EasyChair Preprint

№ 5357

Behaviors Violence Detection of Surveillance Video Using Spatial-Temporal Convolution and Atrous Convolutional

Behzad Lak and Tayebeh Pakaghideh

April 20, 2021

Behaviors Violence detection of surveillance video using spatial-temporal convolution and Atrous Convolutional

[1] Behzad lak, [2]Tayebeh pakaghideh

[1]Assistant Professor Department of Information and Communication Technology Amin University of Law Enforcement Sciences, Tehran, Iran
Behzad_lak@yahoo.com

[2] Master of Software Engineering, Tehran, Iran
t.pakaghideh72@yahoo.com

Abstract

Detecting moving objects in each frame is an essential step in video analysis and violence detection. In this paper, a new method for separating frames containing motion information and detecting violence in them is presented. In the proposed method, frames containing motion information are separated and their roughness is detected at two levels of the network. At level one, Atrose Convolution receives input video to the network and Separates frames containing motion information by applying semantic segmentation to network entry frames then transfers them to the level of the two networks, spatial-temporal convolution, for violence detection. Finally, in order to ensure the correct operation of the network, the regression unit, after checking the output of the information, classifies it into two classes, rough and non-rough, and considers a score for them. The closer the score is to 0, the less violence is detected, and the closer the score is to 1, the more violence is detected. To show the accuracy of the proposed algorithm, two sets of data have been examined, the total accuracy obtained from them is equal to 96% in the ucf-crime data set and also 93% of the surveillance video data set.

Keywords: spatial-temporal Convolutional, Atrous Convolutional , Dvsnet ,flownet , Adaptive Key Frame

تشخیص رفتارهای خشونت آمیز فیلم های نظارتی بااستفاده از کانولوشن زمانی مکانی و آتروس کانولوشن

بهزادلک[1]، طیبه پاک عقیده[2]

[1] استادیار، گروه فناوری اطلاعات و ارتباطات ، دانشگاه علوم انتظامی امین ، تهران،
Behzad_lak@yahoo.com

[2] کارشناس ارشد مهندسی نرم افزار
t.pakaghideh72@yahoo.com

چکیده

تشخیص اشیای متحرک در هر فریم گامی اساسی در تحلیل ویدئو و تشخیص خشونت است.در این مقاله روشی جدید برای جداسازی فریم های حاوی اطلاعات حرکتی و تشخیص خشونت در آنها ارائه شده است.در روش پیشنهادی جداسازی فریم های حاوی اطلاعات حرکتی و تشخیص خشن بودن آنها در دو سطح شبکه صورت می‌گیرد.در سطح یک آتروس کانولوشن ویدئوهای ورودی به شبکه را دریافت می‌کند و با اعمال تقسیم بندی معنایی روی فریم های ورود به شبکه فریم های حاوی اطلاعات حرکتی را جداسازی می‌کند سپس آنها را جهت تشخیص خشن بودن رفتار های صورت گرفته به سطح دو شبکه یعنی کانولوشن زمانی مکانی انتقال می‌دهد.در انتها جهت اطمینان از صحت عملکرد شبکه واحد رگرسیون پس از بررسی خروجی شبکه اطلاعات را در دو کلاس خشن و غیرخشن دسته بندی می‌کند و برای آنها نمره در نظر میگیرد.هرچقدر نمره به دست آمده نزدیک 0 باشد یعنی خشونت تشخیص داده نشده و هرچقدر این نمره نزدیک1 باشد یعنی خشونت تشخیص داده شده است.برای نمایش میزان دقت الگوریتم پیشنهادی دو مجموعه دادگان مورد بررسی قرار گرفته اند که مجموع دقت به دست آمده از آنها برابر با 96٪ در مجموعه داده ucf-crime و همچنین 93٪ از مجموعه داده surveillance video است.

