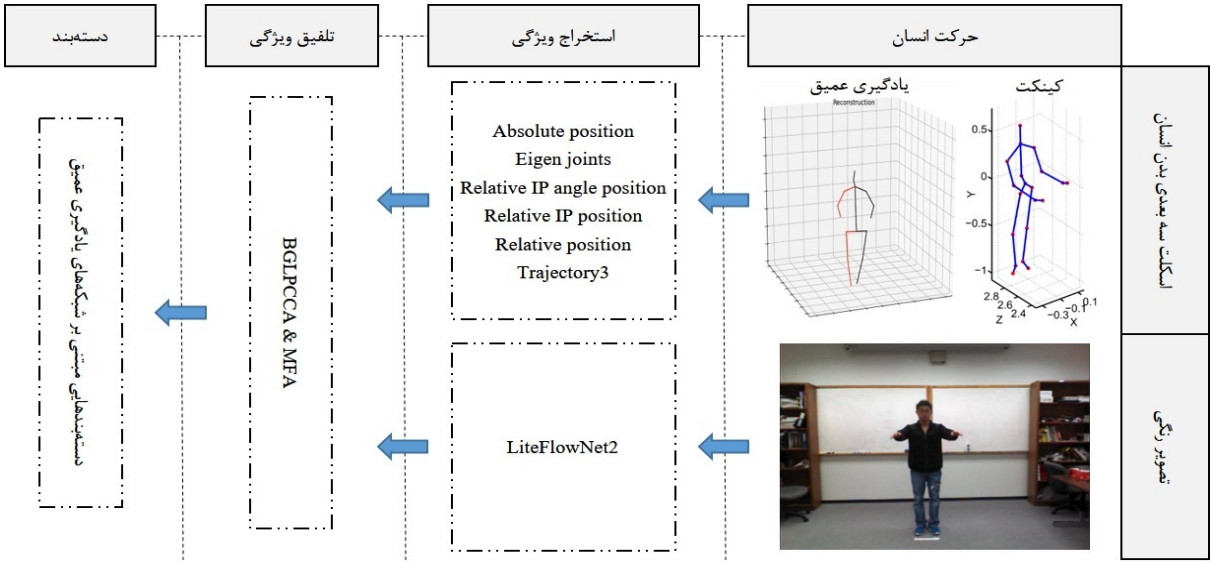
این مقاله، با هدف شناسایی هر چه بهتر حرکت انسان و افزایش دقت و صحت نسبت به پژوهش‌های موجود در این زمینه، انجام شده است. برای این امر، از دو نوع‌داده‌ی پرکاربردتر و مفیدتر یعنی تصاویر رنگی و مختصات ساختار اسکلتی بدن انسان استفاده می‌گردد و با روش‌های نوین یادگیری عمیق، سعی در ایجاد آن‌ها و استخراج ویژگی مفیدتر می‌شود. همچنین ویژگی‌های موجود با یکدیگر تلفیق می‌شوند. در ادامه ساختار مقاله به شرح زیر است: در بخش ۲ روش پیشنهادی مطرح می‌گردد که شامل چندین قسمت ایجاد نوع‌داده‌ها، استخراج ویژگی، روش‌های تلفیق و طراحی دسته‌بندها می‌باشد و جزئیات هر کدام توضیح داده می‌شوند. در بخش ۳ به طور مبسوط چگونگی پیاده‌سازی روش پیشنهادی بیان می‌شود و نتایج به دست آمده مطرح می‌گردد و در نهایت، بخش ۴ به نتیجه‌گیری پژوهش می‌پردازد.

۲- روش پیشنهادی

این پژوهش در راستای بهبود کیفیت و دقت تشخیص حرکات انسانی صورت گرفته است؛ روش پیشنهادی به کار گرفته شده برای این مهم در شکل ۱ نشان داده شده است. روش پیشنهادی یک چارچوب کلی برای تشخیص حرکت می‌باشد. در این پژوهش تشخیص حرکت ابتدا با دو نوع‌داده ویدیو رنگی و ساختار اسکلتی بدن انسان به صورت مجزا انجام می‌شود؛ سپس به تلفیق ویژگی‌های استخراج شده از این دو نوع‌داده پرداخته می‌شود و شناسایی حرکت با نوع‌داده‌های تلفیق شده نیز انجام می‌شود. در ادامه مراحل روش پیشنهادی با جزئیات شرح داده می‌شود.

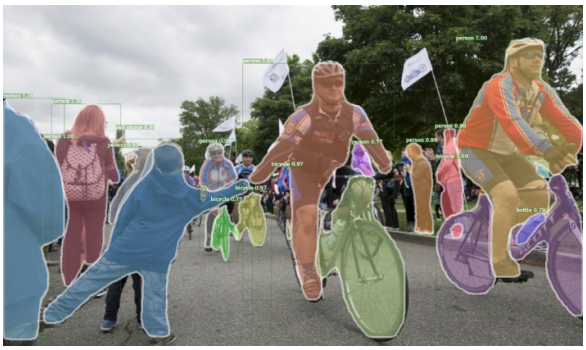
۲-۱- ایجاد نوع‌داده‌ها

برای انجام تحقیق از دو نوع‌داده مفید ویدیو رنگی و ساختار سه‌بعدی اسکلت بدن انسان که نسبت به بقیه نوع‌داده‌ها مزایای بیشتری دارند، استفاده می‌شود. مجموعه‌داده انتخابی، خود دارای ویدیوهای رنگی از حرکات انجام شده و مختصات سه‌بعدی بدن انسان که با کمک دوربین کینکت ایجاد شده است، می‌باشد. بنابراین، برای دستیابی به این دو نوع‌داده از فایل‌های موجود در مجموعه‌داده استفاده می‌گردد. اما علاوه بر دوربین کینکت، برای دستیابی به موقعیت و ساختار سه‌بعدی انسان راه دیگری نظیر بهره‌گیری از روش‌های یادگیری عمیق وجود دارد که در این تحقیق با روشی نوین از یادگیری عمیق، مختصات و موقعیت سه‌بعدی بدن انسان تولید می‌شود و تشخیص حرکت با هر دو مدل اسکلت حاصل از کینکت و اسکلت حاصل از یادگیری عمیق انجام می‌گردد و به مقایسه‌ی آن‌ها نیز پرداخته می‌شود.



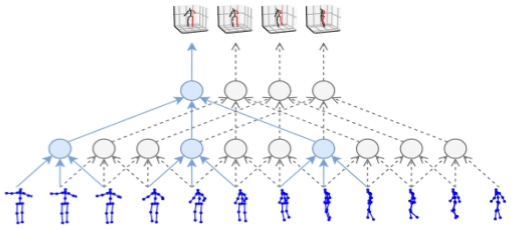
شکل ۱: روش پیشنهادی

روش مبتنی بر الگوریتم‌های جدید شبکه عصبی یادگیری عمیق برای ساخت مختصات اسکلتی از دو بخش اصلی تشکیل شده است: تشخیص اشیا موجود در هر ویدیو و ایجاد مختصات سه‌بعدی ساختار اسکلتی بدن[۱۲] [۱۸]. در بخش اول، تشخیص اشیای موجود در توالی‌های ویدیویی با استفاده از Detectron [۱۹] صورت می‌گیرد. Detectron شامل پیشرفته‌ترین الگوریتم‌های تشخیص اشیا می‌باشد که معروف‌ترین آن ها Mask R-CNN [۲۰] است. در این الگوریتم با استفاده از یک مدل از پیش آموزش دیده، توالی‌های ویدیویی دریافت و نقاط کلیدی دوبعدی[۱۳] بدن انسان در هر فریم به عنوان خروجی ارائه می‌شود. مدل یاد شده از الگوریتم تشخیص شیء شبکه عصبی پیچشی ناحیه‌ای[۱۴] (R-CNN) بهره می‌گیرد [۱۹]. در شکل ۲ نمونه‌ای از عملکرد Detectron در فرایند تشخیص اشیا قابل مشاهده می‌باشد.

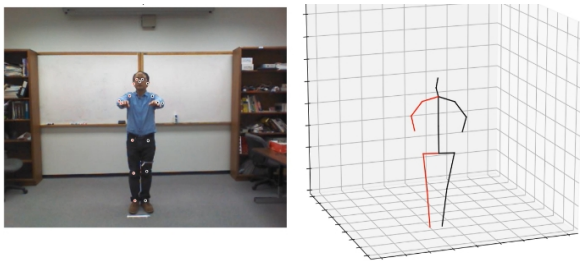


شکل ۲: نمونه‌ای از تشخیص اشیا توسط Detectron [۱۹]

در بخش دوم برای ایجاد ساختار اسکلت سه‌بعدی با یادگیری عمیق، مختصات دوبعدی نقاط کلیدی به مختصات سه‌بعدی تبدیل می‌شوند. در واقع، در این بخش از روی نقاط کلیدی دوبعدی در توالی‌های ویدیویی با استفاده از مدل پیچشی زمانی[۱۵]، مختصات سه‌بعدی اسکلت تخمین زده می‌شود. این مدل، طبق شکل ۳ توالی‌های دوبعدی مفاصل انسان را به‌عنوان ورودی دریافت و تخمینی از موقعیت آن‌ها در فضای سه‌بعدی تولید می‌کند [۲۱]. مدل یاد شده، از قبل بر روی مجموعه‌داده های بزرگی نظیر Human3.6M [۲۲] و HumanEva-I [۲۳] با مختصات واقعی آموزش داده شده است؛ لذا می‌تواند مختصات دقیقی از موقعیت سه‌بعدی با دقت بالا ارائه دهد. ساختار اسکلت سه‌بعدی ایجاد شده برای بدن انسان توسط یادگیری عمیق در شکل ۴ قابل رؤیت است.



شکل ۳: نحوه‌ی ایجاد اسکلت سه‌بعدی از توالی‌های دوبعدی [۲۱]



شکل ۴: ایجاد ساختار سه‌بعدی اسکلت با یادگیری عمیق

۲-۲- استخراج ویژگی

استخراج ویژگی فرآیندی است که در آن، مقادیر و داده‌های خام به اطلاعات با معنی و مفید تبدیل می‌شوند و با انجام عملیاتی بر روی آن‌ها به شکل قابل استفاده‌تری برای پردازش‌های آتی در می‌آیند. در این مرحله از هر دو نوع‌داده ویدیو رنگی و ساختار اسکلت بدن، بسته به نوعشان با روش‌های مختلف ویژگی استخراج می‌شود.

