مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال یازدهم، شماره اول، بهار ۱۴۰۳

ارائه‌ی یک معماری جدید از شبکه‌های باور عمیق برای شناسایی عمل در ویدئو

مجید جودکی[۱]، حسین ابراهیم‌پور کومله[۲]

چکیده

استفاده از یادگیری عمیق در حل مسایل مربوط به تحلیل داده‌های پیچیده و حجیم مانند ویدئوها گسترش یافته است. از جمله پردازش-هایی که روی ویدیوها انجام می‌گیرد، تشخیص عمل‌های انسانی است که کاربردهای مهمی در حوزه نظارت خودکار، تعامل انسان با رایانه و بررسی رفتارهای سالمندان دارد. شبکه‌های باور عمیق از میان انواع مختلف شبکه‌های عمیق، به خاطر ویژگی‌های خاص خود، به ویژه توانایی همگرایی سریع نسبت به دیگر روش‌ها و ساختار یکسان لایه‌ها، مورد توجه قرار گرفته‌اند. لیکن، قدرت شبکه های باور عمیق پایه در پردازش داده‌های پیچیده که مبتنی بر زمان نیز هستند جای تامل دارد.

در این مقاله، یک روش بازگشتی جدید بر مبنای شبکه‌های باور عمیق ارائه شده است. در روش پیشنهادی، توانایی پردازش و تفسیر فریم‌های دوبعدی ویدئو و درک مفهوم زمان به وسیله پیاده‌سازی بازگشتی به شبکه‌های باور عمیق اضافه شده است. این روش قادر به درک مفاهیم کوتاه مدت زمانی با استفاده از ماشین‌های بولتزمن محدود و بلند مدت زمانی بر مبنای پیاده‌سازی بازگشتی می‌باشد. روش پیشنهادی بر روی سه پایگاه داده شناخته‌شده در این حوزه با نام‌های KTH، UCF و HMDB51 ارزیابی شده و به ترتیب به دقت‌های برابر با ۹۵,۰۲، ۹۳,۱٤ و ۷٤,۲۸ دست یافته و با سایر روش‌های محبوب در شرایط مختلف مقایسه گردیده است.

کلید واژه‌ها

یادگیری عمیق، شبکه‌های باور عمیق، ماشین‌های بولتزمن محدود، شناسایی عمل، شبکه‌های عصبی بازگشتی

۱- مقدمه

شناسایی عمل یکی از مهم‌ترین حوزه‌های بینایی ماشین است که به برچسب زدن بر عمل‌های انسان می‌پردازد. در طول دو دهه‌ی گذشته، شناسایی عمل توجه محققین زیادی را به خود جلب کرده است و بسیاری از کاربردهای بینایی ماشین در دنیای واقعی به آن ختم می‌شود [۱]. تجزیه و تحلیل عمل‌های انسان می‌تواند اطلاعات مفیدی را در حوزه‌های ذکر شده در اختیار محققان قرار دهد و این امر می‌تواند به کاربردهای صنعتی متعددی منجر شود. ولی، پیچیدگی‌هایی که در حوزه‌ی کار با ویدئو و شناسایی عمل‌های انسان وجود دارد همواره مشکل آفرین بوده است. از جمله‌ی این موارد می‌توان به پس‌زمینه‌های پیچیده[۱]، جهت‌های[۲] فیلم‌برداری، اشیای مزاحم[۳] یا وضعیت نور[٤] اشاره کرد.

با توجه به چالش‌های این حوزه، تمایل محققین به سمت روش‌های پیچیده‌تر شکل گرفته است. در ابتدا، روش‌های مختلف سعی در استخراج ویژگی‌های[٥] پیچیده‌تر برای یادگیری ویدئو

این مقاله در شهریورماه ۱۴۰۲ دریافت شد در آذرماه بازنگری در دی‌ماه پذیرفته گردید.

[۱] دانشجوی دکتری کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه کاشان
رایانامه: m.joudaki@gmail.com
[۲] گروه کامپیوتر دانشکده برق و کامپیوتر، دانشگاه کاشان
رایانامه: ebrahimpour.kashanu@gmail.com

نویسنده مسئول: حسین ابراهیم‌پور کومله

[1] Intricate Background
[2] Direction
[3] Clutter
[4] Light Condition
[5] Feature

داشتند. به همین سبب، روز به روز بر پیچیده‌تر شدن الگوهای یادگیری و ویژگی‌های استخراج شده افزوده می‌شد [۲].

یادگیری عمیق شامل مجموعه‌ای از مدل‌های یادگیری ماشینی[1] است که در دو حالت یادگیری بانظارت[2] و بدون نظارت در معماری سلسله مراتبی[3] عمیق قرار می‌گیرند. در شبکه‌های عصبی یادگیری با نظارت، ورودی می‌تواند خروجی را کلاسه‌بندی[4] کرده و بر اساس داده‌های قبلی که برچسب داشته‌اند؛ برچسب داده‌ی جدید را پیش‌بینی کند. هدف یادگیری عمیق این است که ماشین در یک روند سلسله مراتبی ویژگی‌های سطح بالا[5] را از ویژگی‌های سطح پایین[6] یاد بگیرد. به این معنا که در لایه‌های ابتدایی ویژگی‌های ساده‌ای مانند لبه‌ها[7] و خطوط و در لایه‌های بعدی ویژگی‌های سطح بالاتر بر مبنای این ویژگی‌ها یاد گرفته شود [۳]. یادگیری عمیق این مشکل اصلی در یادگیری را با معرفی بازنمایی‌هایی[8] که بر حسب سایر نمایش‌های ساده‌تر بیان می‌شوند، می‌شوند، حل می‌کند. یادگیری عمیق به رایانه اجازه می‌دهد تا مفاهیم پیچیده‌ای را از مفاهیم ساده‌تر بسازد. از بین انواع مختلف و محبوب شبکه‌های عمیق، شبکه‌های باور عمیق به دلیل سرعت بالای عملکرد و لایه‌های یکسانی که دارند، محبوب هستند.

با توجه به دو مزیت شبکه‌های باور عمیق در سرعت یادگیری و ساختار یکسان لایه به لایه، تلاشی در جهت درک همزمان سه وجه مکان[9]، مفاهیم کوتاه-مدت زمانی[10] و مفاهیم بلند-مدت زمانی[11] در این نوع شبکه‌ها انجام نگرفته است. این کار کمک می‌کند قدرت شبکه‌های باور عمیق در یادگیری داده‌های پیچیده و مبتنی بر زمان افزایش یابد. استفاده از ماشین‌های بولتزمن محدود دوبعدی، پیاده‌سازی دوبعدی شبکه باور عمیق و اضافه شدن مفهوم بازگشتی را می‌توان از نوآوری‌های روش پیشنهادی این مقاله دانست. درکنار مفهوم زمان، بایستی دریافت داده‌ی خام در این نوع از شبکه‌ها تقویت گردد تا شبکه توان یادگیری و حل مساله‌های مبتنی بر ویدئو را داشته باشد. این موضوع با استخراج فریم‌های مفید انجام گرفته است. در این مقاله تلاش می‌شود هر سه مفهوم مکان، مفاهیم کوتاه-مدت زمانی و مفاهیم بلند-مدت زمانی با ارایه‌ی یک شبکه‌ی بازگشتی[12] بر مبنای ماشین‌های بولتزمن محدود[13] درک و حل گردد.

ساختار مقاله به این صورت خواهد بود. در این بخش به تعریف مساله و پیشینه پژوهش و روش‌های مرتبط پرداخته می‌شود. مطالب اصلی شامل روش پیشنهادی با جزئیات در بخش دوم شرح داده می‌شود. در بخش سوم به ترتیب، آزمایشات انجام شده و نتایج به دست آمده از پیاده‌سازی روش پیشنهادی بیان می‌گردد. در بخش چهارم نتیجه‌گیری از تحقیق انجام شده بیان شده است.

تحقیقات گوناگونی در علوم رایانه به بررسی یادگیری عمیق در فهم ویدئو پرداخته است. یادگیری عمیق، رویکردی برای مواجهه با مشکلات ناشی از نیاز به حجم فراوان داده‌های برچسب خورده و نیاز به ویژگی‌های پیچیده می‌باشد که هدفش کمک به رایانه، برای دسته‌بندی دقیق خواهد بود.

روش‌های ارایه شده در حوزه‌ی شناسایی عمل به دو دسته‌ی کلی تقسیم می‌شوند:

کلاسه‌بندی مستقیم[14]: در این دسته از روش‌ها زمان نادیده گرفته می‌شود. در واقع، این روش‌ها سعی بر آن دارند با ارایه‌ی توصیفی از ویدئو به وسیله‌ی فریم‌های کلیدی یا توصیفات تصویری به شناسایی عمل بپردازند [۸]. نزدیک‌ترین همسایگی[15]، ماشین‌های برداری پشتیبان[16] و انواع دیگر روش‌های دسته‌بندی به ایجاد بستری برای شناسایی عمل در کلاسه‌بندی مستقیم تبدیل شده‌اند.

مدل‌های مبتنی بر زمان[17]: رویکرد روش پیشنهادی در این مقاله به عملکرد این نوع از روش‌ها نزدیک‌تر است. روش‌های مطرح در این دسته، از توصیفات مکان و زمان در کنار هم برای توصیف عمل استفاده می‌کنند. از جمله‌ی تمرکز این روش‌ها مدل‌های مارکوف[18]، فیلدهای تصادفی[19] و شبکه‌های عمیق هستند.

ونگ[20] و سوتر[21] در کار خود از مفهومی با عنوان محافظ پیش‌بینی‌های محلی[22] استفاده کرده‌اند که بعد را کاهش داده و از شبکه‌ی عصبی برای یافتن شباهت استفاده می‌کند. به عبارت دقیق‌تر، در کار آنها از مفهوم محافظ پیش‌بینی‌های محلی برای استخراج ویژگی و ارایه‌ی یک توصیف از ویدئو استفاده می‌گردد. سپس، از شبکه‌ی عصبی برای یافتن این که ویدئوی آزمون به کدام دسته نزدیک‌تر است استفاده خواهد شد [۹]. ژانگ[23] و همکاران روشی را ارایه کرده‌اند که از توصیف نقاط کلیدی بدن استفاده می‌کند. این توصیف به عنوان ویژگی استفاده شده و در اختیار ماشین بردار پشتیبان قرار می‌گیرد تا دسته‌ی مربوط به عمل مشخص گردد [۱۰].

---

[1] Machine Learning
[2] Supervised
[3] Hierarchical
[4] Classification
[5] High-Level
[6] Low-Level
[7] Edges
[8] Representation
[9] Spatial
[10] Short-Term
[11] Long-Term
[12] Recurrent Neural Network (RNN)
[13] Restricted Boltzmann Machines
[14] Direct Classification
[15] K-Nearest Neighbor
[16] Support Vector Machines
[17] Temporal
[18] Markov Models
[19] Random Fields
[20] Wang
[21] Suter
[22] Locality Preserving Projections
[23] Jhuang

باترا[1] و همکاران، یک دیکشنری از عمل‌های انسانی را در اختیار دارند. تلاش می‌شود تا عمل‌های مربوط به هر ویدئو بر اساس عمل‌های موجود در این دیکشنری توصیف گردد. این توصیف در اختیار شبکه‌های عصبی قرار می‌گیرد تا عمل کلاسه‌بندی انجام شود [۱۱].

فتحی و موری[2] از ترکیب ویژگی‌های سطح پایین برای ایجاد توصیف مناسب استفاده کرده‌اند. این توصیف برای هر ویدئو ایجاد شده و در اختیار انواع مختلف Adaboost قرار می‌گیرد [۱۲]. نتایج حاصل از دسته‌بندی با هر کدام از انواع روش‌های Adaboost ارایه شده و مقایسه‌ی مناسبی بین آن‌ها نیز انجام گرفته است.

با توجه به پیچیدگی‌هایی که در ویدئوهای مربوط به شناسایی عمل وجود دارد، نادیده گرفتن زمان باعث از دست رفتن کارایی و دقت می‌شود [۱۳]. به عبارت دقیق‌تر در عمل‌هایی که در عین سادگی از نظر شکل اجرا متفاوت هستند، و همچنین در مورد عمل‌های پیچیده نادیده گرفتن ویژگی زمان می‌تواند به کاهش دقت منجر شود. همچنین، طبق تحقیقات انجام شده، برای شناسایی عمل باید ویدئوی مربوطه در دو طول زمانی کوتاه و بلند تحلیل شود. این دو تحلیل به یادگیری جزئیات و کلیات عمل انجام شده در ویدئو توسط روش کمک می‌کند [۱۴]. با توجه به مطالب مطرح شده، حرکت به سمت استفاده از روش‌های مبتنی بر زمان لازم خواهد بود.لازم به ذکر است تمامی روش‌های ارایه شده در حوزه‌ی زمان از این دو نوع توصیف زمانی بهره نمی‌برند. این موضوع به پایگاه داده و ویدئوهای مورد بررسی بستگی دارد. در صورتی که عمل‌های مربوط به ویدئوها از نظر شکل اجرا (مکان) یا از نظر سرعت اجرا (طول زمانی کوتاه و بلند) بسیار متفاوت باشند می‌توان تمامی توصیفات زمانی کوتاه و بلند را استفاده نکرد.

