
بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

رشته مهندسی رباتیک

پایان‌نامه کارشناسی ارشد

شناسایی رفتارهای ناهنجار در تصاویر ویدئویی

نگارنده : بهنام سبزعلیان

استاد راهنما

دکتر حسین مروی

شهریور ۱۳۹۶

دانشگاه صنعتی شاهرود

دانشکده مهندسی برق و رباتیک

گروه مهندسی رباتیک

پایان نامه کارشناسی ارشد آقای بهنام سبزعلیان به شماره دانشجویی: ۹۳۰۹۲۳۴

تحت عنوان:

شناسایی رفتارهای ناهنجار در تصاویر ویدئویی

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم اثر

ماحصل آموخته‌هایم را تقدیم می‌کنم به آنان که مهر آسمانی‌شان آرام‌بخش آلام زمینی‌ام است

به استوارترین تکیه‌گاهم، دستان پرمهر پدرم

به سبزترین نگاه زندگی‌ام، چشمان پرامید مادرم

به مهربان‌ترین همراه زندگی‌ام، خواهر عزیزم

که هرچه آموختم در مکتب عشق شما آموختم و هرچه بکوشم قطره‌ای از دریای بی‌کران

مهربانی‌تان را سپاس نتوانم گفت.

امروز هستی‌ام به امید شماست و فردا کلید باغ بهشتم،

رضای شما را آوردی گران‌سنگ‌تر از این ارزان نداشتم تا به خاک پایتان نثار کنم، باشد که حاصل

تلاشم، نسیم گونه، غبار خستگی‌تان را بزداید.

بوسه بر دستان پرمهرتان

تشکر و قدردانی

سپاس و ستایش خدای را جل و جلال که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درخشان. آفریدگاری که خویشتن را به ما شناساند و درهای علم را بر ما گشود و عمری و فرصتی عطا فرمود تا بدان، بنده ضعیف خویش را در طریق علم و معرفت بیازماید.

«من لم یشکر المخلوق لم یشکر الخالق»

بر خود لازم می‌دانم از کلیه کسانی که بنده را در تدوین و نگارش این پایان‌نامه یاری نمودند، به‌خصوص از استاد فرزانه جناب آقای دکتر حسین مروی که به‌عنوان استاد راهنما در کلیه مراحل انجام این پژوهش با خوش‌روئی، یاری و راهنمایی‌ام نمودند و همچنین از استاد فرهیخته جناب آقای دکتر علیرضا احمدی فرد که به‌عنوان استاد مشاور وقت خود را در اختیارم گذاشتند و با دقت نظر خاصی مشاوره لازم در این خصوص ارائه نمودند صمیمانه تشکر و قدردانی می‌نمایم.

و در پایان از زحمات خانواده خوبم، دوستان عزیزم و سایر کسانی که در تدوین این تحقیق مرا یاری نمودند متشکرم و از خداوند منان سلامت و سعادت ایشان را خواستارم.

تعهد نامه

اینجانب **بهنام سبزعلیان** دانشجوی دوره کارشناسی ارشد رشته **مهندسی رباتیک** دانشکده **مهندسی برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه **شناسایی رفتارهای ناهنجار در تصاویر ویدئویی** تحت راهنمائی **دکتر حسین مروی** متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

شناسایی رفتار ناهنجار ازلحاظ اهمیت به یک ضرورت در سامانه‌های نظارت بصری و همچنین به‌عنوان یک چالش در تحقیقات بینایی ماشین بدل شده است. گرچه تلاش‌های بسیاری به‌منظور رفع این مشکل انجام‌شده است، اما شناسایی رفتار در یک محیط واقعی و غیرقابل‌کنترل فاصله معناداری تا رسیدن به حد نهایی خود دارد. مشکل اصلی ابهام در تفاوت خصوصیات رفتار عادی و غیر عادی نظیر درگیری، وجود اشیا مشکوک و ... است که تعریف آن معمولاً با توجه به زمینه پیشین تصاویر می‌تواند متفاوت باشد. در این پایان‌نامه یک سیستم شناسایی و موقعیت‌یابی رفتارهای ناهنجار در سکانس‌های ویدئویی ارائه‌شده است. جنبه کلیدی این روش درواقع ترکیب ویژگی‌های شبکه عصبی کانولوشنی زمان-مکانی با مجموعه ویژگی‌های دستی مانند Histograms of Optical Flow (HOF) و Histogram of Oriented Gradients (HOG) به‌منظور شناسایی رفتار ناهنجار در فریم‌های متوالی ویدئویی است. ویژگی‌های دستی مذکور توسط روش نوین اوزان تکرارشونده در تجزیه ماتریس‌های نامنفی Iterative Weighted Non-negative Matrix Factorization (IW-NMF) که بر مبنای روش معمول تجزیه ماتریس‌های نامنفی (NMF) است، تنک می‌شوند. این ویژگی‌ها تنها از حجم پیکسل‌های دارای حرکت استخراج می‌شوند تا بتوانند هزینه محاسبه را کاهش دهند. ساختار مدل شبکه عصبی کانولوشنی به ما اجازه استخراج ویژگی‌های زمان-مکانی را می‌دهد و از ویژگی‌های دستی به‌منظور اطمینان از مقاومت روش ارائه‌شده در برابر نویزهای محلی و افزایش دقت شناسایی بهره گرفته‌شده است.

سیستم ارائه‌شده را توسط مجموعه داده‌های متداول که شامل رفتارها و اعمال ناهنجار متفاوت انسانی در موقعیت‌های گوناگون است، مورد آزمایش و ارزیابی قراردادیم. نتایج به‌دست‌آمده از آزمایش‌های مربوطه نمایانگر این است که سیستم ارائه‌شده در مقایسه با بسیاری از روش‌های متداول پیشین، عملکرد بهتری را دارد و کارایی آن در شناسایی رفتار غیر عادی در مقایسه با روش‌های قبلی بسیار رقابتی است.

کلمات کلیدی

شناسایی رفتارهای ناهنجار در تصاویر ویدئویی، شبکه‌های عصبی کانولوشنی، بینایی ماشین، ویژگی‌های زمان-مکانی CNN، اوزان تکرارشونده در تجزیه ماتریس‌های نامنفی (IW-NMF)، تجزیه ماتریس‌های نامنفی NMF.

فهرست مطالب

فهرست تصاویر

فهرست جدول‌ها

فهرست اختصارات

MEI	Motion Energy Image
MHI	Motion History Image
STV	Space-Time Volume
STIP	Space-Time Interest Point
HOG	Histogram of Oriented Gradients
HOF	Histogram of Optical Flow
MBH	Motion Boundary Histogram
LBP	Local Binary Patterns
VLBP	Volume Local Binary Patterns
SVM	Support Vector Machines
BoV	Bag Of Visual word
FV	Fisher Vector
VLAD	Vector of Locally Aggregate Description
RNN	Recurrent Neural Networks
LSTM	Long-Short Term Memory
LDS	Linear Dynamics Systems
LRCN	Long-term Recurrent Convolutional Networks
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
IW-NMF	Iterative Weighted Nonnegative Matrix Factorization
ROFI	Region Of Foreground Interests
GMM	Gaussian Mixture Model
ROC	Receiver Operation Characteristic
EER	Equal Error Rate
TPR	True Positive Rate
FPR	False Positive Rate
TNR	True Negative Rate
FNR	False Negative Rate

فصل اول

۱- مقدمه

در این فصل به تعریف دقیق مسئله پیش رو، اهمیت، لزوم و اهداف آن می‌پردازیم. از چالش‌های موجود در حل آن سخن می‌گوییم و در انتها با عناوین مطرح‌شده در فصول آتی آشنا می‌شویم و نگاهی گذرا بر مفاد آن‌ها خواهیم داشت. بدین ترتیب به شرح کلیاتی می‌پردازیم که آشنایی با آن‌ها، مسیر حرکت درزمینه جاری را هموارتر می‌نماید.

۱-۱- کلیات

با توجه به گسترش فضای دسترسی به اطلاعات، نیاز به یافتن راه‌حل‌هایی که بتواند با این حجم بزرگ از اطلاعات تعامل داشته باشد بیش‌ازپیش احساس می‌شود. انسان‌ها قادر به یادگیری بازخوردهای پیچیده، تفکیک معنایی و مفهومی صحنه‌ها هستند. ما می‌توانیم به‌سرعت ارتباط بین دو نفر و یا یک نفر با محیط پیرامونش را درک کنیم و نرمال یا غیر نرمال بودن حرکت موردنظر را مشخص می‌کنیم. همچنین انسان‌ها توانایی طبقه‌بندی رویدادهای پیچیده یا تفسیر یک ویدیو را دارند.

اما حجم بزرگ اطلاعات حاضر به همراه مسائلی که در آن وجود دارد بر دشواری‌های این کار افزوده و در برخی موارد حتی انجام آن را غیرممکن ساخته است. بنابراین استفاده از ماشین‌ها برای حل این مشکلات امری ضروری به نظر می‌رسد.

یکی از این مشکلات که در آن حجم داده‌های بسیار زیاد در دسترس است و روش‌های خودکار در حل آن بسیار مفید واقع شدند، مسئله شناسایی ناهنجاری است. شناسایی ناهنجاری در ویدیوهای نظارتی نقش مهمی را در حفظ امنیت عمومی ایفا می‌کند.

تصاویر ویدئویی نظارتی از زمان معرفی آن تا به امروز دستخوش پیشرفت‌های شگفت‌انگیزی شده است. به‌طور مثال دوربین‌های CCTV برای امنیت عمومی شهر لندن در دهه ۶۰ میلادی با قرار دادن دو دوربین در منطقه میدان Trafalgar معرفی شد.

در دهه ۹۰ میلادی زیرساخت‌های امنیتی در شهر لندن به یک شبکه شامل هزاران دوربین افزایش و رشد پیدا کرد. در حال حاضر بیش از هشت میلیون دوربین امنیتی در بریتانیا قرارگرفته است. می‌توان به گزارش‌های مربوط به کشورهای مختلف جهان مانند کانادا، چین و استرالیا نیز اشاره داشت که این افزایش چشمگیر دوربین‌های نظارتی در این کشورها نیز تحقق پیداکرده است. این‌گونه سامانه‌ها در اختیار پلیس و نیروهای امنیتی قرار می‌گیرند تا بتوانند از وقوع جرم جلوگیری کنند. مزایای استفاده از دوربین‌های امنیتی روشن است، به‌جای اینکه نیروهای امنیتی و قانونی در هر گوشه‌ای مستقر شوند، منطقه و نواحی بسیار گسترده‌تری را می‌توان تنها در یک اتاق کنترل مشاهده کرد. اگر عملی در زمان وقوع آن شناسایی نشد، داده‌های ضبط‌شده می‌توانند برای اثبات جرم و تشخیص جنایتکار از قربانی به‌عنوان مدرک مورداستفاده قرار گیرد.

پیشرفت فنّاوری در دهه ۹۰ بسیار مؤثر واقع گشت، به‌طوری‌که پیشرفت فنّاوری و کاهش هزینه‌ها، نه‌تنها دوربین‌های نظارتی را برای دولت‌ها و شرکت‌های بزرگ مقرون‌به‌صرفه کرده است بلکه استفاده از آن را در شرکت‌های کوچک و مصارف خانگی نیز ممکن ساخته است. درنتیجه شمار دوربین‌ها به‌صورت نمایی رشد پیداکرد به‌عنوان‌مثال در سال ۲۰۰۲ تعداد ۴/۲ میلیون دوربین در بریتانیا، به‌طور متوسط یک دوربین برای ۱۴ نفر، مورداستفاده قرارگرفته است.

بنابراین مسئله کنونی این است که بسیاری از دوربین‌های نظارتی توسط مأموران امنیتی مشاهده نمی‌شود و مدل ضبط‌شده آن را تنها بعد از وقوع یک تصادف و یا جرم مورد بازبینی قرار می‌دهند. گرچه زمانی که دوربین‌ها به‌صورت زنده مشاهده می‌شوند، در بسیاری از اتاق‌های کنترل یک اپراتور مأمور مشاهده چندین دوربین در چندین جهت در یک‌زمان است و واقعیت امر این است که توجه و تمرکز آن اپراتور بین چندین دوربین تقسیم می‌شود و ممکن است در مورد رویداد موردنظر در تصاویر غفلت شود. مطالعه انجام‌شده در آزمایشگاه ملی ساندیا در وزارت انرژی ایالات‌متحده [۱] این ادعا را

ثابت می‌کند و نشان می‌دهد که بعد از مشاهده ۲۰ دقیقه از تصاویر و ارزیابی آن، میزان تمرکز به طرز چشم‌گیری کاهش میابد.

با توجه به این واقعیت برخی از مراحل شناسایی پردازش رفتار می‌تواند به‌صورت خودکار انجام شود تا از نادیده گرفتن اعمال مشکوک موردنظر توسط مأمور امنیتی جلوگیری شود.

بسیاری از سامانه‌های تجاری حاوی یک نرم‌افزار است که برخی از سطوح اعمال را تشخیص می‌دهد و می‌تواند یک هشدار را به صدا درآورد. روش ابتدایی که بیشترین اثر را در جلب‌توجه مأموران امنیتی دارد، شناسایی حرکات است. این روش می‌تواند دوره زمانی را که اشیا در آن حرکت می‌کنند، شناسایی کند و همچنین دوره‌ای که هیچ حرکتی در میدان دید دوربین قرار ندارد، نادیده بگیرد.

یک مدل پیشرفته‌تر، توسعه الگوریتم ردگیری‌اشیا است. علاوه بر آن‌که دوره‌های زمانی موردنظر را مشخص می‌کند، اطلاعاتی نیز در مورد مسیر حرکتی یک شی در صحنه می‌دهد.

Lucas و Kanade در سال ۱۹۸۱ یک الگوریتم ردگیری مبتنی بر کرنل ارائه کردند که البته تغییراتی در آن داده‌شده است که هنوز هم به‌عنوان یک روش پرکاربرد در زمینه ردگیری اشیا مورداستفاده قرار می‌گیرند.

فنّاوری آنالیز ویدیو اکنون می‌تواند رفتار و اعمال از پیش تعیین‌شده را در میان انبوه اطلاعات ویدئویی بازیابی کند. این اعمال توسط حرکات در نواحی محدودی تعریف می‌شوند. هشدارهای مختلف نیز می‌توانند توسط اعمال از پیش تعریف‌شده با استفاده از پیام متنی، هشدارهای روی صفحه مانیتور، ایمیل، نقشه‌های مشخص‌شده با نقاط به اجرا درآیند. اگر کاربر تنها زمانی که این اعمال اتفاق می‌افتند را ضبط کند فضای لازم برای ذخیره‌سازی تصاویر ویدئویی به طرز محسوسی کاهش میابد.

مقایسه‌ای در مورد توانایی‌های سیستم آنالیز تصاویر ویدیوهای امنیتی در [۱] ارائه‌شده است. که در جدول ۱-۱ نمایش داده‌شده است.

جدول ۱-۱- سامانه‌های تجاری آنالیز ویدیوهای نظارتی [۱]

منابع	Track	Breach	Crowd	Left or removed Object	Loiter	Fall	Object Classification	Event
Agent Video Intelligence [2]	•	•	•	•	•			
Honeywell [3]	•	•	•	•	•			
IntelliVision [4]	•	•	•	•		•		
Ipsotek [5]	•	•	•	•	•			
Mango [6]	•	•	•	•			•	
ObjectVideo [7]	•	•	•	•			•	
Puretechsystems [8]	•	•	•	•	•		•	
Acic [9]	•	•	•	•	•	•	•	•
VideoIQ [10]	•	•	•	•	•	•	•	•
Iomniscient [11]	•	•	•	•	•	•	•	•

این جدول حاوی لیستی از اسامی شرکت‌های آنالیزکننده در حوزه تصاویر امنیتی است که مورداستفاده مأموران امنیتی قرار می‌گیرند. همچنین این لیست اهداف استاندارد سامانه‌های نظارتی شرکت‌های آنالیز ویدیو که مورد انتظار مأموران امنیتی است را نشان می‌دهد. این اهداف عبارت‌اند از شناسایی تأخیر درحرکت، شناسایی اعمال نامتعارف با محیط و شناسایی‌اشیا نامربوط. علاوه بر این موارد برخی از برنامه‌های ارائه‌شده، اعمال نامتعارف بیشتری مانند شناسایی سقوط، شناسایی رفتار و شناسایی تغییرات موقعیت شی را مورد هدف خود قرار می‌دهند.

با توجه به جدول فوق می‌توان مشاهده کرد که ردگیری اشیا یک بخش اساسی از تمامی سامانه‌های آنالیز ویدیو در تصاویر نظارتی است. شناسایی اعمال نامتعارف با محیط نیز جزو اهدافی است که در تمام این سامانه‌ها موردتوجه قرارگرفته است. اکثر سامانه‌های آنالیز ویدیو دارای توانایی شناسایی اشیا اعم از اشیا جامانده یا حذف‌شده از صحنه و برخی از آن‌ها توانایی آنالیزهای ازدحام را دارا هستند. تعداد اندکی از آن‌ها قابلیت شناسایی تأخیر درحرکت و شناسایی اعمال و رفتار مانند سقوط را دارند.

در حالت کلی الگوریتم شناسایی حرکات نسبت به تغییرات روشنایی، حرکات دوربین، حرکات در پس‌زمینه تصاویر مانند حرکات شاخه‌ها و برگ‌ها، فاصله وسایل نقلیه حساس است. حال این مسئله که چگونه سیستم آنالیز ویدئویی ارتقا پیدا کند و از روش‌های پیچیده اجتناب شود یکی از سؤال‌های اساسی در تحقیقات جاری است. پروژه‌های بسیاری با تمرکز بر سامانه‌های نظارتی، شناسایی انسان، شناسایی اشیا و ... معرفی‌شده‌اند. برخی از این پروژه‌های معتبر در جدول ۱-۲ آورده شده است.

بخش اصلی تحقیقات بر روی این برنامه‌ها به این دلیل که سامانه‌های نظارتی به‌طور گسترده به دوربین‌های ویدئویی وابستگی دارند، در حوزه بینایی ماشین متمرکز است. سامانه‌هایی که اهداف موردنظر در نظارت بصری را تأمین می‌کنند عبارت‌اند از VACE , HID و THIS است.

سازمان‌های نظامی و دفاعی در حمایت مالی از بسیاری پروژه‌های نظارت بصری پیشگام هستند. به‌عنوان‌مثال سازمان تحقیقات پروژه‌های پیشرفته وزارت دفاع ایالات‌متحده (DARPA) در دفتر سامانه‌های اطلاعاتی خود در ۱۹۹۸ از پروژه AVS ، به دنبال آن از یک برنامه سه‌ساله برای توسعه VSAM در سال ۱۹۹۷ و پروژه دوساله برای توسعه VIRAT در سال ۲۰۰۸ پشتیبانی کرده است.

جدول ۱-۲- لیست پروژه‌های انجام‌شده در حوزه نظارت بصری [۱]

نام تجاری	نام پروژه	مدت پروژه	منبع
Smarter Transportation	Smarter Transportation	2014	IBM
PROTECTRAIL	Integrated security of Rail Transport	2010-2014	Dambra
THIS	Transport Hubs Intelligent Surveillance	2009-2010	Saldatos
VIRAT	Video and Image Retrieval Analysis Tool	2008-2010	DARPA

VSAM	Video Surveillance And Monitoring	2007-2010	CMU
AVS	Airborne Visual Surveillance	1998-2002	DARPA
HID	Human Identification at Distance	2000-2004	GVU
VACE I/II	Video Analysis and Content Extraction	2000-2005	CMU
MAVISS	Multi-modal Audio Visible and Infrared Surveillance System	2005	Kumar

هدف ما از انجام این پایان‌نامه، شناسایی رفتارهای ناهنجار و نامتعارف در تصاویر نظارتی است. در این راستا سؤالاتی مطرح می‌شوند که می‌توان به مهم‌ترین این سؤالات پرداخت.

۱. آیا می‌توان رفتارهای نامتعارف و ناهنجار را از طریق داده ویدئویی تشخیص داد؟

۲. چه رفتاری را می‌توان به‌عنوان رفتار ناهنجار تعریف نمود؟

۳. چگونه رفتارهای ناهنجار را از غیر آن، تمیز دهیم؟

۴. چه روش‌هایی برای افزایش کارایی سامانه‌های آنالیز تصاویر نظارتی مورداستفاده قرار می‌گیرند؟

۵. با چه رویکردی غیرمتعارف بودن رفتار از جهات گوناگون را تجمیع کرده و به تصمیمی منطقی برسیم؟

۱-۲- اهداف پایان‌نامه

با توجه به رشد فزاینده دوربین‌های امنیتی و بار هزینه‌ای زیاد نظارت انسانی، انگیزه استفاده از سامانه‌های امنیتی خودکار افزایش می‌یابد. درواقع این سامانه‌ها تلاش دارند کارایی استفاده از دوربین‌های نظارتی را بهبود بخشند. یکی از مزایای استفاده از سامانه‌های تشخیص رفتار ناهنجار پیشرفت استفاده از دوربین‌های نظارتی در تجارت‌های کوچک و یا خانه‌ها است. سیستم نظارت دوربین‌ها علاوه بر این‌که می‌تواند از وقوع جرم جلوگیری کند و به‌عنوان یک ابزار آنالیز حوادث مورداستفاده قرار گیرد، همچنین می‌تواند به‌عنوان یک سامانه کمکی در حوادث ناگوار دخالت داشته باشد. هدف اصلی از این تحقیقات یافتن روشی برای شناسایی رفتارهای ناهنجار است.

یک رفتار ناهنجار در محیط که موردتحقیق قرارگرفته می‌گیرد، رفتاری است که در تصاویر ویدئویی تعریف‌نشده است، اما در محیط موردنظر اتفاق می‌افتد. در حالت کلی این رفتار توسط انسان‌ها واقع می‌گردند. مثال‌هایی از این رفتارها عبارت‌اند از راه رفتن، دویدن، درگیری و ... .

یک رفتار معمولاً از ارتباط بین چند نفر تشکیل می‌شود به‌طور مثال درگیری، راه رفتن و... . یک رویداد می‌تواند حتی توسط انسان انجام نشود بلکه توسط اشیای دیگر مانند وسایل نقلیه یا حیوانات انجام گیرد. تعریف خاصی که می‌توان از رفتارهای ناهنجار داشت بسیار به محیط موردنظر بستگی دارد. درواقع بسیاری از رفتارهای بین افراد در یک صحنه نظارتی جزو موارد قابل‌توجه به شمار نمی‌آید نظیر این رفتارها می‌توان به ملاقات بین افراد، در آغوش گرفتن و ... اشاره داشت. به دلیل اینکه این رفتارها درصحنه‌های بسیاری قابل‌پذیرش است و به‌طورمعمول اتفاق می‌افتد بنابراین می‌توان گفت که رفتارهای ناهنجار به‌صورت معمول اتفاق نمی‌افتند و در اذهان عمومی ممکن است قابل‌پذیرش نباشد مانند رفتار درگیری افراد.

چالش یک سیستم خودکار این است که بین رفتارهای حادث، رفتارهای نرمال از غیر نرمال را مشخص کند. در این پایان‌نامه یک رفتار ناهنجار به همان معنی رفتار غیر نرمال یا عمل غیر نرمال است گرچه

ممکن است در معنای لغوی کمی با یکدیگر متفاوت باشند. سامانه‌های دوربین‌های نظارتی می‌توانند بر مبنای چند-دوربین و یا تک-دوربین سناریوسازی شده باشند.

در این پایان‌نامه سامانه نظارتی دوربین‌ها به‌عنوان یک سیستم نظارتی مبتنی بر یک دوربین تعریف می‌شود که بتواند توسط یک کاربر ساده قابل‌استفاده باشد. از چالش‌هایی که سامانه‌های شناسایی رفتارهای ناهنجار با آن‌ها مواجه هستند می‌توان به یافتن تعریف دقیقی از رفتار نرمال از غیر نرمال در یک محیط خاص، یافتن اطلاعات تصویری مناسب، انتخاب روش‌های مناسب آموزش و طبقه‌بندی و به دست آوردن نتایجی با نرخ False-Positive کم و نرخ True-Positive بالا اشاره داشت. قسمت اعظم تحقیقات در حوزه‌های یادگیری ماشین و بینایی ماشین بر روی بهبود کارایی سامانه‌های نظارت بصری متمرکزشده است. اهداف این تحقیقات در ابتدا شناسایی محدودیت‌هایی است که در محیط‌های واقعی اتفاق می‌افتد و بعدازآن آزمایش و ارائه مجموعه روش‌هایی که می‌توانند تا حدودی بر شرایط چالش‌های طبیعی فائق آیند.

درروش ارائه‌شده در این پایان‌نامه تمامی پروسه موردنظر برای استخراج ویژگی از یک داده خام بصری پوشش داده‌شده است. این ویژگی‌ها کمک می‌کنند که یک تصمیم‌گیری در مورد رفتارهای غیرمتعارف انجام شود. اهداف ما در این پایان‌نامه در موارد زیر خلاصه می‌شوند.

۱. تعریف رفتارهای مشکوک و استفاده از دو رویکرد تشخیص ناهنجاری و تشخیص رفتارهای نامتعارف .

۲. ارزیابی روش‌های ارائه‌شده پیشین در حوزه توصیفگرهای شناسایی رفتار.

۳. تعیین مجموعه ویژگی‌های مؤثر در درک رفتار انسان.

۴. ترکیب روش‌های طبقه‌بندی با ناظر و بدون ناظر به‌منظور بهبود کارایی سیستم شناسایی روش رفتار نامتعارف.

۱-۳- کاربردهای مسئله

همان‌گونه که در بخش‌های پیشین اشاره شد، آنالیز تصاویر نظارتی یکی از اساسی‌ترین مسائل در حوزه تأمین امنیت عمومی است. در سال‌های اخیر به دلیل افزایش موارد سرقت و جنایت و عملیات تروریستی خصوصاً در کلان‌شهرهای جهان میزان استفاده از دوربین‌های نظارتی رشد فزاینده‌ای داشته است. موضوع پایان‌نامه حاضر نیز در جهت پاسخگویی به این نیاز موردبررسی و مطالعه قرارگرفته است. در زیر به مواردی از کاربردهای مسئله فوق اشاره می‌کنیم.

از کاربردهای این سیستم می‌توان به استفاده در بانک‌ها، مؤسسات مالی و بورس اوراق بهادار، موزه‌ها جهت تشخیص سارق، اماکن و معابر عمومی نظیر مراکز خرید، سالن‌های انتظار مترو، راه‌آهن ،فرودگاه، زندان‌ها و مراکز امنیتی جهت تشخیص عبور و مرورهای غیرقانونی، بیمارستان‌ها و تشخیص حالات غیرعادی بیمار، استخرها، کاربردهای خانگی نظیر شناسایی افرادی که قصد ورود غیرقانونی به خانه رادارند، تشخیص و جلوگیری از اعمال تروریستی مانند کار گذاشتن بمب یا درگیری و ... اشاره داشت.

موضوع پایان‌نامه حاضر شناسایی رفتارهای ناهنجار در تصاویر نظارتی، خصوصاً در اماکن عمومی است که در میان اهداف ذکرشده از اهمیت بالایی برخوردار است.

۱-۴- ساختار پایان‌نامه

در فصل اول مقدمه‌ای درباره موضوع پایان‌نامه ارائه شد. در فصل دوم به تعاریف اولیه مربوط به موضوع پایان‌نامه و همچنین مرور کارهای ارائه‌شده پیشین در حوزه نظارت بصری، آنالیز حرکات انسان و شناسایی رفتارها و اعمال غیر نرمال پرداخته می‌شود. در فصل سوم به معرفی روش پیشنهادی و همچنین بیان جزئیات لازم می‌پردازیم. فصل چهارم به بررسی و ارزیابی نتایج و جمع‌بندی مطالب پایان‌نامه اختصاص دارد.

