چهارمین کنفرانس بین المللی
مطالعات نوین در علوم کامپیوتر و فناوری اطلاعات
اسفند ماه 1396
February 2017
International Conference new study on Computer and it

# ارائه‌ی بازنمایی جدید برای شناسایی فعالیت‌های انسانی در ویدیو با استفاده از نقشه‌ی برجستگی

**ویدا عادلی مسبب**
دانشجوی کارشناسی ارشد هوش مصنوعی و رباتیکز، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
vida.adeli@mail.um.ac.ir
**احسان فضل ارثی**
استادیار گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران
fazlersi@um.ac.ir
**احد هراتی**
استادیار گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران
a.harati@um.ac.ir

**چکیده**
امروزه با توجه به گسترش سریع دوربین‌ها و داده‌های ویدیویی در تمامی جوانب زندگی، درک و استخراج اطلاعات از تصاویر و ویدیوها از اهمیت به سزایی برخوردار شده است. در دهه‌های اخیر، محققین زیادی به ارائه‌ی الگوریتم‌هایی برای تجزیه‌وتحلیل محتوایی این داده‌ها برای کاربردهای مختلف پرداخته‌اند. این مسئله زمینه‌ی توسعه‌ی روش‌هایی که توانایی شناسایی فعالیت‌های رخ داده در یک توالی ویدیویی را دارند، فراهم ساخته است. در این مقاله نیز مسئله‌ی شناسایی فعالیت‌های انسانی با ارائه‌ی روشی جدید برای بازنمایی ویدیو مورد بررسی قرار گرفته است. با استفاده‌ی هم‌زمان از اطلاعات ظاهری و حرکتی درون ویدیو، روشی برای استخراج نواحی بالقوه‌ی رخداد فعالیت ارائه شده، که توانایی استخراج مناطق پیش‌زمینه مرتبط با فعالیت را داشته و بازنمایی جدیدی ارائه شده که می‌تواند از این اطلاعات در ترکیب با اطلاعات و ساختار کلی پس‌زمینه، با میزان اهمیت‌های متفاوت استفاده کند. در مدل پیشنهادی برای شناسایی فعالیت، از یک معماری دو جریانه از شبکه‌های عصبی پیچشی (CNN) برای اطلاعات مکانی و زمانی، به منظور استخراج ویژگی‌های فریمی، استفاده شده و سپس این ویژگی‌ها برای ادغام بهینه با در نظر گرفتن روابط زمانی، برای تولید بازنمایی نهایی ویدیو و برچسب‌زنی، در اختیار یک شبکه‌ی حافظه‌ای طولانی کوتاه-مدت (LSTM) قرار می‌گیرند. برای ارزیابی مدل پیشنهادی از مجموعه داده‌ی معتبر و چالشی jHMDB استفاده شده و نشان خواهیم داد که روش پیشنهادی توانسته از دقت بهتری نسبت به سایر روش‌های موجود بر روی این مجموعه داده برخوردار باشد.

**واژگان کلیدی:** شناسایی فعالیت، حرکت، اطلاعات پیش‌زمینه، شبکه‌های عصبی پیچشی، نقشه‌ی برجستگی

**مقدمه و طرح مسئله**

شناخت فعالیت‌های انسانی در ویدیو ازجمله زمینه‌هایی است که در سال‌های اخیر کاربردهای بسیاری را در حوزه‌ی بینایی کامپیوتر به خود اختصاص داده است. یکی از اولین تحقیقات در مورد ماهیت حرکت انسان، توسط دو عکاس معاصر در دهه‌ی 1850 انجام شده است. آن‌ها از سوژه‌های متحرکت عکاسی کردند و این موضوع باعث آشکار شدن جوانب جذابی در مورد حرکت انسان‌ها و حیوانات شد. پس از آن، آزمایش‌های مختلفی در این زمینه صورت گرفت که انگیزه‌ی بزرگی را برای مطالعه، تجزیه‌وتحلیل و درک حرکت انسان در زمینه‌ی علم عصب‌شناسی ایجاد کرد. این امر باعث هموار شدن زمینه برای مدل‌سازی ریاضی فعالیت‌های انسان و شناسایی خودکار آن شد. درنهایت، گستره‌ی این مسئله به حوزه‌ی بینایی کامپیوتر و تشخیص الگو وارد شد. در سال‌های اخیر مسئله‌ی شناسایی فعالیت‌های انسانی در زمینه‌های مختلفی در حوزه بینایی کامپیوتر مورد تحقیق قرار گرفته است.

در حالت کلی روش‌های مبتنی بر بینایی در مسئله طبقه‌بندی ویدیو به دو بخش اصلی تقسیم می‌شوند: مرحله‌ی استخراج ویژگی و ایجاد بازنمایی فریم و یا ویدیو و مرحله‌ی طبقه‌بندی و برچسب‌زنی ویدیو. طبقه‌بندی ویدیو را می‌توان به‌صورت تجزیه‌وتحلیل محتوای بصری یک کلیپ کوتاه با استفاده از روش‌های مختلف و سپس استخراج یک مفهوم کلی از ویدیو و درنهایت اعمال یک برچسب مشخص با توجه به دسته‌های[1] موجود تعریف کرد. با توجه به این تعریف همان‌طور که واضح است، برای استخراج محتوای کلی ویدیو و طبقه‌بندی آن، نیاز است فعالیت‌های افراد درون ویدیو، تشخیص داده شده و با مقایسه فعالیت‌های صورت گرفته در ویدیو موردنظر با محتوای ویدیوهای درون مجموعه داده، برچسب مناسب برای هر ویدیو اختصاص داده شود. به بیانی دقیق‌تر، مسئله شناسایی فعالیت به صورت تشخیص نوع فعالیت رخ داده در یک ویدیو کلیپ، با استفاده از تجزیه و تحلیل محتوای بصری یک دنباله از تصاویر ویدیویی و اختصاص یک برچسب از بین مجموعه‌ای از دسته‌های از پیش تعیین شده اطلاق می‌شود. همان‌قدر که این مسئله برای انسان ساده به نظر می‌رسد، در مقابل طراحی راه‌حل قابل‌قبول برای حل آن توسط یک سیستم بینایی نیز به همان اندازه سخت و پیچیده خواهد بود.

توانایی درک خودکار فعالیت‌های انسانی در ویدیوهای دنیای واقعی کاربردهای فراوانی دارد که می‌تواند دلیل مهمی برای اهمیت این مسئله در چند سال اخیر باشد (Karpathy et al, 2014; Wang et al, 2011). تحلیل خودکار داده‌های ویدیویی در موارد زیادی می‌تواند روند جستجو و بازیابی را کارآمدتر کند. توانایی تشخیص فعالیت‌های پیچیده‌ی انسان‌ها در ویدیوهای مختلف کاربردهای بسیار مهمی را در اختیار قرار می‌دهد و دنیای جدیدی را به حل بسیاری از مسائل حاصل از ارتباط مستقیم انسان با کامپیوتر باز می‌کند. سامانه‌های نظارت خودکار در مکان‌های عمومی مانند فرودگاه‌ها و ایستگاه‌های مترو نیاز به تشخیص فعالیت‌های غیرطبیعی و مشکوک در مقابل فعالیت‌های عادی دارند. به عنوان مثال یک سیستم نظارتی در فرودگاه باید قادر باشد که به طور خودکار فعالیت‌های مشکوک از قبیل انداختن کیف در سطل زباله توسط فرد و یا ردوبدل کردن اشیا مشکوک بین افراد را تشخیص دهد. همچنین شناسایی فعالیت‌های انسانی امکان نظارت بیماران، کودکان و افراد سالمند را به راحتی در زمان بلادرنگ مهیّا می‌سازد. ساخت رابط‌های کاربردی مبتنی بر حرکت انسان برای ساخت محیطی هوشمند مبتنی بر بینایی کامپیوتر از کاربردهای دیگر این سامانه‌هاست که در سال‌های اخیر در کاربردهای واقعیت مجازی نیز مورد توجه قرار گرفته است. درعین‌حال فیلتر کردن محتوایی ویدیوها از دیگر کاربردهای مهم این مسئله است، که در آن محتوای درون فیلم‌ها و ویدیوها مورد پردازش قرار می‌گیرد و با استفاده از این سیستم شناسایی، می‌توان بخش‌هایی از آن را حذف کرده و یا بخش‌هایی از ویدیو را که مورد نظر است، بدون جستجو در کل ویدیو انتخاب کنیم.

به‌طورکلی چالش‌های مختلفی در مسئله‌ی شناسایی فعالیت‌های انسان وجود دارد که حل آن را تا حد زیادی مشکل می‌سازد. اولین و شاید مهم‌ترین چالش، این است که در مرحله‌ی آموزش، یک سامانه‌ی یادگیر برای شناسایی فعالیت نیاز به داده‌های حاشیه‌نویسی‌شده‌ی کامل[2] دارد، به شکلی که هر ویدیو یک یا چند برچسب داشته و برای تک‌تک‌فریم‌ها، مکان فعالیت توسط

[1] Class
[2] Fully Annotated

یک مستطیل محاطی مشخص باشد. واضح است که تولید چنین مجموعه داده‌هایی برای آموزش سامانه‌های یادگیر، کاری طاقت‌فرسا و هزینه‌بر است. یکی دیگر از این چالش‌ها وجود تغییرات زیاد درون دسته‌ای است. برای بسیاری از فعالیت‌ها اختلافات زمانی زیادی در طول اجرای آن توسط افراد مختلف و در شرایط متفاوت وجود دارد. به عنوان مثال در فعالیت "راه رفتن" حرکت‌ها می‌توانند در سرعت اجرا و یا طول گام متفاوت باشند. یک سیستم شناسایی مطلوب باید نسبت به این تغییرات و تفاوت‌های زمانی مقاوم باشد. همچنین بین افراد مختلف نیز تفاوت‌های زیادی به لحاظ جسمی وجود دارد و هر فرد می‌تواند به نحوی در ایجاد تغییراتی در نوع اجرای فعالیت کمک کند. تغییرات فراوان در شرایط تصویربرداری و نورپردازی، تفاوت در زاویه دید و شرایط متفاوت محیطی نیز، چالش‌های زیادی را در مقابل حل این مسئله قرار می‌دهند. در محیط‌های شلوغ و حتی پویا تشخیص موقعیت فرد و شناسایی فعالیت سخت‌تر خواهد شد. حرکت و شلوغی پس‌زمینه‌ی تصویر، استخراج اطلاعات مفید و متناسب با فعالیت رخ‌داده در ویدیو را دشوارتر می‌سازد. در عین حال استفاده از دوربین‌های متحرک نیز تمامی این شرایط‌ها را پیچیده‌تر می‌کند. از طرفی دیگر، ادغام کارآمد ویژگی‌های مختلف فریمی و در نظر گرفتن ساختار سری زمانی به‌صورت مناسب در پردازش‌های ویدیویی ازجمله زمینه‌های باز است، که همچنان بسیار موردتوجه قرار گرفته است.

