مجله ماشین‌بینایی و پردازش‌تصویر
سال ششم، شماره اول، تابستان ۱۳۹۸

شناسایی رفتارهای ناهنجار در تصاویر ویدئویی با استفاده از شبکه عصبی کانولوشنی

بهنام سبزعلیان[۱]، حسین مروی[۲] و علیرضا احمدی‌فرد[۳]

چکیده

شناسایی رفتار ناهنجار ازلحاظ اهمیت یک ضرورت در سامانه‌های نظارت بصری تبدیل‌شده است. همچنین این حوزه 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 به بلوغ رسیدن آن وجود دارد. مشکل اصلی ابهام در تفاوت خصوصیات رفتار غیر نرمال و نرمال است که تعریف آن معمولاً با توجه به زمینه پیشین تصاویر می‌تواند متفاوت باشد. در این مقاله یک سیستم شناسایی و موقعیت‌یابی رفتارهای ناهنجار در سکانس‌های ویدئویی ارائه‌شده است. جنبه کلیدی این روش درواقع ترکیب شبکه عصبی کانولوشنی زمان-مکانی دوبعدی و سه‌بعدی به‌منظور شناسایی رفتار غیر نرمال در فریم‌های متوالی ویدئویی است. همچنین از روش شناساگر Features from Accelerated Segment Test (FAST) به‌منظور افزایش ضریب اطمینان در شناسایی موقعیت‌های موردنظر در تصاویر ورودی به مدل شبکه عصبی کانولوشنی بهره گرفته‌شده است. 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 که شامل استخراج ویژگی‌هایی با حرکات پیچیده نیز هست را می‌دهد.

روش ارائه‌شده توسط مجموعه داده‌ی متداول که شامل رفتارها و اعمال ناهنجار متفاوت انسانی در موقعیت‌های گوناگون است، مورد آزمایش و ارزیابی قرار گرفته است. نتایج حاصل 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 نرمال در مقایسه با روش‌های قبلی بسیار رقابتی است.

کلیدواژه‌ها

بینایی ماشین، شناسایی رفتارهای ناهنجار در تصاویر ویدئویی، شبکه‌های عصبی کانولوشنی، یادگیری ماشین، ویژگی‌های زمان-مکانی Convolutional Neural Network (CNN) .

۱ مقدمه

در طی دو دهه گذشته تشخیص و ردگیری انسان در فریم‌های متوالی ویدئو، بازنمایی و تحلیل فعالیت‌های وی و در آخر شناسایی رفتار سرزده از او یکی از پر چالش ترین مباحث درزمینه ی مطالعات بینایی ماشین و هوش مصنوعی بوده و توجه گروه‌های تحقیقاتی دانشگاه‌های معتبر فراوانی را به خود معطوف نموده است.

این مقاله در مردادماه ۱۳۹۶ دریافت، در خردادماه ۱۳۹۷ بازنگری و در مردادماه همان سال پذیرفته شد.

[۱] دانشجوی کارشناسی ارشد مهندسی رباتیک، دانشکده برق و رباتیک دانشگاه صنعتی شاهرود
رایانامه: behnamsabzalian@gmail.com

[۲] گروه رباتیک، دانشکده برق و رباتیک دانشگاه صنعتی شاهرود
h.marvi@shahroodut.ac.ir و ahmadyfard@shahroodut.ac.ir

نویسنده مسئول: بهنام سبزعلیان

با توجه به رشد فزاینده دوربین‌های امنیتی و بار هزینه‌ای زیاد نظارت انسانی، انگیزه استفاده از سامانه‌های امنیتی خودکار را افزایش می‌دهد. در واقع 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 مورد استفاده قرار گیرد همچنین می‌تواند به‌عنوان یک سامانه کمکی در حوادث ناگوار دخالت داشته باشد. هدف اصلی از این تحقیقات یافتن روشی برای شناسایی رفتار ناهنجار است.

یک رفتار ناهنجار، رفتاری است که در تصاویر ویدیویی از یک محیط مورد تحقیق تعریف‌نشده است، اما در آن اتفاق می‌افتد. در حالت کلی این رفتار توسط انسان‌ها واقع می‌گردند. 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. 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. چالش یک سیستم خودکار این است که بین رفتارهای حادث، رفتارهای نرمال از غیر نرمال را مشخص کند. 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 FP[1] کم و نرخ TP[2] 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 بعد از آن آزمایش و ارائه مجموعه روش‌هایی که می‌توانند تا حدودی بر شرایط چالش‌های طبیعی فائق آیند.

در این مقاله به‌منظور شناسایی و موقعیت‌یابی رفتارهای ناهنجار در تصاویر ویدئویی به معرفی یک مدل شبکه عصبی کانولوشنی زمان-مکانی می‌پردازیم که علاوه بر اینکه به اطلاعات بصری موجود در تصاویر ایستا دسترسی دارد، همچنین می‌تواند اطلاعات حرکتی پیچیده را نیز از فریم‌های متوالی استخراج کند. برای شناسایی رفتارهای نامتعارف در قسمت کوچکی از یک فریم، مدل زمان-مکانی شبکه عصبی کانولوشنی بر روی توده‌های زمان-مکانی[3] از تصاویر اعمال می‌شود.

این شبکه با اعمال بر روی توده‌های زمان-مکانی موردنظر، اطلاعات حرکتی مؤثر را در این توده‌ها استخراج می‌کند و مشخص می‌کند کدام به‌عنوان یک توده نامتعارف شناخته می‌شود. نتایج آزمایش‌های انجام‌شده بر روی مجموعه داده‌های متداول نمایان گر این است که روش پیشنهادی، نتایج بسیار رقابتی را در قیاس با روش‌های پیشین ارائه‌شده کسب کرده است.

این مقاله در ۶ بخش سازمان‌دهی شده است، در بخش ۲ به‌مرور پژوهش‌های پیشین می‌پردازیم، چهارچوب کلی، معرفی روش پیشنهادی و نحوه استخراج ویژگی‌ها در بخش ۳ پرداخته خواهد شد. در بخش ۴ نتایج آزمایش‌ها و مقایسه روش ارائه‌شده با روش پیشین مورد بررسی قرار می‌گیرد. بخش‌های ۵و۶ به ترتیب به جمع‌بندی و تقدیر و تشکر در این مقاله می‌پردازد

۲ مروری بر پژوهش‌های پیشین

مرجع [۱] به شناسایی ناهنجاری در حوزه‌های زمان و مکانی می‌پردازد. یکی دیگر از مشکلات در این حوزه نداشتن یک تعریف مشخص از رفتار ناهنجار است. در این تحقیق رفتارهایی که به‌ندرت در تصاویر ویدئویی اتفاق می‌افتد را به‌عنوان رفتار نامتعارف محسوب می‌کند [۲-۶]. روش‌هایی که برای شناسایی ناهنجاری درگذشته ارائه گردیده است عمدتاً بر مبنای استخراج ویژگی‌ها از مسیرهای شناسایی‌شده اشیا و یا تغییرات زمان-مکانی حادث در تصاویر ویدئویی است.

