دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی

تشخیص فعالیت انسان مبتنی بر اسکلت به همراه تخمین حالت دو بعدی بدن

نگارش

هدیه پورقاسم

استاد راهنما

دکتر محمد رحمتی

خرداد ۱۴۰۲

سپاس‌گزاری

بدین وسیله از زحمات و تلاش بی‌دریغ استاد راهنمای بزرگوارم جناب آقای دکتر رحمتی در طی انجام این پایان‌نامه، صمیمانه سپاس‌گزاری می‌نمایم.

همچنین از استاد گرانقدر، جناب آقای دکتر نیک‌آبادی که زحمت داوری این پایان‌نامه را بر عهده داشتند و به مطالعه‌ی آن پرداختند، نهایت تشکر را دارم.

در نهایت، از سایر همکاران و دوستانی که هر کدام به نحوی در تهیه این گزارش با اینجانب همکاری داشته‌اند تشکر نموده و موفقیت همه آن‌ها را از خداوند متعال خواهانم.

هدیه پورقاسم

خرداد ۱۴۰۲

چکیده

مساله‌ی تشخیص فعالیت انسان به دنبال ایجاد الگوریتم‌ها، روش‌ها و چارچوب‌هایی برای شناسایی خودکار اقدامات انجام شده در یک ویدیو است. از فناوری تشخیص فعالیت انسان می‌توان به طور گسترده در تشخیص‌های پزشکی، کنترل نرخ جرم و جنایت، نظارت بر بیماران یا سالمندان و صنایع دیگر استفاده کرد. در سال‌های اخیر روش‌های تشخیص فعالیت مبتنی بر اسکلت که اعمال را از روی یک دنباله‌ی ورودی از مفاصل اسکلتی شناسایی می‌کنند، توجه زیادی را در جامعه پژوهشی و در حوزه‌ی تشخیص فعالیت به خود جلب کرده‌اند. نمایش مبتنی بر اسکلت استخراج شده از ویدیوی حرکت انسان اطلاعات قابل توجهی منتقل می‌کند. همچنین، داده‌های اسکلتی فشرده‌اند و به طور قابل توجهی هزینه محاسباتی را در مساله‌ی تشخیص فعالیت کاهش می‌دهند. کاهش هزینه محاسباتی، در دسترس بودن داده‌های اسکلتی و بهبود الگوریتم‌های تخمین حالت بدن باعث محبوبیت روش‌های فعلی تشخیص فعالیت مبتنی بر اسکلت شده‌اند.

در این پروژه دو بخش تخمین حالت دو بعدی بدن انسان (بدست آوردن توالی اسکلتی) و تشخیص فعالیت مبتنی بر اسکلت پیاده‌سازی شده‌اند. به صورتی که در ابتدا، با استفاده از مجموعه داده‌های ویدیویی موجود از فعالیت‌های انسان، به استخراج توالی اسکلتی به کمک مدل‌های تخمین حالت دوبعدی پرداخته شده است. سپس دو مدل یادگیری عمیق مبتنی بر شبکه‌های کانولوشنی گرافی برای تشخیص فعالیت مبتنی بر اسکلت راه‌اندازی شده و به کمک داده‌های اسکلتی پیش‌تر استخراج شده، آموزش داده شده‌اند. در نهایت، مدل‌های آموزش داده شده از نظر عملکرد (دقت و هزینه) مورد مقایسه قرار گرفته‌اند. یک سامانه‌ی تحت وب به عنوان رابط کاربری نیز پیاده‌سازی شده است تا کاربران بتوانند یک ویدیو را به عنوان ورودی بارگذاری کرده و در خروجی اسکلت تخمین زده شده و فعالیت تشخیص داده شده را دریافت کنند.

واژه‌های کلیدی:

تشخیص فعالیت انسان مبتنی بر اسکلت، تخمین حالت بدن، شبکه عصبی کانولوشنی گرافی

فهرست مطالب

صفحه

فهرست اشکال

صفحه

فهرست جداول

صفحه

فصل اول

مقدمه

تشخیص فعالیت[۱]، یک مساله‌ی اساسی در مباحث بینایی ماشین[۲] است که وظیفه‌ی آن شناسایی و طبقه‌بندی خودکار اعمال انسان از روی داده‌های بصری است. مطالعات موجود روش‌های مختلفی را برای نمایش ویژگی‌ها در مسائل تشخیص فعالیت مورد بررسی قرار داده‌اند، مانند قاب[۳]های رنگی، جریان‌های نوری، امواج صوتی و اسکلت‌های انسانی. در میان این روش‌ها، تشخیص عمل مبتنی بر اسکلت به دلیل ماهیت فعالیت محور و فشردگی آن در سال‌های اخیر مورد توجه فزاینده‌ای قرار گرفته است. داده‌های اسکلتی فقط اطلاعات ژست گنجانده شده است، دنباله‌های اسکلتی[۴] تنها اطلاعات مربوط به عمل را ضبط می‌کنند و از مزاحمت‌های زمینه‌ای، مانند تغییرات پس‌زمینه و تغییرات نور مصون هستند. در حالی که رویکردهای سنتی برای تشخیص فعالیت به شدت به تصاویر یا ویدیوهای رنگی متکی هستند که اغلب با چالش‌هایی مانند تغییرات دیدگاه و ابعاد بالای داده‌ی رنگی مواجه اند[۱]. در نتیجه، ظهور تشخیص فعالیت مبتنی بر اسکلت[۵] نتایج امیدوارکننده‌ای را در غلبه بر این محدودیت‌ها نشان داده است.

در این بخش مروری بر پیشینه و انگیزه موضوع خواهیم داشت و در این حین به کاربردها و اهمیت مساله میپردازیم. سپس، چالش‌های موجود در مسائل تشخیص فعالیت را بررسی می‌کنیم و بعد به توضیح اهداف این پژوهش و ساختار سامانه مورد نظر را بررسی می‌کنیم. در نهایت مروری بر ساختار این گزارش خواهیم داشت.

۱-۱- پیشینه و انگیزه

تشخیص فعالیت مبتنی بر اسکلت، زیرمجموعه ای از مسائل تشخیص فعالیت است که در آن فعالیت‌های مختلف انسان به کمک دنباله‌ای از اسکلت‌ها (مفصل‌بندی بدن انسان) در واحد زمان تشخیص داده می‌شوند. داده‌های مبتنی بر اسکلت را می‌توان از دستگاه‌های ضبط حرکت یا الگوریتم‌های تخمین حالت بدن از ویدیوها به‌دست آورد. در این مسائل داده‌ها دنباله‌ای از قاب‌ها هستند و هر قاب مجموعه‌ای از مختصات مشترک خواهد داشت.

روش‌های مبتنی بر اسکلت از اطلاعات مفصل اسکلتی به‌دست‌آمده توسط حسگرهای عمق یا داده‌های رنگی برای نمایش اعمال انسان استفاده می‌کنند و امکان تشخیص قوی‌تر و کارآمدتر را فراهم می‌سازند. علاوه بر این، تشخیص عمل مبتنی بر اسکلت پتانسیل قابل توجهی را در سناریوهای دنیای واقعی نشان داده است، جایی که مقاوم بودن[۶] الگوریتم در برابر تغییر ناپذیری دیدگاه بسیار مهم است.

[1] Action Recognition

[2] Computer vision

[3] Frame

[4] Skeleton sequences

[5] Skeleton-based Action Recognition

[6] Robustness

۱-۱-۱- کاربردها

تشخیص فعالیت نقش مهمی در طیف گسترده‌ای از برنامه‌های کاربردی در زمینه‌های مختلف ایفا می‌کند. در زمینه نظارت و امنیت، فرآیند تشخیص فعالیت امکان شناسایی و تجزیه و تحلیل اعمال مشکوک یا غیرعادی را فراهم می‌کند و به پیشگیری از جرایم و افزایش امنیت عمومی کمک می‌کند. در زمینه تعامل انسان و رایانه، تشخیص عمل، رابط‌های طبیعی و بصری را قادر می‌سازد و به کاربران امکان می‌دهد با استفاده از حرکات یا حرکات بدن با رایانه‌ها یا دستگاه‌ها تعامل داشته باشند. همچنین، تشخیص فعالیت کاربرد قابل توجهی در تجزیه و تحلیل ورزشی دارد، جایی که به ردیابی و تجزیه و تحلیل عملکرد ورزشکاران کمک می‌کند و مربیان را قادر می‌سازد تا بینش‌های ارزشمندی جهت بهبود مهارت ارائه دهند. علاوه بر این، در نظارت بر مراقبت‌های بهداشتی، تشخیص فعالیت می‌تواند برای ارزیابی تمرینات فیزیوتراپی برای اطمینان از ایمنی بیمار و پیشرفت توانبخشی استفاده شود. اهمیت تشخیص فعالیت در این کاربردهای مختلف، بر نیاز به الگوریتم‌های دقیق و قوی برای استخراج اطلاعات معنی‌دار از داده‌های بصری و امکان درک خودکار اعمال انسان تأکید می‌کند.

۱-۱-۲- رویکردهای یادگیری عمیق در تشخیص فعالیت مبتنی بر اسکلت

بسیاری از تحقیقات تشخیص فعالیت مبتنی بر اسکلت در سال‌های گذشته بر اساس شبکه‌های عصبی بازگشتی[۷] (RNN) بوده است. فعالیت انسان به تغییرات پویایی بدن انسان در طول زمان اشاره دارد و استخراج ویژگی‌های بدن انسان در مقاطع زمانی مختلف می‌تواند نمایانگر خوبی از فعالیت انسان باشد. بنابراین، تشخیص فعالیت بر اساس توالی‌های اسکلتی اغلب به عنوان یک مساله‌ی سری زمانی در نظر گرفته می‌شود. شبکه‌های عصبی بازگشتی، به صورت بازگشتی به اطلاعات ورودی به ترتیب اتصال پاسخ می‌دهند، به طور گسترده برای حل مسائل ترتیبی[۸] مانند مدل‌سازی زبان و تحلیل ویدیو استفاده می‌شوند. به طور مشابه، RNNها مزایایی در تشخیص فعالیت مبتنی بر اسکلت نشان داده‌اند.

برای رفع کاستی‌های RNN برای استخراج ویژگی‌های مکانی[۹]، بسیاری از مدل‌های تشخیص عملکرد شبکه‌های عصبی کانولوشنی[۱۰] (CNN) بر اساس توالی‌های اسکلتی پدید آمده‌اند. CNNها برای استخراج اطلاعات سطح بالا بسیار سودمند هستند و به طور گسترده برای حذف و یادگیری ویژگی‌های مکانی و زمانی توالی‌های اسکلتی استفاده می‌شوند. یکی از مزیت های کلیدی CNNها در تشخیص فعالیت مبتنی بر اسکلت، توانایی آنها در یادگیری ویژگی‌های سلسله مراتبی است. با استفاده از چندین لایه کانولوشنال می‌توانند الگوهای حرکت محلی را ثبت کنند و به تدریج به نمایش‌های جهانی فعالیت‌ها بپردازند. این استخراج ویژگی سلسله مراتبی به شبکه‌ها اجازه می‌دهد تا تفاوت‌های ظریف و روابط پیچیده بین مفاصل اسکلتی را تشخیص دهند و قدرت تمایز مدل را افزایش دهند.

---

[7] Recurrent Neural Networks

[8] Sequential

[9] Spatial

[10] Convolutional Neural Networks

اسکلت انسان اساساً داده‌ای با ساختار گرافی است، با مفاصل به عنوان گره‌ها در گراف و استخوان‌ها به عنوان یال‌هایی که گره‌ها را به هم متصل می‌کنند. از شبکه‌های کانولوشنی گرافی[11] برای محاسبه پیچیدگی در گره‌های متصل شده توسط یال‌ها به طور مؤثر استفاده می‌کنند. با این حال، مطالعه نشان داده که مناسب‌تر است توالی‌های اسکلتی را به عنوان ساختارهای گراف در نظر بگیریم [۲]. به طور خاص، ظهور روش‌های مبتنی بر گراف (شبکه‌های عصبی کانولوشنی گرافی)، که به خوبی با ویژگی‌های داده‌های اسکلتی مطابقت دارند، آینده امیدوارکننده‌ای را برای تشخیص فعالیت بر اساس توالی‌های اسکلتی ارائه می‌کنند[۳, ۴].

۱-۱-۳- اهمیت الگوریتم استخراج اسکلت

در مسائل تشخیص فعالیت مبتنی بر اسکلت، الگوریتم استخراج اسکلت از اهمیت بالایی برخوردار است، زیرا به طور مستقیم بر کیفیت و دقت نمایش اسکلتی تأثیر می گذارد. یک الگوریتم موثر باید به طور دقیق مفاصل اسکلتی را شناسایی کند. نمایش اسکلت به دست آمده باید ساختار مکانی و زمانی اعمال انسان را به درستی به تصویر بکشد زیرا، این نمایش قوی به عنوان پایه‌ای برای مساله‌ی شناسایی فعالیت به کار می‌رود.

الگوریتم استخراج اسکلت باید تکنیک‌هایی را برای کاهش نویز و فیلتر کردن اطلاعات نامربوط ترکیب کند، بنابراین یک الگوریتم استخراج اسکلت ایده‌آل کیفیت داده‌های اسکلتی را بهبود می‌بخشد و بدست آوردن نتایج تشخیص فعالیت قابل اعتماد را ممکن می‌سازد. یک الگوریتم استخراج اسکلت که به خوبی طراحی شده باشد باید بتواند تنوع در دیدگاه[12] را مدیریت کند و نمایش‌های اسکلتی ثابتی را در زوایای مختلف دوربین ایجاد کند. این تغییر ناپذیری دیدگاه بسیار مهم است، زیرا اعمال انسان را می توان از دیدگاه‌های متعدد مشاهده کرد. با دستیابی به این تغییر ناپذیری، الگوریتم تعمیم و استحکام سیستم را افزایش می‌دهد.

۱-۲- چالش‌ها

از جمله چالش‌های موجود در مساله تشخیص فعالیت می‌توان به تنوع در دیدگاه، انسداد[13]و تنوع کلاس‌ها و تمایزات ظریف آن‌ها اشاره کرد. که در این بخش هر کدام را به مختصر توضیح می‌دهیم.

فعالیت‌های انسان را می‌توان از دیدگاه‌های مختلف مشاهده کرد و افراد ممکن است یک فعالیت را با تغییر در وضعیت بدن انجام دهند. تغییرات دیدگاه باعث ایجاد ابهام در بازنمایی اسکلتی می‌شود و ایجاد ویژگی‌های منسجم و متمایز برای تشخیص عمل را چالش برانگیز می‌کند. علاوه بر این، دیدگاه‌های مختلف ممکن است منجر به تغییر در مقیاس‌ها و روابط فضایی

[11] Graph Convolutional Network

[12] Variation in viewpoint

[13] Occlusion

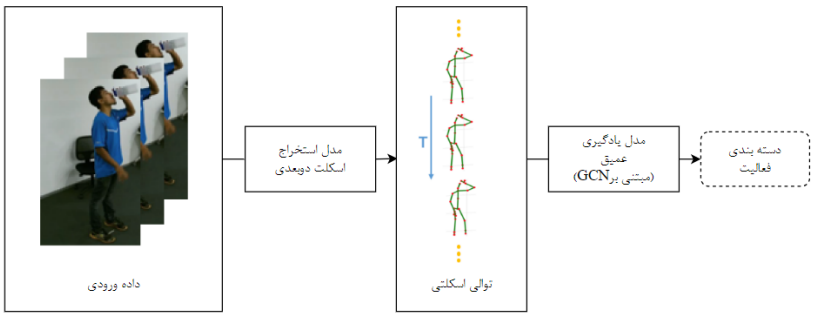
شود که فرآیند تشخیص فعالیت را پیچیده‌تر می‌کند. اکثر روش‌های تشخیص فعالیت فرض می‌کنند که فعالیت از یک دیدگاه ثابت انجام می‌شود. با این حال، در دنیای واقعی، مکان و وضعیت فرد به طور قابل توجهی بر اساس دیدگاهی که عمل از آن گرفته شده، متفاوت است.

انسداد زمانی اتفاق می‌افتد که اشیاء یا قسمت‌های بدن مفاصل خاصی را مسدود کرده و منجر به از دست رفتن یا نادرست بودن داده‌های اسکلتی می‌شود. عملی که باید تشخیص داده شود می‌بایست به وضوح در دنباله‌های ویدیویی قابل مشاهده باشد. این امر در موارد واقعی صادق نیست، به خصوص در یک ویدیوی نظارتی معمولی. انسداد می‌تواند توسط خود شخص یا هر شی دیگری در این زمینه ارائه شود. این می‌تواند اعضایی از بدن که در حال انجام یک فعالیت هستند را نامرئی کند و مشکل بزرگی برای مدل‌های تشخیص فعالیت می‌باشد[۵].

فعالیت‌های متعلق به کلاس‌های مختلف ممکن است حالت‌ها یا الگوهای حرکتی مشابهی از خود نشان دهند و این تمایز تشخیص بین آن‌ها را دشوار می‌سازد. از سوی دیگر، هدف الگوریتم تشخیص فعالیت، متمایز کردن تغییرات ظریف در یک کلاس است. ثبت این تفاوت‌های ظریف در بازنمایی اسکلتی نیازمند استخراج ویژگی‌های و الگوریتم های یادگیری قوی است.

۱-۳- اهداف پژوهش

این تحقیق با هدف بررسی و مقایسه‌ی روش‌های تشخیص فعالیت مبتنی بر اسکلت با استفاده از شبکه‌های کانولوشنی گرافی صورت گرفته است. همچنین، با توجه به اهمیت مرحله‌ی استخراج اسکلت که پیش‌تر ذکر شد، بررسی برخی الگوریتم‌های مطرح این حوزه نیز از اهداف این تحقیق است. به طور کلی در انجام این پروژه ترکیب دو مدل بینایی ماشین یکی برای تخمین حالت دو بعدی بدن انسان و بدست آوردن توالی اسکلتی و دیگری برای تشخیص فعالیت مبتنی بر اسکلت با استفاده از خروجی مدل اول است.



شکل ۱-۱ شمای کلی سیستم ترکیب الگوریتم‌های تخمین حالت بدن و تشخیص فعالیت مبتنی بر اسکلت[۴]

در راستای تحقق این هدف، با استفاده از دو الگوریتم استخراج اسکلت، از داده‌های ویدیویی، اسکلت را استخراج کرده و با به کارگیری الگوریتم بهتر، دو روش مبتنی بر شبکه‌ی کانولوشنی گرافی را راه‌اندازی کرده، آموزش داده و از لحاظ دقت، تعداد پارامترهای قابل یادگیری و زمان یادگیری الگوریتم‌ها را با هم مقایسه خواهیم کرد.

هدف ثانویه این پژوهش بررسی دقت و کارایی سیستم‌های تشخیص فعالیت مبتنی بر اسکلت در سناریوهای دنیای واقعی است. محیط‌های دنیای واقعی اغلب شرایط چالش‌برانگیزی مانند انسداد و تغییرات در حالت انسانی را ایجاد می‌کنند که روش‌های مبتنی بر یادگیری عمیق برای مدیریت موثر آن‌ها تلاش می‌کنند. برای تحقق هدف بررسی سناریوهای دنیای واقعی، مجموعه دادگان Kinetics400 انتخاب شده است که برای هر کلاس از فعالیت‌ها دارای تعداد زیادی داده‌ی ویدیوای از زوایای مختلف، در شرایط و حالات نوری گوناگون است [۶]. همچنین لازم به ذکر است که این ویدیوها توسط افراد معمولی در شرایط زندگی عادی گرفته شده اند و اکثرا شامل نویز و تکان خوردن دوربین هستند. در فصل‌های آینده به معرفی دقیق‌تر مجموعه دادگان می‌پردازیم.

از دیگر اهداف این تحقیق، پیاده‌سازی رابط کاربری مناسب به صورت نرم‌افزار تحت وب است. این رابط کاربری به صورتی خواهد بود که کاربر یک ویدیو مورد نظر را بارگذاری می‌کند. سپس اسکلت‌های خروجی بر ویدیوی ورودی برای وی نمایش داده می‌شود. و پس از آن خروجی و تشخیص نهایی دو مدل تشخیص فعالیت، نمایش داده می‌شود.

۱-۴- ساختار گزارش

در ادامه‌ی این گزارش، به بررسی رویکردها و الگوریتم‌های موجود برای تخمین حالت بدن می‌پردازیم و دو روشی که در پیاده سازی پروژه از آن‌ها استفاده کردیم را با جزییات بیشتری مورد بررسی قرار می‌دهیم. در فصل سوم، نحوه‌ی عملکرد شبکه‌های کانولوشنی گرافی را توضیح می‌دهیم و جزئیات معماری شبکه‌های کانولوشنی گرافی که در این پروژه پیاده‌سازی شده‌اند را بررسی می‌کنیم. در فصل چهارم، به معرفی مجموعه دادگان، جزئیات پیاده‌سازی، مراحل پیش‌پردازش، آموزش و ارزیابی شبکه‌های کانولوشنی گرافی با پارامترهای مختلف می‌پردازیم و این مدل‌ها را در شرایط مختلف با یکدیگر مقایسه و ارزیابی می‌کنیم. همچنین، رابط کاربری تحت وب توسعه داده شده و نحوه استفاده را از آن نیز توضیح می‌دهیم.

فصل دوم

بررسی روش‌های تخمین حالت بدن

با توجه به آنکه این پروژه مرتبط به تشخیص حرکت است، نیاز است که تجزیه و تحلیلی بر دنباله‌ای از تصاویر انجام شود تا بتوان نحوه تغییر نقاط کلیدی[1] بدن در طول الگوی حرکت را استخراج نمود. به منظور انجام این تجزیه و تحلیل در مساله‌ی تشخیص فعالیت انسان می‌توان از الگوریتم‌های تخمین حالت بدن استفاده کرد. تخمین حالت بدن یکی از زیرمجموعه‌های بینایی کامپیوتر است که ژست یک شخص یا شی را در یک تصویر یا ویدئو استنباط می‌کند. از الگوریتم‌های تخمین حالت بدن برای شناسایی و ردیابی حرکت یک فرد یا یک شی در زمان واقعی استفاده می‌شود که در صنایع بسیار مفید است. در عصر رو به رشد فناوری‌های پیشرفته، حالت بدن می‌تواند به ابزاری موثر در بیومکانیک ورزشی، انیمیشن، بازی، روباتیک، توانبخشی پزشکی و نظارت تبدیل شود.

در این فصل، ابتدا به معرفی مفاهیم اولیه الگوریتم‌های تخمین حالت مانند وظایف و نحوه دسته بندی این الگوریتم‌ها، نحوه مدل‌سازی حالت بدن می‌پردازیم. سپس، روش‌های تخمین حالت دو بعدی را مختصراً بررسی می‌کنیم. در نهایت، به معرفی و توضیح دقیق دو الگوریتم تخمین حالت Lightweight OpenPose [۷] و MediaPipe Pose [۸] که در انجام این پروژه استفاده شده اند، می‌پردازیم.

