مقاله پژوهشی

بازشناسی کارای کنش‌های انسانی با محدودکردن فضای جستجو در روش‌های یادگیری عمیق

مریم کوهزادی و نصرالله مقدم چرکری

چکیده: کارایی سیستم‌های بازشناسی کنش‌های انسانی به استخراج بازنمایی مناسب از داده‌های ویدئویی وابسته است. در سال‌های اخیر روش‌های یادگیری عمیق به منظور استخراج بازنمایی فضایی- زمانی کارا از داده‌های ویدئویی ارائه شده است، در حالی که روش‌های یادگیری عمیق در توسعه بعد زمان، پیچیدگی محاسباتی بالایی دارند. همچنین پراکندگی و محدودبودن داده‌های تمایزی و عوامل نویزی زیاد، مشکلات محاسباتی بازنمایی کنش‌ها را شدیدتر ساخته و قدرت تمایز را محدود می‌نماید. در این مقاله، شبکه‌های یادگیری عمیق فضایی و زمانی با افزودن سازوکارهای انتخاب ویژگی مناسب جهت مقابله با عوامل نویزی و کوچک‌سازی فضای جستجو، ارتقا یافته‌اند. در این راستا، سازوکارهای انتخاب ویژگی غیر برخط و برخط، برای بازشناسی کنش‌های انسانی با پیچیدگی محاسباتی کمتر و قدرت تمایز بالاتر مورد بررسی قرار گرفته است. نتایج نشان داد که سازوکار انتخاب ویژگی غیر برخط، منجر به کاهش پیچیدگی محاسباتی قابل ملاحظه می‌گردد و سازوکار انتخاب ویژگی برخط، ضمن کنترل پیچیدگی محاسباتی، منجر به افزایش قدرت تمایز می‌شود.

***کلیدواژه:* بازشناسی کنش‌های انسانی، یادگیری عمیق، فضایی- زمانی، پیچیدگی محاسباتی، سازوکار انتخاب ویژگی.**

۱- مقدمه

روش‌های یادگیری بازنمایی مبتنی بر یادگیری عمیق، هم‌اکنون از بالاترین کارایی در بازشناسی کنش‌های انسانی برخوردارند. از آنجایی که داده‌های ویدئویی در دو بعد فضا و زمان شکل می‌گیرند، ترکیبی از روش‌های یادگیری عمیق بازنمایی داده‌های فضایی و روش‌های یادگیری عمیق بازنمایی داده‌های زمانی برای بازشناسی کنش‌های انسانی در نظر گرفته می‌شود. مهم‌ترین چالش پیش روی مسأله بازشناسی کنش‌های انسانی، چگونگی بازنمایی کنش‌های انسانی در قالب بردار ویژگی است که دربردارنده خصیصه‌های کلیدی متمایزکننده فضایی و زمانی باشد. در یادگیری عمیق بازنمایی داده‌های فضایی روابط موجود در داده‌ها به طور نمایی طی لایه‌های سلسله‌مراتبی افزایش می‌یابند، در حالی که بسیاری از این روابط سودمند نیستند. وجود عوامل نویزی یا داده‌های تکراری، قدرت تمایز و کارایی محاسباتی را کاهش می‌دهد. در یادگیری عمیق بازنمایی داده‌های زمانی نیز، وجود داده‌های تکراری و حجیم در گام‌های زمانی متوالی منجر به ایجاد عوامل نویزی و از دست رفتن جزئیات سودمند می‌شود و به این ترتیب، از قدرت تمایز و کارایی محاسباتی این روش‌ها کاسته می‌گردد. روش‌های یادگیری عمیق در بازشناسی کنش‌های انسانی با چالش‌های اساسی همچون عوامل نویزی فراوان، پیچیدگی محاسباتی [۱] و [۲]، انتساب اشتباه برچسب در بازنمایی زمانی کوتاه‌مدت [۳] تا [۶] و از دست دادن جزئیات سودمند در بازنمایی زمانی درازمدت [۷] مواجه هستند.

در سال‌های اخیر روش‌های متعددی برای یادگیری عمیق بازنمایی فضایی- زمانی کنش‌های انسانی مطرح شده است، اما با بررسی این روش‌ها مشخص می‌شود که اکثر آنها با افزایش هزینه محاسباتی به کمک سخت‌افزارهای قدرتمند توانسته‌اند قدرت تمایزی بازشناسی را افزایش دهند. در حالی که در بسیاری از موارد عدم دسترسی به پردازشگرهای قدرتمند، امکان بهره‌مندی از کارایی بالای آنها را با محدودیت مواجه می‌سازد. در بسیاری از پژوهش‌ها کوچک‌سازی مؤثر فضای جستجو، به منظور کاهش هزینه محاسباتی ضمن حفظ کارایی در نظر گرفته نشده است. بنابراین افزودن سازوکارهای انتخاب ویژگی به شبکه‌های یادگیری عمیق، جهت مقابله با عوامل نویزی و با هدف محدودکردن فضای جستجو، یکی از موضوعات جالب توجه است که علاوه بر کاهش هزینه محاسباتی، می‌تواند منجر به افزایش کارایی بازشناسی کنش‌های انسانی گردد. یکی از روش‌های مؤثری که یادگیری را بر اطلاعات تمایزی متمرکز می‌کند، روش‌های مبتنی بر توجه است [۸] تا [۱۱]. علی‌رغم موفقیت‌های خوبی که روش‌های مبتنی بر توجه در سال‌های اخیر به دست آورده‌اند، این روش‌ها در ضبط روابط زمانی درازمدت و بازشناسی اعمال پیچیده از کارایی قابل قبولی برخوردار نیستند [۹] و اغلب روش‌های مبتنی بر توجه هزینه محاسباتی بالایی دارند [۱۲] تا [۱۷]. با توجه به مطالب ذکرشده، در این مقاله، شبکه‌های یادگیری عمیق فضایی و زمانی با افزودن سازوکارهای انتخاب ویژگی مناسب، جهت متمرکزساختن آنها بر داده‌های تمایزی ضمن کنترل پیچیدگی محاسباتی توسعه یافته‌اند. استفاده از سازوکارهای انتخاب ویژگی می‌تواند بار محاسباتی را با حفظ کارایی مناسب، کاهش دهد و منجر به افزایش کارایی بازشناسی کنش‌های درازمدت و پیچیده گردد. مهم‌ترین نوآوری‌های مطرح‌شده در این مقاله عبارت است از: اعمال سازوکار انتخاب ویژگی غیر برخط و برخط در شبکه‌های فضایی- زمانی عمیق به منظور محدودکردن فضای جستجو و کاهش پیچیدگی محاسباتی و مقایسه کارایی این دو رویکرد.

در ادامه، ابتدا در بخش ۲، پژوهش‌های صورت‌گرفته در حوزه مرتبط مرور می‌شوند و در بخش ۳، روش پیشنهادی مطرح می‌گردد. سپس در بخش ۴، روش پیشنهادی مورد ارزیابی قرار می‌گیرد و در انتها در بخش

این مقاله در تاریخ ۱۳ آذر ماه ۱۳۹۹ دریافت و در تاریخ ۱۲ تیر ماه ۱۴۰۰ بازنگری شد.

مریم کوهزادی (نویسنده مسئول)، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران، (email: Maryam.Koohzadi@modares.ac.ir).

نصرالله مقدم چرکری، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران، (email: Moghadam@modares.ac.ir).

۵ نتیجه‌گیری مطرح می‌شود.

