دوفصلنامه محاسبات و سامانه‌های توزیع شده
سال هفتم، شماره ۱، شماره پیاپی ۱۳، صفحه ۸۶ تا ۹۷، سال ۱۴۰۳

ترکیب ماشین‌های بولتزمن محدود دوبعدی و شبکه‌های LSTM برای شناسایی فعالیت‌های انسانی در ویدئو: یک رویکرد نوین

مجید جودکی(نویسنده مسئول)[۱]، حسین ابراهیم‌پور کومله[۲]
[۱] استادیار، گروه کامپیوتر، دانشکده مهندسی، دانشگاه آیت‌اله بروجردی، بروجرد
[۲] استادیار، دانشکده برق و کامپیوتر، دانشگاه کاشان، کاشان

چکیده

در حوزه تحلیل ویدئو به‌ویژه شناسایی فعالیت‌های انسانی، روش‌های پیشین علی‌رغم موفقیت‌های به دست آمده، در حفظ مستقیم ویژگی‌های فضایی بدون نیاز به پیش‌پردازش پیچیده و مدل‌سازی وابستگی‌های زمانی طولانی دچار محدودیت‌هایی بوده‌اند. در این پژوهش، یک معماری نوین مبتنی بر ترکیب ماشین‌های بولتزمن محدود دوبعدی و شبکه‌های LSTM ارائه می‌شود که با استخراج دقیق الگوهای فضایی از فریم‌های ویدئویی و مدل‌سازی مؤثر روابط زمانی، خلأ موجود در ادبیات پژوهشی را برطرف می‌کند. در روش پیشنهادی،ماشین‌های بولتزمن بدون نیاز به پیش‌پردازش‌های گسترده، ویژگی‌های مهم فضایی را از تصاویر استخراج نموده و شبکه‌های LSTM وابستگی‌های زمانی پیچیده را مدل‌سازی می‌کنند. نتایج تجربی حاصل از آزمایش بر روی مجموعه‌داده‌های KTH، UCF Sports و HMDB51 نشان از بهبود عملکرد با دقت‌های به ترتیب ۹۵.۳٪، ۹۳.۴٪ و ۷۰.۸٪ دارد که نسبت به روش‌های رقابتی موجود، کارایی قابل توجهی به اثبات رسانده است.

کلمات کلیدی:

یادگیری عمیق، ماشین بولتزمن محدود دو بعدی، شبکه‌های LSTM، شناسایی فعالیت انسانی، شبکه‌های عصبی بازگشتی

تاریخچه‌ی مقاله:

تاریخ ارسال: ۱۴۰۳/۰۱/۰۸
تاریخ اصلاحات: ۱۴۰۳/۰۵/۲۹
تاریخ پذیرش: ۱۴۰۳/۰۶/۲۱
تاریخ انتشار: ۱۴۰۳/۰۶/۳۰
ایمیل نویسنده مسئول: m.joudaki@gmail.com

۱. مقدمه

شناسایی فعالیت‌های انسانی[۱] در ویدئو، به دلیل گسترش دامنه محتوای تصویری در عصر حاضر، از اهمیت فزاینده‌ای برخوردار شده و به عنوان یکی از مسائل چالش‌برانگیز در پردازش تصویر و بینایی ماشین[۲] مطرح است. چالش‌هایی نظیر تغییر زاویه‌های دوربین، حضور موانع، شرایط نوری متغیر و پس‌زمینه‌های پیچیده، نیازمند رویکردهای پردازشی نوین و کارآمد هستند که بتوانند به صورت همزمان ویژگی‌های فضایی و وابستگی‌های زمانی را استخراج و مدل‌سازی کنند [۲-۱]. علاوه بر این، توسعه سریع مجموعه‌داده‌های ویدئویی، نیاز به مدل‌هایی با قدرت تفکیک و تعمیم‌دهی بالا را دوچندان کرده است [۳].

با ظهور فناوری‌های یادگیری عمیق و پیشرفت چشمگیر شبکه‌های عصبی مانند AlexNet در سال ۲۰۱۲ [۴]، بسیاری از مسائل موجود در شناسایی فعالیت‌های انسانی تا حد زیادی بهبود یافتند. به طور خاص، مدل‌های مبتنی بر شبکه‌های باور عمیق[۳] (DBN) و شبکه‌های حافظه کوتاه‌مدت بلند[۴] (LSTM) در استخراج ویژگی‌های پیچیده و مدل‌سازی توالی‌های زمانی موفق عمل کرده‌اند [۷-۵]. با این حال، دو محدودیت عمده در رویکردهای موجود مشاهده می‌شود:

- عدم حفظ مستقیم ویژگی‌های فضایی: بسیاری از روش‌ها برای پردازش تصاویر نیاز به پیش‌پردازش‌های پیچیده یا مسطح‌سازی داده‌ها دارند که موجب از دست رفتن اطلاعات مهم فضایی می‌شود.

[1] Human Action Recognition (HAR)
[2] Machine Vision
[3] Deep Belief Networks
[4] Long Short-Term Memory

- مدل‌سازی ناکافی وابستگی‌های زمانی بلندمدت: اگرچه شبکه‌های LSTM در یادگیری وابستگی‌های زمانی موفق بوده‌اند، اما برخی از معماری‌های موجود در شناسایی فعالیت‌های انسانی نتوانسته‌اند به طور جامع تغییرات زمانی کوتاه‌مدت و بلندمدت را مدل‌سازی نمایند.

شبکه‌های باور عمیق که با انباشتن ماشین‌های بولتزمن محدود[5] بر روی یکدیگر ساخته می‌شوند، به دلیل معماری لایه‌ای خود، در شناسایی الگوهای داده‌های پیچیده توانایی بالایی دارند [۹-۸]. در همین حال، LSTMها که نوعی شبکه عصبی بازگشتی[6] هستند، قادر به یادگیری وابستگی‌های بلندمدت بوده و برای وظایف پیش‌بینی توالی مناسب هستند [۱۰]. در این پژوهش، با هدف برطرف کردن خلأهای موجود در ادبیات پژوهشی و افزایش دقت شناسایی فعالیت‌های انسانی در ویدئوها، یک رویکرد نوین ارائه می‌شود که ترکیبی از ماشین‌های بولتزمن محدود دوبعدی (2D-RBM) و شبکه‌های LSTM است. در این مدل، 2D-RBM به صورت مستقیم فریم‌های ویدئویی را پردازش کرده و ویژگی‌های فضایی را بدون نیاز به پیش‌پردازش‌های گسترده استخراج می‌کند؛ سپس با استفاده از قابلیت‌های شبکه LSTM، وابستگی‌های زمانی کوتاه‌مدت و بلندمدت به صورت جامع مدل‌سازی می‌شود.

اهداف اصلی پژوهش به شرح زیر است:

- حفظ و استخراج دقیق ویژگی‌های فضایی از فریم‌های ویدئویی بدون نیاز به پیش‌پردازش پیچیده.
- مدل‌سازی مؤثر وابستگی‌های زمانی کوتاه‌مدت و بلندمدت جهت شناسایی فعالیت‌های پیچیده در توالی‌های ویدئویی.
- کاهش بار محاسباتی از طریق انتخاب هوشمند فریم‌های کلیدی که اطلاعات مفید جهت شناسایی حرکت را ارائه می‌دهند.

