مقاله پژوهشی

بازشناسی کنش انسان از روی تصویر ایستا با استفاده از ژست انسان در شبکه چندشاخه

رقیه یوسفی و کریم فائز

چکیده: امروزه بازشناسی کنش انسان از روی تصویر ایستا به یکی از موضوعات فعال در زمینه بینایی ماشین و شناسایی الگو تبدیل شده است. تمرکز این کار بر روی شناسایی کنش یا رفتار انسان از روی یک تصویر است. برخلاف روش‌های سنتی که از ویدئوها یا دنباله‌ای از تصاویر برای بازشناسی کنش انسان استفاده می‌کنند، یک تصویر ایستا فاقد اطلاعات زمانی است. بنابراین بازشناسی کنش مبتنی بر تصویر ایستا دارای چالش بیشتری نسبت به بازشناسی کنش مبتنی بر ویدئو است. با توجه به اهمیت اطلاعات حرکتی در بازشناسی کنش از روش Im۲flow برای تخمین اطلاعات حرکتی از روی تصویر ایستا استفاده شده است. ساختار پیشنهادی در این مقاله، حاصل ترکیب سه شبکه عصبی عمیق است که تحت عنوان شبکه سه‌شاخه یاد شده است. شبکه اول بر روی تصویر خام رنگی و شبکه دوم بر روی شار نوری پیش‌بینی شده از روی تصویر و شبکه سوم بر روی ژست به دست آمده از انسان موجود در تصویر آموزش می‌بیند. در نهایت تلفیق این سه شبکه عصبی عمیق سبب افزایش دقت بازشناسی کنش انسان شده است. به عبارت دیگر در این مقاله علاوه بر اطلاعات مکانی و زمانی پیش‌بینی شده از اطلاعات ژست انسان نیز برای بازشناسی کنش استفاده شده است زیرا ویژگی ژست برای بازشناسی کنش بسیار حائز اهمیت است. روش پیشنهادی در این مقاله توانسته است به دقت ۹۱٫۸۰ درصد بر روی مجموعه داده Willow۷ action، به دقت ۹۱٫۰۲ درصد بر روی مجموعه داده Pascal voc۲۰۱۲ و به دقت ۹۶٫۸۷ درصد بر روی مجموعه داده Stanford۱۰ دست یابد. با توجه به مقایسه نتایج با روش‌های قبلی متوجه خواهیم شد که روش پیشنهادی بالاترین دقت را بر روی هر سه مجموعه داده نسبت به کارهای اخیر به دست آورده است.

***کلیدواژه:* بازشناسی کنش انسان، پیش‌بینی ژست، شبکه سه‌شاخه، شبکه عصبی عمیق.**

۱- مقدمه

بازشناسی کنش‌های انسان معمولاً به دلیل وجود اطلاعات حرکتی در حوزه ویدئو مطرح می‌شود، زیرا این اطلاعات مزایای زیادی در بازشناسی کنش دارد. علاوه بر این، وجود اطلاعات زمانی سبب مقاوم‌بودن سیستم در برابر نویز پس‌زمینه می‌شود. در حالی که همین کار از طریق تصویر ایستا دارای چالش‌های بیشتری است، زیرا اطلاعات مفیدی که بتواند کنش‌های انسان را از روی تصویر ایستا تشخیص دهد بسیار محدود است و بنابراین منجر به ابهامات درون‌کلاسی می‌شود. علاوه بر این در بعضی از کنش‌های متداول (مانند کتاب‌خواندن، عکس‌گرفتن، کار با کامپیوتر و غیره) تغییرات زیادی وجود ندارد و به عبارت دیگر این کنش‌ها ذاتاً ایستا هستند و نیاز به روش‌های تشخیص بر مبنای نشانه‌های ایستا دارند. در نتیجه این نوع کنش‌ها را می‌توان تنها از روی یک فریم نیز شناسایی کرد. بنابراین روش‌های بازشناسی کنش مبتنی بر تصویر سبب کاهش تعداد فریم‌های مورد پردازش در بازشناسی مبتنی بر ویدئو می‌شود [۱].

اخیراً بازشناسی کنش انسان توجه زیادی در آنالیز رفتار انسان به خود جلب کرده است. علاوه براین تشخیص کنش از روی تصویر ایستا سبب به وجود آوردن داده‌های مفید برای کاربردهای دیگر مانند شاخص گذاری و جستجوی در مقیاس بزرگی از تصاویر آرشیوی، حل مشکلات مربوط به تشخیص شی یا تخمین صحنه، نظارت و بازیابی ویدئو،تشخیص ژست، کاهش تعداد فریم‌های مورد پردازش در ویدئو می‌شود. با توجه به اهمیت و کاربرد این مسأله در مواردی که ذکر شده و همچنین با توجه به این که این مسأله در بینایی ماشین زیاد مورد توجه قرار نگرفته است، در این مقاله به تشخیص کنش از روی تصویر ثابت پرداخته می‌شود [۲].

در روش‌های بازشناسی کنش انسان، هدف پیداکردن ویژگی‌های متمایزکننده از تصویر است و از آنجایی که در تشخیص کنش از روی تصویر ثابت نمی‌توان صرفاً از ویژگی‌های سطح پایین مانند رنگ و بافت یا لبه استفاده کرد، بنابراین از ویژگی‌های سطح بالا مانند بدن، بخش‌های بدن، تعامل انسان و شیء و صحنه برای تشخیص کنش استفاده می‌کنند که این نشانه‌های سطح بالا را می‌توان از طریق ویژگی‌های سطح پایین به دست آورد [۱].

یکی از مشکلات اصلی در بازشناسی کنش انسان از روی تصویر ایستا تغییرات زیاد در ژست و ظاهر انسان و همچنین فقدان اطلاعات زمانی در تصویر است. بنابراین استفاده از روش‌هایی که سبب تخمین اطلاعات حرکتی از روی تصویر ایستا می‌شوند و همچنین استخراج ژست انسان از نشانه‌های مهم در بازشناسی کنش انسان هستند. روش‌های سنتی اغلب به تخمین چنین اطلاعاتی از طریق منابع مختلف تکیه می‌کنند، با این وجود استخراج این ویژگی‌ها همچنان از مسایل حل‌نشده باقی مانده‌اند. در این مقاله برای حل این مشکل از روش یادگیری عمیق استفاده شده است [۲].

موفقیت شبکه‌های پیچشی در بازشناسی، سبب استفاده از این نوع شبکه‌ها در بازشناسی کنش شده و با توجه به اهمیت استفاده از چندین نشانه بصری در معماری‌های چندشاخه، استفاده از این نوع معماری نیز بسیار متداول شده است. بنابراین در این مقاله از ساختار سه‌شاخه[1] برای

این مقاله در تاریخ ۳ اردیبهشت ماه ۱۳۹۸ دریافت و در تاریخ ۲۸ اردیبهشت ماه ۱۳۹۹ بازنگری شد.

رقیه یوسفی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران، (email: rhyousefi83@yahoo.com).

کریم فائز (نویسنده مسئول)، دانشکده مهندسی برق، دانشگاه امیرکبیر، تهران، ایران، (email: kfaez@aut.ac.ir).

1. Three Stream

بازشناسی کنش انسان استفاده شده است. ایده اصلی این مقاله اضافه‌کردن شاخه ژست با الهام از ساختار دوشاخه [۳] برای بازنمایش دقیق‌تر کنش است و از این رو استفاده از روش‌های تخمین ژست در افزایش دقت بازشناسی کنش بسیار مؤثر واقع می‌شود.