برای ذکر نمونه، روش ارائه‌شده در [۷, ۸] که بر روی مسیرهای شناسایی‌شده اشیا متمرکز است بر اساس اینکه کدام‌یک مسیری متعارف در پیش دارد برچسب نرمال و غیر نرمال بودن را به آن اختصاص می‌دهد. این روش‌ها در مواقعی که انسدادی وجود دارد و یا تغییر در روشنایی تصاویر ایجاد می‌شود کارایی خود را به‌طور محسوسی از دست می‌دهند و همچنین در مواردی که صحنه‌های شلوغی در تصاویر پدید می‌آید پیچیدگی محاسباتی بالایی دارند. بنابراین محققان روش‌هایی را ارائه کردند که از ویژگی‌های سطح پایین مانند شارنوری[4] و گرادیان استفاده کنند تا ابعاد و ارتباط زمان-مکانی را در این‌گونه ویژگی‌ها آموزش دهند.

در مقایسه‌ی روش‌های پیشین، این متدها می‌توانند بر مبنای مدل [۵, ۶, ۹] یا خوشه‌بندی بر مبنای ویژگی‌های برجسته [۴, ۱۰, ۱۱] باشند. در تحقیقات اخیر تنک‌سازی رویدادها[5] [۱۲-۱۴] در تصاویر ویدئویی به‌صورت قابل‌ملاحظه ای مورد توجه قرارگرفته است. نتایج مدل‌های ارائه‌شده در [۱۲-۱۷] حاکی از این است که کارایی قابل قبولی در شناسایی ناهنجاری داشته‌اند.

[1] False-Positive
[2] True-Positive
[3] Spatial-Temporal Patches
[4] Optical flow
[5] Sparse events

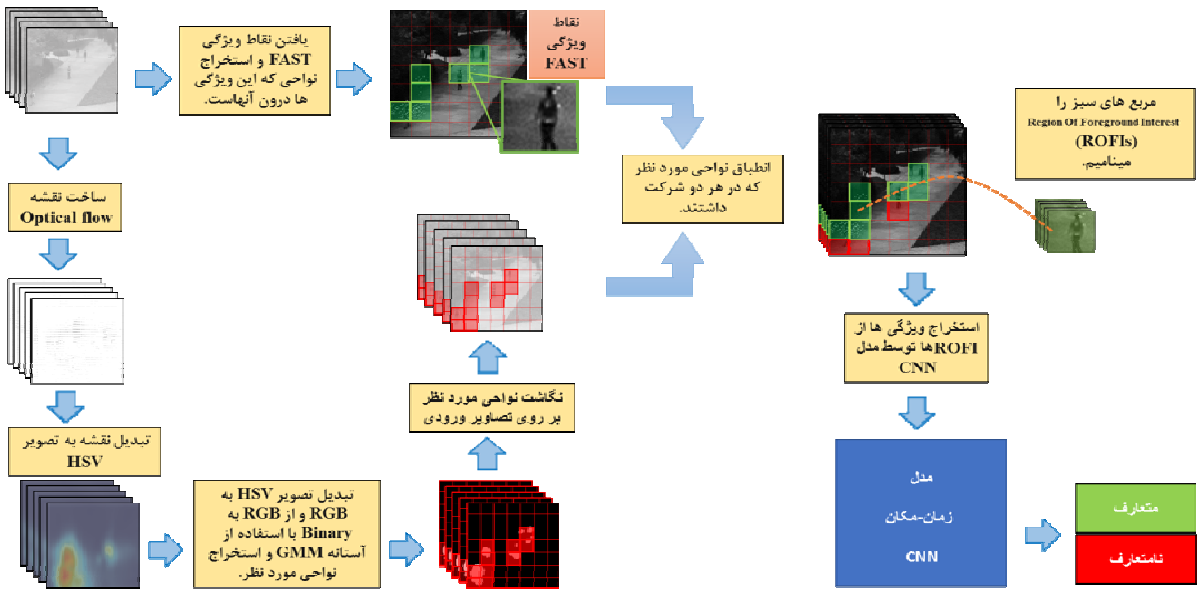
اغلب روش‌های ارائه‌شده در این زمینه از یک چهارچوب کلی برای شناسایی الگوی موردنظر پیروی می‌کنند. این چهارچوب کلی شامل دو مرحله است که مرحله اول آن محاسبه ویژگی‌های دستی از فریم‌های خام ویدئویی است و دومین مرحله شامل آموزش طبقه بندها بر مبنای ویژگی‌های استخراج‌شده است.

در حالت واقعی دانستن این مسئله که کدام ویژگی در شناسایی رفتار غیر نرمال نقش مؤثر دارد بسیار مشکل است. در حالت کلی ویژگی‌هایی اعم از ویژگی‌های زمان-مکان، گرادیان، اطلاعات شارنوری به‌منظور شناسایی موقعیتی که رویداد نامتعارفی در آن رخ‌داده است، مورد استفاده قرار می‌گیرد.

اخیراً کارهای بسیاری ارائه‌شده است که توانایی شبکه‌های عصبی کانولوشنی [۱۸] را در حوزه‌های مختلف بینایی ماشین مانند شناسایی متن [۱۹] و شناسایی طبقه‌بندی اشیا [۲۰, ۲۱] ، شناسایی لبه [۲۲] و شناسایی صورت انسان [۲۳] نشان می‌دهد.

در حوزه طبقه‌بندی ویدیو نیز شبکه‌های عصبی کانولوشنی نقش مؤثری را بازی می‌کنند. شبکه عصبی کانولوشنی سه‌بعدی در [۲۴] برای شناسایی اعمال و رفتار انسان در تصاویر ویدئویی ارائه‌شده است. روش ارائه‌شده در [۲۵] مبتنی بر توده‌های تصویری که با بهره‌گیری از یک شبکه عصبی عمیق سه‌بعدی به‌صورت آبشاری رویدادهای ناهنجار درصحنه‌های ویدئویی را شناسایی کند. اخیرا روشی در [۲۶] ارائه گردیده است که با استفاده از یک شبکه عصبی کانولوشنی سه بعدی بر مبنای Auto-Encoder سعی دارد از طریق ویژگی های بصری و حرکتی رفتارهای ناهنجار را شناسایی کند، همچنین در روشی دیگر [۲۷] از یک Auto-Encoder برای حذف نویز به‌منظور بازسازی تصاویر ورودی و آموزش ویژگی‌های موردنظر در یک شبکه عمیق استفاده‌شده است. گرچه در این مقاله از یک طبقه بند SVM یک کلاسه در لایه‌ی بالایی مدل استفاده کرده است. اما این طبقه بند به‌صورت بهینه با ویژگی‌های شبکه عمیق طراحی‌شده منطبق نشده است. روشی که در [۲۸] ارائه‌شده است به معرفی یک شبکه عصبی کانولوشنی بهبودیافته به‌منظور شناسایی رفتار انسان در تصاویر ویدئویی سه‌بعدی RGB-D می‌پردازد. روش ارائه‌شده در [۲۹] با بهره‌گیری از یک شبکه PCANet به‌منظور استخراج ویژگی‌های حرکتی از گرادیان سه‌بعدی به ساخت یک مدل ترکیبی گوسی عمیق برای مدل کردن الگوی رویدادهای نرمال پرداخته است.در مقاله [۳۰] از یک شبکه عصبی کانولوشنی سه‌بعدی برای شناسایی محل فرود یک بالگرد خودکار استفاده می‌کند.