در این پژوهش، ما ماشین‌های بولتزمن محدود دوبعدی را با شبکه‌های LSTM ترکیب کرده‌ایم تا چالش‌های شناسایی فعالیت‌های انسانی در ویدئوها را برطرف کنیم. رویکرد استفاده از 2D-RBM اطلاعات فضایی را بدون نیاز به پیش‌پردازش گسترده یا مسطح‌سازی داده‌های تصویری حفظ می‌کند. با حفظ این اطلاعات، LSTM توانایی مدل را در شناسایی فعالیت پیچیده‌ای که در طول زمان رخ می‌دهند، افزایش می‌دهد. این یکپارچه‌سازی به مدل اجازه می‌دهد جنبه‌های فضایی، زمانی کوتاه‌مدت و بلندمدت فعالیت انسانی در ویدئوها را به صورت جامع درک کند. به طور سنتی، افزودن یک لایه softmax به معماری پیشنهادی عملکرد طبقه‌بندی را بهینه می‌کند، و ویژگی‌های استخراج‌شده را به کلاس‌های خروجی نگاشت می‌کند [۱۱].

علاوه بر این، برای افزایش کارایی مدل، از روش‌های انتخاب هوشمند فریم استفاده کرده‌ایم. به جای پردازش تمام فریم‌ها در یک توالی ویدئویی که می‌تواند از نظر محاسباتی سنگین باشد و شامل اطلاعات تکراری شود، فریم‌های کلیدی که بیشترین اطلاعات را برای شناسایی حرکت ارائه می‌دهند، انتخاب می‌شوند. این روش نه تنها بار محاسباتی را کاهش می‌دهد، بلکه تمرکز مدل را بر روی بخش‌های زمانی مرتبط‌تر معطوف می‌کند.

این رویکرد، علاوه بر ارتقای توانایی در تفکیک فعالیت‌های انسانی، عملکرد بهتری نسبت به روش‌های رقیب از نظر دقت طبقه‌بندی ارائه می‌دهد. ساختار مقاله شامل مرور ادبیات مرتبط و تحلیل خلأهای موجود، توضیح مفصل رویکرد پیشنهادی، جزئیات پیاده‌سازی شامل استراتژی انتخاب فریم و مراحل آموزش، و ارزیابی تجربی بر روی مجموعه‌داده‌های KTH [۱۲]، UCF Sports [۱۳] و HMDB51 [۱۴] به همراه مقایسه عملکرد با روش‌های موجود می‌باشد.

۲. مبانی تحقیق

در این بخش ابتدا روش‌های متفاوت شناسایی فعالیت انسان را بررسی می‌کنیم و سپس به توضیح فرمول‌ها و روابط ریاضی مرتبط با ماشین بولتزمن محدود خواهیم پرداخت.

۲-۱. مروری بر ادبیات گذشته

در این بخش، مروری جامع بر پیشرفت‌های صورت گرفته در حوزه شناسایی فعالیت‌های انسانی از طریق ویدئو ارائه می‌شود. مطالعات انجام‌شده به چند دسته موضوعی تقسیم می‌شوند که هر کدام نقاط قوت و ضعف خاص خود را دارند:

روش‌های مبتنی بر ویژگی‌های دستی: در مراحل اولیه تحقیقات در حوزه HAR، بیشتر پژوهش‌ها بر استخراج ویژگی‌های دستی و استفاده از الگوریتم‌های سنتی یادگیری ماشین متکی بودند. اگرچه این رویکردها در شرایط ساده موفق عمل می‌کردند، اما در مواجهه با تغییرات پیچیده نظیر تغییر زاویه‌های دوربین، شرایط نوری متغیر و

[5] Restricted Boltzmann Machines - RBMs

[6] Recurrent Neural Network - RNN

پس‌زمینه‌های پیچیده، از توانایی کافی در مدل‌سازی وابستگی‌های زمانی برخوردار نبودند [۱].

شبکه‌های عصبی عمیق: با ظهور یادگیری عمیق، شبکه‌های عصبی نظیر شبکه‌های باور عمیق (DBN)، شبکه‌های عصبی کانولوشنی(CNN) و شبکه‌های حافظه کوتاه‌مدت بلند (LSTM) به عنوان ابزارهایی قدرتمند برای استخراج نمایش‌های سلسله‌مراتبی از داده‌های خام معرفی شدند. به عنوان مثال، LSTM به دلیل قابلیت استخراج وابستگی‌های زمانی بلندمدت و حل مشکل گرادیان محوشونده، توانست عملکرد قابل توجهی در مدل‌سازی توالی‌های ویدئویی نشان دهد [۷]. همچنین، ماشین‌های بولتزمن محدود (RBM) به ویژه در قالب‌های کانولوشنی، توانستند ویژگی‌های فضایی مهم را بدون نیاز به پیش‌پردازش‌های پیچیده استخراج کنند [۶]. با این حال، اگرچه این مدل‌ها نسبت به روش‌های سنتی عملکرد بهتری دارند، در حفظ جامع اطلاعات فضایی و مدل‌سازی وابستگی‌های زمانی بلندمدت همچنان با محدودیت‌هایی مواجه‌اند.

مدل‌های ترکیبی: با هدف بهره‌گیری از نقاط قوت مدل‌های مختلف و رفع ضعف‌های آن‌ها، پژوهشگران به ادغام رویکردهای استخراج ویژگی‌های فضایی و مدل‌سازی توالی‌های زمانی پرداخته‌اند. برخی از مطالعات، با ادغام RBMهای کانولوشنی با مدل‌های زمانی، به استخراج همزمان ویژگی‌های فضایی و پویایی‌های زمانی کوتاه‌مدت دست یافته‌اند [۱۵]. در همین راستا، ترکیب ماشین‌های بولتزمن محدود دوبعدی (2D-Conv-RBM) با شبکه‌های LSTM جهت استخراج ویژگی‌های محلی از فریم‌های ویدئویی و سپس مدل‌سازی وابستگی‌های زمانی ارائه شده است [۹، ۱۶]. اگرچه این رویکردها نتایج بهبود یافته‌ای ارائه می‌دهند، اما همچنان چالش‌هایی مانند انتخاب بهینه فریم‌های ورودی برای کاهش افزونگی و بهبود کارایی مدل باقی مانده است.

رویکردهای انتخاب فریم هوشمند: برای کاهش بار محاسباتی و افزایش دقت، استراتژی‌های انتخاب هوشمند فریم به منظور استخراج فریم‌های کلیدی که اطلاعات بیشتری درباره حرکت ارائه می‌دهند، مطرح شده‌اند [۱۷]. این روش‌ها با حذف فریم‌های تکراری یا نامربوط، نه تنها محاسبات را بهینه می‌سازند بلکه تمرکز مدل را بر روی بخش‌های اطلاعاتی‌تر معطوف می‌کنند.

روش‌های مبتنی بر گراف: در برخی از پژوهش‌های اخیر، داده‌های فضایی-زمانی به‌وسیله مدل‌سازی گرافی مورد بررسی قرار گرفته‌اند. به عنوان مثال، شبکه‌های کانولوشن گرافی فضایی-زمانی جهت مدل‌سازی ساختارهای اسکلتی انسانی ارائه شده‌اند که روابط ساختاری و زمانی بین مفاصل بدن را استخراج می‌کنند [۱۸]. گرچه این رویکرد به‌ویژه در شناسایی فعالیت‌های مبتنی بر داده‌های اسکلتی موفق عمل کرده است، اما کاربرد آن در داده‌های تصویری خام نیازمند بررسی‌های بیشتری است.

