Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره دوم، پاییز و زمستان ۱۴۰۲، صفحات۱۵-۱

مروری بر روش‌های یادگیری عمیق برای تشخیص خشونت

جواد محمودی[۱]*، حسین نظام‌آبادی پور[۲]

[۱]دانشجوی دکتری گروه مهندسی برق، دانشگاه شهید باهنر کرمان، کرمان، ایران

[۲]استادگروه مهندسی برق، دانشگاه شهید باهنر کرمان، کرمان، ایران

چکیده

با رشد بسیار سریع سیستم‌های ویدئویی برای نظارت بر رفتارهای انسانی، تقاضا برای چنین سیستم‌هایی که قادر به تشخیص رخدادهای خشونت‌آمیز به‌صورت خودکار باشند، در حال افزایش است. تشخیص خشونت یکی از حوزه‌های تحقیقاتی فعال در یادگیری ماشین و پردازش تصویر برای جذب محققان جدید است. روش‌های تشخیص خشونت به دودسته عمده تقسیم می‌شوند که عبارت‌اند از: روش‌های یادگیری ماشینی سنتی و روش‌های یادگیری عمیق. در این مقاله، روش‌های یادگیری عمیق مرور شده و گوناگونی روش‌ها و ساختارهای شبکه‌های عصبی عمیق در این خصوص بررسی‌شده است. در ابتدا روش‌های سنتی و عمیق با یکدیگر مقایسه می‌شوند وبرتری روش‌های عمیق بر روش‌های سنتی از جنبه‌های مختلف موردبررسی قرار می‌گیرد. سپس ساختارهای مختلف شبکه‌های عمیق در خصوص تشخیص خشونت بررسی‌شده است. علاوه بر این، مجموعه داده‌های موجود برای تحلیل خشونت در ویدئو نیز معرفی می‌شوند. درنهایت، در مورد تحقیقات انجام‌شده بحث می‌شود که می‌تواند برای گسترش کارهای آینده مفید باشد.

مقاله پژوهشی

تاریخ دریافت:

۱۴۰۱/۰۵/۲۵

تاریخ پذیرش:

۱۴۰۱/۰۸/۳۰

کلیدواژه‌ها:

تشخیص خشونت، یادگیری عمیق، تشخیص رفتار خشونت‌آمیز، سیستم‌های نظارتی، یادگیری ماشین

نویسنده‌مسئول:

Javad.mahmoodi@eng.uk.ac.ir

:10.22034/ABMIR.2022.2900

E-ISSN: 2821-2037

۱- مقدمه

تشخیص خشونت[1] یک موضوع چالش‌برانگیز است که کاربردهای متعدد در بینایی ماشین دارد. این روزها، استفاده از سیستم‌های تشخیص خشونت در شهرهای هوشمند برای نظارت بر شهروندان رو به افزایش است. دوربین‌های نصب‌شده در مکان‌های عمومی مانند مترو، زندان‌ها، ورزشگاه‌ها، مدارس یا خیابان‌ها به نگهبانان کمک می‌کنند تا رفتارهای خطرناک و خشونت‌آمیز را کنترل کنند. علاوه بر این، درخواست رو به رشدی برای طبقه‌بندی ویدئوهای بارگذاری شده از طریق اینترنت برای غلبه بر اثرات مضر صحنه‌های خشونت‌آمیز وجود دارد. به دنبال این کاربردها، سیستم‌های تشخیص خشونت خودکار باید با سریع‌ترین و مؤثرترین روش‌ها پیاده‌سازی شوند.

تشخیص خشونت از میان حجم بالای ویدئوهای ضبط‌شده، کار بسیار دشواری است [۱]. تشخیص خودکار خشونت در سیستم‌های نظارت ویدئویی شامل تشخیص هرگونه برخورد فیزیکی بین دو نفر یا گروهی از افراد است. درواقع، دو رهیافت محبوب برای تشخیص خودکار خشونت وجود دارد: روش‌های سنتی [۲] [۳] [٤] [٥] [٦] [۷] و روش‌های مبتنی بر یادگیری عمیق[2] [۸] [۹] [۱۰] [۱۱]. روش‌های سنتی بر مبنای استخراج ویژگی‌های دستی[3] عمل کرده و به استفاده از ویژگی‌هایی منحصربه‌فرد و مقاوم با توجه به روش خود متکی هستند [۱۲] [۱۳]. از طرف دیگر، روش‌های مبتنی بر یادگیری عمیق بر پایه استفاده از شبکه‌های کانولوشنی سه‌بعدی[4] و حافظه طولانی کوتاه – مدت[5] (LSTM) می‌باشند.

در این مقاله تمرکز ما بر روی روش‌های مبتنی بر یادگیری عمیق برای تشخیص خشونت است و سعی می‌شود این روش‌ها ازنظر ساختار و نحوه عملکرد موردبررسی قرار گیرند و با یکدیگر مقایسه شوند. لازم به ذکر است که تشخیص خشونت زیرشاخه‌ای از روش‌های تشخیص عمل[6] است. بنابراین، مروری جزئی بر روش‌های تشخیص عمل بیان می‌شود و سپس روش‌های مبتنی بر یادگیری عمیق برای تشخیص خشونت به‌طور مفصل مورد ارزیابی قرار می‌گیرند. با توجه به اینکه در طول سالیان متمادی روش‌های بسیار در این حوزه ارائه گردیده است لزوم دسته‌بندی و مقایسه آن‌ها با یکدیگر وجود دارد.

در این مقاله مروری، یک روش سیستماتیک برای مرور روش‌های تشخیص خشونت در ویدئو ارائه می‌گردد که به شرح ذیل است.

- دسته‌بندی روش‌های یادگیری عمیق به‌منظور درک بهتر آن‌ها و مقایسه آن‌ها با یکدیگر.
- بررسی هر مدل ازلحاظ نقاط قوت و ضعف و نوآوری ارائه‌شده در آن.
- بررسی مجموعه داده‌های موجود برای تحلیل روش‌های ارائه‌شده در حوزه تشخیص خشونت و مقایسه آن‌ها با یکدیگر.

ساختار این مقاله این‌گونه است که در بخش دوم مقایسه‌ای بین روش‌های سنتی و عمیق انجام می‌شود که هدف از این مقایسه بیان دلایلی بر لزوم بررسی روش‌های مبتنی بر یادگیری عمیق است. سپس در بخش سوم روش‌های تشخیص فعالیت در انسان بررسی می‌شوند. در بخش چهارم، روش‌های تشخیص خشونت بر مبنای الگوریتم‌های یادگیری عمیق دسته‌بندی‌شده و به‌تفصیل در خصوص ساختار آن‌ها بحث می‌شود. سپس در بخش‌های پنجم و ششم، پایگاه‌های داده مرتبط با تشخیص خشونت و معیارهای ارزیابی معرفی می‌شوند. بحث و پیشنهادهای آتی در بخش هفتم و جمع‌بندی در بخش آخر بیان می‌شود.

۲- مقایسه روش‌های سنتی و یادگیری عمیق

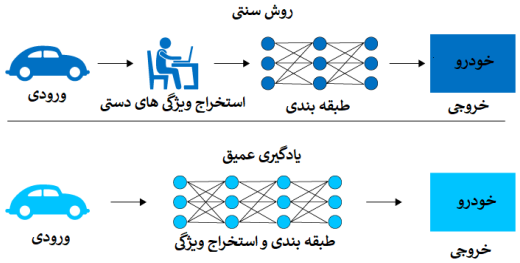
هر مقاله شامل بخش‌های چکیده، کلمات کلیدی، مقدمه، مطالب اصلی، نتیجه‌گیری و مراجع است. سایر بخش‌ها مثل سپاس‌گزاری، ضمایم و زیرنویس‌ها اختیاری است. این بخش‌ها باید در آخر مقاله و قبل از مراجع قرار گیرند.