فصل دوم

۲- مروری بر پژوهش‌های پیشین

۲-۱- مقدمه

حرکات انسان ، از حرکت ساده یک عضو تا حرکات گروهی اعضای بدن گسترده است. به‌عنوان نمونه حرکت پا برای ضربه زدن به توپ فوتبال، حرکتی ساده است اما پریدن و ضربه سر زدن به توپ را می‌توان ترکیبی از حرکات پا ، دست ، سر و بدن دانست که حرکتی پیچیده محسوب می‌شود. باوجوداین تعاریف، توصیف یک Action دشوار خواهد بود. در ادامه نمونه‌هایی از مقالات منتشرشده آورده شده است.

- مقالات [۱۲] و [۱۳] تعریفی اولیه از Action را ارائه می‌دهند و آن را یک حرکت در عضو بدن توصیف می‌کند. بنابراین Action به‌صورت یکسری حرکات با محدوده گوناگون از حالت اولیه و ساده تا حرکات سریالی بدن تعریف می‌گردد. طبق این تعریف به‌عنوان‌مثال حرکت روبه‌جلوی پای چپ به‌عنوان یک حرکت اولیه به‌منظور عمل پیاده‌روی محسوب می‌شود ، درحالی‌که پرش با مانع شامل Action های دویدن و پریدن است.
- در مقاله [۱۴] Action را تحت عنوان یک الگوی حرکتی ساده تعریف می‌کند که توسط یک فرد انجام و در یک‌زمان کوتاه به طول می‌انجامد و رفتار یا فعالیت نیز به یک سکانس از حرکت‌ها بستگی دارد که توسط چندین نفر انسان، که می‌توانند در حوزه محدودی باهم مرتبط باشند ، انجام می‌شود. به‌عنوان‌مثال می‌توان از Action های پیاده راه رفتن یا شنا کردن نام برد و در مورد Activity ها می‌توان به دست دادن دو نفر یا به ثمر رساندن گل توسط یک تیم فوتبال اشاره داشت.
- مقاله [۱۵] معنای درست‌تری از Action ارائه می‌دهد. ازنظر این مقاله اگر تغییر یا تبدیلی در محیط رخ دهد یک Action اتفاق افتاده است. به‌عنوان‌مثال می‌توان ضربه زدن به یک توپ را Action نامید.

ما بهترین تعریفی که اهدافمان را در این پایان‌نامه پوشش می‌دهد را برمی‌گزینیم.

" Action ، بیشترین فعل‌وانفعالات بامعنایی که در اطراف انسان رخ می‌دهد ، است. "

معنایی که به این فعل‌وانفعالات برمی‌گردد را می‌توان رده‌ای از Action نامید. به‌طورکلی Action های انسانی را می‌توان از فرم‌های مختلف فیزیکی به دست آورد. در این تعریف فعل‌وانفعالات را می‌توان توسط حرکات مرتبط باهم درک کرد که ممکن است تغییری در محیط اطراف ایجاد نماید. در برخی از مواقع ممکن است محیط اطراف از اشیای معناداری تشکیل‌شده باشد. این تعریف هم‌راستا با مقاله [۱۵] نیز هست، که یک عمل را با توجه به تغییراتی که در محیط ایجاد کرده است تعریف می‌کند. به‌عنوان‌مثال یک سکانس حرکتی در تصویر ۲-۱ نمایش داده‌شده است.

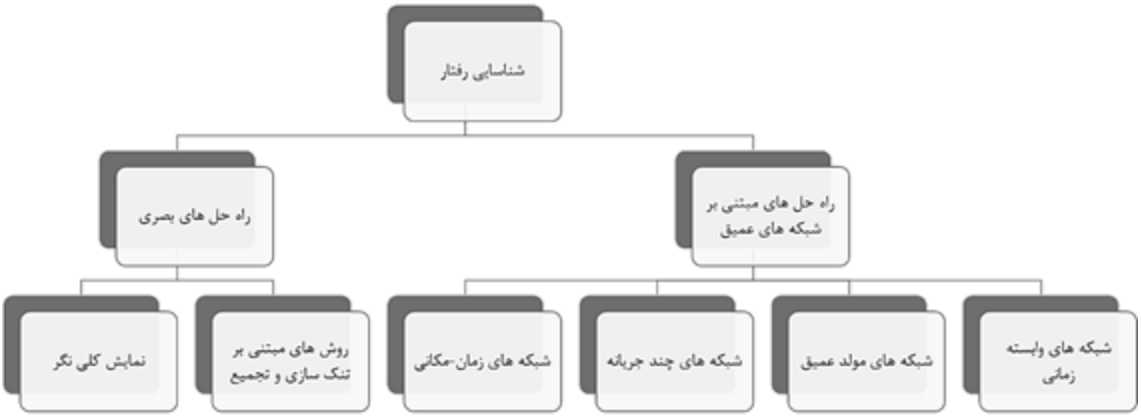
تصویر ۲-۱ - Action هایی که فعل‌وانفعالات بامعنایی بین انسان و محیط دارد.

ابتدا یک حرکت اولیه توسط پای بازیکن برای دویدن انجام می‌شود. این حرکت‌ها یک حرکت به‌هم‌پیوسته است که در محیط انجام می‌شود تا یک معنی را القا نماید. به‌عبارت‌دیگر مجموعه‌ای از حرکات به‌هم‌پیوسته اعضا که دویدن را نتیجه می‌دهد دارای یک معنی است. زمانی که این مجموعه حرکات از بیشترین حرکات اولیه و از پرمعناترین حرکات موردنظر باشد به آن به‌عنوان یک عمل بانام " Running Action " نگاه می‌شود. به‌طور مشابه بدیهی است که ضربه بازیکن و پریدن دروازه‌بان شامل ۲ عمل جدا از هم است که برچسب‌های "ضربه زدن" و "پریدن" را دارد.

در این پایان‌نامه سعی شده است که روش‌های ارائه‌شده بر مبنای درک خواننده طبقه‌بندی شود. به همین دلیل یک بخش جدا برای مطرح کردن تکنیک‌ها و روش‌های مبتنی بر شبکه‌های عمیق[۱]،

[۱] Deep Networks

ساختارهای گوناگون و روش‌های آموزش آن‌ها تدارک دیده‌شده است. تصویر ۲-۲ نگاه جامع به موضوعاتی دارد که در این فصل موردبررسی قرار می‌گیرد.



تصویر ۲-۲- شمای کلی موضوعاتی که در این فصل موردبررسی قرار خواهد گرفت.

آنالیز حرکات و اعمال دارای سابقه طولانی است و برای موضوعات مختلفی نظیر علوم روان‌شناختی، بیولوژی و علوم کامپیوتر بسیار جذاب است.

شناسایی اعمال و رفتار در یک محدوده وسیع‌تری از برنامه‌های اجتماعی، از تصاویر دوربین‌های نظارتی گرفته تا ارتباط بین کامپیوتر- انسان، آنالیز جزئیات، طراحی واسط کاربری، آموزش ربات‌ها، جستجو ، پیمایش تصاویر تحت وب، مسائل پزشکی، بهبود کیفیت زندگی برای افراد مسن، آنالیز ورزشی و ... گسترش پیداکرده است. با توجه به لیست بلندی از برنامه‌ها و متدهایی که ارائه‌شده است می‌توان به این نکته اشاره داشت که آنالیز دستی حرکات و رفتارها بسیار دشوار است.

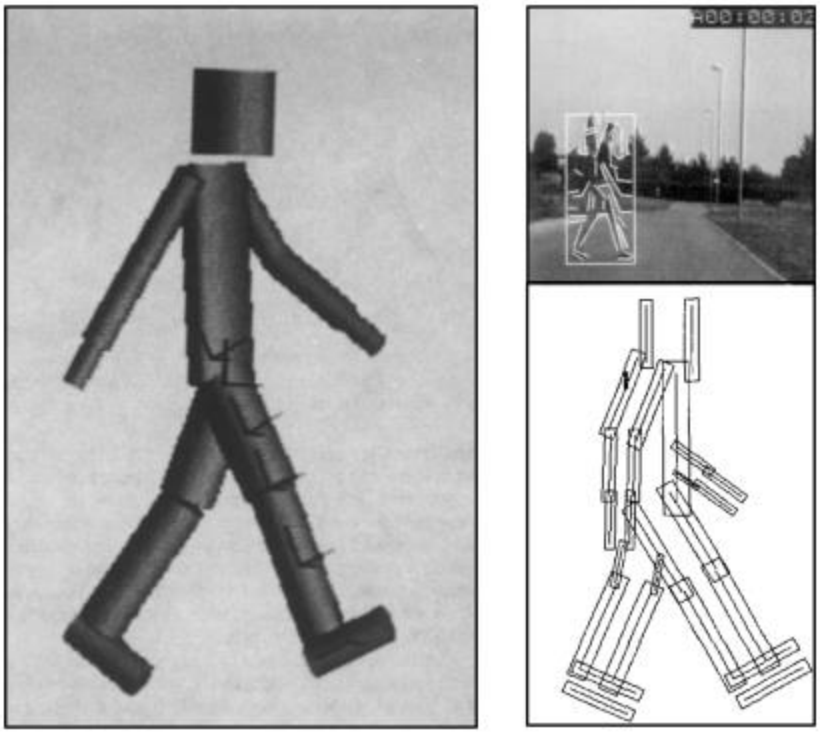
در ابتدا بابیان نقل‌قول از روشی که در دهه ۸۰ میلادی ارائه گردیده است شروع می‌کنیم که گفته است "ابتدا می‌بایست یک سیستم سمبلیک برای ارائه شمایل اطلاعات در مغز وجود داشته باشد و دوم اینکه

مغز می‌بایست شامل یک مجموعه از پردازشگرها باشد که قابلیت استخراج این اطلاعات را از تصاویر داشته باشد." [۱۶].

در زمینه شناسایی رفتار ، یک روش مناسب برای ارائه می‌بایست "دارای محاسبات ساده" و "دارای توصیفی مناسب برای یک مجموعه‌ی عظیم از اعمال" و "بازتاب تشابه‌ها بین دو رفتار همانند" و "مقاوم در برابر تغییرات مختلف (مانند تغییر روشنایی و ...) " باشد.

در تحقیقات جدیدتر در حوزه شناسایی رفتار ، از مدل‌های 3D برای توصیف رفتارها استفاده‌شده است. به‌عنوان یک مثال قابل‌توجه می‌توان به مدل سلسله مراتبی WALKER که در [۱۷] به‌منظور درک و تفسیر حرکات انسانی ارائه‌شده است، اشاره داشت. به‌عنوان مثالی دیگر می‌توان به استفاده از استوانه‌های متصل برای مدل کردن اعضای متصل بدن به‌منظور شناسایی افراد پیاده که توسط [۱۸] ارائه شد، اشاره کرد. در تصویر ۲-۳ این مدل‌های 3D نمایش داده‌شده‌اند. بااین‌حال یافتن دقیق مدل‌های 3D دشوار و پرهزینه است. به همین دلیل توسعه‌دهندگان فعلی از مدل‌سازی 3D اجتناب می‌کنند و در عوض برای نمایش اعمال سطوح Holistic یا Local را برمی‌گزینند.

- **نمایش کلی‌نگر** ؛ شناسایی اعمال بر مبنای استخراج یک نمایش کلی از ساختار بدن انسان، شکل و حرکت آن است.
- **نمایش محلی** ؛ شناسایی اعمال بر مبنای استخراج ویژگی‌های محلی صورت می‌پذیرد.



*تصویر ۲-۳ – مدل‌های 3D که نمایانگر اعمال است. **چپ:** مقاله [۱۷] مدل 3D مربوط به WALKER به‌منظور نمایش عمل راه رفتن ارائه کرده است. **راست:** مقاله [۱۸] مدل WALKER را به‌منظور شناسایی افراد پیاده گسترش داده است. در این مدل از استوانه‌های متصل برای شناخت افراد پیاده بهره برده است.*

۲-۲- نمایش کلی نگر[۲]

ابتدا به بررسی متد قابل‌تأمل در [۱۹] می‌پردازیم. تصویر انرژی حرکت MEI[۳] و تصویر سابقه حرکت MHI[۴] در این مقاله، ارائه گردیده است. با توجه به نام پیشنهادی، ایده اصلی آن کد کردن اطلاعات حرکات مرتبط در یک تصویر است. MEI یک تصویر باینری است که حرکات حادث را توصیف می‌کند و به‌صورت زیر تعریف می‌شود.

$$E_{\tau}^{(x,y,t-i)} = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \qquad (۱-۲)$$

[۲] Holistic

[۳] Motion Energy Image.

[۴] Motion History Image.

در این رابطه $D(x,y,t)$ یک توالی تصویر باینری است که پیکسل‌های شی تشخیص داده‌شده را نمایش می‌دهد ، E_t بیانگر MEI در زمان t است. MHI نشان می‌دهد که چگونه تصویر به حرکت درمی‌آید. هر پیکسل از MHI یک تابعی از سابقه حرکت در آن نقطه است. (شدت بیشتر متناظر با حرکت بیشتر است.).

تصویر ۲-۴ – بالا: توالی حرکت پریدن است، وسط: قالب MEI که توسط [۱۹] ارائه‌شده است. پایین: قالب MHI . MEI مشخص می‌کند که حرکت کجا اتفاق می‌افتد درحالی‌که MHI توصیف می‌کند که تصویر چگونه حرکت می‌کند. قالب‌های انتهایی رفتار مربوطه به‌منظور ارائه و نمایش استفاده می‌شود.

قالب‌های MEI و MHI حاوی اطلاعات مفیدی در مورد محتوای ویدیوهاست. به‌طور مثال از گرادیان قالب MHI به‌منظور حذف کردن حرکات و انسدادها در پشت زمینه توسط [۲۰] مورداستفاده قرارگرفته است. این مطالعه با تعیین نواحی کلیدی حرکت در MHI با بهره‌گیری از روش شناخت نقاط هریس[۵] [۲۱] به دست می‌آید.

[۵] Harris interest point detector.

قالب MEI حجمی[۶] توسط [۲۲] معرفی شد. ایده اصلی آن درواقع نمایش یک رفتار توسط یک شکل 3D است که از محیط‌های مرئی زمان- مکانی توصیف‌شده است. به‌منظور طبقه‌بندی، نتایج 3D، توسط محاسبه متوسط زمانی که هر نقطه نیاز دارد تا به مرز دسترسی داشته باشد به‌دست‌آمده است، به نقشه 2D تبدیل می‌گردند، که در تصویر ۲-۵ قابل‌مشاهده است. مقاله [۲۳] روشی را برای شناخت اعمال ارائه کرده است که بر مبنای اختلاف مشخصات STV[۷] است. یک STV توسط انباره سازی[۸] از کانتورهای[۹] شیء در طول محور زمان ساخته می‌شود. تغییر در سرعت، راستا و شکل یک STV مشخصه یک رفتار را تعیین می‌کند.

*تصویر ۲-۵ – **چپ:** حجم‌های مکان – زمان که در [۲۲] به‌منظور توصیف رفتار مورداستفاده قرارگرفته است. نمایش 3D با محاسبه متوسط زمان دسترسی یک نقطه به مرز، به نقشه 2D تبدیل می‌شود. **راست:** سطوح زمان – مکان در [۲۳] مربوط به سرویس در تنیس و دویدن نمایش داده‌شده است. سطح هندسی (اوج‌ها و دره‌ها) به‌منظور مشخصه یک رفتار مورداستفاده قرار می‌گیرد.*

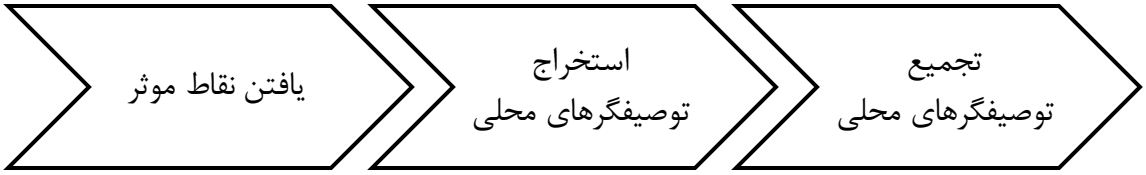
یک ایده دیگر توسط [۲۴] ارائه شد که لاپلاسین فیلترهای گوسی 3D برای ساخت فضای رفتاری مورداستفاده قرار می‌گیرد. هر دو روش معرفی‌شده به‌منظور مقاوم‌سازی در برابر حوزه‌های زمان – مکان از ساختار هرمی استفاده می‌کنند.

---

[۶] volumetric MEI
[۷] Space-Time Volume
[۸] stacking
[۹] contours

۲-۳- نمایش محلی

نمایش محلی برای شناسایی رفتار تحت عنوان یک نتیجه ابتدایی از مقاله [۲۵] بر روی STIP[۱۰] پدیدار شد. در حوزه تصاویر، نمایش محلی برای شناسایی رفتار پروسه زیر را دنبال می‌کند.



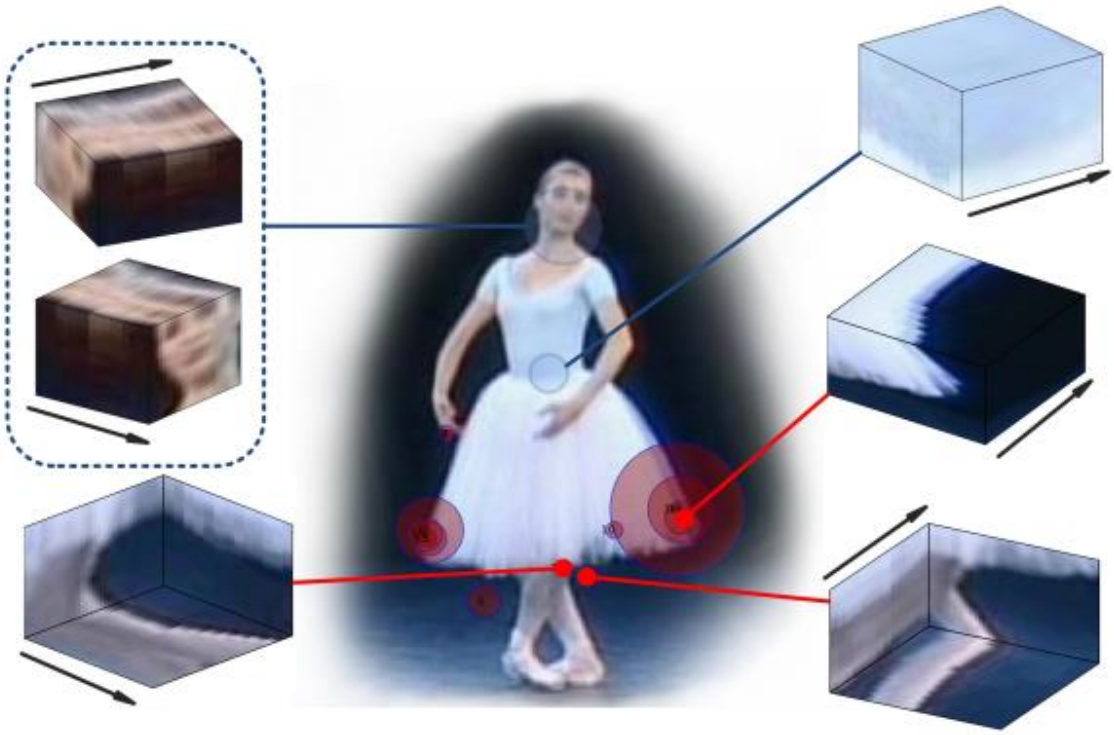
در ادامه به بررسی ایده‌ها و پیشرفت‌های شگفت‌آور در موارد فوق‌الذکر به‌صورت جداگانه می‌پردازیم.

۲-۳-۱- یافتن نقاط مؤثر

برای ساخت یک STIP Detector ، در مقاله [۲۵] روش گوشه یاب هریس[۱۱] [۲۱] گسترش داده‌شده است و روش جدید 3D-Harris Detector را ارائه داده است. در این روش علاوه بر اینکه باید به ساختار مکانی دسترسی داشت، ساختار زمانی نیز برای به‌کارگیری این روش موردنیاز است. ایده اصلی گوشه یاب هریس 2D ، یافتن موقعیت‌های مکانی در یک تصویر با تغییرات اساسی در راستای عمود بر هم است. هریس 3D نقاطی را معرفی می‌کند که تغییرات مکانی مشخص و حرکات غیرثابت داشته باشند. به‌عنوان نمونه تصویر ۲-۶ را می‌توان مثال زد.

[۱۰] Space-Time Interest Point

[۱۱] Harris corner detector



تصویر ۲-۶ – نقاطی که بارنگ قرمز مشخص‌شده است نقاط موردنظر زمان- مکان است که توسط مقاله [۲۵] در قسمت انتهایی تصویر مشخص‌شده است.تغییرات مکانی در انباره فریم‌ها در طول محور زمان با پیکان مشخص‌شده است. همان‌طور که مشخص است به دلیل نداشتن تعداد مناسب از ویژگی‌های مکانی، هیچ نقطه زمان – مکان در قسمت صورت و همچنین کمر یافت نشده است.

برخلاف تصاویر، کلیپ‌های دارای رفتار و عمل مشخص در بیشتر اوقات در محیط غیرقابل‌کنترل اخذ می‌شوند. به همین دلیل می‌بایست برای پردازش ویدیو دقت بیشتری کرد چون ممکن است ویژگی‌های خوب در محیطی با جزئیات زیاد گیر افتاده باشد. افزون بر ویژگی‌های زمان- مکان که از پس‌زمینه به‌دست‌آمده است، که به‌عنوان ویژگی‌های ثابت شناخته می‌شود، ویژگی‌هایی که در کنار نواحی وقوع حرکات هستند برای شناسایی اعمال بسیار کاراست. پس‌زمینه‌ها در انواع خاصی از ویدیوها (مانند ویدیوهای بازی فوتبال) می‌توانند اطلاعات مفیدی را برای شناسایی رفتار مورد نظر در اختیار قرار می‌دهد.

۲-۳-۲- توصیف محلی

ابتدا با یک تعریف ساده شروع می‌کنیم. یک مکعب 3D یا یک مکعب ساده، مکعبی است که از پیکسل‌های اطراف نقاط موردنظر ساخته‌شده است. به‌منظور توصیف محلی در یک نقطه موردنظر همه‌ی مقالاتی که اخیراً منتشرشده است به‌اتفاق آرا مکعب‌ها را برگزیدند [۲۶]،[۲۵].

در ادامه توصیف گرهای محلی مختلف که به‌طور گسترده در شناسایی اعمال و رفتار کاربرد دارند، بررسی می‌شوند، در ضمن یادآوری می‌شود که توصیفگرهای محلی از مکعب‌ها و مسیرها بهره می‌برند.

۲-۳-۲-۱- توصیفگرهای حرکت و لبه

مقاله [۲۷] پیشنهاد داده است که از هیستوگرام گرادیان جهت‌دار به‌عنوان توصیفگر حرکت استفاده شود. این مطالعه با الهام گرفتن از هیستوگرام گرادیان جهت‌دار، به‌منظور مقاوم‌سازی در شناسایی تصویر، این توصیفگر را در حوزه مکان-زمان به کار می‌برد که بانام توصیفگر HoG3D شناخته می‌شود.

OpticalFlow حرکت پیکسل در یک کلیپ ویدئویی را کد می‌کند. با بهره‌گیری از این خاصیت، مقاله [۲۸] هیستوگرام HoF[۱۲] را در نواحی محلی به‌عنوان یک توصیفگر زمان-مکان ارائه داده است.همچنین مقاله [۲۹] با استفاده از هیستوگرام‌ها و اعمال آن‌ها بر روی مکعب‌های محلی، احتمال OpticalFlow را به دست می‌آورد. توصیفگری مقاوم‌تر از HoF در [۳۰] بانام MBH[۱۳] معرفی‌شده است. MBH در حوزه Motion Boundary است که از محاسبه مشتق مکانی به دست می‌آید. تصویر ۲-۷ مثالی از این روش را نمایش داده است. همچنین در مقاله‌های [۳۱] و [۳۲] به ترتیب هیستوگرام‌های تحت عنوان HOS[۱۴] و MHOF[۱۵] به‌منظور مدل کردن دینامیک ازدحام و شناسایی رفتارهای نامتعارف ارائه‌شده

[۱۲] Histogram of Optical-flow

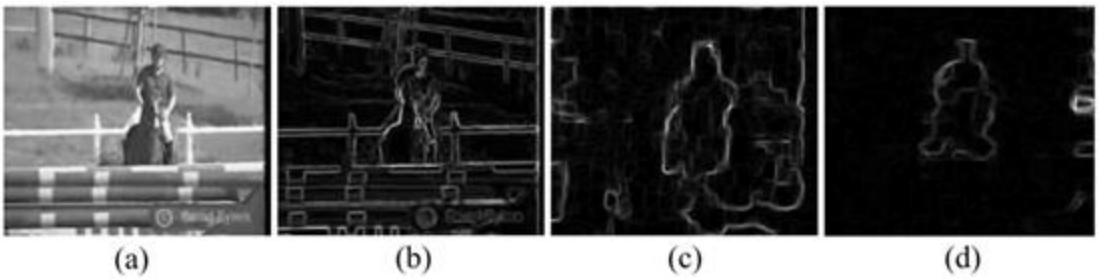
[۱۳] Motion Boundary Histogram

[۱۴] Histogram of Oriented Swarms

[۱۵] Multiscale Histogram of Optical Flow

است. روش ارائه‌شده در [۳۳] به‌منظور مدل کردن OpticalFlow به‌صورت محلی از یک ترکیب‌کننده‌ی احتمالاتی مؤلفه اصلی بهره برده است و برای ثبات بیشتر از روش تصادفی مارکوف استفاده کرده است.

روش‌های مبتنی بر بینایی یکی از تکنیک‌های یادگیری ماشین است که در مسائل مربوط به بینایی موفقیت‌هایی را توانسته به دست آورد. این روش‌ها در شناسایی اعمال ناهنجار محلی اعمال می‌شوند. این‌گونه متدها با استخراج ویژگی‌های بصری و ساختن یک مجموعه‌ای از خوشه بندها سعی دارند یک الگویی از رویدادها، به دست آورند [۳۴].