۲-۱- وظیفه الگوریتم‌های تخمین حالت

اساساً تخمین حالت بر اساس اعضای بدن فرد و موقعیت مفاصل در یک تصویر یا ویدیو، ژست‌های مختلف را پیش‌بینی می‌کند. به عنوان مثال، به کمک این الگوریتم‌ها می‌توان به طور خودکار مفاصل، بازوها و وضعیت ستون فقرات را در حین انجام یک فعالت تشخیص داد. این کار معمولاً با شناسایی، مکان‌یابی و ردیابی تعدادی از نقاط کلیدی روی یک شی یا شخص مشخص انجام می‌شود. برای اشیا، این نقاط می‌تواند گوشه‌ها یا سایر ویژگی‌های مهم باشد و برای انسان، این نقاط کلیدی مفاصل اصلی مانند آرنج یا زانو و به طور کلی اسکلت بدن را نشان می‌دهند. هدف مدل‌های یادگیری ماشین در حل این مساله ردیابی این نقاط کلیدی در تصاویر و ویدیوها است. ورودی یک مدل تخمین حالت بدن معمولاً یک تصویر پردازش شده و خروجی آن اطلاعاتی در مورد نقاط کلیدی است. محل قرارگیری نقاط کلیدی شناسایی شده توسط یک شناسه[2] عضو، به همراه یک امتیاز اطمینان که در بازه‌ی ۰ تا ۱ است، علامت‌گذاری می‌شوند. وظیفه‌ی امتیاز اطمینان[3]، نشان دادن احتمال وجود یک نقطه کلیدی در آن موقعیت خاص است.

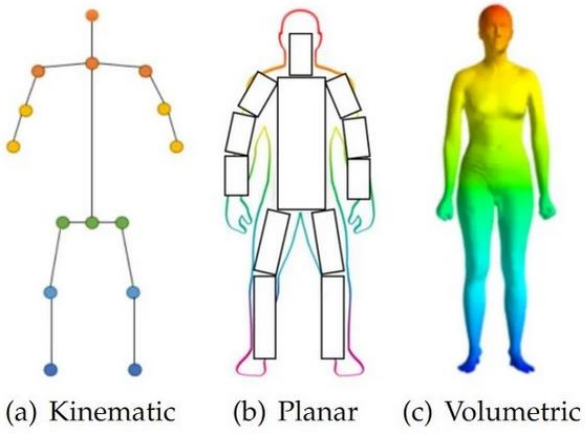
[1] Keypoints

[2] ID

[3] Confidence score

۲-۲- مدل‌سازی حالت بدن انسان

مدل‌سازی بدن انسان جنبه مهمی از الگوریتم‌های تخمین حالت بدن انسان برای نمایش نقاط کلیدی و ویژگی‌های استخراج شده از داده‌های ورودی است. برای مثال، اکثر روش‌های تخمین حالت بدن انسان از یک مدل سینماتیکی[۴] صلب N-مفصلی استفاده می‌کنند. بدن انسان موجودیتی پیچیده با مفاصل و اندام است و شامل ساختار سینماتیک بدن و اطلاعات شکل بدن است. در بسیاری از روش‌ها، یک رویکرد مبتنی بر مدل برای استنباط و ارائه حالت‌های دو بعدی یا سه بعدی بدن انسان استفاده می‌شود. معمولاً سه نوع مدل برای مدل‌سازی بدن انسان وجود دارد، یعنی مدل سینماتیک (قابل استفاده برای تخمین حالت بدن انسان دوبعدی یا سه‌بعدی)، مدل مسطح[۵] (برای تخمین حالت بدن انسان دوبعدی) و مدل حجمی[۶] (برای تخمین حالت بدن انسان سه‌بعدی). مدل سینماتیک شامل مجموعه‌ای از نقاط کلیدی (مفاصل) مانند مچ پا، زانو، شانه ها، آرنج، مچ دست و جهت گیری اندام است که برای ثبت روابط بین اعضای مختلف بدن استفاده می‌شود. این مدل بدن انسان منعطف و شهودی است و با موفقیت در تخمین حالت دو بعدی و سه بعدی بدن انسان استفاده می‌شود [۹، ۱۰]. اگرچه مدل سینماتیکی مزیت نمایش نمودار انعطاف پذیر را دارد، اما در نمایش اطلاعات بافت و شکل محدود است. مدل مسطح، شامل کانتور و عرض ناهموار بدن، تنه و اندام است. اساساً ظاهر و شکل بدن انسان را نشان می‌دهد، در این حالت اعضای بدن با مرزها و مستطیل‌های کانتور یک فرد نمایش داده می‌شوند. مدل حجمی، شامل چندین مدل و ژست‌های محبوب بدن انسان به صورت سه بعدی است که با اشکال هندسی انسان نشان داده شده است، که عموماً برای تخمین ژست سه بعدی انسان مبتنی بر یادگیری عمیق گرفته می‌شود.



شکل ۲-۱ انواع مختلف مدل برای مدلسازی تخمین حالت بدن انسان[۱۱]

[4] Kinematic

[5] Planar

[6] Volumetric

۲-۳- انواع تخمین حالت بدن

تخمین حالت بدن را می توان به دو صورت دو بعدی و سه بعدی انجام داد. الگوریتم‌های تخمین حالت بدن را می‌توان به دو دسته تک نفره و چند نفره نیز تقیسم کرد. در این بخش به معرفی این رویکردها می‌پردازیم.

۲-۳-۱- تخمین حالت بدن دو بعدی

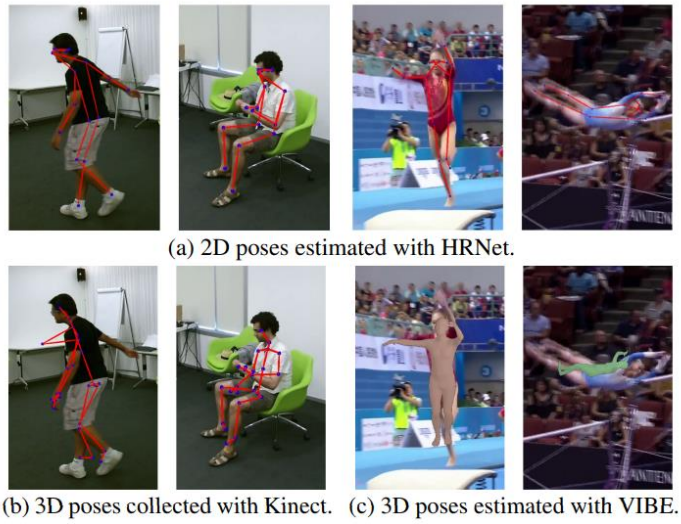
تخمین حالت دوبعدی به استنباط مکان‌های فضایی مفاصل بدن از داده‌های تصویری یا ویدئویی دوبعدی اشاره دارد. این شامل تشخیص و محلی‌سازی نقاط کلیدی مانند سر، شانه‌ها، آرنج، مچ دست، لگن، زانو و مچ پا در فضای دو بعدی است. در این نوع تخمین حالت بدن، مکان‌های مفاصل بدن در فضای دوبعدی نسبت به داده‌های ورودی (یعنی تصویر یا قاب ویدیو) تخمین زده می‌شوند و مکان هر مفصل با مختصات X و Y برای هر نقطه کلیدی نشان داده می‌شود. خروجی تخمین حالت دوبعدی اطلاعاتی در مورد پیکربندی بدن و روابط فضایی بین مفاصل در صفحه تصویر ارائه می‌کند.
روش‌های بینایی کامپیوتری مختلف، مانند روش‌های مبتنی بر یادگیری عمیق یا مدل‌های گرافیکی، می‌توانند برای تخمین ژست دوبعدی استفاده شوند. در روش‌های مبتنی بر یادگیری عمیق، از معماری شبکه‌هایی مانند شبکه‌های عصبی کانولوشنی یا شبکه‌های عصبی تکرارشونده است. این مدل‌ها بر روی مجموعه داده‌های بزرگ برچسب گذاری شده[۷] آموزش داده می‌شوند و می‌توانند الگوها و روابط پیچیده بین مفاصل بدن را بیاموزند که منجر به تخمین حالت بدن دو بعدی دقیق می‌شود. رویکرد دیگر شامل استفاده از مدل‌های گرافیکی، مانند ساختارهای تصویری یا شبکه‌های کانولوشن گرافیکی است که وابستگی‌های بین مفاصل بدن را نشان می‌دهد و محدودیت‌های فضایی حالت‌های انسانی را مدل‌سازی می‌کند.

۲-۳-۲- تخمین حالت بدن سه بعدی

در تخمین حالت بدن سه بعدی، با افزودن تخمین یک بعد Z به دو بعدی حالت قبل، می‌توان یک تصویر دو بعدی را به یک شی سه بعدی تبدیل کرد. تخمین حالت بدن سه بعدی این امکان را به ما می‌دهد تا موقعیت مکانی دقیق یک شخص یا شی نشان داده شده در تصویر یا ویدیو را به صورت سه بعدی پیش‌بینی کنیم. در حالی که مجموعه‌داده‌های انسانی دوبعدی را می‌توان به راحتی به دست آورد، جمع‌آوری مختصات برای تصویر حالت بدن سه بعدی دقیق زمان‌بر است و برچسب زدن دستی عملی نیست. بنابراین، اگرچه ردیابی حالت بدن سه بعدی در سال‌های اخیر پیشرفت‌های چشمگیری داشته است، به‌ویژه به دلیل پیشرفت‌هایی که در تخمین حالت بدن دوبعدی انسان حاصل شده است، هنوز چندین چالش اعم از تعمیم مدل، استحکام در انسداد و کارایی محاسبات وجود دارد که باید بر آن‌ها غلبه کرد.

[7] Annotated

به طور کلی، در روش‌های تخمین حالت بدن انسان پیشنهاد شده تا به امروز، روش‌های دو بعدی کیفیت بهتری نسبت به روش‌های سه بعدی دارند. با توجه به آزمایش‌های انجام شده در مقاله‌ی دوان و همکاران[۱۲]، حالت‌های بدن دوبعدی برآورد شده با HRNet [۱۳] برای ویدیوهای موجود بر دو مجموعه‌داده بررسی شده‌اند و استنباط شده است که ظاهراً کیفیت آن‌ها بسیار بهتر از حالت‌های بدن سه بعدی جمع آوری شده توسط حسگرها یا تخمین زده شده با برآوردگرهای پیشرفته است. به همین دلیل احتمالا استفاده از اسکلت‌های بدست آمده از مدل‌های تخمین حالت بدن دو بعدی برای مساله تشخیص فعالیت نتایج بهتری حاصل می‌شود.


(a) 2D poses estimated with HRNet.

(b) 3D poses collected with Kinect. (c) 3D poses estimated with VIBE.

شکل ۲-۲ مقایسه‌ای از خروجی‌های الگوریتم‌های تخمین حالت بدن دو و سه بعدی[۱۲]

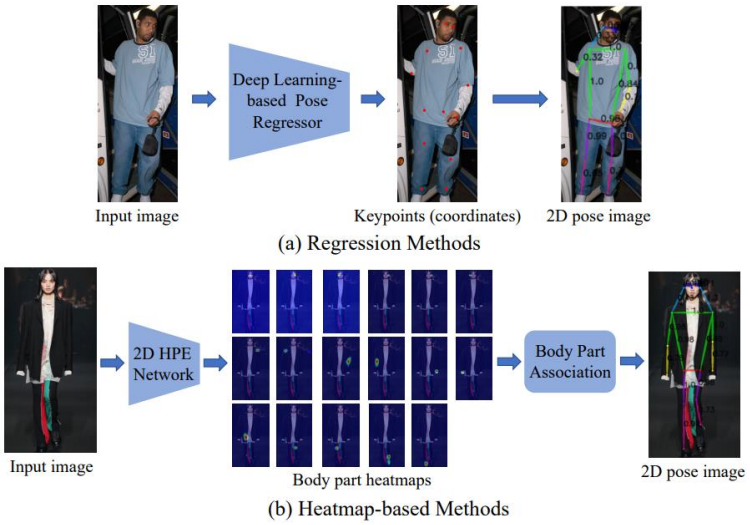
۲-۳-۳- تخمین حالت بدن تک نفره و چند نفره

با توجه به کاربرد، می‌توان حالت‌های بدن را برای یک فرد یا چند نفر تخمین بزنیم. در تخمین حالت بدن تک نفره، مدل حالت بدن را برای یک فرد در یک صحنه معین تخمین می‌زند (در صورتی که چند فرد در صحنه وجود داشته باشند، مدل حالت بدن را برای فردی که با اطمینان بالاتری تشخیص داده شده است، تخمین می‌زند). در مقابل، در مورد تخمین حالت بدن چند نفره، مدل حالت بدن را برای افراد متعدد در توالی ورودی داده شده تخمین می‌زند.

۲-۴- روش‌های تخمین حالت بدن دو بعدی

در تخمین حالت بدن تک نفره دوبعدی، اگر بیش از یک نفر وجود داشته باشد، ابتدا تصویر ورودی برش داده می‌شود تا در هر تصویر فرعی برش داده شده تنها یک نفر وجود داشته باشد. این فرآیند می‌تواند به طور خودکار توسط یک آشکارساز

بالاتنه[8] یا یک آشکارساز تمام بدن[9] به دست آید. به طور کلی، دو روش برای تخمین حالت بدن تک نفره دوبعدی وجود دارد که از تکنیک‌های یادگیری عمیق استفاده می‌کنند: روش‌های رگرسیون[10] و روش‌های مبتنی بر نقشه حرارتی[11]. روش‌های رگرسیون از یک چهارچوب[12] سرتاسری[13] برای یادگیری یک نگاشت از تصویر ورودی به مفاصل بدن انسان استفاده می‌کنند [۱۴]. هدف روش‌های مبتنی بر نقشه حرارتی پیش‌بینی مکان‌های تقریبی اعضای بدن و مفاصل است که توسط نمایش نقشه‌های حرارتی نظارت می‌شود [۱۵]. چهارچوب‌های کلی روش‌های تخمین حالت بدن تک نفره دوبعدی در شکل ۲-۳ نشان داده شده است.



شکل ۲-۳ چهارچوب‌های کلی روش‌های تخمین حالت بدن تک نفره دو بعدی[۱۶]

در مقایسه با تخمین حالت بدن تک نفره، تخمین حالت بدن چند نفره دشوارتر و چالش برانگیزتر است زیرا نیاز به تعیین تعداد افراد، موقعیت آن‌ها و نحوه گروه‌بندی نقاط کلیدی برای افراد مختلف دارد. برای حل این مشکلات، روش‌های تخمین حالت بدن چند نفره را می‌توان به روش‌های بالا به پایین[14] و پایین به بالا[15] طبقه‌بندی کرد. در رویکرد پایین به بالا، مدل هر نمونه از یک نقطه کلیدی خاص (مثلاً همه دست‌های چپ) را در یک تصویر معین تشخیص می‌دهد و سپس تلاش

[8] Upper-body detector

[9] Full-body detector

[10] Regression

[11] Heatmap-based

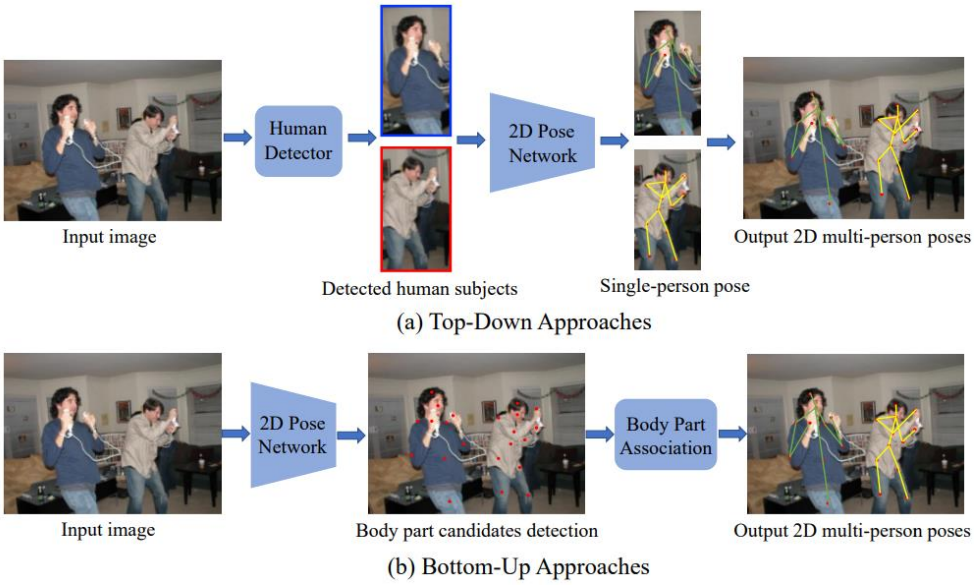
[12] Framework

[13] End-to-end

[14] Top-down

[15] Bottom-up

می‌کند تا گروه‌هایی از نقاط کلیدی را در اسکلت‌هایی برای اشیاء متمایز جمع کند. رویکرد بالا به پایین معکوس این روش است. در حالت بالا به پایین شبکه ابتدا از یک آشکارساز[16] شی برای ترسیم کادری در اطراف هر فرد استفاده می‌کند و سپس نقاط کلیدی را در هر منطقه مشخص شده، تخمین می‌زند. در روش‌های بالا به پایین، تعداد افراد در تصویر ورودی مستقیماً بر زمان محاسبه تأثیر می‌گذارد. سرعت محاسبات برای روش‌های پایین به بالا معمولاً سریع‌تر از روش‌های بالا به پایین است زیرا نیازی به تشخیص حالت بدن برای هر فرد به صورت جداگانه نیست. شکل ۲-۴ چهارچوب‌های کلی برای روش‌های تخمین حالت بدن چند نفره دو بعدی را نشان می‌دهد.



شکل ۲-۴ چهارچوب کلی روش‌های تخمین حالت بدن چند نفره دو بعدی [۱۶]

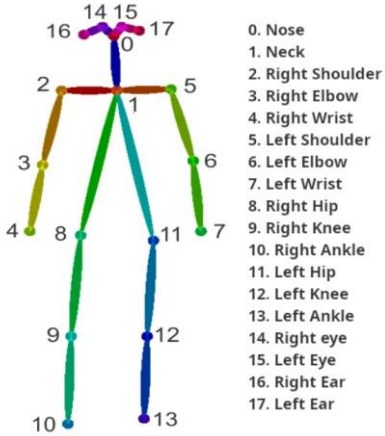
به طور کلی، عملکرد روش‌های تخمین حالت بدن دو بعدی با شکوفایی تکنیک‌های یادگیری عمیق (به خصوص شبکه‌های عصبی کانولوشنی) به طور قابل توجهی بهبود یافته است. در سال‌های اخیر، شبکه‌های عمیق‌تر و قوی‌تر، باعث بهبود عملکرد روش‌های تخمین حالت بدن تک‌نفره دو بعدی مانند DeepPose [۱۱] و Stacked Hourglass Network [۱۲] و همچنین روش‌های تخمین حالت بدن چند نفره دو بعدی مانند AlphaPose [۱۷] و OpenPose [۱۰] شده‌اند.

مدل‌های یادشده و بسیاری از روش‌های دیگری که در سال‌های اخیر معرفی شده‌اند دقت بالا و مطلوبی دارند اما از نظر محاسباتی به اصطلاح مدل‌های سنگینی هستند و منابع سخت‌افزاری و زمان زیادی را برای رسیدن به خروجی مطلوب صرف می‌کنند. به همین سبب، در این پروژه از مدل‌های پیشنهاد شده‌ی سبک‌تر که در طراحی آن‌ها سعی شده هزینه محاسباتی پایین و در عین حال کارآمد باشند، استفاده می‌کنیم.

[16] Detector

Lightweight OpenPose -۵-۲

این روش سعی کرده است تا روش محبوب پایین به بالای OpenPose را بهینه کند و نشان دهد که چگونه می‌توان از تکنیک‌های طراحی مدرن شبکه‌های عصبی کانولوشنی برای مساله‌ی تخمین حالت بدن استفاده کرد. از مزیت‌های این روش این است که برای تعداد افراد متفاوت در داده ورودی، زمان استنتاج تقریباً ثابتی دارد. دقت این نسخه بهینه شده تقریباً با نسخه دو مرحله‌ای OpenPose مطابقت دارد و افت میانگین دقت[۱۷] آن کمتر از ۱ درصد است. خروجی این یک مدلسازی سینماتیک از حالت بدن شامل ۱۸ نقطه کلیدی است که نحوه شماره‌گذاری آن‌ها در شکل ۲-۵ نمایش داده شده است.



شکل ۲-۵ نحوه شماره گذاری نقاط کلیدی در مدل Leightweight OpenPose [۱۰]

۲-۵-۱- معماری مدل OpenPose

مشابه همه روش‌های پایین به بالا، مدل OpenPose از دو بخش تشکیل شده است. بخش اول شامل شبکه عصبی برای ارائه دو تانسور[۱۸] نقشه‌های حرارتی[۱۹] نقاط کلیدی و روابط زوجی آن‌ها (میدان‌های وابستگی بخشی[۲۰]) است. این خروجی ۸ بار نمونه‌برداری کاهشی[۲۱] شده است. بخش دوم شامل گروه‌بندی نقاط کلیدی بر اساس نمونه‌های شخص است که شامل

[17] Average Percision (AP)

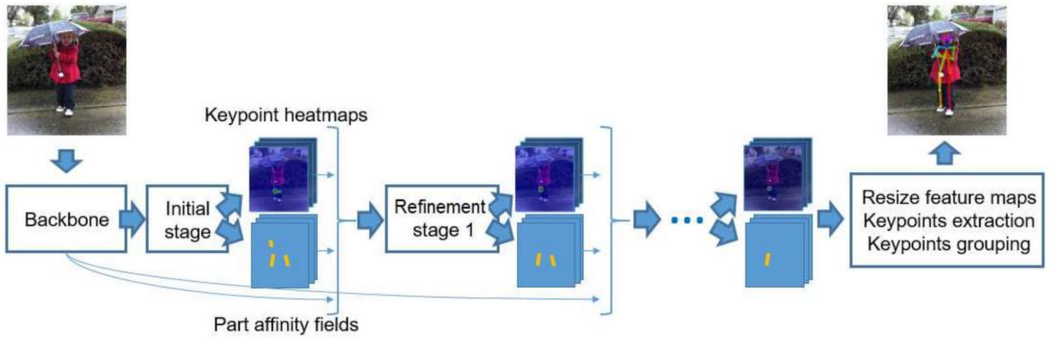
[18] Tensor

[19] Heatmap

[20] Part Affinity Fields

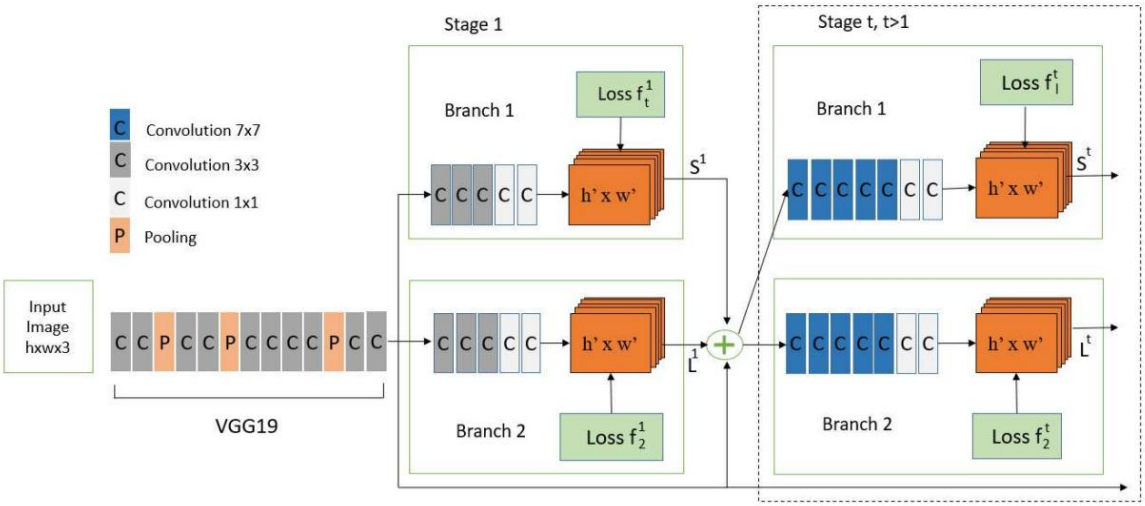
[21] Downsample

نمونه‌برداری افزایشی[۲۲] تانسورها به اندازه تصویر اصلی، استخراج نقاط کلیدی در پیک‌های نقشه حرارتی و گروه‌بندی آن‌ها بر اساس نمونه است.



