مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال دهم، شماره اول، بهار ۱۴۰۲

ارائه یک روش دو جریانه مبتنی بر ویژگی‌های مکمل سنتی و عمیق برای تشخیص فعالیت انسان در ویدئو

عاطفه مرادیانی[۱]، محسن رمضانی[۲]، فردین اخلاقیان طاب[۲]، رحمت الله میرزایی[۳]

چکیده

تشخیص فعالیت انسان، امروزه به‌عنوان یک حوزه مهم در کاربردهای مختلفی مورد استفاده قرار گرفته است و مورد توجه بسیاری از محققان حوزه بینایی ماشین است تا بتوانند فعالیت اجرا شده در یک ویدئو را با دقت بالا طبقه‌بندی نمایند. در این مقاله یک روش دو جریانه با ساختاری جدید معرفی می‌گردد که از دو ویژگی مکانی در هر دو جریان استفاده می‌کند به‌گونه‌ای که این ویژگی-ها بتوانند به پوشش نقاط ضعف همدیگر بپردازند. استفاده از این ساختار در نهایت می‌تواند به صورت دقیق‌تری منجر به پیش‌بینی برچسب فعالیت شود. در جریان اول ضرایب موجک با چندریزگی مناسب و در جریان دیگر ویژگی‌های عمیق از قاب‌ها استخراج می‌شوند. ویژگی‌های حاصل در دو نقشه ویژگی‌های مکانی قرار می‌گیرند و با استفاده از یک شبکه عمیق جدید تغییرات زمانی در نقشه‌ها یاد گرفته می‌شوند و با ترکیب اطلاعات طبقه‌بندی دو جریان برچسب نهایی تعیین می‌گردد. دقت روش پیشنهادی روی ۳ مجموعه داده واقعی UCFYT، UCF-Sport، و JHMDB برابر با ۹۸,۷، ۹۹,۸۳ و ۹۲,۸۶ بوده که عملکرد روش به طور میانگین نسبت به بهترین روش معرفی شده قبلی ۴,۶ درصد بهتر است.

کلیدواژه‌ها

تشخیص فعالیت انسان، روش دو جریانه، نقشه ویژگی‌های مکانی، شبکه عمیق، ترکیب طبقه‌بندها

۱ مقدمه

تشخیص فعالیت انسان به‌عنوان یکی از زمینه‌های چالش برانگیز در بینایی ماشین همواره مورد توجه محققان بوده است و به طور گسترده در کاربردهای گوناگونی مانند جستجوی ویدئو، تعامل انسان و ماشین و نظارت مورد استفاده قرار گرفته است [۱]. فعالیت انسان در حقیقت مجموعه‌ای از حرکت‌ها است که در قاب‌های متوالی اجرا می‌شود و این حرکت‌ها را می‌توان با استفاده از تغییرات مکانی-زمانی نقاط مدل‌سازی کرد [۲]. مدل سازی در حقیقت تولید یک بردار عددی است که نمایانگر فعالیت مورد نظر است و توسط یک روش بازنمایی حاصل می‌شود. برای بازنمایی فعالیت به استخراج ویژگی‌های مکانی و بررسی تغییرات آنها درگذر زمان پرداخته می‌شود که ویژگی حاصل را ویژگی‌های مکانی-زمانی می‌نامند. این ویژگی‌ها در بازنمایی فعالیت انسان مورد استفاده قرار می‌گیرند تا پیش‌بینی برچسب فعالیت در حال اجرا بر اساس مدل حاصل شده انجام شود که از اطلاعات آن برای پشتیبانی برنامه‌های کاربردی مختلف استفاده می‌گردد. استخراج و استفاده از ویژگی‌ها برای بازنمایی فعالیت ، چالش اصلی در

این مقاله در اسفندماه ۱۴۰۰ دریافت، در خردادماه ۱۴۰۱ بازنگری و در تیرماه همان سال پذیرفته شد.

[۱] دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه کردستان
ایمیل: atefeh.moradyani@uok.ac.ir
[۲] گروه علوم کامپیوتر دانشگاه کردستان
رایانامه: {f.akhlaghian , m.ramezani}@uok.ac.ir
[۳] گروه مهندسی برق، دانشگاه کردستان
رایانامه: r.mirzaei@uok.ac.ir

نویسنده مسئول: رحمت‌الله میرزایی
dorl.net/dor/20.1001.1.23831197.1402.10.1.2.0

تحقیقات مختلف بوده است که با توجه به نوع ویژگی مورد استفاده، چالش‌های متفاوتی پیش روی محققان بوده است.

ویژگی‌های مکانی-زمانی مورد استفاده در کارهای پیشین را می‌توان به دو نوع ویژگی سراسری و محلی دسته بندی کرد. ویژگی‌های سراسری به دنبال نشان دادن شکل، ظاهر یا ژست بدن در قاب هستند [۳]. با توجه به اینکه ویژگی‌های سراسری می‌توانند تغییرات ظاهری بدن در طول اجرای فعالیت را مدل نمایند، لذا معنای فعالیت را در مدل نهایی نگهداری می‌نمایند. در حقیقت این ویژگی‌ها فعالیت را به عنوان مجموعه‌ای از ژست‌های بدن تعریف می‌کنند که در تفکیک فعالیت‌هایی که ظاهر بدن در آنها به اندازه کافی متفاوت است، عملکرد خوبی دارد. این دسته از ویژگی‌ها به مسائلی مانند انسداد، تغییر زاویه دید و غیره حساس هستند. بنابراین، ویژگی‌های محلی به‌عنوان جایگزینی برای ویژگی‌های سراسری معرفی شده و مورد استفاده قرار گرفته‌اند که در برابر چنین مواردی پایداری بیشتری از خود نشان می‌دهند [۴]. ویژگی‌های محلی به بازنمایی تغییرات پیکسل‌ها در نواحی خاص می‌پردازند [۵]. در واقع این ویژگی‌ها حرکات مهم در درون ویدئو را مدل کرده و آنها را برای بازنمایی نهایی فعالیت مورد استفاده قرار می‌دهند. اما در این ویژگی‌ها شکل و ژست بدن انسان به‌عنوان عاملی مهم برای تفکیک فعالیت‌هایی با ژست‌های متفاوت اما حرکات محلی مشابه، در نظر گرفته نمی‌شود . به طور کلی ویژگی‌های سراسری و محلی استفاده شده در روش‌های مختلف را می‌توان تحت عنوان ویژگی‌های سنتی شناخت.

در روش‌هایی که از ویژگی‌های سنتی استفاده می‌کنند فرآیندی سه مرحله‌ای شامل استخراج ویژگی، بازنمایی ویژگی و طبقه‌بندی انجام می‌شود [۶]. روش‌های بازنمایی فعالیت انسان با استفاده از ویژگی‌های مکانی-زمانی سنتی معمولا بردارهایی با ابعاد بالایی تولید می‌کنند و خواص معنایی[۱] حرکت‌ها هم مورد استفاده قرار نگرفته‌اند [۵]. علاوه بر این روش‌های بازنمایی روش‌هایی دیگر هم معرفی شده‌اند که مبتنی بر ویژگی‌های عمیق هستند [۱]. در روش‌های تشخیص فعالیت مبتنی بر یادگیری عمیق، سه مرحله‌ی استخراج ویژگی، بازنمایی، و طبقه‌بندی بطور همزمان انجام می‌شود. روش‌های معرفی شده مبتنی بر ویژگی‌های عمیق برای استخراج ویژگی، بازنمایی، و طبقه‌بندی فعالیت ، معماری‌ها و ترکیب‌های مختلفی از شبکه‌های عصبی کانولوشنی[۲]، شبکه‌های - بازگشتی[۳]، و خودرمزنگارها[۴] را مورد استفاده قرار می‌دهند [۱] و - [۷].

اما روش‌های مبتنی بر شبکه‌های عمیق به منظوررسیدن به دقت قابل قبول به داده‌های آموزشی زیاد و زمان آموزش قابل توجه نیاز دارند. بعلاوه ویژگی‌های حاصل از شبکه‌های عمیق، تغییرات معنایی فعالیت‌ها را بخوبی در نظر نمی‌گیرند و همین می‌تواند باعث ایجاد هم پوشانی بین دسته‌های فعالیتی شود که در بخشی از حرکات با هم شباهت دارند [۸]. به همین دلیل روش‌هایی با عنوان دو جریانه[۵] در چند سال گذشته مورد توجه قرار گرفته‌اند که در آن‌ها دو نوع ویژگی به صورت مستقل در دو جریان استخراج شده و برای ایجاد یک بازنمایی مناسب مورد استفاده قرار گرفته‌اند [۲] و [۸]. در تعدادی از روش‌ها سعی شده با بهره‌گیری از مزیت ویژگی‌های سنتی و عمیق، در دو جریان به بازنمایی مناسبی رسید. در واقع هدف اصلی آن است که در طبقه‌بندی نهایی بازنمایی حاصل از دو جریان را مد نظر قرار داد [۹] و [۱۰]. لازم به ذکر است که بهبود حاصل شده در دقت نهایی مناسب و چشمگیر به نظر نمی‌رسند که نشان از ویژگی‌های استفاده شده نامناسب دارد [۱۱] و [۱۲] تا [۱۳].

در این مطالعه با توجه به نتایج بهتر روش‌های دو جریانه، از یک چارچوب دوجریانه با ویژگی‌های جدید برای شناسایی فعالیت انسان استفاده شده است. تلاش بر آن بوده که در یک جریان از یک ویژگی مکانی سراسری که توانایی بازنمایی تغییرات ضبط شده در قاب‌ها را داشته باشد (علی الخصوص تغییرات معنایی) بهره برده شود و در جریان دیگر ویژگی‌های استخراج شده توسط یک شبکه عمیق برای بازنمایی فعالیت مورد استفاده قرار گیرند تا این دو بازنمایی در دو جریان در کنار هم به تشخیص فعالیت دقیقی منجر گردند. در حقیقت، این ویژگی‌ها باید بصورت مکمل هم بوده و در کنار هم بازنمای کامل اطلاعات فعالیت باشند تا با استفاده از یک طبقه‌بند مناسب به تشخیص آنها پرداخت. در این روش از دو نقشه ویژگی[۶] برای بازنمایی اطلاعات مکانی قاب‌های متوالی استفاده می‌شود که شامل نقشه ویژگی‌های سنتی و نقشه ویژگی‌های عمیق هستند. این نقشه‌ها به ترتیب از ویژگی‌های حاصل از ضرایب موجک[۷] و شبکه عمیق ResNet ایجاد می‌گردند.