کلمات کلیدی

کانولوشن زمانی – مکانی ، آتروس کانولوشن ، تقسیم بندی معنایی ویدئو ، شبکه ی جریان ، فریم کلیدی انطباقی

1- مقدمه

درک رفتار انسان و تحلیل فعالیت های آن تا کنون با چالش های زیادی مواجه بوده است[1] بنابراین پردازش فیلم ها وهمچنین درک محتوا از فیلم ها با دقتی مناسب و در مقیاس بزرگ ازاهمیت ویژه ای برخوردار است[2]. در همین راستا با توجه به استفاده از دوربین ها در سطح شهر با حجم بسیار زیادی از ویدئو و محتوا روبرو هستیم که تحلیل این حجم از اطلاعات توسط انسان غیر ممکن است[3] . با تحلیل فیلم های ضبط شده از ویدئو های نظارتی می توان ناهنجاری و خشونت رخ داده را تشخیص داد و در زمان و شرایط مناسب با وقایع, تصمیمات تاثیر گذاری اتخاذ کرد[4]. در این مقاله شبکه دو سطحی طراحی شده است که متشکل از آتروس کانولوشن به عنوان سطح یک شبکه و کانولوشن زمانی –مکانی به عنوان سطح دو شبکه می‌باشد . ابتدا توسط سطح یک این شبکه یعنی آتروس کانولوشن ویدئوهای ورودی دریافت می‌شوند و فریم های حاوی اطلاعات حرکتی از بین فریم های ورودی به شبکه استخراج می‌شوند سپس این فریم های حاوی اطلاعات حرکتی جهت تشخیص خشونت صورت گرفته به سطح دو شبکه انتقال داده

می‌شوند. نکته حائز اهمیت در این شبکه برای تشخیص خشونت استخراج اطلاعات در هر دو بعد زمان و مکان است . این شبکه توانایی تشخیص ناهنجاری و خشونت رخ داده در ویدئو های نظارتی سطح شهر را دارد که مبتنی بر روش یادگیری عمیق است .

تشخیص خشونت به علت کاربردهای گسترده و قابل توجه در زمینه امنیت و آسایش بشر کاربرد خود را در سیستم های نوین مورد استفاده انسان نشان داده است . با توجه به مقالات موجود در زمینه تشخیص خشونت می توان به پژوهش [5] اشاره داشت که با ترکیب ویژگی های زمانی مکانی و شتاب که هریک توسط یک شبکه کانولوشنی دوبعدی و سپس یک شبکه بازگشتی LSTM به دست آمد, به استخراج ویژگی ها پرداخته شد . تغییرات شتاب از مولفه های مهم در تشخیص خشونت نفر به نفر است . در این مقاله با استفاده از شدت تغییرات جریان نوری موجود در سه فریم متوالی این شتاب محاسبه و مدل شده است . برای محاسبه ی ویژگی های مکانی از شبکه ی VGG 19 [6] که بر روی Imagenet آموزش دیده است استفاده شده و ویژگی ها از لایه ی ماقبل آخر به عنوان بردار ویژگی انتخاب شده اند . شبکه ی دیگری که تغییرات جریان نوری را به عنوان ویژگی درنظر می گیرد با استفاده از یک شبکه Tdd [7] است که با استفاده از مجموعه داده ی ucf101 آموزش دیده و ایجاد شده است .

در پژوهشی دیگر [8] ایده TSN یعنی شبکه تقسیم بندی زمانی ارائه شد. که در واقع یک چارچوب جدید برای تشخیص عملکرد مبتنی بر فیلم و مبتنی بر ایده مدل سازی ساختار زمانی دوربرد بود . روش آنها استخراج ویژگی های زمانی به صورت پراکنده بود که شامل نظارت سطحی ویدیو برای فعال کردن یادگیری مؤثر و کاربردی با استفاده از فیلم های دارای صحنه های خشن و اکشن است. که به صورت پراکنده از فریم های ورودی نمونه برداری می کند که به نوعی می توان گفت معماری سگمنتال است. در پژوهش [9] از شبکه 3D ConvNet و فریم کلیدی key frame برای استخراج ویژگی ها از کلیپ هایی که حاوی صحنه های خشونت آمیز هستند استفاده کردند . از3D ConvNet برای کلیپ های کوتاه مدت و از key frame برای کلیپ های طولانی تر استفاده نموده اند . روش فریم کلیدی فیلم را بر اساس فریم های کلیدی استخراج شده تقسیم می کند و سپس شباهت بین فریم های مجاور با تغییر موقعیت مرکز خاکستری را مورد بررسی قرار می دهد .

