مقاله پژوهشی

استخراج ویژگی‌های عمیق بلندمدت برای طبقه‌بندی ویدیو

عباس همدونی اصلی، شیما جاویدانی و علی جاویدانی

چکیده: در این مقاله، رویکردی نوین برای شناسایی کنش های در حال انجام از ویدیوهای بخش‌بندی‌شده ارائه می‌شود. تمرکز اصلی بر استخراج ویژگی‌های بلندمدت از ویدیوها به منظور طبقه‌بندی موثر آنها است. بدین منظور، ابتدا تصاویر جریان نوری میان فریم‌های متوالی محاسبه و با یک شبکه عصبی کانولوشنی از پیش آموزش‌دیده توصیف می‌شوند. برای کاهش پیچیدگی فضای ویژگی و ساده‌سازی یادگیری مدل زمانی، کاهش بعد PCA بر روی بردارهای توصیفی جریان نوری اعمال می‌گردد. سپس به منظور پالایش ورودی، یک ماژول توجه کانالی سبک وزن بر بردارهای کم بعد حاصل از PCA در هر گام زمانی اعمال می‌شود تا مولفه‌های اطلاعاتی تقویت و مولفه‌های کم اثر تضعیف شوند. در ادامه، توصیفگرهای هر ویدیو هم‌راستا شده و در راستای زمان دنبال می‌شوند و استخراج ویژگی‌های بلندمدت با آموزش یک شبکه LSTM دو لایه پشته‌ای انجام می‌پذیرد. پس از LSTM، یک ماژول توجه زمانی به عنوان تجمیع آگاه به زمان به کار گرفته می‌شود تا با وزن دهی داده محور به گام‌های زمانی، لحظات اطلاع‌رسان را برجسته کرده و یک بردار منسجم برای طبقه‌بندی بسازد. نتایج تجربی نشان می‌دهد که ترکیب PCA به همراه توجه کانالی و توجه زمانی ضمن حفظ سبک وزنی مدل، دقت طبقه‌بندی را در هر دو مجموعه داده عمومی UCF۱۱ و jHMDB بهبود می‌بخشد و عملکرد بهتری نسبت به روش‌های مرجع ارائه می‌کند. کد مورد استفاده در این مقاله، به صورت دسترسی باز قابل در دسترس‌است.

کلیدواژه: **طبقه‌بندی ویدیو، شناسایی کنش انسانی، یادگیری عمیق، شبکه‌های عصبی کانولوشنی، شبکه‌های عصبی بازگشتی، حافظه‌ی بلند و کوتاه‌مدت (LSTM).**

۱- مقدمه

تحلیل ویدیو یکی از حوزه‌های پژوهشی در زمینه چندرسانه‌ای است که اخیرا دچار تحولات چشمگیری شده است [۱] و [۲]. در سال‌های اخیر، پژوهش‌های بسیاری با هدف شناسایی فعالیت‌ها و رویدادهای جاری در ویدیوها انجام شده‌اند [۳] و [۴]. شبکه‌های عصبی کانولوشنی[1] (CNN) با یادگیری ویژگی‌ها و بهره‌گیری از مجموعه‌داده‌های بزرگ، در بسیاری از کاربردهای حوزه بینایی ماشین توانسته‌اند نتایجی برتر نسبت به توصیف‌گرهای دستی به‌دست آورند. با این حال، با وجود انجام آزمایش‌های متعدد بر روی ویدیوها نسبت به سایر زمینه‌ها موفق نبوده‌اند. این موضوع می‌تواند ناشی از ویژگی‌های ذاتی ویدیوها نسبت به تصاویر باشد. برای نمونه، تفاوت در طول ویدیوها یک چالش مهم محسوب می‌شود. زیرا یک کنش مشخص ممکن است در ویدیوهایی با طول‌های متفاوت ظاهر شود. عوامل دیگر مانند تغییرات شدید در حرکات دوربین و حضور اجسام متحرک متعدد در پس‌زمینه نیز می‌توانند نقش داشته باشند. به دلیل این گونه تغییرات، دستیابی به مدلی مقاوم برای طبقه‌بندی فعالیت‌های مختلف در ویدیوها نسبتاً دشوار به نظر می‌رسد.

این مقاله روشی نوین برای طبقه‌بندی ویدیوهایی با طول‌های متفاوت معرفی می‌کند. ایده‌ی اصلی، بهره‌گیری از وابستگی‌های بلندمدت موجود در سری‌های زمانی است که نمایانگر حرکت‌های موجود در ویدیو هستند. این در حالی است که با کاهش تعداد کانال‌ها، فضای ویژگی‌ها نیز به شکلی مؤثر حفظ می‌شود. برای تحقق این هدف، الگوریتم مؤثر کاهش بُعد تحلیل مولفه‌های اصلی[2] (PCA) به‌کار گرفته می‌شود تا فضای ویژگی فریم‌محور به فضایی با ابعاد کمتر ولی با حفظ انرژی کافی از فضای اصلی تبدیل شود. به‌منظور یافتن ویژگی‌های بلندمدت، از شبکه‌های[3] LSTM برای یادگیری الگوها و طبقه‌بندی سری‌های زمانی حاصل‌شده استفاده می‌شود. با پیاده‌سازی روش پیشنهادی بر روی دو مجموعه‌داده عمومی UCF۱۱ و jHMDB نشان داده شده است که می‌توان به نتایجی در سطح روش‌های پیشرفته موجود دست یافت. نمای کلی خط لوله‌ی[4] روش پیشنهادی در شکل ۱ قابل مشاهده است.

ساختار ادامه مقاله به‌صورت زیر تنظیم شده است: در بخش ۲، پژوهش‌های پیشین مرتبط با این حوزه و تحلیل آن‌ها ارائه می‌گردد. در بخش ۳، روش پیشنهادی و تئوری پشت آن به‌صورت کامل توضیح داده می‌شود. جزئیات پیاده‌سازی و ارزیابی روش ما در مقایسه با سایر رویکردها نیز در بخش ۴ مورد بحث قرار می‌گیرند.

۲- پژوهش‌های مرتبط

به‌طور کلی، کلید تحلیل ویدیوها استخراج ویژگی‌های مؤثر فضازمانی[5] است. شبکه‌های عصبی کانولوشنی (CNN) برای یادگیری ویژگی‌های فضایی[6] متمایزتر نسبت به توصیف‌گرهای دستی طراحی شده‌اند و به همین دلیل توانسته‌اند نتایج بهتری به‌دست آورند. با این حال، استفاده از آن‌ها برای ویدیوها و درنظرگرفتن بُعد زمانی در شبکه‌های کانولوشنی دوبعدی می‌تواند چالش‌برانگیز باشد [۵] تا [۸].

این مقاله در تاریخ ۱۵ تیر ماه ۱۴۰۴ دریافت و در تاریخ ۲ مهر ماه ۱۴۰۴ بازنگری شد.

عباس همدونی اصلی، موسسه آموزش عالی جهاد دانشگاهی همدان، همدان، ایران، (email: abhamedoni1653@gmail.com).

شیما جاویدانی، موسسه آموزش عالی جهاد دانشگاهی همدان، همدان، ایران، (email: shjavidani1059@gmail.com) .

علی جاویدانی (نویسنده مسئول)، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه بوعلی سینا، همدان، ایران، (email: alijavidani@ut.ac.ir).

1. Convolutional Neural Network
2. Principal Component Analysis
3. Long Short-Term Memory
4. Pipeline
5. Spatio-Temporal
6. Spatial

شکل ۱: خط لوله‌ی کلی روش پیشنهادی.