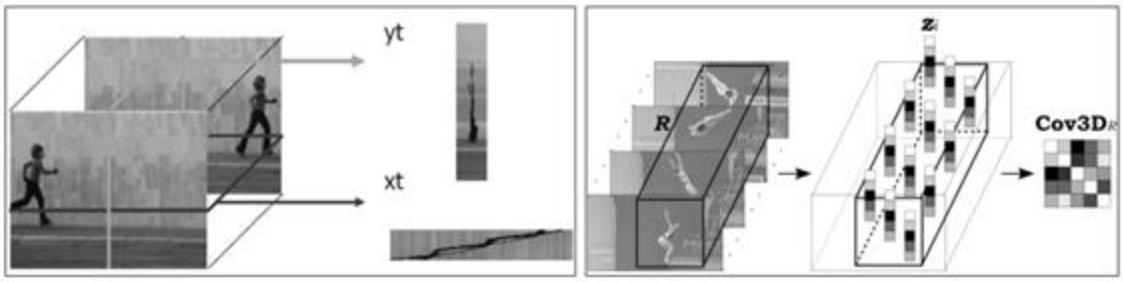
تصویر ۲-۷ – تصویر گرادیان مکانی (b) ، گرادیان افقی (c) و گرادیان عمودی (d) motion boundary مربوط به سوارکاری اسب(a)

۲-۳-۲-۲- توصیف‌کننده‌های الگوی باینری

الگوهای باینری محلی LBP[۱۶] یک توصیفگر 2D مبتنی بر شدت مقدار پیکسل است که در حوزه‌های گوناگون شامل شناسایی چهره و آنالیز بافت با موفقیت مورداستفاده قرارگرفته است [۳۵]. توصیفگر LBP با کوانتیزه کردن همسایگان یک پیکسل با توجه به‌شدت روشنایی آن‌ها محاسبه می‌شود. در مقاله [۳۶] توصیفگرهای مختلف LBP در حوزه زمان-مکان معرفی‌شده‌اند. در LBP حجمی VLBP[۱۷]، حجم‌های محلی توسط هیستوگرام الگوهای باینری کد می‌شوند [۳۶]. بااین‌وجود تعداد الگوهای مجزای تولیدشده توسط (VLBP) ممکن است در همسایگان زیاد بار زیادی را تحمل کند. برای کاهش این فشار در هیستوگرام الگوی باینری محلی از سه صفحه عمود بر هم (LBP-TOP)، که از الحاق الگوهای

[۱۶] Local binary patterns
[۱۷] Volume LBP

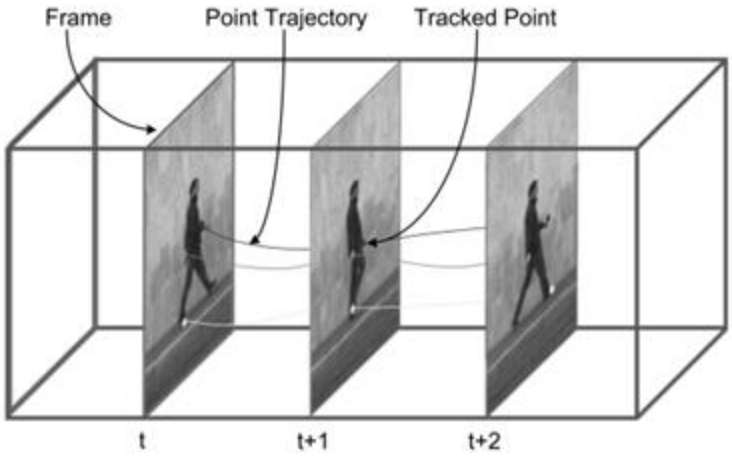
باینری محلی بانام‌های xy,xt,yt به دست می‌آید، استفاده می‌کند. نحوه استخراج این صفحه‌ها در تصویر ۲-۸ نمایش داده‌شده است.



*تصویر ۲-۸ – **چپ:** نحوه استخراج صفحه‌ها برای شناسایی رفتار که از توصیفگر LBP-TOP ارائه‌شده در [۳۶] الهام گرفته‌شده است. **راست:** توصیفگر کوواریانس زمان-مکان ارائه‌شده در [۳۷]*

۲-۳-۲-۳- توصیف‌کننده‌های مبتنی بر مسیر

نقاط موردنظر مکان-زمانی ممکن است دقیقاً در موقعیت مکانی مشابه در مکعب گسترده شده بازمان قرار نگیرد. ازاین‌رو ویژگی‌های استخراج‌شده از مکعب‌ها نمی‌توانند توصیفی از نقاط مؤثر واقعی داشته باشند. مسیر، یک خصیصه است که در طول زمان یک ویژگی را ردگیری می‌کند. تصویر ۲-۹ مسیرهایی که نقاط ویژگی را ردگیری می‌کنند نشان می‌دهد.



تصویر ۲-۹ – مسیرهایی که نقاط ویژگی را در طول فریم‌ها ردگیری می‌کنند.

همان‌طور که در [۳۸] اشاره‌شده است حرکات مرتبط بین مسیرها ،که به‌عنوان‌مثال در راستا، اندازه و موقعیت تفاوت دارند، می‌توانند طبقه‌بندی اعمال و رفتار را با قاطعیت مشخص کند به‌خصوص دسته‌هایی که شامل رفتارهای انسانی یا کنش انسانی (مانند تکان دادن دست) باشد.

در مقاله [۳۹] از لاگرانژ مسیر ذراتی که بر مبنای OpticalFlow است، برای مدل کردن صحنه های ازدحام استفاده کرده است و به محاسبه یک مجموعه ویژگی تغییر ناپذیر با بی نظمی (حداکثر مقدار لیاپانوف) برای شناسایی ناهنجاری ها میپردازد.

۲-۳-۳- تجمیع

با فرض $V = \{v_i\}_{i=1}^{n}, v_i \in R^d$ که یک مجموعه از ویژگی‌های محلی استخراج‌شده از یک ویدیو است، برای شناسایی اعمال نیازمند مکانیزمی برای آموزش مجموعه‌ها و به طبع مقایسه آن‌ها هستیم.

الگوریتم‌های یادگیری مانند ماشین بردار پشتیبان SVM[۱۸] فقط در مواقعی که سایز بردارها ثابت باشد کارایی لازم را دارد؛ اما در مواقعی که سایز مجموعه‌ها متغیر است نمی‌توان از آن استفاده کرد. بدین منظور برای بهره‌مندی از تکنیک‌های یادگیری مختلف به یک مکانیزم برای تجمیع مجموعه ویژگی‌های محلی به توصیفگرهای قابل‌تمایز و با سایز ثابت نیاز داریم. برای این کار انتخاب ماشینی بر اساس BoV[۱۹] که در [۴۰] ارائه گردیده است، بهترین انتخاب است. در BoV ، هیستوگرام وقوع visual word به‌عنوان توصیفگر مورداستفاده قرار می‌گیرد. فرکانس نمایش هر visual word ، به‌عنوان نزدیک‌ترین همتا به ویژگی‌های مجموعه موردنظر، توصیفگر را تعیین می‌کند.

مقاله [۲۶] جز اولین مطالعات انجام‌شده است که از BoV برای شناسایی اعمال استفاده کرده است. در فرم اصلی آن، اطلاعات زمانی توسط BoV نادیده گرفته‌شده است. برای اصلاح این نقطه‌ضعف در مقاله [۲۸] تکه‌های مکان-زمان ارائه‌شده است. ایده اصلی این است که یک ویدیو به چندین زیر ویدیو تقسیم

[۱۸] Support Vector Machines
[۱۹] Bag of Visual Word

می‌شود. توصیفگرهای محلی در هر زیر ویدیو تجمیع می‌گردند که کانال‌ها را تشکیل دهد و ویدیوها بر مبنای کانال توصیفگرها مقایسه شوند. اخیراً تجمیع توسط روش FV[۲۰] در [۴۱] ، [۴۲] برای کد کردن انتخاب می‌شود. روش کد کردن FV در یک متد تجمیع مبتنی بر قاعده Fisher Kernels [۴۳] است که از ترکیب با روش متمایزکننده و مولد برای طبقه‌بندی الگو بهره برده است. به‌طور خلاصه می‌توان تفاوت‌های روش BoV و FV را در موارد زیر بیان کرد.

۱. در BoV عملیات تجمیع دشوارتر از روش FV است.

۲. اگر فرض شود مدل تولید ویژگی‌ها به‌صورت مدل GMM (Gaussian Mixture Model) باشد؛ BoV تنها اطلاعات درجه صفر (اطلاعات مربوط به وقوع) را تجمیع می‌کند درحالی‌که FV از هر دو اطلاعات درجه اول و درجه دوم استفاده می‌کند.

از کد کردن FV در طول مسیرها به‌منظور افزایش کارایی روش ارائه‌شده، استفاده‌شده است.

Stacked FVs را می‌توان روش بسط داده‌شده‌ی تکه‌های زمان-مکانی [۲۸] دانست که با عنوان FVs در [۴۲] معرفی‌شده است. جزئیات مربوط به آنالیز FVs در شناسایی اعمال و رفتار در مقاله [۴۴] ارائه گردیده است. FVs معمولا با ابعاد بالا همراه است. نسخه ساده‌شده‌ای از FV بانام VLAD[۲۱] در [۴۵] ارائه‌شده است که اطلاعات مرتبه دوم را از توصیفگر حذف می‌کند. درنتیجه ابعاد VLAD به نصف ابعاد FV کاهش پیدا می‌کند.

[۲۰] Fisher Vector

[۲۱] Vector of Locally Aggregate Description

۲-۴- معماری‌های عمیق برای شناسایی رفتار

شبکه‌های عصبی عمیق[۲۲] مانند شبکه‌های عصبی کانولوشنی[۲۳] در [۴۶] به‌عنوان متد انتخابی در آموزش محتوای تصویر در [۴۷] ، [۴۸] ، [۴۹] و [۵۰] مورداستفاده قرارگرفته است. ساختارهای عمیق[۲۴] توسط ترکیب سطوح مختلفی از عملگرهای غیرخطی به وجود می‌آید. یافتن فضای پارامتر ساختارهای عمیق در تصمیم‌گیری‌های غیر محدب کار آسانی نیست.

قصد ما از مطرح کردن این قسمت بررسی مدل‌های استفاده‌شده برای یادگیری اعمال و رفتار توسط ویدیوهاست. می‌توان معماری‌هایی که برای شناسایی رفتار ارائه‌شده‌اند را در دسته های زیر قرارداد.

- شبکه‌های زمان-مکانی
- شبکه‌های چند جریانه
- شبکه‌های مولد عمیق
- شبکه‌های وابسته زمانی

در ادامه به بررسی هر یک از این قسمت‌ها خواهیم پرداخت و اشاره‌ای نیز به پیشرفت‌های ممکن خواهیم کرد.

[۲۲] Deep Neural Networks
[۲۳] Convolutional Neural Networks
[۲۴] Deep Architectures

۲-۴-۱- شبکه‌های زمان–مکانی[۲۵]

معماری کانولوشنی با بهره‌گیری مؤثر از ساختار تصویر سعی در کاهش فضای جستجو توسط Pooling و Weight-Sharing در شبکه را دارد. در تصویر ۲-۱۰ قسمت چپ دیاگرام آن نمایش داده‌شده است.

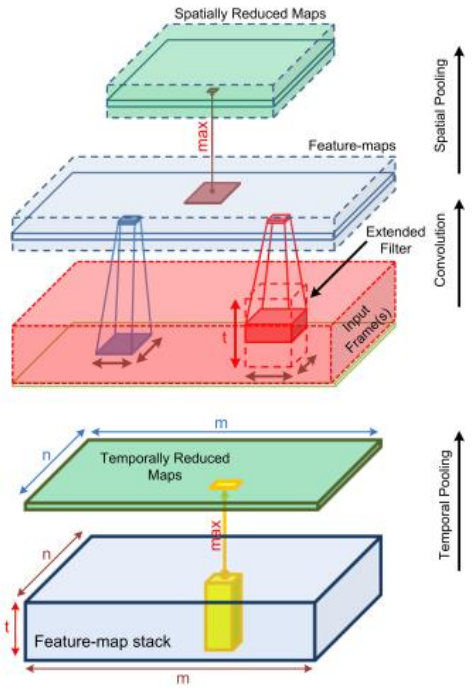
Pooling و Weight-sharing کمک می‌کند که متد مربوطه در برابر تغییرات بزرگنمایی [۲۶] و مکانی مقاوم باشد. فیلترهای آنالیز که توسط معماری‌های CNN آموزش‌دیده می‌شوند نشان می‌دهد که لایه‌های اولیه ویژگی‌های سطح پایین را آموزش می‌بینند درحالی‌که لایه‌های بالایی سطوح معنایی بالایی را آموزش می‌بینند. همین موضوع باعث شده است که از شبکه‌های کانولوشنی به‌عنوان استخراج‌کننده‌های ویژگی استفاده شود.

یک روش مستقیم برای شناسایی رفتار استفاده از شبکه‌های عمیق است که به عملگرهای کانوالوشن به همراه اطلاعات زمانی مسلح است. بدین منظور [۲۷]شبکه‌های کانولوشنی 3D در [۵۱] معرفی‌شده است. یک شبکه کانولوشنی 3D همان‌طور که از نام آن مشخص است، از کرنل‌های سه‌بعدی (فیلترهایی که در طول محور زمان گسترش پیدا کردند) به‌منظور استخراج ویژگی‌ها در ابعاد زمانی و مکانی استفاده می‌کند؛ ازاین‌رو انتظار می‌رود اطلاعات زمان-مکانی و حرکات کد شده‌ی فریم‌های مجاور به دست آید. تصویر ۲-۱۰ دیاگرام مفهومی آن را نشان می‌دهد.

[۲۵] Spatiotemporal Networks

[۲۶] Scale

[۲۷] 3D convolution network



Layer	Parameters
Input	(60 x 40)-7 Frames Grayscale-7 Ch.
Hardwired	Gradient-7x2 Ch. Opt. Flow-6x2 Ch.
Conv 1 - ReLU	2 Groups - (7x7x3)
Pool 1	(2x2) - Spatial
Conv 2 - ReLU	3 Groups- (7x6x3)
Pool 2	(3x3) - Spatial
Conv 3 - ReLU	(7x4)
FC	No. of Classes

تصویر ۲-۱۰ – عملگرهای زمان-مکانی: کانولوشن 2D (آبی)، کانولوشن 3D در [۵۱] (قرمز)، کانولوشن مکانی -max pooling (قهوه‌ای) ، کانولوشن زمانی max-pooling (زرد) .

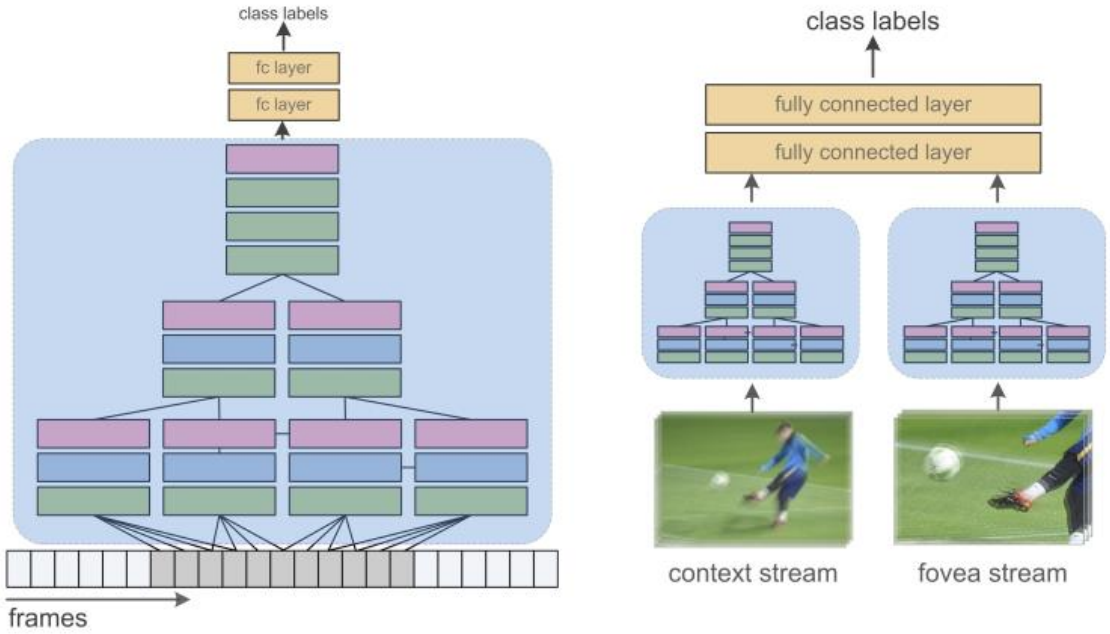
باید توجه داشت که شبکه‌های کانولوشنی 3D دارای ساختار زمانی بسیار مستحکمی است. شبکه تعدادی از فریم‌های از پیش تعریف‌شده را به‌عنوان ورودی می‌پذیرد (برای مثال در [۵۱] ورودی شامل فقط ۷ فریم است.). درحالی‌که داشتن ابعاد مکانی ثابت به نحوی قابل دفاع است ؛ این نکته مبهمی است که چرا این فرضیات مشابه را در برابر حوزه زمانی نیز اعمال می‌کنند. گرچه انتخاب طول زمان برای حرکت‌های طولانی[۲۸] در رفتارهای گوناگون که دارای سرعت مختلفی هستند و به طبع دارای متفاوتی نیز هستند، در ابهام و وضعیت نامشخصی است.

برای پاسخ به اینکه چگونه شبکه‌های کانولوشنی توسط اطلاعات مکانی تغذیه می‌شوند طرح‌های متفاوتی از ادغام[۲۹] موردبررسی قرار می‌گیرد. در مقاله [۵۲] "pooling زمانی"، ارائه‌شده است و به این نتیجه رسیده است که گرفتن بیشینه از pooling در حوزه زمانی بسیار مؤثر است. در [۵۳] مفهوم ادغام

[۲۸] macro motion

[۲۹] fusion

آهسته [۳۰] ارائه شده است که باعث افزایش آگاهی زمانی یک شبکه کانولوشنی می‌شود. در ادغام آهسته، یک شبکه کانولوشنی، تعدادی از قسمت‌های یک ویدیو، که متوالی نیز هستند، را می‌پذیرد و به‌منظور تولید پاسخ‌هایی در حوزه زمانی، آن‌ها را با مجموعه لایه‌های مشابه پردازش می‌کند. این پاسخ‌ها توسط لایه‌های کاملاً متصل [۳۱] پردازش می‌شوند که بتوانند توصیفگر ویدئویی را تولید کنند. تصویر ۲-۱۱جزئیات را نمایش می‌دهد.



تصویر ۲-۱۱ – معماری ارائه‌شده در [۵۳] . رنگ سبز، قرمز و آبی به ترتیب مربوط به لایه‌های نرمال‌سازی ، Spatial-Pooling و کانولوشن است.

فرم‌های دیگری از ادغام شامل ادغام اولیه [۳۲] (مانند شبکه کانولوشنی 3D در [۵۱]) شبکه را با یک مجموعه از فریم‌های مجاور تغذیه می‌کند و ادغام اخیر [۳۳] ویژگی‌های فریم گونه را در لایه‌های کاملاً متصل ادغام می‌کند [۵۳] . در این مقاله همچنین نشان داده‌شده است روش چند رزولوشن [۳۴] که از دو شبکه جدا استفاده می‌کند نه‌تنها دقت را افزایش می‌دهد بلکه باعث می‌شود که تعداد پارامترهایی که

[۳۰] slow fusion
[۳۱] Fully Connected Layers
[۳۲] early fusion
[۳۳] late fusion
[۳۴] multi-resolutional

آموزش می‌بینند نیز کاهش یابد. نتیجه حاصله بستگی به این دارد که هر یک از پایه‌های شبکه، ورودی‌های کوچک‌تری را بپذیرد. (به‌عنوان‌مثال می‌تواند به استریم Foeva و Contex در تصویر فوق اشاره کرد). استریم Fovea نواحی مرکزی یک فریم را تا زمانی که شیء موردنظر، این ناحیه را اشغال کرده باشد دریافت می‌کند.

به‌منظور بهره‌برداری از اطلاعات زمانی، برخی از مطالعات به استفاده از ساختارهای برگشت‌پذیر[۳۵] متوسل شده‌اند. مانند مقالات [۵۴] و [۵۵] که با استفاده از شبکه‌های کانولوشنی آبشاری و یک کلاس شبکه [56] RNN[۳۶] که تحت عنوان شبکه‌های LSTM[۳۷] در [۵۷] شناخته می‌شود به مسئله شناسایی رفتار می‌پردازد. از کلمه بازگشت‌پذیر[۳۸] می‌توان دریافت، در یک RNN (در تصویر ۲-۱۲ نمایش است.) از مدل‌های دینامیکی که حلقه برگشتی[۳۹] دارد، استفاده می‌کند. یک بلاک RNN یک سیگنال خارجی مانند $x^{(t)} \in R^n$ را می‌پذیرد و یک خروجی مانند $z^{(t)} \in R^m$ بر مبنای حالت مخفی $h^{(t)} \in R^r$ توسط روابط زیر تولید می‌کند.

$$h^{(t)} = \sigma(W_x x^{(t)} + W_h h^{(t-1)}) \qquad (۲-۲)$$

$$z^{(t)} = \sigma(W_z h^{(t)}) \qquad (۲-۳)$$

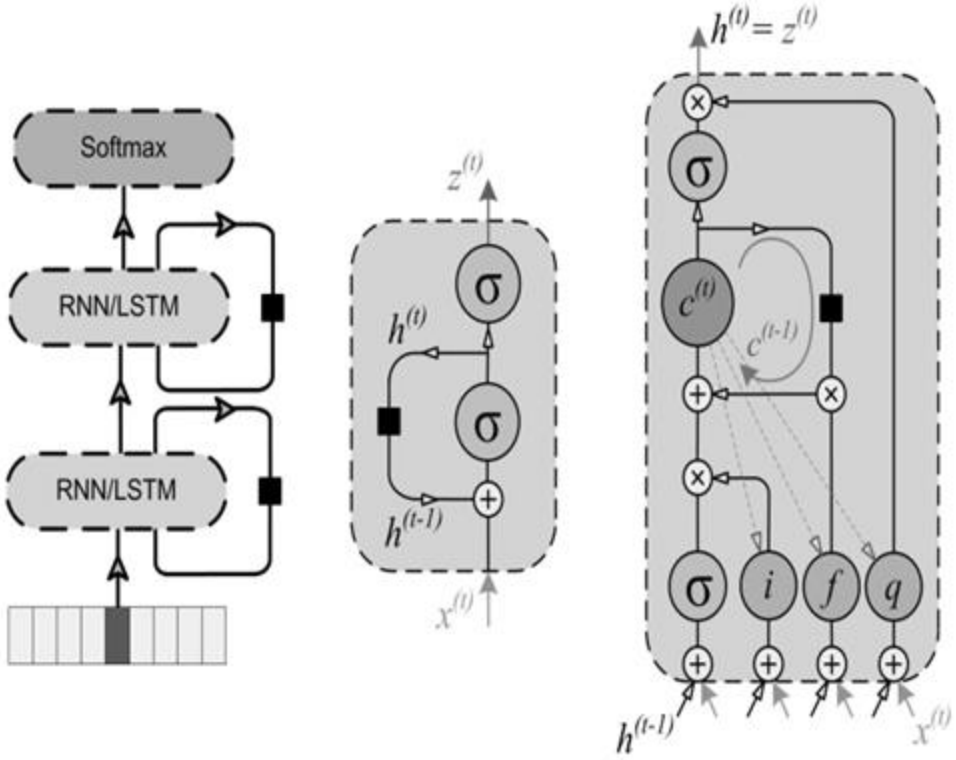
که $W_x \in R^{r\times m}$ ، $W_h \in R^{r\times r}$ و $W_z \in R^{m\times r}$ است. بدیهی است که یک RNN ، تحققی از سیستم دینامیکی خطی LDS[۴۰] است و برای مدل کردن سکانس های ویدیویی کافی به نظر می‌رسد.

برای بررسی بیشتر فرض می‌شود که رابطه بازگشتی یک سلول RNN به فرم $h^{(t)} = w_h h^{(t-1)}$ که $x, h, z \in R$ است. این فرم بازگشتی میتواند به صورت $w_h w_h h^{(t-2)} = w_h^t h^{(0)}$ نوشته شود. به همین

[۳۵] recurrent structures
[۳۶] Recurrent Neural Networks
[۳۷] Long-Short Term Memory
[۳۸] Recurrent
[۳۹] feedback loop
[۴۰] Linear Dynamics Systems

منظور شبکه را زمانی که نتایج مطلوبی حاصل نشود به شرط $W_h < 1$ با ترم های کوتاه و به شرط $W_h > 1$ با ترم های بلند آموزش میدهد.

سلول‌های LSTM که در تصویر ۲-۱۲ نمایش داده‌شده است مشکل محدودسازی حالت‌ها و خروجی‌های سلول RNN همراه با کنترل گیت‌ها را مرتفع می‌کند.

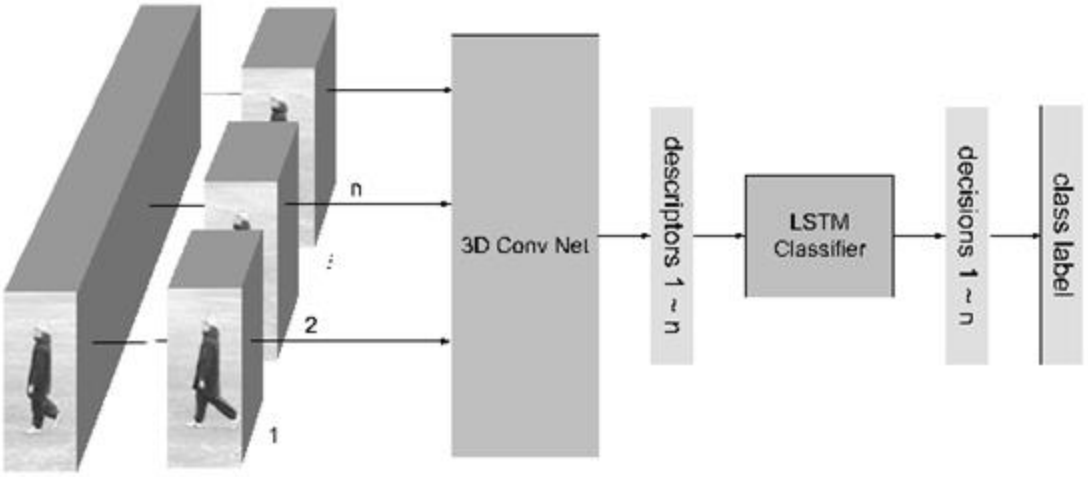


*تصویر ۲-۱۲ – **چپ:** ساختار بازگشت‌پذیر یک شبکه ۲ لایه RNN / LSTM مرکز: ساختار یک سلول RNN که دارای سیستم دینامیکی خطی است. **راست:** سلول LSTM که شامل گیت‌های کنترلی است. تأخیر زمانی با مربع سیاه رنگ مشخص‌شده است.*

برای طبقه‌بندی اعمال در مقاله [۵۴] پیشنهادشده است که یک شبکه LSTM توسط ویژگی‌های استخراج‌شده از یک شبکه کانولوشنی 3D تغذیه می‌شود. دو شبکه کانولوشنی 3D و شبکه LSTM به‌صورت جداگانه آموزش می‌بینند، در ابتدا شبکه کانولوشنی 3D توسط رفتار مورد نظر آموزش می‌بینند و در ادامه که شبکه کانولوشنی 3D به دست آمد از ویژگی‌های کانوالو شده برای آموزش شبکه LSTM استفاده می‌شود. تصویر ۲-۱۳و جدول ۲-۱ جزئیات نتایج پیاده‌سازی را نشان می‌دهد.