روش پیشنهادی در این مقاله بر روی سه مجموعه داده متداول در زمینه بازشناسی کنش مورد ارزیابی قرار گرفته است که شامل مجموعه داده Pascal voc۲۰۱۲ [۴]، مجموعه داده Willow۷ action [۵] و مجموعه داده Stanford۴۰ [۶] می‌باشد که به بهبود قابل توجهی در نتایج نسبت به کارهای اخیر دست یافته‌اند.

این مقاله به صورت زیر سازماندهی می‌شود: در بخش ۲ به مروری بر کارهای قبلی در زمینه بازشناسی کنش می‌پردازیم. در بخش ۳ روش پیشنهادی به طور کامل شرح داده می‌شود. در بخش ۴ مجموعه داده‌ها معرفی و مورد ارزیابی می‌گیرد و در بخش ۵ نتایج حاصل از آزمایش‌ها و مقایسه نتایج با روش‌های قبلی مورد بررسی قرار می‌گیرد. نهایتاً در بخش ۶ نتیجه‌گیری حاصل از روش پیشنهادی بیان می‌شود.

۲- مرور کارهای قبلی

در بازنمایش کنش‌های انسان هدف، پیداکردن ویژگی‌های متمایزکننده از تصویر است و بنابراین روش‌های استخراج ویژگی از روی تصویر را به طور کلی می‌توان به دو دسته تقسیم کرد:

۱) روش‌های استخراج ویژگی دست‌ساز[1]: در این روش استخراج ویژگی توسط یک فرد خبره با استفاده از الگوریتم‌های استخراج ویژگی صورت می‌گیرد. روش‌های استخراج ویژگی دست‌ساز را به طور کلی می‌توان به سه دسته تقسیم‌بندی کرد: دسته اول از ویژگی‌های محلی تصویر مانند SIFT، SURF، HOG، LBP، color و غیره برای استخراج ویژگی استفاده می‌کنند [۷] تا [۹]. دسته دوم مبتنی بر شناسایی ژست انسان از روی تصویر می‌باشند. در این روش‌ها از آشکارسازهایی برای تشخیص بخش‌های بدن انسان و کدگذاری‌کردن آن به ژست‌هایی از انسان استفاده می‌کنند [۱۰]. دسته سوم روش‌های مبتنی بر تعامل انسان و شیء هستند. در این روش‌ها علاوه بر شناسایی انسان برای بازنمایش کنش‌ها، از تعاملات بین انسان و انسان، تعاملات بین انسان و شیء یا روابط میان اشیا هم برای بازشناسی کنش‌ها استفاده می‌کنند.

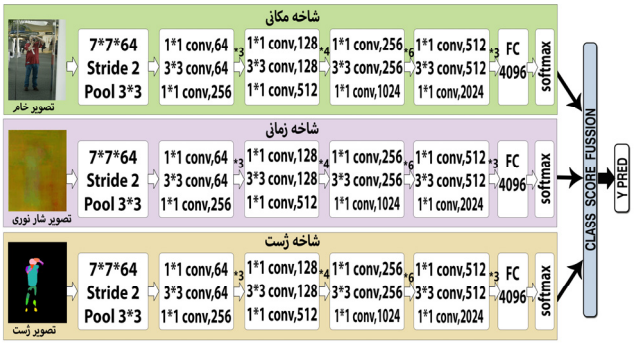
۲) روش‌های یادگیری عمیق: الگوریتم‌های یادگیری عمیق زیرمجموعه‌ای از الگوریتم‌های یادگیری ماشین هستند که هدف آنها کشف چندین سطح از بازنمودهای توزیع‌شده از داده ورودی است. در روش‌های یادگیری عمیق روال استخراج ویژگی به صورت کاملاً اتوماتیک صورت می‌گیرد و نیاز به عامل انسانی برای استخراج ویژگی نمی‌باشد و به عبارت دیگر عمل یادگیری به صورت کاملاً اتوماتیک یا نقطه به نقطه[2] صورت می‌گیرد.

روش‌های قبلی در زمینه بازشناسی کنش از روی تصویر ایستا را می‌توان به سه دسته کلی طبقه‌بندی کرد:

روش‌های مبتنی بر بخش: در این روش‌ها از بخش‌های بدن انسان و بخش‌هایی از اشیای موجود در صحنه برای بازشناسی کنش استفاده می‌شود. یاو و همکارانش [۱۱] صفات کنش را توسط فعل مربوط به کنش‌های انسان مانند سوارکاری و نشستن توصیف می‌کنند. دلایر و همکارانش [۱۲] از تعامل انسان- شیء با استفاده از روش کیسه‌ای از ویژگی‌ها[3] (BOW) برای بازنمایش استفاده کرده‌اند. فی‌فی و همکارانش [۱۳] چندین تعامل را در یک کنش در نظر می‌گیرند که شامل ژست‌های انسان، تعامل انسان و شیء و همچنین رابطه میان اشیا می‌باشد. پرست و همکارانش [۱۴] از تعامل انسان با شیء با استفاده از آشکارسازهای از پیش تعیین شده در محیط تحت نظارت استفاده کرده‌اند و سپس از بخش‌های متفاوتی شامل آشکارساز بالاتنه و صورت برای تشخیص انسان در تصویر استفاده کرده‌اند و سپس اشیایی را که با انسان در تعامل هستند نیز تشخیص داده‌اند. لیانگ و همکارانش [۱۵] برای حل چالش‌هایی مانند تغییرات زیاد در ژست و ظاهر انسان و همچنین فقدان اطلاعات زمانی در تصویر از تلفیق ناحیه انسان و زمینه اطراف آن استفاده کرده‌اند تا اطلاعات را از منابع نویزی متفاوت مانند تشخیص بخش بدن و تشخیص شیء ترکیب کنند. این روش سبب کاهش وابستگی به تخمین دقیق ژست می‌شود ولی نیاز به برچسب‌گذاری دستی بخش‌ها دارد. شارما و همکارانش [۱۶] از تکه‌های محلی تصویر به عنوان بخش استفاده کرده‌اند و با آموزش طبقه‌بندی‌کننده DPM[4] سبب تشخیص کنش‌های انسان شده‌اند. ژانگ و همکارانش [۷] ویژگی‌ها را از طریق روش SIFT به دست آورده‌اند و سپس با بهبود روش VLAD دقت بازشناسی کنش انسان را افزایش داده‌اند. در سال ۲۰۱۶ ژائو و همکارانش [۱۷] ابتدا بخش‌های بدن انسان را توسط یک شبکه Semi-FCN مکان‌یابی می‌کنند، سپس برای هر بخش از بدن از شبکه Res-net به منظور پیش‌بینی معنای کنش مربوط به آن بخش استفاده می‌کنند و در نهایت از طبقه‌بندی‌کننده ماشین بردار پشتیبان برای ترکیب بخش‌های بدن و پیش‌بینی کنش استفاده می‌کنند.