در مجموع، علی‌رغم پیشرفت‌های چشمگیر در هر یک از این حوزه‌ها، چالش‌های اصلی در مدل‌سازی دقیق ویژگی‌های فضایی و وابستگی‌های زمانی بلندمدت در ویدئوها همچنان پابرجاست. رویکرد پیشنهادی این پژوهش، از طریق ادغام ماشین‌های بولتزمن محدود دوبعدی با شبکه‌های LSTM و بهره‌گیری از استراتژی‌های انتخاب فریم هوشمند، سعی در پر کردن این خلأها دارد. این یکپارچه‌سازی امکان استخراج دقیق ویژگی‌های فضایی از فریم‌های خام و مدل‌سازی جامع تغییرات زمانی را فراهم می‌آورد و با بهره‌گیری از منابع به‌روز، نشان‌دهنده عمق و به‌روز بودن مطالعه حاضر نسبت به رویکردهای پیشین است.

۲-۲. روابط ریاضی

معماری یک ماشین بولتزمن محدود باینری شامل واحدهای قابل مشاهده $\mathcal{V} = \{v_i\}, i \in (1, \dots, m\}$ و واحدهای پنهان $\mathcal{H} = \{h_j\}, j \in (1, \dots, n\}$, به نحوی که v_i و h_j به ترتیب حالت دودویی از واحد مرئی i و واحد مخفی j می‌باشند. توزیع احتمال مشترک برای واحدهای قابل مشاهده و پنهان به صورت رابطه (۱) تعریف می‌شود [۱۹]:

$$P(\boldsymbol{V}, \boldsymbol{H}) = \frac{1}{Z} e^{-E(\boldsymbol{V},\boldsymbol{H})} \qquad (۱)$$

که در آن Z تابع پارتیشن است که با جمع‌زدن بر روی انرژی همه حالات ممکن و به شکل رابطه (۲) محاسبه می‌شود همچنین تابع انرژی $\mathrm{E(V, H)}$ به صورت رابطه (۳) تعریف می‌شود:

$$Z = \sum_{\boldsymbol{V},\boldsymbol{H}} e^{-E(\boldsymbol{V},\boldsymbol{H})} \qquad (۲)$$

$$E(\boldsymbol{V}, \boldsymbol{H}) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i,j} v_i w_{ij} h_j \qquad (۳)$$

در اینجا، a_i و b_j بایاس‌های مرتبط واحدهای قابل مشاهده و پنهان هستند و w_{ij} وزن بین واحد قابل مشاهده v_i و واحد پنهان h_j می‌باشند. رابطه احتمال فعال بودن واحد پنهان h_j با توجه به واحدهای قابل مشاهده در رابطه (۴) آمده است:

$$p(h_j = 1|\boldsymbol{V}) = \sigma(b_j + \sum_{i=1}^{m} v_i w_{ij}) \quad (۴)$$

به طور مشابه، احتمال فعال بودن واحد قابل مشاهده v_i با توجه به واحدهای پنهان به صورت (۵) می‌باشد:

$$p(v_i = 1|\boldsymbol{H}) = \sigma(a_i + \sum_{j=1}^{n} h_j w_{ij}) \quad (۵)$$

که در آن $\sigma(x)=1/(1+e^{-x})$ تابع فعال‌سازی سیگموید است.

به دست‌آوردن یک تخمین بدون بایاس از مدل اندکی مشکل به نظر می‌رسد. این عمل باید به وسیله‌ی شروع از یک وضعیت تصادفی در نرون‌های ورودی و انجام نمونه‌برداری گیبس (پس از گذشت زمان زیاد) انجام گردد [۲۰]. یک مرحله‌ی نمونه‌برداری گیبس شامل به‌روزرسانی موازی تمام نرون‌های خروجی بر طبق رابطه‌ی (۴) و به‌روزرسانی موازی تمام نرون‌های مرئی بر طبق (۵) می‌باشد. روند یادگیری سریع‌تری نیز ارائه شده است که البته از نظر نتایج پایانی یکسان هستند [۲۱]. در روش سریع‌تر، داده‌های آموزشی را به نرون‌های ورودی اعمال شده و وضعیت باینری مربوط به نرون‌های خروجی به وسیله‌ی رابطه‌ی (۴) محاسبه می‌شود. هنگامی که بردارهای وضعیت باینری مربوط به نرون‌های مخفی به دست آمدند، عمل بازسازی آغاز می‌گردد. این روش را واگرایی متقابل[۷] می‌گویند. به صورت کلی آموزش RBM به شکل زیر انجام می‌شود:

احتمالی که مدل به بردار قابل مشاهده نسبت می‌دهد از رابطه (۱) به دست می‌آید و در نتیجه تابع هدف را می‌توانیم به شکل رابطه (۶) تعریف کنیم:

$$J_{w_{ij},a_i,b_j} = maximize\left(\frac{1}{m}\sum_{l=1}^{m} log \sum_{h} P\left(v^{(l)}, h^{(l)}\right)\right) \quad (۶)$$

برای محاسبه مقدار بیشینه تابع هدف، طبق رابطه (۷) از تابع هدف نسبت به w_{ij} مشتق گرفته می‌شود:

$$\frac{\partial}{\partial w_{ij}}\left(\frac{1}{m}\sum_{i=1}^{m} log \sum_{h} P\left(v^{(l)}, h^{(l)}\right)\right) = \frac{1}{m}\sum_{l=1}^{m}\sum_{h} X_{il} h_j P(h|v = x) - \sum_{v'}\sum_{h'} v'_i h'_j P(v', h') \quad (۷)$$

در رابطه (۷) تعداد حالات قابل محاسبه $P(v', h')$ نمایی بوده به همین دلیل محاسبه مشتق تابع هدف غیرممکن می‌نماید. بنابراین این مقدار به صورت رابطه (۸) تخمین زده می‌شود:

$$\frac{\partial\, log\, p(\boldsymbol{V})}{\partial w_{ij}} = \langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model} \quad (۸)$$

مقدار $\langle v_i h_j\rangle_{data}$ از حاصلضرب مقادیر حاصل از رابطه‌های (۴) و (۵) به دست می‌آید. برای محاسبه $\langle v_i h_j\rangle_{model}$ می‌توانیم از روش واگرایی متقابل استفاده کنیم که معادل روش نمونه‌برداری گیبز فقط با انجام یک مرحله می‌باشد. پس می‌توان رابطه (۸) را به شکل رابطه (۹) زیر بازنویسی کرد.

$$\frac{\partial\, log\, p(\boldsymbol{V})}{\partial w_{ij}} = \langle v_i h_j\rangle^0 - \langle v_i h_j\rangle^1 \quad (۹)$$

آموزش RBM شامل تنظیم وزن‌ها و بایاس‌ها برای کمینه کردن تفاوت بین توزیع داده و بازسازی داده‌ها توسط مدل است. فرمول‌های گرادیان‌های لگاریتم احتمال نسبت به وزن‌ها و بایاس‌ها به صورت رابطه (۱۰) تا (۱۲) هستند [۲۲]:

$$\Delta w_{ij} = \alpha\left(\langle v_i h_j\rangle^0 - \langle v_i h_j\rangle^1\right) \quad (۱۰)$$

$$\Delta a_i = \alpha(\langle v_i\rangle^0 - \langle v_i\rangle^1) \quad (۱۱)$$

$$\Delta b_j = \alpha\left(\langle h_j\rangle^0 - \langle h_j\rangle^1\right) \quad (۱۲)$$

در روابط بالا پارامتر α نرخ یادگیری بوده و به صورت دستی قابل تنظیم می‌باشد.