جدول ۲-۱- پارامترهای شبکه ارائه‌شده در [۵۴]

	Layer	Paremeters
Colvolutional	Conv 1 - ReLU	7 Filters - (7x7x5)
	Pool 1	(2x2) - Spatial
	Conv 2 - ReLU	35 Filters - (5x5x3)
	Pool 2	(2x2) - Spatial
	Conv 3 - Linear	35 Filters - (5x5x3)
	FC 1 (included only in training)	-
	FC 2 (included only in training)	No. of Classes
RNN	LSTM	50
	FC	No. of Classes

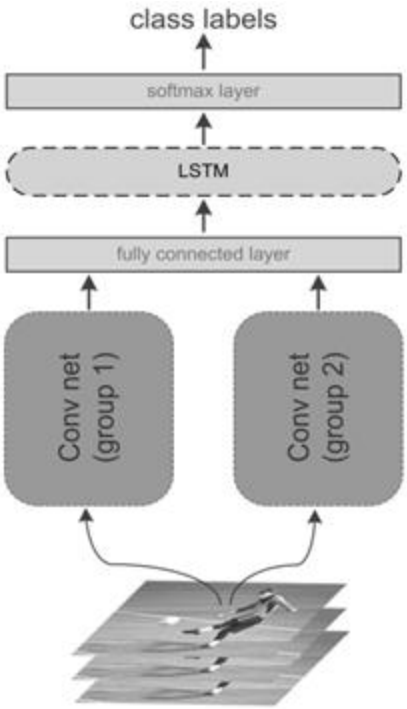


تصویر ۲-۱۳ – ساختار شبکه ارائه‌شده در [۵۴].

ساختار دیگری مبتنی بر LSTM در مقاله [۵۵] ارائه‌شده است که از آموزش End-to-End در شبکه ترکیبی بهره می‌برد. در تصویر ۲-۱۴ ساختار و در جدول ۲-۲ پارامترهای شبکه LSTM نمایش داده‌شده است.

جدول ۲-۲ – پارامترهای *LRCN*

	لایه	پارامترها
Conv	Conv 1 - ReLU	96 Filters - (7x7)
	Pool 1	(3x3) - Spatial
	LRN 1	channels - 5
	Conv 2 - ReLU	384 Filters (2 Groups) - (5x5)
	Pool 2	(3x3) - Spatial
	LRN 2	channels - 5
	Conv 3 - ReLU	512 Filters - (3x3)
	Conv 4 - ReLU	512 Filters (2 Groups)- (3x3)
	Conv 5 - ReLU	384 Filters (2 Groups) - (3x3)
	Pool 5	(3x3) - Spatial
	FC 6 - ReLU	4096
RNN	LSTM	256
	FC	No. of Classes
	Softmax	No. of Classes



تصویر ۲-۱۴ – ساختار شبکه *LRCN* ارائه‌شده در[۵۵].

نتایج ساختار بانام LRCN[۴۱] نشان می‌دهد که نه‌تنها در مورد شناسایی رفتارها بلکه در مورد توضیحات تصاویر و ویدیوها نیز موفق عمل کرده است. با آموزش End-to-End و کانولوشن CNN-LSTM پارامترهای قابل‌قبول فیلتر زمان-مکان محاسبه می‌شود.

۲-۴-۲- شبکه‌های چند جریانه[۴۲]

در مقاله [۵۸] یکی از اولین شبکه‌های عصبی عمیق چند جریانه معرفی شد که ساختار دو شبکه موازی از VGG_16 در [۴۸] برای شناسایی رفتار انتخاب شده است. جزئیات در تصویر ۲-۱۵ نمایش

[۴۱] Long-term Recurrent Convolutional Network

[۴۲] Multiple-stream

داده‌شده است. شبکه جریان مکانی[۴۳] فریم‌های ویدئویی خام را می‌پذیرد درحالی‌که شبکه جریان زمانی[۴۴]، Optical-flow را به‌عنوان ورودی دریافت می‌کند. مطالب اشاره‌شده در مقاله [۵۸] را می‌توان به‌صورت زیر موردبررسی قرارداد.

- آموزش از پیش انجام‌شده برای شبکه جریان مکانی

به‌طور تجربی یک شبکه که به‌طور کلی آموزش می‌بینند می‌تواند دقت بالاتری را به دست آورد.

- ادغام اولیه برای شبکه جریان زمانی

انباره کردن Optical-flow در ورودی شبکه جریان زمانی (مانند ادغام اولیه) بسیار سودمند است.

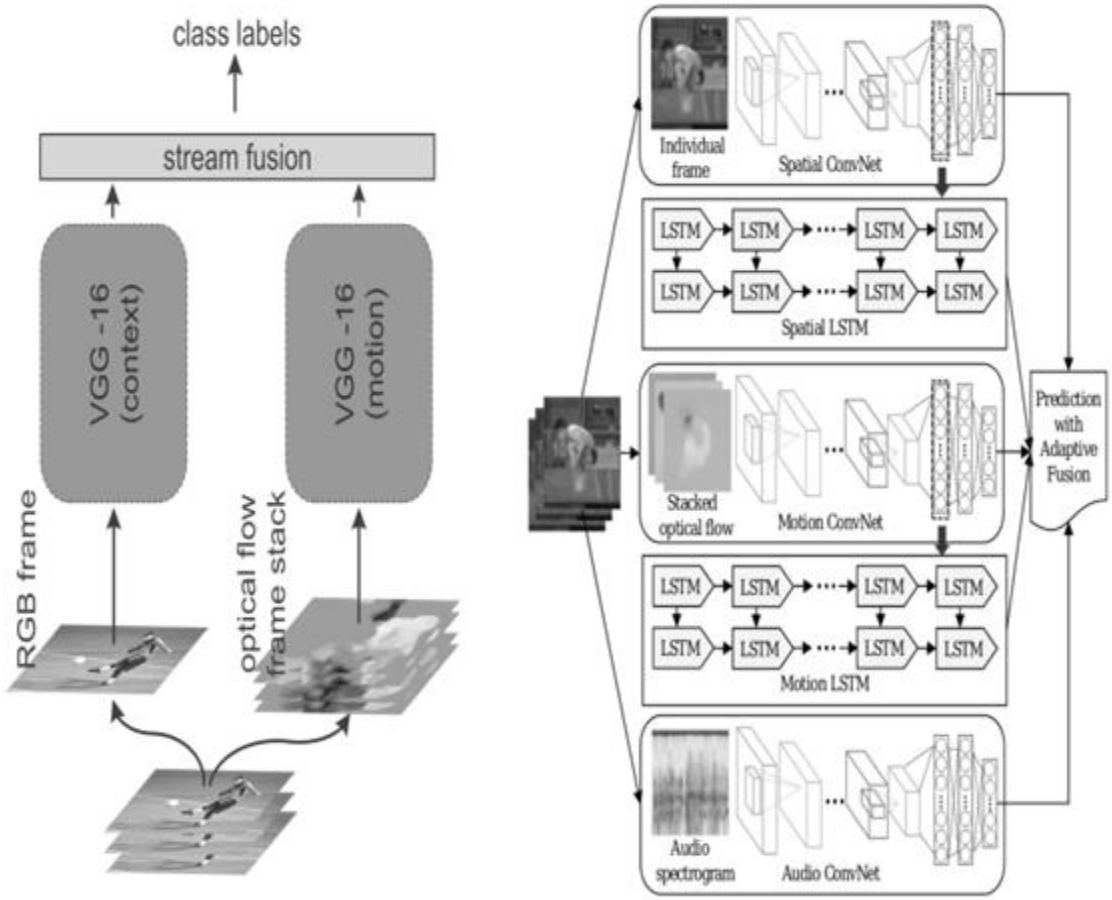
- آموزش Multi-task برای شبکه جریان زمانی

شبکه جریان زمانی می‌بایست به‌طور کامل و دقیق توسط داده‌های ویدئویی موجود آموزش ببینند. این مسئله، چالشی برای مجموعه داده‌های با سایز کوچک و یا متوسط در شبکه‌های عمیق است. برای برطرف نمودن این مشکل شبکه جریان زمانی با بیش از یک‌لایه طبقه‌بندی اصلاح می‌شود. هر لایه طبقه‌بندی بر روی مجموعه داده خاصی عمل می‌کند (یکی روی مجموعه داده HMDB-51 و دیگری بر روی UCF-101 عمل می‌کند [۵۸]) و تنها از همان مجموعه داده مربوطه پاسخش را دریافت می‌کند. این ساختار را می‌تواند ساختار آموزشی Multi-task که سعی در ارائه آموزش دارد ، تصور کرد؛ که نه‌تنها قادر به پاسخ‌گویی به یک وظیفه است بلکه به وظایف دیگر نیز رسیدگی می‌کند.

[۴۳] spatial stream

[۴۴] temporal stream

بسط شبکه دو استریم شامل مقالاتی ازجمله [۵۹] است، که از مسیرهای متراکم معرفی‌شده در [۶۰] ساخته می‌شود ؛ این مسیرها توسط ویژگی‌های کانوالو شده شبکه دو استریم با بهره‌گیری از بردار Fisher تجمیع شده است، و مقاله [۶۱] که استریم سومی را که از سیگنال صدا استفاده می‌کند به شبکه اضافه نموده است.



*تصویر ۲-۱۵ – **چپ:** شبکه دو استریم با رنگ و Optical-flow انباره‌ای ارائه‌شده در [۵۸] **راست:** ساختار multi-stream که توسط [۶۱] ارائه شد.*

۲-۴-۳- مدل‌های شبکه مولد عمیق[۴۵]

پتانسیل فوق‌العاده مدل‌های عمیق ابداع‌شده این است که با توجه به توسعه روزافزون ویدیوهای قابل‌دسترسی از طریق وب، کمترین نظارت را می‌توان برای این مدل‌ها متصور بود.

[۴۵] Deep Generative Network Models

یک مدل مولد مناسب، مدلی است که توزیع داده‌ها را بتواند فراگیرد. مدل‌های مولد برای آنالیز سکانس ([۴۹] و [۶۲]) به‌منظور پیش‌بینی ویژگی یک سکانس مورداستفاده قرار می‌گیرد. این موضوع با در اختیار قرار دادن یک سکانس به‌صورت ($x_1, x_2, x_3 \dots, x_t$) یک مدل را می‌آموزد که ویژگی را پیش‌بینی کند (مانند x_{t+1}). این کار با متدی که در بخش قبل بررسی شد، متفاوت است؛ به دلیل اینکه آن روش نیاز به برچسب برای آموزش ندارد.

به‌هرحال پیش‌بینی با دقت، زمانی به دست می‌آید که محتوا و دینامیک (حرکات اولیه) سکانس را بتوان توسط مدلی که به‌طور مناسب گسترده شده باشد، کسب کرد. ساختارهای مولد عمیق در مقاله‌های [۶۳]، [۶۴]و [۵۷] به دنبال همین هدف هستند؛ مانند یادگیری بدون ناظر از طریق داده‌های زمانی. در آنالیز ویدیو با داده‌هایی که تفسیر آن‌ها هزینه‌بر است، متدهای بدون ناظر به متدهای با ناظر ترجیح داده می‌شوند.

در این قسمت به بررسی مثال‌هایی از ساختارهای مولد عمیق که مستقیم در شناسایی رفتار اعمال‌شده است، می‌پردازیم.

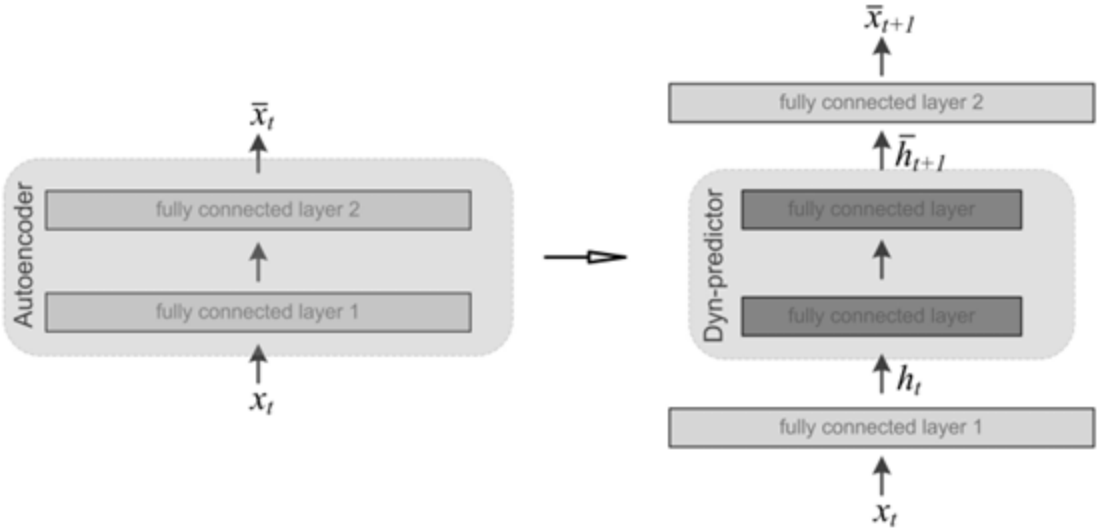
۲-۴-۳-۱- کد کننده یکسویه[۴۶]

با الهام از مدل‌سازی LDS ، مقاله [۶۵] کدکننده یکسویه را که در کلاس رمزگذار خودکار عمیق[۴۷] است، به‌منظور کسب دینامیک حاکم در ویدیو معرفی می‌کند. در اغلب موارد فرم ساده کدکننده یکسویه شامل ۳ لایه است. لایه اول ورودی x_t را به حالت‌های مخفی h_t نگاشت می‌کند. لایه دوم یک‌لایه پیش‌بینی کننده است که حالت‌های مخفی بعدی $\tilde{h}_{t+1}$ را با استفاده از حالت کنونی (مانند h_t) پیش‌بینی می‌کند. لایه آخر نگاشتی از حالت‌های مخفی $\tilde{h}_{t+1}$به فریم‌های تولیدی تخمین زده شده ورودی، $\tilde{x}_{t+1}$ است. به‌منظور کاهش پیچیدگی، پارامترهای شبکه در دو مرحله فراگرفته می‌شود. در مرحله پیش از

[۴۶] Dyn-Encoder

[۴۷] Deep Auto-encoder

آموزش هر لایه به‌صورت مجزا آموزش می‌بینند. بعد از تکمیل این مرحله، آموزش به هم پیوسته انجام می‌گیرد. یک ساختار مفهومی از کدکننده یکسویه در تصویر ۲-۱۶ نمایش داده‌شده است.



*تصویر ۲-۱۶ - ساختار و یادگیری Dynencoder که در [۶۵] ارائه شد. **چپ:** مرحله اول، لایه‌های کد کننده خودکار بیرونی آموزش می‌بینند که فریم ورودی x_t را بازسازی کنند. **راست:** نگاشت اولیه، h_t برای به دست آوردن حالت‌های مخفی ورودی اعمال می‌شود که در آموزش لایه‌های پیش‌بینی کننده دینامیک مورداستفاده قرار می‌گیرد. در پایان لایه کد کننده برای بازسازی اطلاعات فریم از حالت‌های مخفی آموزش می‌بینند، درواقع Dynencoder به‌عنوان یک پیش‌بینی کننده فریم عمل می‌کند.*

کدکننده یکسویه، نشان داد که در سنتز دینامیکی بافت‌های تصویری بسیار موفق است. این کدکننده را می‌تواند به‌عنوان یک‌راه میانبر برای ارائه اطلاعات مکان-زمانی در یک ویدیو دانست. همین‌طور خطای بازسازی یک ویدیو که توسط این کدکننده تولید می‌شود، می‌تواند به‌عنوان یک طبقه بند مورداستفاده قرار گیرد.

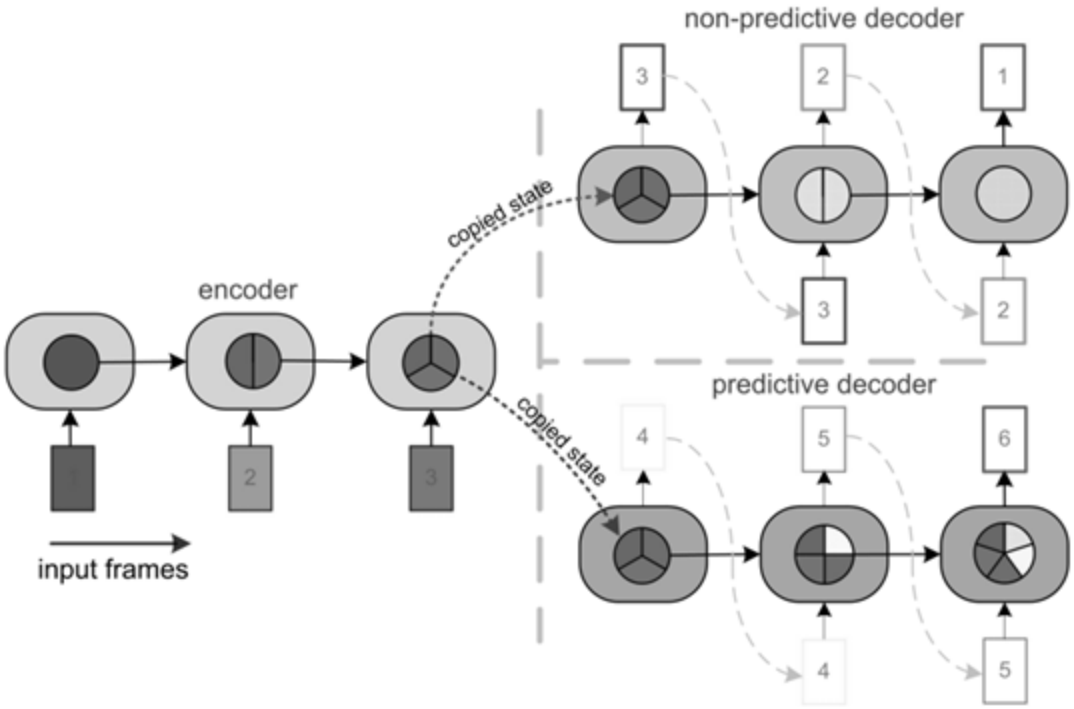
۲-۴-۳-۲- مدل کدکننده خودکار LSTM [۴۸]

مدل‌های مولد برای شناسایی اعمال انتظار دارد که ترم‌های بلند[۴۹] شناخته شوند و توسط مدل‌های عمیق که با سلول‌های LSTM همراه است، انتخاب شوند. در همین رابطه، مقاله [۶۲] مدل LSTM

[۴۸] Long Short Term Memory
[۴۹] long-term

Auto-Encoder را معرفی کرد که در تصویر ۲-۱۷ نمایش داده‌شده است. LSTM Auto-Encoder از دو RNN به نام‌های encoder LSTM و decoder LSTM تشکیل‌شده است.

بخش encoder LSTM یک سکانس را به‌عنوان ورودی دریافت می‌کند و صورت فشرده‌شده آن را یاد می‌گیرد. حالت‌های encoder LSTM شامل شکل و دینامیک سکانس است که فشرده‌شده یک سکانس به‌عنوان حالت‌های encoder LSTM انتخاب می‌شوند. بخش decoder LSTM خروجی آموزش‌دیده بخش قبل را به‌عنوان ورودی دریافت می‌کند که بتواند سکانس ورودی را بازسازی نماید. در تصویر ۲-۱۷ جزئیات را می‌توان مشاهده نمود. LSTM Auto-Encoder می‌تواند برای پیش‌بینی آینده یک سکانس مورداستفاده قرار گیرد.



تصویر ۲-۱۷ – مدل ترکیبی مولد LSTM که توسط مقاله [۶۲] ارائه‌شده است. حالت‌های داخلی (که با دایره در داخل نشان داده‌شده است) encoder LSTM یک قسمت از سکانس ورودی (مانند فریم‌های ۱و۲و۳) را دریافت می‌کنند. این حالات در دو مدل کد کننده که بازسازی و پیش‌بینی می‌کنند، کپی می‌شود. کد کننده باز ساز سعی دارد فریم‌های اصلی را به‌صورت معکوس بازسازی کند. مدل پیش‌بینی کننده به‌منظور پیش‌بینی فریم‌های آینده ۴و۵و۶ آموزش می‌بینند.

۲-۴-۳-۳- مدل‌های خصمانه [۵۰]

برای گریز از مشکلات آموزش مدل‌های مولد عمیق، مقاله [۶۴] شبکه‌های خصمانه را معرفی کرده است که مدل مولد با یک مدل متمایزکننده که بانام adversary شناخته می‌شود، رقابت می‌کند. مدل متمایزکننده یاد می‌گیرد تا تعیین کند که یک نمونه از مدل مولد آمده است یا از داده خودش. در حین آموزش، مدل مولد می‌آموزد نمونه‌هایی را تولید کند که بتواند تشابه بیشتری را با داده اصلی به اشتراک بگذارد تا بتواند در آزمون مدل adversary قبول شود درحالی‌که مدل adversary سعی دارد نوع قضاوت خود را، که با گرفتن یک نمونه و تعیین معتبر یا نامعتبر بودن آن انجام می‌پذیرد، بهبود بخشد.

۲-۴-۴- شبکه‌های وابسته زمانی [۵۱]

قبل از شروع به بررسی این قسمت، می‌خواهیم در مورد ایده و چشم‌انداز وابستگی زمانی بپردازیم. وابستگی زمانی یک فرم از ناظر ضعیف[۵۲] است و حالت‌های آن در فریم‌های متوالی ویدئویی در دو جنبه معنایی و دینامیکی به هم‌وابستگی دارند. یک سکانس، زمانی وابسته است که فریم‌هایش در نظم زمانی درستی قرار داشته باشند. وابستگی زمانی می‌تواند توسط مدل عمیق آموزش ببیند، چنانچه این مدل توسط سکانس‌های منظم و غیرمنظم به‌مانند نمونه‌های مثبت و منفی، به ترتیب، تغذیه شود.

در [۶۶] به بررسی این موضوع می‌پردازد که چگونه وابستگی زمانی می‌تواند در آموزش مدل‌های عمیق برای شناسایی رفتار و تخمین موقعیت[۵۳] مورداستفاده قرار گیرد. یک شبکه Siamese که در تصویر ۲-۱۸ نمایش داده‌شده است، توسط tuple ها آموزش می‌بینند که وابسته بودن یا نبودن سکانس ورودی را تعیین کند. تجربه نشان می‌دهد که

---

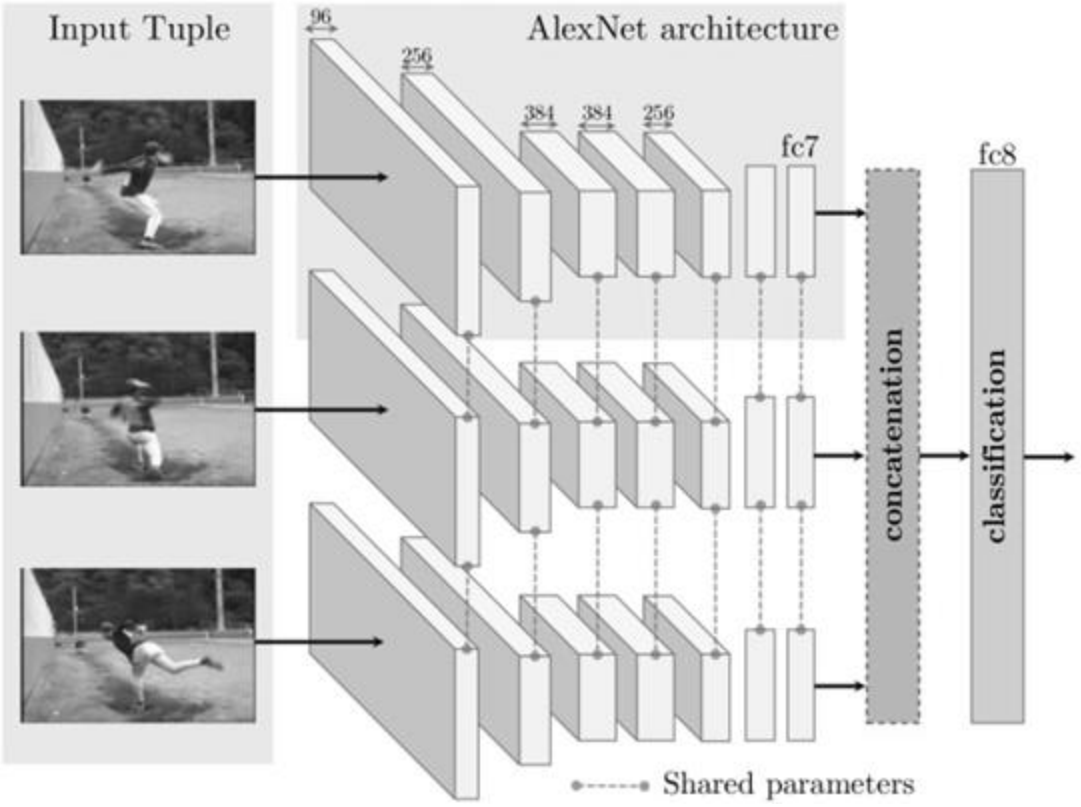
۵۰ Adversarial Model

۵۱ Temporal Coherency Network

۵۲ weak supervision

۵۳ Pose

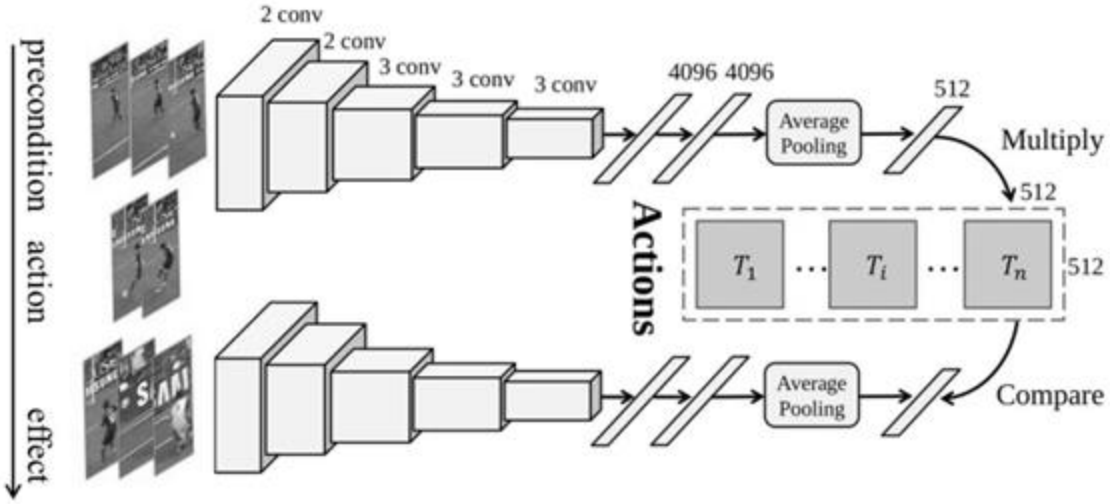
- در مقایسه با دیگر متدهای از پیش آموزش‌دیده شده‌ی برچسب دار مانند ImageNet ، tuple مبتنی بر آموزش توجه بیشتری به موقعیت انسان دارد.
- Tupleهایی که از فریم‌های با حرکت زیاد انتخاب می‌شوند ابهامات بین Tuple های مثبت و منفی را برطرف می‌کنند.
- در مقایسه با شبکه‌هایی که آموزش می‌بینند، شبکه‌هایی که از پیش آموزش‌دیده شده مبتنی بر وابستگی زمانی است، دارای پتانسیل بهبود در دقت می‌باشند.



تصویر ۲-۱۸ – شبکه Siamese Triplet که در [۶۶] مورداستفاده قرارگرفته است. هر یک از این شبکه‌ها حرکت و موقعیت ارائه‌شده در اعمال را در نظر گرفته است.

مطالعه دیگر که توسط [۱۵] انجام‌شده است یک رفتار را به سه قسمت تقسیم می‌کند که ۲ قسمت آن برای طبقه‌بندی اهمیت ویژه‌ای دارد. یک ویدیو با فریم های $(x_1, x_2, \dots x_n)$ به مجموعه شرایط از پیش تعیین‌شده $x_p = (x_1, x_2, \dots x_p)$ و مجموعه مؤثر $x_e = (x_e, x_{e+1}, \dots x_n)$ تقسیم می‌شود. هر دو

مجموعه توسط مدل عمیق آموزش می‌بینند. به‌منظور شناسایی یک رفتار نیاز به تبدیل توصیفگرهای استخراج‌شده سطح بالای X_p به توصیفگرهای استخراج‌شده سطح بالای X_e ، است. توصیفگرهای سطح بالا و عملیات تبدیل با استفاده از شبکه Siamese آموزش می‌بینند. تصویر ۲-۱۹ جزئیات را نمایان ساخته است.