در پژوهش[10] روشی برای کشف سرقت های خشونت آمیز از فیلم های دوربین مداربسته با استفاده از یک مدل توالی عمیق پایان به پایان ارائه شد . آنها برای استخراج ویژگی از VGG-16 و یک CNN از قبل آموزش دیده توسط قاب های ویدیویی اصلی که به استخراج ویژگی می پردازد استفاده نموده اند . آنها با استفاده از لایه های کانولوشنی LSTM که دنباله ای از ویژگی های استخراج شده را دریافت می کند تمام توالی ویژگی ها را توسط دو حافظه طولانی مدت و کوتاه مدت (convLSTM)پردازش می کنند . بعد از این مراحل در انتها از چند لایه ی کاملا متصل برای نتیجه گیری و دسته بندی استفاده نموده اند . در این روش انواع سلاح گرم و سرد موجود در تصویر تشخیص داده می شود و به این ترتیب سرقت هایی که نشان دهنده سطح متفاوتی از پرخاشگری هستند را می توان کلاس بندی نمود .

با توجه به روش های مختلف تشخیص که مورد بررسی قرار گرفت دریافت شد که خلاء موجود در قسمت استخراج صحیح ویژگی ها می باشد بدین ترتیب هریک از روش ها دارای پیچیدگی و هزینه ی محاسباتی بالا و همچنین سر ریز شبکه و از دست دادن ویژگی های حرکتی بودند که بنابراین جهت حل مسئله ی تشخیص خشونت نیاز به اطلاعات صحنه از هر دو سطح به عبارت ساختار صحنه و حرکت صورت گرفته توسط اشخاص حاضر در صحنه می باشد . بنابراین با بررسی روش های مختلف شبکه ای طراحی شده است که متشکل از آتروس کانولوشن به عنوان شبکه پایه و کانولوشن زمانی مکانی به عنوان تشخیص دهنده پایه می باشند که وظیفه استخراج ویژگی ها و تشخیص خشونت را در دوسطح اجرا می کنند . ساختار شبکه طراحی شده به طور کامل در بخش 2 توضیح داده شده است .

2- روش پیشنهادی

در این مقاله روشی پیشنهادی برای استخراج ویژگی های زمانی – مکانی و تشخیص خشونت در دو سطح شبکه ارائه شده است . در این روش ابتدا در سطح یک شبکه با استفاده از شبکه ی تقسیم بندی معنایی, ویدئوهای ورودی به شبکه توسط لایه های اولیه این شبکه یعنی آتروس کانولوشن دریافت می شوند و بعد از اعمال فیلتر ها جهت استخراج ویژگی ها ,فریم ها به قسمت تقسیم بندی معنایی منتقل می‌شوند که در این قسمت از شبکه فریم هایی که در آنها انسان در حال حرکت وجود داشته باشد از فریم های بدون انسان و حرکت جداسازی می‌شوند سپس روی فریم های دارای حرکت, عملیات زمان بندی اعمال می‌شود . بعد از پردازش فریم های زمانبندی شده , فریم های حاوی اطلاعات حرکتی و زمان بندی شده به سطح دو شبکه یعنی کانولوشن زمانی – مکانی جهت تشخیص خشونت انتقال داده می‌شوند .

2-1- ساختار آتروس کانولوشن

این شبکه از دو بخش به عبارت شبکه ی تقسیم بندی معنایی ویدئو و شبکه ی جریان تشکیل شده است . روند کار این سطح از شبکه به این ترتیب است که ابتدا فریم های ویدئویی ورودی توسط آتروس کانولوشن دریافت می شوند سپس فریم ها به شبکه ی تقسیم بندی معنایی ویدئو, انتقال داده شده و بعد از اعمال فرآیند تقسیم بندی, فریم های تقسیم بندی شده به شبکه ی جریان انتقال داده می شوند. در شبکه تقسیم بندی معنایی ویدئو, هر فریم به چندین منطقه تقسیم می‌شود. سپس در مواردی که بیشتر محتوای هرفریم بافریم های بعداز خودش مشابه باشد ومناطقی از فریم که تفاوت های جزئی بین فریم های متوالی داشته باشند به شبکه جریان منتقل می‌شوند و مناطقی که تفاوت زیادی بین فریم های متوالی دارند و در آن اطلاعات حرکتی به طور قابل توجهی تغییر می کند توسط شبکه تقسیم بندی پردازش می‌شوند .