پژوهش‌های متعددی روش‌های مختلف ترکیب اطلاعات[1] را برای استخراج ویژگی‌های زمانی از ویدیو بررسی کرده‌اند. برای نمونه، کارپاتی و همکاران چهار روش ترکیب معرفی کردند که در آن‌ها با تغذیه‌ی چند فریم از ویدیو به یک CNN چندکاناله، تلاش کردند روابط زمانی بین فریم‌ها را شناسایی کنند [۸]. همچنین نگ و همکاران دو روش تجمیع دیگر به این کار افزودند [۹]. نتایج نشان می‌دهد که به دلیل تعداد زیاد پارامترهایی که باید در مدل‌های آن‌ها آموخته شوند، معماری‌های پیشنهادی‌شان نتوانستند ویژگی‌های فضازمانی مؤثری را استخراج کنند. زیسرمن و همکارانش یک شبکه‌ی دومسیره پیشنهاد دادند که تلاش داشت ویژگی‌های فضایی و زمانی را به‌طور مستقل، مشابه سیستم بینایی انسان، شناسایی کند [۶]. آن‌ها در ادامه، مدل پایه‌ی خود را با بررسی روش‌های ترکیب مختلف در بخش‌های متفاوت شبکه‌ی دومسیره بهبود دادند [۱۰]. یکی از چالش‌های آن‌ها کمبود داده برای آموزش مدل بود. برای غلبه بر این مشکل، آن‌ها تعداد فریم‌های ورودی به مدل را محدود کردند که یکی از نقاط ضعف عمده‌ی روششان محسوب می‌شود؛ زیرا نتوانستند کل ویدیو را به‌طور کامل تحلیل کنند. به‌طور مشابه، در پژوهشی دیگر، اطلاعات ظاهری و حرکتی به‌صورت جداگانه استخراج شده و در نهایت دو توزیع احتمالاتی برای پیش‌بینی برچسب ویدیو تلفیق شدند [۱۱].

برای اینکه شبکه‌ی CNN بتواند اطلاعات زمانی را به‌صورت مؤثرتری استخراج کند، عملیات کانولوشن و تجمیع دوبعدی به سه‌بعدی گسترش یافت [۱۲] و [۱۳]. با این حال، عملگرهای کانولوشن سه‌بعدی نیز برای آموزش مناسب به مجموعه‌داده‌های بزرگ نیاز دارند. به‌دلیل پیچیدگی و دشواری در آموزش شبکه‌های C۳D، این مدل حتی بر روی مجموعه‌داده‌های کم‌چالش نیز عملکرد مناسبی ندارد. همچنین از شبکه‌های FCN[2] برای شناسایی کنش‌های انسان نیز استفاده شده است [۱۴]. اکثر روش‌های یادشده نمی‌توانند به‌صورت آنی[3] اجرا شوند؛ چرا که به محاسبات بسیار زیادی نیاز دارند. یکی از سیستم‌های شناسایی کنش آنی توسط لیو و همکاران طراحی شد [۱۵]. همچنین، در روشی دیگر [۱۶] با جایگزین‌کردن بردارهای حرکتی به‌جای جریان نوری و بهره‌گیری از کارت‌های گرافیکی قدرتمند، توانستند بدون افت دقت، عملکردی بلادرنگ ارائه دهند.

در سال‌های اخیر، شبکه‌های عصبی بازگشتی[4] (RNN) توجه پژوهشگران را به خود جلب کرده‌اند [۱۷] تا [۲۱] . از آن‌جا که شبکه‌های LSTM با مشکل ناپدیدشدن یا انفجار گرادیان مواجه نمی‌شوند، در این زمینه مورد آزمایش‌های متعددی قرار گرفته‌اند. شبکه‌ی عمیقی که در پژوهش [۹] معرفی شده، از پنج لایه‌ی پشته‌ای LSTM برای یادگیری ویژگی‌های زمانی مؤثر در طبقه‌بندی ویدیوهای سوم‌شخص بهره می‌برد. همچنین، فیلترهای توجه برای تمرکز بر بخش‌های مختلف ویدیو در راستای بُعد زمانی سودمند هستند. پیرجیوانی و همکاران [۲۲] و لیو و همکاران [۱۶] با کمک LSTMها تلاش کردند تا فیلترهای توجه زمانی را برای ویدیوهای اول‌شخص و سوم‌شخص بیاموزند و نتایجی برتر نسبت به ویژگی‌های ازپیش‌تعریف‌شده کسب کردند. علاوه‌براین، با توجه به نتایج امیدوارکننده شبکه‌های کانولوشنی باقی‌مانده[5] نظیر ResNet-۱۵۲، ، شبکه‌های بازگشتی باقی‌مانده نیز در این حوزه مورد استفاده قرار گرفته‌اند [۲۱] و توانسته‌اند نتایج بهتری نسبت به روش‌های سنتی ارائه دهند. در یکی از پژوهش‌های اخیر، چارچوبی مبتنی بر ترکیب عمیق چندمسیره معرفی شده است که با دریافت لایه‌های مختلف از نگاشت ویژگی‌های یادگرفته‌شده از CNN های ازپیش‌آموزش‌دیده و تغذیه‌ی آن‌ها به LSTM، توانسته نتایج پیشرفته‌ای بر روی مجموعه‌داده‌های کوچک به‌دست آورد [۲۳].

در یک پژوهش اخیر در حوزه‌ی نظارت تصویری صنعتی، اولا و همکاران [۲۴]. ابتدا فریم‌های مهم را با بهره‌گیری از ویژگی‌های برجسته‌ی مبتنی بر CNN انتخاب می‌کند. سپس ویژگی‌های زمانیِ جریان نوری را به‌کمک لایه‌های کانولوشنیِ FlowNet۲ استخراج و در نهایت با یک LSTM چندلایه مدل‌سازی می‌کند. در خط پژوهشی دیگری، جن و همکاران [۲۵] چارچوبی فشرده برای بازشناسی کنش معرفی می‌کنند که در آن همجوشی دووجهیِ RGB و جریان نوری در سطح ویژگی‌های زمانی انجام می‌شود و سپس LSTM با تجزیه‌ی تنسوری فشرده‌سازی می‌گردد تا هزینه‌ی محاسباتی و تعداد پارامترها کاهش یابد. تمرکز اصلی [۲۵] بر فشرده‌سازی وزن‌های LSTM و پیاده‌سازی کارآمد بر روی دستگاه نهایی است.

هر چند [۲۴] و [۲۵] نیز بر جریان نوری و LSTM تکیه دارند، تفاوت‌های کلیدی وجود دارد: در [۲۴] مرحله‌ی انتخاب فریم و استخراج ویژگی مبتنی بر جریان نوری انجام می‌شود اما مکانیزم‌های توجه صریح به‌کار نرفته است. در [۲۵] تمرکز بر فشرده‌سازی وزن‌های LSTM و همجوشی RGB+Flow است. در مقابل، رویکرد حاضر تک‌وجهی و مبتنی بر جریان نوری بوده و فشرده‌سازی را پیش از مدل زمانی و در ورودی LSTM با PCA انجام می‌دهد (کاهش بُعد از ۱۰۲۴ به $d \approx ۵۰$)، سپس با افزودن دو ماژول سبک‌وزنِ توجهِ کانالی (پیش از LSTM) و توجهِ زمانی (پس از LSTM) هم ورودی پالایش می‌شود و

1. Information Fusion
2. Fully Convolutional Networks
3. Real-Time
4. Recurrent Neural Networks
5. Residual CNNs

هم تجمیع آگاه‌به‌زمان شکل می‌گیرد. این طراحی، بی‌نیاز از دست‌کاری معماری LSTM، هزینه‌ی محاسباتی را اندک نگاه می‌دارد. در قسمت تحلیل پیچیدگی، مفصلاً به این موضوع پرداخته شده است.