در این قسمت مقایسه‌ای بین روش‌های سنتی و یادگیری عمیق انجام می‌شود و هدف این است که برتری روش‌های عمیق

[1] Violence detection
[2] Deep learning based methods
[3] Handcrafted features
[4] 3D convolutional neural network
[5] Long short-term memory
[6] Action recognition

موردبررسی قرار گیرد و دلیل تمرکز این مقاله بر روی روش‌های یادگیری عمیق بیان شود.

روش‌های سنتی آن‌هایی هستند که با استفاده از روش‌های دست‌ساز به استخراج ویژگی‌ها می‌پردازند؛ سپس این ویژگی‌ها برای طبقه‌بندی به یک طبقه بند یا شبکه عصبی مصنوعی (کم‌عمق) ارسال می‌شوند. این مدل‌ها برای یادگیری نظارت‌شده مناسب بوده و در هنگام استفاده از آن‌ها می‌بایستی داده‌های به‌خوبی برچسب‌گذاری شوند. عیب اصلی این روش‌ها این است که به‌طور خودکار با تغییرات الگوی ورودی سازگار نیستند. در روش‌های سنتی باید ویژگی‌های کاربردی توسط متخصصان این حوزه شناسایی شوند. سپس از پیچیدگی آن‌ها کاسته شده و درنهایت داده‌ها برای یادگیری به یک طبقه بند ارسال شوند. بزرگ‌ترین مزیت الگوریتم‌های یادگیری عمیق این است که سعی می‌کنند ویژگی‌های سطح بالا را از داده‌ها یاد بگیرند. این امر نیاز به متخصص برای استخراج ویژگی‌ها و کار طاقت‌فرسا برای تحلیل ویژگی‌های استخراج‌شده را از بین می‌برد. **Error! Reference source not found.** نشان‌دهنده مقایسه روش‌های سنتی و یادگیری عمیق است.



شکل (۱): مقایسه روش‌های سنتی و یادگیری عمیق

از دیگر تفاوت‌های روش‌های سنتی و یادگیری عمیق این است که روش‌های عمیق به‌صورت انتها به انتها[1] می‌باشند. از طرف دیگر، روش‌های سنتی نیازمند تقسیم مسئله به بخش‌های مختلف هستند. نتایج بعد از حل هر بخش در مرحله نهایی با یکدیگر ترکیب شوند. به‌عنوان‌مثال، در پیاده‌سازی سیستم‌های تشخیص اشیاء، روش‌های یادگیری عمیق تصویر را به‌عنوان ورودی می‌گیرند و مکان و نام اشیاء را در خروجی ارائه می‌کنند. اما در الگوریتم‌های مبتنی بر یادگیری ماشین که از ماشین‌های بردار پشتیبان[2] (SVMs) استفاده می‌کنند، ابتدا الگوریتم تشخیص شیء، مرزهایی برای اشیاء موجود در تصویر ایجاد می‌کند. سپس تعدادی ویژگی از اشیاء جداشده با مرز استخراج و مورد پردازش قرار می‌گیرد. درنهایت، ویژگی‌ها برای طبقه‌بندی به SVM ارسال می‌شوند.

جنبه دیگری که می‌توان در خصوص مقایسه روش‌های سنتی یادگیری ماشین و یادگیری عمیق بیان کرد، زمان آموزش این دو روش است. الگوریتم‌های یادگیری عمیق به دلیل تعداد زیاد پارامترها زمان زیادی برای آموزش نیاز دارند. درحالی‌که آموزش الگوریتم‌های یادگیری ماشین سنتی چند ثانیه تا چند ساعت طول می‌کشد. این سناریو در مرحله کارایی کاملاً معکوس است. در مرحله کارایی، الگوریتم‌های یادگیری عمیق زمان بسیار کمتری برای اجرا نیاز دارند.

تفسیرپذیری موضوع دیگری است که بین روش‌های یادگیری ماشینی و یادگیری عمیق وجود دارد. درواقع ازنظر ریاضی می‌توانید بفهمید که کدام گره‌ها از یک شبکه عصبی عمیق فعال‌شده‌اند، اما نمی‌دانیم که نورون‌ها قرار است چه چیزی را مدل‌سازی کنند. بنابراین، نمی‌توان نتایج را به‌روشنی تفسیر کرد. این موضوع در مورد الگوریتم‌های یادگیری ماشین صادق نیست و شما توانایی تفسیر رادارید. به‌طورکلی زمانی که قصد دارید روشی را برای پیاده‌سازی یک سیستم با الگوریتم‌های سنتی یادگیری ماشین یا یادگیری عمیق انتخاب کنید به موارد ذیل می‌بایستی توجه داشته باشید.

- مرحله آموزش در روش‌های یادگیری عمیق نیاز به سخت‌افزار قوی‌تر در مقایسه با روش‌های سنتی یادگیری ماشین دارد.
- برای داشتن یک سیستم خوب در حوزه یادگیری عمیق می‌بایستی داده زیاد جهت مرحله آموزش در اختیار داشته باشید.
- قدرت یادگیری عمیق در حل مسائل از قبیل طبقه‌بندی تصویر، تشخیص اشیاء، پردازش زبان در مقایسه با روش‌های سنتی بسیار بالا است.

[1] End to End

[2] Support Vector Machines

در ادامه **Error! Reference source not found.** تعدادی از روش‌های سنتی تشخیص خشونت را با توجه به ویژگی‌های استخراج‌شده بیان می‌کند.

جدول (۱): خلاصه روش‌های سنتی در زمینه تشخیص رفتار خشونت‌آمیز در ویدئو

ردیف	روش پیشنهادشده	نحوه استخراج ویژگی‌ها
۱	توصیف‌گر ViF [۲]	استفاده از دامنه جریان نوری
۲	توصیف‌گر OViF [۳]	استفاده از دامنه و زاویه جریان نوری
۳	توصیف‌گر HOMO [٤]	استفاده از دامنه و زاویه جریان نوری به همراه اعمال مقادیر آستانه متفاوت جهت دودویی کردن تغییرات دامنه و اندازه جریان نوری
٤	توصیف‌گر DiMOLIF [٥]	استفاده از دامنه و زاویه جریان نوری و بدست‌آوردن یک توزیع در اطراف نقاط بااهمیت
٥	روش‌های بینایی کامپیوتر [٦]	تشخیص نقاط موردعلاقه در حوزه زمان و مکان، استفاده از هیستوگرام جریان نوری و هیستوگرام گرادیان جهتی و ترکیب آن‌ها برای استخراج ویژگی از فریم‌ها

به‌طورکلی این روش‌ها سعی در استخراج ویژگی‌ها با استفاده از پردازش دامنه و فاز جریان نوری رادارند. به‌عنوان‌مثال هاسنر و همکاران [۲] با استخراج ویژگی از دامنه جریان نوری توصیف‌گر ViF را ارائه نمودند. روش‌های دیگری از قبیل [۳]، [٤]، [٥] و[٦] با استفاده از دامنه و فاز جریان نوری توصیف‌گرهایی ارائه دادند که به‌مراتب بهتر از روش‌های مبتنی بر دامنه جریان نوری قادر به تشخیص خشونت در ویدئو بودند.

۳- روش‌های تشخیص فعالیت انسان[1] در ویدئو

قبل از پیدایش یادگیری عمیق، روش‌های این حوزه بر مبنای استخراج ویژگی‌های سنتی بودند؛ که از آن جمله می‌توان به روش‌های پیشنهادی در [١٤] و [٧] اشاره کرد. از روش‌های رایج در این حوزه می‌توان به استخراج مسیر[2] و ویژگی در اطراف نقاط موردعلاقه[3] اشاره کرد. سپس، این ویژگی‌ها برای استخراج توصیف‌گر به ازای هر ویدئو استفاده می‌شوند و درنهایت این توصیف‌گر به یک طبقه بند برای دسته‌بندی ارسال می‌شود. ازجمله معایب این روش‌ها می‌توان به زمان‌بر بودن روند پیش‌پردازش اشاره کرد. درواقع، بیشتر این روش‌ها بر مبنای جریان نوری و گرادیان جریان نوری می‌باشند که معمولاً روندهایی با حجم محاسباتی و زمانی بالایی هستند. اما در رویکردهای یادگیری عمیق، آموزش پذیری به‌صورت انتها به انتها است و روند پیش‌پردازش درواقع جای خود را به آموزش شبکه عمیق داده است که یک‌روند زمانی و محاسباتی بالا را دارا است.