در یک جریان ضرایب موجک برای استخراج ویژگی‌های سراسری قاب‌ها استفاده می‌شود. ضرایب موجک به دلیل شباهت به سیستم بینایی انسان و داشتن نمایش چندریزگی[۸] می‌توانند به بازنمایی مکانی بهتر فعالیت منجر شود که با در نظر داشتن معنای فعالیت بهبود تشخیص را به دنبال خواهد داشت. معنای فعالیت را می‌توان توالی حرکت‌هایی تعریف کرد که منجر به آن فعالیت مشخص شده است. در حقیقت زمانی که یک سیستم تشخیص فعالیت همپوشانی بین دسته‌های فعالیتی که هیچ شباهتی به هم ندارند را کاهش دهد و همپوشانی‌ها متمایل به دسته‌های فعالیتی دارای دنباله حرکت‌های مشابه شوند، توانسته معنای فعالیت‌ها را درک و مدل کند. به عبارت دیگر، هدف تشخیص حرکت‌های اجرا شده در طول فعالیت مورد نظر و توانایی تشخیص آنها است. در کنار

---

۱ Semantical features
۲ Convolutional neural network (CNN)
۳ recurrent
۴ Autoencoder
۵ Two-stream
۶ Feature map
۷ Wavelet
۸ Multi-resolution

بررسی سراسری قاب‌ها، بررسی تغییرات محلی نیز در بازنمایی دقیق فعالیت انسان موثر است که ساختار کانولوشنی شبکه‌های عمیق می‌توانند این تغییرات را در مدل نهایی لحاظ نماید. در حقیقت ویژگی‌های حاصل در هر جریان مدل فعالیت را تشکیل می‌دهند که در قالب نقشه‌های ویژگی اطلاعات مکانی قاب‌های ویدئو را نگهداری می‌کنند و مدل مکانی-زمانی نهایی از آنها استخراج خواهد شد. سپس یک شبکه با ساختاری جدید به مدل کردن تغییرات زمانی در نقشه‌های هر جریان می‌پردازد و در لایه آخر طبقه‌بندی انجام خواهد شد. نتایج خروجی این شبکه به ازای هر دو جریان برای طبقه‌بندی نهایی با استفاده از روش ترکیب بیشینه با هم ترکیب خواهند شد تا بتوان به برچسبی با اطمینان بالاتر دست یافت.

به‌طور خلاصه می‌توان دستاوردهای این مقاله را بصورت زیر بیان نمود:

- ارائه یک ساختار دو جریانه با ایجاد نقشه ویژگی‌های مکانی از ویژگی‌های سنتی و عمیق. در این ساختار دو جریانه از ویژگی-های مکانی دست‌ساز و عمیق به موازات هم استفاده می‌شود تا با بهره‌گیری از مزیت‌های هر دو نوع ویژگی برای بازنمایی فعالیت انسان، دقت و اطمینان طبقه‌بندی نهایی افزایش یابد. در این روش دو نقشه ویژگی مکانی از ویژگی‌های سنتی (ضرایب موجک قاب‌ها) و عمیق (خروجی شبکه ResNet برای قاب‌ها) در دو جریان ایجاد می‌شود تا به منظور یادگیری تغییرات زمانی قاب‌ها مورد استفاده قرار گیرند.
- بازنمایی مکانی-زمانی عمیق و طبقه‌بندی در جریان‌ها به صورت مستقل. در این روش بازنمایی مکانی-زمانی فعالیت با استفاده از نقشه ویژگی‌های مکانی هر جریان به صورت مستقل انجام خواهد شد تا برخلاف روش‌هایی که از ترکیب دو جریان در مرحله استخراج ویژگی استفاده می‌کنند، مانع از حذف جزییات ویژگی‌ها برای بازنمایی گردد. بدین‌منظور شبکه‌ی جدیدی برای ایجاد بازنمایی مکانی-زمانی فعالیت در هر جریان معرفی شده است که مدلی از تغییرات مکانی و زمانی ویژگی‌ها تولید می‌کند. به عبارت دیگر، به جای استخراج ویژگی‌های زمانی از ویژگی‌های سنتی (مانند روش-های گذشته)، ویژگی‌های زمانی در دو جریان با استفاده از دو ویژگی مکانی سنتی و عمیق مجزا تولید می‌شوند تا اطلاعات بیشتری برای تفکیک بهتر فعالیت‌های دارای حرکات متفاوت وجود داشته باشند. در لایه آخر همین شبکه نیز طبقه‌بندی فعالیت ورودی بر اساس بازنمایی حاصل در هر جریان انجام خواهد شد. نهایتا پیش‌بینی نهایی برچسب با ترکیب طبقه-بندی جریان‌ها صورت می‌گیرد.
- انجام آزمایش‌های متعدد بر روی مجموعه‌ای از ویدئوهای واقعی. در این پژوهش، از مجموعه داده‌هایی که حاوی ویدئوهایی واقعی در دسته‌های فعالیتی متعدد هستند برای ارزیابی روش پیشنهادی و مقایسه آن با سایر روش‌ها استفاده شده است. مقایسه‌ها نشان می‌دهند که روش پیشنهادی توانسته است به دقت مناسبی نسبت به پژوهش‌هایی که اخیرا ارائه شده‌اند دست یابد.

در ادامه این مقاله، در فصل دو مطالعات پیشین مورد بررسی و تجزیه و تحلیل قرار گرفته اند. در فصل سوم، روش پیشنهادی با جزییات کامل معرفی شده است. فصل چهارم این مقاله به بررسی نتایج آزمایش‌ها و مقایسه روش پیشنهادی با سایر روش‌ها پرداخته است. فصل پنجم نیز به جمع‌بندی و نتیجه‌گیری این مقاله مربوط می‌شود.

۲ پیشینه تحقیق

شناسایی فعالیت انسان[۱] به عنوان یک زمینه تحقیقاتی کاربردی در حوزه های مختلف مانند موتورهای جستجو، دوربین‌های نظارتی، نظارت بر حرکت بیمار و غیره مورد استفاده قرار گرفته است - [۱۴]. در این کاربردها، حرکات بدن انسان اساس تجزیه و تحلیل و تشخیص است [۹]. در حقیقت، حرکت اجزای بدن در این کاربردها باید مدل شود و از مدل برای تشخیص و بازیابی آنها استفاده می‌شود. در تشخیص فعالیت انسان برچسب فعالیت بر اساس مدل ایجاد شده به ویدئوی ورودی نسبت داده می شود، در حالی که در بازیابی ویدئو مدل ایجاد شده برای یک ویدئوی ورودی به عنوان پرسمان[۲] با مدل سایر ویدئوها مقایسه می‌شود تا شبیه‌ترین ویدئوها از لحاظ فعالیت ضبط شده به پرسمان یافته شوند [۱۴]، [۱۵]، و [۱۶]. در زمینه شناسایی فعالیت انسان دو دسته از روش‌ها تا کنون مورد استفاده قرار گرفته‌اند که عبارتند از روش‌های تک جریانه[۳] و روش های دو جریانه[۴] [۲] تا [۱] و [۵]. [۵].

روشهای تک جریانه از یک ویژگی برای شناسایی فعالیت انسان استفاده می‌کنند. به منظور بازنمایی فعالیت‌ها، از ویژگی های[۵] مختلفی استفاده می‌شود که می توان آنها را به ویژگی‌های سنتی[۶] و ویژگی‌های حاصل از شبکه‌های عمیق[۷] تقسیم کرد [۴] تا [۵] و [۲]. ویژگی‌های سنتی خود شامل دو نوع ویژگی سراسری و محلی هستند که هم برای بازیابی و هم برای تشخیص فعالیت مورد استفاده قرار گرفته‌اند. از جمله مهمترین روش‌های استخراج ویژگی‌های محلی می توان به Dollar [۱۷]، SIFT سه بعدی، Harris سه بعدی و ضرایب موجک[۵] اشاره کرد. روش‌هایی هم برای بازنمایی فعالیت بر مبنای ویژگی‌های استخراج شده مورد

---

۱ Human Action Recognition (HAR)
۲ Query
۳ Single Stream
۴ Two Stream
۵ Feature
۶ Traditional
۷ Deep

استفاده قرار گرفته‌اند که از جمله آنها HOG[۱] [۴] و [۱۴]، بردار برآیند [۱۸] و روش فراکتال [۱۸] را می توان نام برد.

از سوی دیگر، شبکه‌های عمیق که به دنبال استخراج ویژگی‌های فعالیت یا طبقه‌بندی فعالیت با استفاده از ویژگی‌های استخراج شده هستند، به دلیل نیاز به آموزش تا کنون در کاربرد بازیابی فعالیت مورد استفاده قرار نگرفته‌اند و فقط در کاربرد تشخیص فعالیت استفاده شده‌اند. به‌عنوان مثال Zare و همکارانش [۶] از ضرایب موجک به‌عنوان ویژگی‌ها استفاده کرده‌اند و هر ویدئو را در یک VSTM[۲] خلاصه نموده‌اند. آنها سپس از یک CNN[۳] در مرحله طبقه‌بندی برای پیش‌بینی برچسب فعالیت ذخیره شده در ویدئو بهره برده‌اند. در روش‌هایی که از شبکه‌های عمیق برای استخراج ویژگی و تعیین برچسب ویدئوی ورودی استفاده می‌کنند، استخراج، بازنمایی و طبقه‌بندی به صورت همزمان انجام می‌شود. در این راستا Khan و همکارانش [۱۹] با استفاده از یک شبکه LSTM شناسایی فعالیت را انجام داده‌اند. روش ارائه شده در این مطالعه یک رویه مبتنی بر LSTM به همراه یک شبکه عصبی کانولوشنی (DCNN) است که به‌طور انتخابی برروی ویژگی‌های مؤثر در قاب ورودی تمرکز می‌کند تا با استفاده از این ویژگی‌ها فعالیت‌های مختلف انسان را تشخیص دهد. در لایه‌های DCNN این ساختار، برای استخراج ویژگی‌های متمایز برجسته، از بلوک‌های residual استفاده می‌شود که اطلاعات بیشتری نسبت به یک لایه کم عمق نگه می‌دارند که البته نیاز به زمان اجرای طولانی دارد.

در مرجع [۷] نیز نویسندگان به دنبال شناسایی فعالیت با استفاده از ویژگی‌های استخراج شده از خودرمزنگارهای انباشته (SAE)[۴] بوده‌اند که مقدار پارامترهای ساختاری SAE را با استفاده از الگوریتم بهینه سازی کلونی زنبور عسل، الگوریتم ژنتیک، الگوریتم تکامل دیفرانسیل، و الگوریتم بهینه سازی ازدحام ذرات تعیین کرده‌اند. از آنجایی که ویدئوها علاوه بر ابعاد مکان، بعد زمان هم دارند و سه بعد را تشکیل می‌دهند، در برخی مطالعات از شبکه‌های عصبی کانولوشنی سه بعدی برای استخراج ویژگی و پیش‌بینی برچسب فعالیت استفاده شده است [۲۰]. شبکه های CNN سه بعدی یک شبکه عصبی کانولوشنی سه بعدی با محاسبات و متغیرهای زیاد هستند که تعدادی از محققان در صدد کاهش بار محاسباتی این روش‌ها در کاربرد تشخیص فعالیت بوده‌اند. در این راستا Fan و همکارانش [۲۱] Bottleneck residual block را معرفی کرده‌اند که با استفاده از آن استخراج ویژگی‌ها و طبقه‌بندی فعالیت با بار محاسباتی کمتر انجام می‌شود. البته این روش نیاز به سخت افزار خاص دارد و قابل اجرا روی تمام سیستم‌ها نیست.

از سوی دیگر روش‌های دوجریانه برای شناسایی فعالیت انسان از دو نوع ویژگی به صورت همزمان در دو جریان مستقل استفاده می‌کنند [۲۲]. در کارهای مختلف از ویژگی‌های متفاوتی در هر جریان استفاده شده است که شامل انواع ویژگی‌های سنتی محلی یا سراسری، و ویژگی‌های عمیق می‌باشند [۲۳] تا [۲۴]. در اکثر روش‌های دو جریانه حتما از ویژگی‌های حاصل از شبکه‌های عمیق در یک جریان استفاده می‌شود تا از توانایی شبکه‌های عمیق در نگاه همزمان و جامع به تغییرات بهره ببرند [۲۵] تا [۲۶] و [۲۷]. از آنجا که این روش‌ها در توجه به تغییرات محلی که می‌تواند در تفکیک فعالیت‌هایی که قالب بدن در طول اجرای آنها مشابه است موثر باشند، عملکرد خیلی مناسبی نداشته باشند [۲۸] تا [۱۶]، لذا در روش‌های دوجریانه، ویژگی‌های سنتی به‌عنوان ویژگی مورد استفاده در جریان دوم مورد استفاده قرار می‌گیرند تا در نهایت به نتایج مناسبی برسند.