به طور معمول در یک ویدیو مناطق هر فریم دارای حرکت های مکانی نسبتاً پراکنده هستند بنابراین, مناطق پردازش شده هرفریم که حاوی اطلاعات حرکتی بودند توسط شبکه تقسیم بندی و شبکه جریان به عنوان مناطق فریم کلیدی و مناطق دارای حرکت مکانی تعیین می شوند . دو مزیت عمده شبکه ی تقسیم بندی معنایی ویدئو را می توان اینگونه بیان کرد :1- افزایش کارایی شبکه, زیرا تقسیم بندی معنایی ویدئو خود را با تفاوت های بین مناطق مختلف از فریم های کلیدی در زمان اجرا وفق می‌دهد . 2- با

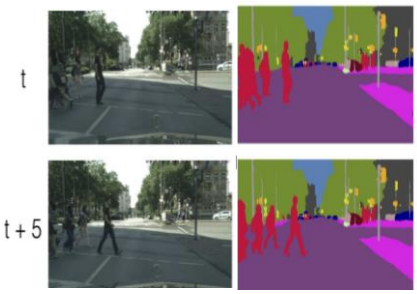
استفاده از شبکه جریان محاسبات قابل توجهی را ذخیره می‌کند . واضح است که این طرح در درجه اول تقسیم بندی معنایی ویدئوهای ورودی به شبکه را مورد هدف قرار داده است بنابراین برای تعیین یک سیاست نظام مند برای اختصاص مناطق فریم به دو شبکه ذکر شده با حفظ انعطاف پذیری و قابلیت تنظیم کارآمد,روش سیاست انطباق زمانبندی فریم کلیدی مطرح است . سیاست زمان بندی فریم کلیدی تطبیقی روشی برای تعیین پردازش و یا عدم پردازش یک منطقه از فریم ورودی توسط شبکه تقسیم بندی است [11] .

2-1-1- شبکه تقسیم بندی معنایی ویدیو

ساختار شبکه ی تقسیم بندی معنایی ویدئومتشکل از یک لایه کانولوشن و سه لایه کانولوشن تماما متصل است . اولین قدم تقسیم فریم های ورودی به مناطق فریم است به این ترتیب که ابتدا اولین فریمی که دارای اطلاعات حرکتی است را دریافت می‌کند و نگه می دارد سپس فریم های بعد از آن را بررسی می‌کند. اولین فریمی که بعد از فریم اول دارای اطلاعات حرکتی بود را مجدد نگه می‌دارد و تفاوت این دو فریم یعنی اولین فریم که فریم اصلی است را با فریم فعلی بررسی می‌کند . به همین ترتیب در ادامه ی بررسی های فریم های بعدی آنها را نسبت به فریم اصلی مطابقت می‌دهد و میزان مطابقت بین فریم های فعلی را با فریم اصلی ارزیابی می‌کند و تعیین می‌کند عملی که در زمان t فریم اصلی انجام شده است چه میزان با عملی که در زمان t+10 فریم فعلی انجام شده است تفاوت دارد. در واقع با این محاسبه تفاوت منطقه ی قاب فعلی با قاب اصلی ارزیابی می‌شود و برای آنها نمره اطمینانی را در نظر می‌گیرد . نمره های اطمینان به دست آمده را با نمره اطمینان از پیش تعیین شده مقایسه می‌کند .اگر نمره اطمینان یک منطقه از حد آستانه پایین تر باشد به شبکه ی تقسیم بندی انتقال داده می‌شود و در غیر این صورت به عنوان فریم دارای حرکت مکانی شناخته می‌شود و به شبکه ی جریان انتقال داده می‌شود . عملکرد این شبکه به این ترتیب است که ارزیابی کند آیا مسیر دریافتی دارای حرکت مکانی است یا خیر که آن را تقسیم بندی کند . هرچه نمره اطمینان مورد انتظار بالاتر باشد به احتمال زیاد مسیر دارای حرکت زمانی است که می‌توان به آن دسترسی داشت .

در مرحله ی بعد مناطق فریم به مسیرهای مختلف هدایت می‌شوند تا تقسیم بندی معنایی را برای منطقه خود ایجاد کنند . برای مسیر دارای حرکت مکانی, یک تابعf برای تقسیم بندی و برای پردازش حرکت مکانی تابع W(*) اعمال می‌شود [12]. با تقسیم بندی منطقه ای فریم کلیدی, یک تقسیم بندی جدید برای آن منطقه ایجاد می‌شود . به همین ترتیب این فرآیند برای تمامی فریم های ورودی به شبکه اعمال خواهدشد . نکته ای که حائز اهمیت است استفاده از همبستگی های زمانی بین فریم ها برای کاهش زمان محاسبه است. در شکل 1,2 نمونه ای از فریم های تقسیم بندی معنایی شده که دارای حرکت و بدون حرکت هستند نمایش داده شده است.