مطالعات [۲۶] تا [۳۰] هرکدام از زاویه‌ای متفاوت به مسئله‌ی شناسایی کنش در ویدیو پرداخته‌اند و مکمل یکدیگرند. در پژوهش [۲۶]، حسن و همکاران چارچوبی برای یادگیری افزایشی فعالیت‌ها در ویدیوها معرفی کردند که بدون نیاز به برچسب‌گذاری کامل داده‌ها، مدل را به‌صورت پیوسته به‌روزرسانی می‌کند و با کمک یادگیری فعال، عملکردی نزدیک به مدل‌های ازپیش‌آموزش‌دیده دارد. در ادامه، ایکیزلر- سینبیس و همکاران در [۲۷] با ترکیب ویژگی‌های چندگانه از شیء، صحنه و انسان در قالب چارچوب یادگیری نمونه‌های چندگانه، نشان دادند که اطلاعات زمینه‌ای می‌تواند مکمل داده‌های بدنی در تشخیص کنش باشد. ونگ و همکاران در [۲۸] با معرفی روش مسیرهای متراکم[1] ویژگی‌های حرکتی را از قاب‌های[2] متوالی با استفاده از جریان نوری متراکم استخراج کردند و نشان دادند که این توصیف‌گر در ویدیوهای واقعی نسبت به روش‌های قبلی مقاوم‌تر است. در پژوهش [۲۹]، شرما و همکاران مدل توجه نرم مبتنی بر LSTM را برای تمرکز خودکار بر بخش‌های مهم ویدیو توسعه دادند که با یادگیری توجه زمانی- مکانی، طبقه‌بندی را با دقت بالاتری انجام می‌دهد. در نهایت، چو و همکاران در [۳۰] با بهره‌گیری از گروه‌تنکی[3] و تمرکز بر حرکات محلی به‌جای حرکات کلی، مدلی مقاوم در برابر حرکت دوربین و پس‌زمینه‌ی شلوغ ارائه کردند که با ترکیب توصیف‌گرهای حرکتی موضعی و کلی، توانست در مجموعه‌داده‌های مرجع عملکرد بهتری از روش‌های موجود به دست آورد.

مطالعات [۳۱] تا [۳۷] به بررسی ابعاد گوناگون بازنمایی ویژگی‌ها و یادگیری ساختارهای سلسله‌مراتبی در بازشناسی کنش می‌پردازند. در پژوهش [۳۱]، روانبخشo و همکاران نشان دادند که ویژگی‌های حاصل از لایه‌ی fc۷ در شبکه‌های CNN که صرفاً بر اساس تصاویر آموزش دیده‌اند، برای تشخیص کنش کافی نیستند؛ از این‌رو ساختاری سلسله‌مراتبی برای نمایش تغییرات زمانی و استخراج زیرکنش‌ها پیشنهاد کردند که در سطوح بالاتر توالی کلی و در سطوح پایین‌تر حرکات جزئی را مدل می‌کند. در ادامه، جاویدانی و همکاران در [۳۲] روشی سبک و کارا برای ویدیوهای سوم‌شخص ارائه دادند که با محاسبه‌ی جریان نوری و فشرده‌سازی ویژگی‌ها توسط PCA، داده‌ی ویدیو را به یک سری زمانی چندکاناله تبدیل کرده و با شبکه‌ی کانولوشنی یک‌بعدی بر روی بُعد زمانی آموزش می‌دهند؛ این روش با کاهش چشمگیر پارامترهای آموزشی، امکان یادگیری روی داده‌های اندک را فراهم کرده است. لو و همکاران در [۳۳] نیز با استفاده از مدل تصادفی مارکوف سلسله‌مراتبی، سازگاری سوپروکسل‌ها را در سطوح مختلف برای بخش‌بندی دقیق‌تر کنش‌ها تضمین کردند.

در پژوهش [۳۴]، گیوکساری و همکاران با الهام از موفقیت‌های تشخیص شیء در تصاویر دوبعدی، مدل Action Tube را معرفی کردند که با استخراج نواحی پیشنهادی مبتنی بر حرکت و اتصال آن‌ها در طول زمان، آشکارسازی مکانی- زمانی کنش را ممکن می‌سازد. پنگ و همکاران در [۳۵] با ارائه‌ی مطالعه‌ای جامع پیرامون مدل Bag of Visual Words (BoVW)، تأثیر مراحل مختلف شامل استخراج، کدگذاری، و همجوشی ویژگی‌ها را تحلیل و در نهایت نمایش ترکیبی مؤثری پیشنهاد کردند. مقاله‌ی [۳۶] (ژوانگ و همکاران) با ارائه‌ی مجموعه‌داده‌ی J-HMDB و تحلیل سیستماتیک روش‌های موجود، نشان داد که ویژگی‌های وضعیت بدن[4] در طول زمان نقش کلیدی در بهبود دقت دارند و توسعه‌ی الگوریتم‌های جریان نوری و تشخیص انسان باید اولویت یابد. نهایتاً، پنگ و همکاران در [۳۷] با معرفی بردارهای فیشر پشته‌ای ساختار چندلایه‌ای از کدگذاری فیشر را ارائه دادند که قادر است اطلاعات سطح میانی را بدون نیاز به استخراج بخش‌های مجزا حفظ کند.

۳- روش پیشنهادی

۳-۱ چارچوب کلی

نمای کلی روش در شکل ۲ آمده است: ابتدا جریان‌نوری میان فریم‌های متوالی محاسبه می‌شود و تصاویر حاصل به یک CNN از پیش آموزش دیده داده می‌شوند تا خروجی FC به عنوان بردار ویژگی nبعدی استخراج شود. برای کاهش پیچیدگی، PCA با حفظ حدود ۸۰ درصد واریانس، بعد ویژگی را به فضای کم‌بعد d (نزدیک به ۵۰) کاهش می‌دهد. پیش از مدل زمانی، یک توجه کانالی سبک وزن بر بردارهای کم‌بعد اعمال می‌شود تا مولفه‌های مفید تقویت و مولفه‌های کم اثر تضعیف شوند. سپس سری زمانی پالایش شده به یک LSTM دو لایه پشته‌ای داده می‌شود تا وابستگی‌های بلندمدت آموخته گردد. پس ازLSTM، توجه زمانی برای تجمیع آگاه به زمان به کار می‌رود و در نهایت بردار به طبقه‌بندی ارسال و برچسب کنش پیش‌بینی می‌شود.

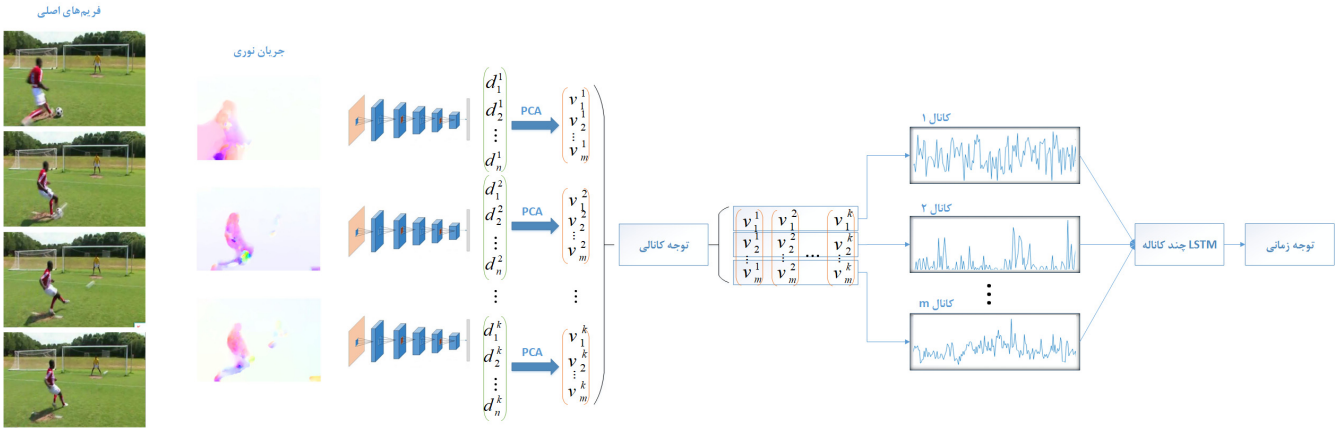
۳-۲ محاسبه‌ی جریان نوری

جریان نوری معمولاً برای توصیف حرکت موجود بین دو فریم مجاور در ویدیو مورد استفاده قرار می‌گیرد. با بهره‌گیری از آن، اشیای ثابت از متحرک قابل تشخیص هستند. به همین دلیل، این تکنیک به‌طور گسترده‌ای در مسائل تحلیل ویدیو به‌کار گرفته می‌شود. با این حال، محدودیت مهم آن این است که تنها در فواصل زمانی کوتاه به‌خوبی عمل می‌کند. در این پژوهش، با توجه به تمرکز اصلی بر جنبه‌ی حرکتی ویدیو، تصاویر جریان نوری بین فریم‌های متوالی برای تمامی ویدیوها در گام اول محاسبه می‌شوند.