ژیا[3] و همکاران از توصیف HOJ3D استفاده کرده‌اند که از خانواده‌ی هیستوگرام‌های شیب‌گرا[4] است. این توصیف توانسته است نقاط کلیدی بدن را توصیف کرده و یک بیان فشرده از ژست هر فرد ارایه کند. توصیف HOJ3D از تغییرات نقاط کلیدی در هر فریم و بازنمایش آن با استفاده از LDA بهره می‌برد. این توصیف در اختیار مدل مخفی مارکوف قرار داده می‌شود [۱۵]. لین[5] و همکاران از توصیف هرم[6] زمانی برای بیان قسمت‌های مختلف یک عمل استفاده کرده‌اند. آن‌ها یک فیلد تصادفی شرطی[7] چندلایه چندلایه ارایه کرده‌اند تا لایه‌های مختلف این هرم را یاد بگیرد [۱۶].

با توجه به افزایش روزافزون ویدئوهای برچسب نخورده و تنوع در رفتارها و ویدئوها، استفاده از شبکه‌های عمیق و روش‌های مبتنی بر یادگیری عمیق افزایش یافته است. تمایل به سمت مدل‌های عمیق نشان از قدرت این نوع از روش‌ها در مدل کردن زمان و مکان ویدئوها دارند. چن[8] و همکاران روشی با عنوان شبکه‌ی باور عمیق مکان-زمان[9] ارایه داده‌اند. آنها توصیف مکان و و زمان را با یکدیگر ترکیب کرده‌اند تا شبکه‌ی باور عمیق بتواند ارتباطات بلندمدت را در ویدئو استخراج کند. این روش یک مدل بدون نظارت است که دقت کمی در مقابل روش‌های بانظارت دارد [۱۷].

ژانگ[10] و همکاران روشی بلادرنگ[11] برای شناسایی عمل‌های انسان ارایه کرده‌اند که بر پایه‌ی شبکه‌های باور عمیق است. البته، آن‌ها از حالت استاندارد این شبکه‌ها استفاده نکرده و تغییراتی در روند یادگیری شبکه ایجاد کرده‌اند. بوسیله این تغییرات؛ شبکه توانسته است ابتدا، وسط و پایان انجام یک عمل را شناسایی کند. این شبکه در هر زمان تعداد محدودی از فریم‌ها را پردازش می‌کند. ویژگی لازم برای شبکه‌ی ارایه شده، بازنمایش اسکلت[12] عمل انجام شده در ویدئو است [۶].

در تلاشی دیگر، عبدلاوی[13] و دوییک[14] بستری را برای تبدیل ویدئو به تعداد زیادی بردار باینری فراهم کرده‌اند. این بردارهای باینری به عنوان توصیفی از ویدئو استفاده می‌گردد و برای یادگیری در اختیار شبکه‌ی باور عمیق استاندارد قرار می‌گیرد. در این مقاله از حالت یک بعدی شبکه‌ی باور عمیق استفاده شده است [۱۸]. نیک فرجام و ابراهیم پور توانسته‌اند مدلی بر پایه‌ی شبکه‌های باور عمیق ارایه دهند که دارای چند ورودی است. این چند ورودی به تنظیم دقیق‌تر ویژگی‌های استخراج شده توسط شبکه انجامیده و دقت را افزایش داده است. ولی، ورودی شبکه کماکان توصیف‌های دستی می‌باشد [۲].

همان طور که مشخص است، در تلاش سایر محققان، که از شبکه‌های باور عمیق به عنوان کلاسه‌بند برای مساله‌ی شناسایی عمل استفاده شده است؛ ابتدا، انواع توصیف‌های لازم استخراج شده و سپس به عنوان ورودی در اختیار شبکه قرار می‌گیرد.

در مقاله‌ای متفاوت، حسینعلی[15] و ونگ[16] از شبکه‌ی باورعمیق استاندارد به عنوان استخراج کننده‌ی ویژگی برای توصیف ویدئوها استفاده کرده‌اند. این توصیف‌ها که به وسیله‌ی شبکه ایجاد شده‌اند به عنوان ورودی در اختیار ماشین برداری پشتیبان قرار می‌گیرند و سپس مرحله‌ی دسته‌بندی انجام می‌گیرد

[1] Batra
[2] Mori
[3] Xia
[4] Histogram Of Oriented Gradient
[5] Lin
[6] Pyramid
[7] Conditional Random Field

[8] Chen
[9] Spatio-Temporal
[10] Zhang
[11] Real-Time
[12] Skeletonize
[13] Abdellaoui
[14] Douik
[15] Hussain Ali
[16] Wang

[۱۹]. سایر ساختارهای عمیق نیز برای شناسایی عمل‌های انسان در ویدئو استفاده شده‌اند که جزئیات آن را می‌توان در مقالات مرتبط یافت [۲۰] تا [۲۳]؛ ولی بررسی آن‌ها، در حوزه‌ی کاری این مقاله قرار نخواهد گرفت. هیچ کدام از نسخه‌ها و گونه‌های شبکه‌های باور عمیق (شبکه‌های باور عمیق استاندارد، شبکه‌های باور عمیق چند وضوحه[1]، شبکه‌های باور عمیق گوسی[2]، شبکه‌های شبکه‌های باور عمیق مکان-زمان یا شبکه‌های باور عمیق کانولوشنی[3]) نتوانسته‌اند نیاز این شبکه به دریافت ویدئو را به عنوان داده‌ی خام برطرف کنند.

۲- روش پیشنهادی

در این بخش به بررسی روش پیشنهادی مقاله خواهیم پرداخت. هر ویدئو شامل فریم‌هایی از انجام یک عمل است که به عنوان ورودی شبکه‌ی پیشنهادی استفاده خواهند شد. فریم‌های این ویدئوها به عنوان داده‌ی خام در اختیار شبکه‌ی پیشنهادی قرار می‌گیرند. به طور کلی، چگونگی پیاده سازی دوبعدی ماشین‌های بولتزمن محدود، روند یادگیری شبکه‌ی باور عمیق دوبعدی، استخراج فریم‌های تاثیرگذار در روند یادگیری و پیاده‌سازی بازگشتی برای فهم مفاهیم زمانی از ویژگی های اصلی روش پیشنهادی این مقاله است.

پس اولین نکته‌ای که مطرح می‌شود چگونگی انتخاب فریم‌های مذکور است. در ادامه حالت‌های مختلفی که برای انتخاب این فریم‌ها وجود دارد آمده است. هر کدام از این حالت‌ها در کارهای مرتبط نیز استفاده شده یا پیاده‌سازی گردیده‌اند:

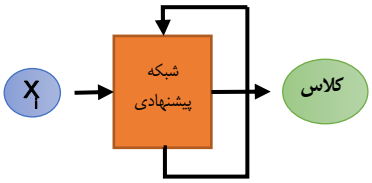
۱) تمام فریم‌های ویدئو در اختیار شبکه قرار می‌گیرند(به ترتیب) [۱۷].

۲) فریم‌ها با یک فاصله‌ی مشخص زمانی انتخاب و در اختیارشبکه قرار می‌گیرند(به عنوان مثال، فریم‌های مربوط به شماره‌های ۱، ۱۱، ۲۱، ۳۱ و ...) [۲۴].

۳) فریم‌های مهم با استفاده از یک الگوریتم مشخص انتخاب شده و به عنوان فریم‌های کلیدی[4] معرفی می‌گردند. سپس این فریم‌ها به عنوان ورودی شبکه استفاده می‌شوند(به ترتیب) [۲۵].

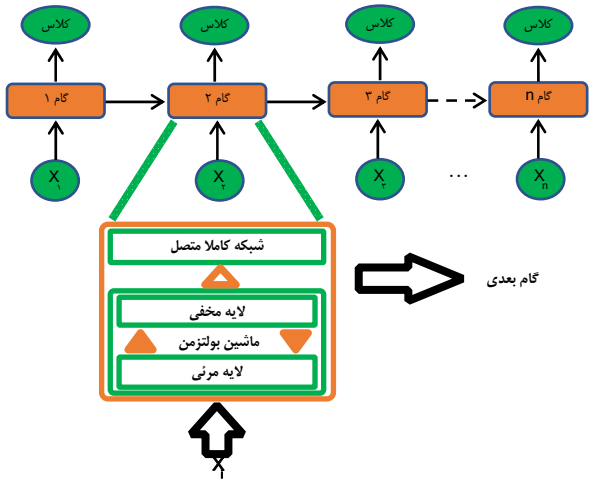
قابلیت ماشین‌های بولتزمن محدود در پیاده‌سازی‌های یک بعدی به صورت کامل آزمون گردیده است [۷]. با توجه به این که ورودی شبکه‌ی پیشنهادی فریم‌های دوبعدی می‌باشند و با تغییر حالت این فریم‌ها به صورت یک بعدی، بسیاری از اطلاعات همسایگی[5] در پیکسل‌های تصویر از میان می‌رود؛ پیاده‌سازی یک بعدی صورت نمی‌گیرد. در حالت یک بعدی از دست رفتن اطلاعات همسایگی پیکسل‌ها در فریم‌های ویدئو در تحقیقات سایر محققین به اثبات رسیده است و بر این موضوع صحه گذارده شده که در صورت توان پردازشی دوبعدی، تبدیل به حالت یک بعدی مناسب نمی‌باشد [۵]. این نکته نیز باید مد نظر قرار گیرد که در استفاده از شبکه‌ی پیشنهادی، قدرت روش‌های یادگیری عمیق بر مبنای پردازش و یادگیری داده‌های خام می‌باشد. پس استخراج ویژگی و ایجاد بردارهای یک بعدی از ویژگی‌های استخراج شده، نادیده گرفتن فلسفه‌ی وجودی این دسته از روش‌ها است. بنابراین، از بردارهای یک بعدی ویژگی نیز استفاده نخواهد شد و پردازش شبکه‌ی پیشنهادی در حالت دوبعدی پیاده‌سازی و ارزیابی می‌شود و آموزش و آزمون شبکه، تنها بر اساس داده‌ی خام(فریم‌های ویدئو) صورت می‌گیرد.

شکل (۱) نمایانگر ساختار کلی روش پیشنهادی این مقاله می‌باشد که ترکیبی از ماشین‌های بولتزمن محدود و شبکه عصبی بازگشتی است.



شکل (۱): ساختار کلی روش پیشنهادی

شکل (۲) عملکرد شبکه پیشنهادی را در طول زمان نشان می‌دهد. همان طور که از شکل (۲) مشخص است، فریم‌های مختلف ویدئو در طول زمان در اختیار شبکه قرار می‌گیرد.



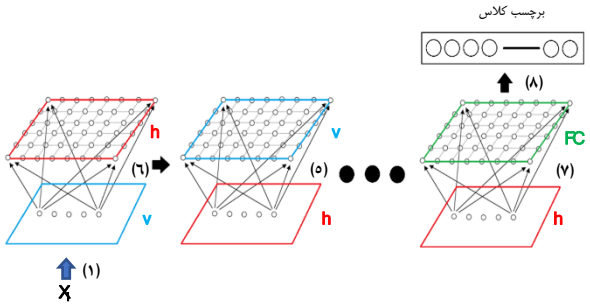
شکل (۲): ساختار روش پیشنهادی در طول زمان

ماشین بولتزمن محدود به کار رفته بر مبنای ورودی خود در لایه‌ی مرئی به حالت تعادل می‌رسد. این حالت تعادل به یادگیری ماشین بولتزمن محدود منجر می‌گردد. خروجی لایه‌ی مخفی مربوط به ماشین بولتزمن محدود در اختیار یک شبکه تماما-متصل قرار گرفته و این شبکه به صورت بانظارت آموزش داده می‌شود.

از آن جا که وزن‌ها اشتراکی هستند؛ با ورود فریم بعدی، ماشین بولتزمن محدود یادگیری خود را ارتقا می‌دهد و یادگیری شبکه تماما-متصل[6] نیز کامل‌تر می‌گردد. این روند آن قدر ادامه می‌یابد که شبکه‌ی پیشنهادی توان دسته‌بندی ویدئوها را داشته

[1] Multi-Resolution
[2] Gaussian
[3] Convolutional
[4] Key-frame.
[5] Neighborhood.
[6] Fully-connected.

باشد. از آن جا که شبکه‌ی مذکور دوبعدی می‌باشد، شماتیک کلی روش پیشنهادی بر مبنای ورودی‌های دوبعدی به همراه شماره روابط استفاده شده در شکل (۳) آمده است. در این شکل لایه مرئی و مخفی ماشین بواتزمن محدود به ترتیب با v و h و لایه کاملا متصل با FC نشان داده شده است.