۲- مروری بر پژوهش‌های مرتبط

در سال‌های اخیر، استفاده از روش‌های یادگیری عمیق در بازشناسی کنش‌های انسانی به شکل چشم‌گیری مورد توجه محققان قرار گرفته است. به طور کلی روش‌های بازشناسی کنش‌های انسانی مبتنی بر یادگیری عمیق را بر اساس چگونگی مدل‌سازی بعد زمان می‌توان در دو دسته کلی مورد بررسی قرار داد: ۱) یادگیری عمیق بازنمایی کوتاه‌مدت و ۲) یادگیری عمیق بازنمایی درازمدت. راهبرد یادگیری بازنمایی سطح ویدئو از عوامل تأثیرگذار در پیچیدگی محاسباتی روش‌های بازشناسی کنش‌های انسانی است. روش‌هایی مانند شبکه‌های پیچشی سه‌بعدی و روش‌های دوجریانه کوتاه‌مدت به دلیل مشکلات محاسباتی، ویدئو را به کلیپ‌های کوتاهی تقسیم می‌کنند و برچسب ویدئو را به این کلیپ‌ها اختصاص می‌دهند (نمونه‌برداری عرضی قاب‌ها[1]). این امر منجر به مشکلاتی مانند از دست دادن اطلاعات مهم، دسترسی محدود زمانی و انتساب اشتباه برچسب می‌شود. از این رو بازنمایی‌های کوتاه‌مدت با وجود پیچیدگی محاسباتی کمتر، به دلیل عدم مشاهده کل داده و انتساب اشتباه برچسب، در بازشناسی کنش‌های انسانی دچار سردرگمی می‌شوند و معمولاً از کارایی کمتری برخوردار هستند. در مواجهه با این مشکل، روش‌های مختلفی سعی دارند تا بازنمایی درازمدت و در سطح ویدئو ارائه دهند. گروهی از روش‌ها، مجموعه‌ای از ویژگی‌های یادگیری‌شده کوتاه‌مدت (نمونه‌برداری با چندین نمونه‌گیری از قاب‌ها) را در قالب یک بازنمایی سطح ویدئو با طول ثابت یکپارچه نموده و پس از آن بازنمایی نهایی در یک فضای بازنمایی جدید با استفاده از روش‌های کدگذاری ایجاد می‌گردد. تصمیم‌گیری در سطح ویدئو بر اساس جمع‌آوری بازنمایی کوتاه‌مدت، ممکن است ناکافی و غیر بهینه باشد، حتی اگر کدگذاری با روشی‌هایی مانند LSTM[2] انجام شود. اما مزیت این گروه از روش‌ها نیاز محاسباتی کمتر آنها نسبت به شبکه‌های یکنواخت پیچشی سه‌بعدی است. گروه دیگر از روش‌ها، شبکه‌های یکنواخت پیچشی سه‌بعدی است. آموزش شبکه‌های پیچشی سه‌بعدی درازمدت به دلیل پیچیدگی و تعداد پارامترهای زیاد، به داده‌های آموزشی بسیاری نیاز دارند و نیاز محاسباتی بالایی دارند. از این رو معمولاً از روش شبکه‌های پیچشی سه‌بعدی برای استخراج بازنمایی فضایی- زمانی محلی و کوتاه‌مدت استفاده می‌شود. روش‌های پیچشی سه‌بعدی درازمدت قابلیت آن را دارند تا بر دنباله طولانی از قاب‌ها اعمال شوند، از این رو می‌توانند بازنمایی فضایی- زمانی مؤثرتری را نسبت به سایر روش‌ها یادگیری نمایند، اما نیاز محاسباتی این روش‌ها بسیار زیاد است. در صورتی که مشکل محاسباتی برطرف گردد، بهترین نتایج متعلق به روش‌های پیچشی سه‌بعدی درازمدت است. این روش‌ها اغلب ناچارند تا رزولوشن فضایی را کاهش دهند که دچار کمبود حافظه نشوند، یا با نمونه‌های ناکافی از ویدئو (نمونه‌برداری تنک قاب‌ها)، بازنمایی را ایجاد نمایند و یا آموزش آنها با شروع از وزن‌های تصادفی و از ابتدا نیست، بلکه شبکه‌های دوبعدی را به عنوان پایه یادگیری خود قرار می‌دهند. روش‌های بازگشتی اگرچه در شبکه‌های بسیار عمیق کارایی خوبی دارند اما آنها نیز با محدودیت طول پنجره به دلیل مشکلات محاسباتی روبه‌رو هستند. همچنین این روش‌ها با چالش از دست دادن جزئیات سودمند به دلیل پراکندگی داده‌های تمایزی محدود در دنباله داده‌های درازمدت مواجه هستند.

از آنجایی که چالش پیچیدگی محاسباتی رابطه مستقیم با داده‌های تکراری یا نویز داشته و رابطه معکوسی با قدرت تمایز بازنمایی فضایی- زمانی دارد، در ادامه تکنیک‌های مطرح برای بهبود این چالش‌ها مورد بررسی قرار می‌گیرد. در جدول ۱ روش‌های مطرح در بهبود پیچیدگی محاسباتی و افزایش کارایی بازشناسی کنش‌های انسانی با یکدیگر مقایسه می‌شوند. از جمله تکنیک‌های تأثیرگذار در پیچیدگی محاسباتی، چگونگی نمونه‌برداری قاب‌ها و راهبرد ایجاد بازنمایی زمانی در سطح ویدئو است و مهم‌ترین تکنیک‌های مطرح در افزایش قدرت تمایز بازنمایی، روش آموزش شبکه و استفاده از سازوکار توجه است.

در این جدول مهم‌ترین راهبردهای نمونه‌برداری از قاب‌های ویدئو برای استخراج ویژگی پیچشی مانند نمونه‌برداری عرضی، تنک، چندین نمونه‌گیری و طولی مطرح شده است. منظور از نمونه‌برداری عرضی، انتخاب چندین قاب متوالی و منظور از نمونه‌برداری طولی، انتخاب بخش‌هایی از هر قاب در ویدئو است. در این جدول همچنین مهم‌ترین راهبردهای بازنمایی بعد زمان در سطح ویدئو مانند بازگشتی، پیچشی سه‌بعدی، خودنظارتی و یا سایر روش‌های رمزگذاری مطرح شده است. همچنین روش‌ها از نظر تکنیک آموزش شبکه مانند آموزش بانظارت و یا بدون نظارت، آموزش انتها به انتها و همچنین استفاده یا عدم استفاده از پیش‌آموزش با یکدیگر مورد مقایسه قرار می‌گیرند. همچنین مهم‌ترین راهبردهای استفاده از سازوکار توجه مانند توجه فضایی، زمانی، هم‌جوشی و همچنین نحوه آموزش آنها مورد توجه قرار می‌گیرد.

همان طور که در جدول ۱ مشاهده می‌شود، در سال‌های اخیر روش‌های مبتی بر توجه، نظر پژوهشگران این حوزه را جلب کرده است. اگرچه سازوکار انتخاب ویژگی به طور مستقیم تا کنون در پژوهشی مطرح نشده است، اما راهبرد چگونگی استفاده از سازوکار توجه، به نوعی بیانگر انتخاب ویژگی‌های مهم هستند.