شکل 1- فریم های بدون عابر که تقسیم بندی معنایی شده اند [11]



شکل 2 -فریم های با عابر پیاده که تقسیم بندی معنایی شده اند[11]

2-1-2 زمانبندی فریم کلیدی انطباقی

زمانبندی فریم کلیدی انطباقی فریم های اصلی با استفاده از معیاری به نام نمره اطمینان مورد انتظار به روز رسانی می‌شود و نمره اطمینان مورد انتظار برای هر منطقه از قاب ها ارزیابی می‌شوند . اگر نمره اطمینان بالاتر از حد آستانه باشد تقسیم بندی ایجاد شده توسط شبکه جریان همانند شبکه تقسیم بندی دسته بندی خواهد شد . در واقع مقدار نمره اطمینان مورد انتظار منعکس کننده اطمینان شبکه جریان برای تولید نتایج مشابه شبکه تقسیم بندی می باشد . هرچه اختلاف منطقه فریم بیشتر باشد , به احتمال زیاد شبکه جریان قادر به استنباط تقسیم بندی صحیح منطقه فریم نیست که در این صورت ابتدا یک منطقه قاب برای نمره اطمینان مورد انتظار را مورد تجزیه و تحلیل قرار می دهد . اگر نمره اطمینان مورد انتظار آن بالاتر از آستانه از پیش تعیین شده باشد توسط شبکه جریان پردازش می‌شود . در غیر این صورت به شبکه تقسیم بندی اختصاص داده می‌شود . در حقیقت عملکردش به این ترتیب است که با برآورد نمره اطمینان مورد انتظار تعیین کند که آیا یک منطقه فریم ورودی به شبکه باید از شبکه تقسیم بندی عبور کند یا عبور نکند . بنابراین با استفاده از قابلیت زمان بندی فریم , فریم های کلیدی پس از مدت زمان مشخصی به روز می شوند . البته پردازش توالی فریم ها از محتویات مشابه با دوره به روزرسانی طولانی بسیار کارآمدتر هستند این در صورتی است که یک دوره به روزرسانی ثابت نمی تواند چنین نتیجه ای را داشته باشد .

فریم هایی که مسیر های دارای حرکت های مکانی دارند مانند ، تقسیم بندی های منطقه ای برای تولید نتایج منطقی کافی هستند . اما وقتی که حرکت های مکانی در صحنه به طرز چشمگیری تغییر می کنند استفاده از شبکه تقسیم منطقی تر است . این به این دلیل است که شبکه جریان قادر به پیش بینی جابجایی اجسامی که در صحنه نیستند و فقط در منطقه فریم کلیدی مربوطه وجود دارند نیست . با توجه به این استدلال ها از معماری تقسیم بندی معنایی پویا و نمره اطمینان مورد انتظار ، سیاست زمانبندی قاب اصلی تطبیقی در این سطح از شبکه اعمال شده است .

که در آن دوره به روز رسانی ثابت نیست و با توجه به نمره اطمینان مورد انتظار آن منطقه تعیین می‌شود .در شبکه تقسیم بندی معنایی ویدئو dnn , Dvsnet به عنوان cnn اجرا می شود. dnn با ارزیابی تعیین می کند که چه زمانی منطقه قاب اصلی R به روز رسانی شود اگر خروجی شبکه ی جریان Fr بتواند تقسیم بندی منطقه ای مطلوبی ایجاد کند Or می شود . که از Or انتظار می رود نزدیک به شبکه تقسیم بندی باشد و Sr , Fr برای تولید Or به تابع محاسبه ی جابه جایی مکانی W(*) انتقال داده می شوند . در غیر این صورت منطقه قاب فعلی به شبکه تقسیم طولانی تر

ارسال می شود و قاب اصلی به روز می شود [13].فرمول محاسبه ی نمره ی اطمینان در زیر آمده است :

$$confidence\ score = \frac{\sum_{p\in P} C(O^r(p), S^t(p))}{p} \quad (1)$$

پس از اتمام مراحل فوق و استخراج ویژگی های حرکتی در فریم های مربوطه آن فریم ها به سطح دوم شبکه یعنی کانولوشن زمانی – مکانی انتقال پیدا می‌کنند و در آن شبکه خشونت های رخ داده تشخیص داده می شوند که روند کار و جزئیات آن در ادامه آمده است .