بعد از سال ۲۰۱٤، رویکرد تحلیل ویدئو با ورود محققین به حوزه یادگیری عمیق تغییر کرد. ازجمله روش‌های این حوزه می‌توان به مراجع [۱٥] و [١٦] اشاره کرد. در [۱٥]، نویسندگان سعی در استفاده از ساختار شبکه‌های کانولوشنی دوبعدی برای تحلیل ویدئو داشتند و موفق به افزایش دقت نسبت به روش‌های مبتنی بر استخراج ویژگی‌های دستی شدند. اما این روش پیشنهادی قادر نبود به‌خوبی اطلاعات و ویژگی‌های حوزه زمان را پردازش نماید. به همین دلیل کارن سیمونیان[4] و همکاران برای رفع این نقص، روشی را مبتنی بر دوشاخه شبکه عمیق پیشنهاد دادند که ویژگی‌های زمانی و مکانی را به‌صورت مجزا پردازش می‌کرد. سپس ویژگی‌های پردازش‌شده توسط این دوشاخه با یکدیگر ادغام می‌شدند. عیب اصلی این روش آموزش جداگانه دوشاخه است. درواقع شاخه‌ای از شبکه کانولوشنی که وظیفه استخراج ویژگی‌های مکانی را دارد از طریق پایگاه‌های داده بزرگ تصویر، آموزش داده می‌شود و شاخه مربوط به استخراج ویژگی‌های زمانی

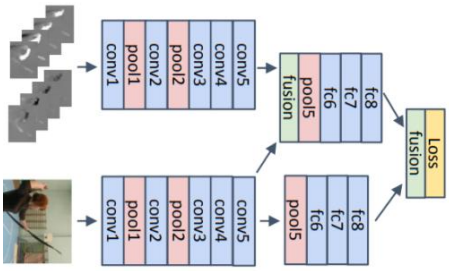
[1] Human action recognition
[2] Trajectory
[3] Interest points
[4] Karen Simonyan

بر روی پایگاه داده‌های مرتبط با ویدئو آموزش داده می‌شود. علاوه بر این، در این شاخه از جریان نوری هر فریم استفاده می‌شود که نیاز به محاسبات زیاد دارد و برای کاربردهای بلادرنگ[1] مناسب نیست. به همین دلیل محققان تمرکز خود را معطوف به روش‌هایی کردند که بدون نیاز به جریان نوری هر فریم قادر به استخراج ویژگی‌های زمانی باشد. ازاین‌رو شبکه‌های عصبی کانولوشنی سه‌بعدی و LSTM پدید آمدند.

ازجمله کارهای اولیه در این حوزه می‌توان به [۱۷] و [۱۸] اشاره کرد. در [۱۷]، از ساختار شبکه‌های عمیق LSTM برای تحلیل ویدئو استفاده‌شده است. ازجمله مزایای این ساختار امکان تحلیل دنباله‌ای از فریم‌های ویدئویی با طول متغیر است. اما این روش پیشنهادی، پیچیدگی محاسباتی بالایی دارد و درنتیجه مدت‌زمان تحلیل یک ویدئو نسبت به روش‌های قبلی در حوزه یادگیری عمیق بیشتر است. در[۱۸]، شبکه کانولوشنی‌سه‌بعدی پیشنهاد گردید؛ که قادر بود علاوه بر استخراج ویژگی‌هایی مکانی، ویژگی‌های زمانی را نیز استخراج کند.

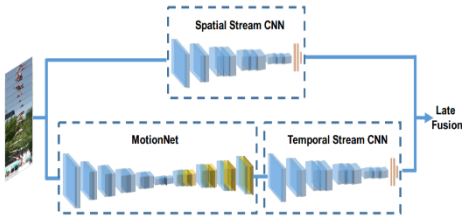
در سال ۲۰۱۶، کریستوف فایختنهوفر[2] و همکاران، شبکه‌های کانولوشنی سه‌بعدی را در ساختار دوشاخه‌ای پیشنهاد دادند [۱۹]. درواقع آن‌ها برای تمایز حرکات مشابه از قبیل شانه کردن سر و مسواک زدن در مکان‌های مختلف از ویدئو این روش را پیشنهاد دادند. در ساختار پیشنهادی ترکیب ویژگی‌های مکانی و زمانی در یک پیکسل مدنظر بود. بنابراین، یک عمل ادغام ویژگی[3] قبل از لایه تمام متصل صورت می‌گرفت که باعث عملکرد بهتر این ساختار ازلحاظ استخراج ویژگی‌های مکانی و زمانی نسبت به روش‌های قبلی شد. اما ضعف بالقوه این ساختار همان استفاده از دو شبکه سه‌بعدی در دوشاخه مجزا است که دارای پیچیدگی‌های محاسباتی و حجم بالای تعداد پارامتر است. **Error! Reference source not found.**) نشان‌دهنده ساختار پیشنهادی آن‌ها است.

در ساختار ارائه‌شده، انواع مختلف ادغام پیاده‌سازی گردیده، که ازجمله می‌توان به ادغام ماکزیمم، ادغام جمع، ادغام دوخطی[4] اشاره کرد. درنهایت روش‌های مختلف ادغام مورد ارزیابی قرارگرفته است.



شکل (۲): ساختار ارائه‌شده در [۱۹]

به‌منظور بهبود ساختار شبکه‌های عصبی کانولوشنی سه‌بعدی، نویسندگان در [۲۰] ساختاری را تحت عنوان شبکه حرکت[5] پیشنهاد دادند، که در این ساختار یک شاخه مخفی برای آموختن جریان نوری معرفی‌شده است. این رویکرد انتها به انتها به محققان این امکان را داد که به‌طور صریح از محاسبات جریان نوری چشم‌پوشی کنند. این موضوع باعث شد که ساختار دوشاخه‌ای برای کاربردهای بلادرنگ مناسب شود. **Error! Reference source not found.**) نشان‌دهنده ساختار ارائه‌شده در [۲۰] است. در این ساختار شبکه حرکت، فریم‌های متوالی ویدئو را به‌عنوان ورودی دریافت می‌کند و یک تخمین از حرکت را به‌عنوان خروجی تولید می‌کند که این خروجی وارد یک شبکه کانولوشنی برای استخراج ویژگی‌های زمانی می‌شود. در ضمن برای استخراج ویژگی‌های مکانی از یک شبکه کانولوشنی موازی با شاخه شبکه حرکت استفاده می‌شود و درنهایت ویژگی‌های استخراج‌شده مکانی و زمانی با یکدیگر ادغام می‌شوند و برای طبقه‌بندی در اختیار لایه آخر قرار می‌گیرند.



شکل (۳): معماری ارائه‌شده در [۲۰]

[1] Real time
[2] Christoph Feichtenhofer
[3] Feature fusion
[4] Bilinear fusion
[5] Motion net

اخیراً تران و همکاران روشی را تحت عنوان شبکه‌های کانولوشنال جداشده از کانال[1] برای تحلیل ویدئو ارائه داده‌اند[۱۰]. نویسندگان درواقع ایده‌ای با استفاده از کانولوشن عمقی[2] و کانولوشن گروهی[3] ارائه داده‌اند؛ که باعث افزایش دقت در تشخیص فعالیت در ویدئو گردیده است. درواقع، ساختار کانولوشن گروهی باعث صرفه جویی در محاسبات می‌شود. با بررسی ساختارهای مختلف برای کانولوشن گروهی نویسندگان به این نتیجه رسیدند که جدا کردن اثر متقابل کانال[4] و اثر متقابل زمانی مکانی[5] می‌تواند باعث افزایش دقت سیستم پیشنهادی شود. از طرف دیگر با به کار بردن ساختار کانولوشن عمقی، به معماری شبکه کانولوشنی جداشده از کانال رسیدند.