۳- روش پیشنهادی

در روش‌های بازشناسی کنش‌های انسانی مبتنی بر یادگیری عمیق، مزیت محاسباتی جدایی جریان‌های اطلاعاتی ظاهر[3] و حرکت[4] نسبت به سایر روش‌ها بدیهی است. در این روش دنباله تصاویر قاب‌ها و دنباله شار نوری قاب‌ها در شبکه‌های موازی، یادگیری شده و در انتها با اعمال روش‌های هم‌جوشی مناسب، بازنمایی فضایی- زمانی واحدی از دو جریان ظاهر و حرکت، برای بازشناسی کنش‌های انسانی ایجاد می‌گردد. روش پیشنهادی سعی دارد تا با توسعه شبکه یادگیری عمیق فضایی و زمانی دوجریانه، بر برخی چالش‌های مطرح‌شده در مسئله بازشناسی کنش‌های انسانی توسط سازوکارهای انتخاب ویژگی فایق آید. در ادامه معماری کلی شبکه عمیق پیشنهادی در شکل ۱ ارائه شده است. در این شکل PSTN سرواژه عبارت Proposed Spatial Temporal Network است. اندیس a سرواژه Appearance و اندیس m سرواژه Motion می‌باشد. همچنین F نشان‌دهنده قاب RGB و $O.F$ نشان‌دهنده Optical Flow است. H_N^a بیانگر حالت پنهان شبکه PSTN ظاهر در گام زمانی N و H_N^m بیانگر حالت پنهان شبکه PSTN در گام زمانی N برای جریان داده حرکت است. *PSTN از شبکه فضایی[5] غیر برخط، قسمت انتخابی

1. Frames
2. Long Short Term Memory
3. Appearance
4. Motion
5. Spatial Network

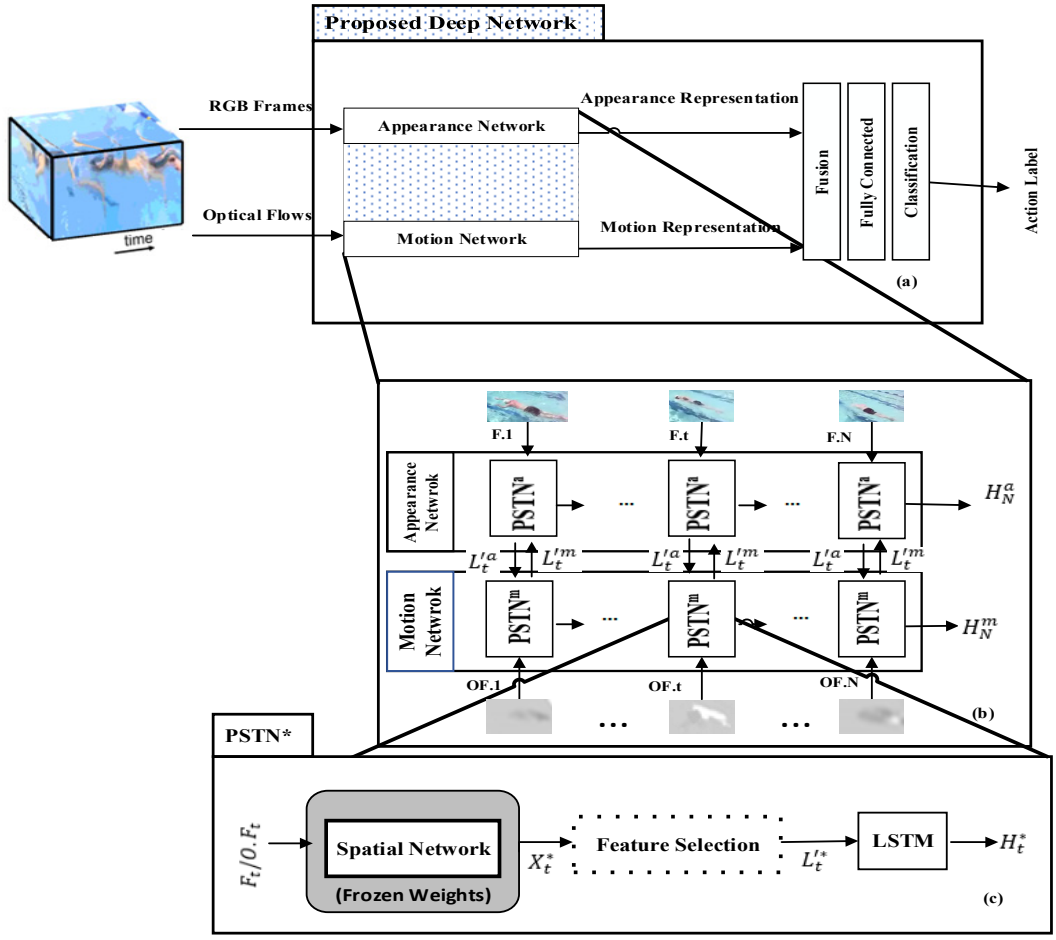
جدول ۱: مقایسه روش‌های اخیر یادگیری عمیق در بازشناسی کنش‌های انسانی بر اساس راهبرد توسعه.

	راهبرد کاهش پیچیدگی محاسباتی										راهبرد افزایش دقت							
	چگونگی نمونه‌برداری قاب‌ها				راهبرد بازنمایی سطح ویدئو						تکنیک آموزش				نوع توجه			
روش	عرضی	تنک	چندین	طولی	بازگشتی	سه‌بعدی	خودنظارتی	سایر کدگذاری	پیش‌آموزش	بدون نظارت	انتها به انتها	فضایی	زمانی	هم‌جوشی	یکپارچه
۲۰۱۷ [۲۱]		*				*			*		*				
۲۰۱۸ [۲۲]		*				*			*		*				
۲۰۱۸ [۲۳]			*					*			*				
۲۰۱۸ [۷]			*					*			*				
۲۰۱۸ [۲۴]			*		*				*		*	*	*		*
۲۰۱۸ [۲۵]			*		*							*	*		
۲۰۱۸ [۱۶]			*			*			*		*		*		*
۲۰۱۸ [۷]		*			*							*	*	*	*
۲۰۱۹ [۲۶]	*					*					*				
۲۰۱۹ [۲۷]		*				*					*				
۲۰۱۹ [۲۸]	*					*			*						
۲۰۱۹ [۲۹]			*		*			*							
۲۰۱۹ [۳۰]					*						*				
۲۰۱۹ [۳۱]		*				*			*		*				
۲۰۱۹ [۳۲]	*	*				*	*			*	*				
۲۰۱۹ [۳۳]			*		*				*	*					
۲۰۱۹ [۱۵]			*		*				*			*	*		*
۲۰۱۹ [۸]		*	*			*			*		*	*	*		*
۲۰۱۹ [۳۴]		*			*							*	*		*
۲۰۱۹ [۱۰]	*				*				*			*			*
۲۰۱۹ [۱۱]		*				*						*	*		
۲۰۱۹ [۱۲]	*		*		*				*			*	*		*
۲۰۱۹ [۱۷]	*				*				*			*			
۲۰۲۰ [۳۵]	*				*							*	*		*
۲۰۲۰ [۹]	*					*					*	*	*		*