روش‌های مبتنی بر ژست: یکی از روش‌های اصلی در بازشناسی کنش انسان استفاده از اطلاعات ژست و شکل می‌باشد. یانگ و همکارانش [۱۸] از بهبود ژست انسان در یک تصویر به عنوان اطلاعات مفید برای تشخیص کنش استفاده کرده‌اند. این الگوریتم در یک روش تلفیقی آموزش می‌بیند که به طور مشترک ژست‌ها و کنش‌ها را در نظر می‌گیرد. این روش نیاز به حاشیه‌نویسی دستی دارد که مانع از استفاده آن برای مجموعه داده‌های بزرگ می‌شود. ژانگ و همکاران، کنش‌های انسان را توسط ترکیب طبقه‌بندی‌کننده‌های مبتنی بر ژست و طبقه‌بندی‌کننده‌های مبتنی بر زمینه تشخیص داده‌اند که طبقه‌بندی‌کننده‌های مبتنی بر زمینه برای هر کنش با استفاده از اطلاعات پس‌زمینه و پیش‌زمینه آموزش می‌بینند [۱۹]. شارما و همکارانش [۲۰] از هر ژست به طور جداگانه از طریق حاشیه‌نویسی تصاویر سه‌بعدی استفاده کرده‌اند. تانی و همکارانش [۸] روش مقاومی برای مقابله با تغییرات زیاد ژست و انسداد ارائه داده‌اند که استفاده از اطلاعات مکانی سبب افزایش دقت طبقه‌بندی‌کننده شده است، زیرا زمانی که انسان دچار انسداد و تغییرات ژست می‌شود، اطلاعات مکانی، بردار ویژگی ناسازگار ایجاد می‌کند و بنابراین با استفاده از کیسه‌ای از ویژگی‌ها می‌توان اطلاعات مکانی را نادیده گرفت و بازنمایی ویژگی سازگاری به دست آورد، حتی اگر ویژگی محلی در مکان‌های مختلف و با تغییرات ژست همراه باشد. علاوه بر این اگر بخشی از انسان دچار انسداد شود روش کیسه‌ای از ویژگی‌ها می‌تواند بازنمایش را تنها از بخش قابل مشاهده ایجاد کند. بیانگچول و همکارانش [۲۱] از تشخیص ژست با استفاده از طبقه‌بندی‌کننده دولایه استفاده کرده‌اند که در لایه اول از جنگل تصادفی و در لایه دوم از طبقه‌بندی‌کننده چندکلاسی استفاده

1. Hand Craft
2. End to End
3. Bag of Word
4. Deformable Parts Model



شکل ۱: ساختار روش پیشنهادی. ورودی شاخه اول تصویر خام RGB، ورودی شاخه دوم شار نوری و ورودی شاخه سوم ماسک ژست است. بعد از آموزش هر یک از شاخه‌ها به طور جداگانه، در مرحله تست میانگین امتیازهای حاصل از سه شاخه در لایه بیشینه هموار به دست می‌آید و کلاسی که بیشترین امتیاز را دارا باشد کلاس مربوط به تصویر ورودی را نشان می‌دهد. معماری به کار رفته در همه شاخه‌ها، معماری Resnet۵۰ است.

می‌کنند. این روش سبب ایجاد ظاهر متمایز می‌شود و علاوه بر این موقعیت مکانی را نیز در نظر می‌گیرد ولی با افزایش تعداد درخت و تعداد ژست در لایه اول علاوه بر این که کارایی تشخیص را بهبود می‌بخشد سبب افزایش زمان اجرا نیز می‌شود. ژانگ و همکارانش [۲۲] بازشناسی کنش را از طریق آنالیز ژست و تعامل با اشیای در صحنه به دست آورده‌اند. این روش سبب به حداقل رساندن استفاده از جعبه محدود می‌شود و تنها در مرحله آموزش از جعبه محدوده استفاده می‌شود که سبب افزایش کاربرد این روش در سیستم‌های زمان واقعی می‌شود. در بسیاری از موارد این روش قادر به جداسازی دقیق مناطقی است که انسان و شیء با هم در تعامل هستند. گریشیک و همکارانش [۲۳] از همه نشانه‌ها برای بازشناسی فعالیت استفاده کرده‌اند زیرا ژست انسان و اشیای اطراف آنها و نحوه ارتباط با این اشیا و صحنه، نشانه‌های حیاتی برای بازشناسی کنش انسان هستند. گریشیک و همکارانش [۱۰] برای بازشناسی کنش از بخش‌های مهم بدن انسان و جعبه محدوده کل بدن انسان استفاده کرده‌اند. در این روش تشخیص کنش انسان، بدون در نظر گرفتن جعبه محدوده در تصاویر تست انجام شده است.

روش‌های مبتنی بر اطلاعات زمانی: یکی از نشانه‌های مهم در بازشناسی کنش انسان اطلاعات زمانی یا حرکتی در تصویر است. در بازشناسی کنش مبتنی بر ویدئو این اطلاعات از طریق شار نوری توسط دو تصویر متوالی به دست می‌آید. بر خلاف ویدئو، هیچ توالی فریمی در تصویر ایستا وجود ندارد و این مسأله، تشخیص کنش انسان را با چالش بیشتری مواجه می‌کند. علاوه بر این با توجه به فقدان اطلاعات زمانی، اگر هیچ اطلاعات متنی به غیر از انسان در تصویر وجود نداشته باشد، این فقدان تشدید می‌شود. مرجان صفایی و حسن فروش [۲۴] برای حل این مشکل از پیش‌بینی ویژگی‌های حرکتی در تصویر برای بازشناسی کنش استفاده کرده‌اند که به این وسیله توانسته‌اند اطلاعات از دست رفته را جبران کنند. با توجه به این که استفاده از نگاشت برجستگی (SM)[1] برای نشان‌دادن اطلاعات مکانی مفید است، در این روش از ترکیب شار نوری پیش‌بینی شده برای هر پیکسل از تصویر و نگاشت برجستگی ایستا استفاده می‌شود. مرجانه صفایی و حسن‌فروش [۲۵] روش TICNN را برای بازشناسی کنش ارائه داده‌اند که در این روش از دو شبکه CNN متوالی استفاده می‌شود. اولین شبکه CNN برای یادگیری انتقالی و شبکه CNN دوم برای استخراج ویژگی‌های زمانی از تصویر استفاده می‌شود.

هدف از این روش، جبران فقدان اطلاعات زمانی در تصویر بیان‌شده است. در سال ۲۰۱۸ مرجانه صفایی و حسن‌فروش [۲۶] روش Zero-shot را برای بازشناسی ارائه داده‌اند که هدف از ارائه این روش استفاده از هر دو اطلاعات زمانی و مکانی در تصویر ایستا برای جبران فقدان اطلاعات زمانی در تصویر است. روحان گائو و همکارانش [۲۷] روش Im۲Flow را برای جبران فقدان اطلاعات زمانی در تصویر ارائه دادند. این روش شار نوری را از تصویر ایستا پیش‌بینی می‌کند و سپس مدل‌های مجزایی را آموزش می‌دهد تا بتواند اطلاعات را از هر دو شاخه به دست آورد.

۳- روش پیشنهادی

شکل ۱ ساختار روش پیشنهادی را نشان می‌دهد. معماری به کار رفته در این روش ساختار سه‌شاخه با سه نوع ورودی متفاوت است. در هر شاخه از یک شبکه عصبی پیچشی عمیق با معماری مشابه تحت عنوان Resnet۵۰ استفاده شده است. هر یک از این شبکه‌ها به طور مستقل با ورودی‌های مربوط آموزش می‌بینند و در نهایت خروجی حاصل از تلفیق نقشه ویژگی‌های سه‌شاخه بعد از آخرین لایه پیچشی با در نظر گرفتن یک ضریب وزنی برای هر شاخه به لایه تمام متصل انتقال می‌یابد، سپس با میانگین‌گیری امتیازها در لایه بیشینه هموار[2]، کلاس مربوط به تصویر ورودی مشخص می‌شود. به عبارت دیگر، کلاسی که بیشترین امتیاز را به دست آورد بیان‌کننده کلاس مربوط به آن تصویر است.