بطور کلی، در این روش‌ها یک جریان ویژگی‌های مکانی حاصل از شبکه های عمیق را در نظر می‌گیرد و یک جریان ویژگی‌های زمانی حاصل از ویژگی سنتی را استخراج خواهد کرد تا با ترکیب آنها بازنمایی مکانی-زمانی فعالیت حاصل شود. در این راستا در مراجع [۲۹] تا [۳۱] برای طبقه‌بندی فعالیت از یک مدل دوجریانه استفاده شده است که در یک جریان ویژگی‌های زمانی و در جریان دیگر ویژگی‌های مکانی استخراج می‌شود. همچنین Dai و همکارانش [۲] یک رویکرد دو جریانه با استفاده از ویژگی‌های MHI و CNN معرفی کرده‌اند. در این ساختار بعد از طبقه‌بندی فعالیت با استفاده از ویژگی‌های هر جریان، برچسب نهایی با توجه به بیشینه امتیاز حاصل شده برای دسته‌های فعالیت در هر جریان تعیین شده است. در کار [۸] یک مدل دوجریان دیگر معرفی شده است که در آن با استفاده از یک شبکه CNN ویژگی‌های مکانی قاب‌های RGB استخراج می‌شود و در جریان دیگر از این روش، جریان نوری[۵] جهت استخراج ویژگی‌های زمانی مورد استفاده قرار گرفته است. در این روش نیز مانند سایر روش‌های معرفی شده و مشابه، طبقه‌بندی فعالیت در هر جریان انجام شده است که بیشینه امتیاز حاصل شده برای دسته‌های فعالیت تعیین کننده برچسب نهایی ویدئوی ورودی خواهد بود.

در برخی روش‌های دوجریانه از شبکه‌های عمیق در هر دو جریان استفاده می‌شود که به‌عنوان مثال Singh و همکارانش [۲۴] یک مدل دو جریانه معرفی کرده‌اند که در هر دو جریان دو شبکه عمیق مورد استفاده قرار گرفته است. این شبکه‌ها که دو نوع شبکه Residual هستند به استخراج ویژگی‌های مکانی قاب‌های RGB می‌پردازند و ویژگی های زمانی نیز بر اساس توالی قاب‌ها استخراج می‌شود. برخی از ساختارهای دو جریانه جدید که از ویژگی‌های حاصل از شبکه‌های عمیق استفاده می‌کنند، برای ایجاد مدل مکانی-زمانی نهایی از شبکه LSTM استفاده می‌کنند که در نهایت

[۱] Histogram of oriented gradients
[۲] Video Spatiotmporal Mapping
[۳] Convolutional neural network
[۴] Stacked Autoencoder
[۵] Optical Flow

شناسایی فعالیت انسان را به دنبال خواهد داشت. به عنوان مثال Yuxuan و همکارانش [۲۰] سعی در طراحی و پیاده سازی یک مدل دو جریان با استفاده از LSTM برای استخراج هر دو ویژگی مکانی و زمانی در قاب های RGB دارند و یک شبکه Dense Net را در جهت استخراج ویژگی‌های زمانی پیاده سازی نموده‌اند تا دقت تشخیص را بهبود ببخشند.

در روشی دیگر Latah و همکارانش [۳۲] یک مدل دوجریان با استفاده از شبکه CNN سه بعدی در هر دو جریان ارائه داده‌اند که ویژگی‌های مورد استفاده از جریان نوری و توالی قاب‌های RGB استخراج شده‌اند. همچنین در روشهایی از چند جریان[۱] جهت استخراج ویژگی و طبقه‌بندی فعالیت انسان استفاده شده است. در این زمینه می‌توان به روش ارائه شده در مرجع [۲۱] توسط Zong وهمکارانش اشاره کرد. در این مطالعه، یک ساختار برای شبکه ResNet دارای چند جریان مبتنی بر برجستگی حرکت[۲] برای تشخیص فعالیت پیشنهاد شده است. مدل MSM-ResNet از سه جریان تعاملی شامل جریان ظاهر ، جریان حرکت و جریان برجستگی حرکت تشکیل شده است. جریان ظاهر و جریان حرکت به ترتیب وظیفه ثبت اطلاعات ظاهر و اطلاعات حرکت را بر عهده دارند، در حالی که جریان برجسته حرکت مسئول ثبت اطلاعات حرکتی برجسته است.

در مطالعه دیگری یک روش دو جریانه معرفی شده است که برخی نواحی ویدئو در آن مورد بررسی قرار گرفته‌اند و کل قاب برای استخراج ویژگی مورد استفاده قرار نمی‌گیرد [۳۳]. در جریان اول، نمود[۳] بدن در ناحیه مورد بررسی و در جریان دوم برجستگی حرکت در ناحیه‌ها به‌عنوان ویژگی استخراج می‌گردند. مدل نهایی مکانی-زمانی فعالیت در این روش با استفاده از یک شبکه کانولوشنی حاصل می‌شود که طبقه‌بندی فعالیت مبتنی بر این مدل خواهد بود. Ma و همکارانش [۳۴] نیز در یک روش چند جریانه از ویژگی‌های موجود در نواحی خاصی تحت عنوان نواحی مورد علاقه[۴]، که مهمترین حرکات فعالیت در آن رخ داده است، برای بازنمایی استفاده می‌کنند. در این روش، این نواحی به شبکه‌ای عمیق داده می‌شود تا خروجی آنها بعد از تجمیع توسط یک خود رمزنگار به بردار نهایی بازنمایی کننده فعالیت انسان نگاشت یابد.

۳ روش پیشنهادی

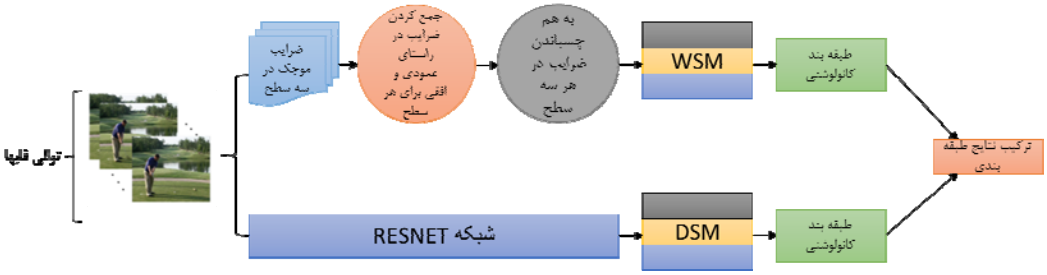
در این بخش، ما با جزئیات معماری روش پیشنهادی برای شناسایی فعالیت آشنا خواهیم شد. مشاهده خواهد شد که در هر جریان از این روش دوجریانه چگونه از نقشه[۵] ویژگی سنتی، و ویژگی‌های عمیق حاصل از شبکه ResNet استفاده خواهد شد. در این مطالعه بر خلاف سایر روشها که به استخراج ویژگی‌های مکانی در یک جریان و ویژگی‌های زمانی در یک جریان دیگر پرداخته‌اند، به ایجاد مدل زمانی از دو ویژگی مکانی مختلف در دو جریان پرداخته می‌شود تا به عملکرد نهایی مناسب با پوشش ضعف‌های بازنمایی هر جریان پرداخته شود. واضح است که فعالیت انسان در قاب‌های متوالی ذخیره می‌شود و برای بازنمایی باید به استخراج ویژگی از این قاب‌های متوالی پرداخت. ما در این مطالعه به جای استفاده از تمام قاب‌ها، از N قاب تصادفی بعنوان قاب‌های کلیدی[۶] بهره خواهیم برد. از آنجایی که فعالیت انسان در طول زمان و قاب‌های متوالی رخ می‌دهد، بررسی تغییرات عمومی رخ داده در قاب‌های متوالی می‌تواند به یک بازنمایی دقیق منجر شود. بنابراین در یک جریان به استخراج یک نقشه ویژگی مکانی توسط ضرایب موجک بعنوان ویژگی سنتی عمومی پرداخته خواهد شد و در جریان دیگر ویژگی‌های عمیق استخراج می‌گردند تا از این دو ویژگی مکانی برای ایجاد مدل مکانی-زمانی نهایی به منظور طبقه-بندی استفاده شود.

لازم به ذکر است که ضرایب موجک به دلیل بهره بردن از نمایش چند ریزگی و داشتن شباهت به سیستم بینایی انسان، بهتر از سایر ویژگی‌ها تغییرات بین قاب‌ها را برای استخراج نهایی مدل زمانی مد نظر قرار می‌دهد. از سوی دیگر ویژگی‌های عمیق که نقش مهمی درکاربردهای مختلف پردازش تصویر داشته‌اند، می‌توانند اطلاعات عمومی و محلی مناسبی را در بازنمایی هر قاب کلیدی نگهداری کنند. ویژگی‌های مکانی حاصل شده در هر جریان توسط یک شبکه کانولوشنی جدید مورد استفاده قرار می‌گیرند تا تغییرات رخ داده میان قاب‌های کلیدی را مدل نمایند. در نهایت براساس تغییرات مدل شده در شبکه، برای هر جریان یک طبقه‌بندی انجام خواهد شد تا با هم‌جوشی[۷] نتایج طبقه‌بندی جریان‌ها، برچسب نهایی ویدئوی ورودی مشخص شود.

شکل ۱ ساختار روش پیشنهادی را نمایش می‌دهد. در جریان اول نقشه ویژگی مکانی که توسط ضرایب موجک در سه سطح استخراج شده‌اند به بازنمایی ویدئوی ورودی می پردازد. در جریان دیگر ویژگی‌های عمیق قاب‌های کلیدی توسط یک شبکه ResNet استخراج شده و در نقشه مکانی عمیق نگهداری می‌گردند. شایان ذکر است که هر ردیف از نقشه‌های ویژگی‌های مکانی حاصل شده در دو جریان به یک قاب کلیدی اختصاص دارد. در ادامه نحوه‌ی ایجاد نقشه‌های مکانی موجک و عمیق به همراه شبکه کانولوشنی مورد استفاده برای یادگیری تغییرات فعالیت‌ها و طبقه‌بندی آنها بطورکامل معرفی خواهند شد.

---

[۱] Multi Stream
[۲] Motion Saliency
[۳] Appearance
[۴] Interest regions
[۵] Map
[۶] keyframe
[۷] fusion

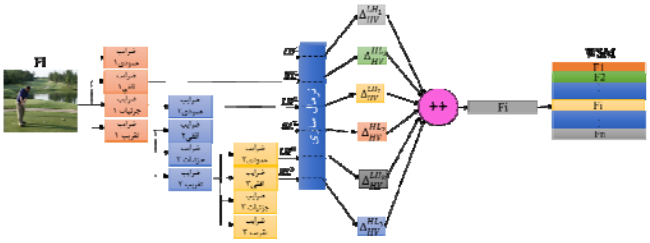


شکل۱ معماری سیستم شناسایی فعالیت انسان با استفاده از روش دو جریانه پیشنهادی

۳-۱ نقشه مکانی موجک[۱] (WSM)

شکل ۲ نحوه محاسبه نقشه ویژگی مکانی سنتی را نشان می‌دهد. نقشه ویژگی مکانی در حقیقت حاوی اطلاعات مکانی قاب‌های کلیدی است که در یک ماتریس جمع آوری شده‌اند تا از آن برای مدل‌سازی تغییر حرکت‌های فعالیت انسان استفاده شود.