شکل ۲-۶ خط لوله[۲۳] مدل OpenPose [۷]

شبکه ابتدا ویژگی‌ها را استخراج می‌کند، سپس تخمین اولیه نقشه‌های حرارتی و میدان‌های وابستگی بخشی را انجام می‌دهد و پس از آن ۵ مرحله اصلاح انجام می‌شود که در هر مرحله قادر به یافتن ۱۸ نوع نقطه کلیدی است. سپس روش گروه‌بندی بهترین جفت (بر اساس وابستگی) را برای هر نقطه کلیدی، از لیست از پیش تعریف‌شده جفت‌های کلیدی، جستجو می‌کند. در طول استنتاج، اندازه تصویر ورودی برای مطابقت با اندازه ورودی شبکه تغییر می‌کند.



شکل ۲-۷ معماری شبکه‌های استفاده شده در مدل OpenPose [۱۸]

[22] Upsample

[23] Pipeline

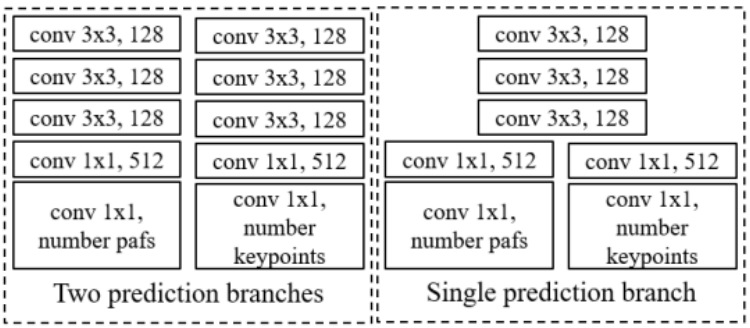
در مدل OpenPose برای استخراج ویژگی‌ها از یک شبکه با معماری VGG19 استفاده شده است [۱۹]. سپس در مرحله اول، تخمین اولیه نقشه‌های حرارتی و میدان‌های وابستگی بخشی در دو شاخه (دارای ساختار یکسان) جدا متشکل از لایه‌های کانولوشنی صورت گرفته است. و در چند مرحله پشت سر هم اصلاح انجام می‌گیرد که در این قسمت نیز از دو شاخه شبکه‌های عصبی کانولوشنی با ساختار یکسان استفاده شده است.

۲-۵-۲- بهینه‌سازی‌های انجام شده در مدل lightweight

طبق بررسی‌های انجام شده، پس از اعمال یک مرحله اصلاح، تکرار مراحل اصلاح به ازای هزینه محاسباتی ثابت بهبود کمتری در دقت و عملکرد الگوریتم دارد [۷]. بنابراین، برای نسخه بهینه‌سازی شده فقط دو مرحله اول را حفظ شده است: مرحله اولیه و یک مرحله اصلاح. همچنین همه مراحل اجرای الگوریتم به جز مرحله گروه‌بندی نقاط کلیدی، از نظر محاسباتی سنگین هستند و نیاز به بهینه‌سازی دارند.

برای بهینه‌سازی شبکه‌ی VGG19، استفاده از یک توپولوژی شبکه سبک وزن با دقت طبقه‌بندی مشابه پیشنهاد شده است. شبکه‌های خانواده MobileNet برای جایگزینی استخراج‌کننده ویژگی ارزیابی شده اند و در نهایت از MobileNet v1 در مدل Lightiweight OpenPose استفاده شده است.

داده ورودی در هر مرحله‌ی تخمین و اصلاح، ترکیبی از ویژگی‌های بدست آمده از شبکه‌ی VGG19 با تخمین قبلی نقشه‌های حرارتی و میدان‌های وابستگی بخشی نقطه کلید است. بنابراین، برای بهینه‌سازی مرحله تخمین اولیه و مرحله اصلاح، بیشترین محاسبات را بین نقشه های حرارتی و میدان‌های وابستگی بخشی به اشتراک گذاشته شده و تبدیل به یک شاخه شدند. به این ترتیب که همه لایه‌ها در یک شاخه مشترک اند، به‌جز دو لایه آخر که مستقیماً نقشه‌های حرارتی و میدان‌های وابستگی بخشی را تولید می‌کنند. همچنین ساختار شبکه‌های کانولوشنی عصبی در دو مرحله تخمین و اصلاح، بهینه شده است و اندازه فیلترها و تعداد لایه‌های کانولوشنی کاهش یافته است.



شکل ۲-۸ ساختار مرحله تخمین دو شاخه و تک شاخه بهینه شده. این معماری تک شاخه برای مرحله اصلاح نیز اعمال می‌شود.

[۷]

بنابراین شبکه‌ی Lightweight OpenPose، شامل یک استخراج‌کننده ویژگی MobileNet v1 است که ویژگی‌های استخراج شده را به یک بلوک تخمین نقشه‌های حرارتی و میدان‌های وابستگی بخشی وارد می‌کند. این بلوک ساخته شده از معماری تک شاخه یک شبکه‌ی عصبی کانولوشنی بسیار سبک است که نقشه‌های حرارتی و میدان‌های وابستگی بخشی اولیه را تولید می‌کند و در نهایت این تخمین ها توسط یک بلوک با ساختار مشابه، اصلاح می‌شوند. این مدل بهینه سازی شده را می‌توان به راحتی روی یک واحد پردازنده مرکزی اجرا کرد و برخلاف مدل OpenPose برای اجرای الگوریتم نیاز به سخت افزار قوی و پردازنده گرافیکی ندارد.

MediaPipe -۶-۲

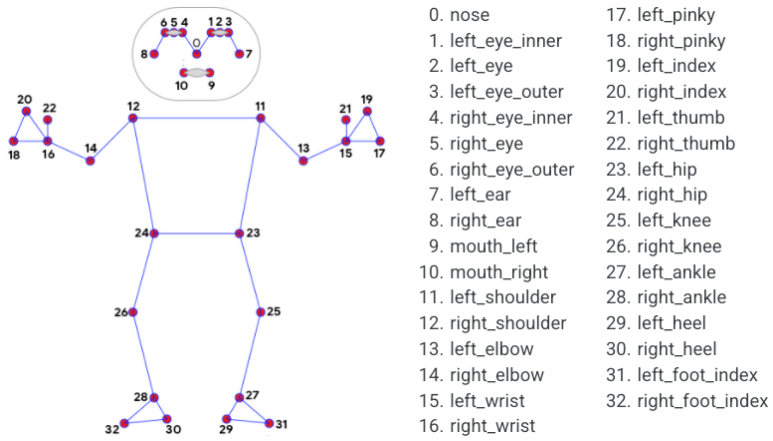
MediaPipe یک چهارچوب متن باز[۲۴] برای ساخت خطوط لوله برای انجام استنتاج بینایی کامپیوتری بر روی داده‌هایی مانند ویدئو یا عکس است. چهارچوب MediaPipe عمدتاً برای پیاده‌سازی برنامه‌های بینایی کامپیوتری در دموها و برنامه‌های کاربردی بر روی سیستم‌های سخت‌افزاری مختلف استفاده می‌شود. این چهارچوب حاوی واسط‌ها برنامه نویسی کاربردی[۲۵] از مدل‌هایی با عملکرد مطلوب برای دامنه وسیعی از مسائل بینایی کامپیوتری است. در این پروژه از قسمت Pose Landmarker این چهارچوب استفاده شده است که امکان تشخیص نقاط کلیدی بدن یک فرد در یک تصویر یا ویدیو را فراهم می‌کند. این ماژول با استفاده از مدل BlazePose [۸]، نقاط کلیدی حالت بدن در مختصات تصویر و در مختصات جهان سه بعدی را استخراج می‌کند و به عنوان یک واسط برنامه نویسی کاربردی در اختیار کاربران قرار گرفته است.

BlazePose یک معماری شبکه عصبی کانولوشنی سبک وزن برای تخمین حالت بدن انسان است که برای استنتاج بلادرنگ[۲۶] شده است. در طول استنتاج، شبکه ۳۳ نقطه کلیدی بدن را برای یک فرد تولید می‌کند و می‌تواند سرعت با بیش از ۳۰ قاب در ثانیه اجرا شود. این امر آن را به ویژه برای موارد استفاده بلادرنگ مناسب می‌سازد. برای خروجی این مدل یک مدلسازی سینماتیکی جدید با استفاده از ۳۳ نقطه روی بدن انسان ارائه شده است. برخلاف مدلسازی‌های OpenPose که دارای ۱۳۵ نقطه کلیدی است، این مدل فقط از تعداد حداقل کافی از نقاط کلیدی روی صورت، دست‌ها و پاها برای تخمین چرخش، اندازه و موقعیت ناحیه مورد نظر برای مدل بعدی استفاده می‌کند. مدل‌سازی معرفی شده در شکل ۲-۹ نشان داده شده است.

[24] Open-source

[25] Application Program Interface (API)

[26] Real-time



شکل ۲-۹ نحوه شماره گذاری نقاط کلیدی در مدل BlazePose [۲۰]

۲-۶-۱- معماری مدل BlazePose

در طول استنتاج، این مدل از یک تنظیم آشکارساز-ردیاب[۲۷] استفاده می‌کند، که عملکرد عالی در زمان واقعی را ممکن می‌سازد. خط لوله این مدل متشکل از یک آشکارساز حالت بدن سبک وزن است که به دنبال آن یک شبکه ردیاب حالت بدن قرار گرفته است. ردیاب مختصات نقطه کلیدی، حضور فرد در قاب فعلی و ناحیه مورد توجه برای قاب[۲۸] فعلی را پیش‌بینی می‌کند. هنگامی که ردیاب نشان دهد که در قاب فعلی هیچ انسانی وجود ندارد، شبکه آشکارساز روی قاب بعدی اجرا می‌شود.

۲-۶-۱-۱- ردیاب شخص در مدل BlazePose

راه‌حل‌های تشخیص اشیا برای آخرین مرحله پس از پردازش خود به الگوریتم Non-Maximum Suppression (NMS) متکی هستند. این برای اجسام صلب با درجه آزادی کم به خوبی کار می کند. با این حال، این الگوریتم برای سناریوهایی که شامل ژست‌های بسیار مفصلی مانند حالت بدن انسان‌ها می‌شود، ناتوان است. برای حل این مشکل، این الگوریتم بر روی تشخیص جعبه محدود کننده[۲۹] یک قسمت نسبتاً سفت و سخت بدن مانند صورت یا نیم تنه انسان تمرکز شده است. طبق مشاهدات در بسیاری از موارد، قوی‌ترین سیگنال به شبکه عصبی در مورد موقعیت نیم تنه، صورت شخص است (زیرا دارای ویژگی‌های کنتراست بالا و تغییرات ظاهری کمتری است). برای اینکه چنین آشکارساز فردی سریع و سبک باشد، فرض

[27] Detector-Tracker

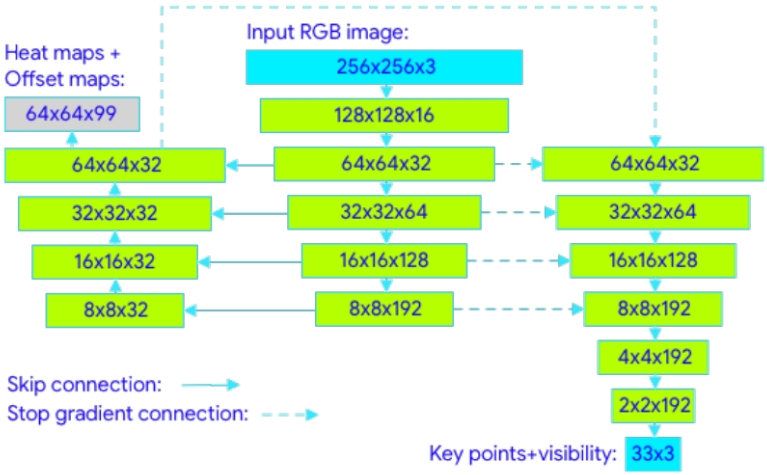
[28] Frame

[29] Bounding box

شده که فرد زمانی در قاب وجود دارد که سر او قابل مشاهده باشد. در نتیجه، از یک آشکارساز صورت سریع به عنوان یک آشکارساز شخص استفاده شده است. این آشکارساز صورت، پارامترهای مرکزی خاص فرد را پیش‌بینی می‌کند.

۲-۶-۱-۲- ردیاب حالت بدن در مدل BlazePose

ردیاب حالت بدن مکان تمام نقاط کلیدی فرد را پیش‌بینی می‌کند و از پارامترهای مرکزی خاص فرد که توسط آشکارساز ارائه شده است، استفاده می‌کند. شبکه عصبی ردیاب حالت بدن، با ترکیب رویکرد نقشه حرارتی، انحراف[۳۰] و رگرسیون ترکیبی طراحی شده است. از نقشه حرارتی و ضیان[۳۱] انحراف فقط در مرحله آموزش استفاده می‌شود و لایه‌های خروجی مربوطه پیش از اجرای استنتاج از مدل حذف می‌شوند. در معماری این مدل یک شبکه عصبی کوچک مبتنی بر نقشه حرارتی رمزگذار-رمزگشا[۳۲] به همراه یک شبکه رمزگذار رگرسیون بر روی هم قرار گرفته‌اند. در معماری این شبکه عصبی، اتصالات پرش[۳۳] بین تمام مراحل شبکه برای دستیابی به تعادل بین ویژگی‌های سطح بالا و پایین استفاده قرار داده شده اند. با این حال، گرادیان‌های رمزگذار رگرسیون به ویژگی‌های آموزش نقشه حرارتی منتشر نمی‌شوند. این رویکرد سبب بهبود پیش‌بینی‌های نقشه حرارتی و افزایش دقت رگرسیون مختصات می‌شود.



شکل ۲-۱۰ معماری شبکه عصبی ردیاب حالت بدن در مدل BlazePose [۸]

[30] Offset map

[31] Loss

[32] Encoder-Decoder

[33] Skip-connection

۲-۷- مقایسه‌ی عملکرد OpenPose و MediaPipe Pose

طبق تحقیقات انجام شده توسط چانگ و همکاران [۲۱]، هنگام مواجهه با چالش هایی مانند موقعیت نامناسب دوربین یا انسداد تصویر، کارایی روش‌های تخمین حالت بدن در تشخیص اعضای بدن کاهش می‌یابد. طبق تحقیقات این گروه، MediaPipe Pose می‌توانند به خوبی با این چالش‌ها مقابله کند، اما OpenPose ضعیف ترین عملکرد را در بین روش‌های شناخته تخمین حالت بدن در این شرایط نشان می‌دهد. همچنین در بین داده‌های ویدیویی مربوط به فعالیت‌های مختلف، درصد مفاصل تشخیص داده شده توسط MediaPipe Pose بالاتر از OpenPose است. در تشخیص حالت بدن در داده‌های ویدیویی، OpenPose کمترین استحکام را داشته است، زیرا زمانی که انسداد در قسمت‌های بدن اتفاق می‌افتد، در دقت تخمین این الگوریتم مشکل پیش می‌آید. در گزارشات ویشنو و همکاران [۲۲]، نیز بیان شده است که MediaPipe در تصاویر با شدت نور کم، در جهت گیری‌های مختلف، فواصل متفاوت از دوربین و در فیلم‌های دارای حرکت، MediaPipe تخمین حالت بدن بهتری را نسبت به OpenPose انجام می‌دهد.

۲-۸- خلاصه

در این فصل، مفاهیم اولیه‌ای را مربوط به انواع روش‌های تخمین حالت بدن معرفی کردیم. سپس به بررسی دو مدل تخمین حالت بدن استفاده شده در انجام این پروژه پرداختیم. روش اول که Lightweight OpenPose بود که حالت سبک شده‌ی مدل محبوب و شناخته شده‌ی OpenPose است. طبق توضیحات داده شده، Lightweight OpenPose یک روش تخمین حالت دو بعدی چند نفره با متد پایین به بالا است که کارایی محاسباتی مناسبی برای اجرا بدون نیاز به پردازشگر گرافیکی دارد. روش دوم استفاده از MediaPipe است که برای کار با الگوریتم‌های یادگیری ماشین واسط‌های برنامه نویسی کاربردی در اختیار کاربران قرار می‌دهد. در این پروژه از ماژول تخمین حالت بدن آن (MediaPipe Pose) استفاده شده است که با استفاده از شبکه عصبی BlazePose (با متد بالا به پایین) یک تخمین سه بعدی تک نفره از حالت بدن انسان استنتاج می‌کند. در نهایت به گزارش نتایج برخی مقایسه‌های انجام شده بین این دو روش پرداختیم و نتیجه آن شد که روش MediaPipe Pose با توجه به بار محاسباتی کمتر، عملکرد مطلوب تری نیز در تخمین حالت دو بعدی روی داده ویدیویی دارد.

فصل سوم

بررسی روش‌های تشخیص فعالیت مبتنی بر اسکلت

اسکلت و سیر مفاصل بدن انسان در برابر تغییر روشنایی و تغییرات صحنه مقاوم هستند و به دلیل سنسورهای عمق بسیار دقیق یا الگوریتم‌های تخمین حالت بدن به راحتی به دست می‌آیند. بنابراین، طیف وسیعی از رویکردهای تشخیص عمل مبتنی بر اسکلت وجود دارد. رویکردها را می‌توان به روش های مبتنی بر ویژگی‌های دست ساز[1] و روش‌های یادگیری عمیق دسته‌بندی کرد. رویکردهای نوع اول چندین ویژگی دست ساز را طراحی می‌کنند تا پویایی حرکت مفصل را به تصویر بکشند. این‌ها می‌توانند ماتریس‌های کوواریانس مسیرهای مشترک و یا موقعیت نسبی مفاصل باشند. موفقیت اخیر یادگیری عمیق منجر به افزایش روش‌های مدل سازی اسکلت مبتنی بر یادگیری عمیق شده است. محققان در ابتدا از شبکه‌های عصبی بازگشتی استفاده کردند و پس از مواجهه با محدودیت‌های این شبکه‌ها در تشخیص فعالیت مبتنی بر اسکلت، شبکه‌های کانولوشنی زمانی[2] برای یادگیری تشخیص فعالیت به روش انتها به انتها[3] معرفی شدند. این روش‌ها پیشرفت‌های امیدوارکننده‌ای را نشان داده‌اند. با این حال، بیشتر روش‌های موجود برای تجزیه و تحلیل الگوهای فضایی به قطعات یا قوانین دست‌ساز متکی هستند. در نتیجه، تعمیم[4] مدل‌های طراحی شده برای یک کاربرد خاص به دیگران دشوار است. محدودیت‌های موجود در این روش‌ها و مطلوب بودن ساختار گرافی برای استخراج ویژگی‌های مهم حرکتی منجر به ظهور مدل‌هایی با استفاده از شبکه‌های کانولوشنی گرافی شد.

به طور کلی، رویکردهای موجود مدل‌های یادگیری عمیق در حوزه تشخیص فعالیت مبتنی بر اسکلت در سه دسته‌ی اصلی شبکه‌های عصبی بازگشتی، شبکه‌های کانولوشنی و شبکه‌های کانولوشنی گرافی قرار می‌گیرند و تمرکز این پژوهش بر مدل‌های مبتنی بر شبکه‌های کانولوشنی گرافی(GCN) است. بنابراین، در این فصل به معرفی مفاهیم پایه این شبکه‌ها می‌پردازیم و به طور مختصر پیشینه‌ای از مدل‌های معرفی شده در این دسته را مطرح می‌کنیم. سپس معماری و جزییات دو مدل مبتنی بر شبکه‌های کانولوشنی گرافی که در این پروژه راه‌اندازی شده‌اند، را شرح خواهیم داد.

۳-۱- مفاهیم پایه GCN

در سالهای قبل، انواع مختلفی از شبکه‌های عصبی گرافی[5] توسعه داده شده‌اند که شبکه‌های کانولوشنی گرافی یکی از آن‌هاست. به طور کلی، مدل‌های شبکه کانولوشنی گرافی می‌توانند از ساختار گراف استفاده کنند و به صورت کانولوشنی به جمع‌آوری اطلاعات گره‌ها از همسایگی‌ها بپردازند. شبکه‌های کانولوشنی گرافی قدرت زیادی برای یادگیری نمایش گراف دارند و در طیف وسیعی از مسائل و کاربردها به عملکرد برتر دست یافته اند. در این بخش مفاهیم، نحوه عملکرد و فرمول سازی یک مساله به صورتی که قابل حل با استفاده از GCNها باشد را معرفی می‌کنیم.