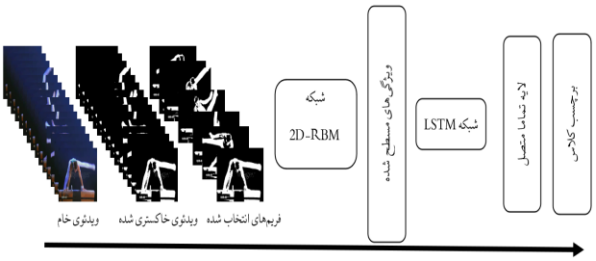
۳. روش پیشنهادی

در این بخش، معماری شبکه پیشنهادی را معرفی می‌کنیم که ماشین‌های بولتزمن محدود دوبعدی را با شبکه‌های حافظه کوتاه‌مدت بلند برای شناسایی فعالیت انسانی در توالی‌های ویدئویی ترکیب می‌کند. هدف اصلی این روش، پردازش مستقیم فریم‌های خام ویدئو بدون نیاز به پیش‌پردازش برای حفظ اطلاعات فضایی و مدل‌سازی کارآمد تغییرات فضایی و زمانی فعالیت انسانی است. دیاگرام کلی روش پیشنهادی در شکل ۱ ارائه شده است.

هر ویدئو از دنباله‌ای از فریم‌ها تشکیل شده است که با پردازش متوالی، اجرای یک فعالیت را نمایش می‌دهند. اولین مرحله در روش پیشنهادی اینست که تمامی فریم‌های ویدئو از رنگی به خاکستری تبدیل می‌شوند تا ابعاد داده کاهش یابد و پیچیدگی محاسباتی کمتر شود. تصاویر خاکستری اطلاعات ساختاری لازم را حفظ می‌کنند و داده‌ها را ساده‌تر می‌کنند. یکی از جنبه‌های کلیدی این روش،

[7] Contrastive Divergence

انتخاب هوشمند فریم‌ها برای ورود به شبکه است که موجب تمرکز شبکه بر روی بخش‌های زمانی مرتبط‌تر می‌شود.



شکل ۱: دیاگرام کلی روش پیشنهادی

۳-۱ : روش هوشمند انتخاب فریم

برای افزایش بهره‌وری محاسباتی، از یک استراتژی هوشمند انتخاب فریم استفاده می‌کنیم. به جای پردازش تمام فریم‌ها در یک توالی ویدئویی، فریم‌های کلیدی بر اساس تحلیل حرکت انتخاب می‌شوند. این الگوریتم شامل محاسبه تفاوت بین فریم‌های متوالی و انتخاب فریم‌هایی است که تفاوت آن‌ها از یک آستانه معین فراتر می‌رود. این روش با کاهش افزونگی، آموزش شبکه را بر روی فریم‌هایی متمرکز می‌کند که بیشترین سهم را در شناسایی حرکت دارند [۲۳]. برای محاسبه تفاوت (D_t) بین دو فریم متوالی F_t و F_{t-1} از رابطه (۱۳) استفاده می‌کنیم:

$$D_t = |F_t - F_{t-1}| \quad (۱۳)$$

سپس از مقدار تفاوت به‌دست آمده، برای محاسبه انرژی حرکت E_t که به شکل جمع مقادیر پیکسلی تفاوت محاسبه می‌شود، در رابطه (۱۴) استفاده می‌کنیم.

$$E_t = \sum_{i,j} D_t(i,j) \quad (۱۴)$$

فریم‌هایی که انرژی حرکت آن‌ها (E_t) از آستانه معین θ بالاتر باشد، به‌عنوان فریم کلیدی انتخاب می‌شوند. این آستانه براساس آماره‌های توزیع انرژی حرکت در ویدئو (E_t) تنظیم می‌شود. یکی از روش‌های رایج برای تنظیم آستانه، استفاده از میانگین و انحراف معیار انرژی حرکت است. برای محاسبه میانگین و انحراف معیار انرژی حرکت از رابطه‌های موجود در (۱۵) استفاده می‌شود :

$$\boldsymbol{\mu_E} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E_t}, \quad \boldsymbol{\sigma_E} = \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{E_t} - \boldsymbol{\mu_E})^2 \quad (۱۵)$$

در روابط بالا T تعداد فریم‌های ویدئو است. سپس، آستانه به‌صورت دینامیک و با توجه به میزان تغییرات حرکت در ویدئو مطابق رابطه (۱۶) تعیین می‌شود:

$$\boldsymbol{\theta} = \boldsymbol{\mu_E} + \boldsymbol{\alpha}.\boldsymbol{\sigma_E} \quad (۱۶)$$

α: یک ضریب قابل تنظیم که مشخص می‌کند چه مقدار بالاتر از میانگین به‌عنوان آستانه در نظر گرفته شود. این فرآیند تضمین می‌کند که تنها فریم‌هایی که تغییرات حرکتی مهمی دارند، وارد مراحل پردازش بعدی شوند. با کاهش تعداد فریم‌های پردازش‌شده، این روش افزونگی اطلاعاتی را کاهش داده و منابع محاسباتی را به سمت فریم‌های مرتبط‌تر هدایت می‌کند. این امر باعث می‌شود که مدل بر روی بخش‌های زمانی حاوی اطلاعات مفید متمرکز شود و در نتیجه دقت شناسایی فعالیت‌ها بهبود یابد.

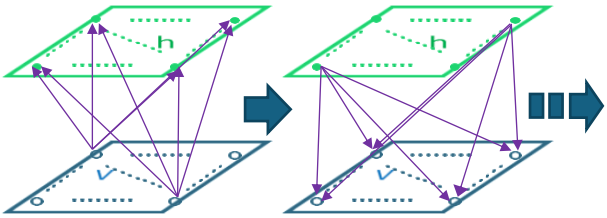
۳-۲ : ماشین بولتزمن محدود دوبعدی

همانطور که در شکل ۲ مشاهده می شود، ماشین بولتزمن محدود دوبعدی شامل یک لایه قابل مشاهده و یک لایه پنهان است که به‌صورت شبکه‌های دوبعدی مطابق با ابعاد فضایی فریم‌های ورودی سازمان‌دهی شده‌اند.

تابع انرژی 2D-RBM به صورت معادله (۱۷) تعریف می‌شود:

$$E(\boldsymbol{V},\boldsymbol{H}) = -\sum_{i,j}\left(a_{i,j}v_{i,j} + b_{i,j}h_{i,j} + \sum_{k,l} v_{i,j}w_{(i,j),(k,l)}h_{k,l}\right) \quad (۱۷)$$

که در آن V و H واحدهای مرئی و پنهان و $a_{i,j}$ و $b_{i,j}$ بایاس‌های لایه‌های مرئی و پنهان هستند. $w_{(i,j),(k,l)}$ وزن بین واحد مرئی $v_{i,j}$ و واحد پنهان $h_{k,l}$ است. با حفظ ساختار فضایی، 2D-RBM به‌طور مؤثری الگوهای محلی و بافت‌های موجود در فریم‌ها را که برای شناسایی فعالیت انسانی حیاتی هستند، استخراج می‌کند.