این نقشه اطلاعات مکانی ضروری را حفظ می‌کند و مانع از دست رفتن اطلاعات حرکت در زمان مدلسازی فعالیت می‌شود. در مرحله اول برای هر قاب، تبدیل موجک در سه مقیاس اعمال می شود. ضرایب موجک اطلاعات ارزشمندی از گرادیان، لبه و بافت تصویررا در برمی‌گیرند. علاوه بر این، ویژگی چند مقیاس و چند ریزگی تبدیل موجک باعث می‌شود که ویژگی حاصل در برابر تغییرات مقیاس قوی گردد. متداول‌ترین روش برای تبدیل موجک گسسته چند سطحی، تجزیه بیشتر زیر باند تقریبی در سطح بعدی است. در داده‌های دو بعدی یا تصاویر، تبدیل موجک گسسته چهار مجموعه ضریب را با استفاده از چهار ترکیب فیلترهای تجزیه موجک تولید می‌کند. این ضرایب شامل ضرایب پایین گذار(ضرایب تقریبی)، بالاگذار(ضرایب جزئیات)، عمودی و افقی هستند. برای سطوح بعدی تجزیه، تنها ضرایب تقریبی بیشتر تجزیه می‌شوند. در اینجا، نگاشت افقی-عمودی موجک از هر قاب کلیدی یک بردار ویژگی مکانی ایجاد می‌کند. ضرایب کمّی، در زیر باند HL و LH در هر مقیاس برای محاسبه نگاشت افقی-عمودی موجک استفاده می‌شود. از آنجا که زیر باندهای HH همبستگی مکانی کمتری دارند در ویژگی مکانی گنجانده نشده است. به عبارت دیگر، در مدل سازی فعالیت انسان باید به دنبال بررسی تغییرات رخ داده در قاب‌های متوالی بود که زیر باندهای HH که عمدتا به لبه‌های مورب می‌پردازند، نمی‌توانند تغییرات قابل توجهی از حرکات در فعالیت‌ها را مدل نمایند. در این کار فرض می‌شود که ضرایب کوچک نزدیک صفر نویز باشند و باید برای ادامه‌ی کار اثرات آنها را کم کرد. در این راستا ضرایب موجک حاصل شده از هر زیرباند LH^i و HL^i در سطح iام را نرمال سازی کرده و به بازه صفر تا یک می‌بریم تا همه تغییرات در زیر باندهای مختلف در نظر گرفته شوند. در نهایت کم اثر کردن مقادیر نویز در محاسبات آتی نیز حاصل می‌گردد. لازم به ذکر است که مقدار i می‌تواند ۱، ۲ یا ۳ باشد.



شکل۲ مراحل تولید نقشه مکانی موجک. i در HL^i نشان دهنده شماره مقیاس است.

. با توجه به مرجع [۵] بردارهای نگاشت افقی-عمودی زیر باندهای نرمال شده‌ی LH و HL به صورت زیر محاسبه می‌گردند:

$$\Delta_{HV}^{HL_i} = \sum_{y=0}^{m_j}(HL_i(x,y)) \oplus \sum_{x=0}^{n_j}(HL_i(x,y)) \qquad (۱)$$

$$\Delta_{HV}^{LH_i} = \sum_{y=0}^{m_j}(LH_i(x,y)) \oplus \sum_{x=0}^{n_j}(LH_i(x,y)) \qquad (۲)$$

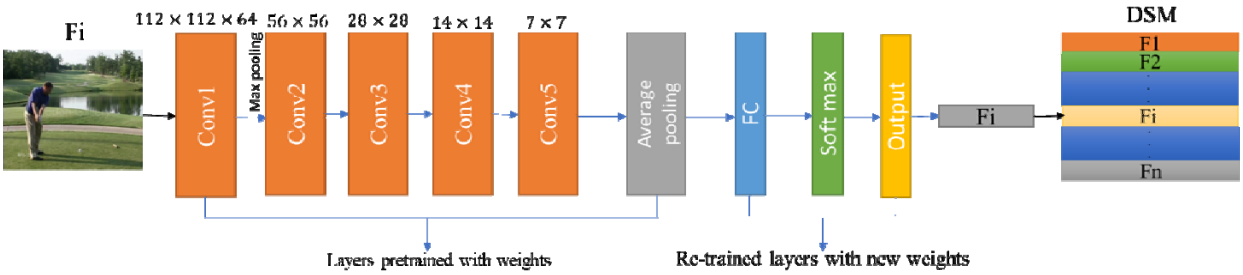
که $\Delta_{HV}^{HL_i}$ و $\Delta_{HV}^{LH_i}$ به ترتیب نمایش افقی-عمودی زیر باندهای HL^i و LH^i را نشان می‌دهند. بعلاوه m_j و n_j تعداد سطرها و ستون‌های زیرباندها در سطح jام هستند و عملگر $\oplus$ نیز ترکیب بردارها را نمایش می‌دهد.

لازم به ذکر است که ترکیب بردارها همان جمع مقادیر متناظر بردارها است. نهایتا نگاشت افقی-عمودی حاصل از ضرایب موجک قاب کلیدی jام در بردار ویژگی F_j قرار خواهند گرفت. بدین منظور، بردارهای $\Delta_{HV}^{Z_i}$ با عملگر ++ به هم چسبانده می‌شوند که Z_i نشان دهنده HL^i و LH^i است. به عبارت دیگر:

$$F_j = \Delta_{HV}^{LH_1} ++ \Delta_{HV}^{HL_1} ++ ... ++ \Delta_{HV}^{LH_3} ++ \Delta_{HV}^{HL_3} \qquad (۳)$$

---

[۱] Wavelet Spatial Map



شکل۳ ساختار شبکه ResNet50

اندازه بردار F_j به‌دست آمده از تبدیل موجک برابر $\sum_{i=1}^{3} 2(m_j + n_j)$ است. در نهایت، نگاشت افقی-عمودی حاصل برای تمام قاب‌های کلیدی در یک نقشه مکانی موجک جمع‌آوری شده‌اند. ارتفاع نقشه مکانی موجک نشان دهنده تعداد قاب‌های کلیدی مورد استفاده و عرض آن نشان دهنده اندازه بردار نگاشت افقی-عمودی موجک است.

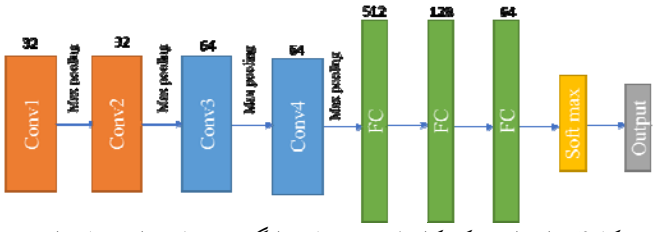
۳-۲ شبکه ResNet

در جریان دوم این روش قاب‌های کلیدی به یک شبکه عمیق برای استخراج مدل مکانی فعالیت داده می‌شوند. دلیل اصلی استفاده از ویژگی‌های مکانی عمیق عملکرد مناسب این ساختارها در وظایف مختلف بینایی ماشین است. در این مطالعه از شبکه ResNet50 استفاده شده که ساختار آن در شکل ۳ به نمایش درآمده است. شبکه مورد استفاده با وزن‌هایی مقداردهی می‌شوند که از آموزش آن توسط مجموعه داده‌ای با اندازه مناسب یعنی ImageNet استخراج گردیده‌اند.

این شبکه با درنظر گرفتن ساختار عمومی هر قاب کلیدی و اطلاعات محلی در آن قاب، به استخراج ویژگی‌های قاب می‌پردازد. ویژگی‌های حاصل شده از این جریان به اندازه کافی غالب هستند تا بتوان از آن برای بازنمایی محتوای بصری قاب‌های کلیدی استفاده کرد. استفاده از این ویژگی‌ها می‌توانند با در نظر گرفتن تغییرات محلی در کنار تغییرات عمومی به یادگیری توالی‌های پیچیده که ممکن است تغییرات عمومی تقریبا مشابهی در جریان اول داشته باشند کمک کند. به‌عنوان مثال، دو فعالیت گلف و تنیس تغییرات عمومی تقریبا مشابهی دارند که ممکن است بازنمایی جریان اول از تفکیک این دو فعالیت ناتوان باشد. این در حالیست که جریان دوم با در نظر گرفتن تغییرات محلی در کنار تغییرات عمومی می‌تواند بازنمایی بهتری از تفاوت آنها داشته باشد. ویژگی‌های حاصل از این جریان در یک نقشه مکانی به نام نقشه مکانی عمیق(DSM) ذخیره می‌گردند. اندازه نقشه مکانی عمیق برای هر قاب ورودی با ابعاد ۲۲۴×۲۲۴ ، یک بردار ۱×۱۰۰۰ است. در نهایت برای یک ویدئو با استخراج ۴۰ قاب کلیدی، یک نقشه مکانی ۴۰×۱۰۰۰ تولید خواهد شد.

۳-۳ استخراج اطلاعات زمانی و طبقه بندی

تا اینجا، در دو جریان به استخراج اطلاعات مکانی قاب‌های کلیدی هر ویدئو با استفاده از یک ویژگی سنتی عمومی و یک ویژگی عمیق پرداخته شده است تا در کنار هم یک مدل مکانی مناسبی از فعالیت انسان ایجاد گردد. در ادامه اطلاعات زمانی فعالیت را از توالی مدل مکانی قاب‌ها استخراج می‌نماییم تا با استفاده از آنها طبقه‌بندی در هر جریان انجام شود. بدین منظور یک شبکه کانولوشنی جدید را معرفی می‌نماییم که خروجی آن برای طبقه‌بندی مستقل هر جریان مورد استفاده قرار خواهد گرفت. ساختار این شبکه در شکل ۴ مشاهده می‌گردد که در مقایسه با شبکه کانولوشنی معمولی به دلیل داشتن فیلترهایی با ابعاد متناسب با ابعاد ورودی در هرلایه، ویژگیهای زمانی کامل‌تری را استخراج می‌کند.



شکل۴ ساختار شبکه کانولوشنی برای یادگیری مدل زمانی و انجام طبقه‌بندی هر جریان

این شبکه از هسته‌هایی با اندازه‌ی ۳×۳ و اندازه گام ۱ بهره می‌برد. در این شبکه از maxpooling هایی با اندازه ی ۲×۲ استفاده می‌شود که ابعاد بردار نهایی را تا حدی کوچکتر نماید و حجم محاسبات بعدی را کاهش دهد. علاوه بر این در ساختار شبکه از لایه‌ی نرمال سازی استفاده شده است تا علاوه بر بهبود عملکرد شبکه، سرعت آموزش آن را نیز افزایش دهد. همچنین از لایه‌ی Dropout نیز در شبکه استفاده شده است تا باعث جلوگیری از بیش‌برازش[1] گردد.