٤- روش‌های تشخیص خشونت در ویدئو

روش‌های تشخیص خشونت به دودسته تقسیم می‌شوند: روش‌های سنتی و روش‌های مبتنی بر یادگیری عمیق. در ادامه به روش‌های ارائه‌شده در زمینه یادگیری عمیق پرداخته می‌شود. روش‌های این حوزه را می‌توان به‌صورت زیر دسته‌بندی کرد که در ادامه هر یک از دسته‌بندی‌ها به‌تفصیل موردبررسی قرار می‌گیرند.

- روش‌های مبتنی بر ساختار شبکه‌های عصبی کانولوشنی سه‌بعدی
- روش‌های مبتنی بر شبکه‌های عصبی دوبعدی و LSTM
- روش‌های مبتنی بر شاخه‌های موازی

٤-١ روش‌های مبتنی بر ساختار شبکه‌های کانولوشنی

ازجمله روش‌های محبوب برای تشخیص خشونت در ویدئو استفاده از شبکه‌های عصبی کانولوشنی سه‌بعدی است. درواقع شبکه‌های عصبی کانولوشنی سه‌بعدی گسترش‌یافته شبکه‌های کانولوشنی دوبعدی می‌باشند [۱۸] [۲۱]. این شبکه‌ها قادر به کدگشایی ویژگی‌های زمانی و مکانی هستند و با انتخاب ویدئو به‌عنوان ورودی قادر به دسته‌بندی آن می‌باشند. به‌عنوان‌مثال، دینگ[6] و همکاران [۲۲] یک شبکه ۹ لایه را بر روی پایگاه داده مبارزه هاکی آموزش دادند و توانستند از این طریق به‌دقت بهتری نسبت به روش‌های سنتی تشخیص خشونت دست یابند. علاوه بر این به‌منظور افزایش دقت شبکه‌های کانولوشنی سه‌بعدی، ژانگ و همکاران، روشی را مبتنی بر مرکز ثقل هر فریم برای استخراج فریم‌های کلیدی در ویدئو پیشنهاد دادند [۲۳]. از دیگر کارهایی که سعی در استخراج فریم‌های کلیدی در ویدئو به‌منظور افزایش دقت شبکه‌های کانولوشنی سه‌بعدی را داشتند، می‌توان به روش پیشنهادی در [٢٤] اشاره کرد. نویسندگان این مقاله با استخراج محل افراد و توجه به فریم‌هایی که فرد در آنجا قرار دارد؛ سعی در استخراج تعداد محدودی فریم برای شبکه کانولوشنی کردند و از این طریق باعث افزایش دقت این شبکه‌ها در تحلیل خشونت در ویدئو گردیدند. از دیگر تحقیقات این حوزه می‌توان به [۲۵] اشاره کرد. در این تحقیق نویسندگان از شبکه‌های عصبی کانولوشنی سه‌بعدی در یک ساختار خود رمزنگار استفاده کرده‌اند. اگرچه شبکه‌های عصبی سه‌بعدی ارتقاء خوبی را در زمینه پیاده‌سازی سیستم‌های تشخیص خشونت نشان داده‌اند؛ منتها ازلحاظ محاسباتی هزینه‌بر هستند و نیازمند تعداد پارامترهای زیادی هستند.

جدول (۲): دقت روش‌های یادگیری عمیق مبتنی بر ساختار شبکه‌های کانولوشنی

مرجع	پایگاه داده مبارزه هاکی	پایگاه داده صحنه‌های شلوغ	پایگاه داده فیلم‌های اکشن
[۲۲]	۹۱	–	–
[۲۳]	۹۹/٦۲	–	۹۹/۹۷
[٢٤]	۹۷	۹۸	۹۹/۷
[۲۵]	۸۲	–	۹۲
[۲٦]	۹٤/٤	۸۰/۹	۹٦/۵
[۲۸]	۹۸/٦	۹۲/۵	–

استفاده از شبکه‌های عصبی دوبعدی نیز یکی از روش‌های محبوب در حوزه تشخیص خشونت است. درواقع، شبکه‌های عصبی دوبعدی برای استخراج ویژگی‌های مورداستفاده قرار می‌گیرند که این ویژگی‌ها به طبقه بند SVM برای طبقه‌بندی ارسال می‌شوند.

[1] Channel separated convolutional networks
[2] Depth-wise convolution
[3] Group convolution
[4] Channel interactions
[5] Spatiotemporal interactions
[6] Ding

ازجمله این روش‌ها می‌توان به روش پیشنهادی در [۲۶] اشاره کرد که از شبکه معروف[۲۷] Alexnet برای استخراج ویژگی‌ها استفاده کرده است. به‌منظور استخراج ویژگی از فریم‌ها ، اطلاعات مربوط به دامنه و جهت جریان نوری به شبکه Alexnet ارسال و سپس به‌وسیله طبقه بند SVM دسته‌بندی می‌شوند . از دیگر تحقیقات این حوزه می‌توان به [۲۸] اشاره کرد که در آن از شبکه کانولوشنی پیشنهادشده در [۲۹] برای استخراج ویژگی‌های مکانی و زمانی استفاده‌شده است.

Error! Reference source not found. نشان‌دهنده دقت روش‌های یادگیری عمیق مبتنی بر ساختار شبکه‌های کانولوشنی بر روی پایگاه‌های داده مختلف است.

۴-۲ روش‌های مبتنی بر شبکه‌های کانولوشنی دوبعدی و LSTM

از دیگر روش‌های تشخیص خشونت در ویدئو می‌توان به استفاده از LSTM در کنار شبکه‌های کانولوشنی دوبعدی اشاره کرد. به‌عنوان‌مثال، ترائوره[1] و همکاران [۳۰]، با استفاده از شبکه پیشنهادشده در [۳۱] ویژگی‌های مکانی فریم‌ها و جریان نوری را استخراج می‌کنند؛ سپس این ویژگی‌ها به شبکه عصبی بازگشتی[2] برای پردازش و استخراج ویژگی‌هایی زمانی فرستاده می‌شوند. در این تحقیق، شبکه عصبی بازگشتی در کنار شبکه‌های کانولوشنی دوبعدی برای طبقه‌بندی ویدئو مورداستفاده قرارگرفته‌اند. از دیگر روش‌هایی که به ترکیب دو نوع شبکه عصبی عمیق برای طبقه‌بندی ویدئو پرداخته‌اند؛ می‌توان به روش پیشنهادی در [۳۲] اشاره کرد. نویسندگان این مقاله از شبکه پیشنهادشده در [۳۳] استفاده کرده‌اند. سپس ویژگی‌های استخراج‌شده از این شبکه به شبکه LSTM منتقل می‌شود و درنهایت از لایه‌های تمام‌اتصال[3] در انتهای ساختار شبکه پیشنهادی برای تصمیم‌گیری در خصوص داده‌های ورودی استفاده می‌شود. به‌منظور افزایش دقت روش‌هایی قبلی، دیتسانتیا[4] و همکاران [۳۴] فریم‌های ورودی را به یکسری سپس ویژگی‌های مکانی آن‌ها از طریق شبکه کانولوشنی دوبعدی استخراج‌شده و برای تصمیم‌گیری به یک LSTM داده می‌شود. از دیگر روش‌های این حوزه، می‌توان به تحقیق انجام‌شده توسط مجتبی اسد[5] و همکاران اشاره کرد[۳۵]. در این تحقیق نویسندگان علاوه بر استفاده از ساختار ترکیبی شبکه‌های کانولوشنی دوبعدی و LSTM، از ساختار ترکیب ویژگی[6] نیز استفاده کرده‌اند. در ضمن یک بلوک باقیمانده[7] به‌منظور ترکیب این ویژگی‌ها پیشنهاد گردیده است. علاوه بر این، به‌منظور عملکرد بهتر شبکه‌های کانولوشنی دوبعدی و LSTM در [۳۶] از ساختار توجه حوزه مکان[8] استفاده‌شده است. درواقع، نویسندگان به ویژگی‌های استخراج‌شده از شبکه کانولوشنی دوبعدی ساختار توجه را اضافه کرده‌اند تا از این طریق دقت سیستم را افزایش دهند.