با وجود چالش‌های فراوان این حوزه، کاربردهای فراوان تشخیص و شناسایی فعالیت، محققین را بر این داشته تا روش‌های مختلفی برای حل مسائل دنیای واقعی در این حوزه ارائه کنند. بسیاری از روش‌های قبلی این چالش‌ها را توسط ارائه‌ی مجموعه ویژگی‌های مختلف، حجم‌های مکانی-زمانی[3] (STV)، یا خط‌سیرها[4] هدف قرار داده‌اند و طبقه‌بندهای متفاوتی را ارائه کرده‌اند. گرچه این روش‌ها نتایج نسبتاً خوبی به دست آورده‌اند، اما اکثر آن‌ها قابلیت تفسیر آنچه را که آموزش می‌بینند، ندارند. به این معنی که دسته ویژگی‌های مختلف از تمامی صحنه را استفاده می‌کنند و مشخص نیست که کدام مجموعه ویژگی‌ها مفید هستند و کدام ویژگی‌ها مفید نیستند. به عبارت دیگر، ویژگی‌های نواحی نامرتبط به فعالیت در فرآیند آموزش دخیل شده و اغلب طبقه‌بندها نیز امکان تمیز دادن آن‌ها را ندارند.

علاوه‌براین، پیچیدگی الگوهای حرکتی موجود در توالی فریم‌های ویدیویی، مسئله شناسایی فعالیت در ویدیو را در مقایسه با پردازش تصاویر ثابت بسیار دشوارتر ساخته است. اطلاعات مکانی و بصری موجود در یک تصویر، اطلاعات مناسبی را در ارتباط با ویژگی‌های ظاهری هر فریم و صحنه در اختیار قرار می‌دهند. اگرچه این ویژگی‌ها می‌توانند منجر به تمایز بین برخی از دسته‌های موجود شوند، در برخی موارد نیز می‌توانند باعث ایجاد خطا و طبقه‌بندی نادرست شوند. به عنوان مثال، الگوهای حرکتی موجود در دو دسته‌ی "دویدن" و "راه رفتن" تنها جنبه‌ی تفاوت بین این دو دسته است، که می‌تواند منجر به برچسب‌گذاری درست ویدیوهای مربوط به این دسته‌ها شود. در حالی‌که اطلاعات ظاهری فریم ممکن است در تشخیص کمک چندان مناسبی نکند. بدون استفاده از اطلاعات بُعد زمانی، این دو دسته ممکن است به‌راحتی با یکدیگر اشتباه گرفته شوند. بنابراین می‌توان گفت که اطلاعات ظاهری و حرکتی، دو عنصر مکمل هستند که ترکیب و ادغام آن‌ها در مراحل اولیه الگوریتم می‌تواند به ارزیابی مناسبی از نواحی بالقوه رخداد فعالیت منجر شود.

به طور کلی، مدل‌سازی بهینه و کارآمد فعالیت‌ها، برای شناسایی فعالیت‌های انسانی از اهمیت به سزایی برخوردار است. برای بدست آوردن ویژگی‌های قابل اطمینان، که نماینده واقعی از فعالیت‌های رخ‌داده هستند، نیاز است که مناطق پس‌زمینه و غیرمرتبط با فعالیت حذف شوند. با این حال، در برخی موارد اطلاعات پس‌زمینه در تشخیص فعالیت رخ‌داده می‌تواند بسیار مؤثر باشد. چراکه محیطی که فعالیت در آن رخ می‌دهد، می‌تواند تعدادی از دسته‌ها را با قطعیت بیشتری از تصمیم‌گیری حذف کند. به عنوان مثال در دسته‌ی "فوتبال بازی کردن"، تشخیص زمین چمن می‌تواند مؤثر باشد. اکثر روش‌های موجود برای مکان‌یابی فعالیت، تأثیر این اطلاعات را به طور کلی در نظر نگرفته و فقط نواحی اصلی مربوط به فعالیت رخ‌داده را به وسیله‌ی یک مستطیل محاطی در اختیار قرار می‌دهند (Gkioxari and Malik, 2015; Wang et al, 2016; Jain et al 2014; Jargalsaikhan et al, 2017). لذا در این مقاله روشی ارائه می‌کنیم که با استفاده هم‌زمان از الگوهای حرکتی

[3] Space-Time Volumes
[4] Trajectory

و ظاهری هر فریم، می‌تواند نواحی با پتانسیل بالا که احتمال وقوع فعالیت در آن‌ها بیشتر است را استخراج کند. در عین حال، بازنمایی‌ای برای ویدیو ارائه می‌کنیم که از اطلاعات مفید و غیرزائد پس‌زمینه در ترکیب با اطلاعات پیش‌زمینه با میزان اهمیت‌های متناسب استفاده خواهد کرد.

علاوه‌براین، چگونگی استفاده از اطلاعات حرکتی در فرآیند شناسایی بسیار تأثیرگذار و مهم است. با توجه به عملکرد چشم‌گیر شبکه‌های عصبی پیچشی[5] (CNN) در کاربردهای حوزه‌ی بینایی کامپیوتر، اکثر روش‌های موجود از رویکردی دو جریانه از این شبکه‌ها برای پردازش اطلاعات بُعد مکانی و زمانی ویدیو استفاده می‌کنند (Karpathy et al, 2014; Gkioxari and Malik, 2015; Simonyan and Zisserman 2014; Yue-Hei et al, 2015; Feichtenhofer et al, 2016).

در این رویکردها اطلاعات حرکتی به صورت جداگانه از اطلاعات ظاهری فریم، برای ایجاد بازنمایی ویدیو مورد پردازش قرار می‌گیرد. در این مقاله با استفاده‌ی همزمان از اطلاعات حرکتی و ظاهری و ایجاد مدلی برای فعالیت در ویدیو با استفاده از این دو نوع داده، بازنمایی‌ای ارائه می‌شود که می‌تواند اطلاعات مرتبط با فعالیت رخ‌داده در صحنه را به خوبی شناسایی کند.

لذا نوآوری‌های اصلی این مقاله را می‌توان در سه مورد اصلی خلاصه کرد:

1. ارائه رویکردی برای استفاده و ترکیب اطلاعات حرکتی و ظاهری در فازهای ابتدایی الگوریتم برای بدست آوردن نواحی مرتبط با فعالیت در فریم. این مسئله باعث عدم نیازمندی روش ارائه شده به مجموعه داده‌های حاشیه-نویسی‌شده‌ی کامل می‌شود و لذا روش پیشنهادی در شرایط یادگیر با ناظر ضعیف نیز، که در آن فقط یک برچسب برای کل ویدیو موجود است، به خوبی عمل خواهد کرد.
2. ارائه بازنمایی‌ای برای ویدیو که از اطلاعات پس‌زمینه در ترکیب با اطلاعات پیش‌زمینه با میزان اهمیت‌های متفاوت استفاده می‌کند.
3. ادغام بهینه توصیفگرهای فریمی و در نظر گرفتن روابط زمانی بین آن‌ها.

درنهایت نشان خواهیم داد که روش ارائه شده توانسته است عملکرد بهتری نسبت به بسیاری از روش‌های موجود برای شناسایی فعالیت و برچسب‌زنی ویدیو، در اختیار قرار دهد.

خلاصه‌ی بخش‌هایی از این مقاله به شرح زیر است: در بخش دوم پیشینه‌ی تحقیق و کارهای مرتبط انجام شده در حوزه‌ی شناسایی فعالیت و طبقه‌بندی ویدیو ارائه شده است. بخش سوم چارچوب اصلی مدل ارائه شده برای شناسایی فعالیت را تشریح می‌کند. در بخش چهارم مجموعه‌داده‌ی مورد استفاده معرفی شده و نتایج آزمایش‌ها مورد بحث قرار می‌گیرند. در نهایت در بخش پنجم، نتایج تحقیق جمع‌بندی و نتیجه‌گیری شده است.

## مروری بر کارهای پیشین

با توجه به کاربردها و چالش‌های گسترده شناسایی فعالیت‌های انسان، تحقیقات فراوانی در این حوزه انجام شده که این تحقیقات منجر به ایجاد و گسترش روش‌های گوناگونی در این زمینه شده است. می‌توان گفت که تشخیص فعالیت‌های انسان نقطه‌ی اشتراک بین دو حوزه‌ی بینایی ماشین و یادگیری ماشین است (Aggarwal and Ryoo, 2011). در حقیقت، توصیف و استخراج ویژگی‌های یک ویدیو در حیطه‌ی بینایی ماشین قرار می‌گیرد و برای مدل کردن، تشخیص و دسته‌بندی از مفاهیم یادگیری ماشین بهره می‌گیریم. مقالات مروری بسیاری در این حوزه به مطالعه جامع و مقایسه‌ی روش‌های مختلف این حوزه پرداخته‌اند (Weinland et al, 2011; Herath et al, 2017; Soomro and Zamir, 2014). در این بخش از مقاله قصد داریم که مروری کلی از روش‌های موجود ارائه دهیم و سپس جایگاه این تحقیق را نسبت به وضعیت فعلی این حوزه مشخص کنیم.

[5] Convolutional Neural Networks (CNNs)

روش‌های موجود شناسایی فعالیت را می‌توان به دو دسته‌ی کلی تقسیم کرد: بازنمایی‌های مبتنی بر ویژگی‌های از پیش طراحی‌شده[6]، که به طور معمول از ویژگی‌های محلی[7] استفاده می‌کنند و مدل‌های مبتنی بر شبکه‌های عصبی عمیق.

**مدل‌های مبتنی بر ویژگی‌های از پیش‌طراحی شده**

به طور کلی، دو دیدگاه اصلی در استفاده از این ویژگی‌ها برای مسئله‌ی شناسایی فعالیت وجود دارد (Herath et al, 2017): برخی از روش‌ها از بازنمایی کلی[8]، که بازنمایی‌ای سراسری از بدن انسان و حرکات آن است، استفاده می‌کنند. برخی دیگر از روش‌ها نیز از ویژگی‌های محلی استفاده می‌کنند. حجم‌های مکان-زمانی (STV) که اولین بار در (Blank et al, 2005) و (Yilmaz and Shah, 2005) ارائه شدند جز دسته روش‌های بازنمایی کلی قرار می‌گیرند. در این روش‌ها، یک حجم سه بُعدی مکان-زمانی به وسیله‌ی انباشتن تصویر فرد، یعنی مناطق پیش‌زمینه که فرد در آنجا وجود دارد، در راستای بُعد زمان بدست می‌آید. این نوع بازنمایی، از یک فضای سه بُعدی (x, y, t) برای بازنمایی اشیا و اشخاص و ارتباط بین اطلاعات مکانی و زمانی استفاده می‌کند. مبنای روش‌های شناسایی با استفاده از فضای مکان-زمان، بر اساس اندازه‌گیری میزان شباهت دو فضای مختلف سه بُعدی با یکدیگر است. به‌جای ساخت یک فضای سه بُعدی برای هر فعالیت، در برخی از روش‌ها هر فعالیت با یک تصویر دو بُعدی الگو بازنمایی می‌شود. تصویر دو بُعدی باینری انرژی حرکت[9] (MEI) و تصویر مقدار عددی پیشینه‌ی حرکتی[10] (MHI) (Bobick and Davis, 2001) نمونه‌هایی از این تصاویر الگو هستند. واضح است که رویکردهای سراسری فقط می‌توانند یک تخمین و الگویی کلی از فعالیت را یاد بگیرند و لذا توانایی یادگیری جزئیات دقیق‌تری از فعالیت را ندارند. چنین مشکلی باعث توسعه‌ی روش‌های مبتنی بر مدل‌های محلی شده است.