شبکه معرفی شده نقشه ویژگی‌های حاصل از هر جریان را به واسطه‌ی یادگیری تغییرات متوالی رخ داده در قاب‌های کلیدی به یک طبقه از فعالیت‌های انسان نگاشت می‌دهد. این کار با محاسبه حاصل ضرب نقطه‌ای فیلترهای مورد استفاده در شبکه و همچنین

[1] overfitting

لایه‌ی ورودی تولید می‌شود. پارامترهای استفاده شده در سیستم پیشنهادی در جدول۱ نشان داده شده‌اند. همانگونه که در ساختار شبکه مشخص است بعد از انجام عملیات کانولوشنی، یک لایه برای نرمال سازی قرار می‌گیرد. این شبکه با یادگیری جریان زمانی که در نقشه ویژگی‌های مکانی نهفته است، به توصیف مجموعه‌ای از حرکات در طول یک دوره زمانی با مدت زمان ثابت می‌پردازد. در واقع یادگیری حرکت با مشاهده تغییرات ویژگی‌های مکانی حاصل می‌شودکه به منظور تشخیص فعالیت انسان مورد استفاده قرار گرفته است.

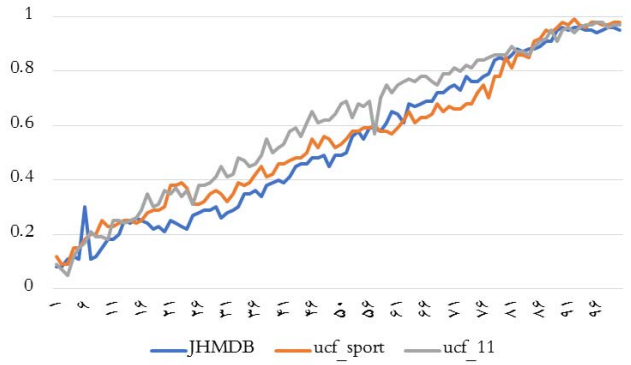
جدول۱ پارامترهای استفاده شده در روش پیشنهادی

نام لایه ها	ابعاد خروجی	اندازه هسته لایه	اندازه گام
Conv1	۳۲×۷۸۴×۴۰	۳×۳	۱×۱
BN	۳۲×۷۸۴×۴۰		
Conv2	۳۲×۷۸۲×۳۸	۳×۳	
Max	۳۲×۳۹۲×۱۹	۲×۲	
Drop	۳۲×۳۹۲×۱۹	-	-
BN	۳۲×۳۹۲×۱۹	-	-
Conv3	۶۴×۳۹۲×۱۹	۳×۳	۱×۱
BN	۶۴×۳۹۲×۱۹	-	
Conv4	۶۴×۳۸۹×۱۷	۳×۳	
Max	۶۴×۱۹۴×۸	۲×۲	
Drop	۶۴×۱۹۴×۸	-	-
BN	۶۴×۱۹۴×۸	-	-
FC	۵۱۲	-	-
Drop	۵۱۲	-	-
FC	۱۲۸	-	-
Drop	۱۲۸	-	-
FC	۶۴	-	-
Drop	۶۴	-	-
Softmax layer	تعداد دسته های کلاس	-	-

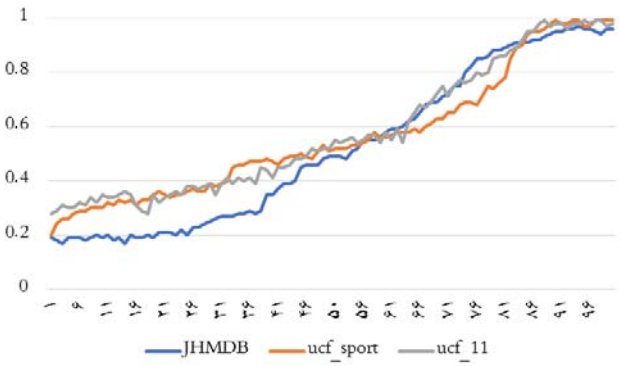
شبکه معرفی شده بعد از یادگیری تغییرات موجود در ویژگی‌های مکانی قاب‌های هر جریان به صورت مستقل، در لایه‌ی softmax به طبقه‌بندی فعالیت در هر جریان خواهد پرداخت. بنابراین در هر جریان از این روش طبقه‌بندی مبتنی بر بیشترین مقدار خروجی شبکه برای هرکدام از کلاس‌ها انجام می‌شود. برای همجوشی نتایج طبقه‌بندی انجام شده در دو جریان و به‌دست آوردن پیش‌بینی نهایی فعالیت، نمرات هر دو جریان به ازای هرکلاس با استفاده از روش‌های مختلف همجوشی، مانند مجموع، حداکثر و میانگین وزنی، می‌توانند ادغام شوند. به عبارت دیگر، فعالیتی که در هر دو جریان بیشترین امتیاز را داشته باشد به‌عنوان برچسب نهایی ویدئو انتخاب می‌گردد. در این روش برای بهره بردن از نقاط قوت هرکدام از جریان‌ها از روش همجوشی حداکثری استفاده می‌شود تا برچسب نهایی فعالیت براساس جریان دارای اطمینان بیشتر تعیین شود.

۴ نتایج تجربی

در این مقاله، نقشه‌ای برای دو ویژگی مکانی مختلف در دو جریان مستقل محاسبه می‌شود که مدل حرکات فعالیت را در بر خواهند داشت و در کنار هم بازنمایی مناسبی از فعالیت اجرا شده را ارائه می‌دهند. در عمل شناسایی، نقشه محاسبه‌شده در هر جریان به طبقه‌بند داده می‌شود تا با در نظر گرفتن تغییرات مکانی و ایجاد مدل مکانی-زمانی نهایی به پیش‌بینی برچسب هر ویدیو ورودی پرداخت. آزمایش‌ها بر روی محیط Google Colab در حالت GPU با حافظه ۸ گیگابایتی اجرا می‌شود. مانند سایر مطالعات در حوزه شناسایی فعالیت انسان، در این آزمایش‌ها ۴۰ قاب کلیدی به صورت تصادفی از هر ویدیو استخراج شده و در آزمایش‌ها مورد استفاده قرار می‌گیرند. در این آزمایش‌ها به ارزیابی دقت طبقه‌بندی بر اساس برچسب‌های پیش‌بینی‌شده پرداخته شده است. لازم به ذکر است که بهبود عملکرد طبقه‌بندی مستقیماً به ویژگی‌های محاسبه شده متکی است، و نتایج بهتر نشان از بازنمایی بهتر فعالیت با استفاده از ویژگی‌های استخراج شده دارد. همچنین، اگر همپوشانی‌ها بین فعالیت‌هایی باشند که حرکت‌ها و توالی تغییرات در راستای زمان در آنها مشابه باشند، بازنمایی معنایی به شکل مناسبی صورت گرفته است. به عبارت دیگر در مدل فعالیت‌ها معنا که همان توالی حرکت‌هاست نگهداری شده است. در این آزمایش‌ها فرآیند آموزش شبکه مورد استفاده برای یادگیری تغییرات زمانی و طبقه‌بندی فعالیت در هر جریان صد epoch تکرار شده است که نحوه همگرایی این شبکه برای نیل به دقت نهایی در شکل ۵ قابل مشاهده است.



(الف)



(ب)

شکل۵ همگرایی شبکه مورد استفاده برای یادگیری تغییرات زمانی در (الف) جریان اول (ویژگی سنتی) و (ب) جریان دوم (ویژگی عمیق)

در این بخش روش معرفی شده روی سه مجموعه داده چالشی با ویدئوهای واقعی که روش‌های قبلی عملکرد ضعیفی روی آنها داشته‌اند، یعنی مجموعه داده‌های UCF Sport، UCF YouTube (UCFYT) و JHMDB ارزیابی خواهد شد. مجموعه داده UCF-Sport شامل ۱۵۰ ویدیو واقعی از ۹ دسته فعالیت است که از صحنه‌های ورزشی ضبط شده است. بازیگران، پس‌زمینه‌ها، دیدگاه‌ها و صحنه‌های مختلفی در ویدیوهای این مجموعه داده وجود دارد. UCF YouTube به‌عنوان دیگر مجموعه داده‌ی مورد استفاده بازیگران، پس زمینه ها، دیدگاه ها، و صحنه های مختلفی دارد. این مجموعه داده شامل ۱۶۰۰ ویدیو در ۱۰ دسته فعالیت است که از فعالیت‌های واقعی در ویدیوهای YouTube ضبط شده است. فعالیت‌های این مجموعه داده شامل کلاس‌های پرتاب بسکتبال، دوچرخه‌سواری، گلف، تاب سواری، اسب‌سواری، فوتبال، تنیس، بالا پایین پریدن، پرش از تخته، و والیبال هستند. از سوی دیگر، مجموعه داده JHMDB برای ارزیابی روش پیشنهادی تشخیص فعالیت در مقایسه با روش‌های دیگر مورد استفاده قرار می‌گیرد. مجموعه داده JHMDB شامل ۹۲۳ ویدیو در ۲۱ دسته است. برای هر مجموعه داده، ۶۷ درصد از ویدیوها به عنوان داده‌ی آموزشی و ۳۳ درصد باقی مانده به عنوان داده‌ی آزمایشی در کار شناسایی در نظر گرفته می شود.

جدول۲ مقایسه نتایج روش پیشنهادی و مرجع [۱۹] روی مجموعه داده‌های UCF11 ، UCF SportوJ-HMDB

معماری مدل	UCF11	UCFSport	JHMDB
Traditional CNN+BiLSTM+RB	۸۲,۲۰	۸۳,۸۰	۷۵,۸۱
Traditional CNN+BiLSTM+RB+Attention	۸۵,۱۸	۸۷,۷۰	۷۷,۷۰
Traditional CNN+BiLSTM+RB+Attention+Center Loss	۸۵,۹۳	۸۹,۵۹	۷۷,۹۰
Dilated CNN+BiLSTM+RB	۸۹,۰۱	۹۲,۶۳	۷۸,۶۳
Dilated CNN+BiLSTM+RB+Attention	۹۶,۳۱	۹۷,۲۴	۷۹,۲۴
Dilated CNN+BiLSTM+RB+Attention+Center loss	۹۸,۳۰	۹۹,۱۰	۸۰,۲۰
Proposed method	۹۸,۷۰	۹۹,۸۳	**۹۲,۸۶**

آزمایش‌های متعددی برای ارزیابی و تأیید کارایی روش پیشنهادی اجرا شده‌اند که نتایج آنها مورد بررسی قرار می‌گیرند. همچنین روش پیشنهادی با روش‌های متعددی مورد مقایسه قرار گرفته است که بیشتر آنها روش‌های دو جریانه جدید هستند. نتایج موجود در جدول ۲ نشان می‌دهند که روش پیشنهادی نسبت به سایر روش‌های دو جریانه‌ای که فقط از ویژگی‌های عمیق بهره می‌برند عملکرد بهتری دارد.