به‌منظور تشخیص رفتار خشونت‌آمیز در یک فرد، نویسندگان در [۳۷] با ترکیب روش‌های مبتنی بر تشخیص اشیاء و شبکه‌های LSTM سعی در استخراج ویژگی‌هایی به‌منظور دسته‌بندی ویدئوهای ورودی کردند. به‌منظور ارزیابی روش پیشنهادی، نویسندگان از دیتاست‌های Weizmann [۳۸]، KTH [۳۹] و دیتاست پیشنهادی خود استفاده کرده‌اند. **Error! Reference source not found.** نشان‌دهنده دقت روش‌های یادگیری عمیق مبتنی بر شبکه‌های کانولوشنی دوبعدی و LSTM است.

جدول (۳): دقت روش‌های یادگیری عمیق مبتنی بر شبکه‌های کانولوشنی دوبعدی و LSTM

مرجع	پایگاه داده مبارزه هاکی	پایگاه داده صحنه‌های شلوغ	پایگاه داده فیلم‌های اکشن
[۳۰]	۹۹/۶۲	۹۶/۷۴	-
[۳۲]	۹۶/۳۳	۸۵/۷۱	۱۰۰
[۳۴]	۸۳/۱۹	۷۵/۷۳	۸۸/۷۴
[۳۵]	۹۸/۸	۹۷/۱	۹۹/۱

[1] Traoré
[2] Recurrent neural network
[3] Fully connected layers
[4] Ditsanthia
[5] Mujtaba Asad
[6] Feature fusion
[7] Residual block
[8] Spatial attention module

٤-٣ روش‌های مبتنی بر شاخه‌های موازی

تعدادی از روش‌ها علاوه بر در نظر گرفتن فریم‌های ویدئو به‌عنوان ورودی، از نوع دیگری از داده به‌عنوان ورودی شبکه‌های عمیق استفاده می‌کنند[٤٠]. به‌طور رایج این داده مبتنی بر جریان نوری است. بنابراین برای استفاده از این دو نوع داده به‌عنوان ورودی، دو شبکه عصبی در شاخه‌های موازی با یکدیگر آموزش داده می‌شوند. که یکی از شبکه‌ها فریم‌های ویدئو و دیگری جریان نوری هر فریم را به‌عنوان ورودی در نظر می‌گیرد. این روش درواقع توسط سیمونیان[1] و همکاران [٤١] ارائه گردید. هدف محققان در این تحقیق تشخیص عمل[2] در ویدئو بود. بعدها این روش با استفاده از شبکه عصبی بازگشتی برای تشخیص خشونت در ویدئو مورداستفاده قرار گرفت[٤٢]. از دیگر روش‌های این حوزه، می‌توان به روش پیشنهادی در [٣٠] اشاره کرد. در این تحقیق ویژگی‌های مکانی موجود در فریم‌ها و جریان نوری از طریق شبکه عمیق پیشنهادی در [٣١] استخراج‌شده و سپس ویژگی‌های استخراج‌شده به یک شبکه بازگشتی برای طبقه‌بندی ارسال می‌شود. استفاده از سه شبکه عمیق به‌صورت موازی نیز ازجمله روش‌های موجود در این حوزه است که نویسندگان در [٤٣] ازاین‌روش استفاده کرده‌اند. درواقع علاوه بر فریم‌های ورودی و جریان نوری از جریان شتاب نیز به‌عنوان ورودی استفاده‌شده است. در مرجع [٩] نیز از سه‌شاخه موازی برای استخراج ویژگی‌ها استفاده‌شده است. نویسندگان در این تحقیق ویژگی‌های مرتبط بازمان، زمانی –مکانی و مکانی را از طریق شبکه‌های عصبی عمیق استخراج می‌کنند.

جدول (٤): دقت روش‌های یادگیری عمیق مبتنی بر ساختار شاخه‌های موازی

مرجع	پایگاه داده مبارزه هاکی	پایگاه داده صحنه‌های شلوغ	پایگاه داده فیلم‌های اکشن
[٤٢]	٩٥/٤	٩٧/٩٧	–
[٣٠]	٩٩/٦٢	٩٦/٧٤	–
[٤٣]	٩٣/٩	–	–
[٩]	١٠٠	٩٩/٣٥	١٠٠
[٤٤]	٩٩/٥	–	–

از دیگر روش‌های این حوزه می‌توان به استفاده از سه شبکه عمیق موازی به همراه سازوکار توجه اشاره کرد؛ که باعث افزایش دقت به‌صورت چشمگیری شده است [٤٤]. درواقع یک شاخه از ساختار پیشنهادی با استفاده از سازوکار توجه نرم[3]، سعی در آموختن و استخراج ویژگی‌های مکانی با احتمال خشونت بالا را دارد. شاخه دیگر جریان نوری را به‌عنوان ورودی در نظر می‌گیرد و سعی در استخراج ویژگی‌های زمانی دارد. درنهایت، شاخه سوم سعی در استخراج ویژگی‌های عمومی مکانی دارد.

٤-٤ مقایسه روش‌های مختلف یادگیری عمیق

به‌طورکلی استفاده از روش‌های مبتنی بر شبکه‌های کانولوشنی سه‌بعدی برای تشخیص خشونت یکی از روش‌های رایج در این حوزه است. ساختارهای مبتنی بر شبکه‌های کانولوشنی سه‌بعدی دارای تعداد پارامترهای زیادی می‌باشند. ازاین‌رو محققان سعی در استفاده از شبکه‌های کانولوشنی دوبعدی برای پردازش ویدئو کردند و ساختارهایی با تعداد پارامتر کمتر در مقایسه با شبکه‌های کانولوشنی سه‌بعدی ارائه دادند. اما در این ساختارها پیش‌پردازش‌هایی بر روی ویدئو انجام می‌گیرد که خود زمان‌بر است و از معایب استفاده از روش‌های یادگیری عمیق مبتنی بر شبکه‌های کانولوشنی دوبعدی است. شبکه‌های LSTM نیز از دیگر ساختارهای مناسب برای پردازش ویدئو می‌باشند. اما در مقایسه با ساختار شبکه‌های کانولوشنی سه‌بعدی از سرعت پایین‌تری در پردازش ویدئو برخوردار می‌باشند[٤٥].

٥- پایگاه داده

با توجه به اینکه محققان بسیاری در حوزه تشخیص خشونت و رفتارهای غیرعادی در ویدئو علاقه‌مند به تحقیق بودند، تقاضای فزاینده‌ای برای پایگاه داده‌های عمومی به‌منظور ارزیابی روش‌های پیشنهادی وجود داشت. این امر باعث به وجود آمدن پایگاه داده‌هایی شد. درواقع، پایگاه داده‌ها این حوزه به دودسته تقسیم‌بندی می‌شوند: پایگاه داده‌های صحنه‌های شلوغ[4] و پایگاه

[1] Simonyan
[2] Action recognition
[3] Soft-attention mechanism
[4] Crowded scenes

داده‌های صحنه‌های بدون شلوغی[1]. نوع اول شامل ویدیوهایی است که اعمال خشونت‌آمیز میان دو نفر رخ می‌دهد. نوع دوم حاوی ویدیوهایی است که درگیری فیزیکی بین گروهی از افراد در حال وقوع است. **Error! Reference source not found.** نشان‌دهنده خلاصه‌ای از ویژگی‌های پایگاه داده‌های معروف در حوزه تشخیص خشونت است.