مطالعات بسیاری، با بهره‌گیری از ویژگی‌های مکان-زمانی به بررسی مسئله شناسایی فعالیت پرداخته‌اند (Bregonzio et al, 2009; Marszalek et al, 2009; Liu et al, 2009; Sun et al, 2009). یکی از پایه‌ای‌ترین روش‌ها در این بین مقاله‌ی (Laptev, 2005) است، که روشی برای استخراج نقاط کلیدی مکان-زمانی، یعنی نواحی‌ای از تصویر که در هر دو بُعد مکان و زمان نواحی برجسته با تغییرات ناگهانی هستند، ارائه می‌دهد. در (Laptev et al, 2008) از هیستوگرام گرادیان[11] (HoG) (Dalal and Triggs, 2005) و هیستوگرام جریان نوری[12] (HoF) برای توصیف ویژگی‌های مکان-زمانی استفاده شده است. سپس توصیفگرهای ویدیویی به کمک مدل کیسه‌ی واژگان[13] (BoW) مدل‌سازی شده و در نهایت پس از ایجاد یک بازنمایی برای ویدیو، از یک طبقه‌بند SVM برای پیش‌بینی دسته‌های مربوط به هر ویدیو استفاده می‌شود. در پژوهش‌های بعدی ویژگی‌های سه بُعدی، شامل اطلاعات زمانی به عنوان بُعد سوم پردازش، ارائه شده‌اند. توصیفگرهای HoG3D (Klaser et al, 2008)، Harris3D (Sipiran and Bustos, 2011) و Cuboids (Dollár et al, 2005) نمونه‌هایی از این روش‌ها هستند.

در (Wang et al, 2009) نشان داده شده است که استفاده از روش‌های نمونه‌برداری متراکم[14]، می‌تواند جایگزین مناسبی برای روش‌های آشکارسازی و استخراج نقاط کلیدی از تصویر باشد. در (Wang et al, 2011) ایده‌ی استفاده از نمونه‌برداری نقاط به‌صورت متراکم برای ایجاد خط‌سیرها تعمیم داده شده است. در ابتدا نقاط کلیدی به صورت متراکم انتخاب شده و سپس این نقاط در طول فریم‌ها به کمک جریان نوری رهگیری شده و خط‌سیرها ایجاد می‌شوند. نسخه‌ی بهبودیافته‌ی این

[6] Handcrafted features
[7] Local features
[8] Holistic representation
[9] Two-dimensional binary motion-energy image
[10] Scalar-valued motion-history image
[11] Histogram of Gradient
[12] Histogram of Optical Flow
[13] Bag of Words
[14] Dense sampling

روش در (Wang and Schmid, 2013) ارائه شده که نسبت به حرکات دوربین نیز مقاوم است. از بین روش‌های مبتنی بر خط‌سیر، هیستوگرام مرز حرکتی[15] (MBH) (Dalal et al, 2006) دارای عملکرد خوبی بوده است.

در نهایت در اکثر روش‌های موجود، برای ایجاد بازنمایی مناسب برای ویدیو از روی بازنمایی‌های فریمی، از مدل کیسه واژگان و کدگذاری بردار فیشر استفاده می‌شود (Wang and Schmid, 2013). با وجود عملکرد نسبتاً قابل‌قبول روش‌های مبتنی بر ویژگی‌های از پیش طراحی‌شده، به دلیل وجود مشکلات زیاد این رویکردها، تحقیقات به سمت استفاده از مدل‌های عمیق پیش رفته است. از جمله‌ی این مشکلات می‌توان به غیرقابل انعطاف بودن نسبت به تنوع‌های درون دسته‌ای فعالیت‌ها اشاره کرد.

**مدل‌های مبتنی بر شبکه‌های عمیق**

در سال‌های اخیر، شبکه‌های عصبی پیچشی در کاربردهای مختلف نظیر تشخیص صوت و طبقه‌بندی تصویر، عملکرد چشم‌گیری داشته‌اند. نتایج خوب بدست آمده با استفاده از این شبکه‌ها، محققین زیادی را به توسعه‌ی آن‌ها برای استفاده در کاربردهای پردازش ویدیویی مانند شناسایی فعالیت ترغیب کرده است. با استفاده از مدل‌های CNN می‌توان مفاهیم پیچیده‌تری از ویدیو را، با در نظر گرفتن توالی‌های فریمی استخراج کرد. تلاش‌های بسیاری برای دخیل کردن اطلاعات زمانی موجود در ویدیو به مدل‌های CNN انجام شده است (Simonyan and Zisserman, 2014; Ji et al, 2013; Wang et al, 2015; Fernando and Gould, 2016; Chéron et al, 2015). به طور کلی، این روش‌ها را در دو دسته می‌توان مورد بررسی قرار داد: شبکه‌های دو جریانه و شبکه‌های مکان-زمانی.

اکثر روش‌های موجود در حوزه‌ی شناسایی فعالیت که از شبکه‌های عمیق استفاده می‌کنند، یک معماری دو جریانه در نظر می‌گیرند (Karpathy et al, 2014; Gkioxari and Malik, 2015; Wang et al, 2015; Feichtenhofer et al, 2016; Yue-Hei et al, 2015). ساختار این معماری‌ها به این صورت است که دو جریان موازی برای اطلاعات مکانی درون فریم و اطلاعات حرکتی بین فریمی در نظر گرفته می‌شود. جریان مکانی، اطلاعاتی درباره‌ی ظاهر فریم و صحنه در اختیار قرار می‌دهد و پردازش‌های لازم را بر روی تصاویر رنگی انجام می‌دهد. در حالی‌که جریان زمانی، اطلاعات حرکتی بین فریم‌های متوالی را پردازش می‌کند و از تصاویر جریان نوری به عنوان ورودی استفاده می‌کند. در نهایت پیش‌بینی و طبقه‌بندی نهایی به وسیله‌ی ادغام خروجی‌های این دو جریان بدست خواهد آمد.

رویکرد دیگر استفاده از شبکه‌های عمیق برای شناسایی فعالیت، به این صورت است که تنها از یک جریان شبکه استفاده می‌شود، با این تفاوت که معماری شبکه به گونه‌ای است که می‌تواند اطلاعات زمانی را نیز مورد پردازش قرار دهد. به این دسته از شبکه‌ها، شبکه‌های مکان-زمانی گفته می‌شود. شبکه‌های پیچشی سه بُعدی[16] نمونه‌ای از این روش‌ها هستند که توسط جی و همکارانش (Ji et al, 2013) معرفی شدند. در این شبکه‌ها از فیلترهای سه بُعدی استفاده شده و عملیات کانولوشن و پولینگ[17]، در بُعد مکان و زمان به‌صورت هم‌زمان انجام می‌گیرد. بدیهی است که این شبکه‌ها تنها می‌توانند تعداد از پیش تعریف‌شده‌ای از فریم‌ها را پردازش کنند. توانایی پردازش ویدیو کلیپ‌هایی با طول زمانی کوتاه و مدل‌سازی کوتاه مدت الگوهای حرکتی از جمله معایب این شبکه‌ها است. به منظور دخیل کردن اطلاعات زمانی درون ویدیو برخی از روش‌ها از دسته‌ی دیگری از شبکه‌های مکان-زمانی مثل شبکه‌های عصبی بازگشتی[18] (RNN) (Du et al, 2015) و یا به طور خاص‌تر حافظه‌های طولانی کوتاه-مدت[19] (LSTM) (Donahue et al, 2015; Srivastava et al, 2015) استفاده می‌کنند.

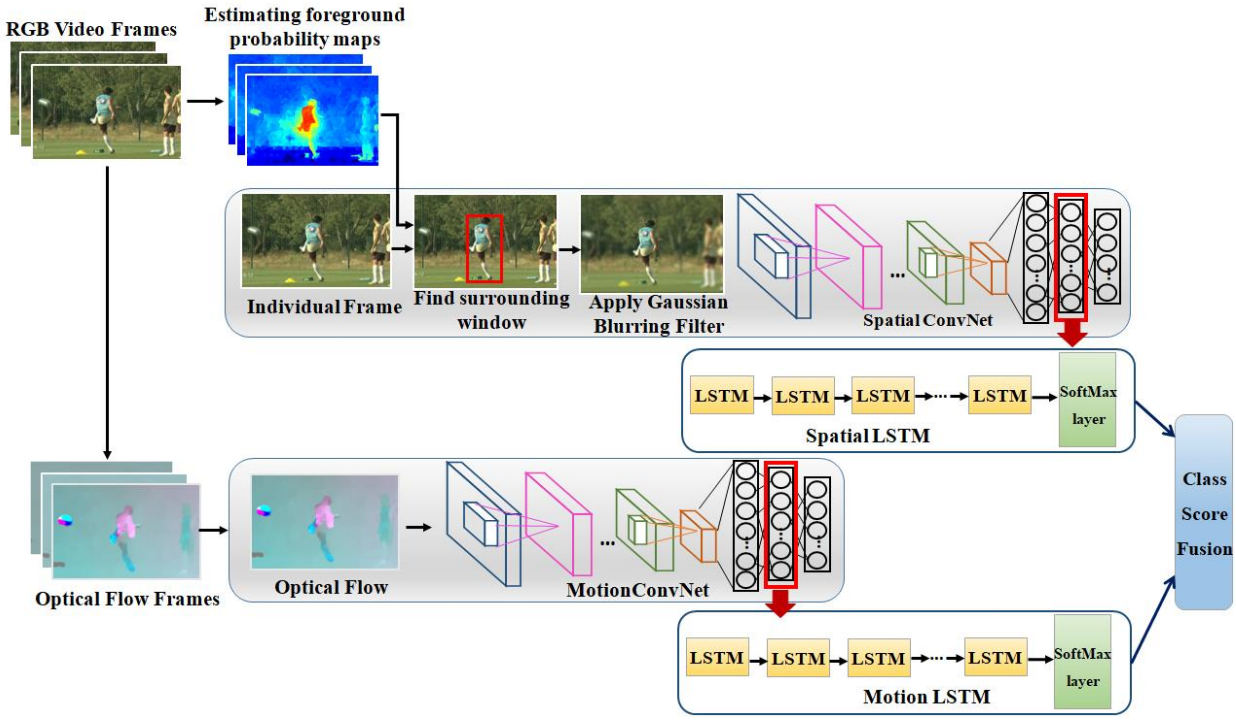
[15] Motion Boundary Histogram
[16] 3D Convolutional Neural Networks
[17] Pooling
[18] Recurrent Neural Networks
[19] Long Short Term Memory