در مقاله پیش رو از ترکیب لایه‌های کانولوشن در ابعاد دوبعدی و سه‌بعدی، به‌منظور استخراج ویژگی‌های زمان-مکانی مقاوم در برابر برخی از مشکلات نظیر تغییر روشنایی شدید، لرزش دوربین و ...، استفاده‌شده و به طراحی یک شبکه عصبی کانولوشنی نوین انجامیده است. همچنین از روش شناساگر FAST به‌منظور افزایش اطمینان در شناسایی موقعیت‌های موردنظر در تصاویر ورودی به مدل شبکه عصبی کانولوشنی بهره گرفته‌شده است. این ویژگی‌ها تنها از پیکسل‌های دارای حرکت استخراج می‌شوند تا بتوانند هزینه محاسبه را کاهش دهند.



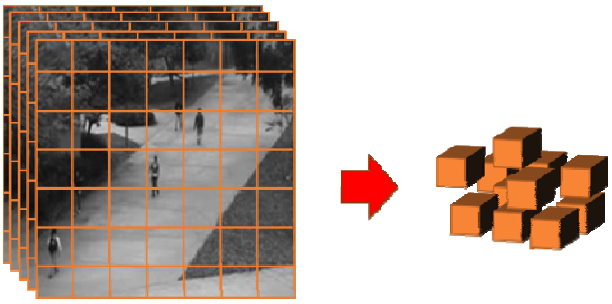
شکل ۱ شمای کلی روش ارائه‌شده در این مقاله

۳ معرفی روش پیشنهادی

شمای کلی از روش پیشنهادی در این مقاله، در شکل ۱ نشان داده‌شده است. جزئیات مربوط به این روش در ادامه شرح داده خواهد شد.

۳-۱ نحوه نمایش ویدیو

به‌منظور آماده‌سازی ویدیو ابتدا می‌بایست هر یک از فریم‌های ویدئویی را به تعدادی سلول‌های مکعبی غیر همپوشان تبدیل کنیم که در شکل ۲ نشان داده‌شده است.



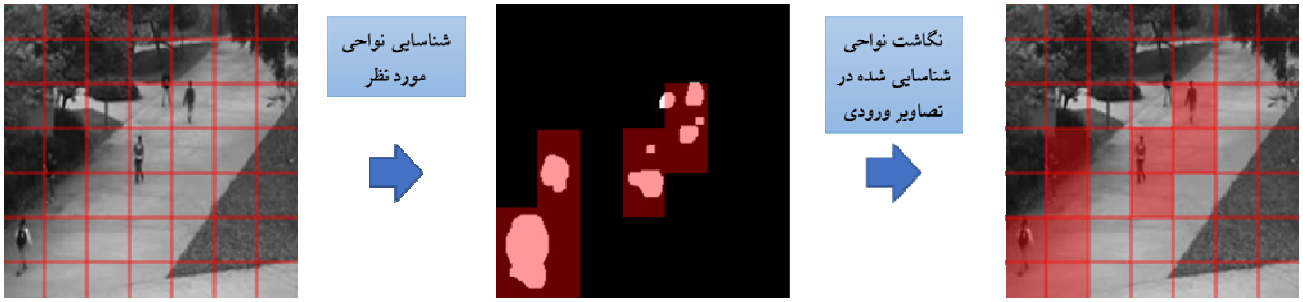
شکل ۲ نحوه نمایش ویدیو. هر یک از فریم‌های ویدیو به‌صورت تعدادی سلول‌های مکعبی غیر همپوشان تبدیل می‌شوند.

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 نواحی پیش‌زمینه و شارنوری

اطلاعات مربوط به پیش‌زمینه تصاویر برای شناسایی رفتار غیرمعمول در ویدیو به‌منظور استخراج نواحی پیش‌زمینه موردنظر Region Of Foreground Interests (ROFIs) بسیار اهمیت دارد. بنابراین نقشه شارنوری 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 پشت زمینه در این نقشه که با استفاده از روش مدل‌های مخلوط گوسی (GMM) Gaussian Mixture Model انجام می‌شود، این نقشه را به تصویر باینری تبدیل می‌کند. در این مرحله نواحی کلیدی در تصویر باینری به دست می‌آید که این نواحی در تصویر ورودی نگاشت می‌شوند. شکل ۳ چگونگی انتخاب نواحی موردنظر را در تصویر باینری نشان می‌دهد.



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

اگر هرکدام از نواحی شناسایی‌شده در تصویر باینری را به‌عنوان Bt 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 ویدئویی مکعبی مقداری ثابت دارد.

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.

$$\sum_{d=1}^{m_t}\sum_{j=1}^{m_y}\sum_{x=1}^{m_x} u(i,j,d) \qquad (۱)$$

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.

۳-۳ شناساگر ویژگی FAST و شناسایی سلول‌های فعال

از شناساگر ویژگی FAST به‌منظور اطمینان از نواحی که در بخش پیشین حاصل شد، بهره گرفته می‌شود.FAST یک متد باینری 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 [۳۱, ۳۲]. حال شناساگر FAST را بر روی تصاویر ویدئویی اعمال می‌شود. برای هر نقطه با موقعیت مکانی (x_p,y_p) در بعد زمانی سلول فعالی شناسایی می‌شود. بعد از انتخاب نقطه موردنظر یک سلول ویدئویی $V\in R^3$ با سایز m_x,m_y,m_t به مرکزیت (x_p,y_p,t_p) ساخته می‌شود. سایز m_x,m_y توسط سایز سلول فعال موردنظر تعیین می‌گردد و m_t در تمام حجم‌های ویدئویی ثابت است. مجموعه سلول‌های به‌دست‌آمده توسط این روش را با مجموعه سلول‌هایی که در بخش پیشین حاصل آمد، در نظر گرفته می‌شود و سلول‌هایی که در هر دو مجموعه ازنظر بعد مکانی با یکدیگر به میزان حداقل ۱۰% اشتراک دارند به‌عنوان یک سلول فعال و یا یک سلول ROFIs شناخته خواهند شد.