۲-۲-۱- استخراج ویژگی نوع‌داده ساختار اسکلت

ساختار اسکلتی انسان از دو مؤلفه‌ی مفصل و استخوان به عنوان اتصال‌دهنده، تشکیل می‌شود. با به‌کارگیری این دو مؤلفه می‌توان ویژگی‌های مطلوبی را برای ساختار اسکلتی بر اساس معیارهای ریاضیاتی نظیر موقعیت‌های مکانی مفاصل[۱۶] (JP) که از الحاق مختصات سه‌بعدی همه‌ی مفاصل به دست می‌آید؛ زوایای بین مفصلی[۱۷] (JA) و ارتباط‌های نسبی مکانی دوبه‌دوی مفاصل[۱۸] (RJP) استخراج نمود [۲۴]. برای انجام فرایند استخراج ویژگی از ساختار اسکلتی نیاز است با توجه به نحوه‌ی به دست آوردن آن، مدل بدن تعریف شود. مدل بدن، نمایانگر شالوده‌ی کلی بدن است؛ همچنین شامل تعداد مفاصل، تعداد ربط دهنده‌ها و اندیس مکان مفاصل می‌باشد. با به‌کارگیری مدل بدن و معیارهای یاد شده، چندین دسته ویژگی برای ساختار اسکلتی حاصل از دوربین کینکت و یادگیری عمیق استخراج می‌گردد.

۲-۲-۲- استخراج ویژگی نوع‌داده ویدیو رنگی

دهه‌های گذشته به منظور استخراج ویژگی از ویدیوهای رنگی، از ویژگی‌ای به نام جریان نوری[۱۹] استفاده می‌کردند. جریان نوری میزان تغییرات آشکار در ویدیو ناشی از حرکت اجسام، می‌باشد. این تغییرات حرکتی به شکل برداری متناسب با دستگاه مختصات تصویر نشان داده می‌شوند. برای محاسبه‌ی جریان حرکتی ابتدا ویدیو به توالی‌های فریمی تبدیل می‌شود؛ سپس میزان تغییر حرکت در فریم‌های پشت سر هم به صورت جریان نوری محاسبه می‌شود.

جریان نوری حرکتی با این که مسیر حرکت را به خوبی نشان می‌دهد اما پردازش آن بسیار سنگین و زمان‌بر است؛ لذا محققان از روش FlowNet2 که مبتنی بر شبکه‌های عصبی پیچشی (CNN) بود استفاده کردند [۲۵]. معماری به‌کار گرفته شده در این روش شامل بیش از ۱۶۰ میلیون پارامتر بود که به همین دلیل کارایی زیاد مطلوبی نداشت [۲۶]. روش LiteFlowNet2 با بهبود FlowNet2 و بهره‌گیری از یک معماری سبک‌تر توانست عملکرد بهتری از خود نشان دهد و همچنین در سرعت اجرا ۳/۱ برابر سریع‌تر عمل نماید [۲۷]. مدل انتخابی در این پژوهش، یک شبکه عصبی پیچشی از پیش آموزش دیده شده روی مجموعه‌داده‌ی Sintel و معماری LiteFlowNet2 می‌باشد. از این مدل به‌منظور استخراج جریان نوری از توالی‌های ویدیویی پشت سر هم استفاده می‌شود.

در شکل ۵ نمونه‌ای از این استخراج ویژگی در یک فریم از ویدیو نشان داده شده است. همان‌طور که بیان گردید ویژگی LiteFlowNet2 یک جریان نوری حرکتی است پس برای نمایش کامل آن نیاز به فضای سه‌بعدی می‌باشد و در شکل ۵ فقط نمونه‌ای از یک فریم حرکت دست در ویدیو و ویژگی‌اش ارائه شده است. این شکل، نشان‌دهنده‌ی با کیفیت و مطلوب بودن ویژگی LiteFlowNet2 است که توانسته جریان حرکت انجام شده در ویدیو رو به خوبی تشخیص دهد.

شکل ۵: عملکرد LiteFlowNet برای استخراج ویژگی از ویدیو رنگی

۲-۳- تلفیق ویژگی

تلفیق ویژگی فرآیند یکپارچه سازی اطلاعات از چندین منبع داده مختلف برای ایجاد و تولید اطلاعات بهتر، مفیدتر نسبت به اطلاعات ارائه شده از هر منبع داده‌ای می‌باشد. در زمینه تشخیص حرکات انسانی، هر یک از نوع‌داده‌ها مزایا و معایب مختص به خود دارند و هر کدام در شرایط خاص این امکان را دارند که از خود عملکرد بهتری نشان دهند. همچنین با یک نوع‌داده به تنهایی، گاهی اوقات نمی‌توان تمام حرکات انجام شده را تشخیص داد؛ لذا روش‌های تلفیق ویژگی برای بهره‌مندی از اطلاعات همه‌ی نوع‌داده‌ها حائز اهمیت می‌شود. با تلفیق ویژگی حاصل از داده‌های گوناگون، می‌توان ضعف‌های هر نوع داده را با دیگری پوشش داد.

در این پژوهش به‌منظور تلفیق ویژگی‌های توالی دو نوع‌داده ویدیو رنگی و ساختار اسکلتی، ابتدا آن‌ها را به یک فضای حالت جدید نگاشت داده و سپس در آن فضا با یکدیگر ترکیب می‌گردند. همچنین، برای انجام این امر از رویکرد روش‌های تلفیق با نظارت ا ستفاده می شود. در روش‌های با نظارت، برای ایجاد فضای مشترک مناسب‌تر بین ویژگی‌ها از برچسب آن‌ها نیز در فرایند ترکیب بهره گرفته می‌شود. به‌کارگیری برچسب ویژگی‌ها، روش‌های تلفیق را قادر می‌سازد که فضای حالت مشترک را به گونه‌ای ایجاد نمایند که داده‌ها در آن‌جا قابلیت تفکیک پذیری مناسبی به خصوص در مسائل طبقه‌بندی داشته باشند و تشخیص حرکت با دقت بالاتری نسبت به حالت استفاده از یک نوع‌داده انجام شود. پژوهش حاضر دو تا از بهترین روش‌های تلفیق با ناظر را به‌کار گرفته است که جزئیات آن‌ها در ادامه به صورت کامل توضیح داده می‌شود.

۲-۳-۱- روش Marginal Fisher Analysis (MFA)

ساده‌ترین روش با نظارت در فرایند تلفیق روش LDA [۲۸] می‌باشد. عملکرد این روش به گونه‌ای است که فضای تلفیق ویژگی را با فرض گوسی بودن توزیع داده‌ها می‌سازد. در واقع در اعمال روش LDA، داده‌ها حول میانگین خود جمع می شوند و بیشترین میزان همبستگی را حول میانگین دارند؛ همچنین فا صله میانگین‌ها از هم بی شترین حد ممکن می‌شود. اما در بسیاری از مسائل دنیای واقعی، فرض گوسی بودن توزیع داده ها برقرار نیست و نمی‌توان معیار های میانگین و

واریانس را برای ایجاد فضای مشترک بین ویژگی‌ها در نظر گرفت؛ لذا همین امر منجر به عدم کارایی مطلوب روش LDA می‌شود.

روش تحلیل حاشیه‌سازی فیشر (MFA) [۲۹]، بدون در نظر گرفتن این فرض و بهره‌گیری از ساختار گراف مانند به ساخت فضای مشترک بین ویژگی‌ها برای تلفیق می‌پردازد و سعی دارد که ویژگی‌های کلاس‌های متفاوت را در این فضا تا بیشترین حد ممکن از یکدیگر تفکیک کند و ویژگی‌های درون یک کلاس تا حد امکان فشرده شوند. این روش با ایجاد تابع هدف موجود در رابطه‌ی (۱)، هدف مورد نظر را برآورده می‌کند.

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{|W^T X^T (ST_{\mathrm{b}} - S_{\mathrm{b}}) XW|}{|W^T X^T (ST_{\mathrm{w}} - S_{\mathrm{w}}) XW|} \qquad (۱)$$

که در این رابطه $ST_w^{kk} = \sum_{kl.k\neq l} S_w^{kl}$ و $ST_b^{kk} = \sum_{kl.k\neq l} S_b^{kl}$ فشرده‌سازی درون کلاس S_w را به صورت زیر تعریف می‌کند.

$$S_w^{kl} = \begin{cases} 1 & where\ k \in R^{k1}(l)\ or\ l \in R^{k1}(k) \\ 0 & otherwise \end{cases} \qquad (۲)$$

که در آن $R^{k1}(l)$ نشان‌دهنده‌ی مجموع نمونه‌هایی از یک کلاس است که در همسایگی $k1$ از نمونه x^l قرار دارند. تابع جداکننده کلاس‌های متفاوت نیز که با S_b نشان داده شده است به صورت زیر تعریف می‌گردد.

$$S_b^{kl} = \begin{cases} 1 & where\ (k.l) \in P^{k2}(c_1)\ or\ (k.l) \in P^{k2}(c_k) \\ 0 & otherwise \end{cases} \qquad (۳)$$

که در آن $P^{k2}(l)$ به مجموعه‌ای از جفت داده‌ها اشاره دارد که در همسایگی $k2$ از کلاس‌های متفاوت قرار دارند.

