تخمین دو طرفه ژست بدن انسان در ویدیو

مینا قدیمی عتیق[۱]، احمد نیک آبادی[۲]

[۱] دانشجوی کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، تهران،
minaghadimi@aut.ac.ir

[۲] استادیار، دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، تهران،
nickabadi@aut.ac.ir

چکیده

تخمین ژست بدن انسان در تصویر که در آن محل قرارگیری اجزاء اصلی بدن انسان در یک تصویر دوبعدی مشخص می‌شود، استفاده‌های فراوانی در کاربردهای مختلف بینائی ماشین دارد. در تخمین ژست در ویدیو، علاوه بر اطلاعات ظاهری موجود در هر فریم می‌توان از اطلاعات زمانی بین فریم‌ها یا ویژگی‌های حرکتی نیز استفاده کرد. اطلاعات زمانی بین فریم‌ها را می‌توان با استفاده از حافظه‌های کوتاه مدت طولانی کانولوشنی مدل‌سازی کرد. با تحلیل حرکت انسان در یک دنباله از فریم‌ها می توان ژست احتمالی انسان در فریم‌های بعدی را پیش‌بینی کرد. در برخی موارد نظیر حالتی که یکی از اندام‌های بدن از یک حالت انسداد خارج و شروع به حرکت می‌کند به دلیل عدم وجود اطلاعات عضو مربوطه در فریم‌های قبلی، اطلاعات حرکتی برای تخمین ژست آن عضو در فریم‌های بعدی موجود نیست. برعکس، در این گونه موارد، اطلاعات فریم‌های بعدی می‌توانند در مورد محل آن عضو در فریم فعلی اطلاعاتی را ارائه نمایند. از این رو با تحلیل رو به عقب فریم‌ها می‌توان به مجموعه جدیدی از اطلاعات حرکتی دست یافت. در این مقاله با تخمین ژست با استفاده از دو مدل مجزای رو به جلو و رو به عقب، دو خروجی متمایز به ازای هر فریم تولید می‌شود. نقشه‌های اطمینان حاصل از این دو مدل با استفاده از یک شبکه‌ی کانولوشنی با یکدیگر ترکیب شده و خروجی نهایی تولید می‌شود. نتایج به دست آمده از اعمال روش پیشنهادی بر روی مجموعه داده‌های Penn Action و Sub-JHMDB نشان‌دهنده برتری این روش بر روش‌های پیشین و استخراج اطلاعات مورد نظر است.

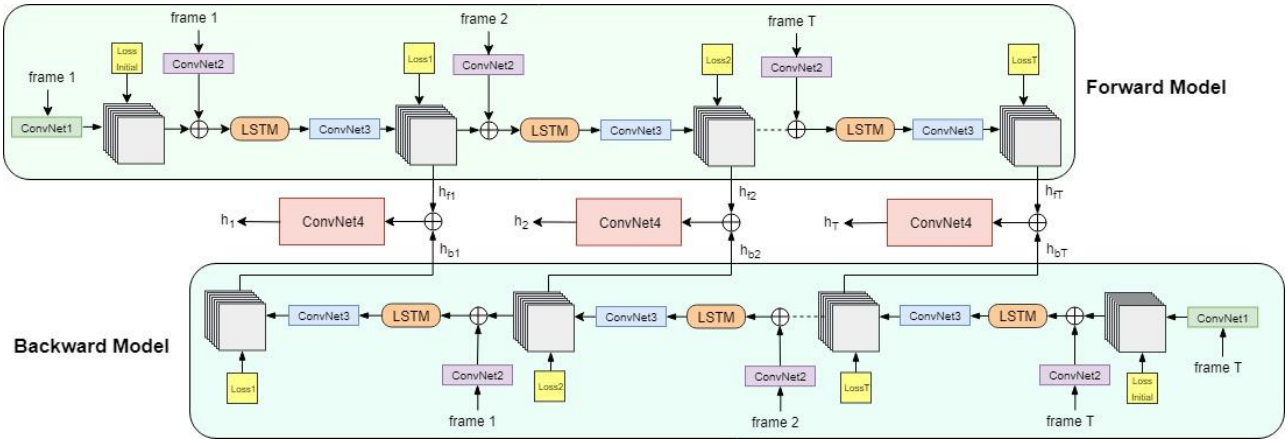
کلمات کلیدی

ژست بدن انسان، ویژگی‌های حرکتی، حافظه کوتاه مدت طولانی کانولوشنی، شبکه‌های عصبی کانولوشنی عمیق