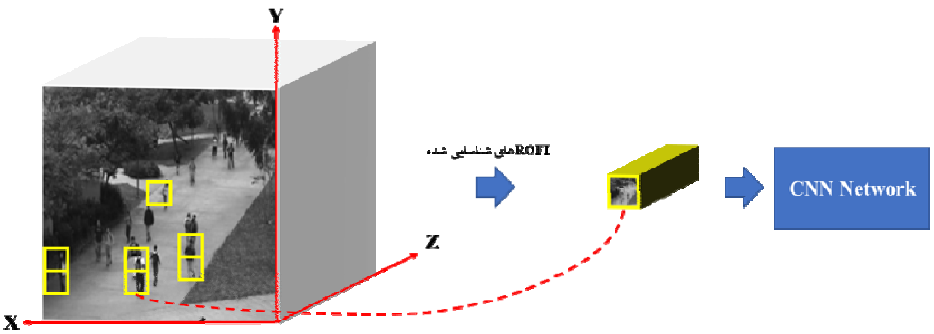
۳-۴ استخراج ویژگی CNN

برای شناسایی و موقعیت‌یابی رفتار غیرمعمول که در نواحی محلی روی می‌دهد روش‌های متفاوتی درگذشته ارائه‌شده‌اند [۱۵] و [۲۹] و [۳۰]. 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 می‌تواند از تکه سلول‌هایی که به‌صورت غیرفعال هستند صرف‌نظر کرد.

درروش پیشنهادی حاضر از شبکه کانولوشنی CNN زمان-مکانی که ارائه‌شده است به‌منظور شناسایی رفتار غیرمعمول در تصاویر ویدئویی استفاده‌شده است. در ادامه به بررسی مدل CNN و ساختار این شبکه 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.

۳-۴-۱ 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 فرایند در شکل ۴ نشان داده‌شده است.



شکل ۴ شمایی از نواحی موردنظر که به‌عنوان ورودی به مدل CNN ارائه‌شده داده می‌شود.

۳-۴-۲ کانولوشن زمان-مکان

لایه کانولوشن قسمت اصلی مدل CNN است. پارامترهای لایه شامل مجموعه‌ای از فیلترها و یا هسته[1]های 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

[1] Kernel

برازش را کنترل کند. در j امین نقشه ویژگی a_{ij} را می‌توان در i امین لایه به‌صورت زیر تعریف کرد.

$$a_i = f(W_n * a_{(i-1)n} + b_{ij}) \quad (۲)$$

که f(.) یک تابع فعال‌ساز است که به‌صورت f(x)=max(0,x) تعریف می‌شود، به آن لایه ReLU گفته می‌شود. W هسته 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 [۲۴]. کانولوشن زمان-مکانی توسط کانوالو یک هسته سه‌بعدی روی سلول‌های مکعبی زمان -مکانی به دست می‌آید. درواقع این کانولوشن سه‌بعدی توسط یک هسته 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.

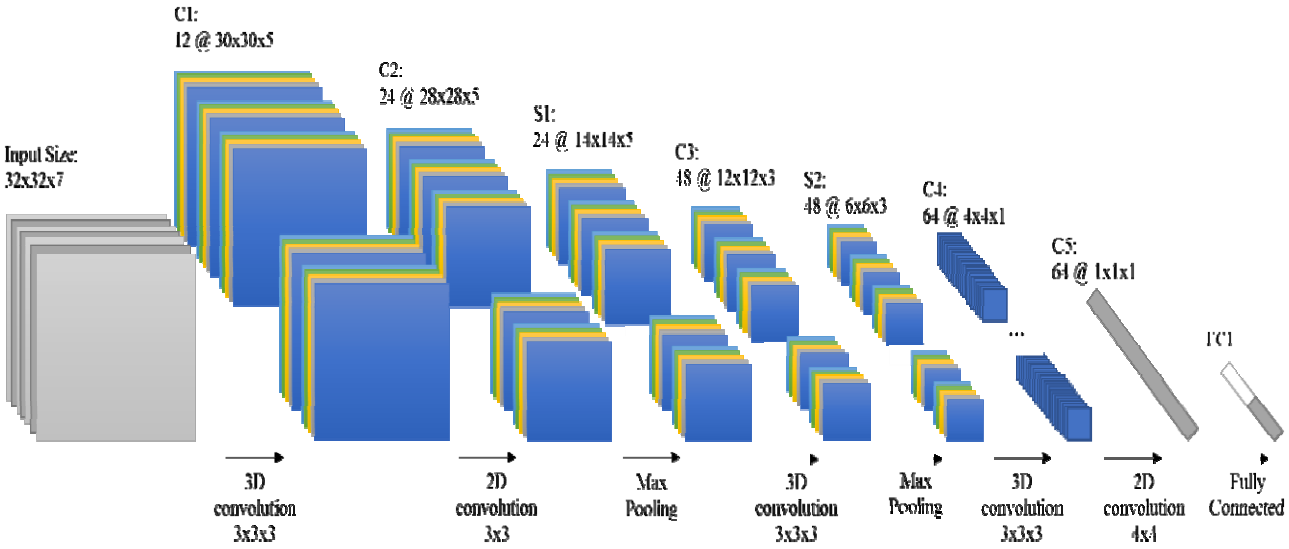
عملگر کانولوشن زمان -مکانی بین هسته سه‌بعدی W_n و سلول مکعبی زمان-مکان $a_{(i-1)n}$ را می‌توان به‌صورت زیر تعریف کرد.

$$[W_n * a_{(i-1)n}](x, y, z) \quad (۳)$$

که

$$\sum_n \sum_{u=0}^{U_i-1} \sum_{v=0}^{V_i-1} \sum_{r=0}^{R_i-1} W_n^{uvr} * a_{(i-1)n}^{(x+u)(y+v)(t+r)} \quad (۴)$$

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



شکل ۵ ساختار مدل CNN برای شناسایی رفتار غیر نرمال. این ساختار شامل سه لایه کانولوشنی سه‌بعدی ، دولایه کانولوشنی دوبعدی ، دولایه Max-pooling و یک‌لایه کاملاً متصل است.

۳-۴-۳ ساختار مدل CNN زمان-مکان

با توجه به توضیحاتی که در قسمت قبل در مورد کانولوشن سه-بعدی 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 ۳۲×۳۲ است را به‌عنوان ورودی مدل CNN سه‌بعدی در نظر می‌گیریم. این مدل را می‌توان در شکل ۵ مشاهده کرد.