شکل-۲: شماتیک ۲ بعدی لایه‌های RBM

ویژگی‌های h_t استخراج‌شده توسط 2D-RBM در هر گام زمانی به شبکه LSTM وارد می‌شوند. LSTM دنباله‌ای از بردارهای ویژگی $\{h_1, h_2, \dots, h_T\}$ را پردازش می‌کند که در آن T تعداد فریم‌های انتخاب‌شده است، و الگوهای زمانی مرتبط با فعالیت مختلف را یاد می‌گیرد. شبکه LSTM از سلول‌های حافظه تشکیل شده است که اطلاعات را در طول زمان حفظ می‌کنند. این سلول‌ها توسط دروازه‌های ورودی (i_t)، خروجی(o_t) و فراموشی (f_t) کنترل می‌شوند و معادلات آن‌ها در (۱۳) تا (۱۷) آمده‌اند:

$$i_t = \sigma(W_i h_t + U_i c_{t-1} + b_i) \quad (۱۸)$$

$$f_t = \sigma\left(W_f h_t + U_f c_{t-1} + b_f\right) \quad (۱۹)$$

$$o_t = \sigma(W_o h_t + U_o c_t + b_o) \quad (۲۰)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh\,(W_c h_t + b_c) \quad (۲۱)$$

$$s_t = o_t \odot tanh\,(c_t) \quad (۲۲)$$

در فرمول های بالا c_t وضعیت سلول در زمان t، W وزن‌های مرتبط با ورودی فعلی h_t ، U وزن‌های مرتبط با وضعیت سلول، b بایاس، s_t وضعیت خروجی پنهان، σ تابع سیگموید و $\odot$ ضرب درایه‌ای می‌باشند. قدرت به خاطرسپاری ورودی‌های قبلی و یادگیری وابستگی‌های بلندمدت در LSTM توانایی شبکه را در شناسایی فعالیت پیچیده‌ای که در طول چندین فریم رخ می‌دهند، افزایش می‌دهد.

۳-۳ : لایه خروجی و طبقه‌بندی

در مرحله نهایی، خروجی‌های LSTM به یک لایه کاملاً متصل با تابع فعال‌سازی softmax -رابطه (۲۳)-منتقل می‌شوند تا توزیع احتمال بر روی کلاس‌های حرکت تولید شود:

$$y = \text{softmax}(W_s s_t + b_s) \quad (۲۳)$$

به نحوی که y بردار خروجی، نشان‌دهنده احتمال هر کلاس می‌باشد و W_s وزن‌های لایه softmax، s_t وضعیت مخفی نهایی شبکه LSTM در زمان t و b_s بایاس لایه خروجی هستند.

۳-۴ : فرآیند آموزش

روش پیشنهادی شامل دو مرحله آموزش است که به ترتیب، پیش‌آموزش بدون نظارت 2D-RBM و آموزش نظارت‌شده شبکه LSTM را در بر می‌گیرد. این مراحل برای بهبود عملکرد مدل در شناسایی حرکات انسانی طراحی شده‌اند. در پیش‌آموزش بدون نظارت ماشین بولتزمن محدود دوبعدی با استفاده از الگوریتم واگرایی متضاد[8] [۲۴] آموزش داده می‌شود تا توزیع ویژگی‌های فضایی در فریم‌های ورودی را یاد بگیرد. هدف این مرحله یادگیری ویژگی‌های باارزش فضایی از فریم‌های خام ویدئویی بدون استفاده از برچسب‌ها می‌باشد. پیش‌آموزش بدون نظارت به مدل کمک می‌کند تا شبکه با ویژگی‌های معنادار مقداردهی اولیه شود. این امر باعث بهبود یادگیری در مرحله آموزش نظارت‌شده بعدی می‌شود. یکی دیگر از مزایای این روش استفاده از وزن‌های مشترک برای کاهش تعداد کل پارامترها می‌باشد. این اشتراک وزن به مدل اجازه می‌دهد تا ویژگی‌های عمومی‌تر و تعمیم‌پذیرتری از داده‌های ورودی را بیاموزد. در مرحله آموزش نظارت‌شده شبکه LSTM، پس از استخراج ویژگی‌های فضایی توسط 2D-RBM، این ویژگی‌ها به شبکه LSTM وارد می‌شوند تا وابستگی‌های زمانی در بین فریم‌های ویدئویی مدل‌سازی شوند. هدف این مرحله یادگیری توالی زمانی حرکات برای شناسایی الگوهای کوتاه‌مدت و بلندمدت می‌باشد. وزن‌های شبکه پیشنهادی ترکیبی (2D-RBM + LSTM) با استفاده از الگوریتم پس‌انتشار خطا در طول زمان[9] [۲۵] تنظیم می‌شوند. در روند آموزش از تابع زیان آنتروپی متقاطع[10] برای کاهش تفاوت بین برچسب‌های پیش‌بینی‌شده و واقعی استفاده می‌شود. فرمول این تابع در رابطه (۲۴) آمده است.

$$L = -\sum_{k=1}^{K} y_k^{(true)}\, log\left(y_k^{(pred)}\right) \quad (۲۴)$$

که K تعداد کلاس‌های فعالیت موجود در دیتاست، $y_k^{(true)}$ مقدار برچسب واقعی برای کلاس k (۱ برای کلاس درست و ۰ برای بقیه) و $y_k^{(pred)}$ احتمال پیش‌بینی شده برای کلاس k می‌باشند.

۴. آزمایش‌ها و بحث

نتایج در این بخش، عملکرد مدل پیشنهادی ارزیابی می‌شود. آزمایش‌ها بر روی سه مجموعه‌داده استاندارد، KHT، UCFS ports و HMDB51 انجام شده‌اند. هدف اصلی این بخش، نشان دادن اثربخشی روش پیشنهادی در مقایسه با رویکردهای موجود تحت شرایط یکسان آزمایش است. برای اطمینان از مقایسه منصفانه، تمامی مجموعه‌داده‌ها به‌صورت یکنواخت پیش‌پردازش شده‌اند. تمام فریم‌های ویدئویی از رنگی به خاکستری تبدیل می‌شوند. این تبدیل رنگ باعث می‌شود داده‌ها از فرمت سه‌کاناله (RGB) به تک‌کاناله (خاکستری) ساده شوند و مدل بتواند بر اطلاعات حرکتی و ساختاری مرتبط با شناسایی فعالیت تمرکز کند.

۴-۱: مجموعه‌داده‌های KTH ، UCF Sports و HMDB51

مجموعه‌داده KTH شامل شش حرکت انسانی است: راه رفتن، دویدن، آهسته دویدن، بوکس، تکان دادن دست، و دست زدن. این مجموعه شامل ۲,۳۹۱ توالی ویدئویی است که با نرخ ۲۵ فریم در ثانیه و وضوح ۱۶۰×۱۲۰ پیکسل ضبط شده‌اند. ویدئوها با دوربین ثابت و پس‌زمینه‌های همگن ثبت شده‌اند و هر توالی تقریباً چهار ثانیه طول می‌کشد. مجموعه‌داده UCF Sports شامل ۱۰ فعالیت ورزشی است .این مجموعه شامل ۱۵۰ توالی ویدئویی با وضوح ۷۲۰×۴۸۰ پیکسل است. ویدئوها از کانال‌های تلویزیونی ضبط شده‌اند و شامل چالش‌های واقعی مانند پس‌زمینه‌های پیچیده،

[8] Contrastive Divergence
[9] Backpropagation Through Time - BPTT
[10] Cross-Entropy Loss Function

شـرایط نوری متغیر، تاری حرکت، انسـداد و زاویه‌های دید مختلف هسـتند. مجموعه‌داده HMDB51 یکی از دیتاسـت‌های پرکاربرد برای شـناسـایی حرکات انسـانی اسـت که شـامل ۵۱ کلاس حرکت مختلف است. این مجموعه‌داده شامل بیش از ۶۷۰۰ کلیپ ویدئویی اسـت که از منابع آنلاین متنوعی مانند فیلم‌ها و آرشـیوهای عمومی ویدئو جمع‌آوری شده‌اند و به همین دلیل از تنوع و پیچیدگی بالایی برخوردار اسـت. هر کلیپ ویدئویی سـناریوهای واقعی را با تنوع قابل‌توجه در پس‌زمینه، حرکت دوربین، نورپردازی، انسـداد و نحوه اجرای حرکات ضبط می‌کند. ویدئوها با وضوح ۲۴۰×۳۲۰ پیکسل و نرخ فریم‌های مختلف ضبط شده‌اند.