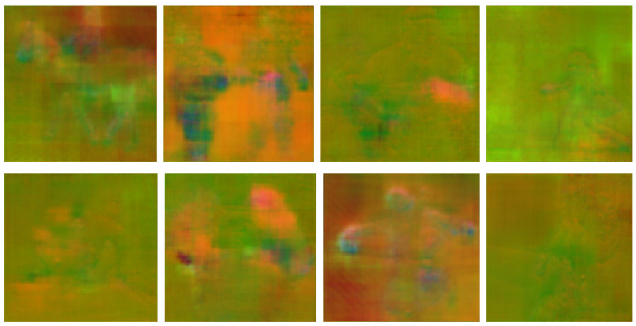
روش تلفیق و نحوه تعیین ضریب وزنی در بخش ۵ توضیح داده خواهد شد. در این ساختار، ورودی شبکه اول، تصویر خام رنگی است. ورودی شبکه دوم شار نوری پیش‌بینی‌شده از تصویر است بدین معنی که پویایی از طریق تصاویر ایستا، تحت عنوان شار نوری است که نحوه استخراج شاری نوری در زیربخش ۳-۱ توضیح داده خواهد شد. ورودی شبکه سوم ژست انسان است که با استفاده از روش SegNet [۲۸] به صورت ماسکی از ژست انسان از تصویر RGB به دست می‌آید که در زیربخش ۳-۲ توضیح داده خواهد شد.

۳-۱ استخراج شار نوری

برآورد حرکت هر پیکسل از دنباله تصاویر یکی از مشکلات در بسیاری از برنامه‌های کاربردی مانند بینایی کامپیوتر و قطعه‌بندی تصویر و طبقه‌بندی شیء و غیره است. به طور کلی شار نوری توصیف‌کننده یک بردار مجزا یا متراکم است که یک بردار جابه‌جایی را به موقعیت خاصی از پیکسل اختصاص می‌دهد که نشان می‌دهد این پیکسل می‌تواند در تصویر دیگر نیز یافت شود. شار نوری عمدتاً از طریق فریم‌های متوالی در ویدئو به دست می‌آید. در این مقاله شار نوری را از روش Im۲Flow [۲۷] پیش‌بینی می‌کنیم. در این روش ابتدا شبکه عصبی، حرکت قبلی را از طریق هزاران ویدئوی بدون برچسب که شامل کنش‌های متفاوت است یاد می‌گیرد و سپس با تزریق تصاویر ایستا به شبکه آموزش دیده شده، تصاویر RGB به شار نگاشت می‌شوند که بیان‌کننده حرکت در تصویر ایستا است. این روش از یک شبکه عصبی پیچشی رمزگذار- رمزگشا استفاده می‌کند که تصاویر ایستا به عنوان ورودی به این شبکه داده می‌شوند و یک شار سه‌کاناله به عنوان خروجی تولید می‌شود. شکل ۲ تصاویر شار نوری برای ورود به شبکه سه‌شاخه را نشان می‌دهد که اگر بخواهیم آن را به صورت بصری مصورسازی کنیم، مطابق شکل ۳ نمایش داده می‌شود.

1. Saliency Map

2. Softmax



شکل ۲: شار نوری پیش‌بینی‌شده از مجموعه داده Willow برای ورود به شاخه زمانی.

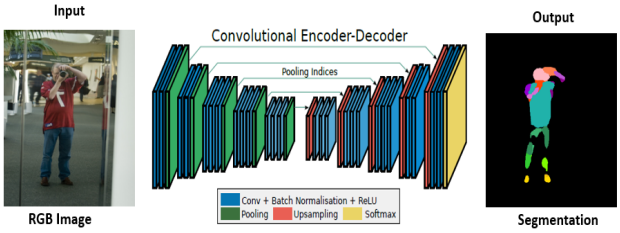
شکل ۳: مصورسازی شار نوری پیش‌بینی شده از تصاویر شکل ۲ از مجموعه داده willow. فلش‌ها بیان‌کننده جهت حرکت در تصویر هستند.

۳-۲ استخراج ماسک ژست از بدن انسان

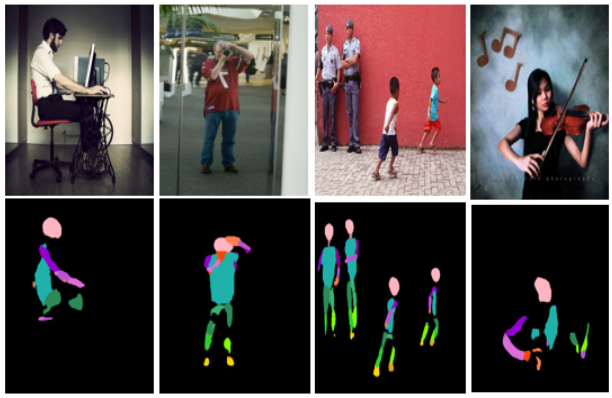
معماری رمزگذار- رمزگشا با بخش up-convolutional برای کارهایی مانند بخش‌بندی معنایی و تخمین عمق و تخمین شار نوری از ویدئو روش مؤثری است. در این کار با استفاده از روش SegNet [۲۸] ماسکی از ژست بدن انسان استخراج می‌شود. از این روش برای قطعه‌بندی بخش‌های بدن استفاده می‌شود که اطلاعات ژست بدن انسان را برای بازشناسی کنش فراهم می‌کند. شکل ۴ نشان‌دهنده معماری SegNet است. بخش مربوط به رمزگذار با شبکه VGG طراحی شده است.

معماری SegNet یک شبکه عصبی عمیق تماماً متصل است که برای قطعه‌بندی معنایی در سطح پیکسل به کار می‌رود. این شبکه دارای یک موتور قطعه‌بندی است که شامل یک شبکه رمزگذار و یک شبکه رمزگشای مربوط است و به دنبال آن یک لایه طبقه‌بندی‌کننده در سطح پیکسل قرار دارد. معماری رمزگذار از نظر توپولوژی برابر با ۱۳ لایه اول پیچشی شبکه VGG۱۶ است که برای طبقه‌بندی شیء طراحی شده است. نقش شبکه رمزگشا نگاشت نقشه ویژگی با وضوح کم را به نقشه ویژگی با وضوح کامل، برای طبقه‌بندی‌کننده در سطح پیکسل است. به عبارت دیگر هدف روش SegNet این است که شبکه رمزگشا نقشه ویژگی‌هایی با وضوح پایین را برای ورودی نمونه‌افزایی می‌کند. خروجی نهایی رمزگشا به یک طبقه‌بندی‌کننده لایه بیشینه هموار تغذیه می‌شود تا حداکثر احتمال را برای هر پیکسل به طور مستقل پیش‌بینی کند.