ساختار این شبکه زمان-مکانی از هشت لایه تشکیل‌شده است. سایز ورودی در ساختار این شبکه به‌صورت ۷×۳۲×۳۲ است، همان‌طور که اشاره شد، از ۷ فریم متوالی به ابعاد ۳۲×۳۲ ساخته‌شده است. نواحی ROFI که در قسمت‌های قبلی به آن اشاره شد می‌بایست قبل از اینکه مورداستفاده مدل CNN قرار گیرد به ابعاد اشاره‌شده ۷×۳۲×۳۲ تغییر سایز پیدا کند.

در ابتدا کانولوشن زمان-مکان سه‌بعدی با هسته با سایز ۳×۳×۳ را بر روی‌داده‌های ورودی اعمال می‌کنیم (۳×۳ بعد مکانی و ۳ بعد زمانی هسته است). می‌بایست به این نکته توجه داشت که یک هسته سه‌بعدی می‌تواند تنها یک نوع از ویژگی را از ROFI به دست آورد، بنابراین به‌منظور اینکه انواع ویژگی متفاوت را استخراج‌کنیم از ۱۲ هسته سه‌بعدی 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 ۵×۳۰×۳۰ است.

بعدازآن یک هسته دوبعدی با سایز ۳×۳ برای لایه کانولوشنی بعدی اعمال می‌شود. سپس یک عملگر ادغام 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 ۲×۲ 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 هسته سه بعدی با سایز ۳×۳×۳ بر روی نقشه ویژگی‌های S1 به دست می‌آید. لایه S2 همانند لایه S1 عملیات مشابهی را انجام می‌دهد. لایه C4 کانولوشن ۳×۳×۳ 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 دوبعدی به‌منظور به دست آوردن ویژگی‌های پیچیده سطح بالا اعمال می‌شود. سایز هسته کانولوشن در این لایه ۴×۴ است، بنابراین سایز نقشه ویژگی‌های خروجی این لایه به ۱×۱ 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[1] گفته می‌شود. خروجی این لایه ۲ واحد است که مطابق با تعداد انواع رفتارها در ویدیو (رفتار نرمال و غیر نرمال) است و هریک از این واحدها احتمالی از یک رفتار را نمایش می‌دهد. این احتمالات توسط روابط زیر به دست می‌آید.

$$p_1 = 1/(1+\exp(u_2 - u_1)) \quad (۵)$$

$$p_1 = 1/(1+\exp(u_1 - u_2)) \quad (۶)$$

که u_i , $i=1,2$ خروجی واحد i ام در لایه FC1 است و به‌منظور نرمال‌سازی احتمالات از روش رگرسیون منطقی بهره گرفته‌شده است. در ادامه این شبکه با توجه به احتمالات موردمحاسبه در مورد اینکه کدام توده از سلول‌های فعال شامل رفتار نامتعارف می‌شود تصمیم‌گیری می‌کند.

۴ ارزیابی و نتیجه‌گیری

روش ارائه‌شده در این مقاله با روش‌های متداول پیشین بر روی مجموعه داده‌ی UCSD [۳۳] که در ادامه در مورد آن توضیحاتی داده خواهد شد، مورد مقایسه قرارگرفته است. این مقایسه با استفاده از معیارهای منحنی Receiver Operation Characteristic (ROC) و نرخ Equal Error Rate (EER) ارزیابی و نتیجه‌گیری شده است. منحنی ROC یک نمودار پراکندگی از حساسیت[2] برای یک سیستم طبقه‌بندی کننده‌ی باینری است که آستانه‌ی آن متغیر است.

به‌منظور تعیین فریم‌های غیر نرمال از دو سطح فریم[3] و سطح پیکسل[4] 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[5] توسط پیکسل‌هایی که الگوریتم به‌عنوان غیرمعمول شناخته است، پوشش داده شود، فریم مربوطه را به‌عنوان فریم غیر نرمال محسوب می‌کنیم.

۴-۱ مجموعه داده[6]

به‌منظور آموزش مدل CNN و ارزیابی روش ارائه‌شده، از مجموعه داده‌ی UCSD که تقریباً در تمامی مقالات ارائه‌شده در این حوزه استفاده‌شده است، بهره گرفته‌ایم.

این مجموعه داده شامل دو زیرمجموعه بانام‌های Peds1 , Peds2 است که از دو فضای باز متفاوت گرفته‌شده است. هر دو زیرمجموعه توسط یک دوربین ایستا در ۱۰ فریم بر ثانیه به ترتیب با رزولوشن ۱۵۸×۲۳۴ و ۲۴۰×۳۶۰ 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. در شکل ۶نمونه فریم‌هایی از این مجموعه داده به نمایش گذاشته‌شده است.

در این مجموعه داده هر یک از ROFI ها به‌عنوان یک فضای مستطیلی به ابعاد ۱۵×۱۵ انتخاب‌شده‌اند که هم برای به دست آوردن یک موقعیت غیر نرمال به‌اندازه کافی کوچک و هم برای استخراج جزئیات تصویری به‌اندازه کافی بزرگ است. طول زمانی ROFIs به‌اندازه ۷ فریم است که درواقع یک توازن بین توانایی شناسایی رفتارهای ناهنجار و فضای حافظه‌ی موردنیاز برای این رفتارهاست.

درنهایت مکعب‌های ROFIs در اندازه ۷×۱۵×۱۵ انتخاب می‌شوند. سایز انتخابی برای این مکعب‌ها بستگی به نوع محیط و نحوه قرارگیری دوربین‌ها دارد که بعد از استخراج به‌اندازه ثابت موردنیاز در شبکه تغییر می‌کند و به شبکه کانولوشنی داده می‌شود.

[1] Fully-Connected

[2] Sensitivity
[3] Frame level
[4] Pixel level
[5] Ground Truth
[6] Dataset

شکل ۶ نمونه تصاویر مربوط به مجموعه داده‌ی مورد استفاده جهت ارزیابی روش ارائه‌شده. ردیف اول و ردیف دوم به ترتیب نمونه تصاویر مربوط به Peds1 و Peds2 از مجموعه داده UCSD است.