شکل (۳): شماتیک دوبعدی لایه‌های روش پیشنهادی به همراه شماره روابط استفاده شده

پس از ورود فریم iام از ویدئو به شبکه، ماشین بولتزمن محدود در رفت و برگشت‌های فراوان به حالت تعادل می‌رسد. از خروجی به وجود آمده در لایه‌ی مخفی(پس از تعادل ماشین) به عنوان ورودی یک شبکه‌ی کاملا متصل استفاده می‌شود.

این روند برای تعداد زیادی فریم تکرار می‌شود و شبکه آموزش می‌بینند. این شبکه در لایه‌ی کاملا متصل خود دارای یک بردار برچسب به تعداد کلاس‌های پایگاه داده‌ی استفاده شده خواهد بود. هنگام آموزش، لایه برچسب مقدار مناسب دارد و در هنگام آزمون، مقدار مناسب در لایه‌ی برچسب تولید می‌شود.

۱-۲- انتخاب فریم‌های مورد نظر از ویدئو

مجموعه داده‌ی متشکل از V ویدئو را در نظر بگیرد که $V = \{V_1, \dots, V_M\}$ می‌باشد. هر کدام از V_iها نشان از یک ویدئو در پایگاه داده دارد ($i \in \{1 \dots M\}$). ویدئو V_i دارای n_i فریم رنگی می‌باشد ($i \in \{1 \dots M\}$). بنابراین، M تعداد کل ویدئوها و n_i تعداد فریم‌های ویدئوی iام است. فریم‌های هر ویدئو را با X نشان می‌دهیم. پس، X_i نشان از iامین فریم ویدئو دارد.

اولین گام روش پیشنهادی پس از تبدیل فریم‌ها از مقیاس رنگی[1] به مقیاس خاکستری[2]، یافتن فریم‌های لازم جهت تغذیه و یادگیری شبکه است. این موضوع از سه جنبه قابل پیاده‌سازی خواهد بود. در روش پیشنهادی هر کدام از این جنبه‌ها، پیاده‌سازی شده و جزئیات آن به همراه نتایج به بحث گذاشته می‌شود:

۱) همه فریم‌های ویدئو بدون تغییر و به ترتیب در اختیار شبکه گذاشته می‌شوند [۱۷].

۲) فریم‌ها در بازه‌های زمانی معین(مثلا بازه‌های زمانی α) انتخاب و در اختیار شبکه قرار می‌گیرند [۲۴]. نکته‌ی حائز اهمیت در این روش انتخاب مناسب عدد α می‌باشد. با توجه به مطالب مطرح شده در این بخش، $\frac{n_i}{\alpha}$ فریم از n_i فریم ویدئوی V_i استخراج شده و در اختیار شبکه قرار می‌گیرد. شبکه برای آموزش و آزمون از این فریم‌ها استفاده خواهد کرد.

۳) با استفاده از یک الگوریتم معین فریم‌های مهم به عنوان فریم‌های کلیدی انتخاب می‌گردند. این فریم‌ها به عنوان ورودی شبکه استفاده می‌شوند[۲۵]. مبنای روش استفاده شده برای یافتن فریم‌های کلیدی روش تصنیم[3] و بائک[4] می‌باشد [۲۶].

نمونه‌برداری فریم روشی است که در آن تعداد ثابتی فریم از میان دنباله‌ای از فریم‌ها بر اساس یک معیار انتخاب می‌شوند. در این روش، از متریک رتبه[5] به عنوان معیار نمونه‌گیری[6] استفاده می‌شود. دلیل نمونه‌گیری تعداد ثابت فریم این است که تعداد زیادی فریم اضافی در یک دنباله عمل وجود دارد. معیار استفاده شده با نام متریک رتبه (برای سازماندهی مجدد فریم‌ها بر اساس مقادیر رتبه) شناخته می‌شود که برای یافتن رتبه هر فریم استفاده می‌شود. پس از یافتن رتبه هر فریم، مجموعه‌های k تایی از فریم‌ها را به عنوان مجموعه معنی‌دار (۲۴، ۲۰، ۱۶ = k) نمونه‌برداری می‌کنیم. برای نمونه‌گیری فریم از معیار SSIM[7] استفاده می‌شود.

متریک SSIM ویژگی‌های ساختاری بین دو فریم مجاور را برای انجام مقایسه‌ها در نظر می‌گیرد. فرض کنید I_t و I_{t+1} دو فریم مجاور در زمان‌های t و t + 1 در یک دنباله عمل باشند. سپس، مقدار SSIM بین I_t و I_{t+1} را که با ψ نشان داده می‌شود، می‌توان به شکل رابطه (۱) تعریف کرد [۲۶]:

$$\psi(I_t, I_{t+1}) = \frac{(2 \times \overline{I_t} \times \overline{I_{t+1}} + C_1)(2 \times var(I_t, I_{t+1}) + C_2)}{((\overline{I_t})^2 + (\overline{I_{t+1}})^2 + C_1)(var(I_t)^2 + var(I_{t+1})^2 + C_2)} \quad (۱)$$

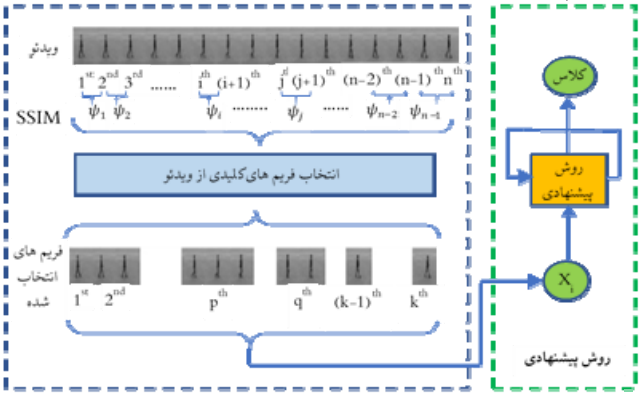
که در آن $\overline{I_t}$ میانگین فریم t و $var(I_t, I_{t+1})$ کواریانس بین فریم‌های I_t و I_{t+1} را نشان می‌دهد. $var(I_t)$ و $var(I_{t+1})$ واریانس‌های درون فریم هستند. C_1 و C_2 ثابت هستند.

شکل (۴) چگونگی انجام روش نمونه‌برداری k فریم رتبه‌بندی شده را نشان می‌دهد. همان‌طور که در شکل (۴) نشان داده شده است؛ ابتدا مقدار SSIM بین فریم فعلی و فریم بعدی برای همه فریم‌های ویدئو محاسبه می‌شود. از آنجا که مقادیر بزرگ‌تر SSIM شباهت بیش‌تری را نشان می‌دهد؛ مقادیر SSIM به دست آمده را به ترتیب صعودی مرتب کرده و k فریم متناظر با k مقدار ابتدای لیست مرتب شده به عنوان فریم‌های کلیدی انتخاب می‌شوند. می‌توان فرض کرد که k فریم نمونه‌برداری شده نیز اطلاعات مکانی و زمانی مشابه دنباله‌های اصلی را با کمی تغییر دارند. به همین ترتیب، آزمایش‌ها و تحلیل‌ها را با مقادیر مختلف k (۲۴، ۲۰،

---

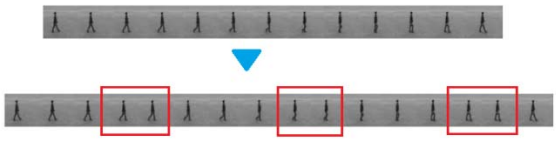
1 RGB.
2 Gray.
3 Tasnim.
4 Baek.
5 Rank Metric.
6 Sampling Criterion.
7 Structural Similarity Index measure

۱۶ = k) برای نشان دادن کلیت روش رتبه‌بندی پیشنهادی انجام می‌دهیم.



شکل (۴): چگونگی انتخاب k فریم کلیدی برای روش پیشنهادی

برای ویدئوهایی با طول L که تعداد فریم‌هایشان از k کمتر است، باید n=k-L فریم تکرار شود تا طول ویدئو به k فریم برسد. شکل (۵) نشان از روند کار روی ویدئوهایی با طول کمتر از k دارد. در شکل (۵)، k=۱۶ در نظر گرفته شده و ویدئو دارای ۱۳ فریم است. بنابراین، فریم‌های ۴، ۸ و ۱۲ تکرار شده‌اند. فریم‌های ۴، ۸ و ۱۲ فریم‌هایی هستند که کمترین مشابهت را با فریم‌های بعدی خود داشته اند. از آن جا که برای رساندن ۱۳ فریم به ۱۶ فریم در ویدئو، نیاز به ۳ فریم دیگر داریم؛ پس از ۳ فریمی استفاده می‌شود که کم‌ترین مقدار SSIM را داشته باشند. این موضوع کمک می‌کند تا بازه ی اطلاعاتی روش پیشنهادی گسترده‌تر شود. مورد دوم آن که پس از انتخاب k فریمی که کم‌ترین مقدار SSIM را در کل فریم‌ها دارند؛ ترتیب فریم‌ها در ورودی به شبکه حفظ می‌شود. با توجه به مطالب مطرح شده در این بخش، k فریم از n_i فریم ویدئوی V_i استخراج شده و در اختیار شبکه قرار می‌گیرد. شبکه برای آموزش و آزمون از این فریم‌ها استفاده خواهد کرد.



شکل (۵): تکرار فریم‌های ۴، ۸ و ۱۲ برای تطبیق با حداقل تعداد فریم کلیدی

۲-۲- ماشین بولتزمن محدود دوبعدی پیاده‌سازی شده

از آنجایی که ماشین‌های بولتزمن محدود (RBM) و شبکه‌های باور عمیق (DBN) در استخراج ویژگی‌های غیرنظارت شده، بسیار موثر هستند و علاوه بر این، در کاربردهایی که نیاز به تفسیرپذیری دارند، می‌توانند بسیار مفید باشند زیرا آن‌ها قادر به یادگیری و نمایش توزیع‌های پیچیده‌ای از داده‌ها هستند، ما بر آن شدیم که از این نوع شبکه برای روش پیشنهادی خود استفاده کنیم. لازم به ذکراست که باتجهیز این شبکه به پیاده‌سازی بازگشتی امکان یادگیری دنباله‌ها نیز فراهم شده است.

ماشین بولتزمن محدود متعلق به دسته روش‌های فیلدهای تصادفی مارکوف[1] است. در مقایسه با ماشین‌های بولتزمن محدود، ماشین‌های بولتزمن استاندارد بیش‌تر به اتصال بین نرون‌های مرئی-مرئی و اتصالات مخفی-مخفی (در درون لایه‌ها) گرایش دارند. لیکن، ماشین بولتزمن محدود یک ساختار دو بخشی است که در آن W ماتریس وزن بین لایه مرئی(v) و لایه مخفی(h) می‌باشد. در ادامه چگونگی پیاده‌سازی دوبعدی آن شرح داده می‌شود.

از مدل نشان داده شده در شکل (۳)، متوجه می‌شویم که ماشین‌های بولتزمن محدود یک مدل گرافیکی مستقیم متشکل از متغیرهای مرئی و متغیرهای پنهان است. علاوه بر این، هیچ ارتباطی بین متغیرهای همان لایه (به عنوان مثال در لایه مرئی یا پنهان) وجود ندارد. تابع انرژی E(v, h) ماشین بولتزمن محدود دوبعدی پیاده‌سازی شده در رابطه (۲) بیان شده است [۵]:

$$E(v,h) = -\sum_i\sum_j b_{ij}v_{ij} - \sum_p\sum_q c_{pq}h_{pq} - \sum_i\sum_j\sum_p\sum_q W_{ijpq}v_{ij}\,h_{pq} \qquad (۲)$$

که برای لایه‌ی مرئی از i و j و لایه ی مخفی از p و q برای اندیس نرون‌های سطر و ستون در هر کدام از لایه‌ها استفاده شده است. تعداد سطر و ستون‌های این دو لایه در پیاده‌سازی روش پیشنهادی یکسان است؛ ولی در نوشتن فرمول‌ها به صورت کلی این تفاوت لحاظ شده است.