۱- مقدمه

تخمین ژست بدن انسان یا تخمین مکان عضوهای بدن یک مسئله چالش‌برانگیز در بینایی ماشین است. این مسئله در بسیاری از کاربردهای دنیای واقعی همچون ویدیوهای نظارتی، تعامل انسان و کامپیوتر، سرگرمی‌های دیجیتال، زمینه‌های پزشکی، صحنه‌های ورزشی و ... نقش مهمی دارد[۱]. عواملی همانند تغییرات گسترده در ظاهر، زاویه‌های دید متفاوت، تغییرات در نحوه ایستادن افراد، پس‌زمینه نامناسب، خود انسدادی و ... باعث افزایش پیچیدگی این مسئله می‌شوند[۲]. هدف در مسئله تخمین ژست بدن انسان، دریافت ورودی به صورت تصویر و یا ویدیو و تخمین مکان عضوهای بدن در ورودی است. درصورتی‌که ورودی ویدیو باشد بایستی در هرکدام از فریم‌های ویدیو ژست بدن انسان تخمین زده شود.

در روش‌های قدیمی موجود در مسئله تخمین ژست بدن انسان در تصویر، بخش بزرگی از بار مسئله بر عهده ساختار درختی یا گرافی تعریف‌شده برای بدن انسان بود. هم‌چنین وظیفه‌ی تخمین عضوهای بدن بر عهده ویژگی‌های سطح پایینی هم‌چون HOG، لبه‌ها، هیستوگرام رنگ و ... بود[۳]. ظهور



شکل (۱): ساختار کلی روش ارائه شده

شبکه‌های عصبی عمیق در راه‌حل‌های ارائه‌شده برای تخمین ژست بدن انسان نیز تغییرات شگرفی ایجاد کرده است. این شبکه‌ها قابلیت تخمین مسقیم مکان اعضای بدن انسان بدون استخراج ویژگی‌های سطح پایین و یا تعریف واضح ساختار بدن انسان را دارند[۴] [۵].

در این راستا راه‌حل‌های ارائه‌شده برای تخمین ژست بدن انسان به سمت استفاده از شبکه‌های عصبی کانولوشنی حرکت کرده است. مقاله [۶] از ترکیبی از شبکه‌های عصبی و مدل‌های احتمالاتی گرافی استفاده کرده است. در روش ارائه‌شده، شبکه‌های عصبی کانولوشنی احتمال وجود هر عضو در هر ناحیه از تصویر را تولید می‌کنند. نتایج حاصل از شبکه عصبی کانولوشنی عمیق، در مدل احتمالاتی گرافی در نظر گرفته شده برای بدن انسان استفاده می‌شود. این روش همچنین به تعریف صریح ساختار گرافی برای بدن می‌پردازد. روش ارائه‌شده در [۷] دارای ساختار چند مرحله‌ای با قابلیت یادگیری ضمنی رابطه‌های مکانی اعضای بدن است. هر کدام از مراحل از شبکه‌های عصبی کانولوشنی تشکیل شده‌اند. شبکه در هر مرحله نقشه‌ی اطمینان هر یک از اعضای بدن را تولید می‌کند. در ساختار ارائه‌شده ویژگی‌های استخراج‌شده از تصویر و نقشه اطمینان تخمین زده‌شده در مرحله قبل به‌عنوان ورودی دریافت می‌شود و نقشه اطمینان بهبود یافته به عنوان خروجی تولید می‌شود. تکرار تخمین ژست بدن انسان در مراحل مختلف در ساختار این شبکه امکان یادگیری ضمنی ساختار گرافی بدن انسان را فراهم می‌کند و نیازی به استفاده مستقیم از ساختار گرافی برای بدن وجود ندارد.