۲-۳-۲- روش Biset Globality Locality Preserving Canonical Analysis (BGLPCCA)

یکی دیگر از الگوریتم‌های با ناظر تلفیق، روش نگه‌داشت تحلیل همبستگی کانونی محلی و عمومی دوطرفه (BGLPCCA) [۳۰] از ترکیب چندین الگوریتم پیشرفته‌ی تلفیق مانند: CCA [۳۱]، LPP [۳۲] و GLPP [۳۳] تشکیل می‌شود. الگوریتم CCA به محاسبه‌ی میزان همبستگی کانونی بین دو مجموعه‌ی چند متغیره می‌پردازد. هر چه دو مجموعه ارتباط بیشتری با هم داشته باشند میزان ضریب همبستگی آن‌ها بیشتر است. الگوریتم خطی LPP با در نظر گرفتن اطلاعات محلی داده‌ها، به کمک بهره‌گیری از مفهوم لاپلاسی گراف، یک ماتریس انتقال برای نگاشت داده‌ها به وجود می‌آورد و با اعمال این ماتریس، داده‌ها را به یک زیرفضا نگاشت می‌دهد. این تبدیل خطی به طور بهینه، اطلاعات و ساختار محلی نوع‌داده‌ها را در این انتقال حفظ می‌کند. در روش GLPP، علاوه بر حفظ اطلاعات محلی به هنگام نگاشت، اطلاعات جهانی و عمومی نوع‌داده‌ها نیز نسبت به یکدیگر حفظ می‌شوند.

در روش تلفیق BGLPCCA، همسایگی‌های محلی و عمومی نوع داده‌ها بر اساس برچسب آن‌ها تعیین می‌گردند. در واقع فضای حالت ثانویه با در نظر گرفتن برچسب داده‌ها و حفظ ساختار محلی و عمومی هر کدام از آن‌ها ساخته می شود. این روش، یک روش جامع است که در آن با نگاشت اطلاعات به فضای حالت جدید با در نظر گرفتن دو معیار گفته شده، تلفیق ویژگی‌ها به گونه‌ای انجام می شود که داده‌های هر کلاس، بیشترین میزان شباهت را به یکدیگر داشته باشند. روش BGLPCCA، به دلیل برخورداری از کارایی بالا و مطلوب می‌تواند به بهترین شکل ممکن ویژگی‌ها را با یکدیگر تلفیق نماید و به حالت ایده‌آل برساند. این روش، طبق رابطه (۴) محاسبه می‌شود.

$$\max_{W_x.W_y} \quad Trace(W_x^T (U_x K_{xy} U_y^T + \beta X L_{xy} Y^T) W_y) \qquad (۴)$$

$$subject\ to \quad W_x^T (U_x K_{xx} U_x^T + \beta X L_{xx} X^T) W_x = I;$$
$$W_y^T (U_y K_{yy} U_y^T + \beta Y L_{yy} Y^T) W_Y = I$$

که در آن $U_x = [u_1^x \ldots u_i^x \ldots u_k^x]$ و $U_y = [u_1^y \ldots u_i^y \ldots u_k^y]$ به ترتیب میانگین فضاهای نمونه کلاس برای X و Y هستند. $L_{xx} = D_{xx} - S_x \circ S_x$، $L_{xy} = D_{xy} - S_y \circ S_y$ و $S_x \circ S_y$ هستند که نماد $\circ$ بیانگر ضرب درایه به درایه بین ماتریس‌ها و D_{xx} ماتریس قطری است. S_x و S_y طبق رابطه‌ی (۵) قابل محاسبه است.

$$S_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{t_s}\right) & i.j \in c\,.c \in C.i \neq j \\ 0 & otherwise \end{cases} \qquad (۵)$$

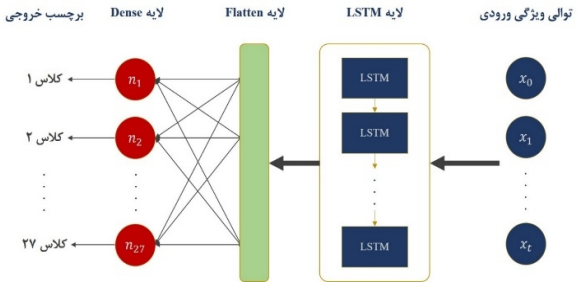
به طور مشابه $K_{xx} = G_{xx} - B_x \circ B_x$، $K_{yy} = G_{yy} - B_y \circ B_y$ و $K_{xy} = G_{xy} - B_x \circ B_y$ هستند که G_{xx} ماتریس قطری است. B_x و B_y طبق رابطه (۶) قابل محاسبه می‌باشند. جزئیات چگونگی حل معادله BGLPCCA در مرجع [۳۰] آورده شده است.

$$B_{ij} = \begin{cases} \exp\left(-\frac{\|U_i - U_j\|^2}{t_B}\right) & .i \neq j \\ 0 & otherwise \end{cases} \qquad (۶)$$

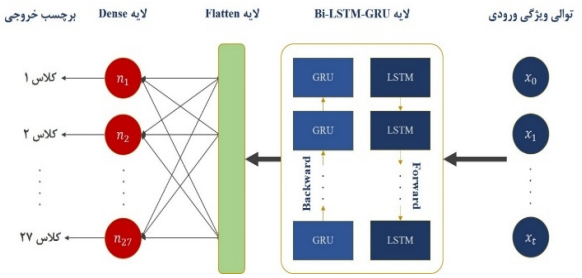
۲-۴- طراحی طبقه‌بندهایی مبتنی بر یادگیری عمیق

از آن‌جا که در چند سال اخیر طبقه‌بندهای یادگیری عمیق به طور کلی بهتر از طبقه‌بندهای یادگیری ماشین سنتی عمل کرده‌اند [۴]؛ در این پژوهش نیز چندین طبقه‌بند به کمک شبکه‌های عصبی برای تشخیص حرکت انسان طراحی می‌شود. تشخیص حرکت در این پژوهش، در سطح فریم انجام شده؛ لذا به طبقه‌بندهایی که ترتیب را متوجه شوند نیاز است.

در ساختار دسته‌بندهای طراحی شده، از شبکه‌های عصبی بازگشتی[۲۰] (RNN)، حافظه کوتاه مدت طولانی[۲۱] (LSTM)، بازجریانی دروازه‌دار[۲۲] (GRU) و حالت‌های دوطرفه‌ی هر یک از این شبکه‌ها و ترکیب آن‌ها با یکدیگر استفاده می‌گردد. در هر دو مرحله‌ی پژوهش، اعم از تشخیص حرکت با تک نوع‌داده‌ها و تلفیق چندین نوع‌داده با هم از طبقه‌بندهای طراحی شده استفاده می‌شود. این طبقه‌بندها را می‌توان به دو حالت یک‌طرفه و دوطرفه تقسیم کرد که در شکل‌های ۶ و ۷ ساختار و جزئیات یکی از طبقه‌بندها برای مثال نشان شده است.



شکل ۶: ساختار مدل LSTM از مدل‌های یک‌طرفه



شکل ۷: ساختار مدل Bi-LSTM-GRU از مدل‌های دوطرفه

۳- پیاده‌سازی و ارزیابی نتایج

در این پژوهش، مراحل پیاده‌سازی با بهره‌گیری از دو زبان قدرتمند برنامه‌نویسی پایتون و متلب انجام شده است. از نرم‌افزار متلب ورژن 2017a برای پیاده‌سازی کدهای استخراج ویژگی ساختار اسکلتی و تلفیق ویژگی ها استفاده می‌گردد. با توجه به سنگینی پردازش ویدیوهای رنگی و محاسبات الگوریتم‌های یادگیری عمیق، از محیط گوگل کولب، به منظور تأمین نیازهای سخت‌افزاری و حافظه کدهای پایتونی استفاده می شود. این محیط لینوکسی با در اختیار داشتن یک فضای پردازشی مبتنی بر ابر نیاز به سخت‌افزارهای قدرتمند برای آموزش مدل‌های یادگیری عمیق را برطرف می‌کند.

به منظور ارز یابی روش پیشنهادی از مجمو عه‌داده معروف و استاندارد UTD-MHAD استفاده می‌گردد. این مجموعه‌داده شامل ۸۶۱ توالی ویدیویی است که با رزولوشن ۶۴۰×۴۸۰ پیکسل توسط دوربین کینکت و با نرخ تقریبا ۳۰ فریم در ثانیه گرفته شده است. همچنین، دارای ۲۷ عمل می‌باشد که توسط هشت نفر (چهار زن و چهار مرد) انجام و هر عمل چهار بار تکرار شده است [۳۴].

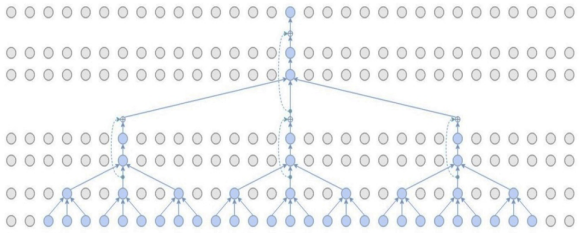
برای پیاده‌سازی روش پیشنهادی، در ابتدا به ساخت نوع‌داده ا سکلتی با الگوریتم‌های مبتنی بر یادگیری عمیق پرداخته می شود. در گام اول این مرحله، از مدل از پیش آموزش دیده شده‌ای در محصول Detecron به نام e2e_keypoint_rcnn_R-101-FPN_s1x استفاده شده است. این مدل از تکنیک end to end بهره می‌برد؛ لذا تمام گام‌ها از فاز اولیه که ورود داده می‌باشد تا زمانی که خروجی نهایی تولید می‌شود را یاد می‌گیرد. همچنین آموزش مدل یاد شده بر روی اجتماع دو مجموعه‌داده‌ی coco_2014_train و coco_2014_valminusminival انجام شده است. سایر مشخصات این مدل در جدول ۱ آورده شده است. در نهایت، به ازای تک‌تک توالی‌های ویدیویی با استفاده از مدل e2e_keypoint_rcnn_R-101-FPN_s1x نقاط کلیدی دوبعدی شخص موجود در تصویر استخراج و در قالب فایل‌هایی با پسوند npz ذخیره می‌شود.

جدول ۱: مشخصات مدل یاد شده برای ایجاد نقاط کلیدی دوبعدی

مبنای مدل	کاربرد مدل	زمان‌بندی نرخ یادگیری	حافظه برای آموزش (GB)	زمان برای آموزش (s/iter)	مجموع زمان آموزش (hr)
R-101-FPN	kps	s1x	۱۰/۶	۰/۹۲۱	۳۳/۳

بعد از ساخت نقاط دوبعدی، با استفاده از یک مدل از پیش آموزش دیده شده‌ی دیگر، این نقاط به مختصات سه‌بعدی اسکلت تبدیل می‌شوند. برای بهره‌گیری از این مدل، نیاز به انجام یک‌سری پیش پردازش و مرتب‌سازی بر روی داده‌های خروجی مرحله‌ی قبل می‌باشد. در واقع، داده ها به فرمت مورد نیاز برای ورود به مرحله‌ی ایجاد مختصات سه‌بعدی تغییر حالت می‌دهند. برای ساخت این فرمت، نقطه شروع و پایان هر فریم به همراه نقاط کلیدی دوبعدی حاصل از مرحله‌ی قبل به صورت دیکشنری در قالب فایلی با پسوند npz ذخیره می‌شوند. نقطه پایان هر فریم برابر با تعداد نقاط کلیدی دوبعدی شخص است.