---

[1] Handcrafted features

[2] Temporal Convolutional Neural Network

[3] End-to-end

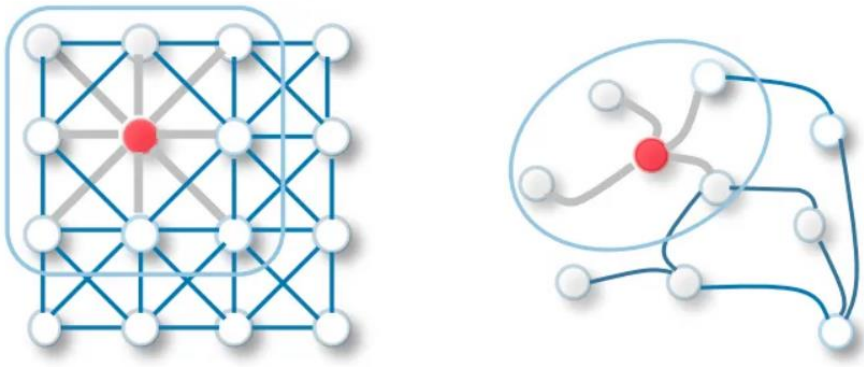
[4] generalization

[5] Graph Neural Network

۳-۱-۱- عملیات کانولوشن گرافی

معنای کلمه‌ی کانولوشن یا پیچیدگی در شبکه‌های کانولوشنی گرافی اساساً همان عملیات کانولوشن یا پیچش در شبکه‌های عصبی کانولوشنی است. این به انجام عملیات ضرب مقدار ورودی در مجموعه‌ای از وزن‌ها اشاره دارد که این وزن‌ها معمولاً به عنوان فیلتر شناخته می‌شوند. فیلترها به عنوان یک پنجره کشویی در کل تصویر عمل می‌کنند و CNNها را قادر می‌سازند تا ویژگی‌های سلول‌های همسایه را بیاموزند. در یک لایه، از فیلتر یکسانی در سراسر تصویر استفاده می‌شود که به آن اشتراک وزن[۶] گفته می‌شود. تفاوت عمده بین CNN و GNN در این است که CNNها به طور ویژه برای کار بر روی داده های ساختاری منظم (اقلیدسی) ساخته شده‌اند، در حالی که GNNها نسخه تعمیم یافته CNN هستند که در آن تعداد اتصالات گره‌ها متفاوت است و گره‌ها نامرتب هستند (داده ساختار یافته غیراقلیدسی).

همانطور که در شکل ۳-۱ مشاهده می‌شود، در CNN هر پیکسل در یک تصویر به عنوان یک گره در نظر گرفته می‌شود که در آن همسایگان با اندازه فیلتر تعیین می‌شوند. پیچیدگی دوبعدی میانگین وزنی مقادیر پیکسل گره قرمز و همسایگان آن را محاسبه می‌کند. همچنین همسایگان یک گره مرتب شده‌اند و اندازه ثابتی دارند. در GCN نیز برای به دست آوردن یک نمایش پنهان از گره قرمز، یک راه حل ساده برای عملیات کانولوشن گراف این است که مقدار متوسط ویژگی‌های گره قرمز را به همراه همسایگان آن در نظر گرفته شود. در این حالت متفاوت از داده‌های تصویر، همسایگان یک گره نامرتب و متغیر هستند.



شکل ۳-۱ کانولوشن در شبکه عصبی کانولوشنی دو بعدی (سمت چپ) و شبکه کانولوشنی گرافی (راست) [۲۳]

به طور کلی می‌توان گفت، انجام عملیات پیچیدگی روی گراف‌ها بسیار چالش برانگیزتر است. گراف‌ها به دلیل ساختار نامنظم شان به سادگی قابل استفاده نیستند. تصاویر در یک شبکه اقلیدسی دو بعدی نشان داده می‌شوند، جایی که یک فیلتر می‌تواند به چپ، راست و غیره حرکت کند. گراف‌ها غیر اقلیدسی هستند و مفهوم جهت‌هایی مانند بالا، پایین و غیره

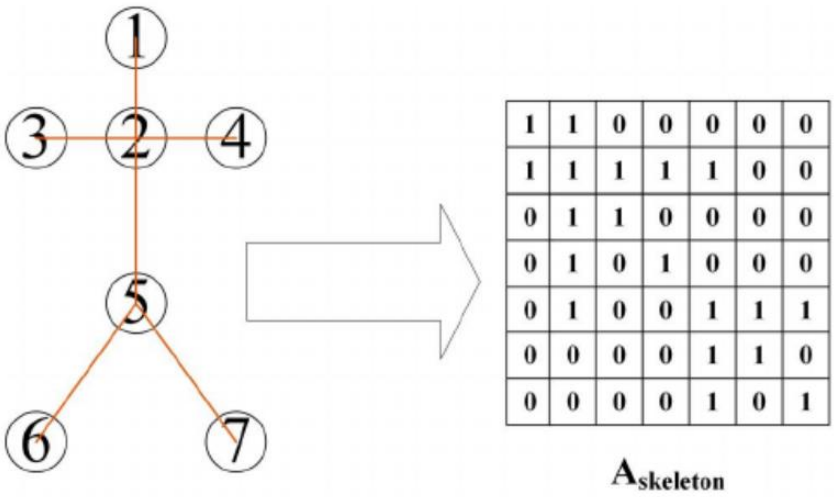
[6] Weight sharing

هیچ معنایی ندارد. گراف‌ها انتزاعی‌تر از تصاویر هستند و متغیرهایی مانند درجه گره، مجاورت و ساختار همسایگی اطلاعات بسیار بیشتری در مورد داده‌ها ارائه می‌دهند.

۳-۱-۲- تبدیل گراف به ورودی مناسب شبکه‌های عصبی

باید ساختار یک گراف را به حالت یک ماتریس مجاورت تبدیل کرد تا بتوان آن را به عنوان داده‌ی ورودی به یک شبکه عصبی داد و به این منظور از سه ساختمان داده‌ی ماتریس مجاورت، ماتریس ویژگی‌های گره و ماتریس ویژگی‌های یال استفاده می‌کنیم که در ادامه به توضیح آن‌ها می‌پردازیم.

روابط و یال‌های بین گره‌ها، به کمک یک ماتریس مجاورت نمایش داده می‌شوند. ماتریس مجاورت یک ماتریس N x N است و با ۰ یا ۱ پر شده است، که در آن N تعداد کل گره‌ها است. ماتریس‌های مجاورت وجود یال‌هایی را نشان می‌دهند که جفت‌های گره را از طریق مقدار موجود در ماتریس به هم متصل می‌کنند. شکل ۳-۲ نمونه‌ی ابتدایی از یک ماتریس مجاورت مربوط به مساله‌ی تشخیص فعالیت مبتنی بر اسکلت را نشان می‌دهد. ماتریس ویژگی‌های گره برخلاف ماتریس مجاورت که رابطه بین گره‌ها را بیان می‌کند، ویژگی[۷]ها یا خاصیت[۸]های هر گره را نشان می‌دهد. اگر N گره وجود داشته باشد و اندازه ویژگی‌های گره F باشد، شکل این ماتریس N x F خواهد بود. گاهی اوقات، یال‌ها نیز می‌توانند مانند گره‌ها ویژگی‌های خاص خود را داشته باشند، که در این صورت این ویژگی‌ها را توسط ماتریس ویژگی‌های یال بیان می‌کنند. اگر اندازه ویژگی‌های یال S و تعداد یال‌های موجود n_edges باشد، شکل این ماتریس n_edges x S است.



شکل ۳-۲ نمایش یک داده‌ی اسکلت گرافی به صورت ماتریس مجاورت [۲۴]

[7] Feature

[8] Attribute

۳-۲- پیشینه GCNها در تشخیص فعالیت مبتنی بر اسکلت

در میان رویکردهای یادگیری عمیق در حوزه‌ی تشخیص فعالیت مبتنی بر اسکلت، بسیاری بر اهمیت مدل‌سازی مفاصل در بین بخش‌هایی از بدن انسان تأکید کرده‌اند. اما پیش از استفاده از شبکه‌های کانولوشنی گرافی، در رویکردهای پیشین(شبکه‌های عصبی بازگشتی و شبکه‌های عصبی کانولوشنی) این بخش‌ها معمولاً با استفاده از دانش تخصصی مربوط به اسکلت به صورت صریح اختصاص داده می‌شدند. شبکه‌ی کانولوشنالی گرافی فضایی-زمانی[۹] (ST-GCN) اولین الگوریتمی بود که به عنوان یک مدل‌های کانولوشنی گرافی برای تشخیص فعالیت مبتنی بر اسکلت پیشنهاد شد[۲]. این روش از رویکردهای پیشین متمایز بود زیرا می‌توانست اطلاعات بخشی از بدن را به طور ضمنی با استفاده از مکانی بودن کانولوشن گرافی همراه با دینامیک زمانی بیاموزد.

پس از آن برای به دست آوردن روابط بین مفاصل دور، برخی از روش‌های وابسته به داده پیشنهاد شده‌اند و مکانیسم افزوده‌ای را برای یادگیری سازگار رابطه بین مفاصل مختلف معرفی کردند [۲۷]. از سوی دیگر، برخی از رویکردها ویژگی‌های ساختاری چند مقیاسی[۱۰] را از طریق توابع چند جمله‌ای مرتبه بالاتر اسکلت استخراج کردند و ماژول‌های پرش چندگانه[۱۱] را برای از بین بردن محدودیت ظرفیت نمایشی ناشی از تقریب یک مرتبه معرفی کردند [۲۸]. متفاوت از این روش‌ها، ژان و همکاران شبکه کانولوشنی گرافی فضایی-زمانی چند مقیاسی[۱۲] (MST-GCN) را معرفی کردند [۲۹]. در این روش از پیچش‌هایی روی زیرمجموعه‌های گرافی[۱۳] که توسط اتصالات باقی‌مانده[۱۴] سلسله‌مراتبی[۱۵] می‌شوند استفاده کردند تا هم وابستگی اتصالات کوتاه برد و هم روابط اتصالات دور را استخراج کنند. همچنین این مدل میدان دریافت[۱۶] زمانی را در یک بلوک واحد بهبود می‌بخشد و اطلاعات زمانی کوتاه برد و بلند مدت را از طریق یک معماری سلسله مراتبی جمع می‌کند.

۳-۳- مدل ST-GCN

در این روش با توجه به توالی اسکلتی مفاصل بدن در قاب مختصات دو بعدی یا سه بعدی، یک گراف فضایی زمانی که در آن مفاصل به عنوان گره‌های گراف و اتصالات طبیعی در ساختار بدن انسان و زمان به عنوان یال‌های گراف است، ساخته

[9] Spatial-Temporal Graph Convolutional Networks

[10] Multi-scale

[11] Multiple-hop

[12] Multi-scale Spatial-Temporal Graph Convolutional Networks

[13] Sub-graph

[14] Residual

[15] Cascaded

[16] Receptive field

می‌شود. بنابراین ورودی این مدل، بردارهای مختصات مفاصل روی گره‌های گراف است. این را می‌توان مشابه CNN های مبتنی بر تصویر در نظر گرفت که در آن ورودی توسط بردارهای شدت پیکسل در چهارچوب تصویر دو بعدی تشکیل می‌شود. چندین لایه از عملیات پیچیدگی گرافی فضایی-زمانی بر روی داده‌های ورودی اعمال می‌شود و نقشه‌های ویژگی[17] سطح بالاتری را روی گراف ایجاد می‌کند. سپس توسط یک طبقه‌بند[18] استاندارد Softmax به دسته فعالیت مربوطه طبقه‌بندی می‌شود. کل مدل به روشی انتها به انتها با استفاده از پس‌انتشار[19] آموزش داده می‌شود. اکنون به بررسی اجزای ساختار مدل ST-GCN می‌پردازیم.

۳-۳-۱- ساخت گراف اسکلتی

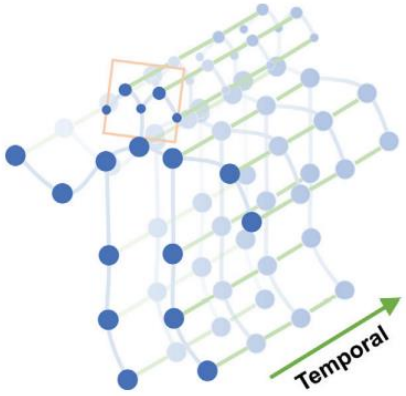
یک توالی اسکلتی معمولاً با مختصات دو بعدی یا سه بعدی هر مفصل انسان در هر قاب نشان داده می‌شود. در این مدل، از گراف فضایی-زمانی برای نمایش سلسله مراتبی توالی‌های اسکلتی استفاده می‌شود. به طور خاص، یک گراف زمانی فضایی بدون جهت $G = (V, E)$ را بر روی یک دنباله اسکلت با N مفصل و T قاب می‌سازیم که دارای اتصال درون بدنه و بین قاب است. در این گراف، مجموعه $V = \{v_{ti} \mid t = 1, \dots, T, i = 1, \dots, N\}$ شامل تمام مفاصل در یک دنباله اسکلت است. به عنوان ورودی ST-GCN، بردار ویژگی در گره $F(v_{ti})$ از بردارهای مختصات و همچنین مقدار اطمینان تخمین مفصل iام در قاب t تشکیل شده است. برای ورودی این مدل، گراف فضایی-زمانی را روی دنباله‌های اسکلت در دو مرحله می‌سازیم. ابتدا، مفاصل درون یک قاب با یال‌هایی مطابق با اتصال ساختار بدن انسان به هم متصل می‌شوند. سپس هر مفصل در قاب متوالی به همان مفصل متصل خواهد شد. بنابراین اتصالات در این تنظیمات به طور طبیعی بدون تخصیص دستی قطعه تعریف می‌شوند. این همچنین باعث می‌شود که معماری شبکه بتواند روی مجموعه داده‌هایی با تعداد اتصالات یا اتصالات مشترک متفاوت کار کند. به عنوان مثال، اگر در مجموعه داده‌ای از نتایج تخمین حالت دوبعدی استفاده شود که ۱۸ مفصل را خروجی می‌دهد، در حالی که در مجموعه داده دیگری از توالی اسکلتی سه بعدی استفاده شود که به عنوان ورودی، ۲۵ مفصل تولید باشد. ST-GCN را می‌توان در هر دو موقعیت تعریف کرد و عملکرد برتر را ارائه داد.

به طور دقیق، مجموعه یال E از دو زیر مجموعه تشکیل شده است، زیرمجموعه اول اتصال درون اسکلتی را در هر قاب به تصویر می‌کشد که با $E_s = \{v_{ti}v_{tj} | (i,j) \in H\}$ نشان داده می‌شود، که در آن H مجموعه مفاصل متصل به طور طبیعی در بدن انسان است. زیرمجموعه دوم شامل یال‌های بین قاب‌ها است که مفاصل مشابه را در قاب‌های متوالی به صورت $E_F = \{v_{ti}v_{(t+1)i}\}$ به هم متصل می‌کند. بنابراین تمام یال‌های E_F برای یک مفصل خاص i نشان دهنده مسیر حرکت آن در طول زمان است.

[17] Feature map

[18] Classifier

[19] Backpropagation



شکل ۳-۳ نحوه تشکیل و اتصال گراف اسکلتی در مدل GCN-ST [۲]

۳-۳-۲- عملیات کانولوشنی گرافی فضایی-زمانی

پیش از بررسی ساختار کامل شبکه‌ی ST-GCN، ابتدا به مدل شبکه کانولوشنی گرافی در یک قاب نگاه می‌اندزیم. در این حالت، در یک قاب τ، N گره مشترک V_t، همراه با یال‌های اسکلت $E_s(\tau) = \{v_{ti}v_{tj}|t = \tau, (i,j) \in H\}$ وجود خواهد داشت.

برای بررسی این حالت ابتدا به بررسی یک عملیات ساده کانولوشنی بر یک تصویر دو بعدی می‌پردازیم. با توجه به این فرض که پس از انجام عملیات پیچش با اندازه گام[۲۰] ۱ و لایه‌گذاری[۲۱] مناسب، نقشه‌های ویژگی خروجی می‌توانند هم اندازه نقشه‌های ویژگی ورودی باشند. یک عملگر پیچشی با اندازه فیلتر $K \times K$ و یک نقشه ویژگی ورودی f_{in} با تعداد کانال c، مقدار خروجی برای یک کانال تنها در مکان فضایی x می‌تواند به صورت نوشته شود:

$$f_{out}(x) = \sum_{h=1}^{k}\sum_{w=1}^{k} f_{in}\big(P(x,h,w)\big) \cdot W(h,w) \qquad (۱-۳)$$

که در آن $P : \mathrm{Z}^2 \times \mathrm{Z}^2 \to \mathrm{Z}^2$ تابع نمونه‌گیری[۲۲] همسایه مکان x را برمی‌شمارد و $W: \mathrm{Z}^2 \to \mathrm{R}^c$ تابع وزن را نشان می‌دهد. در این فرمول یک بردار وزن در فضای واقعی بعد c با بردارهای ویژگی ورودی نمونه برداری شده بعد c ضرب داخلی می‌شود.

[20] Stride

[21] Padding

[22] Sampling function

سپس عملیات پیچیدگی بر روی گراف‌ها با گسترش فرمول بالا به مواردی که نقشه ویژگی‌های ورودی بر روی یک گراف فضایی V_t قرار دارد، تعریف می‌شود. نقشه ویژگی $f_{in}^{t}: V_t \rightarrow R^c$ بر روی هر گره گراف یک بردار دارد. سپس، به تعریف مجدد تابع نمونه‌گیری P و تابع وزن W می‌پردازیم.

۳-۳-۲-۱- تابع نمونه‌گیری

در تصاویر، تابع نمونه‌برداری $P(h,w)$ بر روی پیکسل‌های مجاور با توجه به مرکز x تعریف می‌شود. در گراف‌ها، می‌توانیم به طور مشابه تابع نمونه‌برداری را در مجموعه همسایه برای گره‌ی v_{ti}، $B(v_{ti}) = \{v_{tj} | d(v_{tj}, v_{ti}) \leq D\}$ تعریف می‌شود. در این تعریف $d(v_{tj}, v_{ti})$ نشان دهنده حداقل طول هر مسیر از v_{tj} تا v_{ti} است. در این مدل حداقل فاصله بین دو گره ۱ در نظر گرفته می‌شود. بنابراین تابع نمونه‌گیری برای گره‌ی v_{ti} را می‌توان به صورت زیر نوشت:

$$P(v_{tj}, v_{ti}) = v_{tj} \qquad (۲-۳)$$

۳-۳-۲-۲- تابع وزن

در مقایسه با تابع نمونه برداری، تعریف تابع وزن برای یک گراف دشوارتر است. در پیچیدگی تصویری دوبعدی، پیکسل‌های داخل همسایگی نظم مکانی ثابتی دارند. سپس تابع وزن را می توان با نمایه‌سازی یک بردار با ابعاد (c, K, K) مطابق با نظم مکانی پیاده سازی کرد. برای گراف‌های معمولی مانند آنچه که در این مساله داریم، چنین ترتیب ضمنی وجود ندارد. راه حل این مشکل استفاده از یک فرآیند برچسب‌گذاری گره‌های موجود در همسایگی گره‌ی مرکزی است. در طراحی تابع وزن برای این مدل به‌جای آنکه به هر گره همسایه یک برچسب‌گذاری منحصربه‌فرد داده شود، فرآیند با تقسیم‌بندی مجموعه همسایه $B(v_{ti})$ یک گره مشترک v_{ti} به تعداد K زیرمجموعه، که در آن هر زیر مجموعه دارای یک برچسب عددی است، ساده می‌شود. بدین صورت اگر تابع برچسب گذاری را به صورت $l_{ti}: B(v_{ti}) \rightarrow \{0, \dots, K-1\}$ تعریف کنیم که یک گره در همسایگی را به برچسب زیر مجموعه‌اش نگاشت می‌کند. تابع وزن را می‌توان به حالت زیر بازتعریف کرد:

$$W(v_{tj}, v_{ti}) = W'(l_{ti}(v_{tj})) \qquad (۳-۳)$$

برای تقسیم‌بندی همسایگان یک گره به چند زیرمجموعه چند روش معرفی شده‌است [۲]، که پس از اتمام این بخش معرفی می‌شوند.

۳-۳-۲-۳- کانولوشن گرافی فضایی

حال با قرار دادن تابع وزن و تابع نمونه‌گیری معرفی شده در معادله‌ی (۳-۳)، می‌توان کانولوشن گرافی فضایی را به صورت زیر نوشت:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot W(l_{ti}(v_{tj})) \qquad (۴-۳)$$

که در آن عبارت $Z_{ti}(v_{tj})$ برای نرمال‌سازی اضافه شده است و برابر با تعداد اعضا(کاردینالیتی) زیرمجموعه مربوطه است. این عبارت نقش متعادل کردن سهم زیرمجموعه‌های مختلف در خروجی را دارد.

۳-۳-۲-۳- کانولوشن گرافی زمانی

با آشنایی حاصل شده از کانولوشن گرافی فضایی، اکنون می‌توان مدل‌سازی دینامیک فضایی-زمانی در توالی اسکلت را معرفی کرد. همانطور که قبلاً توضیح داده شد، در ساخت گراف، جنبه زمانی گراف با اتصال همان اتصالات در قاب‌های متوالی ساخته می‌شود. به همین دلیل می‌توان یک استراتژی بسیار ساده برای گسترش کانولوشن گرافی فضایی به حوزه فضایی-زمانی تعریف کرد. کافیست مفهوم همسایگی‌ها گسترش داده شود تا اتصالات زمانی را نیز در بر گیرد. بنابراین معادله تعریف همسایگی به صورت زیر خواهد شد:

$$B(v_{ti}) \rightarrow \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q-t| \leq \lfloor \Gamma/2 \rfloor\} \quad (۵-۳)$$

پارامتر Γ محدوده زمانی را که باید در گراف همسایه گنجانده شود را کنترل می‌کند و می‌توان آن را اندازه فیلتر زمانی نامید. برای تکمیل عملیات کانولوشن در گراف زمانی-فضایی، به تابع نمونه‌گیری نیز نیاز است که مشابه همان حالت تنها فضایی است. تابع وزن یا به طور دقیق‌تر تابع برچسب‌گذاری را نیز باید مشخص کرد. با استفاده از خاصیت ترتیبی بودن محور زمان میتوان تابع برچسب‌گذاری l_{ST} را برای همسایگی‌های گره‌ی v_{ti} به شکل زیر تعریف کرد:

$$l_{ST}(v_{qj}) \rightarrow l_{ti}(v_{tj}) + \left(q - t + \left\lfloor \frac{\Gamma}{2} \right\rfloor\right) \times K \quad (۶-۳)$$

۳-۳-۲-۴- روش‌های تقسیم‌بندی

۱) تک برچسب گذاری[۲۳]: ساده ترین و مستقیم ترین استراتژی تقسیم‌بندی، داشتن زیرمجموعه ای است که خود مجموعه همسایه کل است. در این استراتژی، بردارهای ویژگی در هر گره همسایه با یک بردار وزن یکسان ضرب داخلی خواهند داشت. در تعریف ریاضی داریم: $K = 1$ و $l_{ti}(v_{tj}) = 0, \forall i,j \in V$.