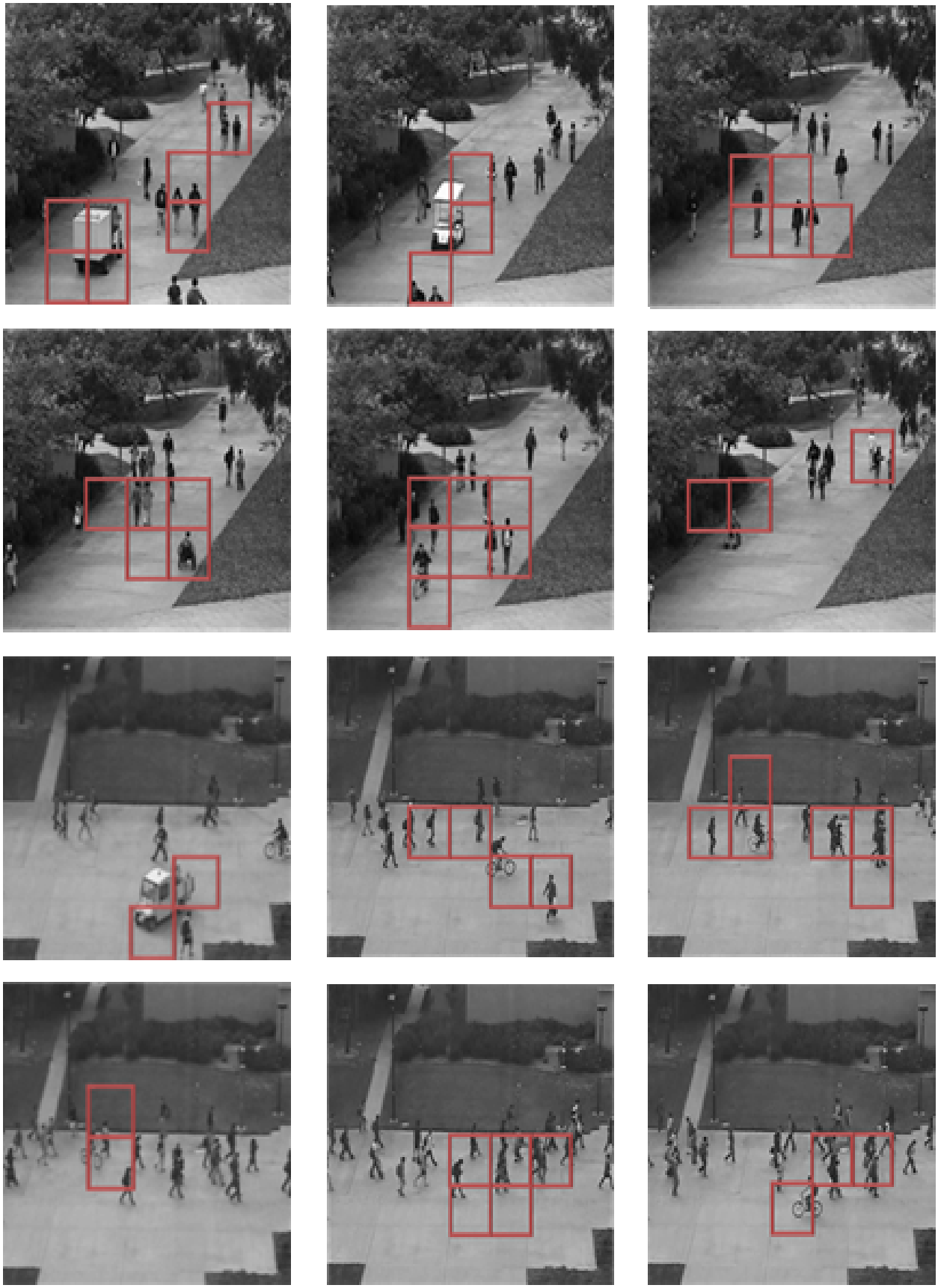
۴-۲ نحوه ارزیابی

روش ارائه‌شده در این مقاله با برخی از روش‌های متداول که تاکنون ارائه‌شده مورد ارزیابی قرارگرفته است [۲, ۸, ۹, ۱۳, ۱۵, ۳۱, ۳۴-۳۸]. به‌منظور ارزیابی روش 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 تغییراتی 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 ناحیه 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 TP و FP می‌توان معیار ROC برای ارزیابی کارایی الگوریتم را به دست آورد.



شکل ۷ نمونه‌هایی از شناسایی نواحی وقوع رفتار غیرعادی توسط روش ارائه‌شده.

۴-۳ نتایج

همان‌طور که در بخش ۴-۱ بیان شد، هر یک از مکعب‌های ROFIs در سایزی معادل ۷×۱۵×۱۵ انتخاب می‌شوند. با توجه به اینکه هیچ‌یک از کلیپ‌های آموزشی مجموعه‌های Ped1 و Ped2 شامل ناهنجاری نیست، به‌صورت تصادفی نصف کلیپ‌های آزمایشی مربوط به Ped1 و Ped2 برای استفاده در مدل آموزشی و مابقی کلیپ‌ها برای آزمایش نمونه‌ها مورداستفاده قرار می‌گیرند.

هر یک از مجموعه‌های Ped1 و Ped2 به‌طور جداگانه آموزش داده می‌شوند. تعداد ۱۴۰۲۴۸ نمونه نرمال و ۳۵۲۱۵ نمونه غیر نرمال از مجموعه Ped1 و تعداد ۶۳۵۷۹ نمونه نرمال و ۲۰۶۳۸ نمونه غیر نرمال از مجموعه Ped2 استخراج‌شده است.

با توجه به اینکه تعداد نمونه‌های غیر نرمال بسیار کمتر از نمونه‌های نرمال است، این امکان وجود دارد که مشکل عدم تعادل در کلاس به وجود آید. برای حل این مشکل با نمونه‌برداری دوباره، سعی شد تعداد نمونه‌های نرمال و غیر نرمال در یک تعادل قرار گیرند.

روش ارائه‌شده در این مقاله توسط یک رابط نرم‌افزاری تحت عنوان Keras [۳۹] که از شبکه‌های عصبی و کانولوشنی با بهره‌گیری از کتابخانه نرم‌افزارهای Theano [۴۰, ۴۱] و Tensor Flow [۴۲] پشتیبانی می‌کند، پیاده‌سازی شده است. این رابط نرم‌افزاری به زبان Python نوشته‌شده است و قابلیت اجرا بر روی CPU و GPU را داراست. روش ارائه‌شده بر روی یک کامپیوتر با CPU i5 Core(TM) Intel(R) 2.4 GHz و 8G RAM اجراشده است. شبکه کانولوشنی در GPU پیاده‌سازی شده است و کارت گرافیکی مورداستفاده در این روش NVIDIA GT620 می‌باشد.

نتایج مربوط به روش ارائه‌شده و همچنین تحقیقات پیشین، در Peds1 و Peds2 از مجموعه داده UCSD در جدول ۱ و جدول ۲ نمایش داده‌شده است.