جدول ۳ مقایسه روش پیشنهادی با روش‌های یک جریانه روی سه مجموعه داده استاندارد UCF11، UCF SportوJ-HMDB

UCF Sports		UCF11		J-HMDB	
روش‌ها	دقت (%)	روش‌ها	دقت (%)	روش‌ها	دقت (%)
Tu [33]	۹۷,۵۳	Afza [28]	۹۴,۵	Tu [33]	۷۱,۱۷
Afza [28]	۹۹,۳	Gharaee [22]	۸۹,۵	Ma [34]	۷۶,۹۰
Dai [2]	۹۸,۶	Dai [2]	۹۶,۹	Dai [2]	۷۶,۳۰
Khan [19]	۹۹,۱۰	Khan [19]	۹۸,۳۰	Khan [19]	۸۰,۲۰
روش پیشنهادی	**۹۹,۸۳**	روش پیشنهادی	**۹۸,۷۰**	روش پیشنهادی	**۹۲,۸۶**

همچنین جدول ۳ به مقایسه روش پیشنهادی با سایر روش‌هایی پرداخته است که در آنها از یک ویژگی عمیق یا سنتی برای بازنمایی فعالیت استفاده شده است. در این دو جدول بهترین نتایج با **ضخامت بیشتر** نشان داده می‌شوند. واضح است که روش معرفی شده از بهترین روش‌هایی که از یک ویژگی استفاده می‌کنند و روش‌های دو جریانه که عمدتا از ترکیب ویژگی‌های مکانی و یا زمانی مختلف بهره می‌برند عملکرد بهتری داشته است. این عملکرد بهتر نشان دهنده کامل بودن بازنمایی نهایی و پیش بینی انجام شده به دلیل استفاده از ویژگی‌های مکمل در دو جریان است.

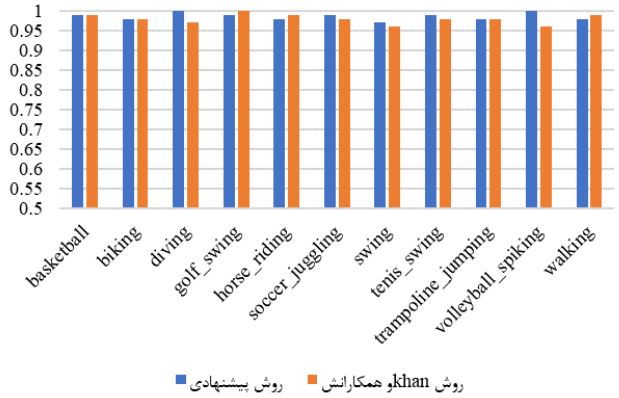
شکل ۶ که به مقایسه دقت روش پیشنهادی و روش معرفی شده توسط Khan و همکارانش (بهترین روش بعد از روش مقاله حاضر) پرداخته است، نشان می‌دهد که روش معرفی شده روی ویدیوهای ورزشی به دقت بالاتری دست یافته است. یکی از دلایل مهم این برتری آن است که این ویدیوها حرکت‌های مشابه بسیاری دارند و تشخیص آنها توسط یک سیستم ساده دشوار است. روش معرفی شده به دلیل استفاده از دو ویژگی سنتی و عمیق در کنار هم و استفاده از مطمئن‌ترین نتیجه حاصل توانسته است که بازنمایی بهتری برای پیش‌بینی نهایی حاصل نماید. مشخص است که روش پیشنهادی ویژگی‌های هر جریان را نهایتا بصورت عمیق می‌آموزد تا درمقایسه با سایر روش‌ها در شناسایی فعالیت به عملکرد مناسبی دست یابد. به‌طور کلی روش پیشنهادی روی دسته‌های مختلف مجموعه داده‌های مورد استفاده، به‌عنوان مجموعه داده‌های چالشی با تعداد کلاس‌های متعدد، عملکرد بهتری دارد.

از سوی دیگر برای ارزیابی عملکرد روش پیشنهادی روی مجموعه داده‌های مختلف، از ماتریس پراکندگی استفاده شده است که در آنها محور x برچسب‌های پیش‌بینی‌شده، و محور y برچسب‌های واقعی را نشان می‌دهند. نتایج آزمایش‌ها نشان می‌دهند که روش پیشنهادی به بهترین شکل ممکن به تفکیک فعالیت‌های مختلف می‌پردازد و هم‌پوشانی دسته‌های مختلف بسیار ناچیز است. شکل ۷ به نمایش ماتریس پراکندگی عملکرد روش پیشنهادی روی مجموعه داده‌های UCFYT11 وUCF Sports پرداخته است. در مجموعه داده UCFYT11 بیشترین هم پوشانی به زوج فعالیت‌های (horse-biking و walking-riding) و (ridingوbiking-walking) به دلیل داشتن حرکات بسیار شبیه به هم مرتبط می‌شود.

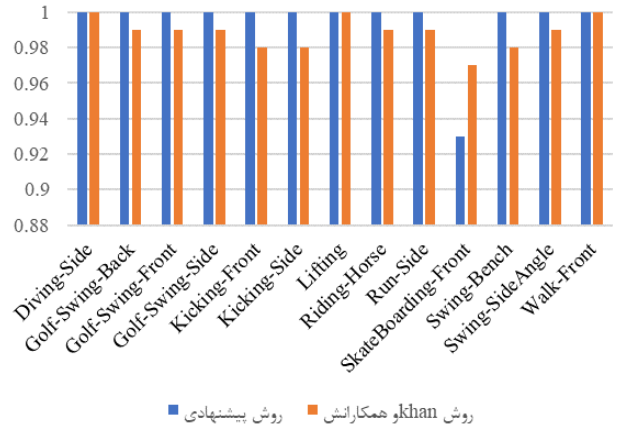
بنابراین می‌توان نتیجه گرفت که عملکرد مناسب روش پیشنهادی در حقیقت ناشی از قدرت تفکیک فعالیت‌های دارای حرکت‌های متفاوت است و هم‌پوشانی بین فعالیت‌های دارای حرکات مشابه نیز نسبت به سایر روش‌ها جزیی بوده و اجتناب ناپذیر هستند. بزرگ بودن اعداد روی قطر اصلی نشان دهنده تشخیص صحیح برچسب ویدئوی ورودی از دسته‌های مختلف فعالیتی است. در مجموعه داده UCF Sports نیز بیشترین روی هم افتادگی فعالیت‌ها به زوج فعالیت‌های (Side-Kicking و Front-Kicking) و (Side-Run و SkateBoarding) مربوط می‌شوند که حرکات مشابه بسیاری دارند. البته نتایج نشان می‌دهند که این هم‌پوشانی‌ها نسبت به سایر روش‌ها کمتر بوده و همچنین هم‌پوشانی اشتباه قابل توجهی بین فعالیت‌های دارای حرکت‌های متفاوت وجود ندارد.

لازم به ذکر است که یادگیری توالی زمانی ویژگی‌ها در هر جریان این روش، اثر بسیار قابل توجهی در دقت نهایی داشته است تا طبقه‌بندی موفق حاصل شود. ویژگی‌های مکانی استخراج شده با وجود آنکه مفید هستند اما به تنهایی کافی نمی‌باشند. در این راستا جدول ۴ به نمایش دقت هر جریان از این روش می‌پردازد. این در حالی است که در این جدول دقت روش برای حالتی که تنها از ویژگی‌های مکانی استخراج شده برای طبقه‌بندی فعالیت استفاده شده باشد نیز وجود دارد. بدین منظور نقشه ویژگی‌های تولید شده در هر جریان به صورت مجزا به یک ماشین بردار پشتیبان quadratic داده می‌شود. در هر دو روشی که در این جدول گزارش شده است، هم‌جوشی ویژگی‌ها قبل از طبقه‌بندی و همچنین هم‌جوشی نتایج طبقه‌بندها هم ارزیابی شده است. مشخص است که هم‌جوشی ویژگی‌ها موجب افت عملکرد روش می‌گردد و دلیل آن هم از دست رفتن اطلاعات مهمی از فعالیت انسان در مرحله هم‌جوشی و ایجاد مدل نهایی خواهد بود.

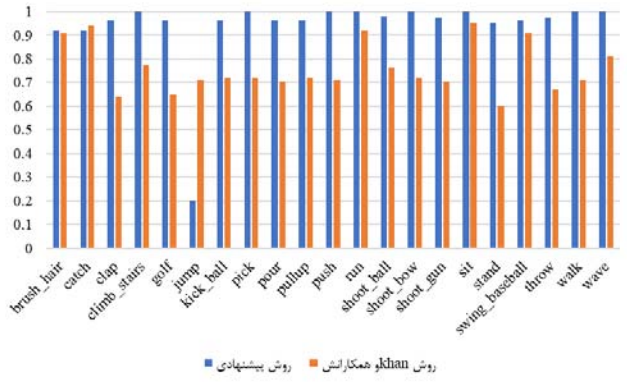
برای بررسی نحوه عملکرد لایه‌های شبکه کانولوشنی معرفی شده در این مقاله، در دو آزمایش لایه‌های ۳۲ و ۶۴ میانی به ترتیب از شبکه حذف می‌شوند که نتیجه نهایی این دو حالت در جدول زیر به نمایش درآمده است. این جدول نشان می‌دهد که با حذف هر کدام از لایه‌های میانی این شبکه افت قابل توجهی در دقت نهایی تشخیص فعالیت رخ می‌دهد. در حقیقت حذف یکی از مقیاس‌های بررسی تغییرات در طول زمان باعث از دست رفتن بخش مهمی از اطلاعات لازم در مدل نهایی مکانی-زمانی فعالیت می‌گردد و با افت دقت توام خواهد بود. حذف لایه ۶۴ میانی در این آزمایش با افت دقت بیشتری همراه بوده است که می‌تواند به دلیل از دست دادن دید کلی به تغییرات و محدود شدن به ویژگی‌های محلی باشد.



الف



ب



ج

شکل ۶ دقت روی (الف) مجموعه داده UCF11، (ب) مجموعه داده UCF Sports و (ج) مجموعه داده HMDB

جدول ۴ ارزیابی عملکرد ویژگی‌های مکانی و مدل مکانی زمانی نهایی و همچنین تاثیر هم‌جوشی جریان‌ها قبل از طبقه‌بندی و بعد از طبقه‌بندی

	UCF Sports	UCF11	J-HMDB
روش	دقت (%)	دقت (%)	دقت (%)
دقت حاصل از ویژگی‌های مکانی با SVM جریان اول	۷۹,۵	۷۵,۲۲	۶۳,۳۱
دقت حاصل از ویژگی‌های مکانی با SVM جریان دوم	۸۱,۴۲	۷۸,۳۸	۶۴,۲۵
دقت روش پیشنهادی با SVM و هم‌جوشی ویژگی‌ها	۶۹,۳۹	۶۶,۵۶	۵۱,۴۵
دقت روش پیشنهادی با SVM و هم‌جوشی نتیجه طبقه‌بند روی دو جریان	۸۳,۸۴	۷۹,۲۵	۶۵,۱۲
دقت حاصل از جریان اول روش پیشنهادی	۹۶,۲۴	۹۴,۳۶	۸۹,۳۷
دقت حاصل از جریان دوم روش پیشنهادی	۹۸,۴	۹۷,۱۲	۹۰,۵۶
دقت روش پیشنهادی با هم‌جوشی ویژگی‌های دو جریان و سپس طبقه‌بندی با شبکه معرفی شده	۹۵,۳	۹۱,۱۹	۸۵,۵۱
روش پیشنهادی	**۹۹,۸۳**	**۹۸,۷۰**	**۹۲,۸۶**