از جمله مشکلات رویکردهای دوجریانه این است که اطلاعات حرکتی به طور جداگانه از اطلاعات ظاهری مورد پردازش قرار می‌گیرند. همچنین چگونگی ادغام این دو جریان اطلاعات برای پیش‌بینی نهایی همچنان از جمله مسئله‌های باز این حوزه است. با این‌حال، در مقایسه با رویکردهای دوجریانه، هر دو نوع شبکه‌ی سه بُعدی و بازگشتی، پردازش‌های بسیار سنگینی داشته و تعداد پارامترهای آن‌ها نیز برای آموزش، بسیار زیاد است. لذا برای آموزش چنین شبکه‌هایی نیاز به مجموعه داده‌های بسیار بزرگ وجود دارد. از آنجا که تولید چنین مجموعه داده‌هایی برای کاربردهای پردازش ویدیو بسیار طاقت‌فرسا و هزینه‌بر است، نیاز به روش‌هایی وجود دارد که بتواند اطلاعات زمانی را به شکلی بهینه و بدون نیاز به داده‌ی آموزشی بزرگ، در روند پردازش دخیل کند. لذا در این مقاله روشی ارائه می‌شود که اطلاعات زمانی را به شیوه مناسب در پردازش‌ها دخیل کرده و در عین حال نیاز به داده‌ی کمتری برای آموزش نسبت به شبکه‌های مکان-زمانی گفته شده دارد. همچنین، از آن‌جایی که این شبکه‌ها توانایی تمایز بین اطلاعات مربوط به پیش‌زمینه، یعنی فعالیت صورت گرفته و اطلاعات پس‌زمینه را ندارند، بازنمایی‌ای ارائه خواهیم کرد که به کمک ترکیب اطلاعات حرکتی و ظاهری، ورودی‌های حاوی اطلاعات مفیدتری را برای پردازش در اختیار این شبکه‌ها قرار خواهد داد.

## چارچوب مدل پیشنهادی برای شناسایی فعالیت

هدف از پژوهش‌های انجام شده در این مقاله، چیزی فراتر از تنها شناسایی و دسته‌بندی فریم‌های ویدیویی بوده است. لذا روشی پیشنهاد شده که با بهره‌گیری از اطلاعات ظاهری و حرکتی فریم‌ها، سعی در شناسایی نواحی‌ای از فریم‌های ویدیو را دارد که احتمال وقوع فعالیت در آن مناطق وجود دارد. چنین روشی، نیاز به مجموعه داده‌های حاشیه‌نویسی شده برای آموزش سیستم، که در آن فعالیت‌های درون هر فریم به صورت دستی توسط کاربر مشخص می‌شود، را از بین می‌برد. با تشخیص مناسب مناطق مربوط به رخداد فعالیت در هر فریم، می‌توان ورودی‌های حاوی اطلاعات مفیدتری را در اختیار روش‌های توصیف ویدیو مثل CNN قرار داده و در نهایت دقت شناسایی نیز نسبت به پردازش فریم‌های خام ویدیو بهتر خواهد بود.

به طور خلاصه، روش پیشنهادی از پنج مرحله اصلی تشکیل شده است: در مرحله اول، به کمک اطلاعات ظاهری و حرکتی درون فریم‌ها، تخمینی از نواحی بالقوه وقوع فعالیت در فریم و یا به بیانی دیگر نقشه‌ی برجستگی فریم بدست آمده و به هر پیکسل درون فریم عددی متناسب با میزان احتمال پیش‌زمینه بودن آن نسبت داده می‌شود. در مرحله دوم، ناحیه‌ی مربوط به رخداد فعالیت به وسیله‌ی میزان احتمالات بدست آمده در مرحله قبل انتخاب شده و سعی می‌شود بازنمایی‌ای ارائه شود که اطلاعات مربوط به پیش‌زمینه با تأثیر بیشتری نسبت به پس‌زمینه، در پردازش‌ها دخیل شوند. سپس در مرحله سوم تصاویر بدست آمده از مرحله‌ی قبل در اختیار یک مدل CNN دو جریانه، یکی برای تصاویر رنگی و دیگری برای تصاویر جریان نوری، برای توصیف قرار گرفته و ویژگی‌های هر فریم استخراج می‌شود. در مرحله چهارم، از یک شبکه عصبی بازگشتی برای ادغام توصیفگرهای فریمی، با در نظر گرفتن روابط زمانی بین آن‌ها استفاده می‌شود و در نهایت در مرحله‌ی آخر، برچسب نهایی مربوط به هر ویدیو اختصاص داده می‌شود. شکل 1 نمایی کلی از چارچوب روش پیشنهادی را نشان می‌دهد.




**شکل 1- نمایی کلی از چارچوب روش پیشنهادی. برای بدست آوردن بازنمایی ویدیو از دو جریان اطلاعاتی ظاهری و حرکتی استفاده می‌شود.**

### بدست آوردن نقشه‌ی احتمالاتی فعالیت

برای برآورد میزان احتمال پیش‌زمینه بودن هر پیکسل درون فریم، از روشی مشابه (Papazoglou and Ferrari, 2013) که برای مسئله‌ی قطعه‌بندی اشیا در ویدیو ارائه شده، استفاده کرده‌ایم. برای این کار، به طور همزمان از اطلاعات و ویژگی‌های برجسته‌ی مکانی و زمانی در طی فریم‌های مختلف ویدیو استفاده می‌شود. در مرحله‌ی اول، بردار جریان نوری بین دو فریم متوالی برای کل ویدیو تخمین زده شده و سپس بر اساس اطلاعات حرکتی هر پیکسل از تصویر و همسایگانش مرزهای حرکتی به‌دست‌آمده و تصمیم گرفته می‌شود که آیا آن پیکسل در درون مرز قرار دارد یا خیر. منظور از مرزهای حرکتی، پیکسل‌هایی از فریم هستند که جریان نوری در آن نقاط تغییرات ناگهانی داشته است. برای بدست آوردن مرزهای حرکتی درون هر فریم، دو شرط کلی در نظر گرفته می‌شود. اولاً، اندازه گرادیان جریان نوری در هر پیکسل، تا حدی می‌تواند میزان احتمال پیش‌زمینه بودن آن نقطه را نشان دهد. اگر $\overrightarrow{f_p}$ بردار جریان نوری در پیکسل $p$ باشد، برای به دست آوردن مرزهای حرکتی می‌توان از اندازه‌ی گرادیان جریان نوری در هر نقطه به‌صورت زیر استفاده کرد (Papazoglou and Ferrari, 2013):

$$b_p^m = 1 - \exp(-\lambda^m \| \overrightarrow{\nabla f_p} \|) \qquad b_p^m \in [0,1] \tag{1}$$

$\lambda^m$ پارامتر کنترلی شیب تابع است و $b_p^m$ شدت تغییرات ناگهانی در جریان نوری هر پیکسل و یا به بیانی دیگر میزان مرز حرکتی بودن هر پیکسل را نشان می‌دهد. هر چه این مقدار به یک نزدیک‌تر باشد احتمال مرز بودن بیشتر خواهد بود. این شرط در برابر نقاطی که دارای اندازه‌ی $b_p^m$ میانه هستند و در اطراف مقدار $0.5$ قرار دارند، دقت خوبی نخواهد داشت. چرا که ممکن است به دلیل خطا در محاسبات جریان نوری، این نقاط به اشتباه مرز و یا پس‌زمینه در نظر گرفته شوند. لذا شرط دومی اعمال می‌شود که تفاوت جهت حرکت هر پیکسل با همسایگی‌های آن را نیز در نظر می‌گیرد. با این ایده که اگر یک پیکسل

جهت حرکتش با تمامی همسایگی‌هایش متفاوت باشد در این صورت این پیکسل احتمالاً مرز حرکتی خواهد بود. شرط دوم به صورت رابطه 2 نوشته می‌شود.

$$b_p^{\theta} = 1-\exp(-\lambda^{\theta}\max(\theta_{p,q}^2)) \qquad b_p^{\theta} \in [0,1] \tag{2}$$

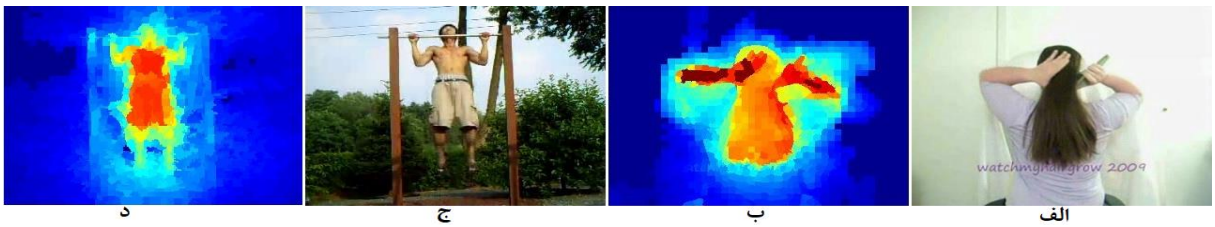
که در آن $\theta_{p,q}$ زاویه بین دو بردار $\overrightarrow{f_p}$ و $\overrightarrow{f_q}$ و $N$ نیز همسایگی‌های پیکسل $p$ را نشان می‌دهد. در نظر گرفتن این رابطه باعث می‌شود که الگوریتم پیشنهادی نسبت به حرکت‌های دوربین نیز مقاوم باشد.

با توجه به اینکه استفاده از هرکدام از این روش‌ها به‌تنهایی ممکن است خطاهایی را موجب شود، لذا از ترکیبی از این دو شرط استفاده می‌شود:

$$b_p = \begin{cases} b_p^m & \text{if } b_p^m > T \\ b_p^m . b_p^{\theta} & \text{if } b_p^m \le T \end{cases} \tag{3}$$

متغیر $T$، میزان آستانه‌ای است که در آن مقدار $b_p^m$ قابل‌قبول فرض شده است. درنهایت با اعمال یک آستانه با مقدار $0.5$ بر روی مقدار $b_p$، تصویر دودویی مرزهای حرکتی به دست خواهد آمد. . استفاده از اطلاعات حرکتی، به‌تنهایی برای به دست آوردن میزان احتمال پیش‌زمینه بودن، با توجه به خطاهایی که در محاسبه جریان نوری وجود دارد، کافی نیست. همچنین ممکن است بخشی از فرد و یا شیء در بخشی از فریم‌های ویدیو ثابت باشد. به‌عنوان‌مثال فقط دست فرد در تعدادی از فریم‌های ویدیو حرکت کند و باقی اعضای بدن فرد ثابت بماند. لذا نیاز به تعریف یک مدل ظاهری نیز برای بدست آوردن مقدار احتمالات پیش‌زمینگی وجود دارد. در نتیجه، یک مدل ظاهری به وسیله‌ی اعمال یک مدل چند گوسی مخلوط[20] (GMM) بر روی مقادیر تصاویر رنگی بدست می‌آید. پس از به دست آمدن مرزهای حرکتی یک نقشه از پیکسل‌های تصویر تهیه می‌شود که مشخص می‌کند آیا پیکسل درون مرز قرار دارد یا خارج مرز. این کار توسط مسئله‌ی نقطه در چندضلعی در هندسه‌ی محاسباتی حل می‌شود. برای این منظور از الگوریتم انتشار اشعه استفاده کرده، با این ایده که اگر نقطه داخل مرز باشد هر انتشار اشعه از آن، مرز را در تعداد فرد نقطه قطع کرده و در مقابل برای پیکسل‌های خارج از مرز، در تعداد زوج نقطه، مرز قطع خواهد شد.