2-2- ساختار شبکه کانولوشنی زمانی – مکانی

کانولوشن زمانی–مکانی برای استخراج اطلاعات از هر دو بعد زمان و مکان یک کرنل سه بعدی به حجم زمان و مکان روی تصویر اعمال می‌کند که توسط آن ویژگی ها از ورودی استخراج می‌شوند و نقشه های ویژگی را تشکیل می‌دهند . نقشه های ویژگی در لایه های کانولوشن به لایه های قبلی به فریم های پیوسته وصل می‌شوند , بدین ترتیب , ویژگی های حرکتی ضبط می‌شوند .

ساختار کانولوشن زمانی – مکانی پیشنهادی در این پژوهش از 8 لایه کانولوشنی و 3لایه تماما متصل تشکیل شده است . در شبکه ی پیشنهادی بعد از هر دو لایه کانولوشن یک لایه 3d average pooling قرار گرفته است که در مجموع از سه لایه 3d average pooling استفاده شده است که سایز هریک از آنها به ترتیب 1*1*1 و 2*2*1 و 2*2*2 می باشد و درانتهای شبکه قبل از لایه های کاملا متصل لایه spp قرار گرفته است که تمام نقشه های ویژگی های استخراج شده و میانگین گرفته شده از هر لایه توسط 3d average pooling به این قسمت یعنیspp انتقال پیدا می‌کند . اندازه ورودی ساختار زمانی – مکانی320 *240*30 می‌باشد که مربوط به 30 فریم پیوسته از 320*240 پیکسل از توالی ویدئویی می‌باشد . در مرحله اول کانولوشن زمانی – مکانی با کرنل سه بعدی سایز 3*3*3 (ابعاد مکانی 3*3 و 3 در بعد زمان) روی داده های ورودی اعمال می‌شوند . یک کرنل سه بعدی فقط می تواند یک نوع ویژگی را ضبط کند.

برای افزایش تعداد انواع ویژگی ها در کل کرنل های سه بعدی در داده های ورودی اعمال می‌شوند که در نهایت نقشه ویژگی ها را در لایه conv1,conv 2 خواهیم داشت که توسط لایه 3d average pooling نقشه های ویژگی به دست آمده میانگین گرفته می شوند همین ترتیب در لایه های conv 4 , conv 3 و conv 5 , conv 6 صورت می گیرد .

که در انتهای شبکه تمام نقشه های ویژگی به دست آمده از تمام لایه های شبکه برای نتیجه گیری نهایی به لایه spp منتقل می شوند بعد از طی این مراحل نتایج به دست آمده از لایه spp به fc1 انتقال پیدا می کند . در لایه fc1 عملیات 2d convolution انجام می شود برای استخراج اطلاعات مکانی در سطحی بالاتر با کرنل سایز 4*4 که نقشه های ویژگی خروجی در fc1 در یک بردار ویژگی قرار می گیرند و در لایه fc2 عملیات 3d convolution انجام می شود جهت استخراج اطلاعات زمانی با کرنل سایز 4*4*4 که نقشه های ویژگی به دست آمده در این لایه در یک بردار ویژگی قرار می گیرند که در نهایت لایه fc3 به طور کامل با هر واحد از بردار های ویژگی ها در لایه های fc2 , fc1 وصل می شوند جهت تعیین رفتارهای حرکتی و غیر حرکتی که تعداد واحدهای خروجی fc3 برابر با 2 است که این همان تعداد انواع رفتار ها یعنی (خشن و غیر خشن)می باشد و هریک از واحدها بیانگر احتمال یک فرضیه رفتار است .