۴-۲ : ارزیابی روش پیشنهادی

عملکرد مدل پیشنهادی بر روی مجموعه‌داده‌های نام‌برده شده ارزیابی شده و با چندین روش مقایسه شده است. برای اطمینان از مقایسه منصفانه بر روی مجموعه داده‌های KTH وUCF Sports ، از تقسیمات آموزشی و آزمایشی مشابه با مطالعات پیشین [۶، ۱۴] استفاده شده است. مجموعه‌داده HMDB51 سه تقسیم‌بندی مجزا برای مقاصد آموزش و آزمون ارائه می‌دهد. نتیجه نهایی با میانگین‌گیری از نتایج طبقه‌بندی بر روی این تقسیم‌بندی‌های آموزشی و آزمایشی محاسبه می‌شود [۱۵]. این مجموعه‌داده به‌دلیل پیچیدگی و تنوع بالای خود از جمله تغییرات چشمگیر در پویایی حرکات، انسداد و نویز پس‌زمینه، چالش‌های قابل‌توجهی را ارائه می‌دهد.

۴-۳ : پارامترهای آموزشی

برای آموزش مدل پیشنهادی با توجه به مطالعات انجام شده در این حوزه، پارامترها مقداردهی شده‌اند. در جدول ۱ پارامترهای روش پیشنهادی به همراه مقادیر تنظیم شده آنها امده است.

۴-۴ : نتایج

در جدول ۲، دقت روش پیشنهادی و سایر روش‌ها بر روی مجموعه‌داده‌های KTH و UCF Sports ارائه شده است. روش پیشنهادی در مقایسه با سایر روش‌ها در هر دو مجموعه‌داده KTH و UCF Sports عملکرد برتری را نشان داده است.

جدول ۱: پارامترهای مدل پیشنهادی

پارامتر	مقدار / توضیحات
حد آستانه حرکت (θ)	منجر به انتخاب ۲۴ فریم شود.
ابعاد فریم	۶۴ در ۶۴
اندازه لایه مرئی 2D-RBM	۶۴ در ۶۴ (بر اساس فریم ورودی)
اندازه لایه مخفی 2D-RBM	۶۴ در ۶۴ (بر اساس ماتریس وزن)
اندازه ماتریس وزن 2D-RBM	۶۴ در ۶۴ در ۶۴ در ۶۴
تابع فعالیت	تابع سیگموید
نرخ یادگیری 2D-RBM	۰.۰۱
تعداد دفعات گیبز سمپلینگ	۲
تعداد لایه‌هایLSTM	۲
تعداد واحدهایLSTM	۲۵۶
نرخ یادگیری(LSTM)	۰.۰۰۱
بهینه‌ساز (LSTM)	بهینه‌سازAdam
اندازه دسته (Batch)	۱۶
تعداد دفعات تکرار آموزش(Epoch)	۱۰۰

بر روی مجموعه‌داده KTH، مدل با دقت ۹۵.۳۲٪ عملکرد بهتری نسبت به روش‌های پیشرفته مانند [۳۲، ۳۴] ارائه کرده است. این موفقیت به‌دلیل استفاده از ماشین بولتزمن محدود دوبعدی برای یادگیری ویژگی‌های فضایی به‌صورت بدون نظارت است که نیاز به مجموعه داده‌های بزرگ برچسب‌خورده را کاهش می‌دهد و شبکه را با ویژگی‌های معنادار اولیه‌سازی می‌کند. حفظ ساختار فضایی داده‌ها توسط 2D-RBM، توانایی مدل در تمایز بین حرکات پیچیده را افزایش داده است. علاوه بر این، بر روی مجموعه‌داده UCF Sportsکه شامل چالش‌هایی نظیر پس‌زمینه‌های پیچیده و زوایای دید متنوع است، مدل پیشنهادی با دقت ۹۳.۴٪ توانست روش‌های پیشرفته‌ای مانند[۲۹، ۳۸] را پشت سر بگذارد.

در جدول ۳، دقت روش پیشنهادی و سایر روش‌ها بر روی مجموعه‌داده‌های HMDB51 ارائه شده است. روش پیشنهادی با دقت ۷۰.۸ عملکرد بسیار رقابتی از خود نشان داده است و در رقابت با روش‌های پیشرفته مانند [۵۱، ۳، ۴۸] قرار دارد. این نتایج بیانگر توانایی روش پیشنهادی در استخراج ویژگی‌های فضایی-زمانی پیچیده و شناسایی حرکات انسانی در شرایط واقعی است. در مقایسه با روش‌هایی مانند [۴۳، ۴۵] مدل پیشنهادی دقت قابل‌توجه بالاتری ارائه می‌دهد که نشان‌دهنده مزایای استفاده از ترکیب یادگیری ویژگی‌های بدون نظارت با مدل‌سازی وابستگی‌های زمانی است. همچنین، روش‌های مبتنی بر یادگیری عمیق مانند [۵۱، ۵۳] دقت بالاتری نسبت به روش پیشنهادی دارند که می‌تواند ناشی از استفاده از معماری‌های پیچیده‌تر یا بهره‌گیری از اطلاعات اضافی باشد. با این حال، نتایج روش پیشنهادی به دلیل ساختار ساده‌تر و استفاده کارآمد از داده‌ها، بسیار قابل‌قبول است و نشان‌دهنده تعادل میان کارایی محاسباتی و دقت در شناسایی حرکات انسانی است.

جدول ۲: مقایسه مدل پیشنهادی با سایر روش‌های رقیب بر روی دیتاست های KHT و UCF Sports

مرجع	متد	دقت ٪	
		KTH	UCF Sports
Leptev et al. [۲۶]	Cuboids+HOG3D	۹۱.۴	۸۲.۶
Nazir et al. [۲۷]	3DHarris+3DSIFT	۹۱.۸۲	NG
Klaser et al. [۲۸]	Dense + HOG3D	NG	۸۵.۶
Rahmani et al. [۲۹]	Deep R-NKTM	NG	۹۰.۰
Niebles et al. [۳۰]	PLSA	۸۳.۳۳	NG
Jhuang et al. [۳۱]	HMAX	۹۱.۷	NG
Taylor et al. [۶]	3D GRBM	۹۰	NG
Le et al. [۳۲]	Hierarchical ISA	۹۳.۹	۸۶.۵
Ji et al. [۳۳]	3D CNN	۹۰.۲	NG
Sun et al. [۳۴]	3D (DL-SFA)	۹۳.۱	۸۶.۶
Zhang et al. [۳۵]	Dual-channel DNN	۹۲.۸	۸۶.۷
Han et al. [۳۶]	2Stream ConvNets	۹۳.۱	NG
Yuan et al. [۳۷]	3D Deep model	NG	۸۷.۳
Wang et al. [۳۸]	LSTM+CNN	NG	۹۱.۸۹
Ahmed et al. [۳۹]	STMEI-PCANet	NG	۸۶.۷
Abdelbaky et al. [۴۰]	ST-VLAD-PCANet	۹۳.۳۳	۹۰
روش پیشنهادی	2D-RBM+LSTM	۹۵.۳	۹۳.۴

NG: اطلاعات در دسترس نیست.