تصویر ۲-۱۹ – ساختارهای موازی کانولوشن که به‌منظور استخراج پیش‌شرط و ویژگی‌ها مورداستفاده قرار می‌گیرند.

در تکمیل این موضوع به‌مرور مقاله [۶۷] می‌پردازیم که مدعی است موفقیت در مدل کردن زبان توسط RNN یکی از نتایج استفاده از فضای اطلاعاتی گسسته است. به همین منظور ساختار گسسته‌ای را برای فریم‌های ویدئویی معرفی کرده است که توسط کوانتیزه کردن آن‌ها به مجموعه‌ای از تکه‌های تصاویر به دست می‌آید. آن‌ها مشاهده کردند که ویدیوها ممکن است توالی موردنظرشان را تأمین نکند، که این مسئله باعث شد یک شبکه عصبی کانولوشنی بازگشتی[۵۴] را با سکانس‌های بلندتر آموزش دهند. با آموزش سکانس‌های بلند و با توجه به پروسه بازگشتی بودن آن، یک شبکه را با زمان طولانی‌تری می‌سازند که ممکن است مقاومت آن را در برابر خطاهای مرحله آزمایش افزایش یابد.

[۵۴] Recurrent Convolutional Neural Network

در این بخش به‌مرور و بررسی برخی از راه‌حل‌های موجود برای شناسایی اعمال و رفتارها پرداختیم. در ابتدا روش‌های مبتنی بر راه‌حل‌های دستی[۵۵] و در ادامه به بررسی روش‌های مبتنی بر راه‌حل‌های شبکه‌های عصبی عمیق پرداختیم. در بخش بعدی به بررسی شناسایی رفتارهای نامتعارف در تصاویر ویدیویی و همچنین روش ارائه‌شده در این پایان‌نامه پرداخته می‌شود.

[۵۵] Hand-crafted

فصل سوم

۳- شناسایی رفتارهای ناهنجار در تصاویر ویدیویی

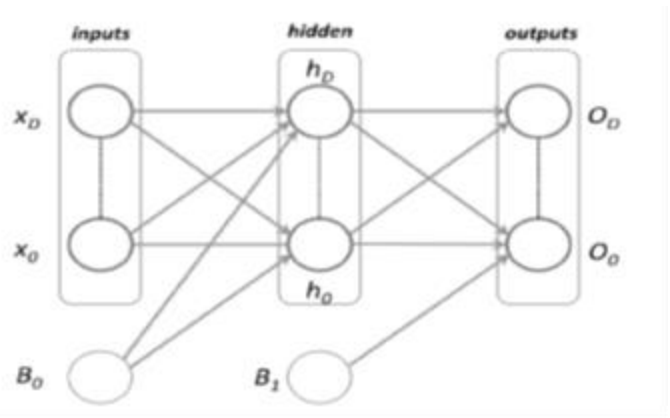
روش ارائه‌شده در این پایان‌نامه ترکیبی از ویژگی‌های شبکه عصبی کانولوشنی ارائه‌شده و همچنین ویژگی‌هایی است که از هیستوگرام‌های HOG و HOF استخراج‌شده‌اند. پیاده‌سازی این روش‌ها در زبان Python و استفاده از نرم‌افزارهای Theano و Keras انجام‌گرفته است. در ادامه به بررسی این ساختارها پرداخته خواهد شد.

۳-۱- معرفی شبکه‌های عصبی، Theano و Keras

در این بخش به مشخصه شبکه‌های عصبی ، شبکه‌های عصبی کانولوشنی و معرفی نرم‌افزارهای Theano و Keras می‌پردازیم.

۳-۱-۱- شبکه عصبی

شبکه‌های عصبی اولین بار توسط Rosenblatt [۶۸] در سال ۱۹۵۸معرفی شد. شبکه‌های عصبی از یک مدل بیولوژیکی شبیه‌سازی‌شده از کارایی مغز انسان الهام گرفته‌شده است. این شبکه بااتصال چند نورون ساخته‌شده است که هر یک از نورون‌ها با اعمال تابع‌های ساده غیرخطی قادر هستند مسائل پیچیده را مدل کنند. یک مدل معمول از این شبکه‌ها از یک‌لایه ورودی، لایه خروجی و یک یا چندلایه مخفی تشکیل‌شده است. در تصویر ۳-۱ ساختار یک شبکه عصبی نمایش داده شده است.



تصویر ۳-۱- ساختار یک شبکه عصبی. نودها در لایه ورودی درواقع همان داده‌های ورودی هستند درحالی‌که نودهای لایه‌های مخفی و خروجی همان پرسپترون ها هستند. هر اتصال بین نودها یک اتصال وزن‌دار را نشان می‌دهد.

لایه‌های مخفی و خروجی از پرسپترون هایی تشکیل‌شده‌اند که اتصالات ورودی را وزن‌دار می‌کنند و زمانی که مقدار آستانه‌ی توابع فعال‌ساز محقق شود فعال می‌شوند. هریک از پرسپترون ها توسط یک الگوریتم بازگشتی بروز می‌شوند تا بتوانند خطای بین خروجی شبکه و هدف را به حداقل ممکن برسانند.

زمانی که وزن‌ها در شبکه آموزش دیدند می‌توان پی برد که کدام ویژگی‌ها در یک مسئله نقش پررنگی را ایفا می‌کنند. اما این ابزار پرقدرت که تنها با پرسپترون ها کار می‌کند نمی‌تواند در مسائلی که هم به اطلاعات مکانی و هم به اطلاعات زمانی وابستگی دارد به‌خوبی ایفای نقش کند که این ضعف باعث توسعه شبکه‌های کانولوشنی شده است.

۳-۱-۲- شبکه عصبی کانولوشنی[۵۶] CNN

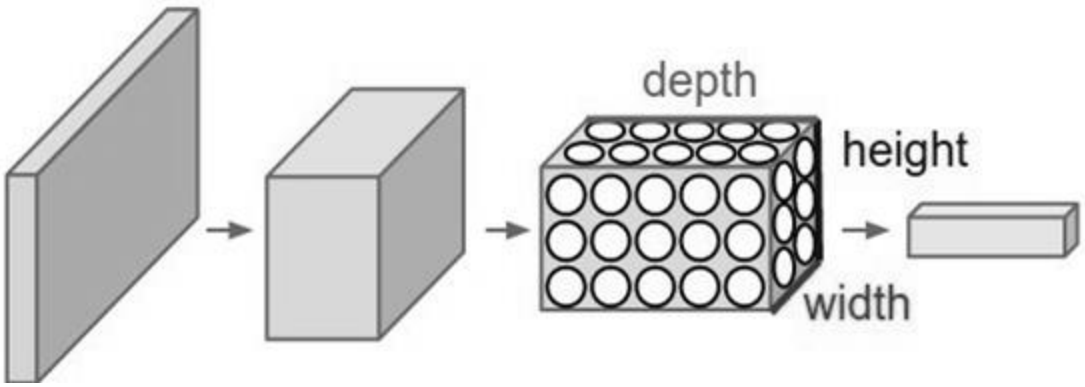
شبکه‌های عصبی کانولوشنی تا حد بسیار زیادی شبیه شبکه‌های عصبی مصنوعی هستند که در بخش قبلی در مورد آن‌ها توضیح داده شد. شبکه‌های عصبی کانولوشنی در ابتدا توسط Lecun [۴۶] ارائه گردید. این نوع شبکه‌ها متشکل از نورون‌هایی با وزن‌ها و بایاس‌های قابل یادگیری (تنظیم) هستند. هر نورون تعدادی ورودی دریافت کرده و سپس حاصل‌ضرب وزن‌ها در ورودی‌ها را محاسبه کرده و در انتها با استفاده از یک تابع فعال‌سازی غیرخطی نتیجه‌ای را ارائه می‌دهد. کل شبکه همچنان یک تابع امتیاز مشتق‌پذیر را ارائه می‌کند, که در یک‌طرف آن پیکسل‌های تصویر ورودی و در طرف دیگر آن امتیازات مربوط به هر دسته قرار دارد. این نوع شبکه‌ها یک تابع هزینه[۵۷] (مانندSoftmax) نیز در لایه آخر دارند.

شبکه‌های عصبی کانولوشنی از این واقعیت که ورودی شامل تصاویر است استفاده کرده و معماری شبکه را به روش معقولی محدود کردند. به‌طور خاص, برخلاف یک شبکه عصبی معمولی, لایه‌های یک شبکه

[۵۶] Convolutional Neural Network

[۵۷] Loss function

عصبی کانولوشن (به‌اختصار ConvNet) شامل نورون‌هایی است که در سه بعد عرض, ارتفاع و عمق[۵۸] قرارگرفته‌اند. هر نورون در هر لایه به‌جای اتصال با تمام نورون‌ها در لایه قبل تنها به ناحیه کوچکی از لایه قبل از خود متصل است. چراکه همگام با رسیدن به انتهای معماری شبکه اندازه تصویر کاهش پیدا می‌کند به‌گونه‌ای که در انتها، تصویر کامل ورودی به یک بردار حاوی امتیاز دسته‌ها کاهش پیدا می‌کند و با یک بردار که حاوی امتیاز هر دسته است مواجه خواهیم بود. تصویر ۳-۲ نمایشی از لایه‌های یک شبکه کانولوشنی را نمایش می‌دهد.



تصویر ۳-۲- لایه‌های یک شبکه ساده کانولوشنی از نورون‌هایی که در سه بعد عرض، ارتفاع و عمق قرارگرفته‌اند، ساخته‌شده است.

همان‌طور که در تصویر بالا می‌بینید هر یک از لایه‌های شبکه عصبی کانولوشنی نورون‌های خود را در سه بعد عرض، ارتفاع و عمق قرار می‌دهند. در این مثال لایه ورودی قرمزرنگ حاوی مقادیر پیکسل‌های تصویر است، بنابراین عرض و ارتفاع آن ابعاد تصویر خواهند بود و عمق آن‌هم برابر با کانال‌های مربوط به تصویر خواهد بود. یک شبکه کانولوشنی از چندلایه تشکیل می‌شود و هر لایه شیوه کار ساده‌ای دارد که در آن‌یک توده سه‌بعدی ورودی دریافت می‌شود و آن را با استفاده از توابع مشتق‌پذیر[۵۹] به یک توده سه‌بعدی خروجی تبدیل می‌کند.

[۵۸] کلمه عمق در اینجا اشاره به بُعد سوم یک توده فعال سازی (activation volume) دارد و به معنای عمق یک شبکه عصبی کامل که به معنای تعداد لایه های موجود در آن است نمیباشد.

[۵۹] Differentiable function

این شبکه‌ها از سه مکانیسم بهره می‌برند.

۱- حوزه‌های محلی به هم مرتبط [۶۰]

۲- اشتراک وزن‌ها [۶۱]

۳- نمونه‌برداری [۶۲]

حوزه‌ی محلی به هم مرتبط توسط یک صفحه که به آن نقشه ویژگی گفته می‌شود، سازمان‌دهی شده است. هر حوزه وزن‌هایش را به اشتراک می‌گذارد و الگوهای محلی را در یک تصویر تنها بااتصال هر نورون به قسمت کوچکی از ورودی به دست می‌آورد. این کار از مزیت همبستگی پیکسل‌های نزدیک به یکدیگر نسبت به پیکسل‌هایی که از یکدیگر دور هستند بهره می‌برد. با اشتراک اوزان حوزه‌های محلی مرتبط، تابع فعال‌ساز در هر نورون، معادل با کانوالو کرنل با ناحیه نمونه‌برداری شده از شدت پیکسل‌های تصویر، عمل می‌کند. ارتباط بین لایه‌های ورودی و لایه‌های دیگر توسط تصویر ۳-۳ نمایش داده‌شده است. نورون‌ها، یک صفحه با نام لایه کانولوشنی تشکیل می‌دهند. با لغزش حوزه‌های محلی در طول تصویر می‌توان ویژگی‌هایی صرف‌نظر از موقعیت آن به دست آورد.

چنانچه اوزان حوزه محلی در هر نورون لایه کانولوشن مشابه باشند، می‌توان مشاهده کرد که لایه کانولوشن تنها یک تصویر کانوالوی از لایه قبلی است. بنابراین اوزان توسط فیلتر کانولوشنی مشخصه دار می‌شوند و در کنار یک بایاس [۶۳] آموزش می‌بینید. زمانی که ورودی از چندین کانال ساخته‌شده باشد نورون به عملگر جمع کننده کانوالوی در تمامی کانال‌ها با ناحیه مشابه تبدیل می‌شود.

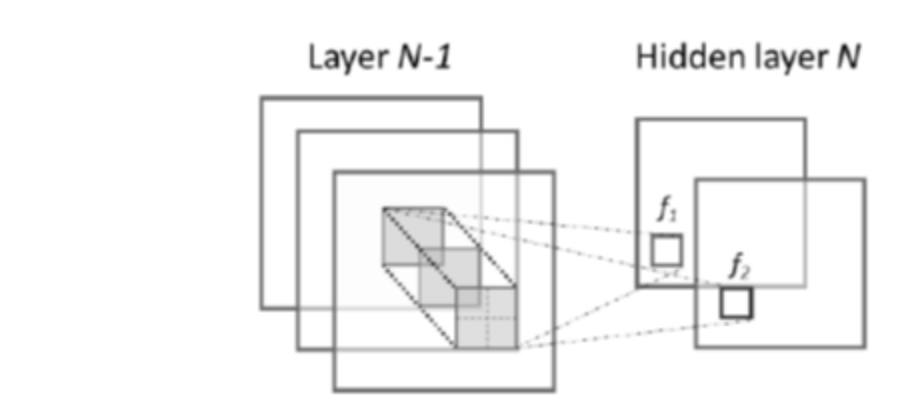
---

[۶۰] Local receptive field

[۶۱] Weight sharing

[۶۲] Sub-sampling

[۶۳] Bias

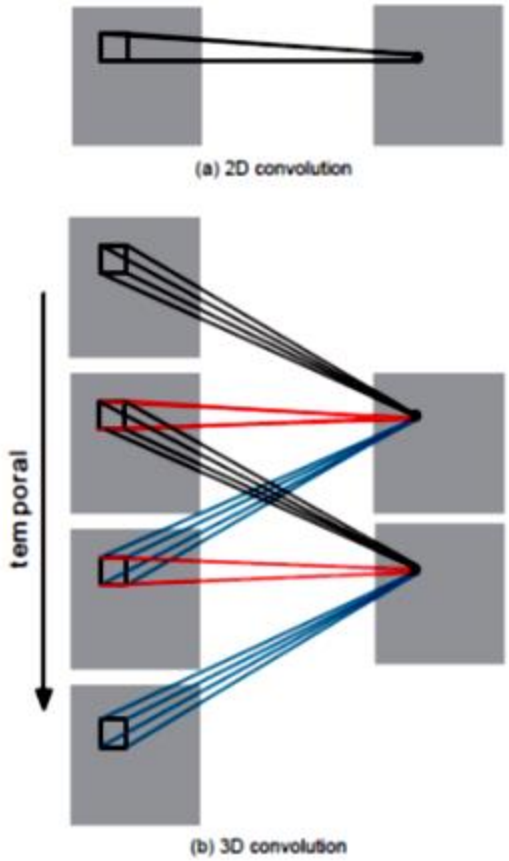


تصویر ۳-۳- دیاگرام نمایش ارتباط بین ورودی و لایه‌های دیگر. هر یک از نقشه ویژگی در لایه‌های مخفی توسط فیلتر کانولوشن لایه خود مورداستفاده قرار می‌گیرد. پردازش فیلترها در ورودی توسط یک پنجره لغزان که نتایج کانولوشن در هر کانال در مختصات مشابه را جمع می‌کند، به دست می‌آید.

در شبکه‌های کانولوشن دوبعدی عملیات کانولوشن بر روی نقشه‌های ویژگی دوبعدی اعمال می‌شد تا فقط ویژگی‌های مکانی را محاسبه کند. اما زمانی که مسئله موردنظر در حوزه آنالیز ویدیو است می‌بایست اطلاعات حرکتی که در فریم‌های متوالی وجود دارد را استخراج کرد. به همین منظور از کانولوشن سه‌بعدی در جایگاه شبکه‌های کانولوشنی برای استخراج ویژگی‌های مکان-زمانی استفاده می‌شود [۵۱]. کانولوشن سه‌بعدی از کانوالو یک کرنل سه‌بعدی به فرم مکعب بر روی چندین فریم متوالی اعمال می‌شود. توسط این ساختار نقشه‌های ویژگی در لایه کانولوشن به چندین فریم متوالی در لایه قبلی متصل شده است تا بتواند اطلاعات حرکتی را استخراج کند. به‌طورمعمول مقدار موقعیت (x,y,z) در j امین نقشه ویژگی در i امین لایه به‌صورت زیر محاسبه می‌شود.

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr}\, v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \qquad (۱-۳)$$

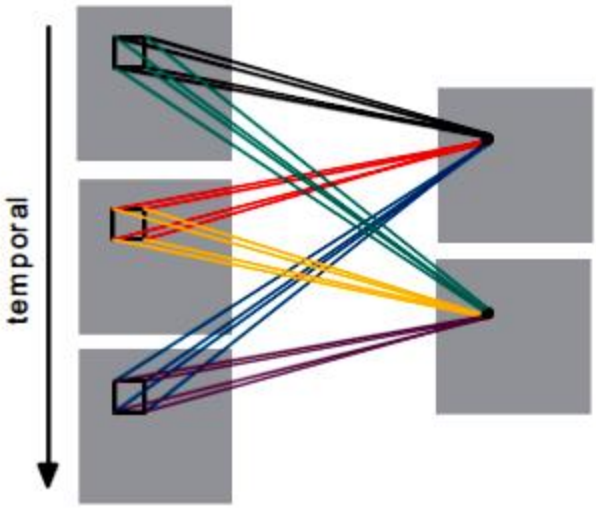
R_i سایز کرنل 3D در طول بعد زمانی است. W_{ijm}^{pqr} مقدار (p,q,r) ام از کرنل متصل به m امین نقشه ویژگی در لایه قبلی است. مقایسه‌ی بین کانولوشن 2D و 3D در تصویر ۳-۴ نمایش داده‌شده است.



تصویر ۳-۴- مقایسه بین کانولوشن 2D و 3D .سایز کرنل کانولوشنی در بعد زمانی ۳ است و مجموعه اتصالات بارنگ کد شده‌اند که وزن‌های اشتراکی با یک رنگ نمایش داده‌شده است. در کانولوشن 3D، کرنل 3D مشابه بر سلول‌های مکعبی در ویدیو اعمال می‌شود تا ویژگی‌های حرکتی را استخراج کند [۵۱].

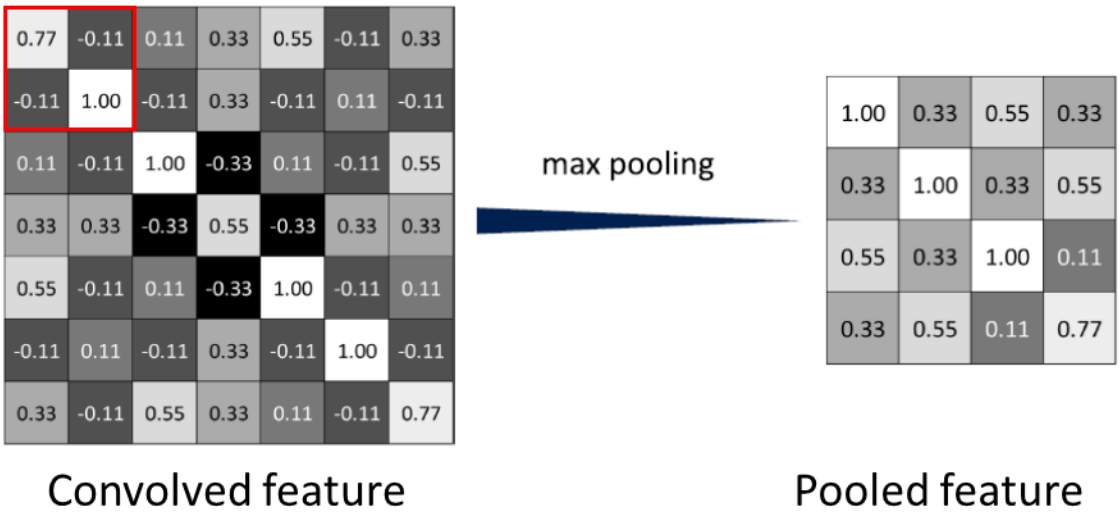
می‌بایست به این نکته توجه داشت زمانی که اوزان کرنل در طول مکعب تکرار شوند، یک کرنل 3D کانولوشنی تنها می‌تواند یک نوع از ویژگی‌ها را از فریم‌های مکعبی استخراج کند. یک روش معمول برای طراحی شبکه‌های مختلف این است که تعداد نقشه‌های ویژگی با تولید انواع مختلف ویژگی از مجموعه نقشه‌های ویژگی سطح پایین[۶۴] در لایه‌های بعدی افزایش یابد. همچون کانولوشن 2D این کار می‌تواند با اعمال کانولوشن های 3D با کرنل‌های مجزا در محل‌های مشابه در لایه قبلی به دست آید. در تصویر ۳-۵ شمایی از آن نمایش داده‌شده است.

[۶۴] Low-level feature



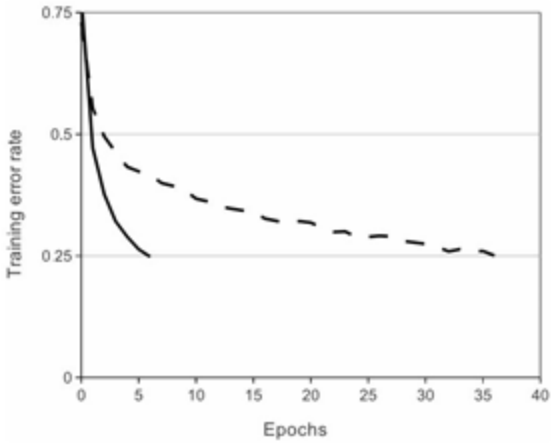
تصویر ۳-۵- استخراج ویژگی‌های مختلف از فریم‌های متوالی. چندین نوع کانولوشن 3D را می‌تواند بر روی فریم‌های متوالی به‌منظور استخراج چندین نوع ویژگی مختلف اعمال کرد [۵۱].

خروجی لایه کانولوشن به لایه نمونه‌برداری داده می‌شود که تابعی را که مربوط به زیر ناحیه‌ای از ورودی می‌شود، محاسبه کند. تابع مذکور معمولاً یک تابع متوسط گیر یا ماکزیمم گیر است. تصویر ۳-۶ نحوه اعمال یک ماکزیمم گیر با اندازه پنجره ۲×۲ را بر روی خروجی لایه کانولوشن نشان می‌دهد.



تصویر ۳-۶- اعمال یک ماکزیمم گیر با اندازه پنجره۲×۲ بر روی ویژگی‌های کانوالوشده

این لایه درواقع با نمونه‌برداری منجر به هموار شدن[۶۵] ورودی می‌شود. در حوزه ماکزیمم گیر[۶۶] به‌طور طور مثال یک تصویر به تکه‌هایی تقسیم می‌شود که مقدار ماکزیمم در هر قطعه محاسبه می‌شود. لایه بعدی ReLU[۶۷] است که چگونگی جریان داده در بین لایه‌ها را کنترل می‌کند. تابع ReLU به‌صورت $f(x)=max(0,x)$ است و همچنین تابع‌های فعال‌ساز دیگری نیز مانند تانژانت هایپربولیک $f(x)=tanh(x)$ و سیگموید $f(x)=(1+e^{-x})^{-1}$ نیز می‌توانند مورداستفاده قرار گیرند. اما به دلیل دارا بودن زمان آموزش کمتر، تابع ReLU بر دیگر توابع ترجیح داده می‌شوند. سرعت همگرایی تابع ReLU نسبت به تابع Tanh در تصویر ۳-۷ نشان داده شده است.

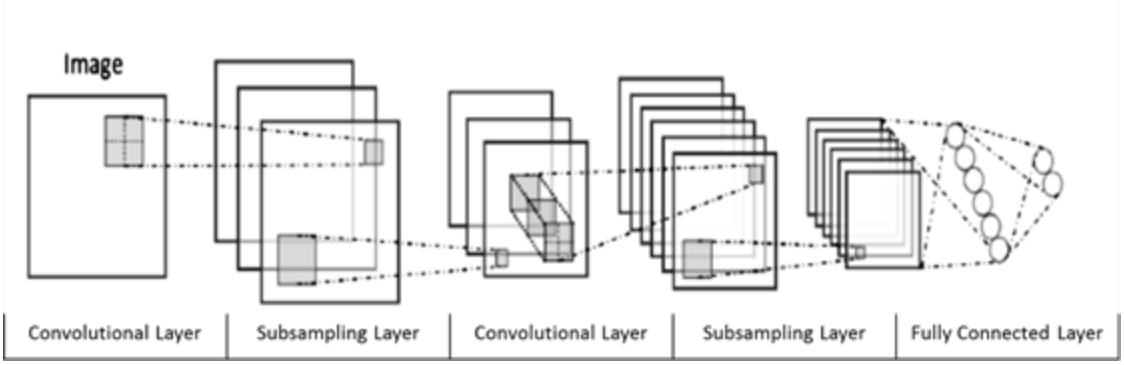


تصویر ۳-۷- مقایسه همگرایی توابع ReLU و Tanh . بهبود شش برابری همگرایی با ReLU (خط ساده) در برابر همگرایی با Tanh (خط چین)

این لایه علاوه بر اینکه می‌تواند باعث کاهش حجم عملیات و پیچیدگی محاسباتی شود از وقوع پدیده برازش[۶۸] نیز جلوگیری می‌کند. خروجی این لایه نیز نسبت به تغییرات کوچک و انتقال مقاوم است [۶۹]. این شبکه از لایه کاملاً متصل[۶۹] در مسائل مربوط به طبقه‌بندی پیروی می‌کند. لایه کاملاً متصل به‌عنوان یک‌لایه کانولوشن ویژه به‌حساب می‌آید که سایز فیلتر با سایز ورودی برابر است و تنها یک

[۶۵] Smooth
[۶۶] Max-Pooling
[۶۷] Rectified Linear Unit
[۶۸] Over-fitting
[۶۹] Fully Connected layer

کانوالو در آن اعمال می‌شود. همان‌طور که قابل‌مشاهده است تصویر ۳-۸ ساختار یک شبکه کانولوشنی را نشان می‌دهد.



تصویر ۳-۸- ساختار یک شبکه کانولوشنی. این شبکه از ترکیب لایه‌های کانولوشن، ادغام کننده و کاملاً متصل تشکیل‌شده است که یک مدل را آموزش دهد تا بتواند مسائل مربوط به طبقه‌بندی را حل کند.