۲) برچسب زدن بر حسب فاصله[۲۴]: یکی دیگر از استراتژی‌های تقسیم‌بندی طبیعی، تقسیم‌بندی مجموعه همسایه بر اساس فاصله یک گره تا گره ریشه v_{ti} است. در این مدل، چون حداکثر فاصله بین دو گره همسایه ۱ فرض شده، مجموعه همسایه به دو زیر مجموعه تقسیم می‌شود، جایی که فاصله صفر است (به خود گره ریشه اشاره می‌کند) و گره‌های همسایه باقی‌مانده در زیر مجموعه جایی که فاصله ۱ است. بنابراین دو بردار وزنی متفاوت خواهیم داشت. در تعریف ریاضی داریم: $K = 2$ و $l_{ti}(v_{tj}) = d(v_{tj}, v_{ti})$.

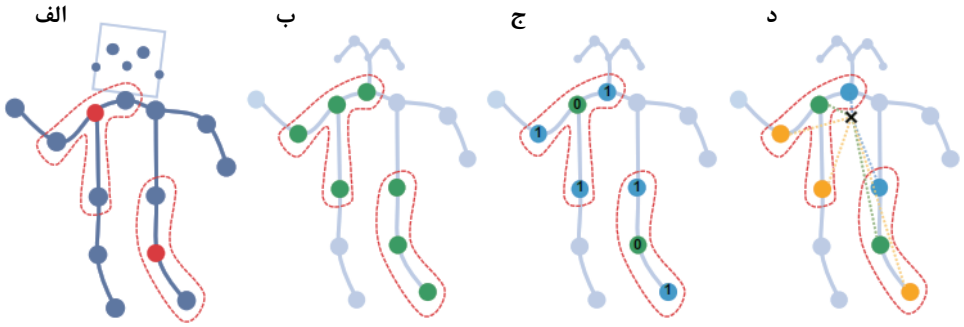
[23] Uni-labeling

[24] Distance partitioning

۳) برچسب زدن برحسب پیکربندی فضایی[۲۵]: از آنجایی که اسکلت بدن از نظر مکانی موضعی است، همچنان می‌توان از این پیکربندی فضایی خاص در فرآیند تقسیم‌بندی استفاده کرد. در این روش مجموعه همسایگان به سه زیرمجموعه تقسیم می‌شود: گره‌ی ریشه به عنوان یک زیرمجموعه در نظرگرفته می‌شود. سپس گروهی از گره‌های همسایه که به مرکز ثقل اسکلت نزدیکتر از گره ریشه هستند، در یک دسته قرار می‌گیرند. در غیر این صورت، گروهی از گره‌های همسایه که از مرکز ثقل اسکلت دورتر از گره ریشه هستند در گروه سوم قرار می‌گیرند. در این حالت مختصات متوسط تمام مفاصل اسکلت در یک قاب به عنوان مرکز ثقل آن در نظر گرفته می‌شود. این استراتژی از این واقعیت الهام گرفته شده است که حرکات اعضای بدن را می توان به طور کلی به عنوان حرکات متحدالمرکز و خارج از مرکز طبقه بندی کرد. در تعریف ریاضی داریم:

$$l_{ti}(v_{tj}) = \begin{cases} 0 & if\ r_j = r_i \\ 1 & if\ r_j < r_i \\ 2 & if\ r_j > r_i \end{cases} \qquad (۳-۷)$$

در فرمول بالا r_i میانگین فاصله مرکز ثقل تا مفصل i در تمام قاب‌های مجموعه آموزشی است. تصویر سه روش تقسیم‌بندی در شکل ۳-۴ نشان داده شده است. از چپ به راست: (الف) یک قاب نمونه از اسکلت ورودی. مفاصل بدن با نقاط آبی ترسیم می شوند. فیلدهای پذیرنده یک فیلتر با حداکثر فاصله‌ی ۱ با نقطه‌چین ترسیم شده‌اند. (ب) روش تک برچسب گذاری، که در آن همه گره ها در یک محله دارای یک برچسب هستند. (ج) تقسیم‌بندی بر حسب فاصله. دو زیر مجموعه عبارتند از خود گره ریشه با فاصله ۰ و سایر نقاط همسایه با فاصله ۱. (د) تقسیم‌بندی پیکربندی فضایی. گره‌ها بر اساس فاصله‌شان تا مرکز ثقل اسکلت (ضربدر سیاه) در مقایسه با گره ریشه برچسب‌گذاری می‌شوند. گره‌های گریز از مرکز فواصل کوتاه تری دارند، در حالی که گره‌های گریز از مرکز نسبت به گره ریشه فواصل طولانی تری تا مرکز دارند.



شکل ۳-۴ روش‌های تقسیم‌بندی برای ساخت عملیات کانولوشن[۲]

با توجه به توضیحات فوق در این قسمت به طور کامل می‌توان نحوه‌ی تعریف و عملکرد یک لایه کانولوشنی گرافی فضایی-زمانی را بر روی یک توالی اسکلتی ورودی متوجه شد.

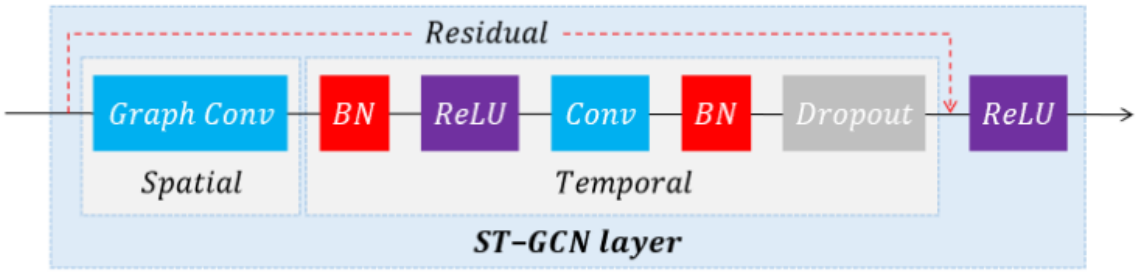
[25] Spatial configuration partitioning

۳-۳-۳- قابلیت یادگیری وزن‌دهی یال‌ها

با توجه به آن که مفاصل در هنگام انجام فعالیت به صورت گروهی حرکت می‌کنند، یک مفصل می‌تواند در چندین گروه از مفاصل بدن ظاهر شود. اما این ظاهرشدن باید در مدل‌سازی پویایی این قسمت‌ها اهمیت متفاوتی داشته باشند. به این معنا، ما یک ماسک قابل یادگیری M را روی هر لایه پیچیدگی نمودار فضایی-زمانی اضافه می‌کنیم. ماسک سهم اهمیت یک گره را به گره‌های همسایه اش بر اساس وزن قابل آموزش در هر یال گراف فضایی مشخص می‌کند. از نظر تجربی، افزودن این ماسک می‌تواند عملکرد تشخیص ST-GCN را بهبود بخشد.

۳-۳-۴- معماری شبکه‌ی عصبی ST-GCN

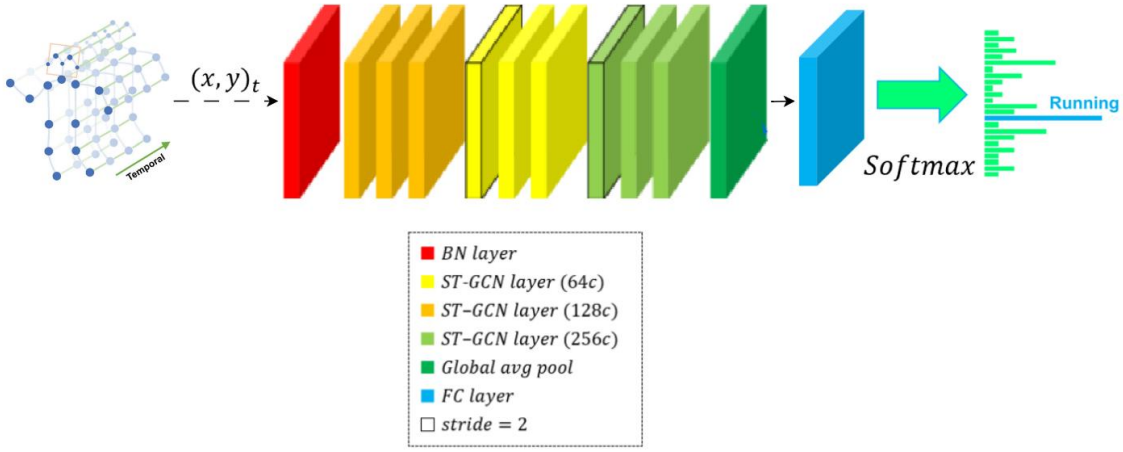
از آنجایی که ST-GCN وزن‌های مشترکی را روی گره‌های مختلف به اشتراک می‌گذارد، حفظ مقیاس داده‌های ورودی در مفاصل مختلف بسیار مهم است. به همین دلیل ابتدا اسکلت‌های ورودی به یک لایه نرمال‌ساز دسته‌ای[۲۶] وارد می‌شوند. مدل ST-GCN از ۹ لایه عملگر کانولوشنی گرافی فضایی-زمانی (واحدهای ST-GCN) تشکیل شده است. سه لایه اول دارای ۶۴ کانال برای خروجی هستند. سه لایه بعدی دارای ۱۲۸ کانال برای خروجی هستند. و سه لایه آخر ۲۵۶ کانال برای خروجی دارند. مکانیسم Resnet بر روی هر واحد ST-GCN اعمال می‌شود و به در انتهای هر لایه یک عملگر حذف تصادفی قرار گرفته است تا از بیش‌برازش جلوگیری شود. در نهایت پس از ۹ لایه واحدهای ST-GCN، یک میانگین‌گیری سراسری[۲۷] بر روی بردار خروجی حاصل انجام می‌شود تا یک بردار ویژگی ۲۵۶ بعدی برای هر دنباله اسکلتی بدست آید. در نهایت، از یک طبقه‌بندی کننده SoftMax برای تشخیص نهایی فعالیت استفاده شده است. تصویر توضیحات فوق در این بخش به صورت واضح و قابل درک در ادامه آورده شده است.



شکل ۳-۵ معماری یک لایه عملگر کانولوشنی گرافی فضایی-زمانی [۳۰]

[26] Batch normalization

[27] Global average pooling



شکل ۳-۶ معماری کلی شبکه عصبی GCN-ST [۳۰]

۳-۴- مدل MST-GCN

از آنجایی که کانولوشن گرافی یک عملیات محلی است، تنها می‌تواند از وابستگی‌های مفصل با برُد نزدیک استفاده کند. پس در مدل‌های همچون ST-GCN روابط مفاصل دور و اطلاعات زمانی دوربرد که برای تشخیص فعالیت‌های مختلف حیاتی هستند، بدون استفاده خواهند ماند. برای حل این مشکل، یک ماژول کانولوشنی گرافی فضایی چند مقیاسی[۲۸] (MS-GC) و یک ماژول کانولوشنی گرافی زمانی چند مقیاسی[۲۹](MS-GC) ارائه شده است. در این حالت، میدان دریافتی مدل در ابعاد فضایی و زمانی بهبود می‌یابد. بطور مشخص، ماژول‌های MS-GC و MT-GC پیچیدگی گراف مربوطه را به مجموعه‌ای از پیچیدگی‌های زیرگراف تجزیه می‌کنند و یک معماری باقی‌مانده سلسله‌مراتبی را تشکیل می‌دهند. در این حالت، ویژگی‌ها با یک سری از کانولوشن‌های زیرگرافی پردازش می‌شوند و هر گره می‌تواند چندین تجمیع فضایی و زمانی را با همسایگان خود کامل کند. بر این اساس، میدان دریافتی معادل نهایی بزرگ‌تر می‌شود، که می‌تواند وابستگی‌های کوتاه‌برد و بلندبرد را در حوزه‌های فضایی و زمانی ثبت کند. با اتصال این دو ماژول به عنوان یک واحد اساسی، یک شبکه کانولوشنی گرافی فضایی- زمانی چندمقیاسی(MST-GCN) تشکیل می‌شود که تعدادی از این واحدها را برای یادگیری نمایش‌های حرکتی مؤثر برای تشخیص فعالیت روی هم می‌گذارد [۲۹].

در این مدل نحوه‌ی تشکیل گراف ورودی از روی توالی اسکلتی همانند مدل قبل است. از دو تعریف ریاضی کانولوشن گرافی فضایی و کانولوشن گرافی زمانی که در مدل قبل توضیح داده شدند، استفاده خواهد شد و در این مدل این تعاریف بسط داده می‌شوند. در ادامه به معرفی و بررسی هر یک از این ماژول‌ها و معماری نهایی این مدل می‌پردازیم.

[28] Multi-Scale Spatial Graph Convolution

[29] Multi-Scale Temporal Graph Convolution

۳-۴-۱- ماژول کانولوشنی گرافی فضایی چند مقیاسی

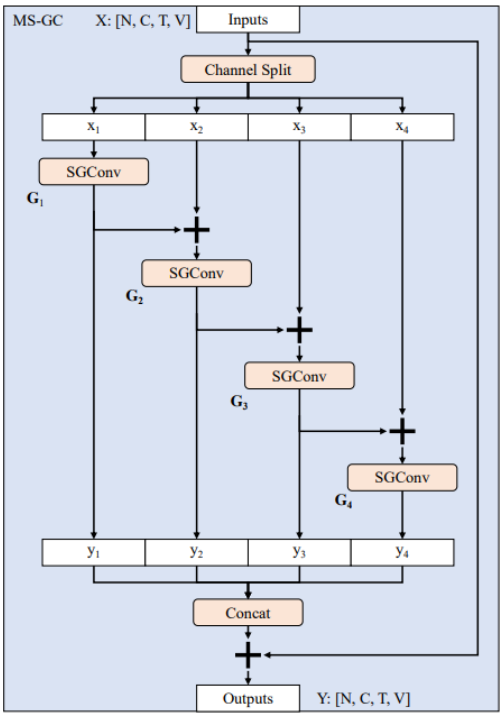
در ساختار این مدل، ماژول MS-GC را پیشنهاد شده است که در آن ویژگی‌های فضایی و عملیات کانولوشن گرافی محلی مربوط به این ویژگی‌ها به گروهی از زیرمجموعه‌ها تقسیم می‌شوند. ایده‌ی اصلی از Res2Net الهام گرفته شده است [۲۹]. Res2Net در سال های اخیر تأثیر مثبت زیادی در بسیاری از زمینه‌های پردازش تصویر داشته و سبب بهبود الگوریتم‌های موجود در مسائل مختلف شده است. در طراحی ماژول‌های این مدل، از ایده‌ی Res2Net برای حل مساله‌ی تشخیص فعالیت مبتنی بر اسکلت به کمک شبکه‌های کانولوشنی گرافی استفاده شده است.

معماری ماژول MS-GC پیشنهادی در شکل ۳-۷ نشان داده شده است. زیرمجموعه‌ها به عنوان یک معماری باقیمانده سلسله‌مراتبی فرموله شده‌اند، بنابراین ویژگی‌ها می‌توانند به صورت سلسله‌مراتبی پردازش شوند. بدین ترتیب، میدان دریافتی معادل بعد فضایی بزرگ شده است و مدل قادر به دریافت روابط بین آن مفاصل دور است.

به طور دقیق‌تر، اگر یک بردار ویژگی ورودی X با ابعاد $[C, T, V]$ به این ماژول وارد شود. در اینجا C تعداد کانال‌های ورودی، T تعداد قاب‌ها در طول زمان و V تعداد گره‌های گراف ورودی است. این بردار ویژگی به s قطعه در امتداد بعد کانال تقسیم می‌شود و هر قطعه را می‌توان x_i نامید، به طوری که$\mathrm{i} \in \{1, 2, \cdots, \mathrm{s}\}$ باشد. بنابراین اندازه‌ی ابعاد هر قطعه $\left[\frac{C}{s}, T, V\right]$ می‌شود. روی هر قطعه‌ی x_i یک عملیات پیچیدگی گرافی فضایی مطابق با معادله‌ی (۳-۴) انجام می‌شود (x_i به جای f_{in} در این معادله جایگذاری می‌شود) که می‌توان حاصل آن را G_i (همان f_{out} در معادله (۳-۴)) نامید. تعداد هر کانولوشن گرافی فرعی $\frac{1}{s}$ تعداد کانال در مقایسه با کانال‌های اصلی است و بر این اساس تنها متغیرهای آن $\frac{1}{s^2}$ پارامترهای کانولوشن کلی خواهد بود. علاوه بر این، در این مدل، اتصالات باقی‌مانده بین دو قطعه مجاور قرار می‌گیرد که تنوع میدان‌های گیرنده را غنی می‌کند تا شبکه عصبی هم وابستگی‌های بین مفاصل محلی و هم غیرمحلی را استخراج کند و یاد بگیرد. به زبان ریاضی می‌توان این قسمت از ماژول MS-GC را این گونه تعریف کرد:

$$y_i = \begin{cases} G_i(x_i) & i = 0 \\ G_i(x_i + y_{i-1}) & i > 0 \end{cases} \qquad (۸-۳)$$

در این معادله y_i خروجی کانولوشن گرافی فرعی قطعه iام است.



شکل ۳-۷ نمایشی از معماری ماژول کانولوشنی گرافی فضایی چند مقیاسی. N اندازه‌ی دسته[۳۰] است. [۲۹]

در این ماژول، قطعه‌ها دارای فیلدهای دریافتی متفاوتی هستند. به عنوان مثال، اگر حداکثر فاصله بین دو گره‌ی همسایه را ۱ فرض کنیم، G_1 می‌تواند اطلاعات همسایگانش با فاصله یکی[۳۱] را جمع‌آوری کند، در حالی که G_2 به طور بالقوه می‌تواند اطلاعات ویژگی‌ها را از همسایگان با فاصله دوتایی به کمک جمع‌آوری اطلاعات از y_1 دریافت کند. بنابراین میدان دریافتی معادل آخرین قطعه y_S چندین بار بزرگ شده است. در نهایت، تمام قطعات به هم متصل می‌شوند و یک اتصال باقیمانده اضافی برای کل ماژول برای کمک به همگرایی مدل در نظر گرفته می‌شود. خروجی‌های ماژول MS-GC را می‌توان به صورت زیر محاسبه کرد:

$$Y = \sigma([y_1; \dots; y_S] + X) \qquad (۹-۳)$$

در معادله فوق σ تابع فعال سازی است. بردار ویژگی خروجی Y شامل اطلاعات ویژگی‌های فضایی بین گره‌ها در فواصل مختلف است که نسبت به نمایش‌های فضایی محلی به دست آمده با استفاده از یک کانولوشن گرافی محلی واحد که در مدل ST-GCN وجود داشت، برتری دارد. ماژول MS-GC می‌تواند تعادل بین پیچیدگی و توانایی نمایش چند مقیاسی مدل را با تنظیم S کنترل می کند. این ماژول قادر است وابستگی‌های بین هر دو اتصال کوتاه و بلند را بدون نیاز معرفی پارامترهای اضافی و عملیات زمان‌بر، استخراج کند.

[30] 1-hop

[31] Batch

۳-۴-۲- ماژول کانولوشنی گرافی زمانی چند مقیاسی

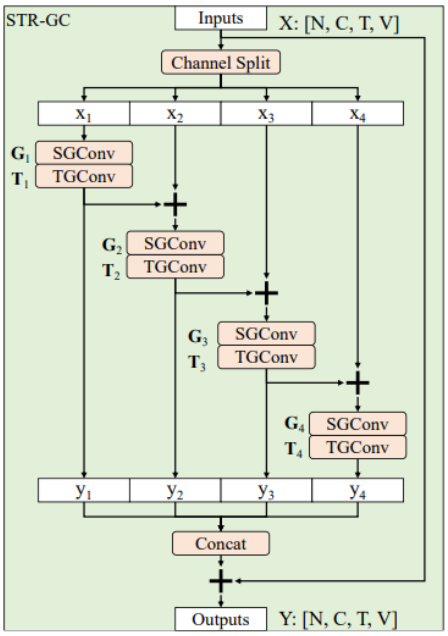
مدل‌سازی زمانی برای تشخیص فعالیت ضروری است، در حالی که مدل‌سازی زمانی دوربرد تا حد زیادی در کارهای قبلی نادیده گرفته شده است. وابستگی زمانی دوربرد نه تنها می تواند ابهام بین فعالیت‌های مختلف را از طریق اطلاعات زمانی کاهش دهد، بلکه اطلاعات کلی را برای کمک به مدل برای یادگیری بهتر ویژگی‌های زمانی مکانی فراهم می‌سازد. بسیاری از کارهای موجود، مدل‌سازی زمانی را با استفاده از کانولوشن‌های زمانی با اندازه فیلتر ثابت در سراسر معماری انجام می‌دهند[۲، ۳۱]. رابطه زمانی دوربرد به طور غیرمستقیم با انباشته شدن مکرر کانولوشن‌های گرافی زمانی محلی در شبکه‌های عمیق مدل‌سازی می‌شود. با این حال، پس از تعداد زیادی عملیات کانولوشن محلی، ویژگی‌های مفید قاب‌های دور تضعیف می‌شوند و نمی‌توان اطلاعات حاصل از آن‌ها را به خوبی ثبت کرد.

برای رفع مشکلی که قبلاً توضیح داده شد، در این مدل به طور طبیعی ماژول MS-GC به دامنه زمانی گسترش داده شده است. ماژول MT-GC پیشنهادی ساختاری مشابه MS-GC دارد، اما اتصالات باقیمانده اضافی را معرفی نمی کند. این ماژول کانولوشن گرافی زمانی محلی را با مجموعه‌ای از پیچیدگی‌های گرافی فرعی جایگزین می‌کند که به صورت اتصالات باقی‌مانده سلسله مراتبی ساخته می‌شوند. پیچیدگی گرافی زمانی در هر زیرمجموعه یکسان است، اما ورودی‌های متفاوتی دارد. مشابه ماژول MS-GC، هنگامی که ویژگی‌های فضایی زمانی از طریق ماژول MT-GC می‌گذرد، یک سری عملیات کانولوشنی گرافی زمانی آبشاری بر روی قطعات مربوطه اعمال می‌شود تا میدان پذیرش زمانی بزرگ شود. بنابراین، خروجی نهایی دارای نمایش زمانی چند مقیاسی است در حالی که هر دو روابط زمانی کوتاه‌برد و بلندبرد به خوبی ثبت شده اند.