۳-۳ محاسبه‌ی جریان نوری

روش‌های مختلفی برای توصیف تصاویر جریان نوری وجود دارد. برای این منظور، توصیف‌گرهای دستی مانند HOG و SIFT مفید هستند. اما گزینه‌ی بهتر، استفاده از شبکه‌های کانولوشنی از پیش آموزش‌دیده است. این شبکه‌ها پیش‌تر بر روی مجموعه‌داده‌های بزرگ تصویری آموزش دیده‌اند و در مقایسه با توصیف‌گرهای دستی، نتایج بهتری در حوزه‌ی شناسایی تصاویر ارائه داده‌اند. با این روش، نه‌تنها نیاز به آموزش CNN که فرآیندی زمان‌بر است از بین می‌رود، بلکه یک بازنمایی قدرتمند نیز حاصل می‌شود. بنابراین، با تغذیه‌ی تصاویر جریان نوری به شبکه‌های کانولوشنی از پیش آموزش‌دیده، نگاشت‌های ویژگی سطح بالا از لایه‌ی کاملاً متصل پایانی به‌عنوان بردارهای توصیفی nبعدی استخراج می‌گردند.

1. Dense Trajectories
2. Frames
3. Group Sparsity
4. Pose



شکل ۲: ساختار دقیق چارچوب پیشنهادی. تصاویر جریان نوری بین فریم‌های مجاور محاسبه شده و توسط یک شبکه‌ی عصبی کانولوشنی از پیش آموزش‌دیده توصیف می‌شوند. الگوریتم کاهش بُعد PCA اجرا می‌شود تا مرحله‌ی یادگیری طبقه‌بند ساده‌تر گردد. سری‌های زمانی چندکاناله‌ی حاصل، به شبکه‌ی LSTM داده می‌شوند تا الگوهای بلندمدت شناسایی شده و طبقه‌بندی آن‌ها به‌صورت مؤثر انجام گیرد.

۳-۳ تحلیل مولفه‌های اصلی

تعداد ابعاد در لایه‌ی کاملاً متصل پایانی شبکه‌های کانولوشنی معمولاً بالا است. همچنین، تصاویر جریان نوری به‌دلیل نمایش حرکت کوتاه‌مدت و عدم حضور اشیاء در آن‌ها، بسیار شبیه به‌هم هستند. این امر منجر به داده‌های پرتراکم و همبسته در فضای با ابعاد بالا می‌شود. لذا، دنبال‌کردن آن‌ها در طول زمان و یادگیری الگوها در چنین فضایی نسبتاً دشوار است. برای کاهش تعداد ابعاد و ساده‌تر کردن فرآیندهای بعدی، از الگوریتم ساده و مؤثر کاهش بعد تحلیل مولفه‌های اصلی (PCA) استفاده می‌شود.

تحلیل مؤلفه‌های اصلی، محورهایی را می‌یابد که با افکندن داده‌ها بر روی آن‌ها، واریانس بیشینه حاصل می‌شود. این کار از طریق محاسبه‌ی بردارهای ویژه‌ی ماتریس کوواریانس ممکن می‌گردد. بردار ویژه‌ی متناظر با بزرگ‌ترین مقدار ویژه، داده‌ها را به فضایی جدید نگاشت می‌کند که در آن بیش‌ترین تغییرات حفظ می‌شود. این روند برای دومین مقدار ویژه و به‌همین ترتیب ادامه دارد.

نگاشت X بر بردار W به صورت زیر تعریف می‌شود

$$Z = W^T X \quad (۱)$$

هدف کلی یافتن W است به‌طوری که واریانس Z بیشینه شود.

$$\mathrm{Var}(Z) = \mathrm{Var}(W^T X) = E\left[(W^T X - W^T \mu)^2\right] \quad (۲)$$

با ساده‌سازی (۲)، رابطه‌ی زیر به‌دست می‌آید

$$\mathrm{Var}(Z) = W^T E\left[(X-\mu)(X-\mu)^T\right] W = W^T \Sigma W \quad (۳)$$

که در آن

$$\mathrm{Var}(X) = E\left[(X-\mu)(X-\mu)^T\right] = \Sigma \quad (۴)$$

برای بیشینه‌سازی واریانس Z تحت شرط $\|W\|=۱$، می‌توان از تابع لاگرانژ با پارامتر α استفاده کرد که به‌صورت زیر نوشته می‌شود

$$L(W,\alpha) = W_۱^T \Sigma W_۱ - \alpha(W_۱^T W_۱ - ۱) \quad (۵)$$

برای یافتن ماکزیمم، مشتق تابع بالا نسبت به W را صفر قرار می‌دهیم که منجر به رابطه‌ی زیر می‌شود

$$\Sigma W_۱ = \alpha W_۱ \quad (۶)$$

در نتیجه، $W_۱$ بردار ویژه‌ی اول ماتریس Σ است که مربوط به بزرگ‌ترین مقدار ویژه می‌باشد. سایر بردارهای ویژه همانند قبل محاسبه می‌شوند.

توصیف‌گرهای تمامی تصاویر جریان نوری با ضرب ساده‌ی آن‌ها در بردارهای ویژه‌ی یافت‌شده از فضای n بعدی به فضای جدید منتقل می‌شوند. در این میان، نسبت واریانس نکته‌ای کلیدی است که به‌صورت تجربی برابر با ۰٫۸ در نظر گرفته شده و برای پیشبرد ایده‌ی بعدی کافی به‌نظر می‌رسد. در نتیجه، تمامی بردارهای توصیفی n بعدی به فضای m بعدی تبدیل می‌شوند، که m تعداد ابعادی است که برای حفظ ۸۰ درصد از اطلاعات اصلی کافی است.

۳-۴ مکانیزم توجه کانالی

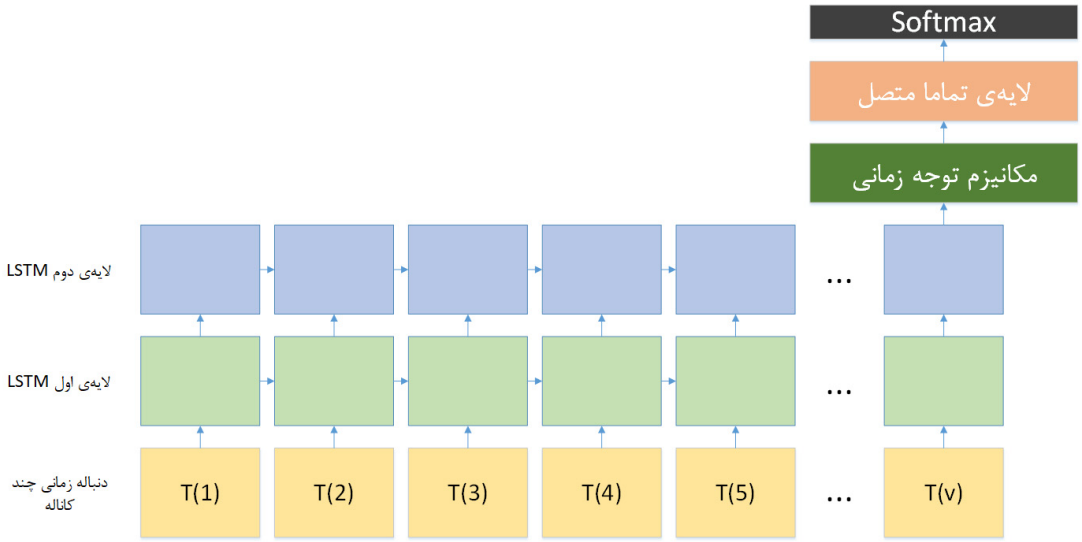
این ماژول وزن‌دهیِ تطبیقی به مؤلفه‌های بردارهای ویژگیِ حاصل از PCA در هر گام زمانی را بر عهده دارد تا مؤلفه‌های اطلاعاتی تقویت و مؤلفه‌های کم‌اثر یا هم‌پوشان تضعیف شوند. فرض کنید پس از کاهش‌بُعد، در هر گام زمانی t، بردار ویژگی $x_t \in \mathbb{R}^d$ در اختیار داریم که در پیاده‌سازی حاضر $d \approx ۵۰$ (برای PoV = ۰٫۸) است. ایده مکانیزم توجه کانالی آن است که پیش از ورود توالی به LSTM، برای هر x_t یک گیت کانالیِ سبک‌وزن اعمال شود تا نسبت به محتوای همان گام، مؤلفه‌های با ارزشِ بیشتر را پررنگ‌تر کند. برای پیاده‌سازیِ توجهِ کانالی از یک شبکه کوچک کاهشی–افزایشی استفاده می‌کنیم. این شبکه ابتدا x_t را به یک فضای میانیِ کم‌بعد فرافکنی می‌کند و سپس با استفاده از یک نگاشتِ افزایشی و تابع سیگموئید، یک ماسک کانالی در بازه (۰،۱) می‌سازد. خروجی، نسخه گیت‌شده از x_t است