در این مقاله از معماری SegNet پیش‌آموزش‌دیده بر روی مجموعه داده‌های ویدئویی J-HMDB و MPII استفاده شده است. در این روش موقعیت بخش‌های بدن انسان توسط این شبکه محاسبه می‌شود. سپس از این شبکه آموزش‌دیده طبق شکل ۴ برای استخراج ماسکی از بخش‌های بدن انسان برای مجموعه داده‌های Pascal voc۲۰۱۲، Stanford۱۰ و Willow استفاده شده است. ماسک‌های استخراج‌شده از این روش به عنوان ورودی به شبکه مربوط به ژست تغذیه می‌شود و ژست انسان از



شکل ۴: معماری قطعه‌بندی بخش‌های بدن انسان. پیچشی‌ها با رنگ آبی، پولینگ با رنگ سبز، لایه نمونه‌افزایی با رنگ قرمز و لایه بیشینه هموار با رنگ زرد نشان داده شد.



شکل ۵: استخراج ماسک ژست انسان توسط شبکه SegNet برای مجموعه داده willow.

این طریق به دست آورده می‌شود (شکل ۵). با توجه به نتایج به دست آمده در این روش استفاده از اطلاعات ژست برای بازشناسی کنش انسان بسیار مؤثر واقع شده است.

هدف از ترکیب اطلاعات مکانی و زمانی، متمایزکردن کنش‌های مشابه (مانند کنش مسواک‌زدن و شانه‌کردن) است. در این کنش‌ها اگر دست در موقعیت مکانی یکسانی حرکت کند اطلاعات زمانی یا حرکتی می‌تواند حرکت را تشخیص دهد و اطلاعات مکانی می‌تواند موقعیت (دندان و مو) را تشخیص دهد و ترکیب آنها به تشخیص کنش انسان می‌انجامد. علاوه بر این، اطلاعات مکانی در بازشناسی کنش‌هایی با ژست‌های مشابه (مانند دویدن و راه‌رفتن) سبب افزایش دقت بازشناسی خواهد شد. به خصوص در کنش‌هایی که انسان با شیء خاصی در تعامل باشد تأثیر بیشتری در افزایش دقت بازشناسی خواهد داشت.

٤- مجموعه داده‌ها

روش پیشنهادی بر روی سه مجموعه داده مورد ارزیابی قرار گرفته که تصاویر هر سه مجموعه داده به صورت رنگی ولی دارای کیفیت و ابعاد متفاوت است.

مجموعه داده Pascal voc۲۰۱۲ [۴] شامل تعداد ۴۵۸۸ تصویر به عنوان داده آموزش و اعتبارسنجی و تعداد ۴۵۶۹ تصویر به عنوان داده تست می‌باشد. این مجموعه داده دارای ۱۰ نوع کنش متفاوت است که در هر کلاس تصویری تعداد ۴۰۰ الی ۵۰۰ تصویر به عنوان آموزش و اعتبارسنجی در نظر گرفته می‌شود. تعداد کل تصاویر آموزش ۲۲۹۴ و تعداد داده‌های اعتبارسنجی نیز ۲۲۹۴ می‌باشد. ولی به دلیل برچسب‌نداشتن داده‌های تست در این آزمایش‌ها از داده‌های اعتبارسنجی به عنوان داده‌های تست استفاده شده است. کلاس‌های تصویری در این مجموعه داده شامل استفاده از کامپیوتر، تلفن‌زدن، نواختن موسیقی، دوچرخه‌سواری، اسب‌سواری، دویدن و راه‌رفتن و پریدن، مطالعه‌کردن و عکس‌گرفتن می‌باشد.

جدول ۱: میانگین متوسط صحت (MAP) بر روی مجموعه داده Willow.

PA (%)	شاخه زمانی	شاخه ژست	شاخه مکانی	تلفیق سه‌شاخه (جمع)
تعامل با کامپیوتر	۳۱٫۴۱	۴۰٫۸۴	۹۴٫۰۴	۹۵٫۰۲
عکس‌گرفتن	۲۴٫۲۹	۴۵٫۱۰	۸۴٫۶۷	۸۶٫۷
نواختن موسیقی	۴۷٫۷۱	۵۳٫۱۹	۹۸٫۱۸	۹۸٫۳
دوچرخه‌سواری	۵۵٫۲۹	۵۶٫۰۹	۹۹٫۴۸	۹۹٫۳۲
اسب‌سواری	۲۷٫۱۴	۵۳٫۲۹	۹۹٫۶۴	۹۹٫۶۲
دویدن	۳۶٫۵	۴۹٫۳۴	۷۹٫۰۹	۸۳٫۲۹
راه‌رفتن	۲۷٫۶۴	۳۷٫۱۴	۷۷٫۰۶	۸۰٫۳۸
MAP (%)	۳۵٫۶۵	۴۷٫۸۶	۹۰٫۳۱	۹۱٫۸

مجموعه داده willow۷ action [۵] دارای تعداد ۹۱۱ تصویر و ۷ نوع کنش است. کلاس‌های تصویری در این مجموعه داده شامل استفاده از کامپیوتر، تلفن‌زدن، نواختن موسیقی، دوچرخه‌سواری، اسب‌سواری، دویدن و راه‌رفتن می‌باشد که تعداد ۴۲۷ تصویر به عنوان داده آموزش و تعداد ۴۸۴ تصویر به عنوان تست استفاده می‌شود.

مجموعه داده Stanford۴۰ [۶] شامل ۴۰ نوع کنش متفاوت می‌باشد. تعداد کل تصاویر این مجموعه داده ۹۵۳۲ تصویر است و هر کلاس تصویری شامل ۱۸۰ تا ۳۰۰ تصویر می‌باشد و دارای ۱۰۰ تصویر به عنوان داده آموزش است. این مجموعه داده را می‌توان به دو دسته تقسیم کرد که ۱۱ کنش دارای حرکات بدن و ۲۹ کنش فاقد حرکت بدن هستند.

۵- نتایج آزمایش‌ها

پیاده‌سازی روش پیشنهادی بر روی جعبه‌ابزار پایتورچ با زبان برنامه‌نویسی پایتون و سیستم عامل ویندوز ۱۰ انجام شده است. سخت‌افزار مورد نیاز برای انجام آزمایش‌ها کارت گرافیک انودیا GeforceGTX۱۰۶۰ با ۶ گیگابایت فضای حافظه و ۱۲۸۰ عدد هسته کودا است. کلیه آزمایش‌ها در دو مرحله انجماد و غیر انجماد انجام می‌شود. در حالت انجماد تمام لایه‌های پیچشی در حالت انجماد هستند و وزن‌ها تغییری نمی‌کنند و فقط لایه‌های تماماً متصل آموزش می‌بینند ولی در حالت غیر انجماد وزن تمام لایه‌های پیچشی و لایه‌های تماماً متصل به روز رسانی می‌شوند. اندازه دسته[1] در شبکه تک‌شاخه ۱۶ ولی در شبکه سه‌شاخه ۶ می‌باشد. به دلیل این که تعداد تصاویر در مجموعه داده‌ها کم است از روش داده‌افزایی (مانند برش، چرخش و ...) برای افزایش تعداد داده‌های آموزشی استفاده می‌شود. تابع هزینه به کار رفته در این آزمایش‌ها تابع هزینه آنتروپی متقاطع است. از تابع فعال‌سازی ReLU و نرخ یادگیری $۱۰^{-۴}$ و ۲×۱۰^{-۴} و تابع بهینه‌سازی Adam و SGD استفاده شده است. کلیه آزمایش‌ها برای حالت انجماد به تعداد ۲۰ دوره متوالی می‌باشد و برای حالت غیر انجماد با توجه به نوع مجموعه داده و نوع شاخه، متفاوت است. در شبکه سه‌شاخه همان طور که در بخش ۳ اشاره شد، برای هر یک از شاخه‌ها یک ضریب وزنی در نظر گرفته می‌شود. در این آزمایش‌ها به صورت تجربی به این نتیجه رسیده‌ایم که برای شاخه مکانی ضریب ۰٫۶۹، برای شاخه زمانی ضریب ۰٫۰۶ و برای شاخه ژست ضریب ۰٫۲۵ در نظر گرفته شود.