برای استفاده‌ی مناسب از مدل حرکتی با توجه به اینکه در فریم‌های میانی ممکن است بخشی از بدن فرد ثابت باشد و فقط بخش کوچکی از فریم‌های میانی به‌عنوان پیش‌زمینه در نظر گرفته شود، پیشنهاد می‌شود که نقشه‌ی به‌دست‌آمده حاصل از داخل و یا خارج بودن هر پیکسل برای تمامی فریم‌های ویدیو انتشار داده شود. درنهایت با ترکیب (جمع احتمالات لگاریتمی) این دو مدل، میزان احتمال پیش‌زمینه بودن هر پیکسل را در اختیار خواهیم داشت. شکل 2 نقشه‌ی احتمالاتی بدست آمده برای دو دسته‌ی مختلف را نشان می‌دهد.




د ج ب الف

**شکل 2 – نمونه‌ای از دو فریم از دو دسته‌ی "Brush_hair" و "Pullup" و نقشه‌ی احتمالاتی بدست آمده. تصویرهای الف و ج فریم اصلی ویدیو و تصویرهای ب و د نقشه‌ی احتمالاتی بدست آمده را نشان می‌دهند.**

[20] Gaussian Mixture Model

**ایجاد بازنمایی‌های فریمی به کمک نقشه‌ی احتمالاتی**

همان‌طور که پیش‌تر نیز اشاره شد، در برخی از موارد، پس‌زمینه نیز می‌تواند تا حدی اطلاعات مناسبی از فعالیت رخ داده را در اختیار قرار دهد. روش‌های موجود در مسئله‌ی شناسایی فعالیت، از منظر استفاده از اطلاعات پس‌زمینه و پیش‌زمینه، به دو صورت عمل می‌کنند. در بخشی از این روش‌ها، اطلاعات مربوط به پس‌زمینه دخیل می‌شوند و در برخی دیگر، این اطلاعات در سیستم شناسایی مورد استفاده قرار نمی‌گیرند. آن دسته از روش‌هایی که اطلاعات پس‌زمینه را دخیل می‌کنند، همان وزنی را که به اطلاعات ناحیه‌ی هدف می‌دهند، برای اطلاعات پس‌زمینه در نظر می‌گیرند. آن دسته از روش‌هایی که این اطلاعات را دخیل نمی‌کنند نیز، باعث وابستگی دقت سیستم موردنظر به روش قطعه‌بندی استفاده شده می‌شوند. به این معنی که، اگر عمل قطعه‌بندی به‌درستی انجام نشود، سیستم با خطا مواجه خواهد شد و در عمل بخشی از اطلاعات پس‌زمینه را، به‌عنوان ناحیه‌ی هدف مورد پردازش قرار خواهد داد. اکثر کارهای انجام‌شده درزمینه‌ی شناسایی فعالیت‌های انسان در ویدیو، از روش‌هایی مثل یابش بدن فرد، ردیابی و یا تفریق پس‌زمینه استفاده می‌کنند. در تمامی این روش‌ها، اطلاعات مربوط به پس‌زمینه، به‌طور کامل از بین رفته و در روند محاسبات و نتیجه‌گیری نقشی ندارند. بااین‌وجود، دخیل کردن اطلاعات مربوط به پس‌زمینه، در استفاده از مجموعه داده‌های دنیای واقعی، بسیار مفید واقع می‌شود. چراکه برخی از ویژگی‌های پس‌زمینه، با فعالیت‌های انجام‌شده در پیش‌زمینه، از هم‌بستگی بسیار بالایی برخوردارند. به‌عنوان‌مثال، پس‌زمینه‌ی مربوط به محیط برفی در شناسایی فعالیت مربوط به "اسکی‌بازی روی یخ"، در شناسایی بسیار مفید خواهد بود. بنابراین، نیاز به سامانه‌هایی وجود دارد که بتوانند محدوده‌ی ROI[21] را به‌صورت هوشمند، به‌گونه‌ای تعیین کنند که وزن بیشتری نسبت به اطلاعات نواحی پس‌زمینه بگیرد.

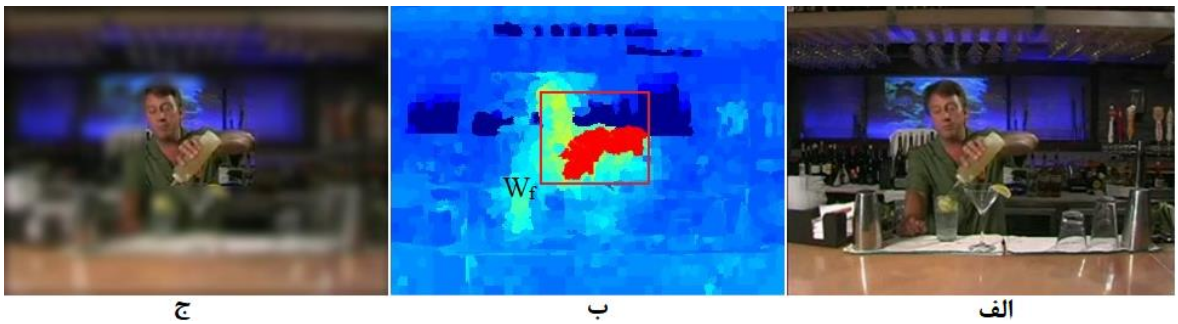
برای این منظور بازنمایی‌ای ارائه می‌کنیم که می‌تواند از اطلاعات پس‌زمینه نیز با وزن مناسبی در تصمیم‌گیری استفاده کند. در روش‌های معمول مکان‌یابی فعالیت، پس از استخراج ناحیه‌ی مربوط به رخداد فعالیت، این نواحی از پس‌زمینه جدا شده و نواحی پس‌زمینه به طور کلی از روند پردازش حذف می‌شوند. در روش پیشنهادی پس از تشخیص مکان وقوع فعالیت در هر فریم، با استفاده از نقشه‌ی احتمالات بدست آمده در مرحله قبل، برعکس روش‌های معمول مکان‌یابی فعالیت، این نواحی از پس‌زمینه جدا نخواهند شد. بلکه بازنمایی ارائه می‌شود که با آگاهی از نواحی مرتبط با فعالیت، از این اطلاعات با تأثیر بیشتری نسبت به پس‌زمینه استفاده می‌کند. در عین حال، اطلاعات کلی پس‌زمینه نیز تا حدی در پردازش‌ها تأثیرگذار خواهند بود.

در این بازنمایی، پیکسل‌هایی با احتمال پیش‌زمینگی $\Pr_i$، بزرگ‌تر از یک مقدار آستانه‌ی $T_p$ استخراج شده و سپس کوچک‌ترین پنجره‌ی $W_f$ که دربرگیرنده‌ی این پیکسل‌هاست، در نظر گرفته می‌شود. در مرحله‌ی بعد بازنمای جدیدی برای تصویر فریم ایجاد خواهد شد، که در آن مقدار هر پیکسل درون پنجره‌ی $W_f$، بدون تغییر عیناً در بازنمایی جدید درج خواهد شد و باقی پیکسل‌ها نیز به وسیله‌ی اعمال یک فیلتر گوسی مطابق با رابطه‌ی 4، با انحراف معیار $\sigma$، متناسب با نسبت عکس احتمال $\Pr_i$ و فاصله‌ی $D_{P_i-W_f}$ تا پنجره‌ی $W_f$ تار می‌شوند.

$$G(x,y)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{x^2+y^2}{2\sigma^2}} \qquad \sigma \sim D_{P_i-W_f} \sim \frac{1}{\Pr_i} \tag{4}$$

در نواحی‌ای که احتمال پیش‌زمینه بودن، از حدی بالاتر است، خود پیکسل و در باقی پیکسل‌ها، پارامتر سیگمای گوسی با نسبت عکس احتمال بزرگ‌تر شده و پیکسل موردنظر تارتر خواهد شد. این کار باعث می‌شود که مناطق پس‌زمینه تا حد زیادی تأثیر و قدرتشان در پاسخگویی به فیلترهای مختلف CNN و یا فیلترهای مربوط به توصیفگرهای ویژگی کم شده و عملاً در این مناطق ساختاری یافت نمی‌شود. در این شرایط جزئیات مربوط به پس‌زمینه‌ی فعالیت در روند شناسایی تأثیر نداشته در حالی‌که، اطلاعات کلی‌ای از آن مثل رنگ پس‌زمینه و ساختارهای کلی آن به روند شناسایی کمک خواهد کرد. شکل 3 بازنمایی وزن‌دهی شده برای دو فریم مختلف از دو دسته را نشان می‌دهد.

[21] Region Of Interests




الف ب ج

**شکل 3- تصویر بازنمایی ایجاد شده برای یک فریم از دسته‌ی "Pour" . الف: تصویر اصلی فریم، ب: پیکسل‌های انتخاب شده به عنوان نواحی با پتانسیل بالا برای رخداد فعالیت، ج: بازنمایی جدید وزن‌دهی شده. در این بازنمایی فقط کلیات مربوط به اطلاعات پس‌زمینه باقی مانده است.**

**ایجاد توصیفگرهای فریمی با استفاده از CNN**

در این پژوهش از CNN به منظور استخراج ویژگی استفاده شده است. به طور کلی، اطلاعات درون ویدیو می‌تواند به دو جریان اصلی، یکی برای اطلاعات مکانی و دیگری اطلاعات حرکتی تقسیم شود. جریان مکانی، اطلاعات مربوط به یک فریم را شامل می‌شود که حاوی اطلاعاتی از ظاهر صحنه و اشیا درون آن است. در حالی که جریان حرکتی، اطلاعاتی درباره نحوه‌ی حرکت اشیا و افراد درون صحنه در چند فریم متوالی را در برمی‌گیرد. در این مقاله نیز از دو شبکه‌ی CNN مشابه شکل 1 برای آموزش این اطلاعات استفاده شده است. به طور کلی، می‌توان گفت که آموزش شبکه‌های CNN به دلیل پارامترهای زیادی که دارند کاری بسیار سخت بوده و نیازمند مجموعه‌های آموزشی بسیار بزرگی است. این مسئله در کاربردهای پردازش ویدیویی نسبت به تصویر پررنگ‌تر بوده و نیاز به داده‌های بسیار زیادی برای آموزش و وزن‌دهی پارامترهای شبکه وجود دارد. لذا با توجه به اینکه آموزش وزن‌های شبکه از ابتدا نیاز به داده‌های بسیار زیادی دارد و ساخت داده‌های آموزشی بزرگ نیز برای مسئله شناسایی فعالیت کاری دشوار است، در اغلب موارد از شبکه‌های از پیش آموزش‌داده‌شده استفاده می‌شود (Varol et al, 2017).