جدول ۳: مقایسه مدل پیشنهادی با سایر روش‌های رقیب بر روی دیتاست HMDB51

مرجع	متد	دقت ٪
Girdhar et al. [۴۱]	Pose regul. Atten. Pooling	۵۲.۲
Meng et al. [۴۲]	Spatio-temporal Atten.	۵۳.۱
Du et al. [۴۳]	C3D	۴۶.۷
Qiu et al. [۴۴]	P3D-199	۶۲.۹
Simonyan et al. [۴۵]	Two-stream CNN	۵۹.۴
Wang et al. [۴۶]	TDD	۶۳.۲
Feichtenhofer et al. [۴۷]	Two-stream fusion	۶۵.۴
Wang et al. [۴۸]	TSN	۶۹.۴
Yudistira et al. [۴۹]	TSN Cornet	۷۰.۶
Zong et al. [۵۰]	MSM-ResNets	۶۶.۷
Soomro et al. [۳]	ARTNet-Res18	۷۰.۹
Liu et al. [۵۱]	AMFNet-C	۷۱.۲
Du et al. [۵۲]	RSTAN (TSN)	۷۰.۵
Liu et al. [۵۳]	STILT	۷۲.۱
روش پیشنهادی	2D-RBM+LSTM	۷۰.۸

این نتایج نشان‌دهنده قدرت مدل در ترکیب یادگیری ویژگی‌های فضایی غنی و مدل‌سازی دقیق ویژگی‌های زمانی توسط شبکه LSTM است. شبکه LSTM توانسته با تحلیل وابستگی‌های کوتاه‌مدت و بلندمدت در توالی‌های ویدئویی، حرکات پیچیده انسانی را به‌طور دقیق شناسایی کند. در مقابل، روش‌های مبتنی بر ویژگی‌های دستی یا معماری‌های کم‌عمق به دلیل ظرفیت محدودشان در مدل‌سازی الگوهای پیچیده فضایی-زمانی، نتایج ضعیف‌تری ارائه کرده‌اند. همچنین، روش پیشنهادی با بهره‌گیری از استراتژی هوشمند انتخاب فریم‌ها، توانسته افزونگی داده‌ها را کاهش داده و بر فریم‌های اطلاعاتی‌تر متمرکز شود که این امر دقت و کارایی محاسباتی را به‌طور قابل‌توجهی بهبود بخشیده است.

۵. نتیجه‌گیری نهایی

ادغام ماشین‌های بولتزمن محدود دوبعدی با شبکه‌های LSTM، مسیر امیدوارکننده‌ای برای پیشرفت شناسایی فعالیت انسانی در ویدئوها ارائه می‌دهد. روش پیشنهادی با بهره‌گیری از نقاط قوت هر دو مدل - یادگیری ویژگی‌های فضایی بدون نظارت و مدل‌سازی توالی زمانی - الگوهای پیچیده فضایی-زمانی را به‌طور مؤثر شناسایی می‌کند.

کاربرد موفق این رویکرد در مجموعه داده‌های معیار، پتانسیل آن را برای وظایف تحلیل ویدئویی در دنیای واقعی نشان می‌دهد. با پرداختن به چالش‌های کلیدی در شناسایی فعالیت انسانی، این روش به توسعه تکنیک‌های یادگیری عمیق مقاوم‌تر و کارآمدتر کمک کرده و فرصت‌های جدیدی برای تحقیقات آینده در زمینه ترکیب یادگیری بدون نظارت و نظارت‌شده، بهینه‌سازی محاسباتی و گسترش کاربرد مدل‌های یادگیری عمیق در تحلیل پیچیده ویدئوها ایجاد می‌کند.

۶. منابع

[1] *R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, vol. 28, no. 6, pp. 976–990, 2010.*

[2] *D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," Computer Vision and Image Understanding, vol. 115, no. 2, pp. 224–241, 2011.*

[3] *K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.*

[4] *A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, 2012, pp. 1097–1105.*

[5] *G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.*

[6] *G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in Proceedings of the 11th European Conference on Computer Vision (ECCV), Heraklion, Greece, Sept. 5–11, 2010, pp. 140–153.*

[7] *J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.*

[8] *M. Joudaki, H. Ebrahimpour Komleh, Introducing a New Architecture of Deep Belief Networks for Action Recognition in Videos. Journal of Machine Vision and Image Processing. 2024; 11(1), 43-58.*

[9] *M. Joudaki, M. Imani and H.R. Arabnia, A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection. Technologies. 2025; 13, 53.*

[10] *S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.*

[11] *H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in Proceedings of the 25th International Conference on Machine Learning (ICML '08), Helsinki, Finland, July 5–9, 2008, pp. 536–543.*

[12] *C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Cambridge, UK, Aug. 23–26, 2004, pp. 32–36, doi: 10.1109/ICPR.2004.747.*

[13] *M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: a spatio-temporal maximum average correlation height filter for action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, June 24–26, 2008.*

[14] *H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in Proceedings of the*

[25] P. J. Werbos, "Backpropagation through time: what it does and how to do it," Proceedings of the IEEE, vol. 78, no. 10, pp. 1550–1560, 1990.

[26] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, June 24–26, 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587580.

[27] S. Nazir, M. H. Yousaf, and S. A. Velastin, "Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition," Computers and Electrical Engineering, vol. 72, pp. 660–669, 2018, doi: 10.1016/j.compeleceng.2018.01.028.

[28] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," International Journal of Computer Vision, vol. 79, no. 3, pp. 299–318, 2008, doi: 10.1007/s11263-007-0122-4.

[29] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, Oct. 14–21, 2007, pp. 1–8, doi: 10.1109/ICCV.2007.4408903.

[30] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, June 21–23, 2011, pp. 3361–3368, doi: 10.1109/CVPR.2011.5995513.

[31] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.

[32] L. Sun, K. Jia, T. H. Chan, Y. Fang, G. Wang, and S. Yan, "DL-SFA: Deeply-learned slow feature analysis for action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, June 23–28, 2014, pp. 2625–2632, doi: 10.1109/CVPR.2014.336.

[33] K. Zhang and L. Zhang, "Extracting hierarchical spatial and temporal features for human action recognition," Multimedia Tools and Applications, 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, Nov. 6, 2011. doi: 10.1109/ICCV.2011.6126543.

[15] N. Srivastava and R. Salakhutdinov, "Discriminative transfer learning with tree-based priors," in Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, Dec. 5–10, 2013, pp. 2094–2102.

[16] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, Feb. 4–9, 2017, pp. 4263–4270.

[17] Y. Zhao, Y. Xiong, and Z. J. Zha, "Recognizing human actions from still images with latent poses," in Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, July 23–27, 2018, pp. 1–6.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, Feb. 2–7, 2018, pp. 7444–7452.

[19] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science, vol. 7441, Springer, Berlin, Heidelberg, 2012, pp. 14–36.

[20] G. E. Hinton, S. Osindero and Y.W. The, (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527-1554.