در این آزمایش‌ها برای ارزیابی دقت از معیار میانگین متوسط صحت (MAP)[2] (۲) استفاده شده است. ابتدا برای تمام کلاس‌ها میانگین صحت (۱) جداگانه محاسبه می‌شود و در نهایت از میانگین صحت میانگین گرفته می‌شود. این معیار نسبت به معیار دقت پایدارتر است زیرا بر اساس اطلاعات بیشتری میزان کارایی سیستم را محاسبه می‌کند. به عنوان مثال این معیار، ترتیبی برای بازشناسی در نظر می‌گیرد و اگر تصویر مرتبط به ابتدای لیست مرتب‌شده بازشناسی نزدیک‌تر باشد به آن امتیاز بیشتری تخصیص داده می‌شود

$$AveP(q_j)=\frac{۱}{|M_j|}\sum_{k=۱}^{|M_j|}precision(C_{jk}) \qquad (۱)$$

M_i تعداد کلاس‌های مرتبطی است که سیستم به درستی آنها را تشخیص داده است. C_{jk} کلاس مرتبط در مرتبه kام و q_j برابر با کلاس jام است

$$MAP=\frac{\sum_{q=۱}^{Q}AveP(q_j)}{Q} \qquad (۲)$$

در این معادله Q تعداد کل کلاس‌ها و q بیان‌کننده کلاس مربوط است.

تابع هزینه آنتروپی متقاطع (۳) برای محاسبه عملکرد یک طبقه‌بندی کننده است که خروجی آن به صورت یک مقدار احتمالاتی بین صفر و یک بیان می‌شود. زمانی مقدار خطا افزایش می‌یابد که احتمال پیش‌بینی از برچسب مقدار واقعی متفاوت باشد. در واقع تابع هزینه آنتروپی متقاطع تفاوت بین دو توزیع احتمال را اندازه‌گیری می‌کند. یعنی توزیع پیش‌بینی شده چقدر به توزیع درست نزدیک است. هدف ما این است که مدلی داشته باشیم که بالاترین احتمال را برای کلاس‌های هدف و کمترین احتمال را برای کلاس‌های دیگر تخمین بزند. بنابراین از تابع هزینه آنتروپی متقاطع استفاده می‌کنیم که برای کارهای طبقه‌بندی چندکلاسی مناسب است

$$J(W_j)=-\frac{۱}{m}\sum_{i=۱}^{m}\sum_{j=۱}^{k}y_j^{(i)}\log(\widehat{y_j^{(i)}}) \qquad (۳)$$

که $\hat{y}$ مقدار پیش‌بینی‌شده، y مقدار درست و m تعداد نمونه است. در این معادله $y_j^{(i)}$ زمانی برابر با ۱ است که کلاس هدف برای iامین نمونه j باشد و در غیر این صورت برابر با صفر است.

در این آزمایش‌ها از روش تلفیق جمع (۴) برای تلفیق نقشه ویژگی‌های سه‌شاخه در آخرین لایه استفاده می‌شود، زیرا با توجه به آزمایش‌های انجام‌شده این روش به دقت بالاتری نسبت به روش‌های دیگر مانند تلفیق الصاق دست یافته است.

در روش تلفیق جمع $y^{sum}=f^{sum}(X^a,X^b,X^c)$[۲] سه نقشه ویژگی را در همان موقعیت مکانی I_j و کانال d با هم جمع می‌کند

$$y_{i,j,d}^{sum}=x_{i,j,d}^{a}+x_{i,j,d}^{b}+x_{i,j,d}^{c} \qquad (۴)$$

که $۱\le d\le D$ ،$۱\le j\le w$ ،$۱\le i\le H$ و $x^a,x^b,x^c,y\in R^{H+W+D}$ هستند [۲۹].

طبق جدول ۱ از مقایسه نتایج حاصل از هر یک از شاخه‌ها بر روی مجموعه داده Willow به این نتیجه می‌رسیم در تصاویری که شیء مشخصی برای انجام کنش مورد نظر وجود داشته باشد شبکه مکانی به دقت بالاتری نسبت به کنش‌هایی مانند دویدن و راه‌رفتن که در تعامل با

1. Batch Size

2. Mean Average Precision

جدول ۲: میانگین متوسط صحت (MAP) بر روی مجموعه داده STANFORD۱۰.

PA (%)	شاخه زمانی	شاخه مکانی	شاخه ژست	تلفیق سه‌شاخه (جمع)
مسواک‌زدن	۵۲٫۳۲	۹۴٫۲۲	۴۰٫۷۲	۹۵٫۹۸
پریدن	۵۳٫۸۰	۹۵٫۷۳	۸۰٫۴۲	۹۶٫۹۱
نواختن گیتار	۶۲٫۵۶	۹۹٫۶۴	۵۹٫۶۶	۹۹٫۹۱
نواختن ویولن	۳۵٫۴۲	۹۹٫۴۴	۳۹٫۷۹	۹۹٫۴۱
دوچرخه‌سواری	۷۱٫۳۳	۹۹٫۹۷	۷۱٫۴۹	۹۹٫۹۵
اسب‌سواری	۵۳٫۳۵	۹۹٫۸۸	۷۱٫۳۸	۹۹٫۹۵
دویدن	۴۸٫۲۱	۹۰٫۶۱	۵۶٫۶۶	۹۳٫۴۶
پرتاب دیسک	۳۵٫۱۳	۹۰٫۸۵	۴۵٫۰۲	۹۴٫۰۵
راه‌رفتن سگ	۵۵٫۰۶	۹۷٫۴۶	۶۰٫۸۰	۹۸٫۵۲
دست‌تکان‌دادن	۲۶٫۴۳	۸۸٫۴۵	۴۰٫۲۸	۹۰٫۵۶
MAP (%)	۴۹٫۳۶	۹۵٫۶۲	۵۶٫۷۲	۹۶٫۸۷

جدول ۳: میانگین متوسط صحت (MAP) بر روی مجموعه داده PASCAL VOC۲۰۱۲.