$$s_t = \mathrm{ReLU}(W_۱ x_t + b_۱) \quad (۷)$$

$$m_t = \sigma(W_۲ s_t + b_۲) \quad (۸)$$

$$x_t' = m_t \odot x_t \quad (۹)$$

در اینجا σ تابع سیگموید و $\odot$ ضرب مؤلفه‌به‌مؤلفه است. r_c کوچک‌تر از d در نظر گرفته می‌شود (برای نمونه r_c=۸). پارامترهای $(W_۱, b_۱)$، $(W_۲, b_۲)$ در طول زمان مشترک هستند. یعنی برای تمام گام‌ها از یک مجموعه پارامتر استفاده می‌شود. بردار m_t نقش اهمیتِ کانالی را بازی می‌کند: مؤلفه‌هایی از x_t که با وظیفه‌ی طبقه‌بندی مرتبط‌تر هستند،



شکل ۳: معماری LSTM چندکاناله‌ی پشته‌ای برای طبقه‌بندی سری‌های زمانی حاصل‌شده استفاده می‌شود. $T(i)$ مقدار سری زمانی چندکاناله را در گام زمانی i ام نمایش می‌دهد.

مقادیر بزرگ‌تری در m_t دریافت می‌کنند و تقویت می‌شوند. در مقابل، مؤلفه‌های نویزی به‌طور تطبیقی تضعیف می‌گردند. از آن‌جا که d پس از PCA کوچک است، این پالایش کانالی با هزینه بسیار اندک انجام می‌شود و ورودی تمیزتری را به بلوک زمانی می‌سپارد.

۳-۵ ایجاد سری‌های زمانی

هم‌راستا کردن توصیف‌گرهای نگاشته‌شده‌ی تصاویر جریان نوری متوالی در کنار یکدیگر، منجر به ایجاد یک ماتریس می‌شود. تعداد سطرهای این ماتریس برابر m و تعداد ستون‌ها وابسته به تعداد فریم‌های هر ویدیو است. می‌توان به این ماتریس از دیدگاهی دیگر نگریست. واضح است که هر سطر از یک ویژگی خاص نشأت گرفته و در طول زمان در امتداد ستون‌های ماتریس پیگیری می‌شود. بنابراین، مجموعه‌ای از n سری زمانی به‌دست می‌آید که هر کدام نمایانگر حرکت یک ویژگی ویدیو هستند و مسئله آن است که این سری‌های زمانی بر اساس برچسب واقعی‌شان طبقه‌بندی شوند.

۳-۶ طبقه‌بندی سری‌های زمانی با استفاده از LSTM

شبکه‌های عصبی بازگشتی قادر به استخراج الگوهای متمایز از داده‌های ترتیبی هستند. با این حال، یکی از ضعف‌های آن‌ها مشکل ناپدیدشدن گرادیان است. LSTM نسخه‌ای از شبکه‌های بازگشتی است که از این مشکل رنج نمی‌برد. لذا، آن‌ها برای استخراج الگوهای بلندمدت از داده‌های ترتیبی با کمک الگوریتم ساده‌ی پس‌انتشار خطا[1] بسیار مناسب‌اند. در چارچوب ما برای یادگیری الگوهای بلندمدت از سری‌های زمانی حرکتیِ حاصل‌شده، از معماری عمیق LSTM استفاده شده است. این معماری با پشته‌کردن دو لایه‌ی LSTM ساخته شده است، به‌طوری که خروجی یک لایه، ورودی لایه‌ی بعدی است. پس از این لایه‌ها، یک لایه‌ی کاملاً متصل با تعداد نورون برابر با تعداد کلاس‌ها قرار گرفته است. در نهایت، برای به‌دست‌آوردن احتمال کلاس‌ها، یک لایه‌ی softmax روی آن‌ها قرار می‌گیرد. شکل ۳ معماری پیشنهادی این چارچوب را نشان می‌دهد.

یکی از مزایای برجسته‌ی شبکه‌های بازگشتی آن است که با داده‌هایی با طول‌های متفاوت سازگار هستند. این ویژگی مطلوبی است چرا که ویدیوها دارای تعداد فریم‌های متفاوتی هستند. مزیت دیگر LSTM ها در مقایسه با شبکه‌های کانولوشنی، تعداد پارامترهای یادگیری به‌مراتب کمتری نسبت به CNNها دارند. بنابراین، قادر هستیم روش پیشنهادی را حتی بر روی مجموعه‌داده‌های کوچک‌تر نیز ارزیابی کنیم. این نکته مزیت بسیار مهمی در مقایسه با شبکه‌های کانولوشنی مانند C۳D [۱۲] است. زیرا پارامترهای یادگیری C۳D در سه بعد (دو بعد فضایی و یک بعد زمانی) تعریف می‌شود که سبب ایجاد تعداد بسیار زیادی پارامتر می‌شود.

۳-۷ مکانیزم توجه زمانی

برای تمرکز تطبیقی بر لحظاتِ مهمِ زمانی و ساخت یک بازنمایی آگاه‌به‌زمان، از یک لایه‌ی توجه افزایشی استفاده می‌کنیم. امتیاز هر گام زمانی به‌صورت زیر محاسبه می‌شود

$$e_t = v^\top \tanh\left(Wh_t + b\right), t = ۱,\ldots,T \qquad (۱۰)$$

که در آن $W \in \mathbb{R}^{r\times h}, v \in \mathbb{R}^r$ و r بعد میانی کوچک است. وزن‌های توجه زمانی با اعمال تابع softmax به‌دست می‌آیند

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=۱}^{T}\exp(e_k)}, \sum_{t=۱}^{T}\alpha_t = ۱ \qquad (۱۱)$$

بردار $c \in \mathbb{R}^h$ به‌صورت میانگینِ وزن‌دار از خروجی‌های زمانی تشکیل می‌شود

$$c = \sum_{t=۱}^{T}\alpha_t h_t \qquad (۱۲)$$

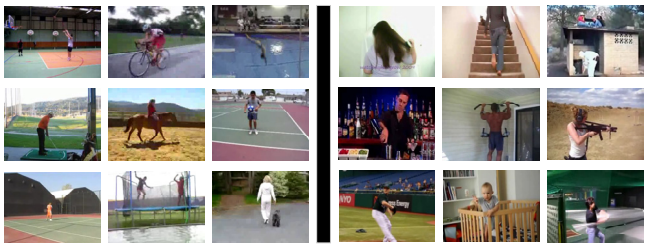
این بردار سپس جهت طبقه‌بندی به softmax اعمال می‌شود

$$z = W_c c + b_c, \hat{y} = \text{softmax}(z) \qquad (۱۳)$$

که در آن $W_c \in \mathbb{R}^{C\times h}$ و C تعداد کلاس‌ها است.

توجه زمانی دو مزیت کلیدی دارد: (۱) افزایش دقت با تمرکز بر بازه‌های اطلاع‌رسان (مثلاً فریم‌های اوج حرکت)، و (۲) تبیین‌پذیری، از طریق نمایش α_t بر حسب زمان.