جدول (۵): خلاصه‌ای از پایگاه‌های داده تشخیص خشونت

پایگاه داده	سال انتشار	مراجع استفاده‌شده
مبارزه هاکی	۲۰۱۱	[۸],[٤],[٤٦],[٤٧],[٤٨]
صحنه‌های شلوغ	۲۰۱۲	[۸],[٤],[٥],[٤٩],[٥٠],[٤٨]
فیلم‌های اکشن	۲۰۱۱	[۸],[٤],[٤٦],[٥١],[٣٦],[٤٨]

۵-۱ پایگاه داده مبارزه هاکی

این پایگاه داده به‌عنوان یک مجموعه داده برای ارزیابی روش‌های تشخیص خودکار خشونت درصحنه‌های غیر شلوغ در نظر گرفته‌شده است و شامل ۱۰۰۰ کلیپ ویدیویی از بازی‌های هاکی با رزولوشن ۳٦۰×۲۸۸ است که به‌طور مساوی به کلیپ‌های خشن و غیر خشن تقسیم‌بندی می‌شود. میانگین تعداد فریم‌ها ٤۱ فریم است. **Error! Reference source not found.** فریم‌هایی از صحنه‌های خشونت‌آمیز و غیر خشونت‌آمیز را در این پایگاه داده نشان می‌دهد.

(الف) (ب) (پ)

(ت) (ث) (ج)

شکل (٤): فریم‌های مختلف از پایگاه داده مبارزه هاکی ردیف بالا فریم‌های خشونت‌آمیز ردیف پایین فریم‌های غیر خشونت‌آمیز

۵-۲ پایگاه داده صحنه‌های شلوغ

این پایگاه داده یک مجموعه داده با رزولوشن ۲٤۰ ×۳۲۰ برای ارزیابی سیستم‌های تشخیص خودکار خشونت درصحنه‌های شلوغ است؛ که شامل ۲٤٦ کلیپ با ۱۲۳ صحنه خشونت‌آمیز و ۱۲۳ صحنه غیر خشونت‌آمیز در محیط شلوغ است. کلیپ‌های ویدیویی در استادیوم‌های فوتبال، خیابان‌ها و مدارس ضبط‌شده‌اند. میانگین تعداد فریم‌ها ۹۰ فریم است. **Error! Reference source not found.** چند فریم مختلف از این مجموعه داده را نشان می‌دهد.

(الف) (ب) (پ)

(ت) (ث) (ج)

شکل (۵): فریم‌های مختلف از پایگاه داده صحنه‌های شلوغ ردیف اول فریم‌های خشونت‌آمیز ردیف دوم فریم‌های غیر خشونت‌آمیز

۵-۳ پایگاه داده فیلم‌های اکشن

این پایگاه داده به‌عنوان یک مجموعه داده برای ارزیابی روش‌های تشخیص خودکار خشونت درصحنه‌های غیر شلوغ در نظر گرفته‌شده است. این مجموعه داده شامل ۲۰۰ کلیپ ویدیویی از فیلم‌های اکشن با رزولوشن ۳۲۰ ×۲٤۰ است؛ که به‌طور مساوی به کلیپ‌های خشن و غیر خشن تقسیم‌بندی می‌شوند. میانگین تعداد فریم‌ها ٤۸ فریم است. **Error! Reference source not**

[1] Uncrowded scenes

found.فریم‌هایی از صحنه‌های خشونت‌آمیز و غیر خشونت‌آمیز را در این پایگاه داده نشان می‌دهد.

(الف) (ب) (پ)

(ت) (ث) (ج)

شکل (٦): فریم‌های مختلف از پایگاه داده فیلم‌های اکشن ردیف اول فریم‌های خشونت‌آمیز و ردیف دوم فریم‌های غیر خشونت‌آمیز

٦- روش‌های ارزیابی

در این بخش، پارامترهای که برای ارزیابی عملکرد سیستم‌های تشخیص رفتار خشونت‌آمیز مورداستفاده قرار می‌گیرند به‌تفصیل بررسی می‌شوند.

٦-١ دقت

دقت نسبت تعداد پیش‌بینی‌های صحیح به تعداد کل نمونه‌های آزمایش است. در حقیقت دقت، میزان صحیح آموزش دیدن یک مدل و نحوه عملکرد سیستم را به‌طوری کلی بیان می‌کند. دقت از رابطه زیر محاسبه می‌شود:

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+TN+FN+FP)} \qquad (١)$$

در فرمول **Error! Reference source not found.**، TP، TN، FP و FN عبارت‌اند از:

TP[1]: ویدئو خشونت‌آمیز، به‌درستی خشونت‌آمیز تشخیص داده شود.

TN[2]: ویدئو بدون خشونت، به‌درستی بدون خشونت تشخیص داده شود.

FP[3]: ویدئو بدون خشونت، به‌اشتباه خشونت‌آمیز تشخیص داده شود.FN[4]: ویدئو خشونت‌آمیز، به‌اشتباه بدون خشونت تشخیص داده شود.

حال محاسبات مربوط به نسبت یا نرخ‌های مختلف برحسب مقادیر بالا مطابق روابط زیر بیان می‌شوند.

$$TPR=\frac{TP}{(TP+FN)} \qquad (٢)$$

$$TNR=\frac{TN}{(TN+FP)} \qquad (٣)$$

در روابط **Error! Reference source not found.**و **Error! Reference source not found.** TPR[5] (حساسیت) به معنی نسبتی از موارد مثبت است که آزمایش آن‌ها را به‌درستی به‌عنوان مثبت علامت‌گذاری می‌کند. TNR[6] (تشخیص) به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به‌درستی به‌عنوان منفی علامت‌گذاری می‌کند.

٦-٢ منحنی AUC-ROC

هنگام پیش‌بینی احتمال، هرچه TPR در مقابل FPR بزرگ‌تر باشد، طبقه‌بندی از کیفیت بهره‌مند است. بنابراین، معیار AUC[7] (سطح زیر منحنی) معرفی گردیده است. در حقیقت مشخص‌کننده احتمال یک شیء به کلاس مثبت است و کیفیت طبقه‌بندی را نشان می‌دهد. در حقیقت معیار AUC نشان‌دهنده سطح زیر منحنی ROC[8] (مشخصهٔ عملکرد سیستم) است. و مقدار آن بین صفر و یک است. منحنی ROC نیز توسط ترسیم نسبت TPR برحسب FPR ایجاد می‌شود.

٧- بحث و پیشنهادهای آتی

به‌طورکلی ویژگی‌های استخراج‌شده از ویدئو تأثیر بسیاری بر روی دقت سیستم تشخیص خشونت دارند. در یک صحنه بدون

[1] True Positive
[2] True Negative
[3] False Positive
[4] False Negative

[5] True Positive Rate
[6] True Negative Rate
[7] Area Under Curve
[8] Receiver Operating Characteristics

خشونت حرکت اشیاء متحرک نسبت به زمان و مکان بسیار کمتر است در مقایسه با صحنه‌هایی که شامل رفتار خشونت‌آمیز هستند. امروزه باوجود گسترش روش‌های تشخیص رفتار خشونت‌آمیز در ویدئو، هنوز این روش‌ها با محدودیت‌هایی روبرو هستند. درواقع، مدل کردن ویژگی‌های یک جسم متحرک چالش اصلی روش‌های تشخیص خشونت است. به‌عنوان‌مثال، مدل کردن یک جسم متحرک که به‌طور ناگهانی وارد صحنه‌ای شلوغ می‌شود ازجمله موارد مشکل‌ساز در پیاده‌سازی این سیستم‌ها است. بنابراین برای جمع‌آوری اطلاعات معنادار در مورد رفتار یکشی، باید از ویژگی‌هایی استفاده کرد که در برابر تغییرات صحنه (چرخش، تاری، پس‌زمینه‌های به‌هم‌ریخته، و غیره) مقاوم بوده و همچنین کمتر در معرض تغییرات ظاهری شیء باشند. برای رفع محدودیت‌های که در بالا ذکر شد، محققان به دنبال استفاده از ابزارهای یادگیری عمیق به‌جای استفاده از ابزارهای سنتی در حوزه تشخیص خشونت شدند. درواقع، روش‌های یادگیری عمیق با بهره‌گیری از مجموعه داده‌های جامع و بزرگ برای آموزش، توانایی حل مسئله باقابلیت بهتری نسبت به روش‌های سنتی رادارند. در یادگیری عمیق، به دلیل ساختار عمیق، دستیابی به ظرفیت یادگیری به میزان قابل‌توجهی افزایش‌یافته است.