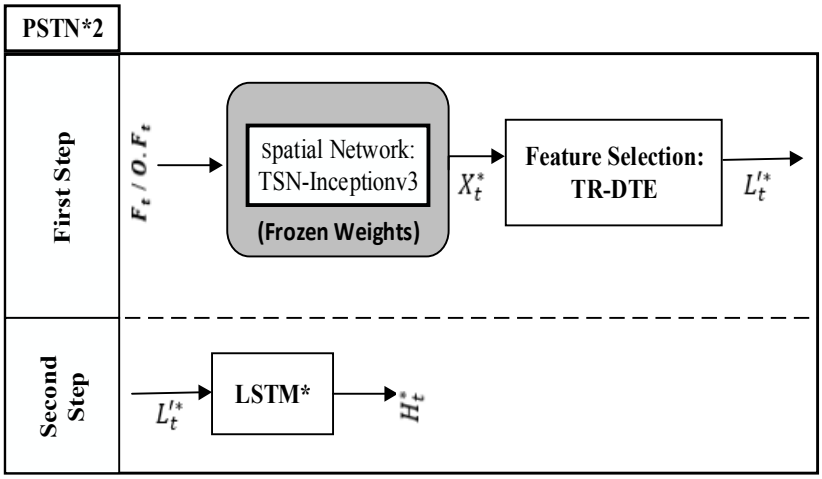
Feature Selection و شبکه LSTM برای یادگیری بازنمایی زمانی تشکیل شده است. این شبکه عمیق، یک شبکه دوجریانه است که از زیرشبکه‌های اصلی ظاهر[1] و حرکت[2] تشکیل شده است. زیرشبکه ظاهر دنباله تصاویر قاب‌ها و زیرشبکه حرکت دنباله شار نوری را در ورودی دریافت نموده و در خروجی بازنمایی فضایی- زمانی کنش را ایجاد می‌نماید. مؤلفه اصلی زیرشبکه‌های ظاهر و حرکت، شبکه پیشنهادی فضایی- زمانی است که در شکل a.۱ ارتباط آنها در زیرشبکه‌های ظاهر و حرکت در گام‌های زمانی متفاوت نشان داده شده و جزئیات مؤلفه اصلی شبکه‌های ظاهر و حرکت در شکل c.۱ آمده است. زیرشبکه فضایی در PSTN بازنمایی فضایی داده‌های خام تصاویر قاب‌ها و یا جریان‌های نوری را ایجاد می‌نماید. پس از آن مؤلفه انتخاب ویژگی[3] قرار دارد که نواحی بااهمیت در بازنمایی فضایی و گام‌های زمانی مهم را مشخص می‌نماید و نهایتاً اطلاعات انتخاب‌شده از بازنمایی‌ها به عنوان ورودی شبکه زمانی در نظر گرفته می‌شود. سازوکار انتخاب ویژگی با محدودکردن فضای جستجو منجر به کاهش نیاز محاسباتی و مقابله با اطلاعات حجیم و تکراری گشته و با تمرکز بر اطلاعات محدود تمایزی، کارایی بازشناسی را بهبود می‌دهد. از این رو نوآوری اصلی این مقاله در راستای تحقق هدف کاهش پیچیدگی محاسباتی با حفظ کارایی مناسب، در قسمت مؤلفه انتخاب ویژگی قرار دارد. در این مقاله، مؤلفه انتخاب ویژگی در دو حالت غیر برخط و برخط در نظر گرفته شده و در قالب شبکه‌های PSTN*۲ و PSTN*۳ طراحی شده و مورد بررسی قرار گرفته است. در حالت غیر برخط دو گام آموزش برای شبکه PSTN*۲ در نظر گرفته می‌شود (شکل ۲). بدین منظور در غالب معماری PSTN*۲، از روش ارائه‌شده در [۱۸] به عنوان سازوکار انتخاب ویژگی غیر برخط استفاده شده است. به این ترتیب پس از آن که مؤلفه انتخاب ویژگی به صورت مستقل آموزش داده شد، اطلاعات بااهمیت انتخاب‌شده به عنوان ورودی گام بعد در نظر گرفته می‌شوند. در صورتی که سازوکار انتخاب ویژگی به صورت غیر برخط آموزش داده شود، نیاز محاسباتی آموزش شبکه کاهش خواهد یافت.

بهره‌مندی از شبکه انتها به انتها، برای بازشناسی کنش‌های انسانی از اهمیت زیادی برخوردار است و اغلب منجر به افزایش کارایی می‌گردد، اگرچه آموزش شبکه به صورت انتها به انتها منجر به افزایش هزینه

1. Appearance Network
2. Motion Network
3. Feature Selection

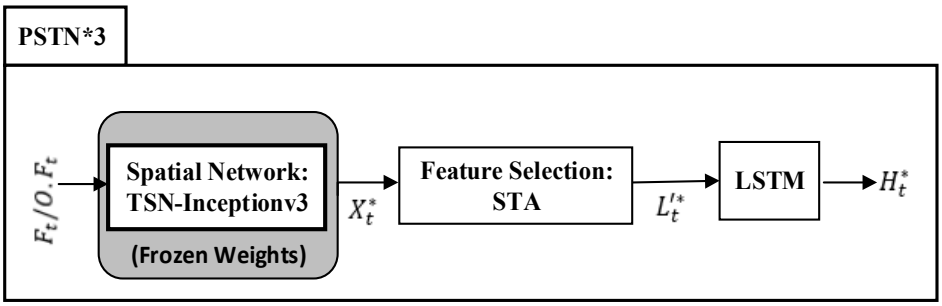


شکل ۱: معماری کلی شبکه عمیق پیشنهادی در شکل a آمده است. مشارکت قسمت‌هایی که با نقطه‌چین نشان داده شده است، در ایجاد بازنمایی و بازشناسی کنش‌های انسانی، انتخابی است. شکل b ارتباط میان زیرشبکه‌های ظاهر، حرکت و همجوشی را در گام‌های زمانی متفاوت نشان می‌دهد. در شکل c، جزئیات شبکه پیشنهادی PSTN* آمده و جریان داده میان قسمت‌های مختلف آن نام‌گذاری شده است. * بیانگر آن می‌باشد که این شبکه بخشی از شبکه ظاهر یا حرکت است.

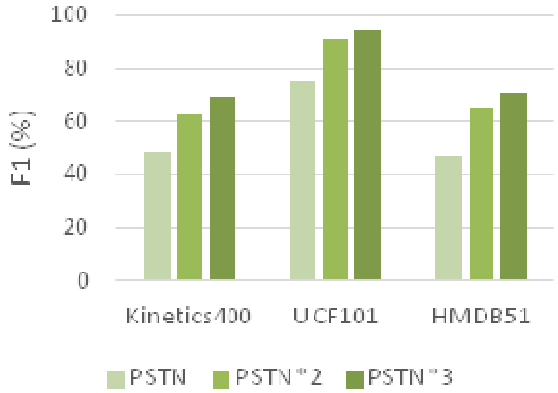


شکل ۲: معماری شبکه PSTN*۲ که از سازوکار انتخاب ویژگی به صورت غیر برخط استفاده شده است. مستطیل خاکستری با عبارت Frozen Weights شامل بخشی از شبکه است که وزن‌های آن در یادگیری انتها به انتهای شبکه ثابت است و در گام جداگانه‌ای آموزش آن انجام شده است. در اینجا ابتدا از شبکه‌ای برای کشف عوامل نویزی استفاده شده و سپس در گام انتخاب ویژگی با نمونه‌برداری بازنمایی فضایی پالایش‌شده فراهم می‌گردد که به عنوان ورودی شبکه LSTM در نظر گرفته می‌شود و شبکه LSTM در قالب معماری کلی ارائه‌شده در شکل ۱ یادگیری می‌شود.