1. Backpropagation



شکل ۴: برخی از فریم‌های نمونه از کلاس‌های مختلف دو مجموعه‌داده: UCF۱۱ (سمت چپ) و jHMDB (سمت راست).

٤- نتایج تجربی

۴-۱ مجموعه داده‌ها

آزمایش‌های روش پیشنهادی بر روی دو مجموعه‌داده عمومی UCF۱۱ و jHMDB انجام شده است. مجموعه‌داده‌ی UCF۱۱ شامل ۱۱ کلاس از دسته‌بندی‌های مختلف ورزشی انسان‌ها مانند شیرجه، والیبال و دوچرخه‌سواری است. ویدیوها از وب‌سایت یوتیوب جمع‌آوری شده‌اند و فرمت آن‌ها Mpeg۴ است. نرخ فریم تمامی ویدیوها برابر با ۲۹٫۹۷ فریم بر ثانیه می‌باشد. به‌دلیل ناسازگاری‌های زیاد ناشی از حرکت دوربین، ظاهر و حالت اجسام و همچنین پس‌زمینه‌های شلوغ، مجموعه‌داده‌ی UCF۱۱ بسیار چالش‌برانگیز به‌شمار می‌رود.

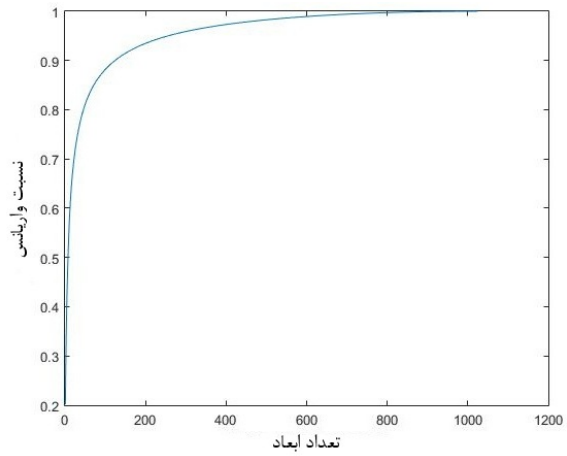
مجموعه‌داده‌ی jHMDB زیرمجموعه‌ای از مجموعه‌ی HMDB۵۱ است. ویدیوهای این مجموعه شامل فعالیت‌های روزمره‌ی انسان‌ها مانند شانه‌زدن مو، پریدن، گلف بازی‌کردن و دویدن هستند. در مجموع، ۹۲۳ کلیپ ویدیویی در ۲۱ کلاس توزیع شده‌اند. طول این کلیپ‌ها بین ۱۵ تا ۴۰ فریم است که بسیار کوتاه‌تر از کلیپ‌های UCF۱۱ می‌باشد. اولین پژوهش ارائه‌شده بر روی این مجموعه‌داده سه تقسیم‌بندی متفاوت برای آموزش و آزمون معرفی کرد تا روش خود را ارزیابی کند. ما نیز از همین الگو پیروی کردیم تا بتوانیم روش پیشنهادی خود را با سایر روش‌های پیشرفته مقایسه نماییم. شکل ۴ نمونه‌هایی از فریم‌های این دو مجموعه‌داده را نشان می‌دهد.

۴-۲ توصیف تصاویر جریان نوری

برای توصیف تصاویر جریان نوری، از شبکه‌ی عصبی کانولوشنی از پیش آموزش‌دیده‌ی GoogLeNet استفاده شده است. این شبکه پیش‌تر بر روی مجموعه‌داده‌ی ILSVRC که شامل بیش از ۱٫۴ میلیون تصویر است، آموزش دیده است. با توجه به نتایج برجسته‌ای که در حوزه‌ی شناسایی تصویر کسب کرده، برتری آن نسبت به سایر شبکه‌های از پیش آموزش‌دیده به‌وضوح قابل مشاهده است. توصیف تصاویر جریان نوری از طریق استخراج ویژگی‌های لایه‌ی کاملاً متصل پایانی انجام می‌شود. GoogLeNet در این لایه دارای ۱۰۲۴ نورون است.

۴-۳ تحلیل مؤلفه‌های اصلی

پس از توصیف تصاویر جریان نوری، به‌منظور کاهش ابعاد آن‌ها، تحلیل مؤلفه‌های اصلی (PCA) بر روی داده‌ها انجام شد. برای این منظور، تمامی فریم‌های جریان نوری از ویدیوهای کلاس‌های مختلف گردآوری شدند. از آن‌جا که این تصاویر تا حد زیادی به یکدیگر شبیه هستند، PCA می‌تواند آن‌ها را به‌طور مؤثری در فضای ویژگی با ابعاد بسیار کمتر نمایش دهد. رابطه‌ی بین تعداد ابعاد و نسبت واریانس حفظ‌شده در عملیات PCA در شکل ۵ نشان داده شده است. همان‌طور



شکل ۵: رابطه‌ی بین نسبت واریانس (PoV) حاصل از اجرای PCA بر روی تصاویر جریان نوری و تعداد ابعاد در توصیف‌های جریان نوری برای مجموعه‌داده‌ی UCF۱۱ نشان داده شده است. با تنظیم مقدار PoV برابر با ۰٫۸، تعداد ابعاد برابر با ۵۰ خواهد بود.

که دیده می‌شود، با حفظ ۲۰۰ بعد از مجموع ۱۰۲۴ بعد، حدود ۹۰ درصد از کل اطلاعات حفظ خواهد شد. ما ۸۰ درصد از انرژی کل را به‌عنوان نقطه‌ی بهینه برای اجرای ایده‌ی بعدی‌مان، یعنی آموزش LSTM در زمانی نسبتاً کوتاه، در نظر گرفتیم. این کار نه‌تنها فضای ویژگی را کارآمدتر می‌سازد، بلکه فرآیندهای بعدی مانند یادگیری طبقه‌بندها، از جمله آموزش LSTM در این پژوهش را نیز بسیار ساده‌تر می‌کند.

۴-۴ تحلیل پیچیدگی

برای یک LSTM پشته‌ایِ دولایه با اندازه نهان ۲۵۶، به‌کارگیریِ کاهش‌بعد با PCA به ۵۰ بعد، موجب کاهشِ حدود ۲٫۱۹ برابر در تعداد پارامترها و هزینه محاسباتی هر گام زمانی نسبت به حالت بدون PCA می‌شود (سطر PCA تنها در جدول ۱).

افزودنِ توجهِ کانالی بر ویژگی‌های PCA، به‌دلیل کوچک‌بودن فضای ورودی، تنها حدود ۰٫۱۰٪ به پارامترها و MACs/گام می‌افزاید و نقش پالایش مؤلفه‌های کم‌اثر را بدون سربار معنادار ایفا می‌کند. همچنین توجهِ زمانی بر خروجی‌های LSTM با افزایش محدودِ ۱٫۹۷٪ در پارامترها و ۱٫۹۶٪ در MACs/گام، امکان وزن‌دهی تطبیقی به لحظات اطلاعاتی توالی را فراهم می‌آورد و معمولاً به بهبود دقت و تبیین‌پذیری منجر می‌شود. ترکیب هر دو توجه همچنان سبک باقی می‌ماند (افزایش ۲٫۰۶٪ در MACs/گام نسبت به PCA تنها) و نسبت MACs به حالت بدون PCA را در حدود ۰٫۴۶۶ برابر حفظ می‌کند. بنابراین کل سامانه از منظر محاسباتی سبک‌وزن باقی می‌ماند، در حالی‌که کیفیت بازنمایی زمانی به‌واسطه مکانیزم‌های توجه بهبود می‌یابد.