معماری مدل استفاده شده، بر اساس شبکه‌های عصبی پیچشی باز شده می‌باشد که توالی‌های ورودی را در لایه‌های اول به بعد، چند در میان می‌بیند. همچنین در ساختار این مدل به هنگام اضافه کردن هر لایه، از مکانیزم residual blocks استفاده شده است [۲۱]. نمایی از این مدل در شکل ۸ قابل رؤیت می‌باشد.



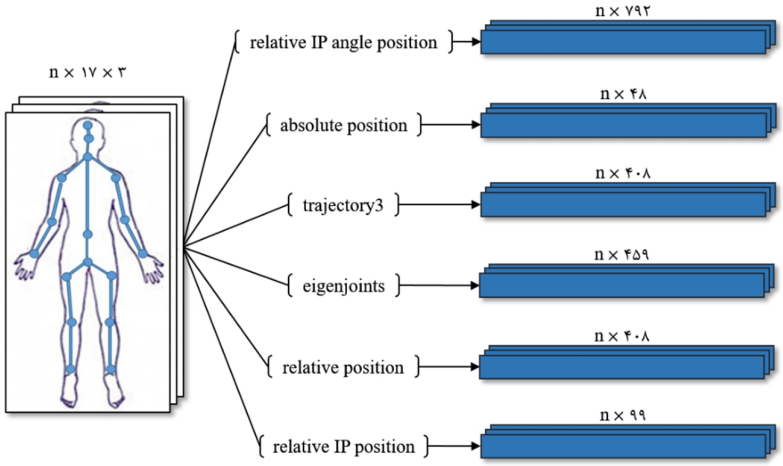
شکل ۸: نمایی از ساختار مدل در ایجاد مختصات سه‌بعدی [۲۱]

در گام بعد، به‌منظور استخراج ویژگی از ساختار اسکلتی، شش ویژگی absolute position، eigenjoints، trajectory3، relative position، relative IP position و relative IP angle position به کار گرفته می‌شود. در ابتدا برای نوع‌داده‌ی اسکلتی با توجه به این که ساختار ا سکلتی از دوربین کینکت یا یادگیری عمیق ایجاد شده، یک مدل بدن ساخته می‌شود. این مدل بر اساس تعداد نقاط مفصلی شناخته شده، تعداد استخوان‌های بین مفاصل و نحوه‌ی قرارگیری آن‌ها ساخته می‌شود. تعداد مفاصل تشخیص داده شده توسط دوربین کینکت بر روی مجموعه‌داده‌ی انتخابی ۲۰ و توسط یادگیری عمیق ۱۷ عدد می‌با شد. در ادامه برای ا ستخراج ویژگی‌های گفته شده، داده‌های اسکلتی در یک ماتریس سه‌بعدی قرار داده می‌شوند که ابعاد این

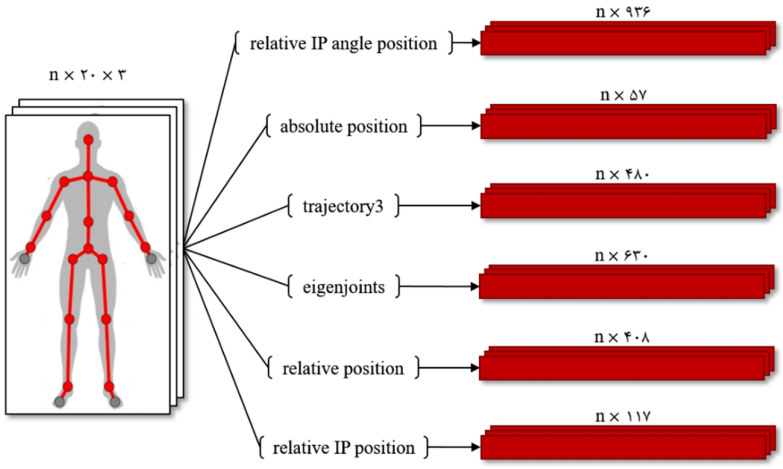
ماتریس به ترتیب برابر با تعداد فریم‌های هر ویدیو، تعداد مفاصل و تعداد محورهای مختصات (x، y و z) می‌باشد.

سپس برای ساخت ویژگی absolute position، مقادیر ماتریس فوق نرمال‌سازی می‌گردد. به تعداد یکی کمتر از تعداد مفاصل، استخوان ایجاد می‌شود که برای ساخت ویژگی eigenjoints مقادیر مکانی این مفاصل از هم کم می‌شود تا طول هر استخوان (رابط بین دو مفصل) به دست آید. به جهت ساخت ویژگی trajectory3، مقدار مکانی هر مفصل در فریم فعلی، منهای مقدار مکانی آن در فریم قبلی می‌شود. برای ایجاد سه ویژگی باقی‌مانده، با بهره‌گیری از مدل بدن ساخته شده، تمامی مفاصلی که دوبه‌دو به یکدیگر وصل هستند در ماتریسی مشخص و نگه‌داری می‌شوند. محاسبه‌ی اختلاف مکانی این نقاط منجر به ساخت ویژگی relative position می‌شود. در ساخت دو ویژگی دیگر، علاوه بر لحاظ کردن تغییرات مکانی مفاصل، تغییرات زمانی آن‌ها در گذر فریم‌ها نیز حائز اهمیت است. ویژگی relative IP position با کنار هم قرار دادن ماتریس مکانی مفاصل به هم متصل و ماتریس تغییرات آن‌ها در طول زمان بعد از عملیات نرمال‌سازی ساخته می‌شود. در ساخت ویژگی relative IP angle position، به جای بهره‌گیری از مکان مفاصل، از ماتریس زوایای مفاصل متصل به هم و ماتریس تغییرات این زاویه‌ها در فریم‌های پشت سر هم استفاده می شود [۲۴]. روند اعمال استخراج ویژگی‌های مذکور بر روی نوع‌داده‌های اسکلتی در شکل‌های شماره ۹ و ۱۰ آورده شده که n تعداد فریم‌ها در هر ویدیو می‌باشد.

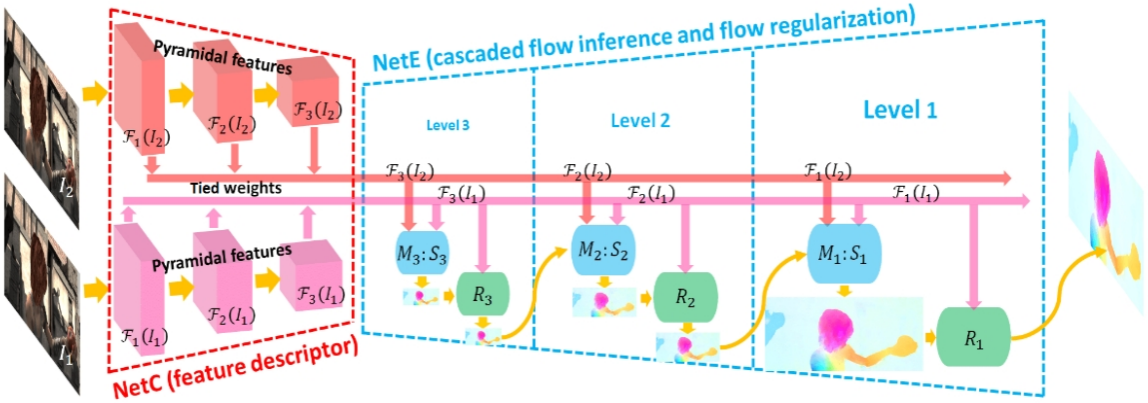
در ادامه به‌منظور استخراج ویژگی از نوع‌داده‌ی ویدیوی رنگی، یک مدل از پیش آموزش دیده شده بر روی مجموعه‌داده‌ی Sintel مورد استفاده قرار گرفته می‌شود. ساختار مدل ذکر شده، مبتنی بر شبکه‌های عصبی پیچشی است که با هدف تخمین جریان نوری بر روی داده‌ها اجرا می‌شود. این مدل، از دو قسمت کدگذاری و کدگشایی تشکیل شده که ورودی آن یک زوج تصویر (دو فریم پشت سر هم) می‌باشد. عملکرد کلی مدل به این صورت است که ابتدا تصاویر با ورود به قسمت Encoder، کدگذاری شده و ویژگی‌هایی از آن‌ها به دست می‌آید؛ سپس این ویژگی‌ها برای کدگشایی به قسمت Decoder فرستاده می شوند تا جریان حرکتی بین دو تصویر استخراج گردد [۲۷]. عملکرد این مدل در شکل ۱۱ قابل مشاهده است.