2-2-1-استفاده از لایه spp در کانولوشن زمانی – مکانی

از آنجایی که لایه های کانولوشن ورودی با اندازه های ثابت را می پذیرند اما خروجی هایی با اندازه متغیر تولید می کنند بنابراین نیاز به یک روش قابل انعطاف تر مانند spp می باشد [13] مدل spp اطلاعات محلی را با استفاده از pooling در مخزن های محلی ذخیره میکند که اندازه این مخزن های محلی که حاوی اطلاعات مکانی هستند برابر با اندازه تصویر می باشند این لایه تصاویر ورودی به شبکه را در هر مقیاسی می پذیرد و در انتها خروجی حاصل از هر لایه را جمع آوری میکند و به لایه کاملا متصل در انتهای شبکه انتقال می دهد [14] بنابراین می توان با استفاده از لایه Spatial pyramid poolingدر ساختار کانولوشن زمانی – مکانی اطلاعات را در چند دامنه و مقیاس به دست آورد که برای حل مسئله ی تشخیص بسیار کاربردی است و می توان آن را برای هر شبکه ای اعمال کرد [15] . و همچنین از این لایه جهت کاهش ابعاد و بهبود بی ثباتی در برابر نویز و تحولات محلی تصویر و همچنین استخراج ویژگی های حرکتی در تصویر ورودی می توان استفاده کرد زیرا این لایه این قابلیت را دارد که بتوان توسط آن قسمت دلخواه از تصویر را پردازش کرد . در الگوریتم طراحی شده این لایه در انتهای شبکه و قبل از لایه های کاملا متصل استفاده می شود که نقشه های ویژگی ها را به سه لایه آخر یعنی لایه های کاملا متصل انتقال می دهد .

3- آزمایش

در این بخش از مقاله به چگونگی پیاده سازی الگوریتم پیشنهادی و آزمایش انجام شده بر روی مجموعه ی دادگان و نتایج به دست آمده پرداخته شده است . برای انجام این آزمایش ها از چارچوب پایتون جهت پیاده سازی استفاده شده است . به این ترتیب که پس از آموزش شبکه های مورد نظرتوسط مجموعه دادگان (شبکه ی آتروس کانولوشن و کانولوشنی زمانی – مکانی) وزن های شبکه های آموزش دیده (فیلتر ها) ذخیره شدند . این وزن های ذخیره شده در آتروس کانولوشن و کانولوشنی زمانی – مکانی در هر لایه کانولوشنی توسط لایه های ادغام که در این شبکه بعداز هر دولایه ی کانولوشنی یک لایه3d average pooling در نظر گرفته شده است, ویژگی ها را با هم ادغام می کند . با توجه به این نکته که لایه های ادغام با توجه به ابعاد ادغام شونده برخی از ویژگی ها که از اهمیت کمتری برخوردار است را حذف می کند سعی شده است از ابعادکوچکتر در شبکه استفاده شود که میزان این اتفاق به حداقل برسد . باید به این نکته توجه داشت که با استفاده از کرنل سایز کوچکتر و داشتن نرخ کمتر می توان میزان محاسبات قابل کنترل تر و همچنین تعداد پارامتر های کمتری داشته باشیم که این امر از سر ریز شدن و انحطاط در شبکه و تشخیص هرچه دقیق تر شبکه کمک می کند .

3-1 جزئیات پیاده سازی

معماری پیشنهادی شبکه ی طراحی شده جهت تشخیص خشونت با استفاده از فریم ورک پایتون و همچنین با استفاده از کتابخانه ی پایتون ازجمله

تنسورفلو جهت پیاده سازی بهره برداری شده است . کدهای پیاده سازی شده بر روی کارت گرافیک NVIDIA GEFORCE GTX و کامپیوتری با 64 گیگابایت فضای حافظه آزمایش شده است . برای تسریع در روند آموزش الگوریتم شبکه ی پیشنهادی وزن های اولیه از طریق مجموعه دادگان ucf [17]به شبکه آموزش داده شده است. برای نمایش میزان دقت شبکه ی طراحی شده مدل های استفاده شده در این الگوریتم بر روی مجموعه داده هایی که شامل ویدئو هایی با صحنه های ناهنجار و خشن می باشند آزمایش شده است و نتایج به دست آمده توسط سایر مقالات و آزمایش های انجام شده در این پژوهش به صورت میزان دقت گزارش شده است . جهت آموزش شبکه ی طراحی شده جهت تشخیص خشونت اندازه فریم های ورودی به شبکه 240*320 می باشد از 30 فریم پیوسته که با نرخ یادگیری 0.003 شروع شده است .