۴-۵ آموزش LSTM

برای آموزش شبکه‌ی LSTM، از الگوریتم پس‌انتشار خطا با اندازه‌ی دسته‌ی ۵۰ برای هر دو مجموعه‌داده استفاده شده است. نرخ یادگیری در ابتدا برابر با ۰/۰۰۱ در نظر گرفته شده و این مقدار در هر ۲۰ دوره[1] کاهش می‌یابد. تعداد سلول‌های حافظه برای هر دو لایه‌ی LSTM برابر با ۲۵۶ تنظیم شده است. لایه‌ی کاملاً متصل، ویژگی‌های یادگرفته‌شده را به برچسب‌های مربوطه نگاشت می‌کند و لایه‌ی softmax نیز احتمال تعلق به هر کلاس را خروجی می‌دهد.

1. Epoch

جدول ۱: مقایسه هزینه محاسباتی و تعداد پارامترها در پیکربندی‌های مختلف (بدون/با PCA و توجه کانالی/زمانی) نسبت به حالت مبنا.

پیکربندی	تعداد پارامتر	MACs/گام	Δ پارامتر (٪)	ΔMACs (٪)	نسبت MACs به حالت مبنا (بدون PCA)
PCA بدون (LSTM دو لایه)	۱٫۸۳۷٫۰۵۶	۱٫۸۳۵٫۰۰۸	+٪۱۱۸٫۷۸	+٪۱۱۹٫۰۷	۱٫۰۰۰
PCA تنها ($d \approx ۵۰$)	۸۳۹٫۶۸۰	۸۳۷٫۶۳۲	٪۰٫۰۰	٪۰٫۰۰	۰٫۴۵۶
PCA + توجهِ کانالی	۸۴۰٫۵۳۸	۸۳۸٫۴۳۲	+٪۰٫۱۰	+٪۰٫۱۰	۰٫۴۵۷
PCA+ توجهِ زمانی	۸۵۶٫۱۹۲	۸۵۴٫۰۸۰	+٪۱٫۹۷	+٪۱٫۹۶	۰٫۴۶۵
PCA + هر دو توجه	۸۵۷٫۰۵۰	۸۵۴٫۸۸۰	+٪۲٫۰۷	+٪۲٫۰۶	۰٫۴۶۶

۴-۶ نتایج روی مجموعه‌داده‌ی UCF۱۱

روش پیشنهادی بر روی مجموعه‌داده‌ی UCF۱۱ با استفاده از طرح اعتبارسنجی Leave-One-Out-Cross-Validation همانند پژوهش اصلی ارزیابی شده است. نتایج درستی طبقه‌بندی و مقایسه با دیگر روش‌های پیشرفته در جدول ۲ گزارش شده است. همان‌طور که مشخص است، روش پیشنهادی عملکرد بهتری نسبت به سایر روش‌ها داشته است.

۴-۷ نتایج روی مجموعه‌داده‌ی jHMDB

در مجموعه‌داده‌ی jHMDB، ارزیابی بر اساس سه تقسیم‌بندی جداگانه‌ی آموزش و آزمون که در پژوهش اصلی پیشنهاد شده‌اند، انجام گرفته است. عملکرد کلی روش پیشنهادی با میانگین‌گیری از نتایج این سه تقسیم‌بندی به‌دست آمده است. درستی طبقه‌بندی روش پیشنهادی و سایر رویکردها برای این مجموعه‌داده در جدول ۳ گزارش شده است.

۵- نتیجه‌گیری

روش پیشنهادی این مقاله بر استخراج الگوهای بلندمدت از ویدیوهای حرکتی تمرکز دارد. پویایی‌های حرکتی از طریق دنبال‌کردن عناصر جریان‌نوری که نمایانگر حرکت‌های کوتاه‌مدت هستند به‌دست می‌آیند. تصاویر جریان‌نوری با یک شبکه عصبی کانولوشنی پیش‌آموزش‌دیده توصیف می‌شوند و به‌دلیل ابعاد بالای بردارهای توصیفی و امکان هم‌بستگی میان آن‌ها، از الگوریتم مؤثر کاهش‌بعد PCA استفاده می‌شود. با هم‌راستا کردن بردارهای حاصل‌شده، یک سری‌زمانی چندکاناله شکل می‌گیرد و یک LSTM دولایه پشته‌ای برای استخراج الگوهای بلندمدت بر آن آموزش داده می‌شود. به‌منظور پالایش ورودی و تجمیع آگاه‌به‌زمان، دو ماژول توجه سبک‌وزن به سامانه افزوده شده است: توجه کانالی پیش از LSTM که مؤلفه‌های کم‌بعد حاصل از PCA را به‌صورت تطبیقی وزن‌دهی می‌کند، و توجه زمانی پس از LSTM که با وزن‌دهی به گام‌های زمانی، لحظات اطلاع‌رسان را برجسته می‌سازد و یک بردار منسجم برای طبقه‌بندی تولید می‌کند. این طراحی ضمن حفظ سبک‌وزنی مدل، با کاهش ابعاد ورودی از ۱۰۲۴ به حدود ۵۰، تعداد پارامترها و هزینه محاسباتی هر گام را به‌طور معنادار کاهش می‌دهد و در عین حال دقت را بهبود می‌بخشد. شواهد عددی این ادعا در تحلیل‌پیچیدگی ارائه شده

جدول ۲: مقایسه درستی روش پیشنهادی با سایر روش‌ها روی مجموعه داده UCF۱۱.

روش	درستی (میانگین ± انحراف معیار)
حسن و همکاران [۲۶]	۵۴/۵٪
لیو و همکاران [۱۵]	۷۱/۲٪
ایکیزلر–سینبیس و همکاران [۲۷]	۷۵/۲٪
مسیرهای پرتراکم [۲۸]	۸۴/۲٪
توجه نرم (Soft Attention) [۲۹]	۸۴/۹٪
چو و همکاران [۳۰]	۸۸/۰٪
Snippets [۳۱]	۸۹/۵٪
LSTM دومسیره (conv-L) [۲۳]	۸۹/۲٪
LSTM دومسیره (fc-L) [۲۳]	۹۳/۷٪
LSTM دومسیره (fu-1) [۲۳]	۹۴/۲٪
LSTM دومسیره (fu-2) [۲۳]	۹۴/۶٪
علی و همکاران [۳۲]	۹۵/۷٪
روش پیشنهادی بدون توجه کانالی و زمانی	**۹۵/۸±۰/۵٪**
روش پیشنهادی با توجه کانالی	**۹۵/۸±۰/٤٪**
روش پیشنهادی با توجه زمانی	**۹٦/۷±۰/۳٪**
روش پیشنهادی با توجه کانالی و زمانی	**۹٦/۷±۰/۳٪**

جدول ۳: مقایسه درستی روش پیشنهادی با سایر روش‌ها روی مجموعه داده jHMDB.

روش	درستی (میانگین ± انحراف معیار)
LSTM دومسیره (conv-L) [۲۳]	۵۴/۵٪
علی و همکاران [۳۲]	۷۱/۲٪
Gd, و همکاران [۳۳]	۷۵/۲٪
گیوکساری و همکاران [۳۴]	۸۴/۲٪
پنگ و همکاران [۳۵]	۸۴/۹٪
LSTM دومسیره (fc-L) [۲۳]	۸۸/۰٪
LSTM دومسیره (fu-1) [۲۳]	۸۹/۵٪
ژوانگ و همکاران [۳۶]	۸۹/۲٪
پنگ و همکاران [۳۷]	۹۳/۷٪
LSTM دومسیره (fu-2) [۲۳]	۹۴/۲٪
روش پیشنهادی بدون توجه کانالی و	**۹۵/۸±۰/۳٪**
روش پیشنهادی با توجه کانالی	**۹۵/۹±۰/۴٪**
روش پیشنهادی با توجه زمانی	**۹۶/۳±۰/۳٪**
روش پیشنهادی با توجه کانالی و زمانی	**۹۶/۳±۰/۲٪**

است. ارزیابی بر روی دو مجموعه‌داده عمومی UCF۱۱ و jHMDB نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های پیشرفته موجود به نتایج برتر دست می‌یابد.

مراجع

[1] H. Qiu and B. Hou, "Multi-grained clip focus for skeleton-based action recognition," *Pattern Recognition,* vol. 148, Article ID: 110188, Apr. 2024.