شکل ۹: روند استخراج ویژگی از نوع‌داده اسکلتی حاصل از یادگیری عمیق



شکل ۱۰: روند استخراج ویژگی از نوع‌داده اسکلتی حاصل از کینکت



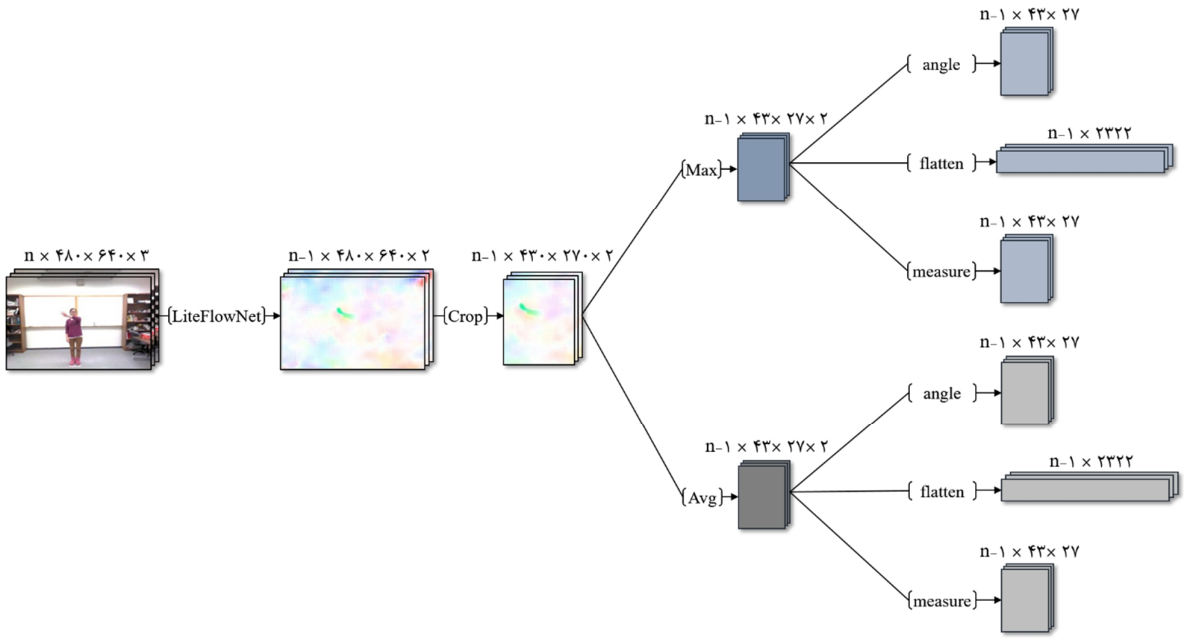
شکل ۱۱: معماری مدل از پیش آموزش دیده شده برای استخراج ویژگی LiteFlowNet [۲۷]

قسمت کدگذاری این مدل در شکل ۱۱، یک زیرشبکه‌ی دو جریانه است که دو تصویر با ورود به این قسمت از ۱۰ لایه‌ی کانوولوشنی عبور می‌کنند و به بردارهای ویژگی تبدیل می شوند. همان‌طور که در شکل ۱۱ دیده می‌شود قسمت کدگشای مدل، از دو ماژول استخراج جریان و منظم‌سازی به ترتیب به نام‌های M:S و R ساخته شده است. ماژول استخراج جریان، خود از دو واحد M و S تشکیل شده است. واحد M، کار تطبیق ویژگی ها با یکدیگر را انجام می‌دهد و واحد S، وظیفه‌ی اصلاح زیرپیکسل‌ها را بر عهده دارد [۲۷]. قسمت Decoder با استفاده از دو ماژول M:S و R، ویژگی های ایجاد شده در قسمت قبل را کدگشایی می‌کند. در واقع، هر سطح در این قسمت از ویژگی‌های به دست آمده در سطح متناظر و هم‌نام خود در قسمت Encoder استفاده می‌کند تا خروجی مربوطه را بسازد. خروجی نهایی مدل، جریان نوری بین دو فریم است که به صورت آبشاری تولید می‌شود [۲۷].

در پژوهش حاضر با بهره‌گیری از مدل یاد شده، ویژگی جریان نوری به اسم LiteFlowNet به ازای فریم های هر و یدیوی رنگی استخراج می شود؛ ابعاد این ویژگی‌ها به صورت (h*w*(a,b)) می‌باشد که h برابر با ارتفاع ماتریس، w برابر با عرض ماتریس و (a,b) دوتایی زوج مرتب است که زاویه و طول حرکت را نشان می‌دهد. ابعاد ماتریس‌های به دست آمده بسیار بزرگ می‌باشد؛ لذا در ادامه‌ی روند استخراج ویژگی از نوع‌داده‌ی ویدیوی رنگی، فضاهای اضافی و محیط ساکن موجود در تصویر برش داده می‌شوند. سپس، با پنجره‌ای به ابعاد ۱۰×۱۰ بر روی ماتریس‌های حاصل حرکت کرده و به جای همه‌ی اعداد موجود در هر مربع، فقط مقدار ماکزیمم یا میانگین آن‌ها به‌عنوان نماینده قرار می‌گیرد. علاوه‌براین، هر یک از مقادیر ویژگی‌های حاصل از ویدیوهای رنگی، ماتریس‌های سه‌بعدی‌ای هستند که بُعد آخر آن‌ها دوتایی زوج مرتب می‌باشد. لذا جهت استفاده از این ماتریس‌ها، در بُعد سوم، یک بار مقدار اول دوتایی زوج مرتب (زاویه‌ی حرکت)، بار بعدی مقدار دوم آن (طول حرکت) و بار سوم حالت باز و مسطح شده‌ی آن‌ها در نظر گرفته می‌شود. در نهایت برای توالی‌های رنگی، شش حالت مختلف زیرمجموعه از ویژگی LiteFlowNet که با ساختاری به صورت "LiteFlow dim10 avg/max angle/measure/flatten" نام‌گذاری شده‌اند، ساخته می‌شود. تمامی مراحل ذکر شده در روند استخراج ویژگی از نوع‌داده‌ی رنگی در شکل ۱۲ نشان داده شده است. مطابق با فرایند قابل مشاهده در شکل ۱۲، در نهایت شش ویژگی متمایز برای نوع‌داده‌ی رنگی حاصل گردیده است. لازم به ذکر است که n نشان‌دهنده‌ی تعداد فریم‌ها در هر ویدیو می‌باشد.

ویژگی‌های این سه نوع‌داده (رنگی، ساختار اسکلتی حاصل از کینکت و یادگیری عمیق) به طبقه‌بندهای طراحی شده در روش پیشنهادی داده می‌شوند و تشخیص حرکت برای هر یک از آن‌ها به صورت جداگانه انجام می‌شود. نتایج تشخیص حرکت فقط با نوع‌داده اسکلتی دوربین کینکت و یادگیری عمیق به ترتیب در جدول های شماره ۲ و ۳ آورده شده است. همچنین نتیجه تشخیص حرکت با استفاده از ویدیو رنگی در جدول شماره ۴ نشان شده است. از معیار صحت[۲۳] برای ارزیابی روش‌های پیاده‌سازی شده استفاده می‌گردد؛ همچنین ماتریس درهم‌ریختگی[۲۴] برای نتایج به دست آمده، محاسبه می‌گردد که در آن، این موارد تعریف می‌شود: مثبت واقعی (TP): مدل به درستی متعلق بودن حرکت را به یک کلاس پیش‌بینی کرده است؛ منفی واقعی (TN): مدل به درستی عدم تعلق حرکت را به یک کلاس پیش‌بینی کرده است. مثبت کاذب (FP): مدل به اشتباه متعلق بودن حرکت را به یک کلاس پیش‌بینی کرده است و منفی کاذب (FN): مدل به اشتباه عدم تعلق حرکت را به یک کلاس پیش‌بینی کرده است؛ صحت بر اساس رابطه (۷) محاسبه می‌شود:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (۷)$$



شکل ۱۲: روند استخراج ویژگی از نوع‌داده ویدیو رنگی

جدول ۲: نتایج تشخیص حرکت با نوع‌داده اسکلتی حاصل از دوربین کینکت

ویژگی / مدل	RNN		GRU		LSTM		Bi-GRU		Bi-LSTM		Bi-LSTM-GRU	
	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا
Kinect absolute position	۵۷/۲۳٪	۲/۷۲	۷۲/۲۵٪	۱/۳۲	۷۶/۸۸٪	۱/۰۳	۸۰/۳۵٪	۰/۹۹	۷۰/۵۲٪	۱/۲۵	۷۳/۴۱٪	۱/۵۶
Kinect eigenjoints	۹۰/۳۴٪	۰/۳۶	۹۳/۷۵٪	۰/۱۸	۹۳/۷۵٪	۰/۲۴	۹۰/۹۱٪	۰/۲۷	۹۲/۰۵٪	۰/۲۶	۹۱/۴۸٪	۰/۲۷
Kinect relative IP angle position	۸۶/۱۳٪	۰/۴۸	۷۹/۷۷٪	۰/۹۳	**۹۵/۳۸٪**	**۰/۱۳**	۹۲/۴۹٪	۰/۲۸	۹۵/۳۸٪	۰/۲۷	۹۴/۸۰٪	۰/۲۴
Kinect relative IP position	۸۹/۶۰٪	۰/۴۴	۹۱/۹۱٪	۰/۳۱	۹۲/۴۹٪	۰/۳۷	۸۹/۶۰٪	۰/۴۳	۹۱/۳۳٪	۰/۲۵	۸۷/۲۸٪	۰/۵۸
Kinect relative position	۷۷/۴۶٪	۱/۱۷	۸۳/۲۴٪	۱/۳۴	۹۱/۳۳٪	۰/۲۹	۸۶/۱۳٪	۱/۰۴	۸۴/۹۷٪	۰/۵۴	۹۱/۹۱٪	۰/۳۴
Kinect trajectory3	۵۳/۴۵٪	۳/۱۲	۹۳/۶۴٪	۰/۲۷	۸۹/۰۲٪	۰/۳۲	۹۱/۳۳٪	۰/۳۶	۹۲/۴۹٪	۰/۲۳	۹۴/۲۲٪	۰/۲۳