[21] A. Fischer and C. Igel, (2014). Training restricted Boltzmann machines: An introduction. Pattern Recognition.

[22] A. M. Nickfarjam and H. Ebrahimpour-Komleh, (2019). Multi-input 1-dimensional deep belief network: Action and activity recognition as case study. Multimedia Tools and Applications, 78, 17739–17761.

[23] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion." Scandinavian conference on Image Analysis. Springer, Berlin, Heidelberg, 2003.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol. 14, no. 8, pp. 1771–1800, 2002.

Korea, Sept. 27, 2019. doi: 10.1109/ICCVW.2019.00368.

[43] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec. 13–16, 2015, pp. 4489–4497.

[44] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, Oct. 22, 2017. doi: 10.1109/ICCV.2017.588.

[45] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, pp. 568–576, 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf.

[46] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition (CVPR), Boston, MA, USA, June 7, 2015. doi: 10.1109/CVPR.2015.7299050.

[47] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 26, 2016. doi: 10.1109/CVPR.2016.213.

[48] L. Wang, X. Yuan, W. Zhe, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, Oct. 11, 2016. doi: 10.1007/978-3-319-46484-8_2.

[49] N. Yudistira and T. Kurita, "Correlation Net: Spatiotemporal multimodal deep learning for action recognition," Signal Process. Image Commun., vol. 115731, 2020. doi: 10.1016/j.image.2020.115731.

[50] M. Zong, R. Wang, and X. Chen, "Motion saliency based multi-stream multiplier ResNets for action vol. 77, no. 13, pp. 16053–16068, 2018, doi: 10.1007/s11042-017-4944-4.

[34] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, "Going deeper with two-stream ConvNets for action recognition in video surveillance," Pattern Recognition Letters, vol. 107, pp. 83–90, 2018, doi: 10.1016/j.patrec.2017.07.004.

[35] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in Proceedings of the 19th British Machine Vision Conference (BMVC), Leeds, UK, Sept. 1–4, 2008, pp. 275.1–275.10, doi: 10.5244/C.22.80.

[36] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 667–681, 2018, doi: 10.1109/TPAMI.2017.2657460.

[37] C. Yuan, X. Li, W. Hu, H. Ling, and S. Maybank, "3D R-transform on spatio-temporal interest points for action recognition," in Proceedings of the IEEE Conference on Pattern Recognition (CVPR), Boston, MA, USA, June 7, 2015. doi: 10.1109/CVPR.2015.7299050.

[38] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," IEEE Access, vol. 6, pp. 17913–17922, 2018, doi: 10.1109/ACCESS.2018.2820079.

[39] A. Ahmed and S. Aly, "Human action recognition using short-time motion energy template images and PCANet features," Neural Computing and Applications, vol. 32, no. 16, pp. 12561–12574, 2020, doi: 10.1007/s00521-020-05189-8.

[40] A. Abdelbaky and S. Aly, "Two-stream spatiotemporal feature fusion for human action recognition," The Visual Computer, vol. 37, pp. 1821–1835, 2021, doi: 10.1007/s00371-020-01913-y.

[41] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/aef4310f1a7bfc5067a8b202d50ff242-Paper.pdf

[42] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, "Interpretable spatio-temporal attention for video action recognition," in Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV), Seoul,

روش ارجاع: م. جودکی، ح. ابراهیم‌پور کومله، ترکیب ماشین‌های بولتزمن محدود دوبعدی و شبکه‌های LSTM برای شناسایی فعالیت‌های انسانی در ویدئو: یک رویکرد نوین، دوفصلنامه محاسبات و سامانه‌های توزیع شده، سال هفتم، شماره۱، شماره پیاپی۱۳، صفحه ۸۶ تا ۹۷، سال ۱۴۰۳.

How to cite: M. Joudaki, H. Ebrahimpour Komleh, Combining 2D Restricted Boltzmann Machines and LSTM Networks for Human Action Recognition in Videos: A Novel Approach. Journal of Distributed Computing and Systems (JDCS), Vol 7, Issue 1, Pages 86 - 98, 2024.

Combining 2D Restricted Boltzmann Machines and LSTM Networks for Human Action Recognition in Videos: A Novel Approach

M.Joudaki[1], H.Ebrahimpour Komleh[2]

[1]Department of Computer Engineering, Faculty of Engineering, Ayatollah Boroujerdi University, Boroujerd, Iran.
[2]Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

Abstract

In the field of video analysis, particularly in human action recognition, previous methods—despite their successes—have encountered limitations in directly preserving spatial features without resorting to complex preprocessing and in modeling long-term temporal dependencies. In this study, we propose a novel architecture based on the integration of 2D Restricted Boltzmann Machines (RBMs) and LSTM networks. This approach addresses the existing gap in the literature by accurately extracting spatial patterns from video frames and effectively modeling temporal relationships. In the proposed method, Restricted Boltzmann Machines extract important spatial features from images without the need for extensive preprocessing, while LSTM networks model the complex temporal dependencies. Experimental results on the KTH, UCF Sports, and HMDB51 datasets demonstrate improved performance with accuracies of 95.3%, 93.4%, and 70.8%, respectively, thereby establishing the significant effectiveness of the proposed approach compared to the current competitive methods.

recognition," Image Vis. Comput., vol. 107, p. 104108, 2021. doi: 10.1016/j.imavis.2020.104108.

[51] S. Liu and X. Ma, "Attention-driven appearance-motion fusion network for action recognition," IEEE Trans. Multimed., 2022. doi: 10.1109/TMM.2021.3088013.

[52] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," IEEE Trans. Image Process., vol. 27, pp. 1347–1360, 2018. doi: 10.1109/TIP.2017.2779836.

[53] *T. Liu, Y. Ma, W. Yang, W. Ji, R. Wang, and P. Jiang, "Spatial-temporal interaction learning based two-stream network for action recognition," Inf. Sci., vol. 606, pp. 864–876, 2022. doi: 10.1016/j.ins.2021.12.065*

مجید جودکی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۸۳ از دانشگاه شهید چمران اهواز و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر-هوش مصنوعی از دانشگاه صنعتی اصفهان در سال ۱۳۸۸ دریافت نمودند. ایشان دکترای کامپیوتر خود را در گرایش هوش مصنوعی از دانشگاه کاشان در سال ۱۴۰۲ اخذ نمودند. زمینه مورد علاقه ایشان یادگیری ماشین، بینایی کامپیوتر، پردازش تصویر و ویدئو با استفاده از شبکه های یادگیر عمیق می‌باشد.
نشانه رایانامه ایشان عبارتند از:

m.joudaki@gmail.com

حسین ابراهیم‌پورکومله مدرک دکترای خود را در سال ۲۰۰۴ از دانشگاه صنعتی کوئیزلند، استرالیا و مدرک پسا دکتری خود را در سال ۲۰۰۷ از دانشگاه نیوکاسل، استرالیا در رشته مهندسی کامپیوتر-هوش مصنوعی اخذ نمودند و در حال حاضر با مرتبه علمی استادیار مشغول به تدریس در دانشکده مهندسی برق و کامپیوتر دانشگاه کاشان می‌باشند. حوزه های تخصصی ایشان یادگیری ماشین، یادگیری ژرف، پردازش تصویر، بینایی ماشین، تئوری یادگیری، کلان داده، فرکتال و تئوری آشوب می‌باشد.
نشانه رایانامه ایشان عبارتند از:

ebrahimpour.kashanu@gmail.com