PA (%)	شاخه زمانی	شاخه مکانی	شاخه ژست	تلفیق سه‌شاخه (جمع)
پریدن	۳۹٫۵۱	۹۰٫۳۹	۶۴٫۰۷	۹۰٫۶۵
تلفن‌زدن	۲۶٫۳۲	۸۷٫۳۱	۳۸٫۶۳	۸۷٫۵۶
نواختن موسیقی	۳۳٫۹۹	۹۴٫۵۳	۳۱٫۹	۹۲٫۳۴
خواندن	۳۱٫۶۲	۸۶٫۷۱	۳۵٫۰۹	۸۸٫۹
دوچرخه‌سواری	۴۵٫۴۹	۹۴٫۷۲	۵۱٫۷۲	۹۵٫۷۵
اسب‌سواری	۳۹٫۳۲	۹۷٫۷۲	۶۲٫۹۲	۹۶٫۷۸
دویدن	۵۷٫۵	۹۱٫۱۱	۶۶٫۸۵	۹۳٫۸۳
عکس‌گرفتن	۲۰٫۸۱	۸۹٫۷۶	۳۵٫۲۹	۹۰٫۱۵
کار با کامپیوتر	۲۷٫۴۹	۸۶٫۳۶	۳۴٫۶۴	۸۶٫۹۴
راه‌رفتن	۳۶٫۳۸	۸۲٫۷۷	۵۴٫۵۲	۸۷٫۲۲
MAP (%)	۳۵٫۸۴	۹۰٫۱۴	۴۷٫۵۴	۹۱٫۰۲

جدول ۴: مقایسه میانگین متوسط صحت (MAP) روش پیشنهادی با روش‌های قبلی بر روی مجموعه داده‌های WILLOW۷ ACTION، PASCAL VOC۲۰۱۲ و STANFORD۱۰.

MAP (%)	Willow۷action	Pascalvoc۲۰۱۲	Stanford۱۰
TICNN [۲۴]	۸۲٫۳	–	۹۶٫۸
Zero-shot [۲۳]	۸۱٫۸	–	۹۵٫۵
POF-SM [۲۲]	۷۶٫۱	–	–
DBN [۱۱]	۸۰٫۴۱	–	–
Zhang, *et al.* [۱۲]	۷۶٫۹۶	۸۳٫۲۳	۸۲٫۶۸
Gkioxari, *et al.* [۱۸]	–	۸۲٫۰۶	–
Im۲flow [۲۶]	۹۰٫۵	۶۶٫۰۱	۸۲٫۳
R*CNN [۱۹]	–	۹۰٫۰۲	–
روش پیشنهادی	۹۱٫۸	۹۱٫۰۲	۹۶٫۸۷

هیچ شیئی نیستند دست یافته است. بنابراین اشیای موجود در صحنه در بازشناسی کنش بسیار حایز اهمیت هستند. به عنوان مثال در کنش دوچرخه‌سواری شناسایی دوچرخه به تشخیص کنش دوچرخه‌سواری بسیار کمک خواهد کرد ولی در کنشی مانند راه‌رفتن و دویدن به دلیل عدم وجود شیء و شباهت بسیار زیاد این دو کنش به هم تشخیص کنش دشوارتر می‌شود. در شبکه زمانی کنش‌هایی که ذاتاً ایستا نیستند مانند سوارکاری، راه‌رفتن، دویدن و غیره به دقت بالاتری نسبت به کنش عکس‌گرفتن که ذاتاً ایستا هستند دست یافته‌اند زیرا اطلاعات حرکتی در تصویر در شناسایی بهتر کنش بسیار مؤثر است.

اهمیت شاخه ژست برای تشخیص کنش به این دلیل است که هر کنشی دارای ژست مخصوص به خود است. به عنوان مثال در تصویری که فرد در کنار یک دوچرخه ایستاده است در مقایسه با تصویری که شخص، سوار دوچرخه است نوع ژست بیان‌کننده این است که آن شخص در حال دوچرخه‌سواری است یا نه. شاخه مربوط به ژست برای همه کنش‌ها در هر سه مجموعه داده از شاخه زمانی بهتر عمل کرده است زیرا شار نوری که از روش Im۲Flow به دست می‌آید دقیق نیست و دارای خطا می‌باشد، بنابراین بعد از ورود تصاویر شار به شبکه زمانی این خطا در کل شبکه پخش می‌شود و دقت این شبکه را کاهش می‌دهد. با توجه به نتایج آزمایش‌ها در جدول ۱ روش پیشنهادی با استفاده از روش تلفیق جمع توانسته است به دقت ۹۱٫۸ درصد بر روی مجموعه داده Willow دست یابد.

جدول ۲ و ۳ نتایج متوسط صحت (AP) و میانگین متوسط صحت (MAP) را برای مجموعه داده Stanford۱۰ و Pascal voc۲۰۱۲ به ترتیب نشان می‌دهد. همان طور که مشاهده می‌شود با توجه به اهمیت ژست، اضافه‌کردن شاخه ژست سبب افزایش دقت در اکثر کنش‌ها شده است ولی در کنش‌هایی مانند اسب‌سواری و دوچرخه‌سواری به دلیل مشابه‌بودن نوع ژست، دقت کمتری در تلفیق جمع نسبت به شاخه مکانی به دست آمده و نهایتاً با روش تلفیق جمع به دقت ۹۶٫۸۷ درصد بر روی مجموعه داده Stanford۱۰ و دقت ۹۱٫۰۲ درصد بر روی مجموعه داده Pascal voc۲۰۱۲ رسیده است.

همان طور که در هر سه مجموعه داده مشاهده می‌شود نتایج حاصل از شاخه مکانی دارای بیشترین دقت نسبت به شاخه زمانی و شاخه ژست است، زیرا در شاخه مکانی علاوه بر شناسایی انسان و استخراج ویژگی‌ها، از اشیا و نشانه‌های موجود در صحنه نیز استفاده می‌کند که به شناسایی کنش کمک می‌کنند. با توجه به این که اطلاعات از تصویر خام رنگی توسط شبکه Resnet۵۰ پیش‌آموزش‌دیده به دست می‌آید و این شبکه توسط تصاویر Imagenet آموزش است و همچنین نتیجه جنس ویژگی‌های تصاویر Imagenet و نوع تصاویر آن مشابه مجموعه داده‌های به کار رفته در این آزمایش است، بنابراین ویژگی‌هایی که از این شبکه به دست می‌آید بازنمایش بهتری از کنش انسان را ارائه می‌دهد و در نتیجه شاخه مکانی به دقت بهتری دست یافته است. در صورتی که تصاویر زمانی و ژست به‌دست آمده را که به عنوان ورودی به دو شاخه زمانی و ژست وارد می‌کنیم دقیق نیستند و دارای خطای اولیه هستند بنابراین دقت به دست آمده از این دو شبکه کمتر از دقتی است که از شبکه مکانی به دست می‌آید.

جدول ۴ به مقایسه میانگین متوسط صحت بین روش پیشنهادی با

روش‌های قبلی می‌پردازد. طبق مقایسه نتایج به این نتیجه می‌رسیم که دقت به دست آمده بر روی مجموعه داده willow نسبت به روش‌های قبلی بر روی این مجموعه داده بیشتر است زیرا اولاً در روش پیشنهادی از اطلاعات ژست نیز برای بازشناسی استفاده گردیده که بسیار مفید واقع شده است. ثانیاً با توجه به این که مجموعه داده پاسکال بسیار چالش‌برانگیز است و مجموعه داده willow دارای تصاویری با کیفیت بالاتر نسبت به مجموعه داده پاسکال است و همچنین دارای پس‌زمینه نویزی و شلوغ در مقایسه با مجموعه داده پاسکال نیست، بنابراین دقت به دست آمده از مجموعه داده willow نسبت به مجموعه داده پاسکال بیشتر است. در مقایسه مجموعه داده willow با مجموعه داده استنفورد با توجه به این که مجموعه داده willow دارای تعداد و تنوع تصاویر کمتری است در نتیجه شبکه با تعداد و تنوع کمتری از تصاویر آموزش می‌بیند و بنابراین دقتی که روی مجموعه داده willow نسبت به مجموعه داده استنفورد به دست آمده است هم در روش‌های قبلی و هم در روش پیشنهادی کمتر است.