جدول ۳: نتایج تشخیص حرکت با نوع‌داده اسکلتی حاصل از روش‌های یادگیری عمیق

ویژگی / مدل	RNN		GRU		LSTM		Bi-GRU		Bi-LSTM		Bi-LSTM-GRU	
	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا
Deep absolute position	۶۶/۴۷٪	۱/۲۷	۷۵/۷۲٪	۰/۹۰	۸۳/۸۲٪	۰/۷۱	۸۲/۶۶٪	۰/۷۳	۸۰/۳۵٪	۰/۷۹	۷۷/۴۶٪	۰/۸۹
Deep eigenjoints	۷۰/۵۲٪	۲/۸۴	۸۰/۳۵٪	۰/۸۲	۸۷/۲۸٪	۰/۵۹	۸۳/۸۲٪	۰/۷۸	۸۳/۸۲٪	۰/۸۰	۸۳/۲۴٪	۰/۷۶
Deep relative IP angle position	۷۶/۸۸٪	۰/۸۴	۸۱/۵۰٪	۰/۷۵	۸۲/۰۸٪	۰/۷۶	۸۲/۰۸٪	۰/۶۰	۸۷/۲۸٪	۰/۵۸	۷۳/۴۱٪	۱/۳۳
Deep relative IP position	۷۳/۹۹٪	۱/۰۹	۸۴/۹۷٪	۰/۷۱	**۸۹/۶۰٪**	**۰/۳۷**	۷۸/۰۳٪	۱/۰۱	۸۶/۷۱٪	۰/۷۳	۸۳/۸۲٪	۰/۸۸
Deep relative position	۶۵/۳۲٪	۱/۵۷	۸۰/۳۵٪	۱/۱۶	۷۹/۱۹٪	۰/۷۲	۷۶/۳۰٪	۱/۵۵	۸۰/۹۲٪	۰/۷۲	۷۷/۴۶٪	۰/۹۹
Deep trajectory3	۶۱/۸۵٪	۱/۴۰	۸۹/۰۲٪	۰/۵۲	۸۶/۷۱٪	۰/۴۴	۸۳/۸۲٪	۰/۵۶	۸۴/۹۷٪	۰/۴۹	۸۳/۲۴٪	۰/۶۱

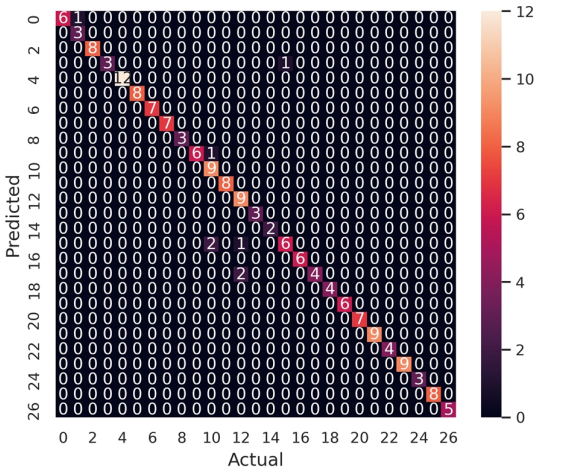
جدول ۴: نتایج تشخیص حرکت با نوع‌داده ویدیو رنگی

ویژگی / مدل	RNN		GRU		LSTM		Bi-GRU		Bi-LSTM		Bi-LSTM-GRU	
	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا
LiteFlow dim10 avg angle	۷۱/۱۰٪	۱/۰۶	۸۷/۸۶٪	۰/۵۴	۹۲/۴۹٪	۰/۳۷	۸۶/۱۳٪	۰/۵۰	۹۳/۰۶٪	۰/۳۱	۹۳/۰۶٪	۰/۳۵
LiteFlow dim10 avg measure	۵۸/۳۸٪	۱/۶۲	۷۲/۲۵٪	۰/۹۸	۸۲/۰۸٪	۰/۶۱	۷۷/۴۶٪	۰/۸۶	۸۵/۵۵٪	۰/۵۶	۷۹/۷۷٪	۰/۶۹
LiteFlow dim10 avg flatten	۶۵/۹۰٪	۱/۱۳	۸۵/۵۵٪	۰/۵۲	۸۷/۸۶٪	۰/۳۹	۸۶/۷۱٪	۰/۵۹	۹۳/۶۴٪	۰/۲۶	۸۹/۶۰٪	۰/۴۸
LiteFlow dim10 max angle	۶۸/۷۹٪	۱/۲۴	۸۵/۵۵٪	۰/۴۹	۹۱/۳۳٪	۰/۲۵	۸۸/۴۴٪	۰/۴۲	۹۳/۰۶٪	۰/۲۴	۹۰/۱۷٪	۰/۳۴
LiteFlow dim10 max measure	۵۸/۳۸٪	۱/۴۳	۷۶/۸۸٪	۰/۸۰	۸۴/۹۷٪	۰/۴۵	۷۷/۴۶٪	۰/۷۴	۸۷/۸۶٪	۰/۴۲	۸۲/۶۶٪	۰/۵۵
LiteFlow dim10 max flatten	۶۸/۷۹٪	۱/۱۲	۸۸/۴۴٪	۰/۴۶	۹۲/۴۹٪	۰/۲۸	۹۱/۳۳٪	۰/۴۲	**۹۴/۲۲٪**	**۰/۲۹**	۹۲/۴۹٪	۰/۴۰

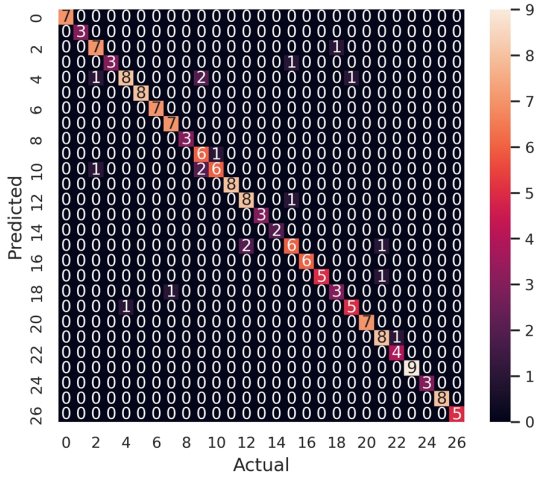
همانطور که از جدول ۲ مشخص است ویژگی Kinect relative IP angle position با صحت ٪۹۵/۳۸ بهترین نتیجه را با دو مدل LSTM و Bi-LSTM کسب کرده است که در این بین، مدل LSTM با مقدار خطای کمتر و برابر ۰/۱۳ برتر است. در بین ویژگی‌های داده‌ی حاصل شده از روش های یادگیری عمیق، طبق جدول ۳ ویژگی Deep relative IP position با صحت ٪۸۹/۶۰ و مقدار خطای ۰/۳۷ به بهترین نتیجه با مدل LSTM رسیده است. همچنین آن‌گونه که در جدول ۴ مشهود است در بین ویژگی های نوع‌داده‌ی ویدیو رنگی، ویژگی LiteFlow dim10 max flatten با مدل Bi-LSTM به صحت ٪۹۴/۲۲ و مقدار خطای ۰/۲۹ دست یافته است. پژوهش حاضر، در زمینه‌ی تشخیص حرکت با تک نوع‌داده‌های ساختار اسکلتی و ویدیو رنگی در مقایسه با پژوهش مشابه [۳۰] از پیشرفت خوب و تأثیرگذاری برخوردار بوده است که در جدول ۵ قابل مشاهده می‌باشد. ماتریس درهم‌ریختگی بهترین نتایج این مرحله، در شکل‌های شماره ۱۳، ۱۴ و ۱۵ آورده شده است. مطابق با این ماتریس های درهم‌ریختگی، مقادیر کلاس های پیش‌بینی شده تطابق بالایی با مقدار واقعی آن‌ها دارد. تفاوت اعداد موجود روی قطر اصلی ماتریس‌ها با درایه‌های دیگر، نشان از عملکرد مطلوب فرایند پیشنهادی برای تشخیص حرکات انسانی با تک نوع‌داده‌های به کار گرفته شده دارد.

جدول ۵: مقایسه ویژگی‌های روش پیشنهادی با دیگر

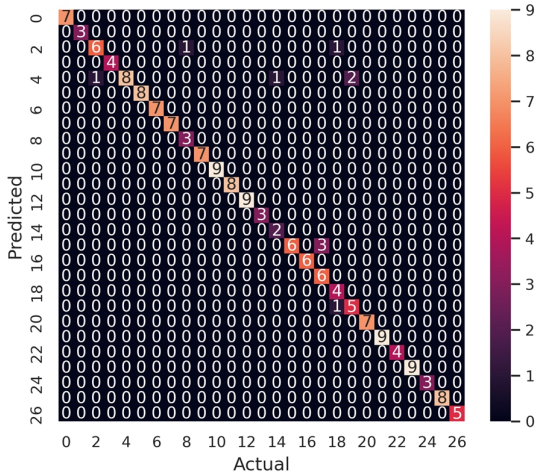
مجموعه ویژگی	نوع‌داده اسکلت		نوع‌داده ویدیو رنگی
	کینکت	یادگیری عمیق	
مجموعه ویژگی مرجع [۳۰]	٪۸۵/۳۵	-	٪۸۲/۵۶
مجموعه ویژگی روش پیشنهادی	**٪۹۵/۳۸**	٪۸۹/۶۰	**٪۹۴/۲۲**



شکل ۱۳: ماتریس درهم‌ریختگی بهترین نتیجه کینکت



شکل ۱۴: ماتریس درهم‌ریختگی بهترین نتیجه یادگیری عمیق



شکل ۱۵: ماتریس درهم‌ریختگی بهترین نتیجه ویدیو رنگی

بعد از تشخیص حرکت با تک نوع‌داده‌ها، به هدف بهره‌گیری از مزایای هر دوی آن‌ها و قابلیت تشخیص دقیق‌تر و بهتر حرکات به پیاده‌سازی روش های تلفیق پرداخته می‌شود. برای انجام تلفیق، فریم‌های هر نوع‌داده با اعمال الگوهای خاص ریاضیاتی با یکدیگر یکسان می‌شوند تا ابعاد ماتریس‌ها با یکدیگر سازگار و قابل تلفیق شوند. تلفیق با دو روش BGLPCCA و MFA مطرح شده در روش پیشنهادی انجام می‌شود. برای انجام این دو روش علاوه بر اطلاعات حرکت در هر فریم، به ماتریس‌های الحاق شده‌ی تمام ویدیوها نیاز است، لذا در هر ویژگی تمام فریم‌های ویدیوها الحاق و در یک ماتریس بزرگ نگهداری می‌شوند؛ همچنین این اقدام برای برچسب‌های هر کلاس نیز صورت می‌گیرد. در انتها، ویژگی‌های تلفیق شده در سطح فریم به دسته‌بندهای طراحی شده ارسال و تشخیص حرکات انسانی انجام می‌شود.