۳-۱-۳- معرفی نرم‌افزارهای Theano و Keras

Theano [۷۰, ۷۱] مجموعه نرم‌افزاری است که توسط گروه یادگیری ماشین MILA در دانشگاه مونترال توسعه داده‌شده است. این نرم‌افزار به‌منظور ارزیابی و بهینه‌سازی روابط ریاضی به زبان python مورداستفاده قرار می‌گیرد. این نرم‌افزار همچنین قادر است در CPU و GPU اجرا شود و در بسیاری از مواقع در GPU به زبان C پیاده‌سازی شده است. Theano به‌منظور انجام محاسبات موردنیاز شبکه‌های عصبی در مقیاس بزرگ طراحی‌شده است و از زمانی که در سال ۲۰۱۰ معرفی‌شده است به‌عنوان نرم‌افزار کاربردی در حوزه‌های تجاری و تحقیقاتی مورداستفاده قرارگرفته است.

Keras [۷۲] یک رابط نرم‌افزاری برای شبکه‌های عصبی سطح بالا به زبان Python نوشته‌شده است که توسط توابع موجود در [۷۳] Tensor flow و Theano اجرا می‌شود و به‌عنوان عضوی از ONEIROS[۷۰] شناخته می‌شود. این کتابخانه شبکه‌های کانولوشنی و برگشت‌پذیر را به‌خوبی پشتیبانی می‌کند. این نرم‌افزار رابطی کاربرپسند و ساده دارد که بر روی CPU و GPU اجرا می‌شود.

[۷۰] Open-ended Neuro-Electronic Intelligent Robot Operating System

۳-۲- معرفی ساختارهای پیشنهادی جهت تشخیص رفتارهای ناهنجار

در این بخش به معرفی روش و ساختارهای به کار گرفته شده در این پایان‌نامه پرداخته خواهد شد.

۳-۲-۱- معرفی روش پیشنهادی

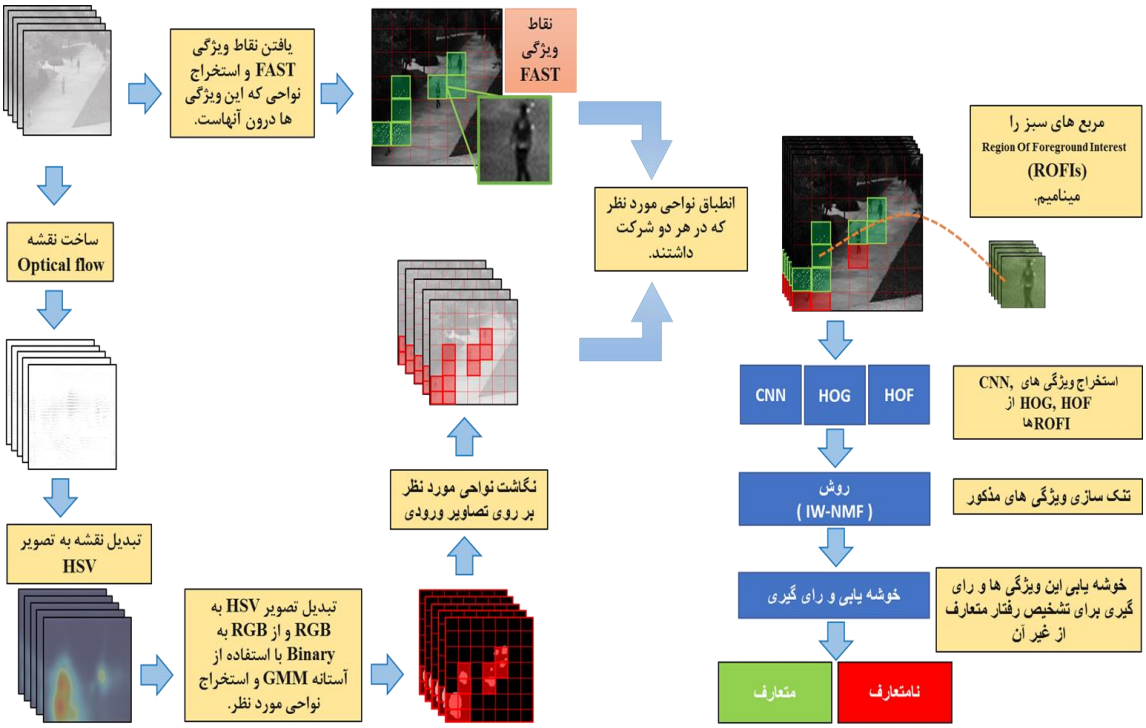
در حال حاضر بسیاری از مقالات ارائه‌شده حاکی از قدرت [۴۶] CNN در حوزه‌های مختلف بینایی ماشین ازجمله شناسایی متن [۷۴] شناسایی و طبقه‌بندی اشیا [۴۷, ۷۵] شناسایی لبه [۷۶] و شناسایی صورت [۷۷] است. در مورد مسائل مربوط به طبقه‌بندی در ویدیو، CNN می‌تواند نقش پررنگ و پراهمیتی را بازی کند. ازجمله روش‌هایی که به این حوزه پرداختند می‌توان به شبکه CNN سه‌بعدی که توسط [۵۱] به‌منظور شناسایی حرکات انسان در سکانس‌های ویدیویی ارائه‌شده است و یا روشی که در [۷۸] به‌منظور شناسایی اعمال انسان به‌صورت 3D در تصاویر RGB-D ارائه‌شده است و روشی که در [۷۹] با استفاده از یک CNN سه‌بعدی محل فرود بالگرد را شناسایی می‌کند، اشاره داشت.

به‌منظور شناسایی و موقعیت‌یابی اعمال غیر نرمال در سکانس‌های ویدیویی موفق به ارائه یک مدل CNN زمان-مکانی شدیم. این مدل می‌تواند اطلاعات ایستا در تصویر و همچنین اطلاعات پیچیده مربوط به حرکات را از فریم‌های متوالی استخراج کند.

برای شناسایی اعمال غیر نرمال در یک قسمت کوچکی از فریم، مدل CNN زمان-مکانی بر روی توده‌های زمان-مکانی اعمال می‌شود. به دلیل اینکه نتایج مقاوم‌تری حاصل شود از ویژگی‌هایی نظیر توصیفگرهای HOG,HOF نیز در این توده‌ها بهره برده شده است که هرکدام از این دو توصیفگر به‌صورت یک مجموعه لغت‌نامه ساخته‌شده است. این مجموعه‌ها به‌صورت لغت‌نامه تنک شده[۷۱] توسط روش مبتنی بر تجزیه ماتریس‌های نامنفی آموزش داده می‌شوند [۸۰-۸۲].

[۷۱] Sparse dictionary coding

بر همین مبنا روش اوزان تکرارشونده در تجزیه ماتریس‌های نامنفی[۷۲] IW-NMF در این پایان‌نامه به‌تازگی ارائه‌شده است که ویژگی‌ها را تنک و آموزش می‌دهد. این روش یکی از روش‌های تجزیه ماتریس‌هاست که یک ماتریس نامنفی به‌صورت ضرب دو ماتریس نامنفی دیگر تجزیه می‌شود. این روش هیچ محدودیتی جز نامنفی بودن ماتریس‌ها ندارد و به دلیل اینکه HOG,HOF دو توصیف‌گر نامنفی هستند بنابراین می‌توان از این متد برای آموزش ویژگی‌ها بهره برد. در ادامه می‌توان با استفاده از خوشه‌یابی هر یک از ویژگی‌های استخراج‌شده را به دو خوشه نرمال و غیر نرمال خوشه‌بندی کرد و برای تعیین و اتخاذ تصمیم از روش رأی‌گیری استفاده‌شده است. شمای کلی روش ارائه‌شده در تصویر ۳-۹ نمایش داده‌شده است.

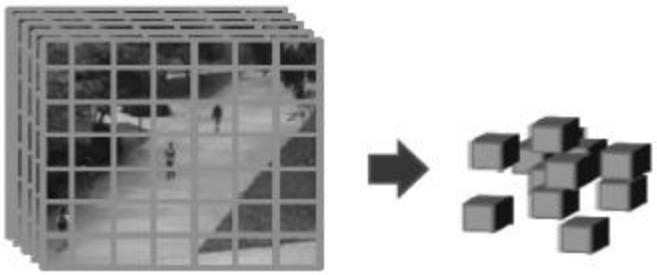


تصویر ۳-۹- شمای کلی روش ارائه‌شده در این پایان‌نامه

[۷۲] Iterative weighted Nonnegative Matrix Factorization

۳-۲-۲- نحوه نمایش ویدیو

به‌منظور آماده‌سازی ویدیو ابتدا می‌بایست هر یک از فریم‌های ویدیویی را به تعدادی سلول‌های مکعبی غیر همپوشان تبدیل کنیم که در تصویر ۳-۱۰ نشان داده‌شده است.



تصویر ۳-۱۰- نحوه نمایش ویدیو. هر یک از فریم‌های ویدیو به‌صورت تعدادی سلول‌های مکعبی غیر همپوشان تبدیل می‌شوند.

اندازه این نواحی غیر همپوشان برابر هستند و با توجه به نواحی به‌دست‌آمده تعداد هفت فریم متوالی با یکدیگر تشکیل یک مکعب زمان-مکانی را می‌دهد. روش ارائه‌شده به‌منظور شناسایی رویدادهای غیرمعمول از همین سلول‌های غیر همپوشان استفاده می‌کند و نواحی موردنظر را در آن‌ها می‌یابد. این مکعب‌ها درواقع نواحی محلی مکان-زمانی موردنظر را نیز توصیف می‌کنند.

نواحی مربوط به اعمال نرمال معمولاً ارتباطات مشابهی بین همسایگان خود دارند و همچنین احتمال وقوع آن‌ها بالاتر است. اما در مورد اعمال غیر نرمال تشابه بین این نواحی و نواحی در همسایگی آن‌ها یک الگوی واحد را نتیجه نمی‌دهد. بدیهی است که احتمال وقوع نواحی غیر نرمال به‌مراتب کمتر از نواحی معمول است.

۳-۲-۳- نواحی پیش‌زمینه و Optical flow

اطلاعات مربوط به پیش‌زمینه تصاویر برای شناسایی رفتار غیرمعمول در ویدیو به‌منظور استخراج نواحی پیش‌زمینه موردنظر(ROFIs[۷۳]) بسیار اهمیت دارد. بنابراین در ابتدا نقشه Optical flow که مزایای آن در شناسایی حرکات و ردگیری اشیا در بسیاری از حوزه‌های بینایی ماشین به اثبات رسیده است، محاسبه می‌شود. در ادامه توسط روش حذف پس‌زمینه در این نقشه که با استفاده از روش GMM[۷۴] انجام می‌شود، این نقشه را به تصویر باینری تبدیل می‌کنیم. در این مرحله نواحی موردنظر در تصویر باینری به دست می‌آید که این نواحی در تصویر ورودی نگاشت می‌شوند. تصویر ۳-۱۱ چگونگی انتخاب نواحی موردنظر را در تصویر باینری نشان می‌دهد.

(a) (b) (c)

تصویر ۳-۱۱- شمایی از استخراج نواحی موردنظر. a) فریم ورودی و سلول‌های غیر همپوشان b) تصویر باینری Opticalflow و شناسایی نواحی موردنظر c) نگاشت این نواحی در فریم ورودی.

اگر هرکدام از نواحی شناسایی‌شده در تصویر باینری را به‌عنوان B_t در فریم t بنامیم، در این صورت سلول مکعبی را به‌عنوان $u \in R^3$ در B_t تعریف می‌کنیم. سلول مکعبی u دارای ابعاد m_x, m_y, m_t است که ابعاد m_x و m_y توسط ابعاد افقی و عمودی هر یک از سلول‌ها تعیین می‌شود و m_t نمایانگر تعداد فریم‌های متوالی در نظر گرفته‌شده است که در تمامی سلول‌های ویدیویی مکعبی مقداری ثابت دارد.

[۷۳] Region Of Foreground Interests

[۷۴] Gaussian Mixture Model

اندازه اشیای یافت شده و همچنین حضور این اشیا در فریم‌های مختلف را می‌توان به‌آسانی با شمارش تعداد پیکسل‌های پیش‌زمینه در هر سلول مکعبی مانند u به دست آورد.

$$F(u) = \sum_{d=1}^{m_t} \sum_{j=1}^{m_y} \sum_{i=1}^{m_x} u(i,j,d) \qquad (۲-۳)$$

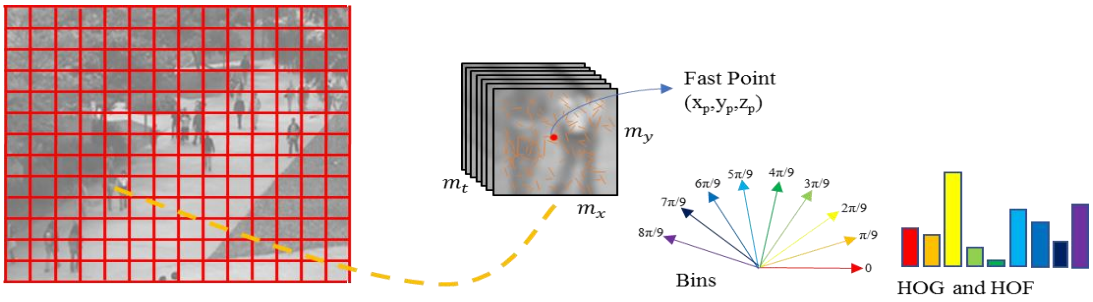
تنها سلول‌هایی که دارای حداقل آستانه اشغال پیش‌زمینه F(u) هستند به‌عنوان سلول‌های فعال به شمار می‌روند. به‌طورمعمول اگر سلولی حداقل ۱۰٪ پیکسل‌هایش مربوط به‌پیش زمینه باشد به‌عنوان سلول فعال شناخته می‌شود.

۳-۲-۴- شناساگر ویژگی FAST و توصیفگرهای HOG,HOF

به‌منظور استخراج ویژگی‌های به‌اصطلاح دستی[۷۵] از داده‌های اصلی، در ابتدا باید نواحی مورد نظر ROFIs ها را با استفاده از شناساگر FAST شناسایی کرد. FAST یک متد شناساگر نقطه است که نقاط موردنظر را با مقایسه شدت یک پیکسل به نام P با همسایه‌هایش، شناسایی می‌کند. اگر تمام شدت پیکسل‌های همسایگان بیشتر یا کمتر از شدت پیکسل P باشد در این صورت P را به‌عنوان یک نقطه موردنظر در نظر می‌گیرد. این شناساگر باینری دارای مزایایی ازجمله سرعت در شناسایی است که در این‌گونه مسائل می‌تواند پراهمیت باشد [۸۳, ۸۴]. حال شناساگر FAST را بر روی سلول‌های مکعبی فعال اعمال می‌کنیم. برای هر نقطه با موقعیت مکانی (x_p,y_p) در بعد زمانی سلول فعالی شناسایی می‌شود. بعد از انتخاب نقطه موردنظر یک سلول ویدیویی $V \in R^3$ با اندازه m_x,m_y,m_t به مرکزیت (x_p,y_p,t_p) ساخته می‌شود. اندازه m_x,m_y توسط سایز سلول فعال موردنظر تعیین می‌گردد و m_t در تمام حجم‌های ویدیویی ثابت است.

[۷۵] Handcrafted

توصیف‌کننده HOG,HOF از هیستوگرام‌های ۹ بخشی گوسی و Optical flow در بازه $(0,8\pi/9)$ محاسبه می‌شود. هیستوگرام‌ها با روش L1 نرمال‌سازی می‌شود که در تصویر ۳-۱۲ چگونگی تشکیل آن نمایش داده‌شده است.



تصویر ۳-۱۲- ویژگی‌های دستی (HOG , HOF) از سلول‌های مکعبی فعالی که نقاط FAST در آن قرارگرفته‌اند، استخراج می‌روند.

۳-۲-۵- معرفی مدل شبکه عصبی کانولوشنی – CNN

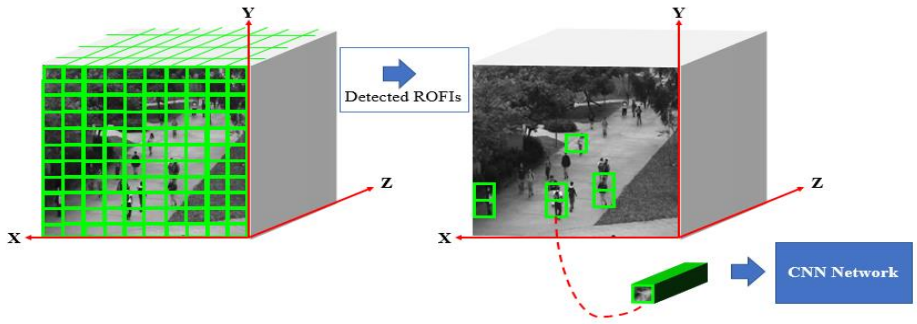
برای شناسایی و موقعیت‌یابی رفتار غیرمعمول که در نواحی محلی روی می‌دهد روش‌های متفاوتی درگذشته ارائه‌شده‌اند [۳۴, ۸۵, ۸۶]. این روش‌ها اغلب تعداد زیادی از تکه‌های غیر همپوشان و یا کل فریم را به‌عنوان یک مجموعه فارغ از اینکه این تکه‌ها شامل اطلاعات حرکتی باشد یا خیر، مورد پردازش قرار می‌دهند. بنابراین این‌گونه متدها هزینه محاسباتی بسیار و همچنین خطای بالای شناسایی را متحمل می‌گردند. به‌منظور به دست آوردن دقت بالاتر و همچنین هزینه محاسباتی پایین می‌توان از تکه سلول‌هایی که به‌صورت غیرفعال هستند صرف‌نظر کرد.

در این پایان‌نامه از شبکه کانولوشنی CNN که ارائه داده‌شده و همچنین ویژگی‌هایی ازجمله HOG,HOF به‌منظور افزایش دقت و مقاومت روش ارائه‌شده در شناسایی رفتار غیرمعمول در تصاویر ویدیویی استفاده کرده‌ایم. در ادامه به بررسی مدل CNN و همچنین ویژگی‌های استخراج‌شده توسط HOG,HOF پرداخته می‌شود.

به‌منظور شناسایی رفتار غیرمعمول در هر دو حوزه مکان و زمان، مدل CNN به‌صورت زمان-مکان ارائه‌شده است که ویژگی‌های حرکتی و بصری این رفتار ناهنجار را استخراج می‌کند.

۳-۲-۵-۱- استخراج ویژگی‌های زمان-مکان

مدل CNN زمان-مکانی ارائه‌شده به‌جای استفاده از کل فریم از اطلاعات موجود در سلول‌های فعال استفاده می‌کند. بنابراین تنها از پیکسل‌هایی که اطلاعات غنی‌تری از رویداد وقوع یافته دارند بهره می‌برد. اندازه مکانی هر یک از تکه سلول‌ها قبل از وارد شدن به مدل CNN می‌بایست طبق ورودی مدل CNN تنظیم و اندازه آن‌ها تغییر پیدا کند. در ادامه مدل CNN زمان-مکانی با اعمال کانولوشن زمان-مکانی بر روی سلول‌ها، ویژگی‌های سطح بالای زمان-مکان که در شناسایی رفتار غیرمعمول مؤثر است، استخراج می‌شوند. این پروسه در تصویر ۳-۱۳ زیر نشان داده‌شده است.



تصویر ۳-۱۳- شمایی از نواحی موردنظر که به‌عنوان ورودی به مدل CNN ارائه‌شده داده می‌شود.

۳-۲-۵-۲- کانولوشن زمان-مکان

لایه کانولوشن هسته اصلی مدل CNN است. پارامترهای لایه شامل مجموعه‌ای از فیلترها و یا کرنل‌های باقابلیت آموزش هستند که در حوزه‌های کوچکی باهم در ارتباط‌اند اما در تمام عمق یک سلول مکعبی گسترش پیداکرده است.

درواقع شبکه، فیلترهایی را که ویژگی‌های خاصی را در برخی موقعیت‌های مکانی ورودی شناسایی می‌کند فعال و آموزش می‌دهد. موضوع دیگری که در مدل‌های CNN از اهمیت خاصی برخوردار است لایه‌های ادغام و یا pooling هستند. درواقع این لایه‌ها به فرم یک نمونه‌بردار غیرخطی عمل می‌کنند.

لایه pooling درواقع باعث کاهش سایز، همچنین باعث کاهش تعداد پارامترها، کاهش حجم محاسبات می‌شود که درنتیجه باعث می‌شود پدیده برازش[۷۶] را کنترل کند. در j امین نقشه ویژگی a_{ij} را می‌توان در i امین لایه به‌صورت زیر تعریف کرد.

$$a_{ij} = f(W_n * a_{(i-1)n} + b_{ij}) \qquad (۳-۳)$$

که f(.) یک تابع فعال‌ساز است که می‌تواند به‌صورت $f(x)=\max(0,x)$ تعریف شود، به آن لایه ReLU گفته می‌شود. W کرنل فلیترهاست ، n نمایه مجموعه نقشه ویژگی است که به نقشه ویژگی در (i-1) امین لایه متصل شده است. * درواقع همان عملگر کانولوشن است و b_{ij} مقدار بایاس نقشه ویژگی کنونی است. W و b می‌بایست آموزش ببینند که بتوانند ویژگی‌های بهتری را استخراج کند.

برای استخراج اطلاعات زمان-مکانی، عملیات کانولوشن زمان-مکانی در لایه‌های کانولوشنی CNN اعمال می‌شود. کانولوشن زمان-مکانی توسط کانوالو یک کرنل سه‌بعدی روی سلول‌های مکعبی زمان – مکانی به دست می‌آید. درواقع این کانولوشن سه‌بعدی توسط یک کرنل سه‌بعدی بر روی سلول‌های مکعبی که از انباره سازی فریم‌های متوالی به دست می‌آید اعمال می‌شود. با این ساختار نقشه‌های ویژگی در لایه کانولوشنی، به چندین فریم متوالی در لایه قبل متصل می‌شود. بنابراین اطلاعات حرکتی را می‌تواند استخراج کرد.

عملگر کانولوشن زمان –مکانی بین کرنل سه‌بعدی W_n و سلول مکعبی زمان-مکان $a_{(i-1)n}$ را می‌توان به‌صورت زیر تعریف کرد

$$\left[W_n * a_{(i-1)n}\right](x,y,z) = \sum_{n} \sum_{u=0}^{U_i-1} \sum_{v=0}^{V_i-1} \sum_{r=0}^{R_i-1} W_n^{uvr} a_{(i-1)n}^{(x+u)(y+v)(t+r)} \qquad (۴-۳)$$

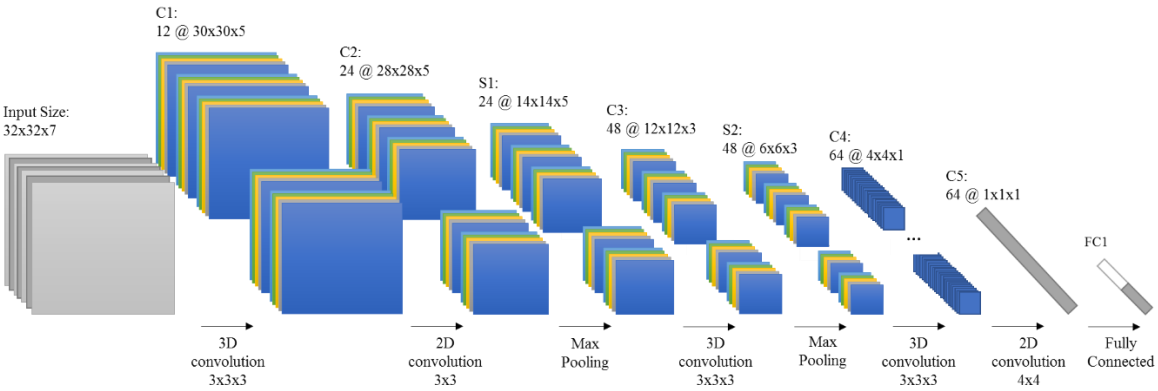
که U_i و V_i و R_i به ترتیب طول، عرض و بعد زمانی کرنل سه‌بعدی است و $x \times y \times t$ سایز زمان-مکان مربوط به سلول مکعبی $a_{(i-1)n}$ است.

[۷۶] Over-fitting

۳-۲-۵-۳- ساختار مدل CNN زمان-مکان

با توجه به توضیحاتی که در قسمت قبل در مورد کانولوشن 3D داده شد، ساختارهای بسیار متفاوتی از مدل CNN را می‌توان به وجود آورد. در ادامه به بررسی ساختار مدل CNN که برای شناسایی رفتار انسان توسعه داده‌ایم، پرداخته می‌شود.

در این شبکه تعداد هفت فریم متوالی که ابعاد هرکدام 32×32 است را به‌عنوان ورودی مدل CNN سه‌بعدی در نظر می‌گیریم. این مدل را می‌توان در تصویر ۳-۱۴ زیر مشاهده کرد.



تصویر ۳-۱۴- ساختار مدل CNN برای شناسایی رفتار غیر نرمال. این ساختار شامل سه لایه کانولوشنی 3D، دولایه کانولوشنی 2D، دولایه Max-pooling و یک‌لایه کاملاً متصل است.

ساختار این شبکه زمان-مکانی از هشت لایه تشکیل‌شده است. سایز ورودی در ساختار این شبکه به‌صورت $7 \times 32 \times 32$ است، همان‌طور که اشاره شد از 7 فریم متوالی به ابعاد 32×32 ساخته‌شده است. نواحی ROFI که در قسمت‌های قبلی به آن اشاره شد می‌بایست قبل از اینکه مورداستفاده مدل CNN قرار گیرد به ابعاد اشاره‌شده $7 \times 32 \times 32$ تغییر سایز پیدا کند.

در ابتدا کانولوشن زمان-مکان سه‌بعدی با کرنل با سایز $3 \times 3 \times 3$ را بر روی‌داده‌های ورودی اعمال می‌کنیم (3×3 بعد مکانی و 3 بعد زمانی کرنل است). می‌بایست به این نکته توجه داشت که یک کرنل 3D می‌تواند تنها یک نوع از ویژگی را از ROFI به دست آورد، بنابراین به‌منظور اینکه انواع ویژگی متفاوت را استخراج‌کنیم از ۱۲ کرنل 3D متفاوت استفاده می‌کنیم و آن‌ها را در ورودی اعمال می‌کنیم

که این امر باعث تولید ۱۲ نقشه ویژگی در لایه C1 می‌شود. سایز هرکدام از نقشه‌های ویژگی 30×30 5×30 است.

بعدازآن یک کرنل 2D با سایز 3×3 برای لایه کانولوشنی بعدی اعمال می‌شود. سپس یک عملگر ادغام[۷۷] بر روی نتایج هر کانولوشن زمان-مکانی اعمال می‌شود. در لایه Pooling بانام S1 در هر یک از نقشه ویژگی در لایه C2 عملیات نمونه‌برداری با یک فاکتور مکانی 2×2 انجام می‌پذیرد که باعث کاهش ابعاد مکانی با همان تعداد نقشه ویژگی می‌شود و باعث به وجود آمدن نقشه‌های ویژگی می‌شود که در برابر اعوجاج‌های کوچک مستحکم‌تر است. به‌منظور تولید مجموعه دیگری از نقشه‌های ویژگی، کانولوشن های زمان-مکانی دیگری را در لایه‌های عمیق‌تر بر روی نقشه‌های ویژگی اعمال می‌کنیم.