٦- نتیجه‌گیری

در این مقاله از معماری سه‌شاخه و شبکه Resnet۵۰ برای بازشناسی کنش‌های انسان از روی تصویر ایستا استفاده شده است. این روش از تلفیق سه نشانه مهم ژست انسان، شار نوری پیش‌بینی‌شده و اطلاعات مکانی در تصویر استفاده کرده است. با توجه به نتایج آزمایش‌ها مشاهده می‌شود که تخمین ژست از طریق شبکه Segnet در بازشناسی کنش بسیار مفید واقع شده است. همچنین روش تلفیق جمع با تعیین ضرایب وزنی توانسته است به دقت بهتری نسبت به روش‌های اخیر در زمینه بازشناسی کنش دست یابد. در نتیجه روش پیشنهادی توانسته است به دقت ۹۱٫۰۲ درصد بر روی مجموعه داده‌های پاسکال ۲۰۱۲، دقت ۹۶٫۸۷ درصد بر روی مجموعه داده Stanford۴۰ و دقت ۹۱٫۸ درصد بر روی مجموعه داده Willow۷ action دست یابد.

مراجع

[1] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343-33612014.

[2] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. IEEE Int. Conf. on Computer Vision, ICCV'17,* pp. 3391-3399, Venice, Italy, 17-21 Jul. 2017.

[3] K. Simonyan and V. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems, NIPS'14,* 9 pp., Montreal, Canada, 8 Dec. 2014

[4] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

[5] https://www.di.ens.fr/willow/research/stillactions

[6] http://vision.stanford.edu/Datasets/40actions.html

[7] L. Zhang, L. Changxi, P. Peipei, X. Xuezhi, and S. Jingkuan, "Towards optimal VLAD for human action recognition from still images," in *Proc. IEEE Int. Acoustics, Speech and Signal Processing Conf., ICASSP'16,* pp. 53-63, Shanghai, China, 20-25 Mar. 2016.

[8] Y. Tani and K. Hotta, "Robust human detection to pose and occlusion using bag-of-words," in *Proc. Inte. Conf. on Pattern Recognition, ICPR''14,* pp. 4376-4381, Stockholm, Sweden, 24-28 Aug. 2014.

[9] F. S. Khan, et al., "Coloring action recognition in still images," *International J. of Computer Vision*, vol. 105, no. 3, pp. 205-221, Dec. 2013.

[10] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. of the IEEE Int. Conf. on Computer Vision,* pp. 2470-2478, Santiago, Chile, 7-13 Dec. 2015.

[11] V. Yao, X. Jiang, and A. Khosla, "Human action recognition by learning bases of action attributes and parts," in *Proc. of ICCV,* pp. 1331-1338, Barcelona, Spain, 6-13 Nov. 2011.

[12] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *Proc. Advances in Neural Information Processing Systems, NIPS'11,* pp. 1503-1511, Granada, Spain, 12-17 Dec. 2011.

[13] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp.1691-1703, Sept. 2012.

[14] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601-614, Mar. 2012.

[15] L. Zhujin, X. Wang, R. Huang, and L. Lin, "An expressive deep model for human action parsing from a single image," in *Proc. IEEE Int. Conf. on Multimedia and Expo, ICME'14,* 6 pp., Chengdu, China, 14-18 Jul. 2014.

[16] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for semantic description of humans in still images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 87-101, Mar. 2016.

[17] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. IEEE Int. Conf. on Computer Vision, ICCV'17,* pp. 3391-3399, Venice, Italy, 22-29 Oct. 2017.

[18] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," *Computer Vision and Pattern Recognition, CVPR'10,* pp. 2030-2037, San Francisco, CA, USA, 15-17 Jun. 2010.

[19] Y. Zheng, Y. J. Zhang, X. Li, and B. D. Liu, "Action recognition in still images using a combination of human pose and context information," in *Proc. 19th IEEE Int. Conf. on Image Processing, ICIP'12,* pp. 785-788, Orlando, FL, USA, 30 Sept.-3 Oct. 2012.

[20] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'13,* pp. 652-659, Portland, ON, USA, 25-27 Jun. 2013.

[21] B. C. Ko, J. H. Hong, and J. Y. Nam, "Human action recognition in still images using action poselets and a two-layer classification model," *J. of Visual Languages & Computing*, vol. 28, no. 1, pp. 163-175, Jun. 2015.

[22] Y. Zhang, L. Cheng, J. Wu, and J. Cai, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. on Image Processing*, vol. 25, no. 11, pp. 5479-5490, Nov 2016.

[23] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r*cnn," in *Proc. of the IEEE Int. Conf. on Computer Vision,* pp. 1080-1088, Santiago, Chile, 7-13 Dec. Dec. 2015.

[24] M. Safaei and H. Foroosh, "Single image action recognition by predicting space-time saliency," arXiv:1705.04641v1, 12 May 2017.

[25] M. Safaei, P. Balouchian, and H. Foroosh, "TICNN: a hierarchical deep learning framework for still image action recognition using temporal image prediction," in *Proc 25th IEEE Int. Conf. on Image Processing, ICIP'18*, pp. 3463-3467, Athens, Greece, 7-10 Oct. 2018.

[26] M. Safaei and H. Foroosh, "A zero-shot architecture for action recognition in still images," in *Proc 25th IEEE Int. Conf. on Image Processing, ICIP'18,* pp. 460-464, Athens, Greece, 7-10 Oct. 2018.

[27] R. Gao, B. Xiong, and K. Grauman, "Im2flow: motion hallucination from static images for action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 5937-5947, Istanbul, Turkey, 18-22 Jun. 2018.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, *Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*, arXiv preprint arXiv:1511.00561, 2015.

[29] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1933-1941, Las Vegas, NV, USA, 27-30 Jun. 2016.

رقیه یوسفی در سال ۱۳۹۰ مدرک کارشناسی مهندسی کامپیوتر خود را در دانشگاه آزاد ساری دریافت نموده است. سپس در سال ۱۳۹۸ تحصیلات کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه آزاد اسلامی قزوین به پایان رسانده است.

کریم فائز در سال ۱۳۵۲ مدرک کارشناسی خود را از پلی تکنیک تهران کسب نمود و درسال های ۱۳۵۲ تا ۱۳۵۴ در همین دانشگاه مشغول به کار گردید. سپس در سال ۱۳۵۵ برای ادامه تحصیل وارد دانشگاه کالیفرنیا در لس‌آنجلس شد و به ترتیب در سال-های ۱۳۵۶ و ۱۳۵۹ مدارک دکترای خود را در معماری کامپیوتر و متدلوژی کامپیوتر از این دانشگاه کسب نمود و عازم ایران گردید. نام‌برده در اواخر سال ۱۳۵۹ در مرکز

تحقیقات مخابرات ایران مشغول بکار شد و از اوایل سال ۱۳۶۲ به دانشکده مهندسی برق دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) پیوست و هم اکنون با درجه استادی عضو هیأت علمی دانشگاه صنعتی امیرکبیر است. زمینه‌های تحقیقاتی نام‌برده بینایی ماشین، شبکه‌های عصبی، سیستم‌های هوشمند، و شبکه‌های کامپیوتری می‌باشد.