محاسباتی می‌شود. از این رو برای ارتقای کارایی شبکه یادگیری عمیق بازنمایی فضایی- زمانی، در ادامه از سازوکار انتخاب ویژگی به صورت برخط، مطابق شکل ۳ استفاده شده است. بنابراین آموزش شبکه فضایی ابتدا در گام جداگانه‌ای انجام شده و سپس بر بخش‌های مهم اطلاعاتی توسط سازوکار انتخاب ویژگی مبتنی بر توجه تمرکز می‌شود. در این شکل جریان‌های داده‌ای نام‌گذاری شده است. مطابق شکل STA از دو ماژول اصلی توجه فضایی که ضریب اهمیت هر یک از مکان‌های بازنمایی ایجادشده و همچنین توجه زمانی که ضریب اهمیت داده در گام زمانی را مشخص می‌کند، تشکیل شده است. داده به دست آمده به عنوان ورودی LSTM در نظر گرفته می‌شود. در انتها بازنمایی فضایی- زمانی کنش‌های انسانی در قالب معماری کلی ارائه‌شده در شکل ۱ یادگیری می‌شود. بدین منظور در غالب معماری PSTN*۳، از سازوکار توجه فضایی و زمانی



شکل ۳: معماری شبکه PSTN*۳ که از سازوکار انتخاب ویژگی به صورت برخط استفاده شده است. مستطیل خاکستری با عبارت Frozen Weights شامل بخشی از شبکه است که وزن‌های آن در یادگیری انتها به انتهای شبکه ثابت می‌باشد و در گام جداگانه‌ای آموزش آن انجام شده است.



شکل ۴: مقایسه کارایی روش پایه و روش پیشنهادی در حالت برخط PSTN*۳ و غیر برخط PSTN*۲ در معیار F_1.

مطرح‌شده در [۷] با نام STA به صورت انتها به انتها استفاده شده است. به این ترتیب پس از آن که مؤلفه انتخاب ویژگی پیش‌آموزش داده شد و وزن‌های آن مقداردهی اولیه شد، در شبکه PSTN*۳ به صورت برخط برای بازشناسی کنش‌های انسانی استفاده می‌شود.

٤- ارزیابی روش پیشنهادی

در این بخش کارایی روش پیشنهادی در مجموعه داده معروف UCF۱۰۱، Kinetics -۴۰۰ و HMDB۵۱ در حوزه بازشناسی کنش‌های انسانی ارزیابی گردیده است. ابتدا تنظیمات پایه شرح داده شده‌اند و پس از آن کارایی روش‌های پیشنهادی در دو حالت انتخاب ویژگی برخط و غیر برخط مورد بررسی قرار گرفته‌اند و در نهایت با سایر روش‌ها مقایسه شده‌اند.

٤-۱ تنظیمات پایه

اندازه دسته کوچک برابر با ۲۵۶ تنظیم گردیده و شبکه با روش بهینه‌سازی ADAM آموزش داده شده است. شبکه عصبی پیچشی دوجریانه TSN را به عنوان یک روش استخراج بازنمایی استاندارد انتخاب نمودیم. بازنمایی تصاویر ظاهر و حرکت یادگیری‌شده در این شبکه (آخرین لایه ادغام در شبکه TSN با ۱۰۲۴ بعد) به عنوان ورودی روش پیشنهادی مورد استفاده قرار گرفته است. ابعاد تمام متغیرهای پنهان در شبکه LSTM روش پیشنهادی PSTN*۳ برابر ۱۰۲۴ است. از مجموعه داده‌های ویدئویی برابر با ۳۲ ویدئو به طور تصادفی برای هر دسته کوچک و تعداد ۶۴ قاب به طور تصادفی از هر ویدئو و با فاصله مساوی انتخاب گردیده و از مرکز هر قاب، پنجره داده‌ای با اندازه ۲۲۴×۲۲۴ برش داده شده است. در انتها نیز روش پیشنهادی با کاهش گرادیان تصادفی به صورت دسته‌ای کوچک و با تکانه[1] برابر با ۰/۹ و نرخ یادگیری 10^{-2} برای هر دو جریان داده و با ضریب کاهش وزن برابر با 5×10^{-4} در نظر گرفته شده و در محیط TensorFlow [۱۹] پیاده‌سازی گردیده است. برای ارزیابی کارایی، دو معیار صحت[2] و F_1 مطابق روابط زیر، مورد توجه قرار گرفته است

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$
$$Precision = \frac{TP}{TP+FP} \quad (۱)$$
$$Recall = \frac{TP}{TP+FN}$$

$$F_1 = \frac{Precision \times Recall}{Precision + Recall} \quad (۲)$$

که TP، TN، FN و FP به ترتیب مقادیر مثبت درست، منفی درست، منفی غلط و مثبت غلط را نشان می‌دهند.

٤-۲ تحلیل و مقایسه پیچیدگی محاسباتی و کارایی

در شکل ۴ کارایی روش پیشنهادی PSTN در حالت انتخاب ویژگی غیر برخط در روش PSTN*۲ و حالت انتخاب ویژگی برخط در روش PSTN*۳ بر اساس معیار F_1 مورد ارزیابی قرار گرفته است. همان طور که مشاهده می‌شود در هر سه مجموعه داده، روش PSTN*۳ منجر به افزایش کارایی بیشتر شده است. در جدول ۲ کارایی روش‌های مذکور بر اساس معیار صحت، به همراه پیچیدگی آنها مورد ارزیابی قرار گرفته است. اگرچه مجموع پارامترهای هر دو مرحله از روش PSTN*۲ در مقایسه با PSTN*۳، حدود ۴۱٪ بیشتر است، اما به دلیل آن که آموزش سازوکار انتخاب ویژگی در روش غیر برخط به صورت جداگانه انجام می‌شود، مقدار GFLOPs[3] درگام دوم (۷/۳)، به دلیل کوچک‌سازی فضای جستجو در داده ورودی حدود ۲۰٪ کاهش و صحت بازشناسی حدود ۱۲٪ بهبود یافته است. روش PSTN*۳ پس از پیش‌آموزش بخش انتخاب ویژگی با صرف محاسبات بیشتر، صحت بالاتری را نسبت به PSTN*۲ به دست آورد، به شکلی که افزایش صحت مشاهده‌شده در مجموعه داده UCF۱۰۱ برابر با ۲/۸٪، در مجموعه داده HMDB۵۱ برابر با ۴/۹٪ و در مجموعه داده Kinetics -۴۰۰ برابر با ۷/۱٪ بود. همچنین تعداد ابرپارامترهای خاص تنظیم‌شده نیز در روش PSTN*۳ در مقایسه با روش PSTN*۲ کاهش یافته است.

1. Momentum
2. Accuracy
3. Giga Floatingpoint Operations per Second

جدول ۲: مقایسه صحت و پیچیدگی روش پیشنهادی PSTN در حالت غیر برخط PSTN*۲ و برخط PSTN*۳. مقادیر #Parameters و GFLOPs بر مبنای برای روش‌های دومرحله‌ای به صورت جمع هر یک از مراحل نشان داده شده است.

Model			Complexity			Accuracy (٪)		
Method	Two-Step	Pre-Training	#Parameters (M)	GFLOPs	#Hyper Parameters	Kinetics ۴۰۰	UCF۱۰۱	HMDB۵۱
PSTN	✗	✗	۲٫۸	۹٫۸	۰	۵۳٫۹	۸۰٫۵	۵۲٫۰
PSTN*۲	✓	✗	۶٫۱+۲٫۸	۱۸٫۴+۷٫۳	۱۶	۶۳٫۱	۹۲٫۵	۶۶٫۱
PSTN*۳	✗	✓	۵٫۳	۱۲٫۱	۴	۷۰٫۲	۹۵٫۳	۷۱٫۰

جدول ۳: مقایسه روش پیشنهادی PSTN در حالت غیر برخط PSTN*۲ و برخط PSTN*۳.