W_{ijpq} وزن بین نرون موجود در سطر i و ستون j از لایه‌ی مرئی(v_{ij}) و نرون موجود در سطر p و ستون q از لایه‌ی مخفی(h_{pq}) می‌باشد. پارامترهای b و c به ترتیب بایاس‌های لایه‌های مرئی و پنهان را نشان می‌دهند. مشخص است با توجه به گسترده‌تر شدن رابطه نسبت به حالت یک بعدی، پیچیدگی زمانی افزایش یافته و زمان بیش‌تری صرف آموزش می‌گردد. توزیع احتمال مشترک به صورت رابطه (۳) محاسبه می‌شود:

$$P(v,h) = \frac{exp(-E(v,h))}{\sum_i\sum_j\sum_p\sum_q exp(-E(v,h))} \qquad (۳)$$

همچنین می‌توان با استفاده از رابطه (۴) توزیع حاشیه‌ای واحدهای مرئی را با جمع‌بندی بر روی واحدهای پنهان محاسبه کرد:

$$P(v) = \frac{\sum_p\sum_q exp(-E(v,h))}{\sum_i\sum_j\sum_p\sum_q exp(-E(v,h))} \qquad (۴)$$

با توجه به ساختار ماشین‌های بولتزمن محدود، واحدهای موجود در لایه‌های مرئی و مخفی به صورت مشروط مستقل هستند. نسخه احتمالی فعال شدن نرون‌های لایه‌های مرئی و مخفی به ترتیب در روابط (۵) و (۶) آمده است:

[1] Markov Random Fields.

$$P(v_{ij}=1|\mathbf{h})=\frac{1}{1+e^{-(b_{ij}+\sum_p\sum_q W_{ijpq}h_{pq})}} \quad (۵)$$

$$P(h_{pq}=1|\mathbf{v})=\frac{1}{1+e^{-(c_{pq}+\sum_i\sum_j W_{ijpq}v_{ij})}} \quad (۶)$$

مزایای استفاده از ماشین‌های بولتزمن محدود عبارتند از [۳۶]:

۱) استفاده از خروجی لایه‌ی مخفی این ماشین‌ها به عنوان ورودی شبکه یا لایه‌ای دیگر قابل پیاده سازی است.

۲) سرعت یادگیری بالایی دارد.

۳) توان یادگیری توزیع‌های مختلف داده‌ای را دارد.

این مزایا برای فهم مفهوم مکان در فریم‌های ویدیو مفید هستند. لیکن، امکان فهم مفاهیم زمانی وجود نداشته است که با پیاده‌سازی ساختار بازگشتی در این مقاله امکان پذیر خواهد بود.

۳-۲- چگونگی ترکیب ماشین بولتزمن محدود دوبعدی و پیاده‌سازی بازگشتی

شبکه‌های عصبی بازگشتی قادر به یادگیری ویژگی‌ها و وابستگی‌های طولانی‌مدت از داده‌های متوالی و سری زمانی هستند. شبکه‌های عصبی بازگشتی دارای پشته‌ای از واحدهای غیرخطی هستند که حداقل یک اتصال بین واحدها، یک چرخه جهت‌دار را تشکیل می‌دهد. یک شبکه‌ی عصبی بازگشتی به خوبی آموزش دیده می‌تواند هر سیستم پویا را مدل کند [۲۷]. با این حال، آموزش شبکه‌های عصبی بازگشتی عمدتا با مسائل مربوط به یادگیری وابستگی‌های طولانی‌مدت مواجه است.

شبکه‌های عصبی بازگشتی دسته‌ای از مدل‌های یادگیری ماشینی بانظارت هستند که از نرون‌های مصنوعی با یک یا چند حلقه بازخورد[1] ساخته شده‌اند [۲۸]. حلقه‌های بازخورد چرخه‌های مکرر در طول زمان یا توالی هستند. آموزش یک شبکه‌ی عصبی بازگشتی به روش بانظارت به مجموعه داده آموزشی از جفت‌های ورودی-هدف[2] نیاز دارد. هدف این است که با بهینه سازی وزن شبکه، تفاوت بین جفت خروجی و هدف (یعنی مقدار خطا) به حداقل برسد.

در روش پیشنهادی لایه ورودی شبکه‌ی کاملا متصل همان لایه‌ی مخفی ماشین بولتزمن محدودی است که به حالت تعادل رسیده و توانسته یک فریم از ویدئو را یاد بگیرد. بنابراین، ورودی‌های شبکه بازگشتی، دنباله‌ای از ماتریس‌ها در طول زمان t (همان فریم‌های تصویر) هستند. ماتریس‌های تولید شده در لایه‌ی مخفی ماشین بولتزمن محدود دوبعدی که به حالت تعادل رسیده بر مبنای فریم در زمان t می‌باشند. به عبارت دقیق‌تر، همان طور که در شکل (۱) آمده است به جز لایه‌ی برچسب و فریم X_i ورودی، ماشین بولتزمن محدود و لایه‌ی کاملا متصل بعد از آن، جزء شبکه‌ی بازگشتی هستند. پس در هر گام از زمان t، ماتریس وزن حاصل از زمان‌ها و گام‌های قبلی برای به روزرسانی در گام t+1 استفاده می‌گردد.

اتصالات بین لایه‌ی مرئی و لایه‌ی مخفی در ماشین بولتزمن محدود با ماتریس وزن W^{VH} تعریف می‌شوند. از آن جا که نرون آمده در سطر p و ستون q از لایه‌ی مخفی ماشین بولتزمن محدود دوبعدی را با h_{pq} و نرون آمده در سطر r و ستون s از لایه‌ی کاملا متصل شبکه عصبی بازگشتی را با O_{rs} نمایش می‌دهیم؛ W^{VH}_{pqrs} بیان‌گر اتصال بین دو نرون ذکر شده از دولایه مورد نظر خواهد بود. لازم به ذکر است؛ طبق نتایج بررسی‌های قبلی، مقداردهی اولیه نرون‌های مخفی با استفاده از یک مقدار کوچک غیرصفر می‌تواند عملکرد و پایداری کلی شبکه عصبی بازگشتی را بهبود بخشد [۲۹].

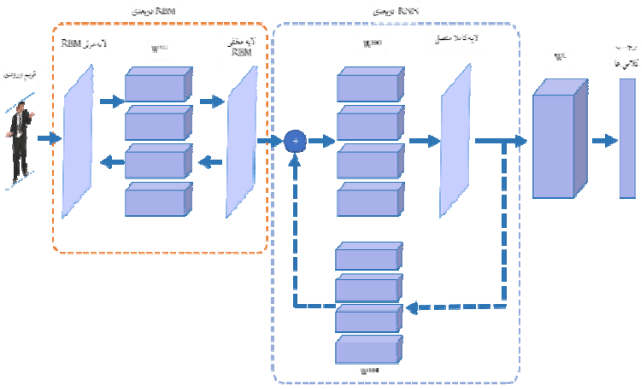
با همین ادبیات، $W^{HH'}_{rsv's'}$ بیان‌گر اتصال بین دو نرون لایه‌ی مخفی شبکه عصبی بازگشتی در دو زمان مختلف است. بنابراین، کارکرد شبکه عصبی بازگشتی مطرح شده به صورت رابطه (۷) تعریف می‌شود:

$$H_t=f_H(W^{VH}X_t+W^{HH'}H_{t-1}+b_H) \quad (۷)$$

f_H تابع فعال‌سازی لایه مخفی H است و b_H بردار بایاس نرون‌های مخفی است. نرون‌های مخفی با اتصالات وزنی W^L به لایه خروجی متصل می‌شوند. W^L_{rsu} بیان‌گر اتصال بین دو نرون r و s از لایه ی کاملا متصل شبکه عصبی بازگشتی و نرون uام از لایه برچسب خروجی است. O_t، مقدار لایه خروجی در زمان t از رابطه (۸) حاصل می‌شود.

$$O_t=f_O(W^LH_t+b_O) \quad (۸)$$

که در آن f_O تابع فعال‌سازی و b_O بردار بایاس در لایه خروجی است. در شکل (۶) بخش‌های مختلف شبکه پیشنهادی و ارتباط بین این بخش‌ها از طریق ماتریس‌های وزن معرفی شده، به تصویر کشیده شده است.



شکل (۶): ساختار کلی روش پیشنهادی با تاکید بر ارتباط بین بخش های مختلف

از آنجایی که جفت‌های ورودی-هدف در طول زمان متوالی هستند؛ مراحل بالا در طول زمان تکرار می‌شوند. بنابراین، پارامتر t در روابط (۷) و (۸) حاکی از حرکت در طول زمان هستند.

[1] Feedback.
[2] Input-target.

در هر مرحله زمانی، وضعیت‌های[1] لایه‌ی مخفی یک پیش‌بینی را در لایه خروجی بر اساس نوع ورودی ارائه می‌کنند. وضعیت لایه مخفی یک شبکه عصبی بازگشتی مجموعه‌ای از مقادیر است که جدا از تأثیر عوامل خارجی، تمام اطلاعات ضروری منحصر به فرد را در مورد وضعیت‌های گذشته شبکه در چندین مرحله خلاصه می‌کند. این اطلاعات یکپارچه می‌تواند رفتار آینده شبکه را تعریف کند و پیش-بینی‌های دقیقی را در لایه خروجی انجام دهد [۳۰].

شبکه عصبی بازگشتی ارایه شده از یک تابع فعال‌سازی غیرخطی ساده در هر واحد استفاده می‌کند. با این حال، چنین ساختار ساده‌ای می‌تواند تغییرات را مدل‌سازی کند؛ اگر در مراحل زمانی مختلف (tهای مختلف) به خوبی آموزش داده شود.

واضح است که توابع غیرخطی، قدرتمندتر از توابع خطی هستند. زیرا می‌توانند مرزهای غیرخطی ترسیم کنند. غیرخطی بودن لایه‌های مخفی این شبکه عصبی بازگشتی دلیل یادگیری روابط بهتر ورودی-هدف است. به همین سبب از تابع Relu[2] استفاده شده است که نسبت به سایر توابع فعال‌سازی در پژوهش‌های اخیر توجه بیشتری را به خود جلب کرده است. Relu تابعی است که برای مقادیر ورودی مثبت باز است و به صورت رابطه (۹) تعریف می‌شود[۳۱]:

$$y(x) = max\,(0, x) \quad (۹)$$

تابع فعال‌سازی Relu منجر به شیب‌های پراکنده‌تر می‌شود و همگرایی را بسیار تسریع می‌کند[۳۱]. Relu از نظر محاسباتی ارزان است، زیرا می‌توان آن را با آستانه‌گذاری مقدار فعال‌سازی در صفر پیاده‌سازی کرد.

۳- ارزیابی

برای آن که شرایط مقایسه‌ای منصفانه داشته باشیم باید شرایط آزمون در روش پیشنهادی و روش‌های رقیب یکسان باشد. به این صورت می‌توان ادعا کرد روش‌های مختلف با یکدیگر در بستر یکسانی مقایسه شده و نتایج به دست آمده قابل اعتماد هستند. این یکسان بودن هم در داده‌ها و هم در روند انجام آزمایشات از نظر آموزش[3] و آزمون[4] باید لحاظ گردد.

۳-۱- پایگاه‌های داده

در استفاده از پایگاه داده‌های استاندارد، ابتدا تمامی فریم‌های تمامی ویدئوها از مقیاس رنگی[5] به مقیاس خاکستری[6] تبدیل شده و در اختیار روش پیشنهادی قرار می‌گیرند. این تبدیل کمک می‌کند تمامی فریم‌ها از حالت سه بعدی به حالت دوبعدی تغییر مقیاس دهند. شرح پایگاه داده‌های استفاده شده در ادامه آمده است.

۳-۱-۱- پایگاه داده‌ی KTH

پایگاه داده ویدئویی KTH شامل شش نوع عمل انسان (پیاده روی[7] (پ‌ر)، دویدن[8] (د)، تندرفتن[9] (ت‌ر)، مشت زدن[10] (م‌ز)، دست تکان دادن[11] (دت‌د) و کف زدن[12] (ک‌ز)) می‌باشد. این عمل‌ها چندین بار توسط ۲۵ نفر در چهار سناریو مختلف انجام شده است: ۱) فضای باز، ۲) فضای باز با تغییر مقیاس، ۳) خارج از منزل با لباس‌های مختلف و۴) در فضای بسته.

در حال حاضر پایگاه داده شامل ۲۳۹۱ ویدئو است. همه ویدئوها با یک دوربین ثابت با نرخ ۲۵ فریم در ثانیه بر روی پس‌زمینه‌های همگن گرفته شده‌اند. هر ویدئو با وضوح ۱۲۰ در ۱۶۰ پیکسل نمونه‌برداری شده و به طور متوسط طول آن چهار ثانیه است. همه دنباله‌ها با استفاده از فرمت فایل AVI ذخیره شده و به صورت آنلاین در دسترس هستند [۳۲].

۳-۱-۲- پایگاه داده‌ی UCF

مجموعه داده‌های ورزشی UCF شامل ۱۰ اقدام ورزشی است: شیرجه[13] (ش)، گلف[14] (گ)، بلند کردن[15] (ب‌ک)، لگد زدن[16] (ل-ز)، اسب سواری[17] (اسب‌س)، دویدن(د)، اسکیت سواری[18] (اسکس)، تخته سواری[19] (ت‌س)، تاب خوردن[20] (ت-و پیاده روی(پ‌ر). این مجموعه شامل۱۵۰ دنباله ویدئویی با وضوح ۷۲۰ در ۴۸۰ است. همه ویدئوها از پخش کانال‌های تلویزیونی مختلف جمع‌آوری و در محیط‌های نامحدود با چندین تنوع درون کلاسی از جمله تغییرات روشنایی، پس‌زمینه پیچیده، تاری حرکت، انسداد و صحنه‌های متنوع ضبط شده‌اند [۳۳].