با در نظر گرفتن ویدیو به صورت دنباله‌ای از تصاویر، برای تخمین ژست بدن انسان در ویدیو می‌توان از روش‌های موجود برای تخمین ژست بدن انسان در تصویر استفاده کرد. با استفاده از این روش‌ها، تخمین اولیه‌ای از ژست بدن انسان در هر فریم به دست می‌آید. روش ارائه‌شده در [۸] به مدل‌سازی رابطه زمانی بین فریم‌ها با استفاده از مدل‌های احتمالاتی گرافی می‌پردازد. در این روش ابتدا با استفاده از یک شبکه عصبی، ژست در هر فریم تخمین زده می‌شود. سپس اطاعات فریم‌های همسایه با استفاده از لایه‌های پیچش شار[۱] به فریم مورد نظر نگاشت می‌شود. همچنین یک لایه‌ی استنتاج زمانی-مکانی در انتهای شبکه‌ی پیشنهادی با بررسی ارتباط زمانی و مکانی بین اعضای بدن در فریم‌های ویدیو به تولید خروجی می‌پردازد. این روش دارای عملکرد بهتری نسبت به روش‌های پیشین ارایه شده است، اما دارای پیچیدگی بسیار بالایی است. سپس، مقاله [۹] به ارایه‌ی روشی برای تخمین ژست بدن انسان در ویدیو پرداخته است. این روش از ساختار پیشنهادی برای تخمین ژست بدن انسان در تصویر در مقاله [۷] به عنوان ساختارپایه استفاده می‌کند. همان‌طور که مطرح شد، در ساختار [۷] در هر مرحله با دریافت ویژگی‌های استخراج شده از ورودی و تخمین به دست آمده در مرحله قبل، تخمین جدیدی حاصل می‌شود. حال در این روش، با ارسال هم‌زمان T فریم به مدل، ژست‌های متناظر هر کدام از فریم‌ها تخمین زده می‌شود. همانند ساختار تکراری ارائه‌شده در [۷]، برای تخمین ژست بدن انسان در فریم‌های $t > 1$ به ویژگی‌های استخراج شده از آن فریم و تخمین ژست در فریم قبل نیاز داریم. تخمین‌های به دست آمده برای فریم قبل و ویژگی‌های استخراج شده از فریم جدید به حافظه کوتاه مدت طولانی کانولوشنی ارسال می‌شوند. با استفاده از این حافظه، تخمین به دست آمده برای فریم‌های قبل در تخمین فریم کنونی دخالت داده می‌شود.

روش‌های بررسی شده در زمینه تخمین ژست بدن انسان در ویدیو، قصد دارند تا با استخراج ویژگی‌های بین فریم‌ها به مدل‌سازی زمانی ژست در دنباله فریم‌های ورودی بپردازند. در این راستا روش ارایه شده در مقاله [۹] با استفاده از حافظه کوتاه مدت طولانی کانولوشنی رابطه بین فریم‌ها را در تخمین ژست دخالت می‌دهد. اما برای مثال در صورت رخداد انسداد، مدل ارائه‌شده قادر نخواهد بود تا ژست فریم بعد از انسداد را به خوبی تخمین بزند. در این راستا به دنبال طراحی ساختاری برای استفاده از اطلاعات فریم‌های بعد علاوه بر اطلاعات فریم‌های قبل هستیم. در ساختار ارائه شده در این مقاله، شبکه با دریافت دنباله‌ی فریم‌ها به دو صورت رو به جلو و رو به عقب می‌تواند از اطلاعات فریم‌های بعد نیز در تخمین حالت بدن در فریم فعلی استفاده کند. هم‌چنین با توجه به کم بودن تعداد فریم‌ها و عدم نیاز به پردازش آنلاین با توجه به کاربرد موردنظر، امکان پیمایش ویدیو تا انتها و سپس تخمین ژست در هر فریم وجود دارد.

در ادامه ابتدا در بخش ۲، مفاهیم پایه مورد نیاز مورد بررسی قرار می‌گیرند. سپس در بخش ۳، روش پیشنهادی برای تخمین ژست بدن انسان در ویدیو ارائه می‌گردد. در بخش ۴ نتایج آزمایش‌های انجام شده در جهت بررسی کارایی مدل ارائه‌شده نمایش داده می‌شود. در بخش پایانی نیز به جمع بندی پرداخته می‌شود.