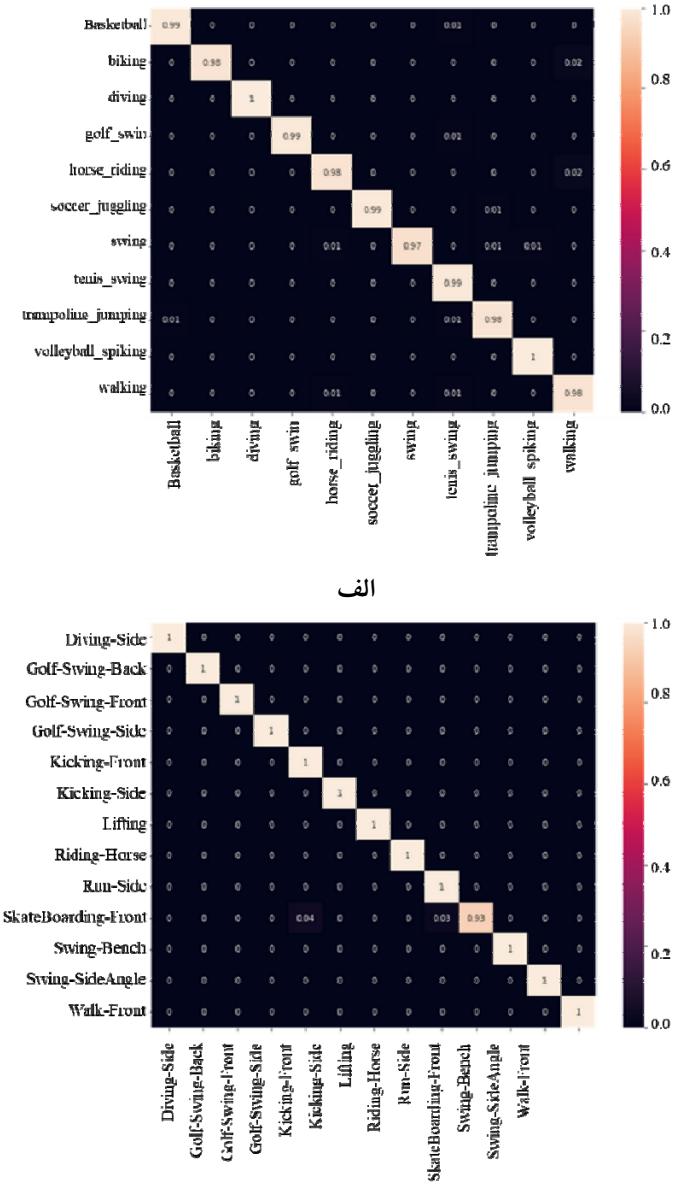
جدول ۵ ارزیابی عملکرد لایه‌های مختلف شبکه کانولوشنی

	UCF Sports	UCF11	J-HMDB
روش	دقت (%)	دقت (%)	دقت (%)
روش پیشنهادی با حذف لایه ۳۲ میانی از شبکه	۹۶,۲۹	۹۱,۱	۸۸,۳۳
روش پیشنهادی با حذف لایه ۶۴ میانی از شبکه	۹۵,۳۸	۹۲,۱۵	۸۹,۴۲
روش پیشنهادی	**۹۹,۸۳**	**۹۸,۷۰**	**۹۲,۸۶**

از نتیجه آزمایش‌ها تا به اینجا اینگونه برداشت می‌شود که جریان دوم دقت بهتری نسبت به جریان اول دارد و در حقیقت جریان اول (استفاده از ضرایب موجک) به عنوان کنترل کننده و کامل کننده جریان دوم (استفاده از ویژگی‌های حاصل از شبکه ResNet) مورد استفاده قرار گرفته است به طوری که دقت نهایی حدود ۲,۲ درصد نسبت به جریان دوم بهبود دارد. جدول زیر نشان می‌دهد که تغییر سطح موجک چه تاثیری بر دقت جریان اول دارد. مشخص است که دقت جریان اول با موجک سطح سه بهترین نتیجه را به دنبال دارد.

جدول ۶ ارزیابی عملکرد جریان اول با سطح‌های مختلف

	UCF Sports	UCF11	J-HMDB
روش	دقت (%)	دقت (%)	دقت (%)
دقت جریان اول با موجک سطح یک	۸۸,۴۹	۸۷,۳	۸۱,۱۴
دقت جریان اول با موجک سطح دو	۹۱,۴۵	۹۰,۳۸	۸۴,۲۷
دقت جریان اول با موجک سطح سه	**۹۶,۲۴**	**۹۴,۳۶**	**۸۹,۳۷**
دقت جریان اول با موجک سطح چهار	۸۶,۳۱	۸۲,۴۷	۷۲,۲۴



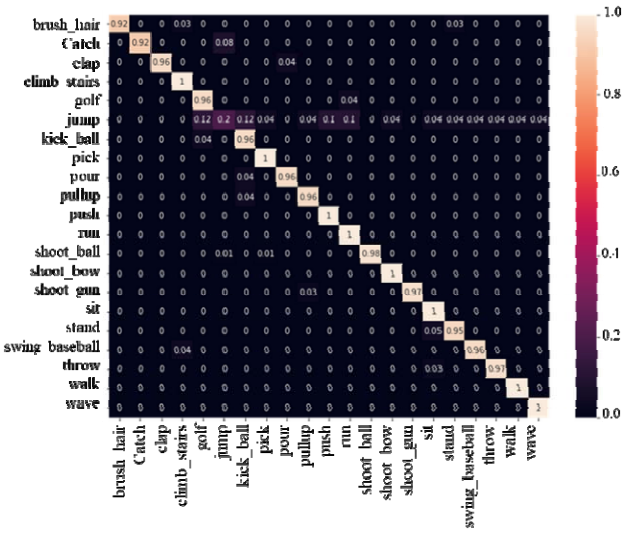
الف

ب

شکل ۷: ماتریس پراکندگی برای (الف) مجموعه داده UCF11 و (ب) مجموعه داده UCF Sports.

شکل ۸ به نمایش ماتریس پراکندگی عملکرد روش پیشنهادی روی مجموعه داده J-HMDB می پردازد. روش‌های مختلف تشخیص فعالیت کمترین دقت را روی این مجموعه داده به‌عنوان یک مجموعه داده چالشی با بیشترین برچسب‌های فعالیتی دارند. ماتریس پراکندگی نشان می‌دهد که روش پیشنهادی نیز هم‌پوشانی بیشتری میان برچسب‌های فعالیتی مختلف این مجموعه داده دارد. البته نتایج موجود در این ماتریس نشان می‌دهد که روش پیشنهادی به هم پوشانی میان فعالیت‌های با حرکت‌های مشابه منجر شده است که این نشان از بازنمایی مناسب فعالیت انسان دارد. واضح است که فعالیت‌هایی مانند catch و jump به دلیل داشتن حرکت-های بسیار نزدیک به هم دارای هم‌پوشانی بیشتری باشند که در روش پیشنهادی هم پوشانی آنها بیشتر از۲ درصد بوده است. این در حالی است که روش‌های پیشین مانند Muhammad و

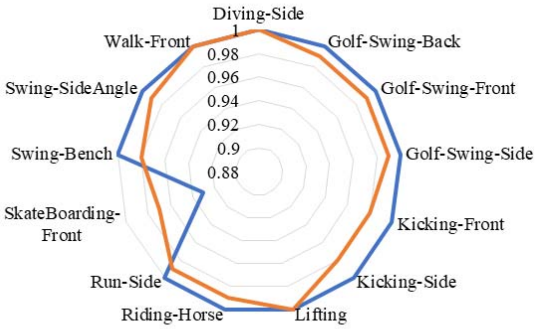
همکارانش [۱۹] به هم پوشانی برابر با ۲ درصد بین این دو دسته فعالیت رسیده‌اند. همچنین در روش khan و همکارانش هم-پوشانی فعالیت‌های stand با sit که حرکات کاملا متفاوتی دارند حدود ۳ درصد است در حالیکه در روش پیشنهادی هم پوشانی این دو دسته صفر است.



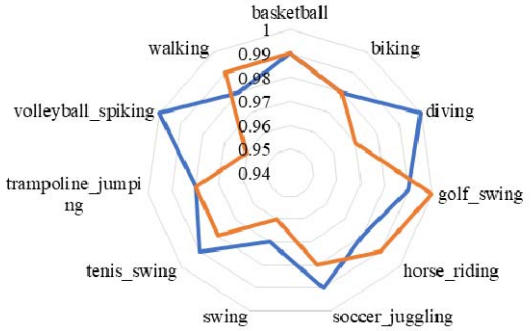
شکل ۸ ماتریس پراکندگی مجموعه داده مجموعه داده J-HMDB

بعلاوه شکل ۹ به نمایش پراکندگی نتایج روش پیشنهادی در مقایسه با روش معرفی شده توسط Khan و همکارانش [۱۹] روی دسته‌های فعالیتی مختلف پرداخته است. نمودارهای پراکندگی نتایج برای دسته‌های مختلف نشان می‌دهند که روش پیشنهادی در اکثریت دسته‌ها عملکرد بهتری دارد. در حقیقت در کنار داشتن عملکرد مناسب روی اکثریت دسته‌های فعالیت، نوسانات منفی قابل توجهی در عملکرد روش پیشنهادی بجز در دسته SkateBoarding در نمای جلو از مجموعه داده‌ی UCFSport و دسته Jump در مجموعه داده JHMDB وجود ندارد. با توجه به شباهت حرکت‌ها در دسته‌های SkateBoarding از نمای جلو و Kicking از نمای جلو، افت دقت روش پیشنهادی در این دسته نسبت به روش مورد مقایسه می‌تواند موجه باشد. همچنین از آنجا که فعالیت Jumping در سایر فعالیت‌ها به عنوان یک زیر فعالیت وجود دارد، افت عملکرد روش پیشنهادی در این دسته موجه به نظر می‌رسد. در سایر دسته‌ها، روش پیشنهادی توانسته است که عملکرد بهتر یا قابل مقایسه ای را ارائه کند که به همین دلیل دقت کلی مناسبی دارد. این عامل باعث می‌شود که روش پیشنهادی را بتوان به عنوان یک روش مطمئن نسبت به سایر روش‌ها در کاربردهای واقعی مورد استفاده قرار داد.

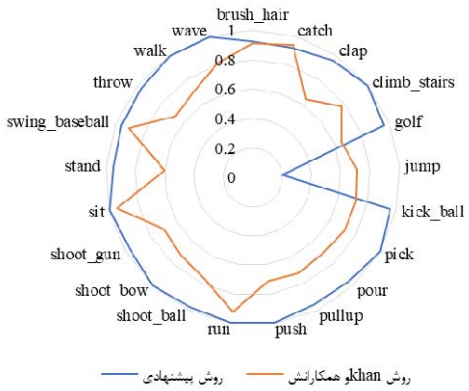
از آنجا که فعالیت‌های انسان تنوع زیادی دارد، پایداری به‌عنوان یکی از مولفه‌های مهم در ارزیابی روش‌های تشخیص فعالیت انسان باید مورد توجه قرار گیرد. لذا در آزمایش دیگری ما به دنبال بررسی تاثیر اضافه شدن دسته‌های فعالیت جدید به مجموعه داده-ها بر دقت نهایی روش هستیم. برای این آزمایش مجموعه داده JHMDB به دلیل داشتن تعداد زیادی دسته رفتاری مورد استفاده قرار گرفته است. در اینجا، عملکرد روش با ۱۳، ۱۵ و ۱۷ دسته رفتاری از این مجموعه داده مورد ارزیابی قرار می‌گیرد.



الف



ب



ج

شکل ۹ نمودار پراکندگی برای روش پیشنهادی و بهترین نتیجه گزارش شده توسط Khan و همکارانش برای (الف) مجموعه داده UCF11، (ب) مجموعه داده UCF Sports و (ج) مجموعه داده JHMDB

شکل ۱۰ نشان می‌دهد که تغییر تعداد کلاس‌های مجموعه داده مورد استفاده تاثیر چندانی در دقت نهایی نداشته است و روش معرفی شده یک روش پایدار برای استفاده در کاربردهای واقعی است.