[2] M. Karim, *et al.*, "Human action recognition systems: A review of the trends and state-of-the-art," *IEEE Access,* vol. 12, pp. 36372-36390, 2024.

[3] J. Xie, *et al.*, "Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 38, no. 6, pp. 6225-6233. Vancouver, Canada, 20-27 Feb. 2024.

[4] Y. Ma and R. Wang, "Relative-position embedding based spatially and temporally decoupled Transformer for action recognition," *Pattern Recognition,* vol. 145, Article ID: 109905, Jan. 2024.

[5] S. Guo, H. Pan, G. Tan, L. Chen, and C. Gao, "A High Invariance Motion Representation for Skeleton-Based Action Recognition," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 30, no. 08, Article ID: 1650018, 2016.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. of the 28th Int. Conf. on Neural Information Processing Systems*, vol. 1, pp. 568-576, Montreal, Canada, 8-13 Dec. 2014.

[7] T. P. Nguyen, A. Manzanera, M. Garrigues, and N. -S. Vu, "Spatial motion patterns: action models from semi-dense trajectories," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 28, no. 07, Article ID: 1460011, 2014.

[8] A. Karpathy, *et al.*, "Large-scale video classification with convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1725-1732, Columbus, OH, USA, 23-28 Jun. 2014.

[9] J. Yue-Hei Ng, *et al.*, "Beyond short snippets: Deep networks for video classification," in *Proc. of the IEEE Con. on Computer Vision and Pattern Recognition*, pp. 4694-4702., Boston, MA, USA, 7-12 Jun. 2015.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1933-1941, Las Vegas, NV, USA, 27-30 Jun. 2016.

[11] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 27, no. 12, pp. 2613-2622, Dec. 2017.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR, abs/1412.0767,* vol. 2, p. 7, 2014.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. on pattern analysis and machine intelligence,* vol. 35, no. 1, pp. 221-231, Jan. 2013.

[14] S. Yu, Y. Cheng, L. Xie, and S. Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision,* vol. 11, no. 8, pp. 744-749, Dec. 2017.

[15] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1996-2003, Miami, FL, USA, 20-25 Jun. 2009.

[16] X. Liu, Y. Li, and Q. Wang, "Multi-view hierarchical bidirectional recurrent neural network for depth video sequence based action recognition," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 32, no. 10, Article ID: 1850033, 2018.

[17] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Hierarchical Multi-scale Attention Networks for action recognition," *Signal Processing: Image Communication,* vol. 61, pp. 73-84, Feb. 2018.

[18] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Computer Vision and Image Understanding,* vol. 166, pp. 41-50, Jan. 2018.

[19] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. on Multimedia,* vol. 20, no. 3, pp. 634-644, Mar. 2018.

[20] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. on Multimedia,* vol. 19, no. 7, pp. 1510-1520, Jul. 2017.

[21] A. Iqbal, A. Richard, H. Kuehne, and J. Gall, "Recurrent residual learning for action recognition," in *Proc. German Conf. on Pattern Recognition*, pp. 126-137, Basel, Switzerland, 12-15 Sept. 2017.

[22] A. Piergiovanni, C. Fan, and M. S. Ryoo, "Learning latent sub-events in activity videos using temporal attention filters," in *Proc. of the 31st AAAI Conf. on Artificial Intelligence*, pp. 4247-4254, San Francisco, CA, USA, 4-7 Feb. 2017.

[23] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition," in Proc. *IEEE Winter Conf. on Applications of Computer Vision*, pp. 177-186, Santa Rosa, CA, USA, 24-31 Mar. 2017.

[24] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Trans. on Industrial Electronics,* vol. 66, no. 12, pp. 9692-9702, Dec. 2019.

[25] P. Zhen, X. Yan, W. Wang, H. Wei, and H.-B. Chen, "A highly compressed accelerator with temporal optical flow feature fusion and tensorized LSTM for video action recognition on terminal device," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,* vol. 42, no. 10, pp. 3129-3142, Oct. 2023.

[26] M. Hasan and A. K. Roy-Chowdhury, "Incremental activity modeling and recognition in streaming videos," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 796-803, olumbus, OH, USA, 23-28 Jun. 2014.

[27] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. of the Computer Vision,* pp. 494-507, Crete, Greece, 5-11 Sept. 2010.

[28] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in Proc. *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3169-3176, olorado Springs, CO, USA, 20-25 Jun. 2011.

[29] S. Sharma, R. Kiros, and R. Salakhutdinov, *Action Recognition Using Visual Attention*, arXiv Preprint arXiv:1511.04119, 2015.

[30] J. Cho, M. Lee, H. J. Chang, and S. Oh, "Robust action recognition using local motion and group sparsity," *Pattern Recognition,* vol. 47, no. 5, pp. 1813-1825, May 2014.

[31] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, *Action Recognition with Image Based CNN Features*, arXiv Preprint arXiv:1512.03980, 2015.

[32] A. Javidani and A. Mahmoudi-Aznaveh, "Learning representative temporal features for action recognition," *Multimedia Tools and Applications,* vol. 81, no. 3, pp. 3145-3163, Jan. 2022.

[33] J. Lu, R. Xu, and J. J. Corso, "Human action segmentation with hierarchical supervoxel consistency," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3762-3771, Boston, MA, USA, 7-12 Jun. 2015.

[34] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 759-768, Boston, MA, USA, 7-012 Jun. 2015.

[35] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding,* vol. 150, pp. 109-125, Sept. 2016.

[36] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. on Computer Vision*, pp. 3192-3199, Sydney, Australia, 1-8 Dec. 2013.

[37] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Proc. European Conf. on Computer Vision*, pp. 581-595, Zurich, Switzerland, 6-12 Sept. 2014.

عباس همدونی اصلی در سال ۱۳۶۹ مدرک کارشناسی مهندسی برق الکترونیک را از دانشگاه صنعتی خواجه نصیرالدین طوسی دریافت نمود. وی تحصیلات خود را در مقطع کارشناسی ارشد مهندسی برق قدرت در دانشگاه تهران ادامه داد و سپس در سال ۱۴۰۳ موفق به اخذ درجه دکتری در رشته مهندسی برق از دانشگاه بوعلی سینا همدان گردید. نام‌برده از سال ۱۳۸۳ تاکنون به‌عنوان عضو هیأت علمی در مؤسسه آموزش عالی جهاد دانشگاهی همدان به فعالیت مشغول است. زمینه‌های پژوهشی مورد علاقه ایشان شامل سیستم‌های قدرت، پخش بار، بهینه‌سازی و کاربرد روش‌های محاسباتی در مهندسی برق می‌باشد.

شیما جاویدانی مدرک کارشناسی خود را در رشته ریاضیات محض در سال ۱۳۸۰ از دانشگاه بوعلی سینا همدان دریافت کرد و در همان دانشگاه تحصیلات خود را در مقطع کارشناسی ارشد در همان رشته به پایان رساند. وی از سال ۱۳۸۵ به‌عنوان عضو هیأت علمی در مؤسسه آموزش عالی جهاد دانشگاهی همدان به تدریس و پژوهش مشغول است. حوزه‌های تحقیقاتی مورد علاقه ایشان شامل جبر خطی، محاسبات نرم و کاربردهای ریاضی در هوش مصنوعی و علوم داده می‌باشد.

علی جاویدانی در سال ۱۳۹۴ مدرک کارشناسی مهندسی کامپیوتر را از دانشگاه صنعتی اصفهان دریافت نمود و در سال ۱۳۹۷ کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی را از دانشگاه شهید بهشتی اخذ کرد. وی در سال ۱۳۹۸ به‌عنوان دستیار پژوهشی در دانشگاه کوییز کانادا به فعالیت علمی پرداخت و سپس در سال ۱۴۰۴ موفق به دریافت درجه دکتری تخصصی مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه تهران گردید. زمینه‌های پژوهشی مورد علاقه ایشان شامل یادگیری ماشین و یادگیری عمیق، پردازش تصویر و ویدیو و یادگیری خودنظارتی بازنمایی‌ها است. وی هم‌اکنون در دانشگاه بوعلی سینا همدان فعالیت دارد.