Convolutional Networks	Two-Step	GFLOP	UCF۱۰۱	HMDB۵۱
[۳۶] TSN	–	۳۳	۹۴٫۲	۶۹٫۴
[۲۱] I۳D	–	۳۴۰	۹۸٫۰	۸۰٫۷
Attention Based LSTM Networks (RGB+Flow)	**Two-Step**	**GFLOP**	**UCF۱۰۱**	**HMDB۵۱**
[۱۴] Collaborative	✗	–	۹۴٫۰	۶۸٫۷
[۱۵] Unified Spatio-Temporal Attention	✗	–	۹۲٫۸	–
[۲۰] STDAN+RGB Difference	✗	۱۸٫۱	۹۱٫۰	۶۰٫۴
[۱۲] TAMNet	✗	–	۹۵٫۷	۷۵٫۳
[۱۷] Attention Mechanism	✓	۱۶٫۵۶	۹۲٫۸	۶۷٫۱
[۱۶] Temporal Attention	✓	–	۹۱٫۸	۶۶٫۱
PSTN*۲	✓	۷٫۳	۹۲٫۵	۶۶٫۱
PSTN*۳	✗	۱۲٫۱	۹۵٫۳	۷۱٫۰

٤-۳ کارایی روش پیشنهادی در شناسایی کنش‌های انسانی درازمدت

در این بخش کارایی انتخاب ویژگی در بازشناسی کنش‌های انسانی با بازه‌های زمانی مختلف مقایسه شده است. در شکل ۵، نتایج در کلیپ‌ها با طول‌های متفاوت مورد ارزیابی قرار گرفته و هر یک از آزمایش‌ها ۱۰ مرتبه تکرار شده است. نتایج نشان می‌دهند که روش پیشنهادی مبتنی بر انتخاب ویژگی در کلیپ‌ها با طول مختلف پایدارتر عمل کرده و در کلیپ‌های درازمدت بهتر از روش پایه عمل کرده است. قابل توجه است که روش پیشنهادی در حالت انتخاب ویژگی غیر برخط PSTN*۲ کارایی بالاتری نسبت به روش پیشنهادی در حالت انتخاب ویژگی برخط در بازشناسی کنش‌های انسانی درازمدت دارد. همچنین مشاهده می‌شود که توانایی بهره‌برداری از اطلاعات تمایزی، در روش‌های با قابلیت انتخاب ویژگی با افزایش طول کلیپ‌ها، افزایش یافته است.

٤-٤ کارایی روش پیشنهادی در شناسایی کنش‌های انسانی پیچیده

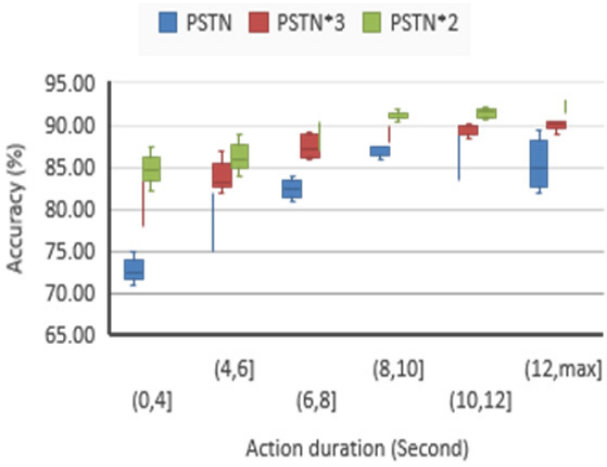
کنش‌های انسانی پیچیده از اطلاعات حرکتی خاصی برخوردارند و همچنین شامل حرکاتی با شباهت زیاد در سایر کلاس‌ها می‌باشند. میزان بهبود کارایی در حالت انتخاب ویژگی برخط و غیر برخط در بازشناسی کنش‌های پیچیده از UCF۱۰۱ در شکل ۶ آمده است. نتایج نشان می‌دهند که میزان بهبود کارایی در حالت انتخاب ویژگی برخط PSTN*۳ در بازشناسی کنش‌هایی که در نواحی کوچک رخ می‌دهند مانند Apply Eye، Apply Lips، shaving bread و brushing teeth، بیشتر از روش پیشنهادی PSTN*۲ است. از طرفی PSTN*۲ با قابلیت بهتر ضبط روابط زمانی درازمدت، در سایر کلاس‌های پیچیده، صحت بالاتری را نسبت به حالت برخط به دست آورده است. اگرچه روش PSTN*۲ کارایی خوبی در بازشناسی کنش‌های انسانی درازمدت از خود نشان داده است، اما در بازشناسی کنش‌های انسانی پیچیده از کارایی بالایی برخوردار نیست. بنابراین نتایج نشان می‌دهند که انتخاب ویژگی برای بازشناسی کنش‌های انسانی پیچیده که اطلاعات حرکتی خاصی را دربردارند، منجر به بهبود کارایی مناسبی می‌شود.

٤-۵ مقایسه با سایر روش‌ها

در جدول ۳ روش پیشنهادی با شبکه‌های رایج سه‌بعدی پیچشی و شبکه‌های بازگشتی مبتنی بر توجه نزدیک به روش پیشنهادی مقایسه شده است. بالاترین کارایی متعلق به روش‌های پیچشی سه‌بعدی با پیچیدگی محاسباتی بسیار بالا است.

روش‌های بازگشتی مبتنی بر توجه از بازنمایی یادگیری‌شده در شبکه‌های پیچشی سه‌بعدی در ورودی خود استفاده می‌کنند. نتایج نشان می‌دهند که صحت روش PSTN*۲ که از سازوکار انتخاب ویژگی غیر برخط برخوردار است، با بهترین روش‌های مبتنی بر توجه دومرحله‌ای قابل رقابت است، اگرچه ماژول توجه در این روش‌ها (مراجع [۱۶] و [۱۷]) با مدل‌های پیچیده‌تری محاسبه می‌شود که هزینه محاسباتی بیشتری را منجر می‌گردد. روش [۱۷] شبکه بسیار پیچیده‌تری نسبت به PSTN*۲ در گام دوم خود دارد، چنان که از شبکه ۵ لایه Convolutional LSTM برای طبقه‌بندی استفاده نموده و ماژول توجه در آن ترکیبی از شبکه تبدیل‌کننده فضایی[1] و LSTM است و بنابراین پیچیدگی محاسباتی آن بالاتر از روش PSTN*۲ می‌باشد. در [۱۶]، توجه زمانی با استفاده از یک شبکه رمزگذار- رمزگشای بازگشتی دولایه استخراج شده که نسبت به روش مطرح در گام اول PSTN*۲، هزینه محاسباتی مشابهی دارد. در گام دوم [۱۶]، از خروجی توجه زمانی در یک شبکه پیچشی عمیق برای بازشناسایی کنش‌ها استفاده شده که نسبت به شبکه LSTM ساده که در گام دوم PSTN*۲ قرار دارد، پیچیدگی محاسباتی بسیار بالاتری دارد.

1. Spatial Transformer



شکل ۵: مقایسه صحت روش پیشنهادی PSTN در حالت برخط PSTN*۳ و غیر برخط PSTN*۲ در کنش‌ها با بازه زمانی متفاوت.