۳-۱-۳- پایگاه داده‌ی HMDB51

پایگاه داده‌ی HMDB51 دارای اعمال پیچیده‌ای است که از منابع مختلفی جمع‌آوری گردیده‌اند. ویدئوهای این پایگاه داده از دو بخش تشکیل شده‌اند: بخش اول از فیلم‌های ساخته شده در دنیا و بخش دیگر از پایگاه‌های عمومی ویدئو در دنیا مانند آرشیو پری-لینگر[21]، یوتیوب[22] یا ویدئوهای گوگل. این پایگاه داده دارای ۶۸۴۹ ویدئو می‌باشد که ۵۱ کلاس را در بر می‌گیرد. هر کلاس

---

1. State.
2. Rectified Linear Unit
3. Train.
4. Test.
5. Color-scale.
6. Gray-scale.
7. Walking.
8. Running.
9. Jogging.
10. Boxing.
11. Hand Waving.
12. Hand Clapping.
13. Diving.
14. Golf.
15. Lifting.
16. Kicking.
17. Riding Horse.
18. Skate-boarding.
19. Swinging-bench.
20. Swinging-side.
21. Prelinger Archive.
22. YouTube.

بیش از ۱۰۱ ویدئو را شامل می‌شود [۳].از آن جا که کلاس‌های این پایگاه داده زیاد است، برای رسم ماتریس درهم‌ریختگی از تکنیک ماتریس خاکستری استفاده شده است. چون تعداد کلاس-ها در این مجموعه داده زیاد است از ذکر نام آنها در اینجا صرفنظر می‌کنیم.

۲-۳- ارزیابی روش پیشنهادی

ارزیابی روش پیشنهادی روی پایگاه داده‌های استاندارد شرح داده شده در بخش‌های قبلی مورد بحث قرار می‌گیرد. لازم به ذکر است، برای پایگاه داده‌های KTH و UCF شرایط پیاده سازی از جهت چگونگی انتخاب ویدئوهای آموزش و آزمون به همراه نتایج سایر روش‌های رقیب برگرفته از [۳۴] می‌باشد. همچنین، برای پایگاه داده‌های HMDB51 شرایط گفته شده برگرفته از [۳۵] می-باشد. بنابراین، مقایسه‌ی تمامی روش‌ها به صورت منصفانه صورت گرفته است. همچنین عملکرد الگوریتم پیشنهادی روی چند سطح از وجود نویز در فریم‌های ویدیو با پیاده‌سازی استفاده از فریم‌های کلیدی بررسی می‌شود.

در جداول (۱)، (۲) روش‌های رقیب به همراه الگوریتم استفاده شده توسط آنها به ترتیب بر روی پایگاه داده KTH و UCF معرفی شده است. در جدول (۵) علاوه بر دقت، الگوریتم استفاده شده در روش‌های رقیب بر روی پایگاه داده HMDB51 آمده است.

جدول (۱): روش‌های رقیب به همراه الگوریتم مورد استفاده بر روی پایگاه داده KTH

روش	الگوریتم استفاده شده
Laptev et al.	Cuboids+HOG3D
Nazir et al.	3DHarris+3DSIFT
Niebles et al.	PLSA
Jhuang et al.	HMAX
Taylor et al.	3D GRBM
Le et al.	Hierarchical ISA
Ji et al.	3D CNN
Sun et al.	3D (DL-SFA)
Zhang et al.	Dual-channel deep network
Han et al.	Two-stream ConvNets
Abdelbaky and Aly	ST-VLAD-PCANet

جدول (۲): روش‌های رقیب به همراه الگوریتم مورد استفاده بر روی پایگاه داده UCF

روش	الگوریتم استفاده شده
Laptev et al.	Dense + HOF
Klaser et al.	Dense + HOG3D
Rahmani et al.	Deep R-NKTM
Le et al.	Hierarchical ISA
Zhang et al.	Dual-channel DNN
Sun et al.	3D (DL-SFA)
Yuan et al.	3D Deep model
Wang et al.	LSTM+CNN
Ahmed and Aly	STMEI-PCANet
Abdelbaky and Aly	ST-VLAD

۱-۲-۳- استفاده از همه فریم‌ها به عنوان ورودی شبکه‌ی پیشنهادی

در این بخش ابتدا ویدئوهای مربوط به آموزش انتخاب و تمامی فریم‌های آن‌ها به همراه برچسب عمل انجام گرفته در اختیار شبکه پیشنهادی قرار می‌گیرد تا آموزش صورت گیرد. پس از آن، در مرحله‌ی آزمون می‌توان تمامی فریم‌های ویدئو را برای تولید برچسب مذکور در اختیار شبکه قرار داد.

همان طور که از جداول (۳)، (۴) و (۵) ستون " همه فریم‌ها "مشخص است، روش پیشنهادی دقت قابل قبولی را روی هر سه پایگاه داده‌ی KTH، UCF و HMDB51 به دست آورده است؛ ولی روش‌های قدرتمندتر از روش پیشنهادی نیز وجود داشته و قدرت روش پیشنهادی در میانه‌ی راه بوده است. به نظر می‌رسد دلیل این امر را می‌توان در ماشین‌های بولتزمن دوبعدی پیاده‌سازی شده جستجو کرد. از آن جا که این ماشین‌ها در حالت دوبعدی دارای تعداد بسیاری وزن برای فهم ارتباط بین پیکسل‌ها هستند و این موضوع با پیاده‌سازی بازگشتی و اشتراکی شدن این وزن‌ها وخیم‌تر شده؛ یادگیری مختل شده است. به عبارتی یادگیری بیش از حد[1] اتفاق افتاده است. با آزمایشات بعدی این موضوع بیش‌تر بررسی خواهد گردید.

۲-۲-۳- استفاده از فریم‌های ویدئو با فاصله‌ی زمانی مشخص (α) به عنوان ورودی شبکه‌ی پیشنهادی

در این بخش فریم‌های مختلف ویدئو با فاصله‌ی زمانی مشخص α انتخاب و برای آموزش در اختیار شبکه قرار داده می‌شوند. این اقدام باعث می‌شود تعداد تصاویر استفاده شده به شدت کاهش یابد و تئوری یادگیری بیش از حد شبکه، مورد بررسی دقیق‌تر قرار گیرد. همچنین، در روند آزمون نیز همین اقدام جهت تولید برچسب ویدئو صورت خواهد گرفت.

همان طور که از جدول (۳) مشخص است، میزان دقت بر روی پایگاه داده KTH نسبت به انتخاب همه فریم‌ها افزایش یافته است و این موضوع فرضیه‌ی آموزش بیش از حد ماشین بولتزمن محدود دوبعدی که به صورت بازگشتی پیاده‌سازی شده است را تایید می‌کند. تایید فرضیه آموزش بیش از حد در جداول (۴) و (۵) نیز بر مبنای پایگاه داده ی UCF و HMDB51 آزمون شده است. البته، در جدول (۴) میزان α برابر با ۱۶ و در جدول (۵) برابر با ۸ در نظر گرفته شده است.

چگونگی انتخاب مقدار α در این قسمت چالش اساسی است. مقدارهای ۱۳، ۱۶ و ۸ انتخاب شده در جداول (۳) تا (۵) بر مبنای آزمایش‌های انجام گرفته برای مقادیر مختلف α در بازه‌ی ۰ تا ۲۰ می‌باشد. دقت‌های مختلف به دست آمده برای مقادیر مختلف α با اجرا روی پایگاه داده‌های KTH، UCF و HMDB51 به ترتیب در شکل‌های (۷) تا (۹) آمده است.

[1] Overfitting.

۳-۲-۳- استفاده از فریم‌های کلیدی به عنوان ورودی شبکه‌ی پیشنهادی

دلیل استفاده از فریم‌های کلیدی در واقع حذف فریم‌های اضافی در یک دنباله عمل است. منظور از فریم‌های اضافی فریم‌های مجاورند که در اکثر موارد، حاوی انواع مشابهی از اطلاعات از نظر تغییرات مکانی و زمانی هستند. بنابراین، بسیار مهم است که فریم‌های غیر ضروری که اطلاعات عملی یکسانی دارند کنار گذاشته شوند. گاهی اوقات، تعیین بهترین مجموعه فریم‌هایی که اطلاعات عملی بهتری دارند، برای تشخیص موثر عمل‌های انسان بسیار مهم است.

تاثیر استفاده از فریم‌های کلیدی در پردازش ویدئوها در پایگاه داده‌ها در جداول (۳) تا (۵) نمایان است. همان‌طور که مشخص است دقت روش پیشنهادی افزایش یافته و در مقایسه با روش‌های رقیب از دقت بهتری برخوردار است. با اشتراکی شدن وزن‌ها، ماشین بولتزمن محدود دوبعدی پیاده‌سازی شده از حالت تطبیق بیش از حد خارج شده و به کارایی مطلوب خود رسیده است.

ایده استخراج فریم‌های کلیدی و تاثیر آن بر روی شبکه های دیگر در مقاله [۲۶] بررسی شده است. هر چند که شبکه پیشنهادی در کلاسه‌بندی ویدئوها تواناست، ولی زمانی که فریم-های با اطلاعات مفید و بدون افزونگی داده در اختیار این شبکه قرار می‌گیرد، کارآیی واقعی شبکه نمایان شده و نتایج بسیار خوبی در مقایسه با سایر روش های رقیب تولید می‌کند.

جدول (۳): روش‌های متفاوت انتخاب فریم به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده KTH

	همه فریم‌ها	فریم‌ها با فاصله α= ۱۳	فریم‌های کلیدی
روش	دقت (%)	دقت (%)	دقت (%)
Laptev et al.	۹۱,۴۰	۹۱,۴۰	۹۱,۴۰
Nazir et al.	۹۱,۸۲	۹۱,۸۲	۹۱,۸۲
Niebles et al.	۸۳,۳۳	۸۳,۳۳	۸۳,۳۳
Jhuang et al.	۹۱,۷۰	۹۱,۷۰	۹۱,۷۰
Taylor et al.	۹۰,۰۰	۹۰,۰۰	۹۰,۰۰
Le et al.	**۹۳,۹۰**	**۹۳,۹۰**	۹۳,۹۰
Ji et al.	۹۰,۲۰	۹۰,۲۰	۹۰,۲۰
Sun et al.	۹۳,۱۰	۹۳,۱۰	۹۳,۱۰
Zhang et al.	۹۲,۸۰	۹۲,۸۰	۹۲,۸۰
Han et al.	۹۳,۱۰	۹۳,۱۰	۹۳,۱۰
Abdelbaky and Aly	۹۳,۳۳	۹۳,۳۳	۹۳,۳۳
روش پیشنهادی	۸۱,۵۰	۸۵,۱۲	**۹۵,۰۲**

همان طور که از شکل‌های (۷) تا (۹) برمی‌آید، هیچ الگوی خاصی برای مقدار مناسب α وجود ندارد و این موضوع با آزمایشات مختلف مشخص گردیده است. همچنین، مشخص است که مقدارهای زیاد α باعث شده تا اطلاعات لازم در اختیار شبکه قرار نگیرد و آموزش به درستی صورت نگیرد. از طرفی دیگر، مقدارهای کم α باعث شده است اتفاق مربوط به آموزش بیش از حد به دلیل تعداد بالای فریم‌های ورودی دوباره تکرار شود. به همین سبب، راهی برای یافتن مقدار مناسب α(با توجه به پایگاه داده‌های مختلف) وجود ندارد و باید آزمون صورت گیرد.