۳-۴-۳- معماری شبکه‌ی عصبی MST-GCN

در قسمت‌های پیشین به معرفی ماژول‌های استفاده شده در مدل MST-GCN پرداختیم. حال به توضیح چگونگی اتصال این ماژول‌ها در این مدل و ساختار کلی این مدل می‌پردازیم. این مدل در واقع مدل ST-GCN که پیش‌تر معرفی شد را گسترش داده است و در همان معماری تغییراتی ایجاد می‌کند. برای ترکیب ماژول MS-GC با ماژول MT-GC در این مدل یک واحد کانولوشنی گرافی باقیمانده فضایی زمانی (STR-GC) ساخته می‌شود که در شکل ۳-۸ نشان داده شده است. در این واحد دو ماژول به طور پشت سرهم در یک بلوک واحد قرار می‌گیرند. برای راحتی، کانولوشن گرافی زمانی به عنوان T_i نشان داده می‌شود. در این ماژول، ویژگی‌های مکانی و زمانی به‌طور متناوب در هر زیر مجموعه به‌روزرسانی می‌شوند و میدان دریافتی مکانی و زمانی مربوطه بزرگ‌تر می‌شود. به زبان ریاضی واحد حاصل از ترکیب این دو ماژول را می‌توان به صورت تعریف می‌شود:

$$y_i = \begin{cases} T_i(G_i(x_i)) & i = 0 \\ T_i(G_i(x_i + y_{i-1})) & i > 0 \end{cases} \qquad (۱۰-۳)$$



شکل ۳-۸ نحوه‌ی اتصال ماژول‌های فضایی و زمانی در مدل GCN-MST، ساختار یک بلوک GC-STR [۲۹]

در نهایت این مدل با کنار هم قرار دادن پشت سر هم ۱۰ بلوک STR-GC ساخته شده است، چهار بلوک اول دارای ۶۴ کانال برای خروجی است. در بلوک های ۵ و ۸، تعداد کانال‌ها دو برابر می‌شود. در نهایت هم یک میانگین‌گیری سراسری بر روی بردار خروجی بلوک‌ها انجام و پس از آن از یک شبکه عصبی کاملا متصل[۳۲] برای تشخیص نهایی فعالیت استفاده شده است.

۳-۵- خلاصه

در این فصل، مروری بر پیشینه‌ی شبکه‌های کانولوشنی گرافی و کاربرد آن‌ها در مساله‌ی تشخیص فعالیت انسان داشتیم. به توضیح مفاهیم اولیه‌ی شبکه‌های کانولوشنی گرافی پرداختیم و پس از آشنایی با مفاهیم اولیه معماری و منطق ساختاری دو مدل استفاده شده در این پروژه را شرح دادیم. ابتدا مدل ST-GCN را بررسی کردیم که به عنوان اولین مدل مبتنی بر شبکه‌های کانولوشنی گرافی برای تشخیص فعالیت معرفی شده بود. سپس به بررسی مدل MST-GCN پرداختیم که تعمیمی از مدل اول است، با این تفاوت که این مدل قابلیت یادگیری وابستگی‌های دوربرد بین مفاصل را نیز دارد.

[32] Fully Connected Neural Network

فصل چهارم

آزمایش‌ها و ارزیابی

در این فصل به بررسی مدل‌های تشخیص فعالیت مبتنی بر اسکلت پیشنهادی و مقایسه آن‌ها با یکدیگر با استفاده از داده اسکلتی بدست آمده از MediaPipe Pose می‌پردازیم. در ابتدای این فصل مجموعه داده‌های استفاده شده را معرفی می‌کنیم و مشخصات کلی آن به همراه مزیت و معایبی که دارد را بیان می‌کنیم. سپس چند نمونه از خروجی‌های دو مدل از قبل آموزش داده شده‌ی تخمین حالت بدن را نمایش می‌دهیم. با توجه به عملکرد کلی و هزینه محاسباتی این دو مدل، تصمیم بر آن شد تا تنها از مدل MediaPipe Pose در آموزش و راه‌اندازی مدل‌های تشخیص فعالیت مبتنی بر اسکلت بر روی مجموعه‌داده انتخابی، استفاده کنیم.

برای پیاده‌سازی مدل‌های تشخیص فعالیت از کتابخانه پایتورچ در زبان برنامه نویسی پایتون استفاده شده است. برای پیاده‌سازی این مدل‌ها از معماری و اصول طراحی معماری ذکر شده در مقاله‌های آن‌ها استفاده شده است. آزمایش‌ها مختلفی را بر که روی مدل‌ها پیشنهادی انجام گرفته است را توضیح می‌دهیم و تغییرات حاصل در زمان اجرا و عملکرد آن‌ها را گزارش می‌کنیم. این آزمایش‌های مختلف شامل تغییراتی در مراحل پیش پردازش، تنظیم نرخ یادگیری، تغییر نوع تابع وزن‌دهی در شبکه‌ی کانولوشنی گرافی و تغییر سایر پارامترهای موثر می‌شود. تمام ارزیابی‌ها و آزمایش‌های انجام شده دارای مقدار ۳۲ برای اندزاه دسته[1]های یادگیری هستند. در تمام آزمایش‌ها، ماکسیمم تعداد ایپاک[2] در مرحله آموزش مدل‌ها ۲۰۰ فرض شده است و یک روش توقف زودهنگام برای جلوگیری از بیش‌برازش[3] در نظر گرفته شده است. همچنین عملکرد نهایی مدل‌ها به صورت رتبه‌بندی گزارش خواهد شد؛ به شکلی که عملکرد مدل را به دو صورت برترین شناسایی و ۳ شناسایی برتر[4] بررسی می‌کنیم. در فرآیند آموزش و اعتبارسنجی مدل‌ها، معیارهای دیگری نیز گزارش شده است که در ادامه توضیح خواهیم داد.

۴-۱- مجموعه دادگان

در این پروژه از مجموعه دادگان Kinetics400 استفاده کرده‌ایم این مجموعه داده به صورت عمومی در اختیار همگان قرار دارد و می‌توان آن را از این لینک دریافت کرد [٦]. این مجموعه دادگان که توسط گروه محققان DeepMind در سال ۲۰۱۷ به عنوان یک مجموعه دادگان بزرگ برای تشخیص فعالیت انسان ارائه شده است، شامل ۴۰۰ کلاس فعالیت انسانی است. هر کلیپ حدود ۱۰ ثانیه است و از یک ویدیوی یوتیوب[5] متفاوت گرفته شده است. فعالیت‌ها متمرکز بر انسان هستند و طیف وسیعی از طبقات را شامل می‌شوند که شامل تعاملات انسان و اشیاء مانند نواختن سازها و همچنین تعاملات انسان و انسان مانند دست دادن است.

[1] Batch

[2] Epoch

[3] Overfitting

[4] Top 3 accuracy

[5] YouTube

مجموعه دادگان Kinetics را می‌توان به عنوان جانشین دو مجموعه داده ویدئویی فعالیت انسان HMDB-51 [۳۲] و UCF-101 [۳۳] که پیش‌تر به عنوان معیارهای استاندارد برای تشخیص فعالیت انسان معرفی شده بودند، دانست. این دو مجموعه داده قدیمی‌تر به خوبی به جامعه علمی خدمت کرده‌اند، اما سودمندی آن‌ها اکنون در حال انقضاء است. این به این دلیل است که آن‌ها به اندازه کافی بزرگ نیستند یا دارای تنوع کافی برای آموزش و آزمایش نسل فعلی مدل‌های طبقه‌بندی فعالیت انسان بر اساس یادگیری عمیق نیستند. از دیگر نقاط ضعف مجموعه دادگان پیشین تنوع محدود ویدیوهای موجود در آن‌هاست؛ به عنوان مثال، ۷ کلیپ از یک ویدیو از یک شخص در حال برس زدن موهای خود وجود دارد. این بدان معنی است که تنوع نسبت به زمانی که عمل در هر کلیپ توسط شخص دیگر و در شرایط نوری و محیط متفاوتی انجام شود، بسیار کمتر است. این مشکل در Kinetics وجود ندارد زیرا هر کلیپ از یک ویدیوی متفاوت گرفته شده است.

در Kinetics کلیپ‌ها از ویدیوهای یوتیوب تهیه شده‌اند. در نتیجه، در بیشتر موارد، آن‌ها به صورت حرفه‌ای ویدئو و مطالب ویرایش شده نیستند (مانند فیلم‌های تلویزیونی). بنابراین، در ویدیوهای این مجموعه دادگان می‌تواند حرکت/ لرزش قابل‌توجه دوربین، تغییرات نور، سایه‌ها، به هم ریختگی پس‌زمینه، و غیره وجود داشته باشد. مهم‌تر از آن، تعداد زیادی از اجراکنندگان (زیرا هر کلیپ از یک ویدیوی متفاوت است) با تفاوت‌هایی در نحوه انجام فعالیت وجود دارد (مثلاً سرعت)، لباس، حالت بدن و شکل بدن، سن، و کادربندی دوربین و دیدگاه. این شرایط سبب می‌شود که داده‌های Kinetics به شدت شبیه به سناریوهای دنیای واقعی باشند و این مساله‌ی تشخیص فعالیت انسان را بسیار چالش برانگیز می‌کند.

۴-۱-۱- محتوای مجموعه دادگان Kinetics 400

این مجموعه دادگان بر اعمال انسان (به جای رویدادها) متمرکز است. فهرست کلاس‌های فعالیت‌ها این موارد را شامل می‌شود: فعالیت‌های شخصی (مفرد)، به‌عنوان مثال. نقاشی، نوشیدن، خندیدن، مشت زدن. فعالیت‌های شخص-شخص، به عنوان مثال. در آغوش گرفتن، بوسیدن، دست دادن؛ و، فعالیت‌های شخص-اشیاء، به عنوان مثال. پخت غذاهای مختلف، چمن زنی، شستن ظروف. برخی از اقدامات ریزدانه هستند و برای تشخیص نیاز به استدلال زمانی دارند، به عنوان مثال انواع مختلف شنا. اقدامات دیگر برای تمایز نیاز به تأکید بیشتری بر روی اشیاء دارد، به عنوان مثال نواختن انواع مختلف سازهای بادی. این مجموعه دادگان دارای ۴۰۰ کلاس فعالیت انسانی است، با ۴۰۰-۱۱۵۰ کلیپ برای هر فعالیت، که هر کدام از یک ویدیوی منحصر به فرد است. هر کلیپ حدود ۱۰ ثانیه طول می‌کشد. نسخه فعلی دارای ۳۰۶۲۴۵ ویدیو است و به سه گروه تقسیم شده است، یکی برای آموزش دارای ۲۵۰ تا ۱۰۰۰ ویدیو در هر کلاس، یکی برای اعتبارسنجی با ۵۰ ویدیو در هر کلاس و دیگری برای آزمایش با ۱۰۰ ویدیو در هر کلاس. آمار در جدول ۱ آورده شده است. کلیپ‌ها از ویدیوهای یوتیوب هستند و وضوح و نرخ قاب متغیری دارند.

جدول ۴-۱ تعداد کلیپ ها برای هر کلاس در قسمت‌های آموزش/ اعتبارسنجی/تست

داده آموزش	داده اعتبارسنجی	داده تست
۱۰۰۰ - ۲۵۰	۵۰	۱۰۰

۴-۱-۲- زیرمجموعه دادگان انتخاب شده برای پیاده‌سازی

آموزش شبکه‌ها و انجام آزمایش‌های مختلف در آموزش شبکه‌ها بر روی کل مجموعه دادگان Kinetics (۴۰۰ کلاس فعالیت مختلف که هر کلاس حداقل ۴۰۰ ویدیو دارد) بسیار هزینه بر است. همچنین، تعداد زیادی از این کلاس‌ها فعالیت‌های مربوط به انسان-شی هستند به طوری که نمیتوان از اسکلت بدن اطلاعات مفیدی برای تشخیص فعالیت بدست آورد. به همین دلایل، در انجام این پروژه از ۳۸ کلاس از ویدیوهای این مجموعه دادگان استفاده می‌کنیم. سعی شده است کلاس‌هایی انتخاب شوند که اسکلت بدن در انجام آن فعالیت‌ها نقش به خصوصی داشته باشد. از لحاظ آماری همچنان تعداد کلیپ‌ها برای هر کلاس در قسمت‌های آموزش، اعتبارسنجی و تست مانند جدول ۱ است. زیرمجموعه انتخاب شده برای انجام مراحل این پروژه، شامل کلاس‌هایی است که در شکل ۴-۱ آورده شده‌اند.

```
['archery', 'bench pressing', 'bouncing on trampoline', 'bowling', 'clapping',
 'climbing a rope', 'cracking neck', 'crawling baby', 'dancing macarena',
 'disc golfing', 'doing aerobics', 'dribbling basketball',
 'dunking basketball', 'grinding meat', 'hammer throw', 'high jump',
 'high kick', 'hockey stop', 'hurdling', 'jogging', 'jumping into pool',
 'kicking soccer ball', 'playing drums', 'playing tennis', 'playing ukulele',
 'playing violin', 'pole vault', 'presenting weather forecast', 'pull ups',
 'recording music', 'riding mechanical bull', 'riding or walking with horse',
 'robot dancing', 'running on treadmill', 'shearing sheep', 'skiing slalom',
 'sword fighting', 'tying bow tie']
```

شکل ۴-۱ کلاس‌های انتخاب شده برای انجام مراحل پروژه

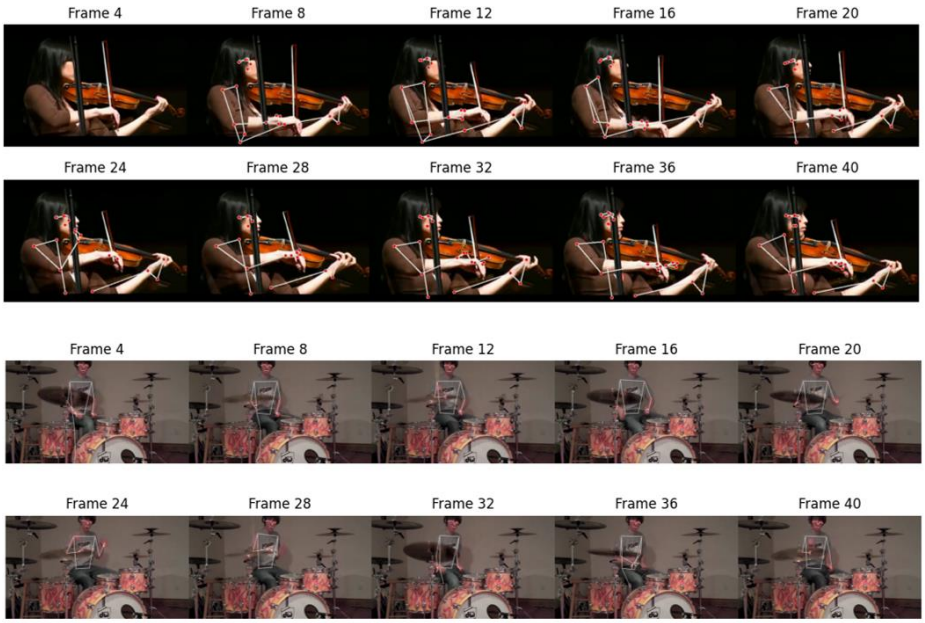
۴-۲- تخمین حالت بدن روی داده‌های ویدیویی

همان طور که قبلاً توضیح داده شد در این پروژه از دو مدل Leightweight OpenPose و MediaPipe Pose برای تخمین حالت استفاده کرده‌ایم. به منظور آنکه اجرای هر دوی این الگوریتم‌ها برای استخراج داده اسکلتی بر روی کل داده‌های مجموعه دادگان بسیار زمان‌بر و هزینه‌بر بود، تصمیم بر آن شد که یکی از این دو الگوریتم را بر روی کل دادگان اجرا کنیم. به این منظور در ابتدا، تخمین حالت بدن و استخراج داده اسکلتی را روی تعدادی از نمونه داده‌های موجود با هر دو الگوریتم انجام دادیم. در راه‌اندازی Leightweight OpenPose از مدل از قبل آموزش داده شده که در گیت‌هاب[۶] موجود بود استفاده شد و برای راه‌اندازی MediaPipe Pose از واسط برنامه نویسی کاربردی چهارچوب MediaPipe استفاده کردیم.

به صورت کیفی، داده‌های حاصل از MediaPipe Pose دقیق‌تر و کاربردی‌تر به نظر آمد. این روش همچنین سرعت استنتاج بالاتری نسبت به Leightweight OpenPose داشت. به همین سبب، همه دادگان را با به الگوریتم MediaPipe Pose دادیم تا داده اسکلتی حاصل از تخمین حالت بدن مجموعه دادگان مورد نظر بدست آید.

[6] https://github.com/Daniil-Osokin/lightweight-human-pose-estimation.pytorch

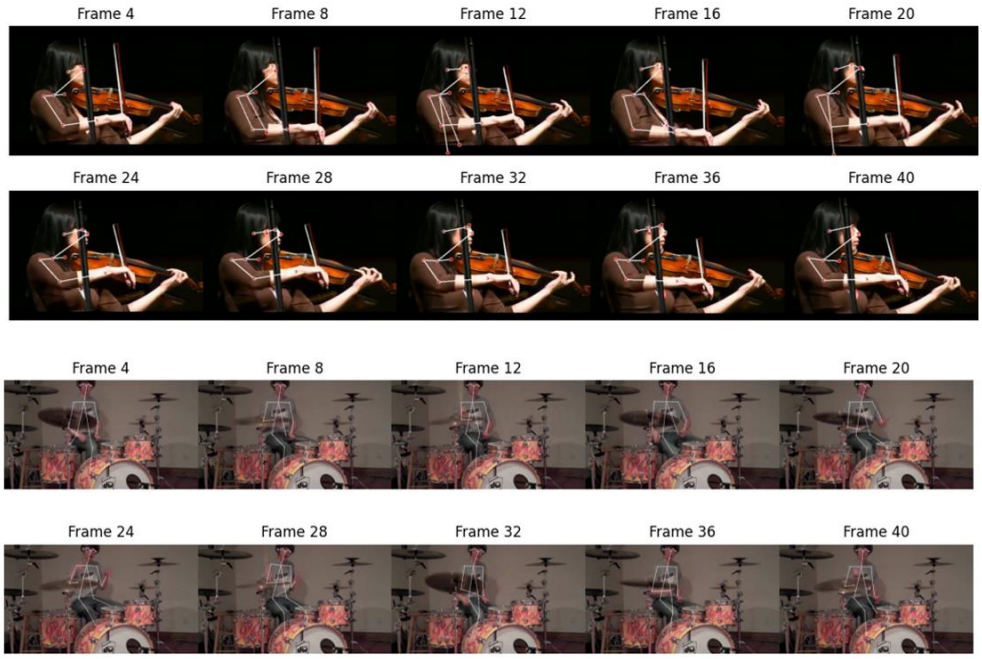
خروجی الگوریتم MediaPipe Pose برای هر نمونه داده ویدیویی، لیستی به طول تعداد قاب‌های ویدیو است که هر عضو این لیست شامل یک دیکشنری حاوی اطلاعات مقدار X و Y هر مفصل در قاب مدنظر است. دامنه مقدار X و Y عددی بین ۰ تا ۱ است. علاوه بر مختصات هر مفصل، این الگوریتم امتیازی به عنوان امتیاز پدیداری[7] را به خروجی هر مفصل اختصاص می‌دهد. این امتیاز نشان دهنده احتمال قابل مشاهده بودن هر مفصل در تصویر است. دامنه این امتیاز هم عددی بین ۰ تا ۱ است. معمولا زمانی که این امتیاز کمتر از ۰.۵ باشد مفصل در تصویر قابل مشاهده نیست. در مرحله‌ی بعدی که آموزش مدل تشخیص فعالیت مبتنی بر اسکلت است، می‌توان این عدد امتیازی پدیداری را هم در کنار مختصات مفصل به عنوان یک ویژگی به مدل وارد کرد. اگر الگوریتم قادر نباشد در یک قاب هیچ مفصلی را تشخیص دهد یا به عبارتی حضور انسان را در قاب نفی کند، خروجی برای آن قاب None خواهد بود. این روش به طور خودکار پیش از شروع تخمین حالت، داده‌ها را از نظر ارتفاع و عرض قاب تصویر هم اندازه می‌کند. مقدار قاب بر ثانیه[8] در ویدیوها پیش از اجرای الگوریتم مقدار ثابت ۳۰ تنظیم شده است. برخی نمونه خروجی‌های این دو مدل تخمین حالت بدن بر چند نمونه از مجموعه دادگان را می‌توان در اشکال ۴-۲ و ۴-۳ مشاهده کرد.



شکل ۴-۲ نمونه خروجی داده اسکلتی بدست آمده توسط مدل MediaPipe Pose

[7] Visibility score

[8] Frame Per Second(FPS)



شکل ۴-۳ نمونه خروجی داده اسکلتی بدست آمده توسط مدل Lightweight OpenPose

۴-۳- پیش‌پردازش داده‌های اسکلتی

در این پروژه مراحل مختلفی برای پیش پردازش داده‌ی اسکلتی انجام داده‌ایم. برخی از مراحل پیش پردازش یک بار روی تمام دادگان انجام شده است و در آزمایش‌ها مختلف بر روی شبکه‌های پیاده‌سازی شده ثابت است. برخی پیش پردازش‌ها نیز به منظور انجام آزمایش‌ها مختلف صورت گرفته و نتایج اعمال آن‌ها در بخش‌های پیش‌رو گزارش شده است.

۴-۳-۱- پیش‌پردازش‌های انجام شده‌ی ثابت در آزمایش‌ها

پیش پردازش ثابت انجام شده بر روی کل مجموعه دادگان در سه مرحله انجام شده است. ابتدا دادگانی که تعداد زیادی قاب بدون اسکلت داشتند را از مجموعه‌دادگان حذف کرده‌ایم. در این مرحله از پیش‌پردازش دادگانی که در ویدیوی پردازش شده، در بیش از ۷۰٪ قاب‌های آن‌ها حضور فرد تشخیص داده نشده بود را به طور کامل از مجموعه دادگان حذف کردیم. به طور مثال اگر ویدیو از ۳۰۰ قاب تشکیل شده بود و برای بیش از ۲۱۰ قاب آن هیچ اسکلتی تخمین زده نشده بود، این داده را از مجموعه دادگان حذف کردیم.

در مرحله دوم، مفاصلی که در تصویر وجود قابل شناسایی نبودند را بی اثر کردیم. مفاصلی که مقدار امتیاز پدیداری آن‌ها کمتر از ۰.۵ تخمین زده شده بود را به عنوان مفاصلی که در تصویر قابل رویت نیستند در نظر گرفتیم و مقادیر X، Y و امتیازی پدیداری را برای این مفاصل برابر با صفر قرار دادیم. علت انجام این کار این بود که اکثر الگوریتم‌های تخمین حالت اگر مفصلی

را در تصویر تشخیص ندهند مقادیر آن را NaN گزارش می‌دهند اما در الگوریتم MediaPipe Pose مفاصلی که قابل تشخیص نیستند با X و Y منفی گزارش می‌شدند و برای بی اثر شدن این مفاصل مقادیر آن‌ها را به صفر تغییر دادیم.

در مرحله سوم، با مشاهده آنکه در اکثر مواقع در قاب‌هایی که برای آن‌ها هیچ اسکلتی تخمین زده نشده است، واقعا شخصی وجود ندارد؛ برای هر داده‌ی نمونه قاب‌هایی که در آن‌ها هیچ اسکلتی تشخیص داده نشده است را حذف نمودیم.