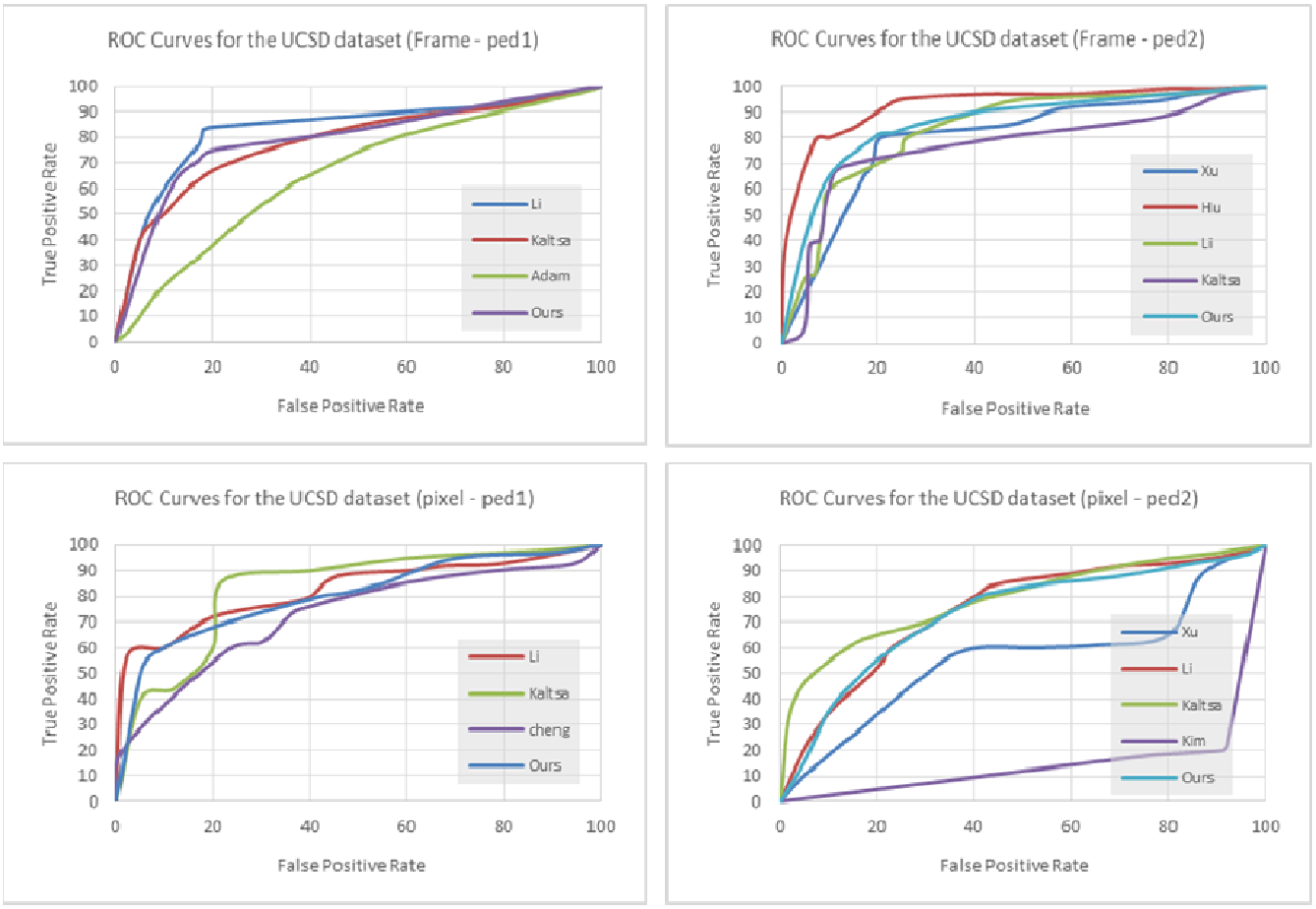
جدول ۱ نرخ ERR در زیرمجموعه Peds1 از مجموعه داده UCSD

نام نویسنده مقالات	نرخ ERR در سطح فریم	نرخ ERR در سطح پیکسل
[۳۱] Cheng et al.	۱۹/۹	۳۸/۸
[۱۳] Cong et al.	۲۳	۵۱/۲
[۲] Adam	۳۲/۶	۳۸/۹
[۹] Kim	۱۹/۶	۳۹/۶
[۳۵] Kaltsa	۲۱/۱	۲۷
[۲۶] Conv-AE	۲۷/۹	-
روش ارائه‌شده	۲۱/۳	۲۸

جدول ۲ نرخ ERR در زیرمجموعه Peds2 از مجموعه UCSD

نام نویسنده مقالات	نرخ ERR در سطح فریم	نرخ ERR در سطح پیکسل
[۲] Adam	۲۲/۴	۴۳/۸
[۹] Kim	۲۲/۴	۳۱/۱
[۳۵] Kaltsa	۲۵/۱	۲۶/۹
[۲۶] Conv-AE	۲۱/۷	-
روش ارائه‌شده	۱۹/۲	۲۷/۶

نتایجی که از سایر روش‌ها آورده شده، از مراجع مربوطه که این روش‌ها در آن‌ها معرفی‌شده‌اند، اقتباس‌شده است. در شکل ۷ نتایج شناسایی رفتارهای غیرعادی آورده شده است. منحنی‌های مربوط به ROC در مجموعه داده UCSD در شکل ۸ 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.



شکل ۸ 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 روش ارائه‌شده کارایی قابل‌رقابتی را در مقایسه با دیگر روش‌ها دارد. در حالت کلی می‌توان گفت که روش ارائه‌شده، کارایی بسیار رقابتی در مجموعه داده UCSD با نتایج متدهای دیگر را دارا است.

۵ جمع‌بندی

در این مقاله روشی برای شناسایی رفتارهای ناهنجار در تصاویر ویدئویی بخصوص درصحنه‌های پیچیده و شلوغ ارائه‌شده است. این روش ارائه‌شده مجموعه ویژگی‌هایی را بر مبنای اطلاعات شارنوری ، 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 شارنوری 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 تصاویر ورودی 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 شارنوری 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 مقاله به این سلول‌ها ROFIs گفته می‌شود. این سلول‌های ROFIs درواقع همان سلول‌هایی هستند که می‌بایست برای شناسایی رفتار ناهنجار مورد بررسی قرار گیرند.

با توجه به آنچه ذکر شد، هر یک از این سلول‌ها به‌عنوان ورودی مدل CNN برای شناسایی رفتار ناهنجار مورد استفاده قرار می‌گیرند. مدل CNN پیشنهادی با توجه به محاسبه احتمال متعارف و یا نامتعارف بودن این سلول‌ها تصمیم می‌گیرد که کدام سلول حاوی اطلاعات ناهنجار است و آن را به‌عنوان یک سلول ناهنجار معرفی می‌کند.