۲- مفاهیم پایه

۲-۱- حافظه کوتاه مدت طولانی کانولوشنی

شبکه‌های عصبی بازگشتی[۲] دارای قابلیت دریافت دنباله‌ای از ورودی‌ها و تولید دنباله‌ای از خروجی‌ها هستند. اتصال‌های بازخورد موجود در ساختار این شبکه‌ها، قابلیت دریافت دنباله را ممکن کرده است. اما در هنگام مدل کردن

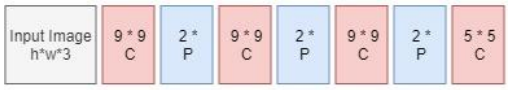
وابستگی‌های طولانی مدت مشکل‌های ناپدید شدن گرادیان و انفجار گرادیان مشاهده می‌شود[۱۰]. در راستای حل مشکلات موجود، معماری جدیدی از شبکه‌های بازگشتی به نام حافظه کوتاه مدت طولانی ارائه شده است که مشکل شبکه‌های بازگشتی را با اضافه کردن سلول حل می‌کند. هر سلول دارای سه گیت ورودی، خروجی و فراموشی است. گیت ورودی تعیین می‌کند که چه اطلاعاتی در حافظه ذخیره شوند، گیت خروجی کنترل می‌کند که اطلاعات تا چه زمانی در حافظه ذخیره شوند و گیت فراموشی نیز کنترل می‌کند که اطلاعات تا چه زمانی در حافظه ذخیره شده و سپس در چه زمانی از حافظه پاک شوند. در ادامه جزییات عملکرد حافظه کوتاه مدت طولانی بررسی می‌شود.

$$g_t = tanh(W_{xg} * X_t + W_{hg} * h_{t-1} + \epsilon_c)$$
$$f_t = sigmoid(W_{xf} * X_t + W_{hf} * h_{t-1} + \epsilon_f)$$
$$i_t = sigmoid(W_{xi} * X_t + W_{hi} * h_{t-1} + \epsilon_i)$$
$$o_t = sigmoid(W_{xo} * X_t + W_{ho} * h_{t-1} + \epsilon_o)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot tanh(C_t)$$

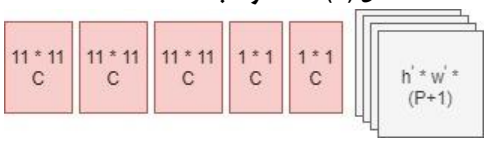
که h_t وضعیت پنهان حافظه کوتاه مدت طولانی در زمان t، o_t وضعیت گیت فراموشی، i_t مقدار گیت ورودی و f_t مقدار گیت فراموشی در زمان t است. C_t مقدار حافظه، g_t مقدار کاندید جدید برای حافظه در زمان t، عبارات ϵ عبارت بایاس و ضرایب W وزن‌های شبکه است.

در حافظه کوتاه مدت طولانی عملگر * نمایش‌دهنده ضرب ماتریس‌ها است، اما در حافظه کوتاه مدت طولانی کانولوشنی نمایش‌دهنده عملگر کانولوشنی است. تعریف عملگرها به صورت کانولوشنی باعث می‌شود تا گیت‌های تعریف شده به جای اطلاعات کلی به اطلاعات ناحیه‌ای توجه بیشتری داشته باشند. در نتیجه اطلاعات عضوها در ناحیه‌ی کوچکی مورد توجه قرار می‌گیرد.

شکل (۲): ساختار شبکه ConvNet1



شکل (۳): ساختار شبکه ConvNet2



شکل (٤): ساختار شبکه ConvNet3

۳- روش پیشنهادی

ساختار کلی روش پیشنهادی در شکل (۱) نشان داده شده است. در این روش فریم‌های ویدیو به مدل‌های رو به جلو و رو به عقب فرستاده شده و هر کدام از دو مدل، ژست بدن انسان را تخمین می‌زنند. خروجی به دست آمده از مدل‌های رو به جلو و رو به عقب به شبکه‌ی ترکیب ارسال می‌شود. خروجی شبکه‌ی ترکیب، ژست نهایی تخمین‌زده شده توسط مدل است. هدف این مقاله بهبود ژست تخمین‌زده شده برای فریم‌های ویدیوی ورودی با پیمایش رو به جلو و رو به عقب ورودی است. در ادامه بخش‌های مختلف ساختار فوق و راهکارهای پیشنهادی برای بهبود تخمین ژست شرح داده می‌شود.