۴-۳-۲- پیش پردازش‌های متغیر بررسی شده در آزمایش‌ها

دو نوع پیش پردازش دیگر پیش از ورود داده توالی اسکلتی به مدل‌های تشخیص فعالیت، می‌توان انجام داد. سپس طبق نحوه انجام آزمایش در هر دو مقاله‌ی ST-GCN و MST-GCN، تعداد کل قاب‌های هر داده‌ی نمونه را برابر با ۳۰۰ قرار می‌دهیم. این کار را با پخش مجدد قاب‌ها از ابتدا تا زمانی که تعداد آن‌ها به ۳۰۰ قاب برسد پیاده‌سازی می‌کنیم. در واقع نوعی لایه گذاری[9] در داده توالی اسکلتی انجام می‌دهیم. از دیگر مراحل پیش پردازش پیشنهاد شده در مقاله‌ی مدل MST-GCN، آن است که پس از انجام لایه گذاری، یک پنجره به اندازه‌ی ۱۵۰ قاب از داده‌ی حاوی ۳۰۰ قاب، نمونه برداری کنیم. در واقع، به صورت تصادفی یک بازه‌ی حاوی ۱۵۰ قاب از داده نمونه را انتخاب کنیم. این کار به شدت در سرعت پردازش مدل تاثیر گذار خواهد بود زیرا تعداد قاب‌های کل دادگان نصف می‌شوند. در آزمایش‌ها و ارزیابی‌های انجام شده، این دو نوع پیش پردازش را امتحان کرده‌ایم و تاثیر هر یک را در قسمت‌های پیش‌رو بررسی خواهیم کرد.

۴-۴- آموزش و ارزیابی مدل‌های ST-GCN و MST-GCN

در این قسمت به بررسی آزمایش‌های انجام شده می‌پردازیم و نتایج آن‌ها را گزارش می‌کنیم. در این بخش ابتدا نتایج آموزش یک مدل شبکه کانولوشنی گرافی سه لایه ساده را بر روی مجموعه دادگان گزارش می‌کنیم. سپس همان آزمایش را بر روی دو مدل معرفی شده‌ی ST-GCN و MST-GCN، تکرار می‌کنیم و با بررسی نتایج به قدرت و نحوه عملکرد دو مدل پیش‌تر معرفی شده، در مقایسه با یک شبکه‌ی عصبی کانولوشنی گرافی ساده قابل درک باشد. برای پیاده‌سازی هر دو مدل ST-GCN و MST-GCN به ترتیب مخزن کد st-gcn[10] و MST-GCN[11] موجود در گیت‌هاب استفاده کردیم و برای پیاده سازی مدل GCN سه لایه از کتابخانه‌ی DIG استفاده کردیم[۳٤]. همه مدل‌ها با استفاده از چهارچوب پایتورچ پیاده‌سازی شده اند.

همچنین آزمایش‌ها مختلفی در شرایط آموزش مختلف، با تغییراتی در ورودی (نحوه پیش پردازش) و همچنین تغییراتی در معماری مدل‌ها صورت گرفته است. در بخش دوم به توضیح و تفسیر این آزمایش‌ها بر روی شبکه‌های ST-GCN و MST-

[9] Padding

[10] https://github.com/yysijie/st-gcn

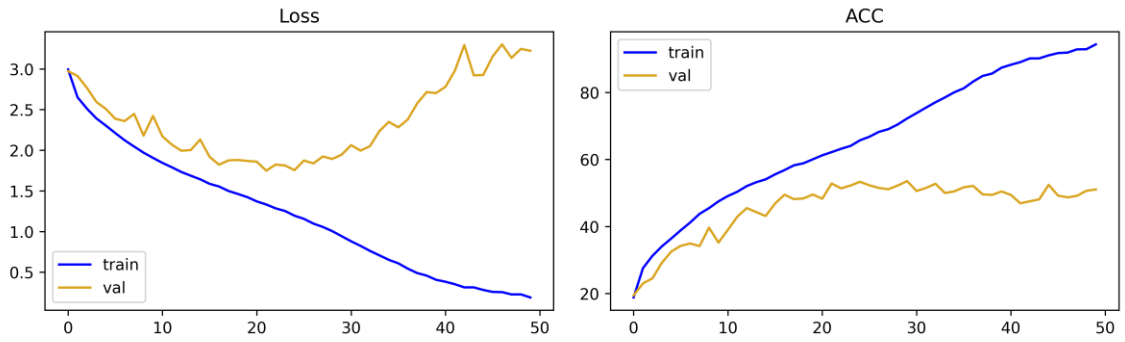
[11] https://github.com/czhaneva/MST-GCN

GCN می‌پردازیم و نتایج را بررسی می‌کنیم. در این آزمایش‌ها تاثیر پیش پردازش‌های مختلف، حالت‌های مختلف نرخ آموزش و تابع بهینه‌ساز، تاثیر افزودن حذف تصادفی[12] و استفاده از توابع وزن‌دهی مختلف در این مدل‌ها را بررسی کرده ایم.

در تمامی آزمایش‌ها ویژگی‌های ورودی به صورت یک تانسور با ابعاد (N, T, V, C) به مدل داده می‌شوند. در این تانسور N برابر اندازه دسته است که در تمام آزمایش‌ها برابر ۳۲ قرار گرفته است. T تعداد قاب نمونه‌هاست که در اکثر آزمایش‌ها برابر ۳۰۰ است و اگر اینطور نباشد در آزمایش مورد نظر در ادامه ذکر شده است. V تعداد مفاصل یا همان گره‌های گراف است که با توجه به خروجی MediaPipe Pose برابر ۳۳ است. در نهایت، C تعداد کانال‌های ورودی یا همان بعد ویژگی‌های هر گره‌ی گراف است که در مساله‌ی ما می‌تواند مقدار ۲ یا ۳ بگیرد. در حالتی که (X, Y) را به عنوان ویژگی هر مفصل به مدل بدیم دو کانال داریم و در حالتی که (X, Y, visibility) را به عنوان ویزگی هر مفصل به مدل بدهیم سه کانال خواهیم داشت. در اکثر آزمایش‌ها از حالت سه کاناله استفاده کرده‌ایم و تنها در دو حالت است که برای بررسی تاثیر حذف امتیاز دیدگانی، حالت دو کاناله امتحان شده است.

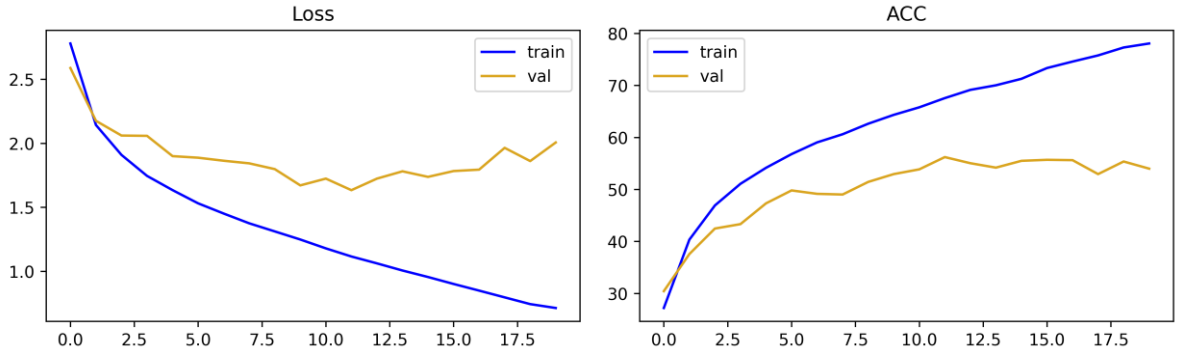
۴-۴-۱- بررسی نتایج سه مدل مبتنی بر کانولوشن گرافی

در این سه آزمایش از بهینه‌ساز ADAM استفاده کردیم و مقدار نرخ یادگیری برابر 0.003 قرار دادیم. داده‌های در پیش پردازش لایه‌گذاری شده‌اند، قاب‌های بدون اسکلت از آن‌ها حذف شده، و تعداد کل قاب ها ۳۰۰ است. همچنین در این مدل‌های کانولوشنی گرافی از روش تقسیم بندی همسایگان به صورت تک برچسب گذاری استفاده کردیم که بر تابع وزن‌دهی تاثیر می‌گذارد. بردار ویژگی ورودی در این آزمایش‌ها سه کاناله در نظر گرفته شده است و از حذف تصادفی در آموزش مدل‌ها استفاده نکردیم. در ادامه نمودارهای دقت و هزینه مربوط به سه مدل کانولوشنی گرافی ST-GCN، MST-GCN و شبکه کانولوشنی گرافی ساده سه لایه قابل مشاهده اند.

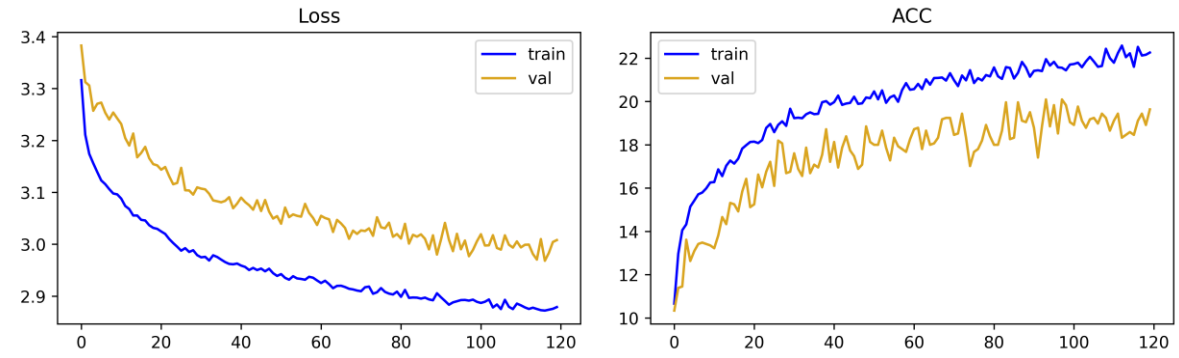


شکل ۴-۴ نمودارهای دقت و هزینه مدل GCN-ST

[12] Dropout



شکل ۴-۵ نمودارهای دقت و هزینه مدل GCN-MST



شکل ۴-۶ نمودارهای دقت و هزینه مدلGCN ساده سه لایه

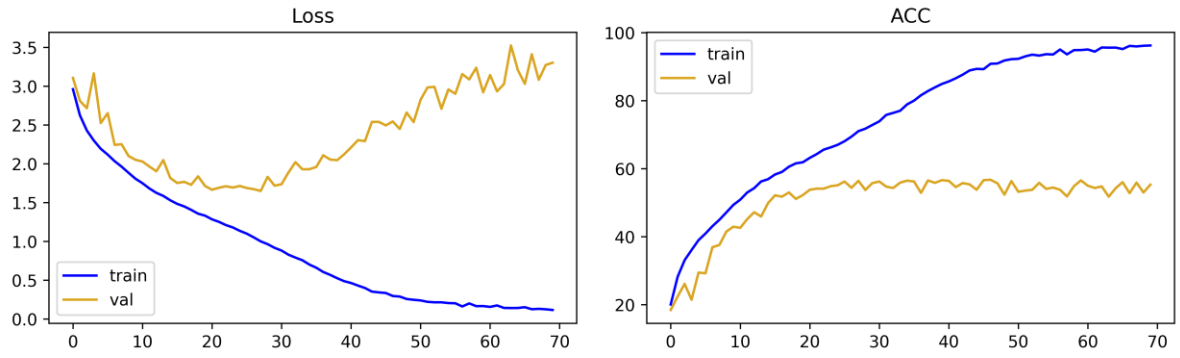
نتایج کمی آموزش و تست این سه مدل در جدول ۲ قابل مشاهده است. این نتایج را می‌توان اینطور تحلیل کرد که از نظر دقت، مدل MST-GCN از سایر مدل‌ها عملکرد بهتری دارد، اما زمان آموزش آن در هر ایپاک بسیار زیادتر از زمان آموزش دیگر مدل‌ها در هر ایپاک است. در واقع می‌توان گفت که با توجه به حجم محاسباتی و زمان آموزش زیاد، دقت را به میزان چشمگیری در مقایسه با مدل ST-GCN افزایش نداده است. طبق نتایج مدل GCN ساده می‌توان، می‌بینیم که این مدل تعداد ایپاک خیلی زیاد و زمان آموزش خیلی زیادی نیاز دارد تا به بالاترین دقت برسد و دقت نهایی آن با اختلاف معناداری کمتر از دقت دو مدل پیشرفته دیگر است.

جدول ۴-۲ مقایسه کمی مدل‌های کانولوشنی گرافی

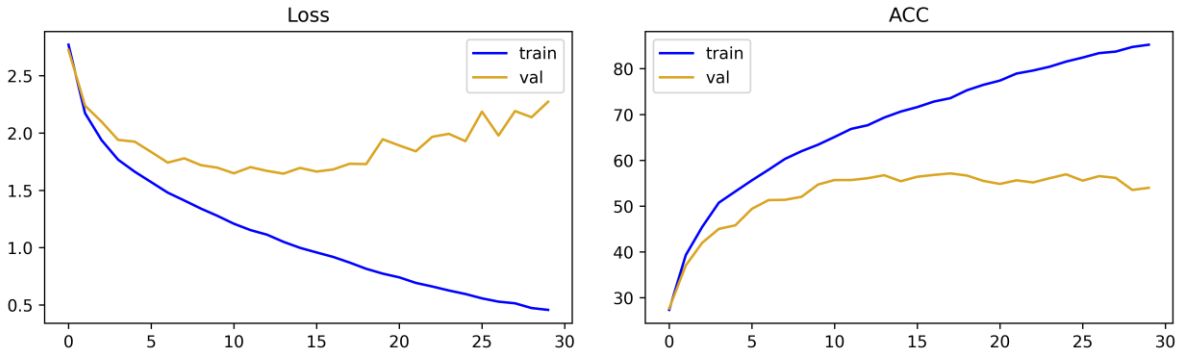
مدل	تعداد پارامترهای قابل یادگیری	زمان آموزش (ثانیه)	دقت (درصد)	دقت ۳ برترین (درصد)	تعداد ایپاک تا رسیدن به بهترین نتیجه
ST-GCN uniform	۲,۶۴۷,۹۹۰	۱۵۷۱/۶۵۵۲	۵۲/۹۵	۷۱/۳۶	۲۹
MST-GCN uniform	۳,۱۸۶,۸۴۳	۲۹۱۲/۴۲۷۶	۵۳/۶۲	۷۱/۸۶	۱۲
GCN	۲۱,۱۵۸	۳۴۳۴/۵۵۸۶	۲۱/۰۷	۳۵/۹۱	۹۳

۴-۴-۲- بررسی تغییر روش تقسیم بندی همسایگان

همانطور که در فصل قبل توضیح داده شد، روش تقسیم بندی همسایگان یک گره در تابع وزن شبکه‌ی کانولوشنی گرافی تاثیر به سزایی دارد. در آزمایش قبل از روش تک برچسب گذاری در آموزش هر دو مدل ST-GCN و MST-GCN استفاده کردیم. در این آزمایش برای مدل ST-GCN از روش برچسب زدن برحسب فاصله برای تقسیم بندی همسایگان استفاده می‌کنیم. همچنین برای مدل MST-GCN از روش برچسب زدن برحسب پیکربندی فضایی استفاده می‌کنیم. در این حالات به گره‌هایی که در فاصله‌های متفاوتی از گره‌ی مرکزی قرار دارند وزن‌های متفاوتی داده می‌شود؛ این سبب یادگیری اطلاعات بیشتری طی انجام عملیات کانولوشن گرافی می‌شود. در نتیجه، انتظار می‌رود در نتایج این آزمایش، دو مدل نسبت به آزمایش قبل به دقت بالاتری دست یابند. سایر شرایط آموزش، در این آزمایش‌ها نسبت به آزمایش‌های قسمت قبل تغییری نکرده است. نتایج این دو آزمایش در جدول ۳ گزارش و نمودارهای دقت و هزینه در مرحله آموزش در اشکال ۲۶ و ۲۷ نمایش داده شده‌اند.



شکل ۴-۷ نمودارهای دقت و هزینه مدل GCN-ST با روش برچسب زدن برحسب فاصله



شکل ۴-۸ نمودارهای دقت و هزینه مدل GCN-MST با روش برچسب زدن برحسب پیکربندی فضایی

با توجه به نتایج گزارش شده در جدول ۳، و مقایسه مقادیر دقت‌ها با نتایج گزارش شده در جدول ۲ که مربوط به آموزش این دو مدل با روش تقسیم بندی همسایگان به صورت تک برچسب گذاری بود. می‌توان مشاهده کرد که با بهبود روش تقسیم بندی همسایگان، دقت در مدل ST-GCN، ۱/۸۸ و دقت ۳ برترین، ۰/۹۲ درصد بهبود داشته است. همچنین در مدل MST-GCN، دقت ۱/۸۹ و دقت ۳ برترین، ۱/۱ درصد بهبود داشته است.

جدول ۴-۳ نتایج کمی دو شبکه‌ی GCN-ST و GCN-MST با بهبود در روش تقسیم بندی گره در گراف

مدل	تعداد پارامترهای قابل یادگیری	زمان آموزش (ثانیه)	دقت (درصد)	دقت ۳ برترین (درصد)	تعداد ایپاک تا رسیدن به بهترین نتیجه
ST-GCN distance	۲,۸۷۷,۵۶۸	۱۸۴۶/۶۵۴۵	۵۴/۸۳	۷۲/۲۸	۴۶
MST-GCN spatial	۳,۲۲۸,۸۴۳	۴۲۵۶/۹۸۸۱	۵۵/۵۱	۷۲/۹۶	۱۷

۴-۴-۳- بررسی تغییر حضور امتیاز پدیداری در ویژگی‌های ورودی

در آزمایش‌های پیشین از ورودی سه کاناله حاوی اطلاعات مختصات X، Y و امتیاز پدیداری هر مفصل در مراحل آموزش، اعتبارسنجی و تست استفاده کردیم. حال می‌خواهیم بررسی کنیم که حذف امتیاز پدیداری از بردار ویژگی ورودی شبکه چه تاثیری در دقت و عملکرد شبکه‌ها دارد. در این آزمایش دو شبکه‌ی ST-GCN distance وMST-GCN spatial را با داده ورودی دو کاناله در شرایط آموزش و با هایپرپارامترهایی مشابه با دو آزمایش قبل آموزش دادیم.

امتیاز پدیداری در واقع میزان اطمینان مدل تخمین حالت دو بعدی در ارتباط با اعداد تخمین زده شده برای مختصات مفاصل را نشان می‌دهد. بنابراین می‌توان گفت، این کمیت می‌تواند اطلاعات مفیدی راجع به هر مفصل در اختیار مدل تشخیص فعالیت مبتنی بر اسکلت قرار دهد. با توجه به این توضیحات، می‌توان انتظار داشت که با حذف این کمیت از بردار ویژگی ورودی، عملکرد مدل‌های تشخیص فعالیت کمی افت پیدا کند. همانطور که در جدول ۴ مشاهده می‌شود، شاهد مقدار کمی افت دقت در هر دو شبکه‌ی ST-GCN و MST-GCN هستیم.

جدول ۴-۴ نتایج کمی دو شبکه‌ی GCN-ST و GCN-MST با حذف امتیاز پدیداری از بردار ویژگی

مدل	دقت (درصد)	افت دقت (درصد)	دقت ۳ برترین (درصد)	افت دقت ۳ برترین (درصد)
ST-GCN distance, filter visibility score	۵۱/۰۴	۳/۷۹	۷۰/۵۷	۱/۷۱
MST-GCN spatial, filter visibility score	۵۴/۳۸	۱/۱۳	۷۲/۶۲	۰/۳۴

۴-۴-۴- بررسی تغییر هایپرپارامترها و نحوه پیش پردازش در مدل ST-GCN

در آزمایش‌های این بخش، برخی هایپرپارامترها و پیش پردازش‌های مختلف را در آموزش و تست مدل ST-GCN امتحان کردیم. با تحلیل نتایج موجود در جدول ۵، می‌توان تاثیر تغییر هر یک از هایپرپارامترها و پیش‌پردازش‌ها را مشاهده کرد. با توجه به نتایج عملکرد مدل در شرایط مختلف می‌توان نتیجه گرفت که در مدل ST-GCN، آموزش با بهینه‌ساز ADAM،

نرخ یادگیری ۰/۰۰۳ و بدون استفاده از حذف تصادفی به دقت بالاتر و عملکرد بهتری دست می‌یابد. راجع به تاثیر انجام پیش پردازش‌ها نیز می‌توان گفت که انجام لایه گذاری تاثیر مثبتی بر یادگیری بهتر این مدل دارد و انجام برش پنجره‌ای ۱۵۰ قابی از بین ۳۰۰ قاب، نه تنها به یادگیری بهتر این مدل کمکی نمی‌کند، بلکه باعث می‌شود مدل افت دقت داشته باشد.

جدول ۴-۵ نتایج کمی تغییر هایپرپارامترها و نحوه پیش پردازش در مدل GCN-ST

	انواع هایپرپارامترها			انواع پیش‌پردازش‌ها		نتایج (درصد)	
مدل	بهینه ساز	نرخ یادگیری	حذف تصادفی	لایه گذاری	برش پنجره‌ای	دقت	دقت ۳ برترین
ST-GCN	ADAM	۰/۰۰۳	×	✓	×	**۵۲/۹۵**	**۷۱/۳۶**
	SGD	۰/۰۰۳	×	✓	×	۴۴/۷۹	۶۲/۷۱
	SGD	۰/۰۰۱	×	✓	×	۴۴/۶۹	۶۲/۲۴
	SGD	۰/۰۱	×	✓	×	۴۳/۷۰	۶۲/۲۴
	ADAM	شروع از ۰/۱ و کاهش تصاعدی در هر ۱۰ ایپاک	×	✓	×	۴۱/۱۱	۶۰/۴۳
	ADAM	۰/۰۰۳	با احتمال ۰/۵	✓	×	۵۱/۲۸	۷۰/۱۳
	ADAM	۰/۰۰۳	×	✓	✓	۴۷/۲۹	۶۷/۰۵
	ADAM	۰/۰۰۳	×	×	×	۵۲/۶۸	۷۰/۹۱

۴-۴-۵- بررسی تغییر هایپرپارامترها و نحوه پیش پردازش در مدل MST-GCN

در آزمایش‌های این بخش نیر مانند بخش قبل برخی هایپرپارامترها و پیش پردازش‌های مختلف را در آموزش و تست مدل MST-GCN امتحان کردیم. اکثر آزمایش‌ها همانند بخش قبل است. با این تفاوت که تاثیر حذف تصادفی را در این مدل بررسی نکردیم زیرا در معماری مدل اصلی ماژولی برای اعمال حذف تصادفی در نظر گرفته نشده است. با تحلیل نتایج موجود در جدول ۶ می‌توان تاثیر تغییر هر یک از هایپرپارامترها و پیش‌پردازش‌ها را مشاهده کرد.