4- اعتبارسنجی

یکی از بخش های مهم در ارائه ی یک شبکه جدید مربوط به ارائه ی گزارش کاملی از کارایی شبکه ی طراحی شده و صحت و دقت عملکرد شبکه در شرایط مختلف است . به همین منظور جهت اثبات کارایی بهتر و میزان دقت شبکه ی طراحی شده نسبت به شبکه های طراحی شده ی مشابه که توسط سایر محققان ارائه و پیاده سازی شده است نتایج به دست آمده از عملکرد این شبکه با سایر شبکه ها مقایسه شده اند . شباهت شرایط آزمایش به یکسان بودن مجموعه دادگان آزمایش و یکسان بودن پارامترهای ارزیابی کیفیت مربوط می شود . به همین دلیل و با گذشت زمان تعدادی از روش های آزمون شبکه و پارامترهای ارزیابی تبدیل به معیار استانداردی برای مقایسه ی روش ها با یکدیگر شده اند که این معیار های استفاده شده در این مقاله عبارت اند از Recall , Accuracy ,Precision .

جدول1 :نتایج ارزیابی معیارها بر روی مجموعه دادگان

	UCF - crime	Surveillance Video
Accuracy	96%	93%
Recall	95%	98%
Precision	98%	96%

5- نتیجه گیری

شبکه ی پیشنهادی مورد ارزیابی قرار گرفت و مزایا و معایب آن بیان شد . همانطور که در مورد شبکه ی پیشنهادی توضیح داده شد شبکه ی آتروس کانولوشن وظیفه ی تقسیم بندی معنایی ویدئو را دارد و همچنین شبکه ی کانولوشن زمانی مکانی با استخراج ویژگی های زمانی و مکانی به صورت هم زمان به حل مسائل و تشخیص خشونت کمک می کند . و به دلیل پراهمیت بودن حرکت به جای ویژگی های ظاهری, در فریم های ویدئویی, این شبکه گزینه مناسبی جهت استخراج ویژگی ها می باشد . برای مجموعه داده surveillance video[16] دقتی برابر با 93٪ محاسبه شد و برای مجموعه داده ی UCF-crime[17] دقتی برابر با 96٪ محاسبه شده است . این اختلاف در دقت دو مجموعه ناشی از ماهیت ویدئوهای موجود در دو مجموعه است .

مراجع

[1] Ryoo, Michael S., Brandon Rothrock, and Larry Matthies. "Pooled motion features for first-person videos." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[2] Mabrouk, Amira Ben, and Ezzeddine Zagrouba. "Abnormal behavior recognition for intelligent video surveillance systems: A review." *Expert Systems with Applications* 91 (2018): 480-491.

[3] Ryoo, M. S., et al. "Robot-centric activity prediction from first-person videos: What will they do to me?." *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015.

[4] Xia, Lu, et al. "Robot-centric activity recognition from first-person rgb-d videos." *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015.

[5] Yu, Wei, et al. "Visualizing and comparing AlexNet and VGG using deconvolutional layers." *Proceedings of the 33 rd International Conference on Machine Learning*. 2016.

[6] Cao, Congqi, et al. "Action Recognition with Joints-Pooled 3D Deep Convolutional Descriptors." *IJCAI*. Vol. 1. 2016.

[7] Si, Chenyang, et al. "An attention enhanced graph convolutional lstm network for skeleton-based action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

[8] Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." *European conference on computer vision*. Springer, Cham, 2016.

[9] R-C3D: Region Convolutional 3D Network for Temporal Activity Detection.IEEE,2017

[10] Rezaee, Mohammad, et al. "Using a vgg-16 network for individual tree species detection with an object-based approach." *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*. IEEE, 2018.

[11] Xu, Yu-Syuan, et al. "Dynamic video segmentation network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[12] Zhu, Xizhou, et al. "Deep feature flow for video recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[13] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015): 1904-1916.

[14] Yang, Wanli, et al. "Video-Based Human Action Recognition Using Spatial Pyramid Pooling and 3D Densely Convolutional Networks." *Future Internet* 10.12 (2018): 115.

[15] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017): 834-848.

[16] Weixin Li,(2010). Anomaly Detection in Crowded Scenes . IEEE.

[17] Sultani, W. (2019). Real-world Anomaly Detection in Surveillance Videos. IEEE.