لایه کانولوشنی C3 با اعمال کانولوشن با یک کرنل 3D با سایز $3 \times 3 \times 3$ بر روی نقشه ویژگی‌های S1 به دست می‌آید. لایه S2 همانند لایه S1 عملیات مشابهی را انجام می‌دهد. لایه C4 کانولوشن $3 \times 3 \times 3$ را اعمال می‌کند که ۶۴ نقشه ویژگی را نتیجه می‌دهد. بعد از اعمال سه لایه کانولوشن زمان-مکانی، بعد زمانی نقشه‌های ویژگی به‌دست‌آمده به یک کاهش پیدا می‌کند. به دنبال لایه C4 یک‌لایه کاملاً متصل C5 قرار دارد. در لایه C5 یک کانولوشن 2D به‌منظور به دست آوردن ویژگی‌های پیچیده سطح بالا اعمال می‌شود. سایز کرنل کانولوشن در این لایه 4×4 است، بنابراین سایز نقشه ویژگی‌های خروجی این لایه به 1×1 کاهش پیدا می‌کند که هرکدام از آن‌ها به‌تمامی ۶۴ نقشه‌های ویژگی در لایه C4 متصل می‌شود. نقشه‌های ویژگی خروجی لایه C5 با یکدیگر ادغام می‌شود که به هر واحد بردار ویژگی کاملاً متصل[۷۸] گفته می‌شود. خروجی این لایه ۲ واحد است که مطابق با تعداد انواع رفتارها در ویدیو است (رفتار نرمال و غیر نرمال) و هریک از این واحدها احتمالی از یک رفتار را نمایش می‌دهد.

[۷۷] Pooling

[۷۸] Fully Connected

از مجموعه ویژگی‌های تولیدشده در لایه C5 که ابعاد آن ۶۴ است در فاز آزمایش استفاده می‌شود که در ادامه توضیح داده خواهد شد.

۳-۲-۶- متد اوزان تکرارشونده در تجزیه ماتریس‌های نامنفی (IW-NMF)

درصورتی‌که ویژگی‌ها به‌صورت $(x_1,\ldots,x_n)$ در یک ناحیه از سکانس‌های ویدیویی استخراج شوند یک لغت‌نامه از الگوی نرمال مانند $W \in R^{p\times q}$ با اعمال تنک‌سازی آموزش می‌بینند. به‌منظور تنک‌سازی و آموزش ویژگی‌ها روشی را با عنوان اوزان تکرارشونده در تجزیه ماتریس‌های نامنفی[۷۹] ارائه کرده‌ایم که در این بخش موردبررسی قرار می‌گیرد. درروش NMF کلاسیک داریم

$$X \approx W \times H \qquad (۵-۳)$$

که $X \in R^{p\times n}$ یک ماتریس از ویژگی‌هاست که هر ویژگی p بعد دارد و n نیز درواقع تعداد ویژگی‌هایی است که تشکیل یک مجموعه را داده‌اند، $W \in R^{p\times q}$ به آن بردارهای پایه‌ای گفته می‌شود که در هر ستون قرار دارد و $H \in R^{q\times n}$ شامل بردارهای ضرایب است. در این‌گونه روش‌ها می‌بایست $q < p$ باشد که عملیات تنک‌سازی و کاهش بعد انجام پذیرد. نکته قابل‌توجه این است که ماتریس‌های X,W,H همگی نامنفی هستند و ستون‌های W (بردارهای پایه‌ای) نیز نرمال‌سازی شده‌اند. روش ارائه‌شده بر مبنای بهینه‌سازی تابع هزینه طراحی‌شده است تا بتواند تنک‌سازی را به‌صورت مؤثرتری انجام دهد. در فرم اوزان تکرارشونده که با یک پارامتر کنترل می‌شود مدل ارائه‌شده به‌صورت زیر تعریف می‌گردد.

$$X \approx W \times S \times H \qquad (۶-۳)$$

ماتریس‌های X,W,H مانند ماتریس‌های روش NMF کلاسیک تعریف می‌شوند که در مورد آن‌ها توضیح داده شد. ماتریس متقارن $S \in R^{q\times q}$ درواقع یک ماتریس هموار کننده[۸۰] است. الگوریتم ارائه‌شده در

[۷۹] Iterative Weighted Non-negative Matrix Factorization

[۸۰] Smoothing matrix

این روش که در هر دو ماتریس پایه و ضرایب عمل تنک‌سازی را انجام می‌دهد در زیر نمایش داده‌شده است.

الگوریتم پیشنهادی IW-NMF

Input: *feature set* $X = [x_1, \ldots, x_n]$. *Initialize W, H and* max_{iter}.
Repeat until convergence:
For $a = 1 \ldots q$ ***do begin***
For $b = 1 \ldots n$ ***do***

$$H_{ab} = H_{ab} \frac{\sum_{i=1}^{p}((W_{ia}S)X_{ib}) \Big/ \sum_{k=1}^{q}(W_{ik}S)H_{kb}}{\sum_{i=1}^{p}(W_{ia}S)} \qquad (۷\text{-}۳)$$

For $c=1 \ldots p$ ***do begin***

$$W_{ca} = W_{ca} \frac{\sum_{j=1}^{n}\left((H_{aj}S)X_{cj}\right) \Big/ \sum_{k=1}^{q}(W_{ck}S)H_{kj}}{\sum_{j=1}^{n}(SH_{aj})} \qquad (۸\text{-}۳)$$

$$W_{ca} = \frac{W_{ca}}{\sum_{j=1}^{n} W_{ja}} \qquad (۹\text{-}۳)$$

End
End

ماتریس متقارن مثبت S که به‌عنوان ماتریس هموار کننده عمل می‌کند به‌صورت زیر تعریف می‌شود.

$$\boldsymbol{S = (1 - iter_{\theta_i})I + \frac{iter_{\theta_i}}{q}11^T} \qquad (۱۰\text{-}۳)$$

$iter_{\theta_i}$ پارامتر هموارکننده است که در هر تکرار توسط رابطه زیر به دست می‌آید.

$$\boldsymbol{iter_{\theta i} = i\left(\theta / Max_{iter}\right).} \quad \boldsymbol{i = 0, 1, \ldots, Max\_iter} \qquad (۱۱\text{-}۳)$$

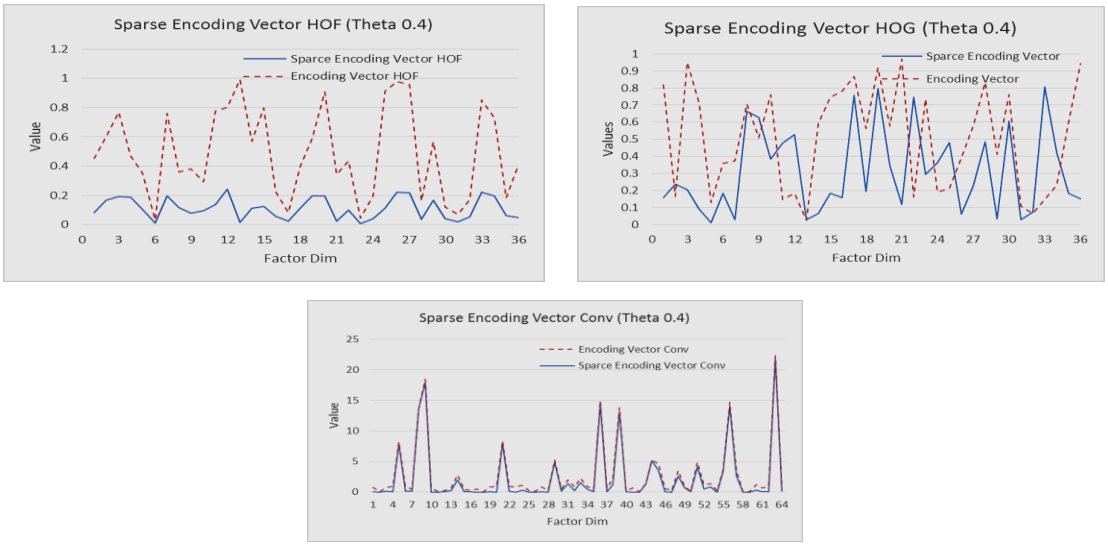
I ماتریس واحد و 1 برداری که تمام المان‌هایش یک است، پارامتر Θ پارامتر ثابت هموار کننده است که در بازه $0 < \theta < 1$ تغییر می‌کند، i و max_{iter} به ترتیب شماره تکرار و بیشترین تعداد تکرار الگوریتم می‌باشد.

همان‌طور که اشاره شد، $iter_{\theta_i}$ در بازه $(0,\theta)$ در تکرارها تغییر می‌کند که این تغییر باعث تغییر در ماتریس S نیز می‌شود. نکته‌ای که باید به آن اشاره داشت این است که روش ارائه‌شده در تکرارهای الگوریتم نقاط کلیدی و متمایزکننده را حفظ می‌کند و آن را قدرت می‌بخشد و بالعکس آن ویژگی‌هایی که ضعیف و یا نویز هستند را در این تکرارها ضعیف‌تر و یا حذف می‌کند.

بعد از استخراج HOG,HOF آن‌ها را در مجموعه‌های جداگانه قرار می‌دهیم و به دلیل اینکه هر دو دارای مقادیر نامنفی هستند می‌توانیم از این روش ارائه‌شده برای آموزش لغت‌نامه الگوی نرمال W بهره ببریم. در حالت آزمایش ویژگی‌های جدید $X^{36\times N}$، را با المان‌های تنک شده $W^{36\times 36}$ و $H^{36\times N}$ بازسازی می‌کنیم.

$$min_H \|X - WH\|_2^2 \rightarrow \begin{cases} R^{HOG} = W\_HOG \times H\_HOG \\ R^{HOF} = W\_HOF \times H\_HOF \\ R^{CNN} = W\_CNN \times H\_CNN \end{cases} \qquad (۱۲-۳)$$

R^{HOG} و R^{HOF} و R^{CNN} به ترتیب ویژگی‌های بازسازی‌شده‌ی توصیف‌کننده‌های HOG و HOF و شبکه CNN ارائه شده هستند. این ویژگی‌های بازسازی‌شده در مقابل نویز مقاوم و همچنین قابلیت تمایز بیشتری دارند، بنابراین بهتر می‌توان آن‌ها را جداسازی کرد. برای نمونه مقایسه بین بردار ضرایب معمول و تنک شده در تصویر ۳-۱۵ نمایش داده‌شده است.



تصویر ۳-۱۵مقایسه‌ای بین بردار ضرایب معمولی و تنک شده

۳-۲-۲- شناسایی اعمال غیرمعمول در مقادیر ویدیویی

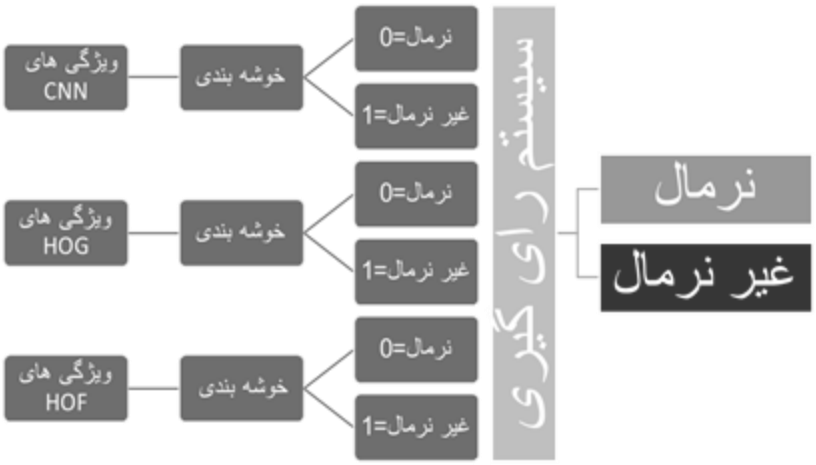
برای شناسایی و تصمیم‌گیری در مورد اینکه یک سلول فعال شامل رفتار غیرعادی است یا خیر، علاوه بر اینکه از خوشه یابی ویژگی‌های بازسازی‌شده R^{CNN} که توسط مدل CNN برای هر سلول تولیدشده، بهره می‌برد، برای اخذ نتیجه مطمئن‌تر، از خوشه یابی در ویژگی‌های بازسازی‌شده R^{HOG} و R^{HOF} استفاده کرده‌ایم.

برای به دست آوردن ویژگی‌های بازسازی‌شده در فاز آزمایش می‌توان از بردارهای پایه‌ای آموزش‌دیده شده $W_{trained}$ در روابط زیر استفاده نمود.

$$H_{test} = inv(W_{trained}) \times X$$
$$WH_R = W_{trained} \times H_{test} \qquad (۱۳-۳)$$

که WH_R ویژگی بازسازی‌شده است، $inv(W_{trained})$ ماتریس معکوس $W_{trained}$ ، X و H_{test} به ترتیب ویژگی جدید دیده نشده و ماتریس بردار ضرایب ویژگی‌های جدید است. حال که WH_R به دست آمد توسط الگوریتم K-means [۸۷] خوشه یابی می‌شود.

در خوشه یابی از الگوریتم K-means که شامل ۲ خوشه است استفاده کرده‌ایم و به طبع تعداد خوشه‌ها می‌بایست با تعداد حالت‌های تصمیم‌گیری (رفتار نرمال و غیر نرمال) برابر باشد. هر یک از خوشه یاب های K-means ویژگی جدید را برچسب‌گذاری می‌کنند. حال با پیش رو داشتن سه تصمیم گیر که شامل خوشه یابی ویژگی‌های R^{CNN} ، خوشه یابی مربوط به R^{HOG} و خوشه یابی مربوط به R^{HOF} است، می‌بایست یک تصمیم جامع را اتخاذ کنیم.



تصویر ۳-۱۶- نمایش جزئیات مربوط به بلوک خوشه یابی. هر یک از مجموعه ویژگی‌های در سه خوشه نرمال و غیر نرمال خوشه یابی شده‌اند. سیستم رأی‌گیری تصمیم می‌گیرد که این سلول مکعبی شامل رفتار غیر نرمال است یا خیر.

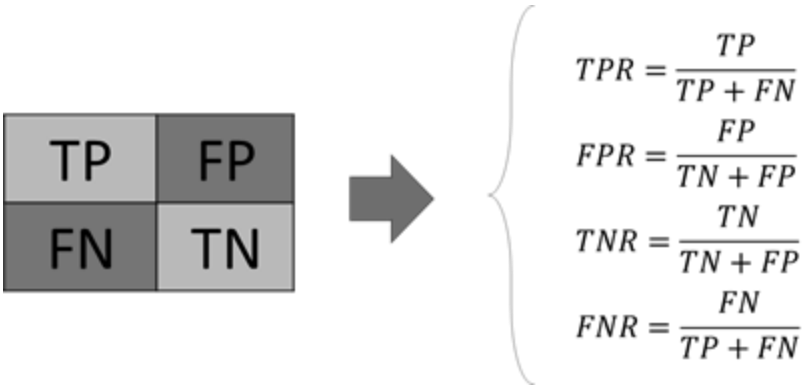
بنا بر مطالعات انجام‌شده می‌توان از یک سیستم رأی‌گیری ساده در این مورد بهره برد، به‌طوری‌که اگر حداقل ۲ تصمیم گیر از ۳، به غیر نرمال بودن یک سلول فعال رأی دهند آن سلول به‌عنوان سلول غیرعادی تلقی می‌گردد و بالعکس، که در تصویر ۳-۱۶ نشان داده‌شده است.

فصل چهارم

۴- ارزیابی و نتیجه گیری

۴-۱- مقدمه

روش ارائه‌شده در این پایان‌نامه با روش‌های متداول پیشین بر روی مجموعه داده‌های UCSD [۸۸] و UMN [۸۹] که در ادامه در مورد آن‌ها توضیحاتی داده خواهد شد، مورد مقایسه قرارگرفته است. این مقایسه با استفاده از معیارهای منحنی ROC[۸۱] و نرخ ERR[۸۲] ارزیابی و نتیجه‌گیری شده است. منحنی ROC یک نمودار پراکندگی از حساسیت[۸۳] برای یک سیستم طبقه‌بندی کننده‌ی باینری است که آستانه‌ی آن متغیر است. برای تشکیل نمودارهای ROC به دو طبقه‌بندی واقعی و طبقه‌بندی پیش‌بینی‌شده بر اساس مدل نیاز است. این طبقه‌بندی‌ها منجر به تشکیل یک جدول 2×2 می‌شود که بر اساس این جدول آماره‌هایی تعریف می‌شوند که عبارت‌اند از:



تصویر ۴-۱- آماره‌های موردنیاز برای ترسیم نمودار ROC

برای ترسیم نمودار ROC باید دو محور افقی فرونشست (FPR) و محور عمودی حساسیت (TPR) تعریف شوند. هر چه نقاط به سمت بالا و چپ نزدیک‌تر باشد مناسب‌تر است و مدل پیش‌بینی به حالت ایده آل خود نزدیک‌تر است. درصورتی‌که مدل به‌صورت تصادفی عمل کند نقاط در اطراف خط $TPR = FPR$ قرار می‌گیرند. مساحت زیر منحنی‌های ROC بیانگر عملکرد خوب مدل در تمایز بین دو گروه تشخیص است. بر اساس مساحت بین منحنی ROC و خط $TPR = FPR$ قدرت منحنی‌های ROC

[۸۱] Receiver Operation Characteristic
[۸۲] Equal Error Rate
[۸۳] Sensitivity

مشخص می‌شود. هر چه این مساحت کمتر شود نشان می‌دهد که مدل پیش‌بینی قدرت ضعیف‌تری دارد و هرچه مساحت افزایش یابد نشان از قدرت مدل پیش‌بینی و نزدیکی آن به مدل حقیقی است. نرخ خطای برابر (EER) به‌صورت آستانه برابری این دو نرخ FPR=FNR تعریف می‌شود و غالباً به‌عنوان یک ویژگی نشان‌دهنده کارایی سیستم مطرح می‌گردد. مقدار کاهشی مربوط به مقدار EER نمایانگر افزایش کارایی دقت در سیستم است.

به‌منظور تعیین فریم‌های غیر نرمال از دو سطح فریم[۸۴] و سطح پیکسل[۸۵] بهره گرفته‌شده است. این دو مقدار اندازه‌گیری به‌صورت زیر تعریف می‌شوند.

الف) اندازه‌گیری در سطح فریم : اگر یک پیکسل از هر فریم به‌عنوان غیرمعمول شناخته شود، آن فریم به‌عنوان یک فریم غیر نرمال محسوب می‌شود.

ب) اندازه‌گیری در سطح پیکسل: اگر حداقل ۴۰٪ از پیکسل‌های زمینه درست[۸۶] توسط پیکسل‌هایی که الگوریتم به‌عنوان غیرمعمول شناخته است، پوشش داده شود، فریم مربوطه را به‌عنوان فریم غیر نرمال محسوب می‌کنیم.

۴-۲- مجموعه داده‌ها[۸۷]

به‌منظور آموزش مدل CNN و ارزیابی روش ارائه‌شده در این پایان‌نامه از دو مجموعه داده‌یUCSD و UMN که تقریباً در تمامی مقالات ارائه‌شده در این حوزه استفاده‌شده است، بهره گرفته‌ایم. در ادامه به معرفی این مجموعه داده‌ها پرداخته می‌شود.

---

[۸۴] Frame level

[۸۵] Pixel level

[۸۶] Ground Truth

[۸۷] Datasets

۴-۲-۱- مجموعه داده UCSD

این مجموعه داده شامل ۲ زیرمجموعه بانام‌های Peds1 , Peds2 است که از دو فضای باز متفاوت گرفته‌شده است. در حالت نرمال عابران پیاده در محل پیاده‌روی حرکت می‌کنند ولی در حالت غیر نرمال وسایل نقلیه مانند خودرو، دوچرخه، ویلچر و یا عابران پیاده از محل چمن‌کاری شده اطراف پیاده راه از محل عابر پیاده عبور می‌کنند. هر دو زیرمجموعه توسط یک دوربین ایستا در ۱۰ فریم بر ثانیه به ترتیب با رزولوشن 234×158 و 360×240 ضبط‌شده است. فایل‌های زمینه درست در آن اجازه ارزیابی سطوح فریم و پیکسل را می‌دهد.

در زیرمجموعه Peds1 از ۳۶ ویدیو برای آزمایش و ۳۴ ویدیو برای آموزش و در زیرمجموعه Peds2 از ۱۶ ویدیو برای آزمایش و ۱۲ ویدیو برای آموزش استفاده‌شده است.

۴-۲-۲- مجموعه داده UMN

در مجموعه داده UMN سه صحنه متفاوت وجود دارد. در حالت نرمال عابران پیاده در حالت پیاده‌روی یا حالت ایستاده هستند اما در حالت غیر نرمال عابران پیاده به اطراف باحالت وحشت فرار می‌کنند. ویدیوها در سه سناریو مختلف بدون لرزش و حرکت دوربین ضبط‌شده است. این مجموعه از ویدیوها در ۳۰ فریم بر ثانیه گرفته‌شده است.

با توجه به اینکه در این مجموعه داده فایل‌های زمینه درست وجود ندارند پس لزوماً می‌بایست از روش اندازه‌گیری در سطح فریم استفاده نمود. در تصویر ۴-۲ نمونه فریم‌هایی از این مجموعه داده‌ها به نمایش گذاشته‌شده است.

تصویر ۴-۲- نمونه تصاویر مربوط به مجموعه داده‌های مورداستفاده در این پایان‌نامه. ردیف اول و ردیف دوم به ترتیب نمونه تصاویر مربوط به Peds1 و Peds2 از مجموعه داده UCSD، ردیف سوم نمونه تصاویر مربوط به مجموعه داده UMN.

۴-۳- نحوه ارزیابی و نتایج اخذشده

همان‌طور که اشاره شد روش ارائه‌شده در این پایان‌نامه در قیاس با متدهایی که در زمینه شناسایی رفتارهای غیر نرمال معرفی‌شده‌اند مورد ارزیابی قرار می‌گیرد [۲۹, ۳۱-۳۴, ۳۹, ۸۳, ۹۰, ۹۱]. به‌منظور ارزیابی متد ارائه‌شده در مجموعه داده UCSD دو معیار برای ارزیابی میزان دقت شناسایی رفتارهای غیرمعمول، معیار سطح پیکسل و سطح فریم، مورداستفاده قرارگرفته است.

معیار سطح فریم تنها بر روی تغییرات متمرکز است که پیش‌بینی می‌کند کدام فریم شامل رفتار غیرمعمول است بدون اینکه محل وقوع آن را مشخص کند. معیار سطح فریم، یک فریم را چنانچه شامل حداقل یک رفتار غیرعادی باشد، فریم غیر نرمال می‌داند و به جایگاه وقوع رفتار غیر نرمال هیچ حساسیتی ندارد. در سویی دیگر معیار سطح پیکسل معیاری است که موقعیت زمان-مکانی در فریم را مشخص می‌کند. همان‌طور که اشاره شد اگر حداقل ۴۰٪ از پیکسل‌های زمینه درست توسط پیکسل‌هایی که الگوریتم به‌عنوان رفتار غیرمعمول شناخته است، پوشش داده شود آن فریم را به‌عنوان فریم غیرمعمول شناسایی می‌کند. در ادامه با محاسبه نرخ‌های True Positive Rate (TPR) و False Positive Rate (FPR) می‌توان معیار ROC برای ارزیابی کارایی الگوریتم را به دست آورد.

الگوریتم ارائه‌شده با زبان پایتون پیاده‌سازی شده است. در این پیاده‌سازی از کتابخانه Keras به همراه توابع Theano بهره برده‌ایم که کتابخانه Keras نیز با پایتون توسعه‌یافته است. نتایج اخذشده بر روی یک پردازنده i5 با ۸ گیگ رم و توسط پردازنده گرافیکی Nvidia 1050Ti به‌دست‌آمده است.

مدل ارائه‌شده بر روی مجموعه داده UCSD Peds1 به دلیل مناسب بودن تعداد داده‌ها برای تعریف یک عمل نرمال بسیار کارا عمل می‌کند. این مجموعه رفتارهای متنوع پیاده‌روی، از پیاده‌روی یک فرد گرفته تا پیاده‌روی گروهی، را شامل می‌شود. علاوه بر این رفتارهای غیر نرمال کمی در داده‌های آموزش وجود دارند. مجموعه داده UCSD Peds2 به دلیل اینکه اندازه اشیای غیر نرمال در طول سکانس تقریباً ثابت مانده است، از خطای کمتری نسبت به مجموعه‌ی قبلی برخوردار است.

جدول ۴-۱- نرخ EER در زیرمجموعه Peds1 از مجموعه داده UCSD

نام نویسنده مقالات	نرخ EER در سطح فریم	نرخ EER در سطح پیکسل
Cheng et al. [۸۳]	۱۹/۹	۳۸/۸
Zhu et al. [۳۲]	۱۵	-
Adam [۲۹]	۳۲/۶	۳۸/۹
Kim [۳۳]	۱۹/۶	۳۹/۶
Li et al. [۳۴]	۲۳/۵	۳۷/۸
Kaltsa [۳۱]	۲۱/۱	۲۷
روش ارائه‌شده	۲۰/۱	۲۷/۵

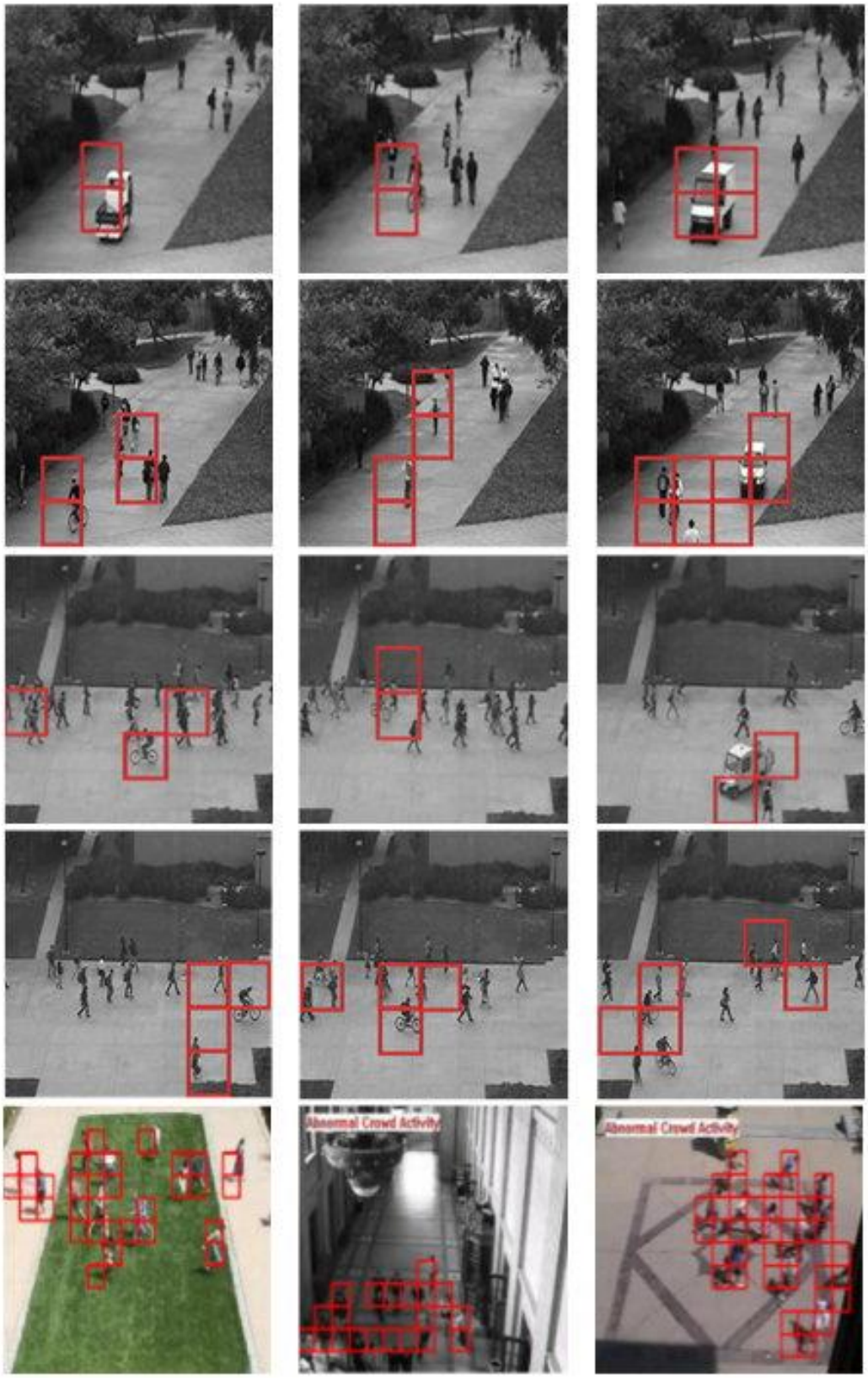
نتایج مربوط به Peds1 و Peds2 در مجموعه داده UCSD در جدول ۴-۱ و جدول ۴-۲ نمایش داده‌شده است. در این جداول میزان نرخ EER در الگوریتم ارائه‌شده با دیگر پژوهش‌های متداول پیشین به مقایسه گذاشته‌شده است. هر چه نرخ EER در روش‌های مقایسه شده کمتر باشد بهبود کارایی آن روش را نشان می‌دهد.