در این مقاله نیز از مدل CNN ارائه شده در (Gkioxari and Malik, 2015) استفاده شده است. ساختار این شبکه ساختار شبکه‌ی AlexNet بوده که از پنج لایه‌ی کانولوشنی و سه لایه‌ی تماماً متصل[22] تشکیل شده است. در مدل پیشنهادی، از خروجی لایه‌ی تماماً متصل FC7 به منظور استخراج ویژگی‌های فریمی استفاده می‌شود. ورودی شبکه‌ی جریان مکانی، تصاویر رنگی فریم‌های مربوط به ویدیوها و ورودی شبکه‌ی زمانی نیز تصاویر جریان نوری بین فریم‌های ویدیو است. در نهایت ویژگی‌های استخراج شده از لایه‌ی FC7 مربوط به دو جریان، در اختیار یک شبکه‌ی LSTM قرار می‌گیرد، تا این شبکه بتواند با در نظر گرفتن روابط قوی‌تر زمانی بین فریم‌های مختلف، این ویژگی‌ها را به صورت بهینه با یکدیگر ادغام کرده و در نهایت توصیفگر مربوط به ویدیو ساخته شود.

**مدل‌سازی روابط زمانی و ادغام توصیفگرهای فریمی**

پس از تولید یک بازنمایی برای فریم‌های ویدیو مرحله‌ی بعدی، طبقه‌بندی و برچسب‌زنی ویدیو خواهد بود. همانطور که واضح است برای طبقه‌بندی ویدیو نیاز است که تمامی ویدیوها به بازنمایی‌هایی با طول یکسان تبدیل شوند. در اکثر کارهای انجام شده در این حوزه، پس از استخراج توصیفگرهای فریمی، برای ساخت بازنمایی ویدیو از عملگرهای آماری مانند میانگین‌گیری استفاده می‌شود. این کار باعث از دست رفتن حجم زیادی از اطلاعات می‌شود. لذا در این مقاله سعی شده روابط زمانی بین

[22] Fully-connected layer

فریم‌ها به منظور ادغام بهینه‌ی آن‌ها در نظر گرفته شود. برای مدل‌سازی بهتر روابط زمانی بین توصیفگرهای مختلف فریمی و ادغام آن‌ها از شبکه‌ی LSTM استفاده می‌کنیم.

از آن‌جایی که مدل CNN مربوط به جریان حرکتی، تنها قادر به تحلیل الگوهای حرکتی کوتاه مدت بین دو یا چند فریم متوالی است، از مدل LSTM که نوعی از شبکه‌های بازگشتی است، برای پردازش الگوهای حرکتی بلندمدت‌تر استفاده کرده‌ایم. شبکه‌های عصبی بازگشتی، نوعی از شبکه‌های عصبی مصنوعی هستند که ساختار مناسبی برای پردازش داده‌های متوالی مانند فریم‌های ویدیویی دارند. وجود حلقه‌های جهت‌دار در اجزا مختلف این شبکه‌ها، وجه تمایزی است که آن‌ها را مناسب برای چنین کاربردهایی می‌سازد. شبکه‌های LSTM نسخه‌ی بهبودیافته‌ی این شبکه‌های بازگشتی هستند که در سال 1997 برای اولین بار ارائه شده‌اند (Hochreiter and Schmidhuber, 1997). مدل LSTM می‌تواند از روابط و اطلاعات زمانی مربوط به داده‌های متوالی استفاده کرده و به وسیله‌ی ساختار بازگشتی‌ای که در واحدهای پنهان خود دارد، یک توالی از داده‌ی ورودی را به یک برچسب مشخص در خروجی تبدیل کند. هر کدام از واحدهای این مدل دارای یک سلول حافظه‌ی داخلی هستند که اطلاعات را در توالی زمان‌های مختلف در خود ذخیره کرده و قادر به بررسی تغییرات و تأثیرات واحدهای حافظه در طول زمان هستند. تمامی این سلول‌ها مدل LSTM را قادر به یادگیری روابط پیچیده و طولانی‌مدت زمانی می‌سازد که شبکه‌های بازگشتی معمولی قادر به یادگیری آن‌ها نیستند.

ساختار شبکه‌ی LSTM به گونه‌ای است که می‌تواند ایرادات اصلی شبکه‌های عصبی پیچشی را به خوبی پوشش دهد. عدم توانایی یادگیری روابط زمانی بین داده‌ها، از جمله‌ی این کاستی‌هاست. با این حال برای آموزش این شبکه‌های، با توجه به تعداد پارامترهای بسیار زیاد آن‌ها، نیاز به مجموعه داده‌های بزرگ‌تری وجود دارد. برای رفع این مشکل، در این پژوهش رویکردی ارائه شده که مدل‌های CNN و LSTM را با یکدیگر ترکیب کرده و سعی در استفاده از مزایای هر دوی این مدل‌ها دارد. به این معنی که برای پردازش تصاویر و ایجاد بازنمایی از مدل‌های CNN استفاده شده، چرا که تعداد پارامترهای این شبکه به نسبت شبکه‌های بازگشتی بسیار کمتر بوده و برای آموزش نیاز به داده‌های کمتری دارند. این شبکه‌ها می‌توانند با داده‌های آموزشی بسیار کمتر از شبکه‌های بازگشتی خروجی قابل‌قبول و خوبی را در اختیار قرار دهند. سپس برای در نظر گرفتن روابط زمانی بین فریم‌ها و ادغام بهینه‌ی این بازنمایی‌ها از شبکه‌ی LSTM استفاده می‌کنیم. درنتیجه مدل ایجاد شده هم در بُعد پردازش اطلاعات مکانی و هم زمانی عمیق بوده و داده‌ها مفهمومی‌تری در اختیار قرار می‌دهد.

برای ادغام CNN و LSTM از خروجی فعال‌ساز لایه‌ی FC7 شبکه‌ی CNN به ازای هر فریم استفاده می‌کنیم. سپس ویژگی‌های استخراج شده در اختیار شبکه‌ی LSTM قرار گرفته و خروجی نهایی خروجی لایه‌ی Softmax خواهد بود. اگر خروجی لایه‌ی FC7، برای توالی فریم‌های ویدیوی iام را به صورت رابطه‌ی 5 در نظر بگیریم:

$$x_{fc7}^{i} = [x_{fc7}^{i,1}, x_{fc7}^{i,2}, ..., x_{fc7}^{i,T}] \quad (5)$$

که در این رابطه $x_{fc7}^{i}$ خروجی لایه‌ی تماماً متصل برای ویدیوی iام و $T$ طول گام زمانی از فریم‌های ویدیو است که در هر بار مورد پردازش قرار می‌گیرند، دراین‌صورت خروجی نهایی را به صورت رابطه‌ی 6 می‌توان نوشت:

$$h_{fc7}^{i} = \mathrm{LSTM}(x_{fc7}^{i}) \quad (6)$$

در این رابطه $h_{fc7}^{i}$ خروجی نهایی LSTM پس از پردازش $T$ فریم متوالی است. در نهایت برچسب‌های نهایی، با استفاده از رابطه‌ی 7 بدست خواهد آمد.

$$y^{i} = \mathrm{softmax}(h_{fc7}^{i}) \quad (7)$$

مقدار $y^{i}$، میزان احتمال تعلق ورودی به هرکدام از دسته‌های موجود را، در اختیار قرار می‌دهد.

در نهایت برای ادغام دو جریان مکانی و زمانی از روش ادغام در سطح امتیاز دسته‌ها[23] استفاده شده است. برای این کار، امتیازهای خروجی لایه‌ی Softmax دو جریان را با یکدیگر جمع کرده و دسته‌ای با بیشترین احتمال، به عنوان برچسب داده‌ی مورد نظر انتخاب شده است.

### ارزیابی مدل پیشنهادی

ارزیابی روش پیشنهادی و تجزیه و تحلیل شیوه‌ی عملکرد آن در کاربردهای دنیای واقعی، مستلزم استفاده از مجموعه داده‌ای معتبر است. لذا در این تحقیق مجموعه داده‌ی jHMDB که یکی از مجموعه داده‌های چالشی در مسئله شناسایی فعالیت است، مورد استفاده قرار گرفته است. همچنین، برای سنجش روش پیشنهادی نیز، رویکردهایی مشابهی که بر روی این مجموعه داده بهترین نتایج را داشته‌اند، مورد بررسی قرار گرفتند (Wang and Schmid, 2013; Peng et al, 2014; Gkioxari and Malik, 2015; Chéron et al, 2015). در بخش‌های بعد، این مجموعه داده و نتایج روش پیشنهادی بر روی آن به تفصیل شرح داده می‌شوند.

### مجموعه داده‌ی jHMDB

اکثر مجموعه‌داده‌های موجود در حوزه‌ی شناسایی فعالیت، حاوی تصویرهای فریمی از فرد انجام‌دهنده‌ی فعالیت هستند که به صورت کامل و واضح قابل رؤیت بوده و مقیاس تقریبی فرد نیز در تصویرها تا حد زیادی یکسان و قابل پیش‌بینی است. در این مجموعه‌داده‌ها غیر قابل رؤیت بودن فرد انجام‌دهنده‌ی فعالیت توسط مانعی، به عنوان پیش‌فرض مسئله در نظرگرفته نشده است (Ferrari et al, 2008). مجموعه‌داده‌ی jHMDB یک زیرمجموعه از تمام تغییرات احتمالی در ظاهر و مقیاس فرد در ویدیو را تشکیل می‌دهد. این مجموعه‌داده، زیرمجموعه‌ای از مجموعه داده‌ی HMDB51 است و از 21 دسته‌ی مختلف تشکیل شده است. در این مجموعه داده ویدیوها به مدت زمان انجام فعالیت محدود شده‌اند. تعداد ویدیوهای موجود شامل 928 ویدیو است، که به طور تقریبی بین 36 تا 55 ویدیو برای هر دسته در نظر گرفته شده است. هر ویدیو کلیپ شامل 15 تا 40 فریم با ابعاد 240 × 320 است. به طورکلی این مجموعه‌داده به سه بخش مختلف برای آموزش و آزمون تقسیم شده، که در تمامی مقالات نیز این سه بخش مورد استفاده قرار گرفته و ارزیابی نهایی به صورت میانگین دقت بدست آمده بر روی این سه بخش است. در این مقاله نیز از این سه بخش برای آموزش و آزمون مدل پیشنهادی استفاده شده است. معیار ارزیابی مورد استفاده نیز میانگین دقت شناسایی هر ویدیو در تمامی دسته‌ها، بر روی سه بخش داده‌ی آموزش و آزمون است. در نهایت به هر ویدیو یک برچسب، متناسب با حداکثر مقدار امتیازی که توسط طبقه‌بندهای فعالیت در اختیار قرار گرفته، انتساب داده می‌شود.