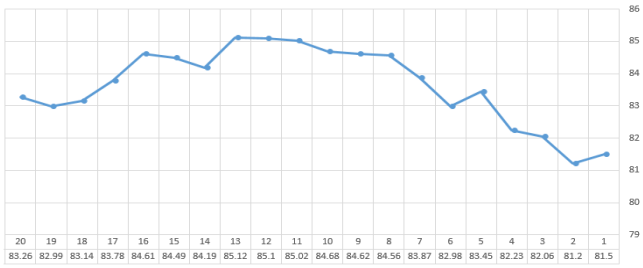
جدول (۴): روش‌های متفاوت انتخاب فریم به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده UCF

	همه فریم‌ها	فریم‌ها با فاصله α= ۱۶	فریم‌های کلیدی
روش	دقت (%)	دقت (%)	دقت (%)
Laptev et al.	۸۲,۶۰	۸۲,۶۰	۸۲,۶۰
Klaser et al.	۸۵,۶۰	۸۵,۶۰	۸۵,۶۰
Rahmani et al.	۹۰,۰۰	۹۰,۰۰	۹۰,۰۰
Le et al.	۸۶,۵۰	۸۶,۵۰	۸۶,۵۰
Zhang et al.	۸۶,۷۰	۸۶,۷۰	۸۶,۷۰
Sun et al.	۸۶,۶۰	۸۶,۶۰	۸۶,۶۰
Yuan et al.	۸۷,۳۰	۸۷,۳۰	۸۷,۳۰
Wang et al.	**۹۱,۸۹**	**۹۱,۸۹**	۹۱,۸۹
Ahmed and Aly	۸۶,۷۰	۸۶,۷۰	۸۶,۷۰
Abdelbaky and Aly	۹۰,۰۰	۹۰,۰۰	۹۰,۰۰
روش پیشنهادی	۷۸,۳۱	۸۸,۷۹	**۹۳,۱۴**

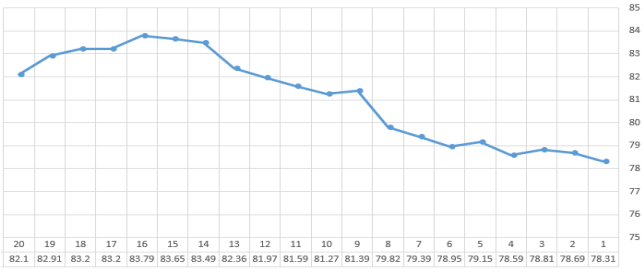
جدول (۵): روش‌های متفاوت انتخاب فریم به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده HMDB51

	همه فریم‌ها	فریم‌ها با فاصله α= ۸	فریم‌های کلیدی
روش	دقت (%)	دقت (%)	دقت (%)
Girdhar et al.	۵۲,۲	۵۲,۲	۵۲,۲
Meng et al.	۵۳,۱	۵۳,۱	۵۳,۱
C3D	۴۶,۷	۴۶,۷	۴۶,۷
P3D-199	۶۲,۹	۶۲,۹	۶۲,۹
Two-stream CNN	۵۹,۴	۵۹,۴	۵۹,۴
TDD	۶۳,۲	۶۳,۲	۶۳,۲
Two-stream fusion	۶۵,۴	۶۵,۴	۶۵,۴
TSN	۶۹,۴	۶۹,۴	۶۹,۴
TSN Cornet	۷۰,۶	۷۰,۶	۷۰,۶
MSM-ResNets	۶۶,۷	۶۶,۷	۶۶,۷
ARTNet-Res18	۷۰,۹	۷۰,۹	۷۰,۹
AMFNet-C	۷۱,۲	۷۱,۲	۷۱,۲
RSTAN (TSN)	۷۰,۵	۷۰,۵	۷۰,۵
STILT	**۷۲,۱**	**۷۲,۱**	۷۲,۱
روش پیشنهادی	۶۰,۲۳	۶۳,۲۰	**۷۴,۲۸**

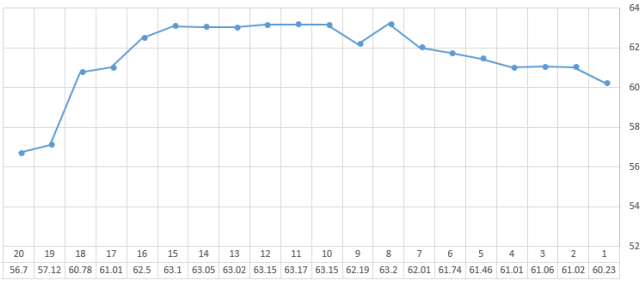
مورد دیگر آن که افزایش دقت روی پایگاه‌های داده UCF و HMDB51 در دو حالت استفاده از تمام فریم‌ها و همچنین استفاده از فریم‌های ویدئو با فاصله‌ی مشخص زمانی(α) به عنوان ورودی شبکه‌ی پیشنهادی کم‌تر از پایگاه داده‌ی KTH بوده است. به عبارت دقیق‌تر، روند آمده در این بخش کمک کم‌تری به آموزش روی پایگاه‌های داده‌ی مذکور کرده است که این موضوع را می‌توان در پیچیدگی بیش‌تر پایگاه داده‌های UCF و HMDB51 جستجو کرد. در کل نیز می‌توان این موضوع را متوجه شد که به دلیل پیچیدگی بیش‌تر پایگاه داده، افزایش دقت و کارایی سخت‌تر خواهد بود.



شکل (۷): دقت‌های مختلف به دست آمده برای مقادیر مختلف α با اجرا روی پایگاه داده‌های KTH

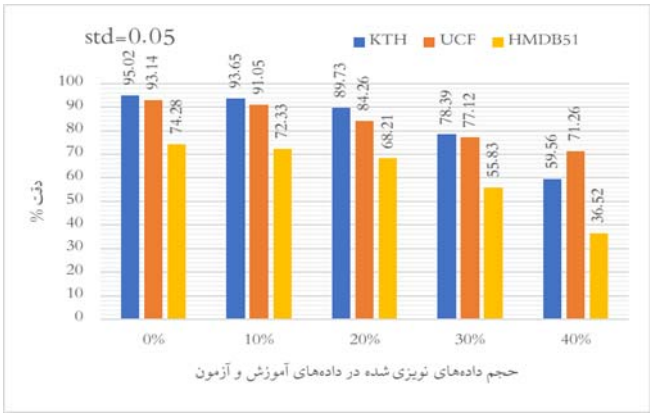


شکل (۸): دقت‌های مختلف به دست آمده برای مقادیر مختلف α با اجرا روی پایگاه داده‌های UCF

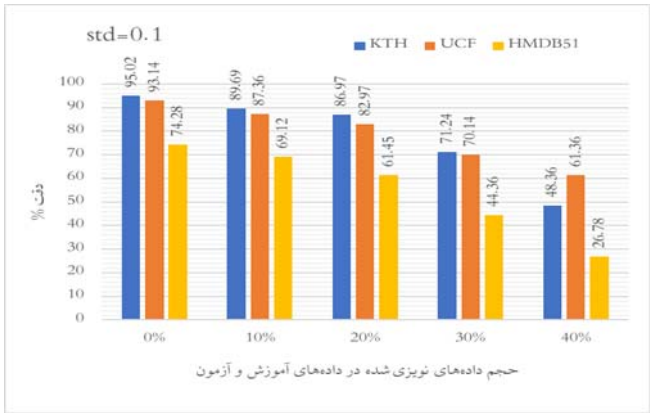


شکل (۹): دقت‌های مختلف به دست آمده برای مقادیر مختلف α با اجرا روی پایگاه داده‌های HMDB51

با توجه به بررسی انجام شده و مشخص شدن این موضوع که روش پیشنهادی در حالت استفاده از فریم کلیدی بهترین بازدهی را داشته است؛ عملکرد الگوریتم پیشنهادی روی چند سطح از وجود نویز در فریم‌های ویدیو با پیاده‌سازی استفاده از فریم‌های کلیدی بررسی می‌شود. نویز گوسی با میانگین صفر و انحراف معیارهای مختلف به صورت ساختگی روی فریم‌های ویدیو اعمال گردیده تا دقت روش پیشنهادی در کلاسه‌بندی ویدیو را با وجود نویز بسنجد. نتایج در شکل‌های (۱۰) تا (۱۳) آمده است. این اشکال دربرگیرنده نتایج آمده در جداول (۳) تا (۵) مربوط به سنجش دقت بدون وجود نویز نیز می‌باشند. چهار حالت دیگر که ۱۰%، ۲۰%، ۳۰% و ۴۰% داده های آموزش و آزمون نویزی باشند نیز در شکل‌های (۱۰) تا (۱۳) گزارش شده است. هر حالت، دارای چهار پیاده‌سازی برای نویز گوسی با انحراف معیارهای ۰,۰۵، ۰,۱، ۰,۱۵ و ۰,۲ است. مشخص است که با افزایش میزان داده‌های نویزی شده یا افزایش انحراف معیار از دقت کاسته می‌شود. همچنین، با وجود ۴۰% داده‌ی نویزی دقت کاهشی داشته است که بررسی درصدهای بالاتر مناسب نمی‌باشد. با توجه به داده‌های شکل‌های (۱۰) تا (۱۳) می‌توان برداشت کرد که روش پیشنهادی با ۱۰% یا ۲۰% داده‌ی نویزی کارایی قابل قبولی را با انحراف معیارهای مختلف در نویز گوسی ارایه داده است و برای ۳۰% نیز کارایی متوسط است. لیکن، با افزایش این مقدار به ۴۰% که تعداد بالایی داده‌ی نویزی را شامل می‌شده است؛ طبیعتا کارکرد روش پیشنهادی مختل شده است.



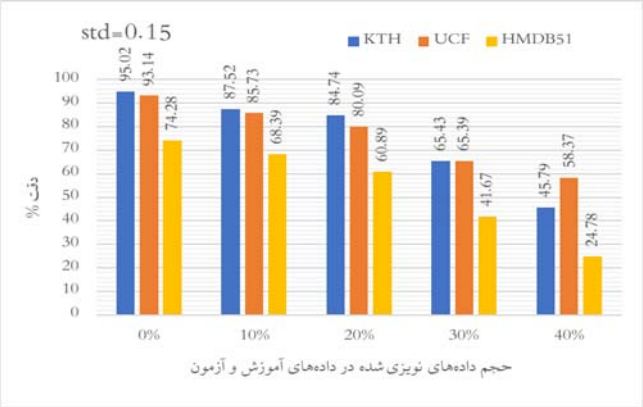
شکل (۱۰): دقت به دست آمده بر روی داده‌های نویزی شده از صفر تا ۴۰ درصد با نویز گوسی با میانگین صفر و انحراف معیار ۰,۰۵



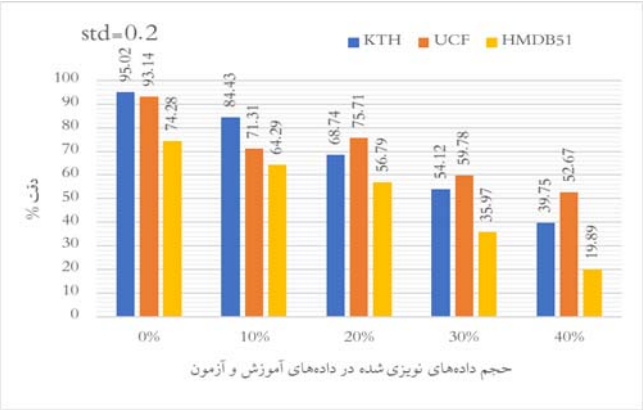
شکل (۱۱): دقت به دست آمده بر روی داده‌های نویزی شده از صفر تا ۴۰ درصد با نویز گوسی با میانگین صفر و انحراف معیار ۰,۱

۳-۳- ماتریس‌های درهم‌ریختگی

در زمینه یادگیری ماشین و به طور خاص مسئله‌ی طبقه‌بندی، یک ماتریس درهم‌ریختگی به یک طرح جدول خاص اطلاق می‌گردد که امکان تجسم عملکرد یک الگوریتم را دارد. این ماتریس در یادگیری بدون نظارت معمولا ماتریس تطبیق نامیده می‌شود. هر ردیف از ماتریس نمونه‌های یک کلاس واقعی را نشان می‌دهد؛ در حالی که هر ستون نشان دهنده‌ی نمونه‌های یک کلاس پیش‌بینی‌شده هستند (و یا برعکس). برای سادگی در نمایش جداول (۶) تا (۸) از نام اختصاری برای عمل‌های انجام شده در هر پایگاه داده که در بخش‌های ۳-۱-۱، ۳-۱-۲ و ۳-۱-۳ آورده شده‌اند، استفاده شده است.



شکل (۱۲): دقت به دست آمده بر روی داده‌های نویزی شده از صفر تا ۴۰ درصد با نویز گوسی با میانگین صفر و انحراف معیار ۰,۱۵



شکل (۱۳): دقت به دست آمده بر روی داده‌های نویزی شده از صفر تا ۴۰ درصد با نویز گوسی با میانگین صفر و انحراف معیار ۰,۲

۱-۳-۳- پایگاه داده‌ی KTH

در این بخش به ارایه ماتریس‌های درهم‌ریختگی جداول (۶) تا (۸) برای سه نوع پیاده‌سازی بر روی پایگاه داده‌ی KTH می‌پردازیم و مقدارهای صحت[1] و بازآوری[2] مربوط هر کلاس از پایگاه داده را ارایه می‌کنیم.