۳-۱- مدل رو به جلو

در این مدل دنباله‌ای از T فریم ویدیو به عنوان ورودی مرحله‌های مختلف به شبکه ارسال می‌شوند. شبکه با دریافت ورودی با اندازه‌های $h \times w \times 3$ خروجی‌های $h_{fi}, 1 \leq i \leq T$ را تولید می‌کند. لایه‌هایی که با C مشخص‌شده‌اند، لایه‌های کانولوشنی و لایه‌هایی که با P مشخص شده‌اند، لایه‌های تجمعی[۳] هستند. خروجی حاصل از این شبکه دارای ابعاد $h' \times w' \times (P+1)$ است که P نشان‌دهنده تعداد اعضایی است که مکان آن‌ها تخمین زده شده است. در خروجی به ازای هر عضو یک نقشه اطمینان تولید می‌شود. همچنین یک نقشه اطمینان برای پس‌زمینه نیز در نظر گرفته می‌شود. در نقشه‌های اطمینان، مقدار هر مکان نشان‌دهنده‌ی میزان اطمینان رخداد عضو (یا پس‌زمینه) در آن مکان است. در نتیجه هر نقطه‌ای که مقدار بزرگتری داشته باشد، احتمال رخداد عضو در آن مکان بیشتر است.

در صورتی که $t = 1$ باشد، فریم ورودی که فریم اول در دنباله ورودی است به شبکه ConvNet1 ارسال می‌شود. این شبکه به منظور دریافت فریم اول در دنباله‌ی ورودی و تخمین یک ژست اولیه برای بدن انسان طراحی شده است. برای بهبود تخمین به دست آمده در مرحله اول، ابتدا فریم $t = 1$ به ConvNet2 ارسال می‌شود. این شبکه وظیفه استخراج ویژگی از فریم ورودی را بر عهده دارد. نقشه‌های اطمینان به دست آمده از ConvNet1 تخمین اولیه بوده و دارای اطمینان بالایی نیستند، از این رو نتایج به دست آمده از ConvNet1 و ویژگی‌های استخراج شده از فریم توسط ConvNet2 الحاق شده و به حافظه کوتاه مدت طولانی کانولوشنی ارسال می‌شوند. خروجی حاصل از این حافظه به ConvNet3 ارسال شده و نقشه‌های اطمینان متناظر با فریم اول تولید می‌شود. با دریافت فریم‌های دیگر، نقشه‌های اطمینان به دست آمده برای فریم قبل با ویژگی‌های استخراج شده از فریم جدید الحاق شده و به حافظه کوتاه مدت طولانی کانولوشنی ارسال می‌شوند. با ارسال خروجی به دست آمده از حافظه کوتاه مدت طولانی کانولوشنی به ConvNet3 نقشه اطمینان فریم جدید تخمین زده می‌شود. در ساختار چندمرحله‌ای، حافظه کوتاه مدت طولانی دارای قابلیت فراموشی اطلاعات قدیمی، جذب اطلاعات جدید و تولید خروجی است. در نتیجه در تخمین ژست هر فریم، ژست تخمین‌زده شده در فریم‌های قبل نیز به ایفای نقش می‌پردازند. در ادامه به بررسی دقیق‌تر ساختار شبکه‌های ConvNet1، ConvNet2 و ConvNet3 می‌پردازیم.

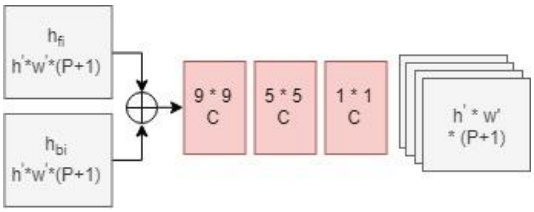
ساختار شبکه ConvNet1 در شکل (۲) نمایش داده شده است. ورودی شبکه، فریم اول دنباله با ابعاد $h \times w \times 3$ است که h و w ارتفاع و عرض فریم هستند. خروجی این شبکه نیز دارای ابعاد $h' \times w' \times (P+1)$ است که هم‌بعد با خروجی نهایی شبکه برای هر فریم است. این شبکه دارای وظیفه تولید تخمین اولیه برای فریم شروع کننده دنباله است. از آنجایی که شبکه ConvNet1 قصد دارد تا به تنهایی تخمینی از نقشه اطمینان ارائه دهد، دارای ساختار بزرگتری است. البته تخمین به دست آمده از این شبکه میزان اطمینان بالایی نداشته و به عنوان تخمین اولیه مورد استفاده قرار می‌گیرد.