جدول ۴-۲- نرخ EER در زیرمجموعه Peds2 از مجموعه UCSD

نام نویسنده مقالات	نرخ EER در سطح فریم	نرخ EER در سطح پیکسل
Li et al. [۳۴]	۱۸/۵	۲٤/۷
Adam [۲۹]	۲۲/۴	۴۳/۸
Kim [۳۳]	۲۲/۴	۳۱/۱
Kaltsa [۳۱]	۲۵/۱	۲۶/۹
روش ارائه‌شده	۱۷/٤	۲۷/۱

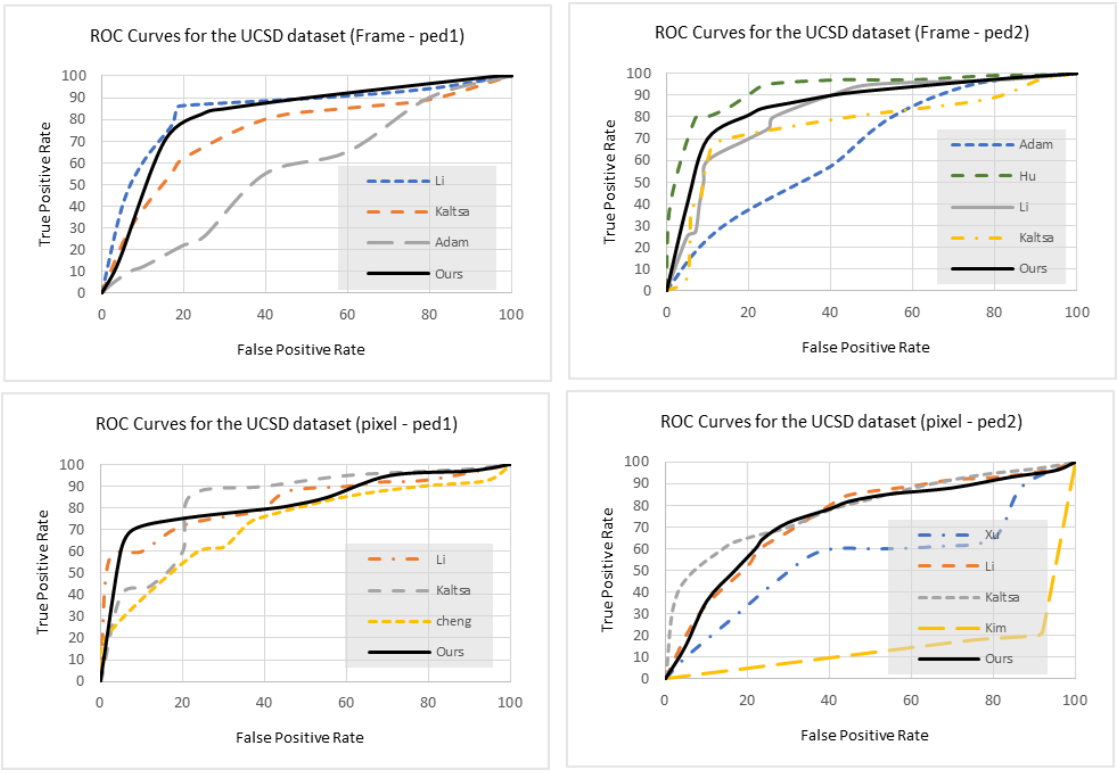
نتایجی که از دیگر متدها برای مقایسه آورده شده‌اند از مقاله‌هایی که این روش‌ها معرفی‌شده‌اند اقتباس‌شده است. در تصویر ۴-۳ نتایج شناسایی رفتارهای غیرعادی آورده شده است.

تصاویر ردیف‌های اول و دوم که متعلق به مجموعه داده UCSD Peds1 است یک مسیر پیاده‌روی را نشان می‌دهد که افراد نسبت به دوربین دور و یا به آن نزدیک می‌شوند. در این مجموعه داده عبور وسایل نقلیه مانند خودرو، دوچرخه، اسکوتر و ... به‌عنوان یک رفتار غیر نرمال تلقی می‌شود و در الگوریتم ارائه‌شده توانسته است به‌خوبی این نواحی را شناسایی کند. در ردیف‌های سوم و چهارم که متعلق به مجموعه داده UCSD Peds2 است. این مجموعه مانند مجموعه مذکور قبلی دارای یک پیاده راه است که تفاوت آن با مجموعه قبلی ثابت بودن تقریبی فاصله افراد نسبت به دوربین بوده و مسیر حرکت آن‌ها از چپ به راست است. که روش پیشنهادی در تشخیص ناهنجاری‌هایی نظیر حرکت خودرو و یا دوچرخه موفقیت‌هایی را کسب نموده است. تصاویر موجود در ردیف پنجم متعلق به مجموعه داده UMN است که افراد به‌طور ناگهانی به اطراف در حال فرار هستند. در این مجموعه داده نیز روش معرفی‌شده به‌خوبی توانسته است افراد در حال فرار را شناسایی کند.



تصویر ۴-۳- نمونه‌هایی از شناسایی نواحی وقوع رفتار غیرعادی توسط روش ارائه‌شده.

منحنی‌های مربوط به ROC در مجموعه داده UCSD در تصویر ۴-۴ گزارش داده‌شده است. با توجه به تصویر زیر می‌توان مشاهده کرد که متد ارائه‌شده کارایی قابل‌رقابتی را در مقایسه با روش‌های پیشین ارائه داده است.



تصویر ۴-۴- منحنی‌های ROC در مجموعه داده UCSD در سطوح پیکسل و فریم.

به‌خصوص نتایج حاصله در سطح فریم درروش ارائه‌شده بسیار شبیه به نتایج بهترین متدهایی است که در این زمینه معرفی‌شده‌اند. در سطح پیکسل نیز متد ارائه‌شده کارایی قابل‌رقابتی را در مقایسه با دیگر روش‌ها دارد. در حالت کلی می‌توان گفت که روش ارائه‌شده در این پایان‌نامه کارایی بسیار رقابتی در مجموعه داده UCSD با نتایج متدهای دیگر را دارا است.

در مجموعه داده UMN نتایج به‌صورت AUC[۸۸] گزارش‌شده است. به دلیل اینکه در این مجموعه داده، فایل‌های مربوط به زمینه درست در سطح پیکسل موجود نیست بنابراین الگوریتم ارائه‌شده درصورتی‌که

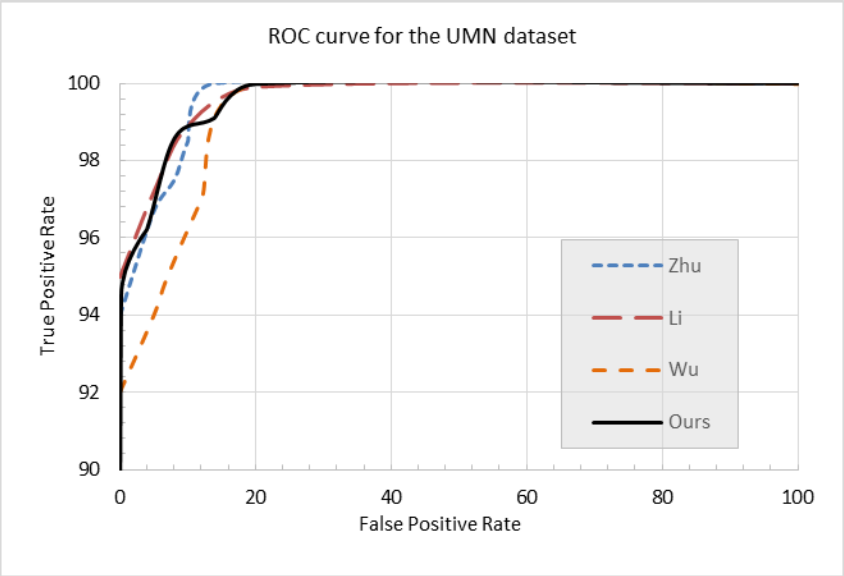
[۸۸] Area Under Curve

تنها یک ناحیه را به‌عنوان غیرمعمول شناسایی کند، کل فریم را غیرمعمول در نظر می‌گیرد. متوسط مقادیر AUC در مجموعه داده UMN در جدول ۴-۳ نمایش داده‌شده است. نتایجی که از متدهای دیگر گزارش‌شده است از مقالاتی که این روش‌ها در آن‌ها معرفی شدند گرفته‌شده است.

جدول ۴-۳- نرخ AUC در مجموعه داده UMN

نام نویسنده مقالات	نرخ AUC
Li et al. [۳۴]	۰/۹۹۶
Zhu et al. [۳۲]	۰/۹۹۷
Wu [۳۹]	۰/۹۹٤
Kaltsa [۳۱]	۰/۹۸۱
روش ارائه‌شده	۰/۹۹۵

همان‌طور که مشاهده می‌شود الگوریتم ارائه‌شده به دلیل استفاده از ویژگی های زمان-مکانی که از اعمال کانولوشن سه بعدی در تصاویر بدست آمده است و همچنین بهره گیری از روش تنک سازی اطلاعات حاصل از روش های دستی و ترکیب این دو روش با یکدیگر، نتایج قابل‌توجهی را نسبت به رقبای خود کسب نموده است. همچنین گزارشی از منحنی ROC در سطح فریم نیز در مجموعه داده UMN ارائه‌شده است. این نتایج در تصویر ۴-۵ قابل‌مشاهده است. همان‌طور که مشاهده می‌شود روش ارائه‌شده از بسیاری از روش‌های پیشین کارایی بهتری را داراست و نتایج کاملاً قابل‌رقابتی را در مقایسه با متدهایی که تنها برای کارایی طراحی‌شده‌اند رقم‌زده است.



تصویر ۴-۵- منحنی‌های ROC در مجموعه داده UMN

۴-۴- جمع‌بندی

در این پایان‌نامه روشی برای شناسایی رفتارهای ناهنجار در تصاویر ویدئویی بخصوص درصحنه‌های پیچیده و شلوغ ارائه‌شده است. این روش ارائه‌شده مجموعه ویژگی‌هایی را بر مبنای اطلاعات Optical flow ، نقاط ویژگی الگوریتم FAST و شبکه عمیق CNN استخراج می‌کند.

این روش از یک مدل CNN زمان-مکانی نوین برای شناسایی و موقعیت‌یابی رفتارهای ناهنجار درصحنه‌های مختلف ویدئویی بهره برده است. مدل زمان-مکانی CNN به‌منظور تولید ویژگی‌ها در دو بعد زمانی و مکانی توسط کانولوشن های زمان-مکانی طراحی‌شده است. در این روش تصاویر ویدئویی به مکعب‌های غیر همپوشان تقسیم می‌شود و هر یک از این مکعب‌های زمان-مکانی که دارای اطلاعات حرکتی غنی‌تری باشد به‌عنوان ورودی به مدل CNN داده می‌شود. مدل زمان-مکانی CNN به‌منظور ساخت ویژگی‌های مقاوم در دو بعد زمان و مکان با اعمال کانولوشن سه‌بعدی طراحی‌شده است، بنابراین ویژگی‌های ظاهری به‌خوبی اطلاعات حرکتی موجود در فریم‌های متوالی استخراج می‌شوند.

برای به دست آوردن مکعب‌های غیر هم‌پوشان که حاوی اطلاعات حرکتی مؤثر هستند ابتدا نقشه Optical flow مربوطه را به دست می‌آوریم. این نقشه حاوی اطلاعات حرکتی پیکسل‌های موجود در تصاویر ویدئویی است و به‌منظور استخراج نواحی که اطلاعات حرکتی در آن پررنگ‌تر است ابتدا این نقشه را به فضای تصویر HSV تبدیل می‌کنیم و با استفاده از متد GMM توسط دو گوسی آستانه مربوط به باینری کردن تصویر را به دست می‌آوریم. توسط این آستانه تصویر مربوطه را به یک تصویر باینری بدل می‌کنیم. حال می‌توان نواحی که اطلاعات حرکتی مؤثری را دارا هستند، مشاهده نمود. این نواحی به‌دست‌آمده نیز توسط مکعب‌های غیر همپوشان تقسیم می‌شوند.

با توجه به این‌که نواحی به‌دست‌آمده می‌توانند تحت تأثیر عواملی دارای خطا باشند، از روش شناساگر FAST کمک گرفته‌شده است تا میزان خطای شناسایی این نواحی به طرز محسوسی کاهش یابد.

FAST یک شناساگر باینری است که نقاط موردنظر را با مقایسه شدت یک پیکسل با همسایه‌هایش شناسایی می‌کند که اگر شدت تمامی پیکسل‌های همسایه بیشتر یا کمتر از پیکسل موردنظر باشد، آن را به‌عنوان یک نقطه موردتوجه در نظر می‌گیرد. شناساگر FAST همچنین در شناسایی نقاط از سرعت عمل بالایی برخوردار است.

شناساگر FAST بر روی سلول‌های مکعبی اعمال می‌شود. در هر نقطه‌ای که به دست می‌آید یک سلول مکعبی به مرکزیت مختصات نقطه به‌دست‌آمده در نظر گرفته می‌شود. حال می‌توان با اشتراک سلول‌های مکعبی که از باینری کردن نقشه Optical flow به دست آمد و همچنین سلول‌هایی که توسط نقاط شناساگر FAST حاصل شد، مجموعه‌ای تحت عنوان سلول‌های فعال را به دست آورد که در این پایان‌نامه به این سلول‌ها ROFIs[۸۹] گفته می‌شود. سلول‌های ROFIs درواقع همان سلول‌هایی هستند که می‌بایست برای شناسایی رفتار ناهنجار موردبررسی قرار گیرند.

[۸۹] Region Of Foreground Interests

با توجه به آنچه ذکر شد، هر یک از این سلول‌ها به‌عنوان ورودی مدل CNN برای شناسایی رفتار ناهنجار مورداستفاده قرار می‌گیرند و در ضمن به‌منظور افزایش دقت در شناسایی از توصیف گرهای HOG و HOF نیز در این سلول‌ها بهره گرفته‌شده است.

به دلیل اینکه اطلاعات استخراج‌شده توسط توصیفگرهای HOG و HOF و همچنین ویژگی‌های تولیدشده توسط مدل شبکه CNN ، از تمایز بیشتری در شناسایی رفتار نامتعارف برخوردار باشند از روش نوین IW-NMF که در این پایان‌نامه ارائه‌شده است جهت تنک‌سازی و آموزش چنین ویژگی‌هایی استفاده‌شده است.

این روش ارائه‌شده همچنان که از نام آن مشخص است بر مبنای روش تجزیه ماتریس‌های نامنفی (NMF) است، که در این روش یک ماتریس نامنفی به ضرب دو ماتریس نامنفی دیگر تبدیل می‌شود. روش IW-NMF علاوه بر اینکه ماتریس نامنفی اولیه را تجزیه می‌کند، هر دو ماتریس حاصله را نیز تنک‌سازی می‌کند. این تنک‌سازی ماتریس‌ها باعث می‌شود که نویزها و اطلاعات کم‌ارزش‌تر نادیده گرفته شوند و بالعکس اطلاعات ارزشمند تقویت گردند. این مشخصه کمک می‌کند که در طبقه‌بندی، این ویژگی‌ها قابلیت تمایز بیشتری داشته باشند.

۴-۴-۱- پیشنهادهایی برای ادامه کار

پیشنهادهایی برای استخراج ویژگی‌های زمان-مکان

برای به دست آوردن ویژگی‌های زمان-مکان می‌توان از یک الگوی پیش‌بینی کننده برای تشخیص رفتارهای ناهنجار استفاده نمود. این الگوی پیش‌بینی کننده می‌تواند توسط روش‌های مختلف آماری به دست آید. مکعب‌های نواحی موردنظر که از تصاویر فریم‌های متوالی به‌دست‌آمده است، می‌توانند توسط ترکیب یک سیستم کد کننده خصمانه و روش LSTM برای پیش‌بینی حالت بعدی وقوع رویداد در تصاویر ویدئویی مورداستفاده قرار گیرند. به‌طور خلاصه LSTM انتخاب می‌کند که کدام داده قبلی ثبت

و ضبط گردد و کدام به فراموشی سپرده شود. استفاده از این خاصیت LSTM در کد کننده خصمانه باعث می‌شود که ویژگی‌هایی با مقاومت بیشتر از تصاویر ویدئویی استخراج شود.

پیشنهادهایی برای یادگیری اطلاعات در فضای برخط

با توجه به اینکه یکی از موارد موردنیاز در سامانه‌های واقعی و برخط یادگیری این سیستم در زمان عملکرد است. پیشنهاد می‌شود که مانند روش ارائه‌شده در ابتدا به آموزش سیستم موردنظر با استفاده از تصاویر مهیا شده پرداخته شود و در ادامه سیستم بتواند در هر رویداد جدید موردبررسی خود را به‌روزرسانی کند و اطلاعات آموخته خود را ارتقا بخشد. این کار را می‌توان با بهره‌گیری از یک سیستم آموزش لغت‌نامه‌ای و استفاده از رابطه بازگشتی گرادیان نزولی محقق کرد.

این متد ارائه‌شده توسط مجموعه داده‌های UMN و UCSD که توسط دوربین‌های نظارتی تحت شرایط محیطی چالش‌برانگیزی ضبط‌شده است، مورد ارزیابی و بررسی قرارگرفته است. در ارزیابی‌های انجام‌گرفته می‌توان مشاهده کرد که کارایی این الگوریتم در برابر روش‌های دیگر بهتر و بسیار رقابتی است. این روش درواقع با یادگیری رفتارهای معمول درصحنه‌های مختلف می‌تواند هرگونه رفتاری که در آن محیط جایز نیست را شناسایی و موقعیت آن را نیز درصحنه به نمایش بگذارد.

منابع

.۱ Goldgof, D. B., Sapper, D., Candamo, J., & Shreve, M. (2009). *Evaluation of Smart Video for Transit Event Detection* (No. Report No. 2117-7807-00).

.۲ *Agent Video Intelligence*. 2017; Available from: http://www.agentvi.com./

.۳ *Honeywell*. 2017; Available from: http://www.honeywellvideo.com/products/video-systems/ias/index.html.

.٤ *IntelliVision*. 2017; Available from: https://www.intelli-vision.com/smart-security/security-surveillance./

.٥ *Ipsotek*. 2017; Available from: https://www.ipsotek.com./

.٦ *Mango*. 2017; Available from: http://mangodsp.com./

.۷ *ObjectVideo*. 20 ;۱۷Available from: http://www.objectvideo.com./

.۸ *Puretechsystems*. 2017; Available from: http://www.puretechsystems.com/video-analytics.html.

.۹ *Acic*. 2017; Available from: http://www.acic.eu/en/products/acic-activity-detection.html.

.۱۰ *VideoIQ*. 2017 ;Available from: http://www.videoiq.com./

.۱۱ *Iomniscient*. 2017; Available from: http://www.iomniscient.com./

.۱۲ Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, *104*(2), 90-126.

.۱۳ Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*(6), 976-990.

.۱٤ Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473-1488.

.۱٥ Wang, X., Farhadi, A., & Gupta, A. (2016). Actions~ transformations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2658-2667).

.۱٦ Marr, D., & Vaina, L. (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, *214*(1197), 501-524.

.۱۷ Hogg, D. (1983). Model-based vision: a program to see a walking person. *Image and Vision computing*, *1*(1), 5-20.

.۱۸ Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *CVGIP: Image understanding*, *59*(1), 94-115.

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

.۲۰ Tian, Y., Cao, L., Liu, Z., & Zhang, Z. (2012). Hierarchical filtered motion for action recognition in crowded videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(3), 313-323.

.۲۱ Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference* (Vol. 15, No. 50, pp. 10-5244).

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

.٢٣ Yilmaz, A., & Shah, M. (2005). Actions sketch: A novel action representation. In *Computer Vision and Pattern Recognition, CVPR*, (Vol. 1, pp. 984-989).

.٢٤ Shao, L., Zhen, X., Tao, D., & Li, X. (2014). Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Transactions on Cybernetics*, *44*(6), 817-827.

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

.٢٦ Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance.*(pp. 65-72).

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

.٢٨ Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition* (pp. 1-8).

.٢٩ Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, *30*(3), 555-560.

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

.٣١ Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L. J., & Strintzis, M. G. (2015). Swarm intelligence for detecting interesting events in crowded environments. *IEEE transactions on image processing*, *24*(7), 2153-2166.

.٣٢ Zhu, X., Liu, J., Wang, J., Li, C., & Lu, H. (2014). Sparse representation for robust abnormality detection in crowded scenes. *Pattern Recognition*, *47*(5), 1791-1799.

.٣٣ Kim, J., & Grauman, K. (2009). Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *Computer Vision and Pattern Recognition.* (pp. 2921-2928).

.٣٤ Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, *36*(1), 18-32.

.٣٥ Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971-987.

.٣٦ Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 915-928.

.٣٧ Sanin, A., Sanderson, C., Harandi, M. T., & Lovell, B. C. (2013). Spatio-temporal covariance descriptors for action and gesture recognition. In *Applications of Computer Vision.* (pp. 103-110).

.٣٨ Jiang, Y. G., Dai, Q., Xue, X., Liu, W., & Ngo, C. W. (2012). Trajectory-based modeling of human actions with motion reference points. *Computer Vision–ECCV* , 425-438.

.٣٩ Wu, S., Moore, B. E., & Shah, M. (2010). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR),* (pp. 2054-2060).

.٤٠ Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision,* (Vol. 1, No. 1-22, pp. 1-2).

.٤١ Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2011). Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR),* (pp. 3169-3176).

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

.٤٣ Jaakkola, T., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems* (pp. 487-493).

.٤٤ Oneata, D., Verbeek, J., & Schmid, C. (2013). Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of the IEEE international conference on computer vision* (pp. 1817-1824).

.٤٥ Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR),* (pp. 3304-3311).

.٤٦ LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

.٤٧ Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

.٤٨ Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

.٤٩ Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

.٥٠ Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

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

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

.٥٣ Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

.٥٤ Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding* (pp. 29-39).

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

.٥٦ Robinson, A. J., & Fallside, F. (1988). Static and dynamic error propagation networks with application to speech coding. In *Neural information processing systems* (pp. 632-641).

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

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

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

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

.٦١ Wu, Z., Jiang, Y. G., Wang, X., Ye, H., Xue, X., & Wang, J. (2015). Fusing multi-stream deep networks for video classification. *arXiv preprint arXiv:1509.06086*.

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

.٦٣ Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). ACM.

.٦٤ Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

.٦٥ Yan, X., Chang, H., Shan, S., & Chen, X. (2014, September). Modeling video dynamics with deep dynencoder. In *European Conference on Computer Vision* (pp. 215-230).

.٦٦ Misra, I., Zitnick, C. L., & Hebert, M. (2016). Unsupervised learning using sequential verification for action recognition. *CoRR, abs/1603.08561*.

.٦٧ Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., & Chopra, S. (2014). Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*.

.٦٨ Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

.٦٩ Bishop, C. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*.

.٧٠ Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., ... & Bengio, Y. (2010, June). Theano: A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf* (pp. 1-7).

.٧١ Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., ... & Bengio, Y. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint*.

.٧٢ Chollet, F. *keras*,. 2015; Available from: https://github.com/fchollet/keras.

.٧٣ Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *OSDI* (Vol. 16, pp. 265-283).

.٧٤ Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., & Shet, V. (2013). Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*.

.٧٥ Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

.٧٦ Shen, W., Wang, X., Wang, Y., Bai, X., & Zhang, Z. (2015). Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3982-3991).

.٧٧ Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

.٧٨ Wang, K., Wang, X., Lin, L., Wang, M., & Zuo, W. (2014). 3D human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 97-106).

.٧٩ Maturana, D., & Scherer, S. (2015). 3d convolutional neural networks for landing zone detection from lidar. In *International Conference on Robotics and Automation (ICRA),* (pp. 3471-3478).

.٨٠ Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111-126.

.٨١ Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788.

.٨٢ Liu, W., Zheng, N., & Lu, X. (2003). Non-negative matrix factorization for visual coding. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Vol. 3, pp. III-293).

.٨٣ Cheng, K. W., Chen, Y. T., & Fang, W. H. (2015). Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. *IEEE Transactions on Image Processing*, *24*(12), 5288-5301.

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

.٨٥ Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* (pp. 1446-1453).

.٨٦ Li, N., Wu, X., Xu, D., Guo, H., & Feng, W. (2015). Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. *Neurocomputing*, *155*, 309-319.

.٨٧ Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web* (pp. 1177-1178). ACM.

.٨٨ *UCSD Dataset*. Available from: www.svcl.ucsd.edu/projects/anomaly/dataset.html.

.٨٩ *UMN Dataset*. Available from: www.mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.
.٩٠ Xu, D., Song, R., Wu, X., Li, N., Feng, W., & Qian, H. (2014). Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, *143*, 144-152.
.٩١ Saligrama, V., Chen, Z. (2012). Video anomaly detection based on local statistical aggregates. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2112-2119).

Abstract

Unusual behavior detection is critically important for visual surveillance. It is also a challenging research topic in computer vision. Although much effort has been devoted to tackle this problem, such detection task in a realistic and uncontrolled environment is still far from mature. The major difficulty lies in the ambiguous characteristic in differentiating normal and abnormal behaviors, whose definitions often vary according to the context of video's history .

We propose a framework for detecting and locating abnormal activities in video sequences of crowded scenes. The key aspect of our method is the pairing of the spatial-temporal Convolutional Neural Networks (CNN) with handcrafted feature sets such as HOF and HOG for anomaly detection in contiguous video frames. Handcrafted features learned sparse by using our propose method IW-NMF based on sparse NMF. These feature extracted only from volumes of moving pixels that reduce the computational costs. The architecture of CNN model allow us to extract spatial-temporal features and using handcrafted features to ensure robustness to local noise, and increase detection accuracy. We test our framework on popular benchmark datasets containing various human abnormal activities and situations. Evaluation results show that our method outperforms most of other methods and achieves a very competitive detection performance compared to state-of-the-art methods.

Keywords:
Video anomaly detection, Convolutional Neural Networks, Machine vision, Spatial-temporal CNN, Iterative weighted non-negative matrix factorization (IW-NMF), Non-negative Matrix Factorization (NMF).

Shahrood University of Technology

Faculty of Electrical Engineering and Robotic

M.Sc. Thesis in Robotic Engineering

Abnormal behavior detection in video frames

By: Behnam Sabzalian

Supervisor:

Dr. Hossein Marvi

September 2017