جدول (۶): جدول درهم‌ریختگی برای استفاده از همه فریم‌ها به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده KTH

		کلاس‌های واقعی						
		پر	د	تر	مز	دتد	کز	صحت%
نتایج طبقه بندی	پر	۰,۸۱	۰,۰۷	۰,۱	۰	۰	۰	۸۲,۶۵
	د	۰,۰۸	۰,۸۴	۰,۱۱	۰	۰	۰,۰۱	۸۰,۷۶
	تر	۰,۱۱	۰,۰۸	۰,۷۹	۰	۰	۰	۸۰,۶۱
	مز	۰	۰	۰	۰,۸۵	۰,۱۵	۰,۰۷	۷۹,۴۳
	دتد	۰	۰,۰۱	۰	۰,۱	۰,۷۸	۰,۱	۷۸,۷۸
	کز	۰	۰	۰	۰,۰۵	۰,۰۷	۰,۸۲	۸۷,۲۳
	بازآوری	%۸۱	%۸۴	%۷۹	%۸۵	%۷۸	%۸۲	

جدول (۷): جدول درهم‌ریختگی برای استفاده از فریم‌ها با فاصله زمانی α=13 به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده KTH

		کلاس‌های واقعی						
		پر	د	تر	مز	دتد	کز	صحت%
نتایج طبقه بندی	پر	۰,۹۵	۰	۰,۰۵	۰	۰	۰	۹۵
	د	۰,۰۴	۰,۹۷	۰,۰۵	۰	۰	۰	۹۱,۵
	تر	۰,۰۱	۰,۰۳	۰,۹۰	۰	۰	۰	۹۵,۷۴
	مز	۰	۰	۰	۰,۹۴	۰,۰۴	۰	۹۵,۹۸
	دتد	۰	۰	۰	۰,۰۴	۰,۹۲	۰,۰۵	۹۱,۰۸
	کز	۰	۰	۰	۰,۰۲	۰,۰۴	۰,۹۵	۹۴,۰۵
	بازآوری	%۹۵	%۹۷	%۹۰	%۹۴	%۹۲	%۹۵	

جدول (۸): جدول درهم‌ریختگی برای استفاده از فریم‌های کلیدی به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده KTH

		کلاس‌های واقعی						
		پر	د	تر	مز	دتد	کز	صحت%
نتایج طبقه بندی	پر	۰,۹۷	۰	۰,۰۵	۰	۰	۰	۹۵,۰۹
	د	۰,۰۲	۰,۹۷	۰,۰۵	۰	۰	۰	۹۳,۲۶
	تر	۰,۰۱	۰,۰۳	۰,۹۰	۰	۰	۰	۹۵,۷۴
	مز	۰	۰	۰	۰,۹۵	۰	۰	۱۰۰
	دتد	۰	۰	۰	۰,۰۳	۰,۹۶	۰,۰۵	۹۲,۳۰
	کز	۰	۰	۰	۰,۰۲	۰,۰۴	۰,۹۵	۹۴,۰۵
	بازآوری	%۹۷	%۹۷	%۹۰	%۹۵	%۹۶	%۹۵	

۲-۳-۳- پایگاه داده‌ی UCF

در این بخش به ارایه ماتریس‌های درهم‌ریختگی در جداول (۹) تا (۱۱) برای سه نوع پیاده‌سازی بر روی پایگاه داده‌ی UCF می‌پردازیم.

۳-۳-۳- پایگاه داده ی HMDB51

در این بخش به ارایه ماتریس‌های درهم ریختگی شکل‌های (۱۴) تا (۱۶) برای سه نوع پیاده سازی روی پایگاه داده ی HMDB51 می‌پردازیم. از آن جا که تعداد کلاس‌های این پایگاه داده زیاد است، برای رسم ماتریس درهم ریختگی از تکنیک ماتریس خاکستری استفاده می‌شود.

جدول (۹): جدول درهم‌ریختگی برای استفاده از تمام فریم‌ها به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده UCF

		کلاس‌های واقعی										
		ش	گ	لز	ب‌ک	اسب‌س	د	اسک‌س	ن‌ش	ن‌خ	ج	صحت %
نتایج طبقه بندی	ش	۰,۸۴	۰,۰۴	۰,۰۲	۰	۰,۰۸	۰	۰	۰	۰	۰	۸۵,۷۱
	گ	۰,۰۷	۰,۷۱	۰,۰۸	۰	۰,۱	۰,۰۴	۰,۰۹	۰	۰	۰	۶۵,۱۳
	لز	۰	۰,۰۵	۰,۸	۰	۰	۰,۰۳	۰	۰	۰	۰	۹۰,۹۰
	ب‌ک	۰,۰۵	۰	۰	۰,۸۶	۰	۰	۰	۰	۰,۰۸	۰	۸۶,۸۶
	اسب‌س	۰	۰,۰۶	۰	۰,۰۹	۰,۵۹	۰,۰۴	۰	۰,۰۶	۰	۰,۰۴	۶۷,۰۴
	د	۰	۰,۰۸	۰,۰۹	۰	۰,۱۱	۰,۷۸	۰	۰	۰	۰	۷۳,۵۸
	اسک‌س	۰,۰۴	۰	۰	۰,۰۵	۰	۰,۰۳	۰,۸۴	۰	۰,۱۱	۰,۰۷	۷۳,۶۸

[1] Precision.
[2] Recall.

	ت‌س	۰	۰	۰	۰	۰,۰۵	۰	۰	۰,۸۴	۰	۰,۰۸	۸۶,۵۹
	ت‌خ	۰	۰	۰	۰	۰	۰,۰۸	۰	۰	۰,۸۱	۰	۹۱,۰۱
	پ	۰	۰,۰۶	۰,۰۱	۰	۰,۰۷	۰	۰,۰۷	۰,۱	۰	۰,۸۱	۷۲,۳۲
	بازآوری	%۸۴	۷۱ %	۸۰ %	۸۶ %	۵۹ %	۷۸ %	%۸۴	۸۴ %	۸۱ %	%۸۱	

جدول (۱۰): جدول درهم‌ریختگی برای استفاده از فریم‌ها با فاصله زمانی α=16 به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده UCF

		کلاس‌های واقعی										
		ژ°	گ	ل‌ز	ب‌ک	اسب‌س	د	اسک‌س	ت‌س	ت‌خ	پ	صحت %
نتایج طبقه بندی	ش	۰,۸۵	۰	۰,۰۲	۰	۰,۰۸	۰	۰	۰	۰	۰	۸۹,۴۷
	گ	۰,۰۶	۰,۸۳	۰,۰۵	۰	۰,۰۹	۰,۰۳	۰,۰۹	۰	۰	۰	۷۲,۱۷
	ل‌ز	۰	۰,۰۵	۰,۸۸	۰	۰	۰,۰۳	۰	۰	۰	۰	۹۱,۶۶
	ب‌ک	۰,۰۵	۰	۰	۰,۸۹	۰	۰	۰	۰	۰,۰۵	۰	۸۹,۸۹
	اسب‌س	۰	۰,۰۴	۰	۰,۰۶	۰,۶۳	۰,۰۳	۰	۰,۰۶	۰	۰,۰۴	۷۳,۲۵
	د	۰	۰,۰۴	۰,۰۴	۰	۰,۰۹	۰,۸۵	۰	۰	۰	۰	۸۳,۳۳
	اسک‌س	۰,۰۴	۰	۰	۰,۰۵	۰	۰,۰۲	۰,۸۵	۰	۰,۰۷	۰,۰۶	۷۷,۹۸
	ت‌س	۰	۰	۰	۰	۰,۰۵	۰	۰	۰,۸۶	۰	۰,۰۴	۹۰,۵۲
	ت‌خ	۰	۰	۰	۰	۰	۰,۰۴	۰	۰	۰,۸۸	۰	۹۵,۶۵
	پ	۰	۰,۰۴	۰,۰۱	۰	۰,۰۶	۰	۰,۰۶	۰,۰۸	۰	۰,۸۶	۷۷,۴۷
	بازآوری	%۸۵	۸۳ %	۸۸ %	۸۹ %	۶۳ %	۸۵ %	%۸۵	۸۶ %	۸۸ %	%۸۶	

جدول (۱۱): جدول درهم‌ریختگی برای استفاده از فریم‌های کلیدی به عنوان ورودی شبکه‌ی پیشنهادی در پایگاه داده UCF

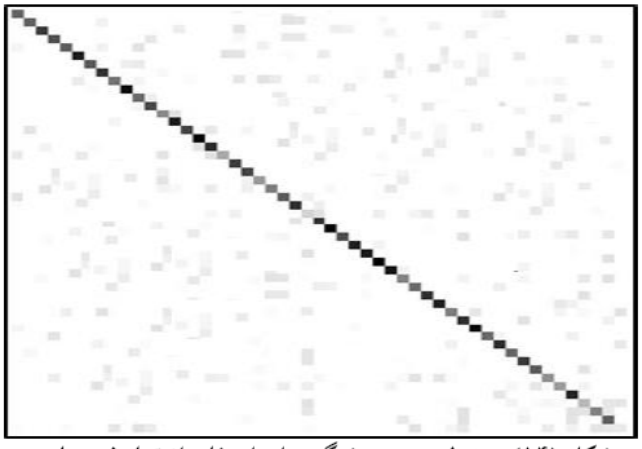
		کلاس‌های واقعی										
		ژ°	گ	ل‌ز	ب‌ک	اسب‌س	د	اسک‌س	ت‌س	ت‌خ	پ	صحت %
نتایج طبقه بندی	ش	۰,۹۵	۰	۰,۰۲	۰	۰,۰۴	۰	۰	۰	۰	۰	۹۴,۰۵
	گ	۰,۰۳	۰,۹۶	۰,۰۱	۰	۰,۰۱	۰,۰۳	۰,۰۵	۰	۰	۰	۸۸,۰۷
	ل‌ز	۰	۰,۰۲	۰,۹۴	۰	۰	۰,۰۲	۰	۰	۰	۰	۹۵,۹۱
	ب‌ک	۰,۲	۰	۰	۰,۹۶	۰	۰	۰	۰	۰,۰۳	۰	۹۵,۰۵
	اسب‌س	۰	۰	۰	۰,۰۲	۰,۸۴	۰,۰۳	۰	۰,۰۳	۰	۰	۹۱,۳
	د	۰	۰,۰۲	۰,۰۲	۰	۰,۰۳	۰,۹۱	۰	۰	۰	۰	۹۲,۸۵
	اسک‌س	۰	۰	۰	۰,۰۲	۰	۰	۰,۹۲	۰	۰,۰۳	۰,۰۲	۹۲,۹۲
	ت‌س	۰	۰	۰	۰	۰,۰۲	۰	۰	۰,۹۳	۰	۲	۹۵,۸۷
	ت‌خ	۰	۰	۰	۰	۰	۰,۰۱	۰	۰	۰,۹۴	۰	۹۸,۹۴
	پ	۰	۰	۰,۰۱	۰	۰,۰۶	۰	۰,۰۳	۰,۰۴	۰	۰,۹۶	۸۷,۲۷
	بازآوری	%۹۵	%۹۶	%۹۴	%۹۶	%۸۴	%۹۱	%۹۲	%۹۳	%۹۴	%۹۶	

۴-۳- جزئیات شبیه‌سازی

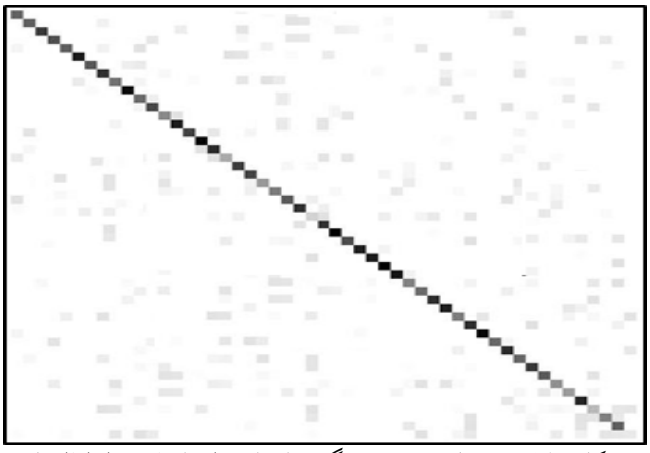
در این قسمت به پارامترهای روش پیشنهادی، تنظیمات و مقادیری که در پیاده سازی روش استفاده شده است می پردازیم. این موارد در جدول (۱۲) آمده است.