روش پیشنهادی PSTN*۳ در مقایسه با اغلب روش‌های مبتنی بر توجه یکپارچه صحت بالاتری را به دست آورده است. روش ارائه‌شده در [۲۰] از دو مسیر جداگانه برای پردازش بازنمایی تماماً متصل و بازنمایی پیچشی جهت بازشناسی کنش‌ها استفاده می‌نماید که پیچیدگی معماری هر یک از آن مسیرها مشابه روش PSTN*۳ است، اما صحت این روش در مجموعه داده UCF۱۰۱، ۴/۱٪ و در مجموعه داده HMDB۵۱، ۵/۷٪ کمتر از روش PSTN*۳ می‌باشد. در این روش از تفاضل قاب‌های RGB به جای جریان شار نوری استفاده شده است. در [۱۵] سلول‌های عصبی توجه بر چندین حالت از داده‌های ورودی برای کاوش توجه فضایی- زمانی اعمال شده که نیازمند هزینه حافظه بسیار زیادی علاوه بر نیاز محاسباتی است. معماری شبکه به کار رفته در روش [۱۴] در ساختار خود از زیرشبکه‌های عمیق فضایی، زیرشبکه عمیق زمانی و زیرشبکه Collaborative learning به صورت انتها به انتها بهره می‌برد و در نتیجه پیچیدگی محاسباتی بالاتری نسبت به روش پیشنهادی دارد.

اگرچه روش [۱۲] در مقایسه با PSTN*۳ از صحت بالاتری برخوردار است اما معماری بسیار پیچیده‌تری دارد، به طوری که معماری کلی آن علاوه بر شبکه BiDirectional LSTM از یک لایه شبکه پیچشی و دولایه شبکه بازگشتی برای محاسبه توجه استفاده نموده است. از این رو روش PSTN*۳ در یک شبکه بازگشتی دوجریانه با پیچیدگی محاسباتی پایین قادر به بازشناسی نشانه‌های کلیدی در میان اطلاعات فضایی- زمانی و یادگیری بازنمایی با قدرت تمایز بالا است. با توجه به موارد مطرح‌شده، روش‌های پیشنهادی PSTN*۲ و PSTN*۳ با محدودساختن فضای جستجو، تأثیر قابل توجهی در افزایش کارایی و کاهش پیچیدگی محاسباتی شبکه بازگشتی داشته‌اند.

۵- نتیجه‌گیری

بهره‌مندی از سازوکارهای محلی مناسب جهت مقابله با عوامل نویزی و محدودکردن فضای جستجو به طور قابل توجهی در کارایی یادگیری عمیق بازنمایی فضایی- زمانی کنش‌های انسانی تأثیرگذار است. در این مقاله، اعمال روش‌های انتخاب ویژگی به عنوان سازوکار محلی مناسب، جهت کنترل پیچیدگی محاسباتی و کارایی، در دو حالت برخط و غیر برخط در شبکه عمیق مورد بررسی قرار گرفت. مشاهده گردید چنان که از سازوکار انتخاب ویژگی به صورت غیر برخط استفاده شود، با محدودکردن فضای جستجو، منجر به کاهش پیچیدگی محاسباتی و با حذف عوامل نویزی منجر به افزایش کارایی خصوصاً در بازشناسی کنش‌های درازمدت می‌شود. در صورتی که از سازوکار انتخاب ویژگی که پیش‌آموزش داده شده به صورت برخط استفاده شود، منجر به افزایش کارایی بیشتری به خصوص در بازشناسی کنش‌های پیچیده می‌شود. در حالی که روش برخط هزینه محاسباتی بالاتری نسبت به روش غیر برخط دارد.

مراجع

[1] A. Karpathy, *et al.*, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'14,* pp. 1725-1732, Columbus, OH, USA, 23-28 Jun. 2014.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 4489-4497, Santiago, Chile, 7-13 Dec. 2015.

[3] L. Wang, et al., Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, Springer, 2016.

[4] L. Wang, *et al.*, "Temporal segment networks for action recognition in videos," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740- 2755, Nov. 2018.

[5] A. Diba, V. Sharma, and L. Van Gool, *Deep Temporal Linear Encoding Networks*, 2017.

[6] Z. Lan, *et al.*, Deep local video feature for action recognition, 2017.

[7] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. on Image Processing*, vol. 27, no. 3, pp. 1347-1360, Mar. 2017.

[8] Q. Liu, X. Che, and M. Bie, "R-STAN: residual spatial-temporal attention network for action recognition," *IEEE Access*, vol. 7, pp. 82246-82255, 2019.

[9] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3D ConvNets with attention for action recognition," *Pattern Recognition*, vol. 98, Article ID: 107037, Feb. 2020.

[10] Y. Quan, Y. Chen, R. Xu, and H. Ji, "Attention with structure regularization for action recognition," *Computer Vision and Image Understanding*, vol. 187, Article ID: 102794, Oct. 2019.

[11] J. Zhang, H. Hu, and X. Lu, "Moving foreground-aware visual attention and key volume mining for human action recognition," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3, Article ID:. 74, 16 pp., Aug. 2019.

[12] H. Sang, Z. Zhao, and D. He, "Two-level attention model based video action recognition network," *IEEE Access*, vol. 7, pp. 118388-118401, 2019.

[13] S. Sharma, R. Kiros, and R. Salakhutdinov, *Action Recognition Using Visual Attention*, arXiv preprint arXiv:1511.04119, 2015.

[14] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773-786, Mar. 2018.

[15] D. Li, *et al.*, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. on Multimedia*, vol. 21, no. 2, pp. 416-428, Feb. 2018.

[16] H. Zhang, *et al.*, "End-to-end temporal attention extraction and human action recognition," *Machine Vision and Applications*, vol. 29, no. 7, pp. 1127-1142, Oct. 2018.

[17] H. Ge, *et al.*, "An attention mechanism based convolutional LSTM network for video action recognition," *Multimedia Tools and Applications*, vol. 78, pp. 20533-20556, Mar. 2019.

[18] M. Koohzadi and N. M. Charkari, "A context based deep temporal embedding network in action recognition," *Neural Processing Letters*, no. 1, 34 pp., 2020.