جدول ۶: بهترین نتایج حاصل از دو روش تلفیق روی دو نوع‌داده ویدیو رنگی و اسکلت سه‌بعدی در تشخیص حرکات انسانی

تلفیق ویژگی / مدل	RNN		GRU		LSTM		Bi-GRU		Bi-LSTM		Bi-LSTM-GRU	
	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا	صحت	خطا
BGLPCCA - LiteFlow dim10 max angle & Kinect relative IP angle position	۷۷/۳۸٪	۰/۸۱	۸۹/۲۹٪	۰/۳۵	**۹۷/۰۲٪**	**۰/۱۲**	۸۹/۸۸٪	۰/۳۷	۹۴/۶۴٪	۰/۱۵	۸۱/۵۵٪	۰/۹۴
BGLPCCA - LiteFlow dim10 avg angle & Kinect relative IP angle position	۷۰/۲۴٪	۲/۳۹	۸۴/۵۲٪	۱/۱۱	۹۷/۰۲٪	۰/۱۳	۸۸/۶۹٪	۰/۴۲	۹۱/۶۷٪	۰/۴۸	۹۵/۲۴٪	۰/۲۰
BGLPCCA - LiteFlow dim10 avg angle & Deep trajectory3	۷۵/۶۰٪	۰/۹۰	۹۲/۸۶٪	۰/۳۲	۹۷/۰۲٪	۰/۱۴	۸۹/۲۹٪	۰/۴۰	۹۵/۸۳٪	۰/۱۶	۹۴/۰۵٪	۰/۲۳
MFA - LiteFlow dim10 max measure & Kinect relative IP position	۸۳/۳۳٪	۰/۵۰	۸۱/۵۵٪	۰/۷۵	۹۵/۲۴٪	۰/۱۷	۹۱/۶۷٪	۰/۲۹	۹۱/۶۷٪	۰/۲۸	۸۹/۸۸٪	۰/۲۷
MFA - LiteFlow dim10 max flatten & Kinect relative IP position	۸۷/۵۰٪	۰/۳۸	۹۲/۲۶٪	۰/۲۷	۹۵/۲۴٪	۰/۲۳	۹۵/۸۳٪	۰/۱۳	۹۱/۶۷٪	۰/۳۰	۸۹/۸۸٪	۰/۳۶

معیار ارزیابی در این پژوهش، صحت می‌باشد که طبق فرمول (۷) قابل محاسبه است. برای تلفیق ویژگی‌ها، یک مرتبه ویژگی‌های نوع‌داده اسکلتی حاصل از دوربین کینکت و یک مرتبه هم ویژگی‌های نوع‌داده اسکلتی حاصل از روش‌های یادگیری عمیق با نوع‌داده ویدیو رنگی تلفیق می‌شوند. بعد از تلفیق دو به دوی تمام ویژگی‌ها با یکدیگر و به دست آمدن نتایج به کمک طبقه‌بندها، بهترین نتیجه مربوط به روش BGLPCCA در تلفیق ویژگی LiteFlow dim10 max angle با Kinect relative IP angle position با صحت مطلوب ٪۹۷/۰۲ و خطای ۰/۱۲ می‌باشد. همچنین با همین روش، تلفیق ویژگی LiteFlow dim10 avg angle با ویژگی‌های Kinect relative IP angle position و Deep trajectory3 نیز به صحت مطلوب ٪۹۷/۰۲ و به ترتیب به میزان خطای ۰/۱۳ و ۰/۱۴ رسیدند که جزء بهترین نتایج هستند. به دلیل تعداد زیاد نتایج، از هر یک از دو روش تلفیق، بهترین نتیجه‌ی آن‌ها با مدل‌های مختلف در جدول ۶ آورده شده است. همان‌طور که از این جدول مشخص است روش BGLPCCA عملکرد بهتری نسبت به روش MFA داشته است و به میزان صحت بالا و به‌سزایی رسیده است.

شکل‌های ۱۶ و ۱۷ نشان‌دهنده‌ی نمودار تغییرات صحت و خطای بهترین نتیجه‌ی حاصل از این پژوهش در روند آموزش می‌باشد و در شکل ۱۸ ماتریس درهم‌ریختگی بهترین نتیجه‌ی مقاله مذکور نشان داده شده است. در جدول ۷ نتیجه کار با پژوهش مشابه [۳۰] و سایر پژوهش‌های مشابه و به‌روزی که تا کنون در این زمینه و بر روی مجموعه‌داده‌ی مورد استفاده در این مقاله، انجام شده است مورد مقایسه قرار گرفته است. همان‌طور که از جدول ۷ مشهود است پژوهش حاضر از پژوهش‌های دیگر برتر بوده است. جدول مذکور، نشان‌دهنده‌ی افزایش چشم‌گیر میزان صحت و کیفیت بالای پژوهش حاضر در تشخیص حرکات انسانی می‌باشد.

۴- نتیجه‌گیری

مسئله تشخیص حرکات انسانی به دلیل گسترش روزافزون و ورود هوش مصنوعی به زندگی‌های امروزه، مسئله‌ی مهمی می‌باشد. این مسئله، در زمینه‌های بسیاری نظیر ساخت ربات، مباحث امنیتی و دستگیری سارقان، مسائل پزشکی و فیزیوتراپی، مراقبت‌های بهداشتی، سرگرمی، ساخت وسایل ورزشی و اصلاحی، بازی‌های کامپیوتری، رانندگی هوشمند و خودروی خودران، متاورس و واقعیت مجازی، دوربین‌های نظارتی و مدار بسته، معماری، صنعت، آموزش و مدیریت کاربرد دارد.

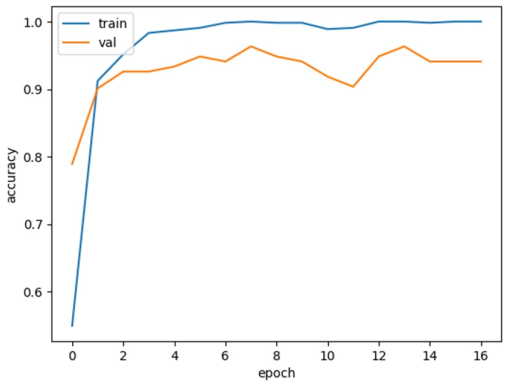
در این مقاله به‌منظور افزایش مقدار صحت در تشخیص حرکات انسانی، از دو نوع‌داده‌ی ساختار اسکلتی و ویدیو رنگی و تلفیق ویژگی‌های آن‌ها استفاده شده که باعث بهبود قابل توجهی نیز شده است. این میزان بهبود عملکرد ناشی از استفاده صحیح از ابزارهای به‌کار گرفته شده از جمله نوع‌داده‌های کارا، به‌کارگیری روش‌های نوین در تخمین نقاط اسکلتی، روش‌های مناسب در استخراج ویژگی، تلفیق ویژگی‌ها به صورت کارآمد و طراحی طبقه‌بندهای قوی با یادگیری عمیق می‌باشد که همگی نقاط قوت کار محسوب می‌شوند.

برای تخمین نقاط اسکلتی، علاوه بر به‌کارگرفتن دوربین کینکت از یادگیری عمیق نیز استفاده شد به‌گونه‌ای که مدل به‌روز شبکه عصبی مورد استفاده بر روی مجموعه‌داده خود سفارشی سازی شد و ساختار جدیدی از مختصات اسکلت بدن را ارائه داد. ساختار اسکلتی به دست آمده از یادگیری عمیق به دلیل از پیش آموزش دیده بودن مدل بر روی مجموعه‌داده‌های بزرگ با مختصات واقعی مفاصل، از دقت خوبی برخوردار است و در زمانی که شرایط محیطی برای استفاده از دوربین کینکت مهیا نباشد، جایگزین بسیار مناسبی برای آن است. همچنین، در کاهش هزینه دستگاه‌هایی نظیر کینکت موثر واقع می‌شود.

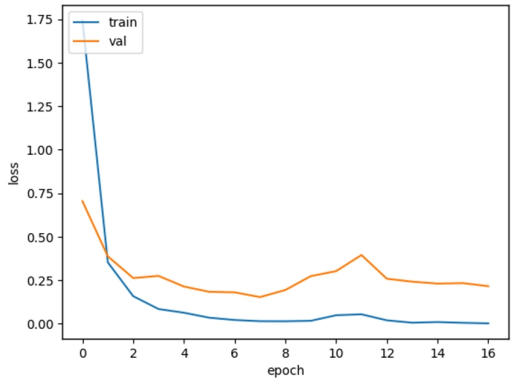
به‌منظور استخراج ویژگی برای نوع‌داده‌ی اسکلتی از روش‌های ریاضیاتی و برای نوع‌داده‌ی ویدیوی رنگی از شبکه‌های عصبی نوین استفاده شد و تشخیص حرکت با صحت بالا انجام گردید. پژوهش، در پردازش ویدیوهای رنگی با محدودیت‌های بسیار نظیر احتیاج به حافظه موقت[۲۵] زیاد و پردازش سنگین مواجه بود که با به‌کارگیری فضای ابری مرتفع گشت. در ادامه برای بهره‌مندی از مزایای حداکثری هر نوع‌داده و کسب اطلاعات بیشتر، از روش‌های جدید تلفیق در فضای مشترک به هنگام تشخیص حرکت استفاده شد. مقدار صحت قابل توجه حاصل از این مقاله برابر ٪۹۷/۰۲ می‌باشد که با افزایش ۳/۵۳ درصدی نسبت به

پژوهش مشابه [۳۰] و افزایش ۱/۷۲ درصدی نسبت به آخرین پژوهش انجام شـــده [۴۲] بر روی مجموعه‌داده‌ی UTD-MHAD، باعث بهبود کیفیت در شناسایی حرکت‌های انسانی شده است.