جدول (۱۲): مقادیر پارامترهای اشتفاده شده در شبکه

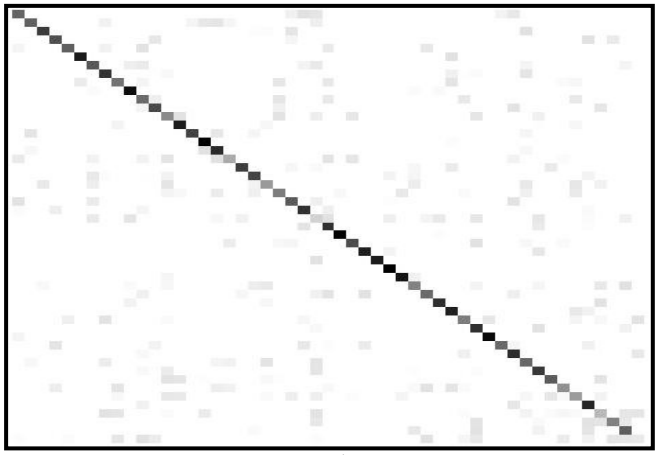
پارامتر یا تنظیم	شرح	مقدار	پایگاه داده
α	فاصله ی زمانی مشخص بین فریم ها	۱۳	KTH
α	فاصله ی زمانی مشخص بین فریم ها	۱۶	UCF
α	فاصله ی زمانی مشخص بین فریم ها	۸	HMDB51
P	تعداد سطرهای لایه های شبکه	۱۲۰	KTH
Q	تعداد ستون های لایه های شبکه	۱۶۰	KTH
P	تعداد سطرهای لایه های شبکه	۴۸۰	UCF
Q	تعداد ستون های لایه های شبکه	۷۲۰	UCF
P	تعداد سطرهای لایه های شبکه	۲۵۶	HMDB51
Q	تعداد ستون های لایه های شبکه	۳۴۰	HMDB51
L	تعداد نرون های لایه کلاس	۶	KTH
L	تعداد نرون های لایه کلاس	۱۰	UCF
L	تعداد نرون های لایه کلاس	۵۱	HMDB51
M	تعداد ویدیوهای پایگاه	۲۳۹۱	KTH
M	تعداد ویدیوهای پایگاه	۱۵۰	UCF
M	تعداد ویدیوهای پایگاه	۶۸۴۹	HMDB51
K	تعداد فریم های کلیدی استفاده شده	۱۶	KTH
K	تعداد فریم های کلیدی استفاده شده	۱۶	UCF
K	تعداد فریم های کلیدی استفاده شده	۱۶	HMDB51
نرخ یادگیری ماشین بولتزمن محدود		۰,۰۵	KTH
نرخ یادگیری ماشین بولتزمن محدود		۰,۰۵	UCF
نرخ یادگیری ماشین بولتزمن محدود		۰,۰۵	HMDB51
ممنتوم ماشین بولتزمن محدود		۰,۵	KTH
ممنتوم ماشین بولتزمن محدود		۰,۵	UCF
ممنتوم ماشین بولتزمن محدود		۰,۵	HMDB51



شکل (۱۴): جدول درهم ریختگی برای استفاده از تمام فریم‌ها به عنوان ورودی شبکه ی پیشنهادی در پایگاه داده HMDB51



شکل (۱۵): جدول درهم ریختگی برای استفاده از فریم‌ها با فاصله زمانی α=8 به عنوان ورودی شبکه ی پیشنهادی در پایگاه داده HMDB51



شکل (۱۶): جدول درهم ریختگی برای استفاده از فریم‌های کلیدی به عنوان ورودی شبکه ی پیشنهادی در پایگاه داده HMDB51

۴- نتیجه‌گیری

در این مقاله روشی بر مبنای ترکیب مفاهیم مربوط به ماشین بولتزمن محدود در شبکه‌های باور عمیق و شبکه‌های بازگشتی ارایه شده است. شبکه‌ی پیشنهادی بر مبنای ایده‌ی پشته‌سازی ماشین-های بولتزمن محدود در شبکه‌های باور عمیق بنا شده و همچنین در روش پیشنهادی این پشته‌سازی بر اساس شبکه‌های بازگشتی می‌باشد. ایده بازگشتی‌سازی پشته‌های ماشین بولتزمن محدود در شبکه‌های باور عمیق کمک می‌کند ساختار پیشنهادی به درک مفهوم بلند-مدت زمانی اهتمام ویژه‌ای داشته باشد. به همین سبب، یادگیری ویدئو برای این شبکه امکان پذیر خواهد بود. با توجه به دو موضوع زیر ایده‌ی بازگشتی‌سازی این نوع از شبکه‌ها کاملا به ذهن نزدیک است:

- هر چه لایه‌های ماشین‌های بولتزمن محدود به کار رفته در شبکه‌های باور عمیق بیش‌تر باشد، یادگیری بهتر خواهد شد.
- از نظر پتانسیل‌های محاسباتی و پیچیدگی‌های زمانی، امکان استفاده از تعداد بسیار بالای این ماشین‌ها در بدنه‌ی شبکه وجود ندارد.

پس، یکی از بهترین راه‌هایی که می‌توان قدرت شبکه‌های باور عمیق را با استفاده از اجرای بالای ماشین‌های بولتزمن محدود افزایش داد، بازگشتی‌سازی شبکه‌ی باور عمیق است.

شبکه‌ی بازگشتی پیشنهادی با دریافت یک توالی از فریم‌ها در ورودی خود تصمیم‌گیری می‌کند. زمانی که ورودی تغییر می‌کند؛ با توجه به اشتراکی بودن وزن‌ها در طول زمان، یادگیری در شبکه-ای ثابت انجام می‌شود. از آن جا که یک توالی، ورودی شبکه خواهد بود، برای به‌روزرسانی هر وزن از خطای به دست آمده از هر ورودی استفاده می‌شود. بنابراین، وزن‌ها در طول زمان و بر مبنای ورودی‌های مختلف به‌روزرسانی می‌شوند. این نوع از شبکه-ی پیشنهادی به دریافت مفاهیم کوتاه-مدت زمانی و بلند-مدت زمانی کمک شایانی می‌کند. نتایج به دست آمده حاکی از آن است که روش پیشنهادی قدرت پردازش و درک شبکه‌های باور عمیق را در حوزه‌ی ویدئو بالابرده و به تغییری مهم برای درک مفهوم زمان منجر شده است.

مراجع

[1] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black, "Towards Understanding Action Recognition," In Proceedings Of The IEEE International Conference On Computer Vision ،pp 3192-3199 ،2013.

[2] C. F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, Q. Fan, "Deep Analysis Of CNN-Based Spatio-Temporal Representations For Action Recognition," In Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition, pp. 6165-6175, 2021.

[3] A.M. Nickfarjam, H. Ebrahimpour-Komleh, "Multi-Input 1-Dimensional Deep Belief Network: Action And Activity Recognition As Case Study," Multimedia Tools and Applications, vol. 78, pp. 17739-17761, 2019.

[۴] فیروزه رضوی, محمد جعفر تارخ, محمود البرزی. "شناسایی آلزایمر با استفاده از شبکه عصبی یادگیری عمیق," مجله تحقیقات علوم رفتاری, جلد ۱۸, شماره ۲, صفحات ۲۶۰-۲۶۹, ۲۰۲۰.

[5] A. A. Tehrani, A. M. Nickfarjam, H. Ebrahimpour-Komleh, D. Aghadoost, "Multi-Input 2-Dimensional Deep Belief Network: Diabetic Retinopathy Grading As Case Study," Multimedia Tools And Applications, vol. 80, no. 4, pp. 6171-6186, 2021.

[6] J. Zhang, C. Ling, S. Li, "EMG Signals Based Human Action Recognition Via Deep Belief Networks," IFAC-Papersonline, vol. 52, no. 19, pp. 271-276, 2019.

[7] A. Fischer, C. Igel, "Training Restricted Boltzmann Machines: An Introduction," Pattern Recognition, 2014.

[8] R. Poppe, "A Survey On Vision-Based Human Action Recognition," Image And Vision Computing, vol. 28, no. 6, pp. 976-990, 2010.

[9] L. Wang, D. Suter, "Learning And Matching Of Dynamic Shape Manifolds For Human Action

Recognition," IEEE Transactions On Image Processing, vol. 16, no. 6, pp. 1646-1661, 2007.

[10] H. Zhang, F. Zhou, W. Zhang, X. Yuan, Z. Chen, "Real-Time Action Recognition Based On A Modified Deep Belief Network Model," in 2014 IEEE International Conference On Information And Automation (ICIA), pp. 225-228, July 2014.

[11] D. Batra, T. Chen, R. Sukthankar, "Space-Time Shapelets For Action Recognition," in 2008 IEEE Workshop On Motion And Video Computing, pp. 1-6, January 2008.

[12] A. Fathi, G. Mori, "Action Recognition By Learning Mid-Level Motion Features," in 2008 IEEE Conference On Computer Vision And Pattern Recognition, pp. 1-8, June 2008.

[13] N. Twomey, T. Diethe, X. Fafoutis, A. Elsts, R. Mcconville, P. Flach, I. Craddock, "A Comprehensive Study Of Activity Recognition Using Accelerometers," Informatics, vol. 5, no. 2, p. 27, June 2018.

[14] G. Varol, I. Laptev, C. Schmid, "Long-Term Temporal Convolutions For Action Recognition," IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 40, no. 6, pp. 1510-1517, 2017.

[15] L. Xia, C. C. Chen, J. K. Aggarwal, "View Invariant Human Action Recognition Using Histograms Of 3D Joints," in 2012 IEEE Computer Society Conference On Computer Vision And Pattern Recognition Workshops, pp. 20-27, June 2012.

[16] S. Y. Lin, Y. Y. Lin, C. S. Chen, Y. P. Hung, "Learning And Inferring Human Actions With Temporal Pyramid Features Based On Conditional Random Fields," in 2017 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 2617-2621, March 2017.

[17] B. Chen, "Deep Learning Of Invariant Spatio-Temporal Features From Video," Doctoral Dissertation, University Of British Columbia, 2010.

[18] M. Abdellaoui, A. Douik, "Human Action Recognition In Video Sequences Using Deep Belief Networks," Traitement Du Signal, vol. 37, no. 1, pp. 37-44, 2020.

[19] K. H. Ali, T. Wang, "Learning Features For Action Recognition And Identity With Deep Belief Networks," in 2014 International Conference On Audio, Language And Image Processing, pp. 129-132, July 2014.

[20] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, "Sequential Deep Learning For Human Action Recognition," in International Workshop On Human Behavior Understanding, pp. 29-39, November 2011.

[21] L. Wang, Y. Qiao, X. Tang, "Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors," in Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition, pp. 4305-4314, 2015.

[22] A. Ullah, K. Muhammad, I. U. Haq, S. W. Baik, "Action Recognition Using Optimized Deep Autoencoder And CNN For Surveillance Data Streams Of Non-Stationary Environments," Future Generation Computer Systems, vol. 96, pp. 386-397, 2019.

[23] N. Jaouedi, N. Boujnah, M. S. Bouhlel, "A New Hybrid Deep Learning Model For Human Action Recognition," Journal Of King Saud University-Computer And Information Sciences, vol. 32, no. 4, pp. 447-453, 2020.

[24] W. Wu, D. He, X. Tan, S. Chen, S. Wen, "Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6222-6231, 2019.

[25] R. V. Vamsi, D. Subburaman, "A Review on Video Summarization," in Proceedings of International Conference on Deep Learning, Computing and Intelligence, pp. 495-504, Springer, Singapore, 2022.

[26] N. Tasnim, J. H. Baek, "Deep Learning-Based Human Action Recognition with Key-Frames Sampling Using Ranking Methods," Applied Sciences, vol. 12, no. 9, p. 4165, 2022.

[27] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, S. Valaee, "Recent advances in recurrent neural networks," arXiv preprint arXiv:1801.01078, 2017.

[28] V. S. Lalapura, J. Amudha, H. S. Satheesh, "Recurrent neural networks for edge intelligence: a survey," ACM Computing Surveys (CSUR), vol. 54, no. 4, pp. 1-38, 2021.

[29] I. Sutskever, J. Martens, G. Dahl, G. Hinton, "On the importance of initialization and momentum in deep learning," in International conference on machine learning, pp. 1139–1147, 2013.

[30] I. Sutskever, J. Martens, G. E. Hinton, "Generating text with recurrent neural networks," in Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 1017–1024, 2011.

[31] S. R. Dubey, S. K. Singh, B. B. Chaudhuri, "Activation Functions in Deep Learning: A comprehensive Survey and Benchmark," Neurocomputing, 2022.

[32] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local svm approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., vol. 3, pp. 32–36, IEEE, 2004.

[33] M. D. Rodriguez, J. Ahmed, M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, IEEE, 2008.

[34] A. Abdelbaky, S. Aly, "Two-stream spatiotemporal feature fusion for human action recognition," The Visual Computer, vol. 37, no. 7, pp. 1821-1835, 2021.

[35] T. Liu, Y. Ma, W. Yang, W. Ji, R. Wang, P. Jiang, "Spatial-temporal interaction learning based two-stream network for action recognition," Information Sciences, vol. 606, pp. 864-876, 2022.

[36] Upadhya, Vidyadhar, and P. S. Sastry. "An overview of restricted Boltzmann machines." Journal of the Indian Institute of Science, vol. 99, pp. 225-236, 2019.

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. در حال حاضر دانشجوی دکترای کامپیوتر در گرایش هوش مصنوعی در دانشگاه کاشان مشغول به تحصیل هستند. زمینه مورد علاقه ایشان یادگیری ماشین، بینایی کامپیوتر، پردازش تصویر و ویدئو با استفاده از شبکه های یادگیر عمیق می‌باشد.

حسین ابراهیم‌پور کومله 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.