### جزئیات پیاده‌سازی

برای پیاده‌سازی مدل پیشنهادی از دو کتابخانه‌ی Caffe و Keras استفاده شده است. همانطور که پیش‌تر نیز گفته شد برای استخراج ویژگی‌های فریمی از مدل دو جریانه‌ی CNN استفاده کرده‌ایم. ساختار شبکه‌ی مورد استفاده برای هر دو جریان اطلاعاتی، ساختار شبکه‌ی AlexNet است. با توجه به نتایج تحقیقاتی که در (Varol et al, 2017) نشان داده شده، آموزش شبکه‌ی عصبی از ابتدا، بر روی تصاویر رنگی نتایج خوبی را در اختیار قرار نمی‌دهد. لذا در این پژوهش، از شبکه‌ی از پیش آموزش داده‌شده‌ی AlexNet مشابه مدل (Gkioxari and Malik, 2015) برای مدل جریان مکانی استفاده شده است و به منظور انطباق شبکه با مسئله، وزن‌های لایه‌های آن تنظیم دقیق[24] شده‌اند. برای استخراج الگوهای حرکتی درون ویدیو نیز، از شبکه‌ی دیگری به عنوان جریان دوم اطلاعاتی استفاده شده که ساختار آن مشابه مدل حرکتی ارائه شده در

[23] Score level fusion
[24] Fine-tuning

مقاله‌ی (Gkioxari and Malik, 2015) است. پس از ارائه این مدل که بر روی تصویرهای جریان نوری مجموعه داده‌ی UCF101 آموزش داده شده، توسط پژوهش‌های مختلفی مورد استفاده قرار گرفته و توانسته اطلاعات مناسبی را با استفاده از پردازش تصاویر جریان نوری در اختیار قرار دهد. برای ساخت ورودی این شبکه، بردار جریان نوری بین دو فریم متوالی توسط روش (Brox et al, 2004) استخراج شده و سپس با استفاده از مقادیر مؤلفه‌های x و y و اندازه جریان نوری، تصویری سه بُعدی ایجاد می‌شود. در نهایت این تصویر در اختیار شبکه‌ی CNN مربوط به اطلاعات حرکتی قرار می‌گیرد. سپس ویژگی‌های فریمی، به وسیله‌ی استخراج خروجی لایه‌ی FC7 مربوط به دو شبکه‌ی مکانی و زمانی بدست خواهند آمد. خروجی این لایه، یک بردار با 4096 ویژگی برای هر کدام از دو بازنمایی ظاهری و حرکتی، به ازای هر فریم در اختیار قرار می‌دهد.

برای ادغام توصیفگرهای فریمی، از مدل LSTM با طول گام زمانی 10 فریم استفاده شده است. برای بدست آوردن برچسب نهایی ویدیو نیز، از مجموع امتیازهای دو طبقه‌بند[25] استفاده کردیم. حداکثر تعداد تکرار برای آموزش این مدل 15000 تکرار در نظر گرفته شده است.

**یافته‌های پژوهش**

همانطور که پیش‌تر نیز گفته شد، برای ارزیابی روش پیشنهادی از مجموعه داده‌ی jHMDB استفاده شده است. برای ارزیابی نیز از میانگین دقت سیستم در شناسایی هرکدام از دسته‌ها استفاده شده است.

به منظور ارزیابی جامع روش پیشنهادی، سعی شده، نتایج بدست آمده با روش‌های جدیدی که بهترین نتایج رو بر روی مجموعه داده‌ی مورد استفاده دارند، مقایسه شود. جدول **Error! Reference source not found.**1 دقت بدست آمده حاصل از اعمال روش پیشنهادی در طبقه‌بندی ویدیوهای مربوط به 21 دسته‌ی مجموعه داده‌ی jHMDB را در مقایسه با سایر روش‌ها نشان می‌دهد. دقت بیان شده در این جدول میانگین دقت شناسایی بر روی سه بخش مختلف داده‌های آزمون این مجموعه داده است.

**جدول 1- میانگین دقت شناسایی فعالیت‌های ویدیو بر روی سه بخش مجموعه داده‌ی jHMDB**

| | دقت شناسایی سیستم (٪) | | | | |
|---|---|---|---|---|---|
| روش پیشنهادی | P-CNN (Chéron et al, 2015) | Action Tube (Gkioxari and Malik, 2015) | FV+SFV (Peng et al, 2014) | IDT+FV (Wang and Schmid, 2013) | |
| **69.31** | 61.1 | 62.5 | 69.03 | 65.9 | jHMDB |

همانطور که از نتایج بدست آمده مشخص است، سیستم ارائه شده توانسته از دقت خوبی نسبت به سایر روش‌های موجود بر روی این مجموعه‌داده برخوردار باشد. بهبود نتایج، حاصل از چند مسئله است که در مدل پیشنهادی در نظر گرفته شده است. در روش پیشنهادی، برخلاف بسیاری از روش‌های موجود (Wang and Schmid, 2013, Peng et al, 2014)، به جای پردازش فریم‌های خام، سعی شده است تا اطلاعات مربوط به پس‌زمینه و نواحی بالقوه رخداد فعالیت شناسایی شوند. در عین‌حال، برخلاف روش‌های دیگر (Gkioxari and Malik, 2015; Chéron et al, 2015) اطلاعات پس‌زمینه از پردازش‌ها به طور کامل حذف نشده‌اند، بلکه بازنمایی‌ای ارائه کردیم که می‌تواند از اطلاعات و ساختار کلی پس‌زمینه به خوبی در تصمیم‌گیری استفاده کند. همچنین در نظر گرفتن روابط زمانی بین فریم‌ها به منظور ادغام بهینه‌ی آن‌ها، برای ایجاد بازنمایی نهایی ویدیو نیز، از دلایل دیگر بهبود روش ارائه شده است.

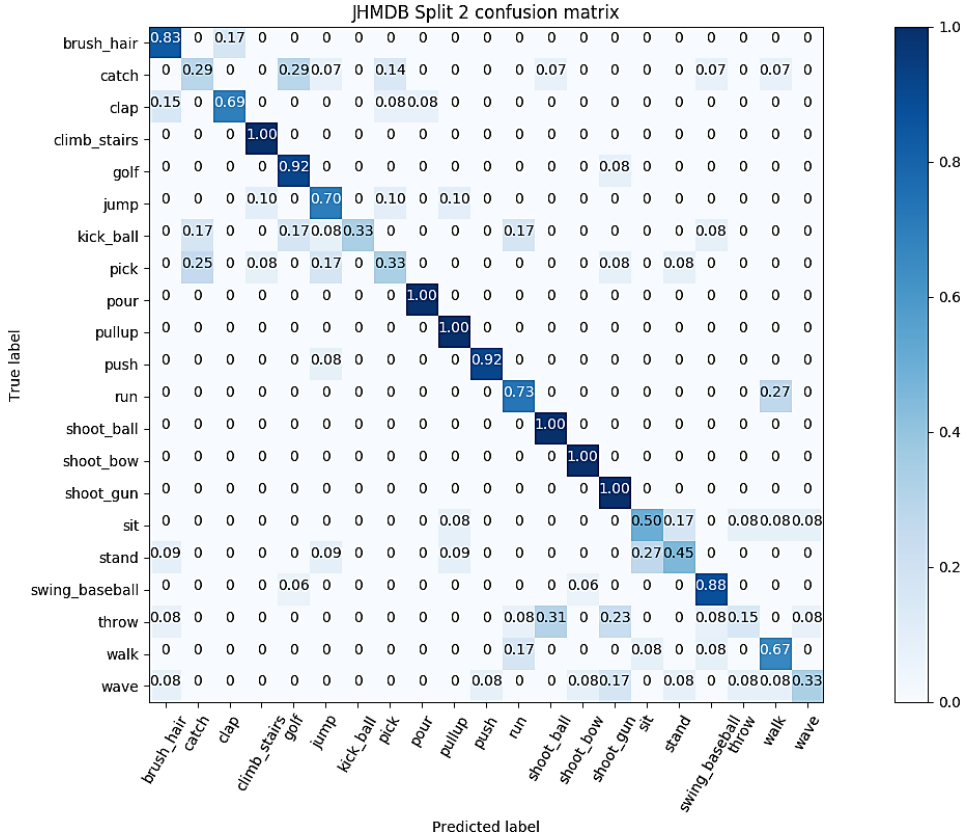
**Error! Reference source not found.** دقت سیستم در شناسایی فعالیت‌ها، به تفکیک برای دو بازنمایی ارائه شده و ترکیب آن‌ها، بر روی سه بخش مجموعه داده را نشان می‌دهد.

[25] Sum fusion method

**جدول 2- نتایج شناسایی فعالیت به تفکیک بازنمایی‌های مختلف بر روی بخش‌های مختلف مجموعه داده.**

| | Split1 | Split2 | Split3 | میانگین سه بخش |
|---|---|---|---|---|
| بازنمایی مدل ظاهری (تصویر وزن‌دهی شده) | 61.19 | 63.18 | 57.82 | 60.73 |
| بازنمایی مدل حرکتی (تصویر جریان نوری) | 53.05 | 53.67 | 54.05 | 53.59 |
| بازنمایی ادغام شده‌ی دو مدل | 67.91 | 71.48 | 68.53 | **69.31** |

بازنمایی مدل ظاهری، نتیجه‌ی دقت سیستم شناسایی را با استفاده از اطلاعات مربوط به جریان شبکه‌ی مکانی، یعنی استفاده از تصاویر وزن‌دهی تار شده به تنهایی، نشان می‌دهد. در مقابل ردیف دوم این جدول دقت شناسایی با استفاده از اطلاعات حرکتی درون ویدیو را نشان می‌دهد. در ردیف سوم نیز نتایج ادغام دو بازنمایی بر روی سه بخش از مجموعه داده نشان داده شده است. همانطور که از نتایج این جدول مشخص است، هر کدام از این بازنمایی‌ها به تنهایی دقت چندان مناسبی نداشته و ترکیب بهینه‌ی این دو نوع جریان اطلاعاتی توانسته دقت خوبی را در اختیار قرار دهد. شکل 4 نیز ماتریس سردرگمی[26] روش پیشنهادی حاصل از ادغام دو بازنمایی را بر روی بخش دوم مجموعه داده نشان می‌دهد.