ساختار شبکه ConvNet2 در شکل (۳) نمایش داده شده است. ورودی شبکه فریم‌های ویدیو با ابعاد $h \times w \times 3$ است که h و w ارتفاع و عرض فریم هستند. این شبکه به ازای هر کدام از فریم‌های ورودی تکرار شده و بردار ویژگی استخراج می‌کند. در شکل (۴) نیز ساختار شبکه ConvNet3

نمایش داده شده است. این شبکه دارای وظیفه تولید تخمین نهایی برای هر فریم است و دارای خروجی با ابعاد $h' \times w' \times (P+1)$ است. پارامترهای شبکه‌های ConvNet2، ConvNet3 و حافظه‌ی کوتاه مدت طولانی کانولوشنی در مرحله‌های مختلف مدل رو به جلو مشترک در نظر گرفته شده است.

۳-۲- مدل رو به عقب

ساختار مدل رو عقب همانند مدل رو به جلو است. این مدل فریم‌های ورودی را با ترتیب برعکس دریافت می‌کند. در این مدل نیز دنباله‌ای از T فریم ویدیو به عنوان ورودی مرحله‌های مختلف به شبکه ارسال می‌شود. شبکه با دریافت ورودی با اندازه‌های $h \times w \times 3$ خروجی‌های $h_{bi}, 1 \le i \le T$ را تولید می‌کند. با توجه به دریافت ورودی‌ها با ترتیب برعکس، مدل ابتدا به تخمین ژست در فریم $t = T$ می‌پردازد. پس از تخمین اولیه ژست در فریم اول دنباله توسط شبکه ConvNet1 و استخراج ویژگی از فریم $t = T$ توسط شبکه ConvNet2، تخمین ژست بدن انسان برای این فریم توسط شبکه ConvNet3 به دست می‌آید. سپس شبکه به تخمین ژست برای فریم $T-1$ می‌پردازد و با پیش‌روی برعکس در دنباله فریم‌های ورودی، برای هر فریم از دنباله تخمین ژست به دست می‌آید. ساختار شبکه‌های ConvNet1، ConvNet2 و ConvNet3 همانند مدل رو به جلو است. پارامترهای این شبکه‌ها نیز در مرحله‌های مختلف مدل رو به عقب مشترک بوده و با پارامترهای شبکه‌های موجود در مدل رو به جلو متفاوت است.

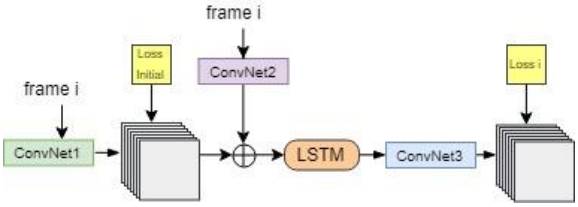


شکل (۵): ساختار شبکه ترکیب ConvNet4

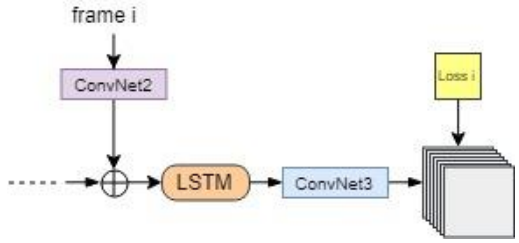
۳-۳- شبکه ترکیب

پس از ارسال فریم‌های ویدیو به شبکه‌های رو به جلو و رو به عقب، دو مجموعه نقشه اطمینان توسط هر کدام از این شبکه‌ها تولید می‌شوند. حال قصد داریم تا برای هر فریم یک نقشه اطمینان نهایی تولید کنیم. در نتیجه به دنبال شبکه‌ای هستیم که دو مجموعه نقشه اطمینان دریافت کرده و نقشه اطمینان نهایی را تولید کند. در این راستا مدل ترکیبی با نام ConvNet4 طراحی شده است. ساختار این شبکه در شکل (۵) نمایش داده شده است. ورودی این شبکه به صورت دو نقشه‌ی اطمینان h_{fi} و h_{bi} با اندازه‌های $h' \times w' \times (P+1)$ است که توسط مدل‌های رو به جلو و رو به عقب به دست آمده است. این شبکه دارای وظیفه‌ی تولید نقشه اطمینان نهایی با توجه به اطلاعات موجود در نقشه اطمینان‌های h_{fi} و h_{bi} است. همان‌طور که در بخش مدل رو به جلو توضیح دادیم، این مدل دنباله فریم‌های ورودی را با ترتیب موجود در ویدیو دریافت می‌کند.