شکل ۱۰ بررسی تاثیر تغییر تعداد کلاس‌های مجموعه داده JHMDB بر دقت نهایی

۵ نتیجه گیری و جهت گیری آینده

ویژگی‌های مکانی-زمانی نقش اساسی در تشخیص فعالیت‌های مختلف در داده‌های ویدیویی دارند. در این مقاله، ما یک روش دو جریانه جدید برای تشخیص فعالیت انسان، با استفاده از ویژگی‌های سنتی و عمیق دنباله‌ای از قاب‌های کلیدی را پیشنهاد کردیم. برای این منظور، ویژگی‌های مکانی مبتنی بر ضرایب موجک قاب‌های کلیدی در یک جریان استخراج شدند و در نقشه مکانی موجک ذخیره گشتند. در جریان دیگر ویژگی‌های عمیق قاب‌ها با استفاده از شبکه ResNetحاصل شده و در نقشه ویژگی-های عمیق قرار گرفتند. در این دو نقشه ویژگی‌های مکانی موجک که دارای چندریزگی و ویژگی‌های معنایی مناسب هستند، و ویژگی‌های عمیق که دید سراسری و محلی مناسبی از فعالیت دارند، مدل شده‌اند. این دو نقشه با استفاده از یک CNN جدید به مدل نهایی مکانی-زمانی هر جریان نگاشت —یافته‌اند که مدل نهایی تغییرات مکانی رخ داده در طول زمان را بازنمایی کرده است. در شبکه CNN پیشنهادی، بعد از ایجاد مدل خاص هر جریان، طبقه‌بندی و پیش‌بینی برچسب انجام شده است. در گام آخر نتیجه طبقه‌بندی جریان‌ها با هم ترکیب شده‌اند تا برچسب نهایی ویدئوی ورودی تعیین گردد. ما آزمایش‌های گسترده‌ای را روی سه مجموعه داده واقعی و چالشی UCF11، UCF Sports و J-HMDB انجام دادیم. روش پیشنهادی به دقت تشخیص ۹۸,۷٪ در مجموعه داده UCF11، ۹۹,۸ ٪ در مجموعه داده UCF Sports، و ۹۲,۸٪ در مجموعه داده J-HMDB دست یافت که نسبت به روش‌های HAR پایه برتری محسوسی دارد. چارچوب تشخیص فعالیت پیشنهادی در این مقاله از یک استراتژی یادگیری دو جریانه بهره برده و با استفاده از مجموعه داده‌های تشخیص فعالیت ارزیابی شده است. در ادامه این کار، سعی بر آن است که ویژگی‌های معنایی را به صورت برجسته‌تر در دو جریان فعلی یا به صورت مجزا در یک جریان سوم در نظر بگیریم. در حقیقت می توان تفکیک فعالیت به زیر فعالیت‌ها را در کارهای آتی انجام داد و با مدل کردن این زیر فعالیت‌ها و توالی آنها در یک جریان مجزا یا به‌صورت نهفته در دو جریان فعلی به بهبود نتیجه نهایی روش پیشنهادی فعلی پرداخت یا حتی از آن برای پیش‌بینی حرکت‌ها و فعالیت‌ها استفاده نمود.

مراجع

[1] Ullah, Amin, et al. "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments." Future Generation Computer Systems 96 (2019): 386-397.

[2] Dai, Cheng, Xingang Liu, and Jinfeng Lai. "Human action recognition using two-stream attention based LSTM networks." Applied soft computing 86 (2020): 105820.

[3] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2022). Action recognition via pose-based graph convolutional networks with intermediate dense supervision. Pattern Recognition, 121, 108170.

[4] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. Sensors, 19(5), 1005.

[5] Zare, Amin, Hamid Abrishami Moghaddam, and Arash Sharifi. "Video spatiotemporal mapping for human action recognition by convolutional neural network." Pattern Analysis and Applications 23.1 (2020): 265-279.

[6] Gao, Z., Xuan, H. Z., Zhang, H., Wan, S., & Choo, K. K. R. (2019). Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet of Things Journal, 6(6), 9280-9293.

[7] Ramezani M, Yaghmaee F. A review on human action analysis in videos for retrieval applications. Artificial Intelligence Review. 2016 Dec 1;46(4), pp:485-514.

[8] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." arXiv preprint arXiv:1406.2199 (2014).

[9] Ramezani M, Yaghmaee F. Content-based human actions retrieval by a novel low complex action representation. In2014 4th International Conference on Computer and Knowledge Engineering (ICCKE) 2014 Oct 29 (pp. 204-208). IEEE.

[10] Ren, Z., Zhang, Q., Gao, X., Hao, P., & Cheng, J. (2021). Multi-modality learning for human action recognition. Multimedia Tools and Applications, 80(11), 16185-16203.

[11] Zong, Ming, et al. "Motion saliency based multi-stream multiplier ResNets for action recognition." Image and Vision Computing 107 (2021): 104108.

[12] Khan, M. A., Alhaisoni, M., Armghan, A., Alenezi, F., Tariq, U., Nam, Y., & Akram, T. (2021). Video analytics framework for human action recognition.

[13] Guha, R., Khan, A. H., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2021). CGA: A new feature selection model for visual human action recognition. Neural Computing and Applications, 33(10), 5267-5286.

[14] Zhao S, Chen L, Yao H, Zhang Y, Sun X. Strategy for dynamic 3D depth data matching towards robust action retrieval. Neurocomputing. 2015 Mar 5;151, pp:533-43.

[15] Li, Z., & Li, D. (2022). Action recognition of construction workers under occlusion. Journal of Building Engineering, 45, 103352.

[16] Wang, H., Yu, B., Xia, K., Li, J., & Zuo, X. (2021). Skeleton edge motion networks for human action recognition. Neurocomputing, 423, 1-12.

[17] Dollár, Piotr, et al. "Behavior recognition via sparse spatio-temporal features." 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. IEEE, 2005.

[18] Ramezani, M., & Yaghmaee, F. (2018). Motion pattern based representation for improving human action retrieval. Multimedia Tools and Applications, 77(19), 26009-26032.

[19] Muhammad, K., Ullah, A., Imran, A. S., Sajjad, M., Kiran, M. S., Sannino, G., & de Albuquerque, V. H. C. (2021). Human action recognition using attention based LSTM network with dilated CNN features. Future Generation Computer Systems, 125, 820-830.

[20] Zhao, Yuxuan, et al. "Improved two-stream model for human action recognition." EURASIP Journal on Image and Video Processing 2020.1 (2020): 1-9.

[21] Fan, Hongxiang, et al. "F-E3D: FPGA-based acceleration of an efficient 3D convolutional neural network for human action recognition." 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP). Vol. 2160. IEEE, 2019.

[22] Gharaee, Z., Gärdenfors, P., & Johnsson, M. (2017). First and second order dynamics in a hierarchical SOM system for action recognition. Applied Soft Computing, 59, 574-585.

[23] Zhou, Y., Sun, X., Zha, Z. J., & Zeng, W. (2018). Mict: Mixed 3d/2d convolutional tube for human action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 449-458).

[24] Singh, Roshan, et al. "A Dual Stream Model for Activity Recognition: Exploiting Residual-CNN with Transfer Learning." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2020): 1-11.

[25] Naeem, H. B., Murtaza, F., Yousaf, M. H., & Velastin, S. A. (2021). T-VLAD: Temporal vector of locally aggregated descriptor for multiview human action recognition. Pattern Recognition Letters, 148, 22-28.

[26] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2022). Action recognition via pose-based graph convolutional networks with intermediate dense supervision. Pattern Recognition, 121, 108170.

[27] Wu, H., Song, C., Yue, S., Wang, Z., Xiao, J., & Liu, Y. (2022). Dynamic video mix-up for cross-domain action recognition. Neurocomputing, 471, 358-368.

[28] Afza, F., Khan, M. A., Sharif, M., Kadry, S., Manogaran, G., Saba, T., ... & Damaševičius, R. (2021). A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. Image and Vision Computing, 106, 104090.

[29] Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2017, March). Two stream lstm: A deep fusion framework for human action recognition. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 177-186). IEEE.

[30] Xiong, Q., Zhang, J., Wang, P., Liu, D., & Gao, R. X. (2020). Transferable two-stream convolutional neural network for human action recognition. Journal of Manufacturing Systems, 56, 605-614.

[31] Ben Tanfous, A., Zerroug, A., Linsley, D., & Serre, T. (2022). How and What To Learn: Taxonomizing Self-Supervised Learning for 3D Action Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2696-2705).

[32] Latah, M. (2017). Human action recognition using support vector machines and 3D convolutional neural networks. Int. J. Adv. Intell. Informatics, 3(1), 47.

[33] Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B., & Yuan, J. (2018). Multi-stream CNN: Learning representations based on human-related regions for action recognition. Pattern Recognition, 79, 32-43.

[34] Ma, M., Marturi, N., Li, Y., Leonardis, A., & Stolkin, R. (2018). Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognition, 76, 506-521.

عاطفه مرادیانی در سال ۱۳۹۸ مدرک کارشناسی خود را از دانشگاه کردستان دریافت نمود. ایشان در حال حاضر در مقطع کارشناسی ارشد رشته هوش مصنوعی دانشگاه کردستان در حال تحصیل هستند. زمینه‌های کاری مورد علاقه ایشان شبکه‌های عمیق و تشخیص رفتار انسان می باشد.

محسن رمضانی در سال ۱۳۹۷ مقطع دکتری خود را در دانشگاه سمنان به پایان رسانید. ایشان از همان سال به عنوان عضو هیات علمی دانشگاه کردستان شروع به فعالیت کردند. ایشان در دو حوزه پردازش ویدئو و داده‌کاوی به صورت تخصصی مشغول به پژوهش هستند. از جمله موضوعات مورد علاقه ایشان می‌توان به تشخیص رفتار انسان، بازیابی رفتار انسان، سیستم‌های توصیه‌گر و انتخاب ویژگی اشاره کرد.

فردین اخلاقیان طاب سال ۱۳۴۴ در شهر سنندج متولد گردید. او مقاطع کارشناسی، کارشناسی ارشد و دکترای خود را در به‌ترتیب در دانشگاه‌های صنعتی اصفهان، تربیت مدرس تهران و دانشگاه ولنگونگ استرالیا گذراند. ایشان هم اکنون دانشیار گروه کامپیوتر در دانشگاه کردستان می‌باشند. زمینه های تحقیقاتی مورد علاقه ایشان شامل یادگیری ماشین، یادگیری عمیق و ماشین بینایی می باشد. از نامبرده بیش از ۱۰۰ مقاله در مجلات و کنفرانس‌های معتبر بین المللی و داخلی منتشر گردیده است.

رحمت الله میرزایی سال ۱۳۴۰ در شهر سنندج متولد گردید. او مقاطع کارشناسی، کارشناسی ارشد و دکترای خود را به ترتیب در دانشگاه‌های تبریز، تهران و انستیتو علوم هند بنگلور گذراند. ایشان هم اکنون استادیار گروه برق در دانشگاه کردستان می‌باشند. زمینه های تحقیقاتی مورد علاقه ایشان در مهندسی برق و کامپیوتر می باشد. از نامبرده بیش از ۴۰ مقاله در مجلات و کنفرانس‌های معتبر بین المللی و داخلی منتشر گردیده است.