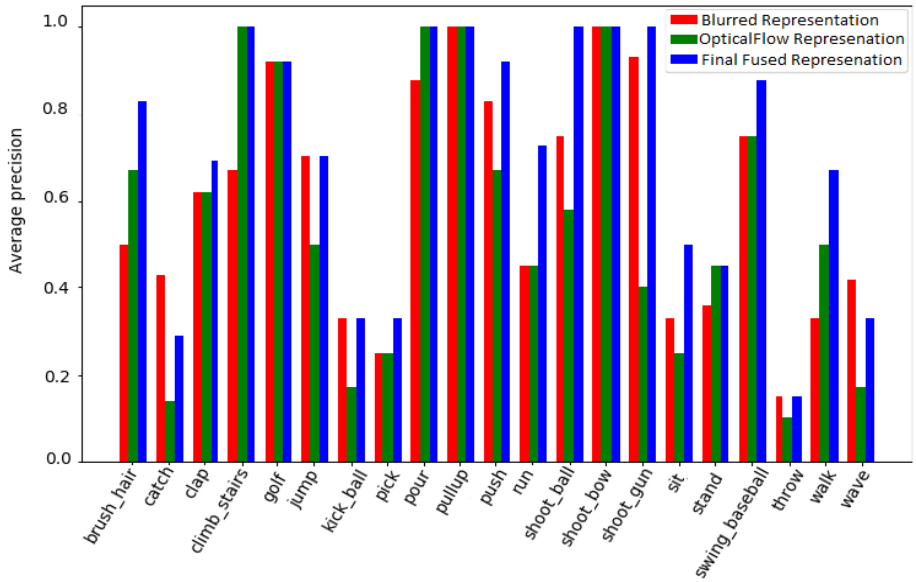



**شکل 4- ماتریس سردرگمی روش پیشنهادی بر روی بخش دوم مجموعه داده.**

نمودار 1 نیز، نمودار میله‌ای دقت شناسایی در هر دسته برای سه بازنمایی مختلف وزن‌دهی شده، بازنمایی حرکتی جریان نوری، و ادغام‌شده‌ی این دو را نشان می‌دهد. همانطور که از این نمودار مشخص است ادغام دو بازنمایی توانسته در اکثر دسته‌ها دقت شناسایی را به نسبت دو بازنمایی دیگر افزایش دهد. دلیل این مسئله نیز این است که این دو بازنمایی تا حد

[26] Confusion matrix

زیادی مکمل یکدیگر بوده و استفاده همزمان از اطلاعات ظاهری و حرکتی درون فریم‌های ویدیو می‌تواند در تصمیم‌گیری سیستم بسیار موثر باشد.




**نمودار 1- نمودار میله‌ای دقت شناسایی در هر دسته برای سه بازنمایی ارائه شده. قرمز: بازنمایی وزن‌دهی شده، سبز: بازنمایی حرکتی جریان نوری، آبی: ادغام شده‌ی مدل دو بازنمایی.**

## بحث و نتیجه‌گیری

در این مقاله مدلی برای حل مسئله شناسایی فعالیت‌های انسانی در داده‌های ویدیویی ارائه شده است. نتایج پژوهش‌های این مقاله، اهمیت استفاده از اطلاعات حرکتی درون ویدیو را برای شناسایی نواحی بالقوه‌ی رخداد فعالیت، در رویکردهای مبتنی بر یادگیری با ناظر ضعیف، که در آن فقط یک برچسب برای کل ویدیو در اختیار است، نشان می‌دهد. برای ایجاد یک بازنمایی حاوی اطلاعات مفید برای ویدیو، رویکردی ارائه شد که از اطلاعات حرکتی، در کنار اطلاعات ظاهری فریم‌ها به خوبی استفاده کرده و با استفاده از این اطلاعات توانستیم نواحی بالقوه را که احتمال وقوع فعالیت در آن‌ها وجود دارد، به خوبی تخمین بزنیم. در مدل پیشنهادی از دو جریان اطلاعاتی مکانی و زمانی از مدل‌های CNN استفاده شده است. برای ورودی جریان مکانی، بازنمایی‌ای ارائه کردیم که با استفاده از اطلاعات پیش‌زمینگی بدست آمده از نقشه‌ی برجستگی، به خوبی توانسته اطلاعات پیش‌زمینه را به طور کامل حفظ کرده و از ساختار و اطلاعات کلی پس‌زمینه در تصمیم‌گیری استفاده کند. از طرفی اطلاعات حرکتی باری دیگر با استفاده از تصاویر جریان حرکتی بین دو فریم متوالی به عنوان ورودی جریان زمانی، به طور مستقل مورد پردازش قرار گرفتند. در نهایت برای ادغام توصیفگرهای فریمی، از روشی استفاده کردیم که به خوبی توانست با در نظر گرفتن روابط زمانی بین فریم‌ها، این اطلاعات را به صورت بهینه با یکدیگر ادغام کند. برای این منظور از مدل LSTM استفاده کردیم. نتایج بدست آمده نشان داده است که استفاده‌ی بهینه از این دو نوع اطلاعات و ترکیب بهینه‌ی دو بازنمایی ارائه شده، توانسته دقت خوبی در شناسایی فعالیت‌های انسانی بر روی مجموعه داده‌ی jHMDB به نسبت سایر روش‌های موجود بر روی این مجموعه، در اختیار قرار دهد.

به عنوان پیشنهادی برای کارهای آینده، می‌توان از رویکردی جمعی[27] به منظور ادغام نتایج دو طبقه‌بند مختلف به جای استفاده از مجموع امتیازهای طبقه‌بندها، استفاده کرد. همچنین انباشتن تعدادی از فریم‌های جریان نوری، به عنوان ورودی

[27] Ensemble approach

جریان زمانی مدل CNN، می‌تواند نتایج این طبقه‌بند را بهتر کند. چرا که در این شرایط، شبکه ورودی‌هایی با الگوهایی حرکتی بلند مدت‌تری را دریافت کرده و لذا تحلیل الگوهای حرکتی درون ویدیو در این شرایط بهتر انجام خواهد شد.

**منابع**

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks", Paper presented at the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). "Action recognition by dense trajectories", Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.

Gkioxari, G., & Malik, J. (2015). "Finding action tubes", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 759-768).

Wang, L., Qiao, Y., Tang, X., & Van Gool, L. (2016). "Actionness estimation using hybrid fully convolutional networks", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2708-2717).

Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., & Snoek, C. G. (2014). "Action localization with tubelets from motion", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 740-747).

Jargalsaikhan, I., Little, S., & O'Connor, N. E. (2017, August). "Action localization in video using a graph-based feature representation", In Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on (pp. 1-6). IEEE.

Simonyan, K., & Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos", In Advances in neural information processing systems (pp. 568-576).

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). "Beyond short snippets: Deep networks for video classification", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). "Convolutional two-stream network fusion for video action recognition", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1933-1941).

Aggarwal, J. K., & Ryoo, M. S. (2011). "Human activity analysis: A review", ACM Computing Surveys (CSUR), 43(3), 16.

Weinland, D., Ronfard, R., & Boyer, E. (2011). "A survey of vision-based methods for action representation, segmentation and recognition", Computer vision and image understanding, 115(2), 224-241.

Herath, S., Harandi, M., & Porikli, F. (2017). "Going deeper into action recognition: A survey", Image and Vision Computing, 60, 4-21.

Soomro, K., & Zamir, A. R. (2014). "Action recognition in realistic sports videos", In Computer Vision in Sports (pp. 181-208). Springer International Publishing.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005, October). "Actions as space-time shapes", In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 2, pp. 1395-1402). IEEE.

Yilmaz, A., & Shah, M. (2005, June). "Actions sketch: A novel action representation", In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 984-989). IEEE.

Bobick, A. F., & Davis, J. W. (2001). "The recognition of human movement using temporal templates", IEEE Transactions on pattern analysis and machine intelligence, 23(3), 257-267.

Bregonzio, M., Gong, S., & Xiang, T. (2009, June). "Recognising action as clouds of space-time interest points", In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1948-1955). IEEE.

Laptev, I. (2005). On space-time interest points. International journal of computer vision, 64(2-3), 107-123.

# International Conference new study on Computer and it

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009, September). "Evaluation of local spatio-temporal features for action recognition", In BMVC 2009-British Machine Vision Conference(pp. 124-1). BMVA Press.

Liu, J., Luo, J., & Shah, M. (2009, June). "Recognizing realistic actions from videos "in the wild"", In Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on (pp. 1996-2003). IEEE.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008, June). "Learning realistic human actions from movies", In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.

Marszalek, M., Laptev, I., & Schmid, C. (2009, June). "Actions in context", In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2929-2936). IEEE.

Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T. S., & Li, J. (2009, June). "Hierarchical spatio-temporal context modeling for action recognition", In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2004-2011). IEEE.

Klaser, A., Marszałek, M., & Schmid, C. (2008, September). "A spatio-temporal descriptor based on 3d-gradients", In BMVC 2008-19th British Machine Vision Conference (pp. 275-1). British Machine Vision Association.

Dalal, N., & Triggs, B. (2005, June). "Histograms of oriented gradients for human detection", In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

Sipiran, I., & Bustos, B. (2011). "Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes", The Visual Computer, 27(11), 963-976.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005, October). "Behavior recognition via sparse spatio-temporal features", In Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on (pp. 65-72). IEEE.

Wang, H., & Schmid, C. (2013). "Action recognition with improved trajectories", In Proceedings of the IEEE international conference on computer vision(pp. 3551-3558).

Dalal, N., Triggs, B., & Schmid, C. (2006, May). "Human detection using oriented histograms of flow and appearance", In European conference on computer vision (pp. 428-441). Springer, Berlin, Heidelberg.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). "3D convolutional neural networks for human action recognition", IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

Wang, L., Qiao, Y., & Tang, X. (2015). "Action recognition with trajectory-pooled deep-convolutional descriptors", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4305-4314).

Fernando, B., & Gould, S. (2016, June). "Learning end-to-end video classification with rank-pooling", In International Conference on Machine Learning(pp. 1187-1196).

Chéron, G., Laptev, I., & Schmid, C. (2015). "P-CNN: Pose-based CNN features for action recognition", In Proceedings of the IEEE international conference on computer vision (pp. 3218-3226).

Du, Y., Wang, W., & Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1110-1118).

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). "Long-term recurrent convolutional networks for visual recognition and description", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015, June). "Unsupervised learning of video representations using lstms", In International Conference on Machine Learning (pp. 843-852).

Papazoglou, A., & Ferrari, V. (2013). "Fast object segmentation in unconstrained video", In Proceedings of the IEEE International Conference on Computer Vision (pp. 1777-1784).

Varol, G., Laptev, I., & Schmid, C. (2017). “Long-term temporal convolutions for action recognition”, IEEE transactions on pattern analysis and machine intelligence.

Hochreiter, S., & Schmidhuber, J. (1997). “Long short-term memory”, Neural computation, 9(8), 1735-1780.

Peng, X., Zou, C., Qiao, Y., & Peng, Q. (2014, September). “Action recognition with stacked fisher vectors”, In European Conference on Computer Vision (pp. 581-595). Springer, Cham.

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008, June). “Progressive search space reduction for human pose estimation”, In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). “High accuracy optical flow estimation based on a theory for warping”, Computer Vision-ECCV 2004, 25-36.