با توجه به ساختار تعریف شده برای این شبکه، در تخمین ژست بدن انسان در هر کدام از فریم‌ها اطلاعات فریم‌های قبل نقش ایفا می‌کنند. با ارسال دنباله‌ی فریم‌ها با ترتیب برعکس به مدل رو به عقب، اطلاعات فریم‌های بعد در تخمین ژست بدن انسان دارای نقش خواهند بود. در نتیجه در هنگام استفاده همزمان از مدل‌های رو به جلو و رو به عقب و ارسال تخمین‌های به دست آمده از این دو مدل به شبکه ترکیب، اطلاعات حاصل از پیمایش فریم‌های قبل و فریم‌های بعد دارای نقش خواهند بود. در طراحی شبکه ترکیب از ساختار ارائه شده در [۷] کمک گرفته شده است.



شکل (٦): ساختار شبکه deploy1



شکل (۷): ساختار شبکه deploy2

۳-٤- آزمایش شبکه

در زمان آزمایش به ازای هر فریم، خروجی مدل رو به جلو و مدل رو به عقب تولید می‌شود. ابتدا برای فریم‌های $t = 1$ و $t = T$ که T در این بخش برابر با تعداد فریم‌های ویدیو است، ژست اولیه تخمین زده می‌شود. برای تخمین نقشه‌های اطمینان فریم‌های شروع کننده دنباله از شبکه deploy1 که در شکل (۶) نمایش داده شده است، استفاده می‌شود. این شبکه از بخش ابتدایی شبکه‌های رو به جلو و رو به عقب تشکیل شده است و نقشه‌های اطمینان فریم ورودی را به عنوان خروجی تولید می‌کند. پس از به دست آوردن خروجی برای فریم‌های اول و آخر ویدیو، به سراغ فریم‌های بعدی یعنی $t = 2$ و $t = T - 1$ می‌رویم. برای تخمین ژست توسط مدل‌های رو به جلو و رو به عقب در این فریم‌ها از شبکه deploy2 که در شکل (۷) نمایش داده شده است، استفاده می‌کنیم. این شبکه ژست تخمین زده شده برای فریم‌های $t = 1$ و $t = T$ را دریافت می‌کند. پس از پیمایش تمامی فریم‌های ویدیو و تخمین ژست توسط مدل‌های رو به جلو و رو به عقب، زمان تخمین نهایی ژست فرا می‌رسد. در این مرحله، دو تخمین به دست آمده از مدل‌های رو به جلو و رو به عقب را در دست داریم. با ارسال تخمین‌های به دست آمده به شبکه ConvNet4، این شبکه به تولید نقشه اطمینان نهایی می‌پردازد.

٤- آزمایش‌ها

٤-۱- معرفی مجموعه داده‌ها

مجموعه داده Penn Action [۱۱] یک مجموعه داده ویدیویی شامل ۲۳۲۶ ویدیو است. در این مجموعه داده ۱۲۵۸ ویدیو برای آموزش و ۱۰۶۸ ویدیو برای آزمایش در نظر گرفته شده است. به طور متوسط هر ویدیو دارای ۷۰ فریم است، اما تعداد فریم‌ها در هر ویدیو در مجموعه داده دارای گستردگی زیادی است. این مجموعه داده دارای مختصات مکانی ۱۳ عضو

شامل سر، شانه‌ها، آرنج‌ها، مچ‌ها، مفاصل ران، زانوها و مچ‌های پا در هر فریم از هر ویدیو است. هم‌چنین یک برچسب اضافی برای نمایش یا عدم نمایش هر عضو در هر فریم وجود دارد. بر اساس کارهای قبلی، نتیجه تنها بر روی عضوهایی که دیده می‌شوند انجام می‌شود.

مجموعه داده Sub-JHMDB [۱۲] مجموعه داده JHMDB یک مجموعه داده ویدیویی برای تخمین ژست بدن انسان در ویدیو است. برای مقایسه نتایج به دست آمده بر روی این مجموعه داده، تنها از بخشی از مجموعه داده که Sub-JHMDB نام دارد، استفاده می‌شود. در این زیر مجموعه، در تمامی فریم‌ها بدن فرد به طور کامل وجود دارد و هیچ عضوی پنهان نیست. این مجموعه داده دارای سه بخش مختلف است. روش‌های ارائه شده بایستی جداگانه بر روی هر کدام از سه مجموعه داده آموزش داده شده و آزمایش شوند. نتیجه نهایی از میانگین‌گیری نتایج سه زیر مجموعه به دست می‌آید.