اگرچه در چند سال گذشته پیشرفت‌های قابل‌توجهی در استفاده از شبکه عمیق برای تشخیص خشونت صورت گرفته است، اما هنوز چالش‌های زیادی در این خصوص وجود دارد. اصلی‌ترین چالش مدل‌های یادگیری عمیق، نیاز به حجم زیادی از داده‌ها است تا این مدل‌ها به عملکرد مطلوب خود برسند. بااین‌حال، حجم زیادی از داده همیشه آسان نیست. و در بعضی از مواقع که مقادیر زیادی داده در مورد یک موضوع وجود دارد، اغلب اوقات برچسب‌گذاری نشده می‌باشند، بنابراین نمی‌توان از آن‌ها برای آموزش هر نوع الگوریتم یادگیری نظارت‌شده‌ای استفاده کرد. لذا استفاده از یادگیری عمیق با عملکرد عالی در کنار داده‌های آموزشی کم ازجمله چالش‌های این حوزه است. البته پیشرفت‌هایی در این زمینه حاصل‌شده است که ازجمله آن‌ها می‌توان به یادگیری انتقالی اشاره کرد، اما هنوز کافی نیست.

از دیگر چالش‌های یادگیری عمیق در حوزه پردازش ویدئو می‌توان به حجم زیاد پارامترهای اشاره کرد. در حقیقت با توجه به این نکته که ورودی شبکه‌های عمیق در این حوزه، ویدئو است تعداد پارامترهای شبکه به طرز چشمگیری افزایش می‌یابد. لذا پیشنهاد استفاده از روش‌هایی برای کاهش تعداد پارامترهای شبکه احساس می‌شود. به‌عنوان‌مثال در [۵۲] ساختاری پیشنهاد گردیده است که با تجزیه عملگرهای خطی به‌عنوان حاصل‌ضرب مجموع عملگرهای خطی ساده‌تر، تعداد پارامترهای شبکه عمیق کاهش‌یافته است.

از طرف دیگر به‌منظور افزایش دقت در شبکه‌های عمیق، مکانیزم توجه[1] نیز موردتوجه محققان است. درواقع، آن‌ها به مدل‌های یادگیری عمیق اجازه می‌دهند تا به بخش خاصی از ویدئو که دارای خشونت است توجه کنند. این بدان معناست که برای یک ویدئو، شبکه عصبی به‌صورت پویا انتخاب می‌کند که به کدام قسمت‌های ویدئو وزن بیشتری دهد و کدام بخش‌ها را برای استفاده بعدی به خاطر بسپارد. ازجمله ایرادات این روش می‌توان به افزایش تعداد پارامترها و افزایش حجم محاسبات اشاره کرد. پیشنهاد می‌شود که در این خصوص ماژول‌هایی با تعداد پارامتر کمتر و سرعت بیشتر پیشنهاد گردد.

۸- جمع‌بندی

با توسعه روزافزون دوربین‌های نظارتی برای کنترل رفتار انسان، نیاز به سیستم‌هایی که به‌طور خودکار رویدادهای خشونت‌آمیز را شناسایی کنند، افزایش‌یافته است. تشخیص رفتار خشونت‌آمیز به موضوع برجسته‌ای در بینایی ماشین تبدیل‌شده است و محققان بسیاری را به خود جذب کرده است. بسیاری از محققین روش‌های مختلفی را برای تشخیص چنین اقداماتی در ویدئوها پیشنهاد کرده اند. هدف اصلی این مقاله بررسی جدیدترین مطالعات در زمینه تشخیص خشونت بود. انواع مختلف روش‌های تشخیص خشونت ویدیویی، که با استفاده از یادگیری عمیق انجام می‌شوند، در این مطالعه موردبررسی قرارگرفته است. در ابتدا، روش‌های سنتی و یادگیری عمیق از دیدگاه کلی موردبررسی قرار گرفتند، سپس روش‌های یادگیری عمیق در این حوزه بر اساس ساختار شبکه

[1] Attention mechanism

عمیق دسته‌بندی شدند. علاوه بر این، پایگاه‌های داده معروف برای ارزیابی روش‌های تشخیص خشونت معرفی گردید. سپس، معیارهای ارزیابی برای روش‌های پیشنهادی موردبحث قرار گرفت و درنهایت، چالش‌ها، مسائل باز و مسیرهای آینده برای تشخیص رفتار خشونت‌آمیز در ویدیو موردبررسی قرار گرفت.

References

[1] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in 2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2014, vol. 2, pp. 478–485.

[2] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 1–6, DOI: 10.1109/CVPRW.2012.6239348.

[3] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using Oriented VIolent Flows," Image and Vision Computing, vol. 48–49. Elsevier Ltd, pp. 37–41, Apr. 01, 2016, DOI: 10.1016/j.imavis.2016.01.006.

[4] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," Expert Syst. Appl., vol. 127, pp. 121–127, Aug. 2019, DOI: 10.1016/j.eswa.2019.02.032.

[5] A. Ben Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," Pattern Recognit. Lett., vol. 92, pp. 62–67, Jun. 2017, DOI: 10.1016/j.patrec.2017.04.015.

[6] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6855 LNCS, no. PART 2, pp. 332–339, DOI: 10.1007/978-3-642-23678-5_39.

[7] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3169–3176, 2011, DOI: 10.1109/CVPR.2011.5995407.

[8] J. Mahmoodi, H. Nezamabadi-pour, and D. Abbasi-Moghadam, "Violence detection in videos using interest frame extraction and 3D convolutional neural network," Multimed. Tools Appl., vol. 81, no. 15, pp. 20945–20961, 2022, DOI: 10.1007/s11042-022-12532-9.

[9] S. M. Mohtavipour, M. Saeidi, and A. Arabsorkhi, "A multi-stream CNN for deep violence detection in video sequences using handcrafted features," Vis. Comput., no. 0123456789, 2021, DOI: 10.1007/s00371-021-02266-4.

[10] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video Classification with Channel-Separated Convolutional Networks," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-October, pp. 5551–5560, Apr. 2019, DOI: 10.1109/ICCV.2019.00565.

[11] M. S. Kang, R. H. Park, and H. M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," IEEE Access, vol. 9, pp. 76270–76285, 2021, DOI: 10.1109/ACCESS.2021.3083273.

[12] G. M. Basavaraj and A. Kusagur, "Vision based surveillance system for detection of human fall," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017, pp. 1516–1520, DOI: 10.1109/RTEICT.2017.8256851.

[13] P. A. Dhulekar, S. T. Gandhe, N. Sawale, V. Shinde, and S. Khute, "Surveillance System for Detection of Suspicious Human Activities at War Field," in 2018 International Conference On Advances in Communication and Computing Technology (ICACCT), 2018, pp. 357–360, DOI: 10.1109/ICACCT.2018.8529632.

[14] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in 2013 IEEE International Conference on Computer Vision, 2013, pp. 3551–3558, DOI: 10.1109/ICCV.2013.441.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks." pp. 1725–1732, 2014, Accessed: Jan. 12, 2022. [Online]. Available: http://cs.stanford.edu/people/karpathy/deepvideo.