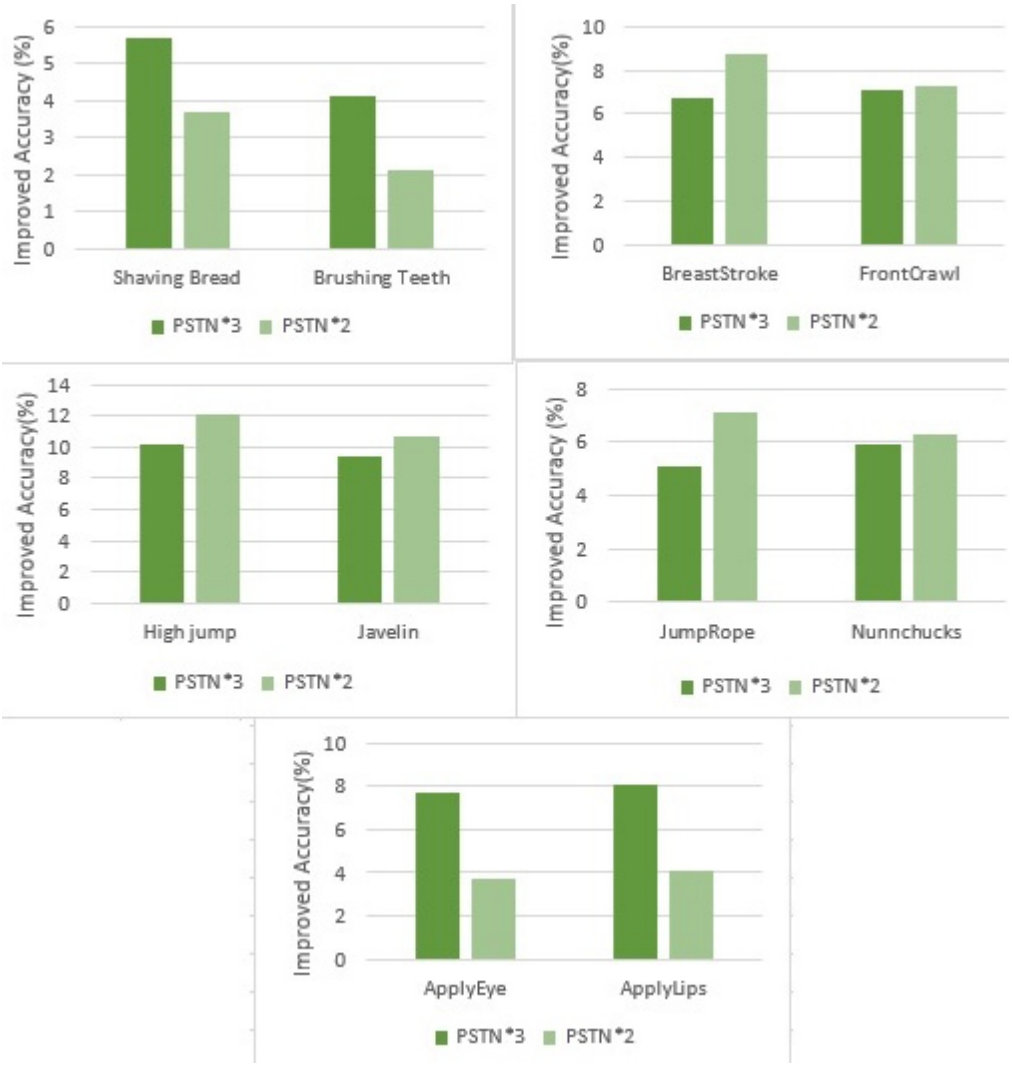
[19] M. Abadi, *et al.*, "Tensorflow: a system for large-scale machine learning," in *Proc. of the 12th USENIX Conf. on Operating Systems Design and Implementation,* pp. 265-283, Savannah, GA, USA, 2-4 Nov. 2016.

[20] Z. Zhang, Z. Lvm C. Gan, and Q. Zhu, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," *Neurocomputing*, vol. 410, pp. 304-316, 14 Oct. 2020.

[21] J. Carreira, A. Zisserman, and Quo Vadis, *Action Recognition? A New Model and the Kinetics Dataset*, arXiv preprint arXiv:1705.07750, 2017.

[22] A. Diba, *et al.*, Spatio-temporal channel correlation networks for action classification, 2018.

[23] J. Zhu, W. Zou, Z. Zhu, and L. Li, "End-to-end video-level representation learning for action recognition," in *Proc. 24th Int. Conf on Pattern Recognition*, pp. 645-650, Beijing, China, 20-24 Aug. 2018.



شکل ۶ مقایسه میزان بهبود صحت روش پیشنهادی در حالت انتخاب ویژگی برخط و غیر برخط در بازشناسی کنش‌های انسانی پیچیده.

[24] Z. Li, K. Gavrilyuk, E.Gavves, M. Jain, C G. Snoekab, "VideoLSTM convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41-50, 20-24 Jan. 2018.

[25] T. Yu, *et al.*, "Joint spatial-temporal attention for action recognition," *Pattern Recognition Letters*, vol. 112, pp. 226-233, Jul. 2018.

[26] Z. Qiu, T. Yao, C. W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 12056-12065, Long Beach, CA, USA, 15-20 Jun. 2019.

[27] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 6202-6211, Seoul, South Korea, 27 Oct.-2 Nov. 2019.

[28] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: motion-augmented RGB stream for action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7874-7883, Long Beach, CA, USA, 15-20 Jun. 2019.

[29] C. Y. Ma, M. H. Chen, Z. Kirab, and G.n AlRegib, "TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition," *Signal Processing: Image Communication*, vol. 1, pp. 76-87, 2019.

[30] B. Pang, K. Zha, H. Cao, C. Shi, and C. Lu, "Deep RNN framework for visual sequential applications," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 423-432, Long Beach, CA, USA, 15-20 Jun. 2019.

[31] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo, "Evolving space-time neural architectures for videos," in *Proc. of the IEEE In. Conf. on Computer Vision*, pp. 1793-1802, Long Beach, CA, USA, 15-20 Jun. 2019.

[32] C. Zhuang, A. Andonian, and D. Yamins, *Unsupervised Learning from Video with Deep Neural Embeddings*, arXiv preprint arXiv:1905.11954, 2019.

[33] N. Sayed, B. Brattoli, and B. Ommer, Cross and Learn: Cross-Modal Self-Supervision, arXiv preprint arXiv:1811.03879, 2018.

[34] L. Meng, *et al.*, "Interpretable spatio-temporal attention for video action recognition," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision Workshops*, , pp. 1513-1522, Seoul, South Korea, 27-28 Oct. 2019.

[35] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Applied Soft Computing*, vol. 86, Article ID: 105820, Jan. 2019.

[36] L. Wang, *et al.*, "Temporal segment networks: towards good practices for deep action recognition," in *Proc. 14th European Conf.,* pp. 20-36, Amsterdam, The Netherlands, 11-14 October, 2016.

مریم کوهزادی دکتری مهندسی نرم‌افزار را از دانشگاه تربیت مدرس در سال ۱۳۹۹ دریافت نموده، همچنین مدرک کارشناسی ارشد و کارشناسی را به ترتیب در رشته هوش مصنوعی و مهندسی نرم‌افزار از دانشگاه الزهرا (س) در سال‌های ۱۳۹۰ و ۱۳۸۸ دریافت کرده است. نامبرده در مقاطع کارشناسی ارشد و دکتری از طریق برگزیدگان علمی پذیرش شده است. علایق اصلی تحقیقاتی و مقالات ایشان در زمینه تحلیل و درک تصاویر، یادگیری عمیق و هوش مصنوعی است.

نصرالله مقدم چرکری مدرک کارشناسی خود را در سال ۱۳۶۵ و در رشته مهندسی کامپیوتر از دانشگاه شهید بهشتی تهران دریافت نمود. همچنین ایشان مدرک ارشد و دکتری را در رشته مهندسی سیستم‌های اطلاعاتی دانشگاه یاماناشی ژاپن به ترتیب در سال‌های ۱۳۷۰ و ۱۳۷۳ دریافت کرد. نامبرده هم اکنون دانشیار دانشکده مهندسی برق و کامپیوتر دانشگاه تربیت مدرس تهران است. ایشان بیش از ۱۲۰ مقاله در کنفرانس‌های بین‌المللی و مجلات معتبر منتشر کرده است. علایق اصلی تحقیقاتی ایشان عبارتند از: تحلیل و بازیابی تصاویر، شبکه‌های پیچیده، الگوریتم‌ها و پردازش‌های موازی و بیوانفورماتیک.