سپاسگزاری

در اینجا لازم است از تمامی اعضای پژوهشگاه پردازش سیگنال و تصویر دانشگاه صنعتی شاهرود و همچنین پژوهشگاه دانش‌های بنیادی به‌منظور حمایت از این پروژه تشکر و قدردانی به عمل‌آورده شود.

مراجع

[1] Cong, Y., J. Yuan, and J. Liu, Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition, 2013. 46(7): p. 1851-1864.

[2] Adam, A., et al., Robust real-time unusual event detection using multiple fixed-location monitors. IEEE transactions on pattern analysis and machine intelligence, 2008. 30(3): p. 555-560.

[3] Basharat, A., A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. in IEEE Conference on Computer Vision and Pattern Recognition. CVPR. . 2008.

[4] Boiman, O. and M. Irani, Detecting irregularities in images and in video. International journal of computer vision, 2007. 74(1): p. 17-31.

[5] Wang, X., X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. in IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. 2007.

[6] Li, J., S. Gong, and T. Xiang. Global Behaviour Inference using Probabilistic Latent Semantic Analysis. in BMVC. 2008.

[7] Jiang, F., et al., Anomalous video event detection using spatiotemporal context. Computer Vision and Image Understanding, 2011. 115(3): p. 323-333.

[8] Wu, S., B.E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[9] Kim, J. and K. Grauman. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). . 2009.

[10] Zhong, H., J. Shi, and M. Visontai. Detecting unusual activity in video. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR . 2004.

[11] Hamid, R., et al. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. . 2005.

[12] Lu, C., J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. in Proceedings of the IEEE International Conference on Computer Vision. 2013.

[13] Cong, Y., J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011.

[14] Cong, Y., J. Yuan, and Y. Tang, Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Transactions on Information Forensics and Security, 2013. 8(10): p. 1590-1599.

[15] Li, W., V. Mahadevan, and N. Vasconcelos, Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence, 2014. 36(1): p. 18-32.

[16] Mehran, R., A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. in IEEE Conference on Computer Vision and Pattern Recognition. CVPR. . 2009.

[17] Roshtkhari, M.J. and M.D. Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Computer vision and image understanding, 2013. 117(10): p. 1436-1452.

[18] LeCun, Y., et al., Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998. 86(11): p. 2278-2324.

[19] Goodfellow, I.J., et al., Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082, 2013.
[20] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.
[21] Sermanet, P., et al., Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.
[22] Shen, W., et al. 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. 2015.
[23] Taigman, Y., et al. Deepface: Closing the gap to human-level performance in face verification. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
[24] Ji, S., et al., 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2013. 35(1): p. 221-231.
[25] Sabokrou, M., et al., Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 2017. 26(4): p. 1992-2004.
[26] Hasan, M., et al. Learning temporal regularity in video sequences. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[27] Xu, D., et al., Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 2017. 156: p. 117-127.
[28] Wang, K., et al. 3D human activity recognition with reconfigurable convolutional neural networks. in Proceedings of the 22nd ACM international conference on Multimedia. 2014. ACM.
[29] Feng, Y., Y. Yuan, and X. Lu, Learning deep event models for crowd anomaly detection. Neurocomputing, 2017. 219: p. 548-556.
[30] Maturana, D. and S. Scherer. 3d convolutional neural networks for landing zone detection from lidar. in IEEE International Conference on Robotics and Automation (ICRA). . 2015. IEEE.
[31] Cheng, K.-W., Y.-T. Chen, and W.-H. Fang, Gaussian Process Regression-Based Video Anomaly Detection and Localization With Hierarchical Feature Representation. IEEE Transactions on Image Processing, 2015. 24(12): p. 5288-5301.
[32] Laptev, I. and T. Lindeberg. Space-time interest points. in IEEE conference proceedings 9th International Conference on Computer Vision, Nice, France. 2003..
[33] UCSD Dataset. Available from: www.svcl.ucsd.edu/projects/anomaly/dataset.html.
[34] Hu, Y., Y. Zhang, and L. Davis. Unsupervised abnormal crowd activity detection using semiparametric scan statistic. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2013.
[35] Kaltsa, V., et al., Swarm intelligence for detecting interesting events in crowded environments. IEEE transactions on image processing, 2015. 24(7): p. 2153-2166.
[36] Xu, D., et al., Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing, 2014. 143: p. 144-152.
[37] Zhu, X., et al., Sparse representation for robust abnormality detection in crowded scenes. Pattern Recognition, 2014. 47(5): p. 1791-1799.
[38] Saligrama, V. and Z. Chen. Video anomaly detection based on local statistical aggregates. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012..
[39] Chollet, F. keras,. 2015; Available from: https://github.com/fchollet/keras.
[40] James, B., et al. Theano: a CPU and GPU math expression compiler. in Proceedings of the Python for Scientific Computing Conference (SciPy).
[41] Al-Rfou, R., et al., Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688, 2016.
[42] Abadi, M., et al. TensorFlow: A system for large-scale machine learning. in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA. 2016.

بهنام سبزعلیان مدرک کارشناسی خود را در رشته نرم افزار از دانشگاه سمنان اخذ نموده است و مقطع کارشناسی ارشد را در رشته مهندسی رباتیک در دانشگاه صنعتی شاهرود به پایان رسانده است.زمینه های مورد علاقه ایشان پردازش تصویر و ویدیو، بینایی‌ماشین و یادگیری عمیق است.

حسین مروی مدرک کارشناسی خود را در رشته برق الکترونیک از دانشگاه فردوسی مشهد اخذ نمودند .ایشان مقطع کارشناسی ارشد را در دانشگاه شیراز در رشته برق گرایش مخابرات به پایان رساندند. مدرک دکتری تخصصی خود را از کشور انگلستان دانشگاه Surrey مرکز CVSSP و در زمینه پردازش گفتار اخذ نمودند. نامبرده هم اکنون به عنوان هیات علمی با مرتبه دانشیاری در دانشکده مهندسی برق و رباتیک دانشگاه صنعتی شاهرود مشغول فعالیت می‌باشند. زمینه‌های مورد علاقه ایشان پردازش سیگنال‌ها، زمینه‌های مختلف پردازش گفتار و پردازش سیگنال‌های حیاتی می باشد.

علیرضا احمدی‌فرد مدرک کارشناسی را از دانشگاه صنعتی اصفهان در مهندسی الکترونیک و کارشناسی ارشد را از دانشگاه صنعتی امیرکبیر در مهندسی مخابرات اخذ نمودند. ایشان مدرك دکتری تخصصی را در زمینه پردازش‌تصویر و ماشین‌بینایی از مرکز CVSSPدانشگاه Surrey در سال ۲۰۰۲اخذ نمودند. زمینه‌های تحقیقاتی مورد علاقه ایشان پردازش سیگنال، پردازش تصاویر و شناسایی الگو می‌باشد.