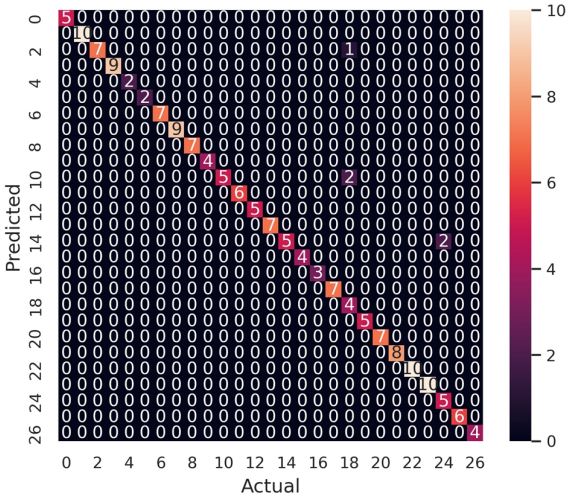
به‌عنوان کارهای پیشنهادی آینده پژوهش می‌توان تعداد متنوع‌تری از نوع‌داده‌های مختلف را به کار گرفت؛ همچنین اســـتفاده از مدل‌های بزرگ و پیشـــرفته‌تر مبتنی بر شـــبکه‌های عصـــبی به هنگام اســـتخراج ویژگی می‌تواند در بهبود هر چه بیشـــتر این فرایند مفید واقع شـــود؛ علاوه بر این در قســـمت روش تلفیق اســـتفاده شـــده در چارچوب پیشنهادی پژوهش نیز می‌توان از سبک‌های دیگر تلفیق نظیر روش‌های مبتنی بر یادگیری عمیق بهره گرفت.



شکل ۱۶: نمودار صحت بهترین نتیجه‌ی پژوهش



شکل ۱۷: نمودار خطای بهترین نتیجه‌ی پژوهش



شکل ۱۸: ماتریس درهم‌ریختگی بهترین نتیجه تلفیق به روش BGLPCCA

جدول ۷: مقایسه صحت نهایی پژوهش با سایر پژوهش‌ها

پژوهش	تشخیص حرکت بر روی مجموعه‌داده UTD-MHAD
مرجع [۳۰] (۲۰۱۸)	۹۳/۴۹٪
مرجع [۳۵] (۲۰۲۱)	۹۳/۵۷٪
مرجع [۳۶] (۲۰۲۱)	۹۵/۰۷٪
مرجع [۳۷] (۲۰۲۲)	۹۶/۳۰٪
مرجع [۳۸] (۲۰۲۳)	۹۱/۶۱٪
مرجع [۳۹] (۲۰۲۳)	۹۳/۹۰٪
مرجع [۴۰] (۲۰۲۳)	۹۴/۴۱٪
مرجع [۴۱] (۲۰۲۳)	۹۵/۱۰٪
مرجع [۴۲] (۲۰۲۳)	۹۵/۳۰٪
روش پیشنهادی	**۹۷/۰۲٪**

مراجع

[1] Jegham, Imen, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. "Vision-based human action recognition: An overview and real world challenges." *Forensic Science International: Digital Investigation* 32 (2020): 200901.

[2] Mimouna, Amira, Anouar Ben Khalifa, and Najoua Essoukri Ben Amara. "Human action recognition using triaxial accelerometer data: selective approach." In *2018 15th International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 491-496. IEEE, 2018.

[3] Li, Meng, Howard Leung, and Hubert PH Shum. "Human action recognition via skeletal and depth based feature fusion." In *Proceedings of the 9th International Conference on Motion in Games*, pp. 123-132. 2016.

[4] Li, Chuankun, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. "Skeleton-based action recognition using LSTM and CNN." In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 585-590. IEEE, 2017.

[5] Das, Srijan, Michal Koperski, Francois Bremond, and Gianpiero Francesca. "Action recognition based on a mixture of RGB and depth based skeleton." In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1-6. IEEE, 2017.

[6] Ghojogh, Benyamin, Hoda Mohammadzade, and Mozhgan Mokari. "Fisherposes for human action recognition using kinect sensor data." *IEEE Sensors Journal* 18, no. 4 (2017): 1612-1627.

[7] Liang, Bin, and Lihong Zheng. "A survey on human action recognition using depth sensors." In *2015 International conference on digital image computing: techniques and applications (DICTA)*, pp. 1-8. IEEE, 2015.

[8] Wang, Heng, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. "Dense trajectories and motion boundary descriptors for action recognition." *International journal of computer vision* 103, no. 1 (2013): 60-79.

[9] Gkioxari, Georgia, and Jitendra Malik. "Finding action tubes." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 759-768. 2015.

[10] Chaaraoui, Alexandros Andre, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. "Evolutionary joint selection to improve human action recognition with RGB-D devices." *Expert systems with applications* 41, no. 3 (2014): 786-794.

[11] Li, Bo, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN." In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 601-604. IEEE, 2017.

[12] Wang, Pichao, Wanqing Li, Chuankun Li, and Yonghong Hou. "Action recognition based on joint trajectory maps with convolutional neural networks." *Knowledge-Based Systems* 158 (2018): 43-53.

[13] Chi, Hyung-gun, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. "InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20186-20196. 2022.

[14] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "Fusion of depth, skeleton, and inertial data for human action recognition." In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2712-2716. IEEE, 2016.

[15] Khaire, Pushpajit, Praveen Kumar, and Javed Imran. "Combining CNN streams of RGB-D and skeletal data for human activity recognition." *Pattern Recognition Letters* 115 (2018): 107-116.

[16] Elmadany, Nour El Din, Yifeng He, and Ling Guan. "Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis." *IEEE Transactions on Image Processing* 27, no. 11 (2018): 5275-5287.

[17] Yu, Jiahui, Hongwei Gao, Wei Yang, Yueqiu Jiang, Weihong Chin, Naoyuki Kubota, and Zhaojie Ju. "A discriminative deep model with feature fusion and temporal attention for human action recognition." *IEEE Access* 8 (2020): 43243-43255.

[18] Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli. "3d human pose estimation in video with temporal convolutions and semi-supervised training." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753-7762. 2019.

[19] Girshick, R., Radosavovic, I., & Gkioxari, G. (2018). Piotr Doll´ar, and Kaiming He, "Detectron,".

[20] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.

[21] Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli. "3d human pose estimation in video with temporal convolutions and semi-supervised training." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7753-7762. 2019.

[22] Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." *IEEE transactions on pattern analysis and machine intelligence* 36, no. 7 (2013): 1325-1339.

[23] Sigal, Leonid, Alexandru O. Balan, and Michael J. Black. "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." *International journal of computer vision* 87, no. 1-2 (2010): 4-27.

[24] Vemulapalli, Raviteja, Felipe Arrate, and Rama Chellappa. "Human action recognition by representing 3d skeletons as points in a lie group." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588-595. 2014.

[25] Ilg, Eddy, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. "Flownet 2.0: Evolution of optical flow estimation with deep networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462-2470. 2017.

[26] Hui, Tak-Wai, Xiaoou Tang, and Chen Change Loy. "A lightweight optical flow cnn—revisiting data fidelity and regularization." *IEEE transactions on pattern analysis and machine intelligence* 43, no. 8 (2020): 2555-2569.

[27] Hui, Tak-Wai, Xiaoou Tang, and Chen Change Loy. "Liteflownet: A lightweight convolutional neural network for optical flow estimation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8981-8989. 2018.

[28] Belhumeur, Peter N., Joao P. Hespanha, and David J. Kriegman. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." *IEEE Transactions on pattern analysis and machine intelligence* 19, no. 7 (1997): 711-720.

[29] Yan, Shuicheng, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. "Graph embedding and extensions: A general framework for dimensionality reduction." *IEEE transactions on pattern analysis and machine intelligence* 29, no. 1 (2006): 40-51.

[30] Elmadany, Nour El Din, Yifeng He, and Ling Guan. "Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis." *IEEE Transactions on Image Processing* 27, no. 11 (2018): 5275-5287.

[31] Hardoon, David R., Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods." *Neural computation* 16, no. 12 (2004): 2639-2664.

[32] He, Xiaofei, and Partha Niyogi. "Locality preserving projections." *Advances in neural information processing systems* 16 (2003).

[33] Huang, Sheng, Ahmed Elgammal, Luwen Huangfu, Dan Yang, and Xiaohong Zhang. "Globality-locality preserving projections for biometric data dimensionality reduction." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 15-20. 2014.

[34] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor." In *2015 IEEE International conference on image processing (ICIP)*, pp. 168-172. IEEE, 2015.

[35] Wu, Hanbo, Xin Ma, and Yibin Li. "Spatiotemporal multimodal learning with 3D CNNs for video action recognition." *IEEE Transactions on Circuits and Systems for Video Technology* 32.3 (2021): 1250-1261.

[36] Weiyao, Xu, et al. "Fusion of skeleton and RGB features for RGB-D human action recognition." *IEEE Sensors Journal* 21.17 (2021): 19157-19164.

[37] Cha, Junuk, et al. "Learning 3D skeletal representation from transformer for action recognition." *IEEE Access* 10 (2022): 67541-67550.

[38] Li, Xing, et al. "PointMapNet: Point Cloud Feature Map Network for 3D Human Action Recognition." *Symmetry* 15.2 (2023): 363.

[39] Zhong, Zhuokun, et al. "Multimodal cooperative self□attention network for action recognition." *IET Image Processing* 17.6 (2023): 1775-1783.

[40] Li, Xing, Qian Huang, and Zhijian Wang. "Spatial and temporal information fusion for human action recognition via Center Boundary Balancing Multimodal Classifier." *Journal of Visual Communication and Image Representation* 90 (2023): 103716.

[41] Sun, Yaohui, et al. "Integrating vision transformer-based bilinear pooling and attention network fusion of rgb and skeleton features for human action recognition." *International Journal of Computational Intelligence Systems* 16.1 (2023): 116.

[42] Zhang, Man, Xing Li, and Qianhan Wu. "Spatio-Temporal Information Fusion and Filtration for Human Action Recognition." *Symmetry* 15.12 (2023): 2177.

زیرنویس‌ها

[۱] Human Action Recognition
[۲] Noise
[۳] RGB
[۴] Kinect
[۵] Skeleton
[۶] Conciseness
[۷] Robustness
[۸] View-Independent
[۹] Motion Boundary
[۱۰] Action Tubes
[۱۱] Joint Trajectory Maps
[۱۲] 3D Human Pose
[۱۳] 2D Keypoints
[۱۴] Region-based CNN
[۱۵] Temporal Convolution
[۱۶] Joint Positions
[۱۷] Joint Angels
[۱۸] Relative Joints Positions
[۱۹] Optical Flow
[۲۰] Recurrent Neural Networks
[۲۱] Long Short-Term Memory
[۲۲] Gated Recurrent Unit
[۲۳] Accuracy
[۲۴] Confusion Matrix
[۲۵] RAM