با توجه به نتایج عملکرد مدل در شرایط مختلف می‌توان نتیجه گرفت که در مدل MST-GCN، آموزش با بهینه‌ساز ADAM، نرخ یادگیری ۰/۰۰۳ به دقت بالاتر و عملکرد بهتری دست می‌یابد. راجع به تاثیر انجام پیش پردازش‌ها نیز می‌توان گفت که انجام لایه گذاری برش پنجره‌ای ۱۵۰ قابی از بین ۳۰۰ قاب، نه تنها به یادگیری بهتر این مدل کمکی نمی‌کند و در

هر دو حالت باعث می‌شود مدل افت دقت و عملکرد داشته باشد. در حالتی که این دو نوع پیش پردازش انجام نشدند به بهترین نتیجه برای این مدل دست یافتیم.

جدول ۴-۶ نتایج کمی تغییر هایپرپارامترها و نحوه پیش پردازش در مدل GCN-MST

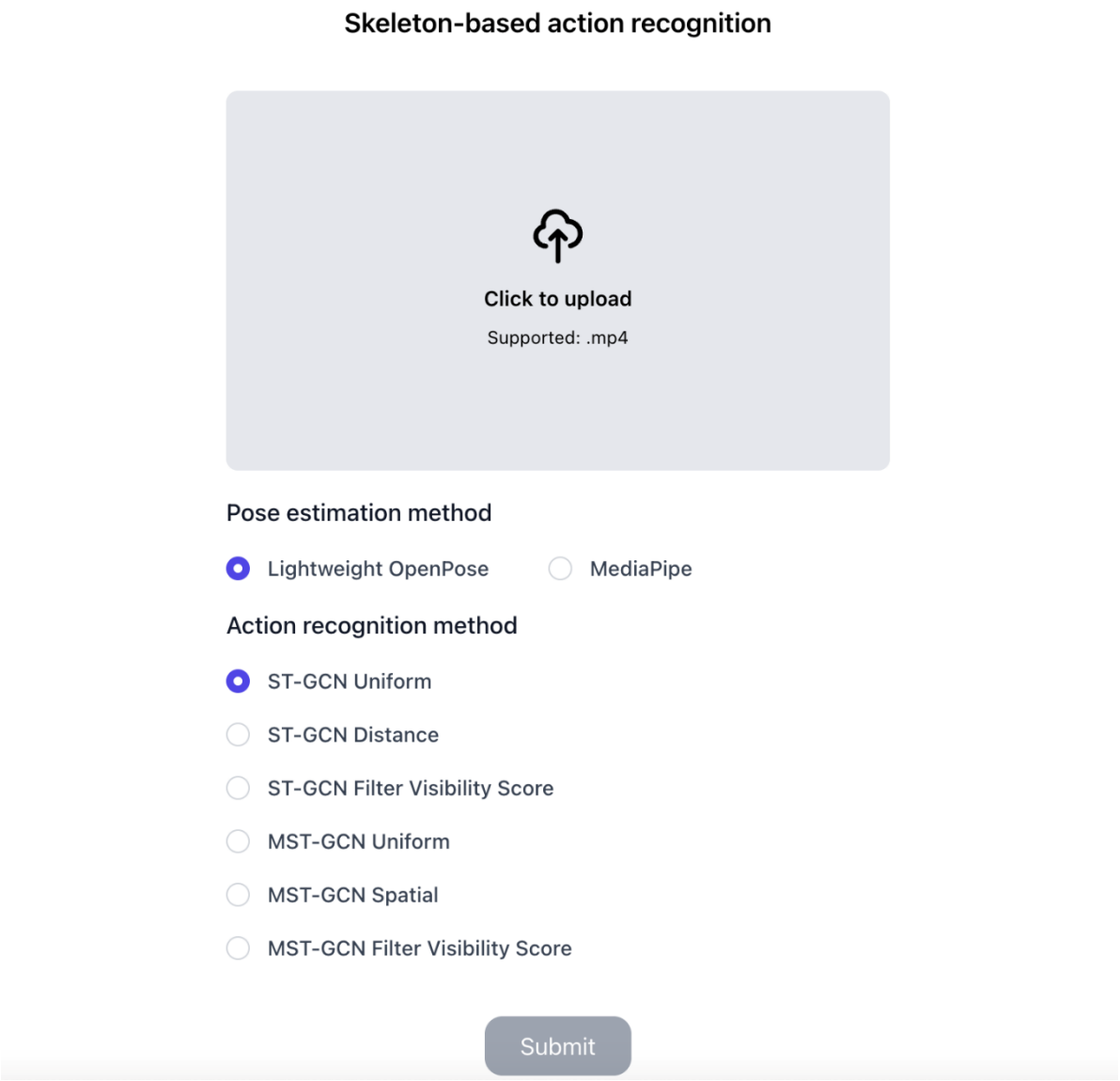
	انواع هایپرپارامترها		انواع پیش‌پردازش‌ها		نتایج (درصد)	
مدل	بهینه ساز	نرخ یادگیری	لایه گذاری	برش پنجره‌ای	دقت	دقت ۳ برترین
MST-GCN	ADAM	۰/۰۰۳	✓	×	۵۵/۵۱	۷۲/۹۶
	SGD	۰/۰۰۳	✓	×	۵۳/۶۰	۶۹/۹۲
	ADAM	شروع از ۰/۱ و کاهش تصاعدی در هر ۵ ایپاک	✓	×	۵۰/۵۳	۶۸/۴۹
	ADAM	۰/۰۰۳	✓	✓	۴۹/۱۲	۶۸/۱۸
	ADAM	۰/۰۰۳	×	×	**۵۶/۲۰**	**۷۴/۳۳**

۴-۵- رابط کاربری

تا این قسمت، با نحوه عملکرد مدل‌ها و تاثیر تغییر پارامترهای مختلف در عملکرد آن‌ها آشنا شدیم. در این بخش، نحوه عملکرد و پیاده‌سازی رابط کاربری تحت وب توسعه داده شده برای دو مدل تخمین حالت بدن و دو مدل تشخیص فعالیت انسان مبتنی بر اسکلت را شرح می‌دهیم. هدف ما از توسعه رابط کاربری این است که کاربر بتواند یک ویدیو کوتاه به عنوان ورودی در سامانه آپلود کند و به عنوان خروجی، سامانه هم اسکلت فرد و هم فعالیت انجام شده در ویدیو را تشخیص دهد.

رابط کاربری پیاده سازی شده برای این پروژه از طریق این لینک قابل دسترسی است. صفحه‌ی اصلی این رابط کاربری به کاربر این امکان را می‌دهد تا بتواند یک ویدیوی دلخواه را آپلود کند و سپس، از بین مدل‌های موجود یک مدل برای تخمین حالت بدن و یک مدل برای تشخیص فعالیت انسان انتخاب کند. در این رابط کاربری دو مدل Lightweight OpenPose و MediaPipe Pose را به عنوان مدل‌ها تخمین حالت پیاده‌سازی کردیم. برای مدل‌های تشخیص فعالیت سه حالت مختلف در دو مدل ST-GCN و MST-GCN را برای هر مدل پیاده سازی کردیم. حالت اول پیاده سازی معماری این دو مدل با روش تقسیم بندی همسایگان به صورت تک برچسب گذاری (uniform) است. حالت دوم پیاده سازی معماری مدل ST-GCN با روش برچسب زدن برحسب فاصله و مدل MST-GCN با روش برچسب زدن برحسب پیکربندی فضایی

استفاده است؛ حالت سوم حذف کانال حاوی اطلاعات امتیاز پدیداری و پیاده سازی این دو مدل به صورت دو کاناله است. پس از انتخاب روش و ارسال ویدیو از سمت کاربر، رابط کاربری همان ویدیو به همراه اسکلت تخمین زده شده توسط مدل تخمین حالت بدن را خروجی می‌دهد. و پس از طی شدن مراحل استنتاج مدل تشخیص فعالیت انسان، فعالیت تشخیص داده شده را نیز به کاربر نمایش می‌دهد. شکل ۴-۹ نمای کلی رابط کاربری پیاده‌سازی شده را نشان می‌دهد.



شکل ۴-۹ صفحه‌ی اصلی رابط کاربری پیاده‌سازی شده

۴-۵-۱- جزییات پیاده‌سازی رابط کاربری

بخش عقبی[۱۳] رابط کاربری با استفاده از زبان جاوا اسکریپت و پایتون و بخش جلویی[۱۴] با استفاده از چهارچوب ری‌اکت توسعه داده شده است. با استفاده از جاوااسکریپت در توسعه این رابط کاربری توانسته‌ایم یک سیستم چند هسته‌ای[۱۵] و چند نخی[۱۶] توسعه دهیم. اتصال و پیام‌رسانی بین این دو بخش، با استفاده از وب‌سوکت و polling صورت می‌گیرد. پس از آپلود ویدیو و انتخاب گزینه‌ها توسط کاربر، ویدیو بر روی سرور آپلود می‌شود. پس از آن، نحوه ارتباط این بخش جلویی و عقبی پس از درخواست کاربر به این صورت است که بخش جلویی موارد انتخاب شده‌ی کاربر را در قاب یک FormData به بخش عقبی ارسال می‌کند. بخش عقبی پس از دریافت درخواست از سمت بخش جلویی، یک پردازش[۱۷] برای استخراج اسکلت حالت بدن را شروع می‌کند. پس از استخراج اسکلت توسط مدل تخمین حالت بدن در بخش عقبی، نتایج در قاب رشته حاوی آدرس URLای که ویدیوی حاوی اسکلت در آن قرار دارد به صورت یک رخداد[۱۸] در وب‌سوکت به بخش جلویی فرستاده می‌شود. در همین حین، بخش عقبی یک پردازش دیگر برای اجرای مدل تشخیص فعالیت روی داده‌ی اسکلتی را شروع می‌کند و پس از اتمام فرآیند استنتاج توسط مدل تشخیص فعالیت، نام فعالیت تشخیص داده شده نیز به صورت یک رخداد در وب‌سوکت در قاب یک رشته به بخش جلویی فرستاده می‌شود.

۴-۶- خلاصه

در این فصل به معرفی مجموعه دادگانی استفاده شده در انجام این پروژه پرداختیم. سپس جزییات راه‌اندازی دو مدل تخمین حالت بدن MediaPipe Pose و Lightweight OpenPose را توضیح دادیم. در قسمت اصلی آزمایش‌های انجام شده بر روی مدل‌های تشخیص فعالیت مبتنی بر اسکلت را شرح دادیم و نتایج آن‌ها را تحلیل کردیم. از مهم‌ترین قسمت‌های این فصل، مقایسه عملکرد سه مدل مبتنی بر شبکه‌های کانولوشنی گرافی ST-GCN، MST-GCN و GCN سه لایه ساده بود؛ در این مقایسه مشاهده شد که عملکرد مدل MST-GCN از لحاظ دقت بهتر از ST-GCN بوده و هر دو این مدل‌ها عملکرد به مراتب بهتری نسبت به یک شبکه‌ی ساده‌ی سه لایه کانولوشنی گرافی داشتند. تاثیر مقادیر مختلف هایپر پارامترها و پیش پردازش‌های مختلف بر این دو مدل نیز در آزمایش‌ها بررسی و گزارش شد. در نهایت رابط کاربری تحت وب توسعه داده شده در این پروژه را همراه با جزییات پیاده سازی آن شرح دادیم.

---

[13] Backend

[14] Frontend

[15] Multi-core

[16] Multi-thread

[17] Process

[18] Event

فصل پنجم

نتیجه‌گیری و پیشنهادات

۵-۱- جمع‌بندی و نتیجه‌گیری

مساله‌ای که در این پروژه به آن پرداختیم، تشخیص فعالیت انسان مبتنی بر اسکلت در دادگان ویدئویی، با تمرکز بر استفاده از شبکه‌های کانولوشنی گرافی است. این کار شامل شناسایی دقیق اسکلت افراد توسط مدل‌های تخمین حالت بدن و سپس طبقه‌بندی فعالیت‌های مختلف انجام شده توسط افراد بر اساس اطلاعات توالی اسکلتی به دست آمده است. چالش اصلی در این زمینه توانایی تشخیص قوی و کارآمد اقدامات با در نظر گرفتن عواملی مانند انسداد، نویز، تغییرات در وضعیت انسان و تغییرات دیدگاه است. هدف نهایی این پروژه، پیاده‌سازی یک سیستم تشخیص فعالیت انسان با استفاده از اسکلت دو بعدی تولید شده از یک مدل تخمین حالت بدن بود.

برای رسیدن به این هدف، در فصل ابتدایی به ارائه مقدمه‌ای بر توضیح این مساله، اهمیت و معرفی چالش‌های موجود پرداختیم. همچنین رویکردهای مختلف شبکه‌های عصبی عمیق برای تشخیص فعالیت مبتنی بر اسکلت را معرفی کردیم و از اهمیت روش‌های تخمین حالت بدن و تاثیر آن در عملکرد مدل تشخیص فعالیت گفتیم. در فصل دوم به بررسی رویکردها و الگوریتم‌های موجود برای تخمین حالت بدن پرداختیم و دو روش lightweight OpenPose و MediaPipe Pose را با جزئیات بررسی کردیم. در فصل سوم جزییات معماری و نحوه عملکرد شبکه‌های کانولوشنی گرافی پیاده سازی شده در این پروژه را توضیح دادیم. در نهایت در فصل چهارم به جزئیات پیاده‌سازی، مراحل پیش‌پردازش، آموزش و ارزیابی شبکه‌های کانولوشنی گرافی با پارامترهای مختلف پرداختیم و معیارهای ارزیابی را برای هر کدام محاسبه کردیم. همچنین به بررسی رابط کاربری گرافیکی طراحی شده و نحوه کار با آن نیز پرداختیم.

در این پژوهش به بررسی و پیاده‌سازی دو نوع شبکه کانوولوشنی گرافی با نام‌های ST-GCN و MST-GCN پرداختیم و این دو شبکه را با استفاده از توالی اسکلتی بدست آمده از مدل MediaPipe Pose بر زیرمجموعه‌ای مجموعه دادگان Kinetics 400 آموزش دادیم. در آموزش و ارزیابی هر کدام از مدل‌ها به بررسی تاثیر استفاده از توابع فعالسازی، پیش پردازش‌ها و هایپرپارامترهای مختلف پرداختیم. طبق ارزیابی‌های انجام شده، عملکرد مدل MST-GCN با دستیابی به دقت ۵۳/۶۲، بهتر از مدل ST-GCN بود؛ باید توجه داشت که این مدل به منابع و زمانی بیشتری نیز برای آموزش نیاز داشت. طبق آزمایش‌های انجام شده، تغییر روش تقسیم بندی همسایگان به روش‌هایی مثل تقسیم‌بندی بر اساس فاصله یا بر اساس پیکربندی فضایی نهایت در عملکرد این دو مدل تاثیر مثبتی دارد. این اتفاق به این دلیل است که در این روش‌ها تابع وزن دهی بین گره‌های مختلف تمایز بیشتری می‌گذارد و مدل جزئیات و اطلاعات بیشتری را یاد می‌گیرد. همچنین در نتایج آزمایش‌ها مشاهده شد که حذف امتیاز پدیداری از ویژگی‌های هر مفصل در توالی اسکلتی، باعث کاهش دقت تشخیص مدل می‌شود.

در نهایت از هر نوع شبکه‌ی کانولوشنی گرافی بهترین مدل را ذخیره کرده تا در رابط کاربری از آن استفاده کنیم. در رابط کاربری، کاربر ویدیوی مورد نظر خود را بارگذاری کرده و مدل‌های تخمین حالت بدن و تشخیص فعالیت مد نظرش را

انتخاب می‌کند. پردازش‌های لازم روی داده ویدیویی انجام شده، سپس به ویدیو همراه با اسکلت تخمین زده شده و برچسب تشخیص داده شده توسط مدل کانولوشنی گرافی را به کاربر نمایش می‌دهیم.

۵-۲- پیشنهادات

در این بخش به ایده‌هایی می‌پردازیم که می‌توانند موضوع این پروژه را در آینده گسترش دهد و شاهد کارهایی باشیم که به نتایج بسیار بهتر و کاربردی‌تری برسند. یکی از اقداماتی که می‌توان برای بهبود این پروژه انجام داد این است که از توالی اسکلتی سه بعدی برای آموزش مدل‌های تشخیص فعالیت انسان استفاده شود. در بسیاری از فعالیت‌های انسان و همچنین در برخی زوایا و دیدگاه‌های خاص در تصویر مقدار بعد سوم هر مفصل، حاوی اطلاعات فراوانی است که مدل تشخیص فعالیت می‌تواند با یادگیری این اطلاعات تشخیص دقیق‌تر داشته باشد.

همچنین می‌توان از مدل‌های دیگری برای قسمت تخمین حالت بدن انسان و بدست آوردن توالی اسکلتی استفاده کرد. در این پژوهش تنها از یکی از این مدل‌ها برای آموزش شبکه‌های کانولوشنی گرافی استفاده کردیم. در کارهای آینده، می‌توان از الگوریتم‌های تخمین حالت بدن مختلف دیگری استفاده کرد و بررسی شود که کدام الگوریتم برای آموزش یک شبکه‌ی عصبی کانولوشنی گرافی در حل مساله‌ی تشخیص فعالیت انسان مبتنی بر اسکلت بهتر عمل می‌کند.

علاوه بر آن، می‌توان رویکردهای مختلف یادگیری عمیق در مساله‌ی تشخیص فعالیت انسان را با یکدیگر ترکیب کرد و دقت و عملکرد مدل‌های حاصل را بررسی نمود. هر یک از رویکردهای یادگیری عمیق که در مساله‌ی تشخیص فعالیت انسان استفاده می‌شوند نقاط قوت و ضعف مختلفی دارند و می‌توان با ترکیب این مدل‌ها به دقت و عملکرد بهتری دست یافت.

منابع و مراجع

[1] Duan, Haodong, et al. "Revisiting skeleton-based action recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[2] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

[3] Wang, Wei, and Yu-Dong Zhang. "A short survey on deep learning for skeleton-based action recognition." *Proceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing Companion*. 2021.

[4] Ren, Bin, et al. "A survey on 3d skeleton-based action recognition using learning method." *arXiv preprint arXiv:2002.05907* (2020).

[5] Al-Faris, Mahmoud, et al. "A review on computer vision-based methods for human action recognition." *Journal of imaging* 6.6 (2020): 46.

[6] Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

[7] Osokin, Daniil. "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose." *arXiv preprint arXiv:1811.12004* (2018).

[8] Bazarevsky, Valentin, et al. "Blazepose: On-device real-time body pose tracking." *arXiv preprint arXiv:2006.10204* (2020).

[9] Chen, Xianjie, and Alan L. Yuille. "Parsing occluded people by flexible compositions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[10] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[11] Zheng, Ce, et al. "Deep learning-based human pose estimation: A survey." *arXiv preprint arXiv:2012.13392* (2020).

[12] Lin, Jintao, et al. "Ocsampler: Compressing videos to one clip with single-step sampling." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[13] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[14] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[15] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer International Publishing, 2016.

[16] X. Ying, "An Overview of Overfitting and its Solutions," *J. Phys.: Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[17] Fang, Hao-Shu, et al. "Rmpe: Regional multi-person pose estimation." *Proceedings of the IEEE international conference on computer vision*. 2017.

[18] Huu, Phat Nguyen, Ngoc Nguyen Thi, and Thien Pham Ngoc. "Proposing posture recognition system combining MobilenetV2 and LSTM for medical surveillance." *IEEE Access* 10 (2021): 1839-1849.

[19] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[20] MediaPipe. https://github.com/google/mediapipe. Accessed: April 2023.

[21] Chung, Jen-Li, Lee-Yeng Ong, and Meng-Chew Leow. "Comparative Analysis of Skeleton-Based Human Pose Estimation." *Future Internet* 14.12 (2022): 380.

[22] Vishnu, J. G., and S. J. Divya. "A Comparative Study of Human Pose Estimation."

[23] Wu, Zonghan, et al. "A comprehensive survey on graph neural networks." *IEEE transactions on neural networks and learning systems* 32.1 (2020): 4-24.

[24] Fang, Zheng, et al. "Spatial-temporal slowfast graph convolutional network for skeleton-based action recognition." *IET Computer Vision* 16.3 (2022): 205-217.

[25] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* (2016).

[26] Henaff, Mikael, Joan Bruna, and Yann LeCun. "Deep convolutional networks on graph-structured data." *arXiv preprint arXiv:1506.05163* (2015).

[27] Shi, Lei, et al. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[28] Liu, Ziyu, et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[29] Chen, Zhan, et al. "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 2. 2021.

[30] Two Stream ST-GCN. https://github.com/littlepure2333/2s_st-gcn. Accessed: Jun 2023.

[31] Peng, Wei, et al. "Learning graph convolutional network for skeleton-based human action recognition by neural searching." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 03. 2020.

[32] Kuehne, Hildegard, et al. "HMDB: a large video database for human motion recognition." *2011 International conference on computer vision*. IEEE, 2011.

[33] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." *arXiv preprint arXiv:1212.0402* (2012).

[34] Liu, Meng, et al. "DIG: A turnkey library for diving into graph deep learning research." *The Journal of Machine Learning Research* 22.1 (2021): 10873-10881.

Abstract

The problem of human action recognition seeks to create algorithms, methods and frameworks for automatically identifying the actions performed by a human in a video. Human action recognition technology can be widely used in medical diagnosis, crime rate control, patient or elderly monitoring and other industries. In recent years, skeleton-based action recognition methods that detect actions from an input sequence of skeletal joints have attracted a lot of attention in the research community. Skeleton-based representation extracted from human motions in a video, conveys significant information. Moreover, skeletal data are compact; hence, significantly reduce the computational cost in the action recognition problem. Reducing computational cost, availability of skeletal data, and recent improvements in body pose estimation algorithms have made current skeleton-based action recognition methods popular.

In this project, two algorithms, one for 2D pose estimation (obtaining the skeletal sequence) and the other for skeleton-based action recognition have been implemented. At first, the skeletal sequence was extracted on the kinetics400 dataset with the help of 2D pose etimation models. Then, two deep learning models based on convolutional networks have been implemented and trained with the help of previously extracted skeletal data. At last, the trained models are compared in terms of performance (accuracy and cost). A web-based system is also implemented as a user interface; so that, users can upload a video as input and receive the estimated skeleton and detected activity as output.

Keywords: Skeleton-based human action recognition, Pose estimation, Graph convolutional neural networks

Amirkabir University of Technology
(Tehran Polytechnic)

Computer Engineering

BSc Thesis

Skeleton-based human action recognition with the help of 2D pose estimation

By
Hedieh Pourghasem

Supervisor
Dr. Mohammad Rahmati

June 2023