[16] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Adv. Neural Inf. Process. Syst., vol. 1, no. January, pp. 568–576, Jun. 2014, Accessed: Apr. 25, 2021. [Online]. Available: http://arxiv.org/abs/1406.2199.

[17] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 677–691, 2017, DOI: 10.1109/TPAMI.2016.2599174.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and

ACM, 2017, vol. 60, no. 6, pp. 84–90, DOI: 10.1145/3065386.
[28] Z. Meng, J. Yuan, and Z. Li, "Trajectory-Pooled Deep Convolutional Networks for Violence Detection in Videos," in Computer Vision Systems, 2017, vol. 10528 LNCS, pp. 437–447, DOI: 10.1007/978-3-319-68345-4_39.
[29] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9912 LNCS, pp. 20–36, 2016, DOI: 10.1007/978-3-319-46484-8_2.
[30] A. Traore, M. A. Akhloufi, A. Traoré, M. A. Akhloufi, A. Traore, and M. A. Akhloufi, "Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, vol. 2020-Octob, pp. 154–159, DOI: 10.1109/SMC42975.2020.9282971.
[31] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 10691–10700, 2019.
[32] A.-M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," in 2019 2nd Scientific Conference of Computer Sciences (SCCS), 2019, pp. 104–108, DOI: 10.1109/SCCS.2019.8852616.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14, 2015.
[34] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj, "Video Representation Learning for CCTV-Based Violence Detection," in 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), 2018, pp. 1–5, DOI: 10.1109/TIMES-iCON.2018.8621751.
[35] M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. He, "Multi-frame feature-fusion-based model for violence detection," Vis. Comput., vol. 37, no. 6, pp. 1415–1431, 2021, DOI: 10.1007/s00371-020-01878-6.
[36] S. A. Sumon, R. Goni, N. Bin Hashem, T. Shahria, and R. M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches," Vietnam J. Comput. Sci., vol. 7, no. 1, pp. 19–40, Oct. 2020, DOI: 10.1142/S2196888820500013.
[37] A. J. Naik and M. T. Gopalakrishna, "Deep-M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 4489–4497, 2015, DOI: 10.1109/ICCV.2015.510.
[19] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 1933–1941, Apr. 2016, DOI: 10.1109/CVPR.2016.213.
[20] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden Two-Stream Convolutional Networks for Action Recognition," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11363 LNCS, pp. 363–378, Apr. 2017, DOI: 10.1007/978-3-030-20893-6_23.
[21] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 4724–4733, 2017, DOI: 10.1109/CVPR.2017.502.
[22] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence Detection in Video by Using 3D Convolutional Neural Networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8888, pp. 551–558, 2014, DOI: 10.1007/978-3-319-14364-4_53.
[23] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3D convolutional neural networks," IEEE Access, vol. 7, pp. 39172–39179, 2019.
[24] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," Sensors (Switzerland), vol. 19, no. 11, Jun. 2019, DOI: 10.3390/s19112472.
[25] T. Z. Ehsan, M. Nahvi, and S. M. Mohtavipour, "DABA-Net: Deep Acceleration-Based AutoEncoder Network for Violence Detection in Surveillance Cameras," Iran. Conf. Mach. Vis. Image Process. MVIP, vol. 2022-Febru, no. February, 2022, DOI: 10.1109/MVIP53647.2022.9738791.
[26] A. S. Keçeli and A. Kaya, "Violent activity detection with transfer learning method," Electron. Lett., vol. 53, no. 15, pp. 1047–1048, Jul. 2017, DOI: 10.1049/el.2017.0970.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Communications of the

pp. 2945–2956, 2017, DOI: 10.1109/TIFS.2017.2725820.
[47] K. Deepak, L. K. P. Vignesh, and S. Chandrakala, "Autocorrelation of gradients based violence detection in surveillance videos," ICT Express, vol. 6, no. 3, pp. 155–159, Sep. 2020, DOI: 10.1016/J.ICTE.2020.04.014.
[48] L. Ye, T. Liu, T. Han, H. Ferdinando, T. Seppänen, and E. Alasaarela, "Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences," Remote Sens. 2021, Vol. 13, Page 628, vol. 13, no. 4, p. 628, Feb. 2021, DOI: 10.3390/RS13040628.
[49] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image Vis. Comput., vol. 48, pp. 37–41, Apr. 2016, DOI: 10.1016/j.imavis.2016.01.006.
[50] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines," https://doi.org/10.1080/08839514.2020.1723876, vol. 34, no. 4, pp. 329–344, Mar. 2020, DOI: 10.1080/08839514.2020.1723876.
[51] S. R. Dinesh Jackson et al., "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," Comput. Networks, vol. 151, pp. 191–200, Mar. 2019, DOI: 10.1016/J.COMNET.2019.01.028.
[52] C. W. Wu, "ProdSumNet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions," no. 1, pp. 1–10, 2018, [Online]. Available: http://arxiv.org/abs/1809.02209.
violence: individual person violent activity detection in video," Multimed. Tools Appl., vol. 80, no. 12, pp. 18365–18380, 2021, DOI: 10.1007/s11042-021-10682-w.
[38] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, vol. 2, pp. 1395-1402 Vol. 2, DOI: 10.1109/ICCV.2005.28.
[39] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 2004, vol. 3, pp. 32-36 Vol.3, DOI: 10.1109/ICPR.2004.1334462.
[40] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11219 LNCS, no. 1, pp. 318–335, 2018, DOI: 10.1007/978-3-030-01267-0_19.
[41] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Adv. Neural Inf. Process. Syst., vol. 1, no. January, pp. 568–576, 2014.
[42] X. Xu, X. Wu, G. Wang, and H. Wang, "Violent Video Classification Based on Spatial-Temporal Cues Using Deep Learning," in 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 2018, vol. 01, pp. 319–322, DOI: 10.1109/ISCID.2018.00079.
[43] Z. Dong, J. Qin, and Y. Wang, "Multi-stream Deep Networks for Person to Person Violence Detection in Videos," Commun. Comput. Inf. Sci., vol. 662, pp. 517–531, 2016, DOI: 10.1007/978-981-10-3002-4_43.
[44] H. Li, J. Wang, J. Han, J. Zhang, Y. Yang, and Y. Zhao, "A novel multi-stream method for violent interaction detection using deep learning," Meas. Control (United Kingdom), vol. 53, no. 5–6, pp. 796–806, 2020, DOI: 10.1177/0020294020902788.
[45] H. Weytjens and J. De Weerdt, "Process Outcome Prediction: CNN vs. LSTM (with Attention)," Lect. Notes Bus. Inf. Process., vol. 397, pp. 321–333, 2020, DOI: 10.1007/978-3-030-66498-5_24.
[46] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 12,

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 2, Autumn & Winter 2023, pp. 1-15

A Review through Deep Learning Techniques for Violence Detection

Javad Mahmoodi[1*], Hossein Nezamabadi-pour[1]

[1]Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

Article Information

Original Research Paper

Received:
2022 August 16

Accepted:
2022 November 21

Keywords:
violence detection, deep learning, violent behavior detection, surveillance systems, machine learning

Corresponding Author*:
Javad.Mahmoodi@eng.uk.ac.ir

Abstract

With the rapid growth of video systems to monitor human behaviors, demands are increased on such systems which can detect violence events automatically. The violence detection is one of the active research area in machine learning and image processing to attract new researchers. The methods of violence detection are divided into two major categories which are traditional machine learning techniques and deep learning methods. In this article, deep learning methods have been reviewed and the variety of methods and structures of deep neural networks have been examined in this area. First, traditional and deep methods are compared with each other, and the superiority of deep methods over traditional methods is investigated from different aspects. Then, different structures of deep networks have been investigated regarding the detection of violence. Moreover, the available datasets for the analysis of violence in video are also introduced. Finally, it is discussed about the conducted research that can be useful for the development of future works.

: 10.22034/ABMIR.2022.2900

E-ISSN: 2821-2037