٤-٢- معیار ارزیابی

برای ارزیابی تخمین‌های حاصل در هر فریم از ویدیو، از معیار ارزیابی PCK[۴] که معیاری استاندارد در مقایسه پایگاه داده‌های موجود است، استفاده می‌شود. بر اساس این معیار، تخمین به دست آمده در صورتی درست تلقی می‌شود که فاصله‌ی مکان عضو تخمین زده شده از مکان واقعی، کمتر از $\alpha.\max(h,w)$ باشد. h و w در این معیار ارتفاع و عرض مستطیل پوشش‌دهنده بدن هستند. برای سازگاری نتایج به دست آمده در آزمایش‌ها با آزمایش‌های سایر مقاله‌ها از $\alpha = 0.2$ استفاده می‌شود.

٤-٣- نتایج آزمایش‌ها

برای آموزش شبکه طراحی شده از پارامترهای شبکه ارائه شده در[۸] استفاده می‌شود. همانطور که در بخش روش پیشنهادی دیدیم، شبکه با دریافت $t = T$ فریم ورودی آموزش می‌بیند. برای مشخص کردن مقدار T آزمایش‌هایی در [۹] انجام شده است. شبکه با ارسال تعداد متفاوتی فریم به ورودی، آموزش داده شده و نتیجه‌ی حاصل از آزمایش بررسی شده است. بر اساس نتایج به دست آمده $T = 5$ بهترین گزینه برای آموزش شبکه است. از این رو ساختار مدل رو به جلو و رو به عقب هم با در نظر داشتن $T = 5$ طراحی می‌شود.

هدف در مسئله تخمین ژست بدن انسان، تخمین مکان x و y هر عضو مورد نظر در فریم ورودی است، اما شبکه‌ی طراحی شده دارای خروجی در قالب نقشه اطمینان است. حال بایستی از نقشه‌های اطمینان به دست آمده، مکان x و y هر عضو را استخراج کنیم. در این راستا نقطه دارای بیشترین اطمینان در هر نقشه اطمینان، به عنوان خروجی استخراج می‌شود. پس از استخراج مکان‌های تخمین زده شده برای هر عضو، به محاسبه دقت با معیار PCK می‌پردازیم. نتایج حاصل بر روی مجموعه داده Penn Action در جدول (۱) قابل مشاهده است. همچنین در جدول (۲) دقت به دست آمده در تخمین هر کدام از عضوهای موردنظر نیز گزارش شده است.

جدول (۱): نتایج به دست آمده بر روی مجموعه داده Penn Action

دقت	روش
۴۵/۳	[۱۳]
۴۸/۰	[۱۴]
۸۱/۱	[۱۵]
۹۱/۸	[۱۶]
۹۶/۵	[۸]
۹۷/۱	[۷]
۹۷/۷	[۹]
۹۷/۹٤	روش ارائه شده

جدول (۲): مقایسه نتایج به دست آمده در تخمین اعضای مختلف در مجموعه داده Penn Action با پارامتر $\alpha = 0.2$

روش	H	S	E	W	Hip	K	A
[۱۳]	۶۲/۸	۵۲/۰	۳۲/۳	۲۳/۳	۵۳/۳	۵۰/۲	۴۳/۰
[۱۴]	۶۴/۲	۵۵/۴	۳۳/۸	۲۴/۴	۵۶/۴	۵۴/۱	۴۸/۰
[۱۵]	۸۹/۱	۸۶/۴	۷۳/۹	۷۳/۰	۸۵/۳	۷۹/۹	۸۰/۳
[۱۶]	۹۵/۶	۹۳/۸	۹۰/۴	۹۰/۷	۹۱/۸	۹۰/۸	۹۱/۵
[۸]	۹۸/۰	۹۷/۳	۹۵/۱	۹۴/۷	۹۷/۱	۹۷/۱	۹۶/۹
[۷]	۹۸/۶	۹۷/۹	۹۵/۹	۹۵/۸	۹۸/۱	۹۷/۳	۹۶/۶
[۹]	**۹۸/۹**	**۹۸/٦**	۹۶/۶	۹۶/۶	۹۸/۲	۹۸/۲	۹۷/۵
ارائه شده	**۹۸/۹**	۹۸/۵۵	**۹۷/۰**	**۹۷/۰**	**۹۸/٤**	**۹۸/۵**	**۹۷/۷**

همان‌طور که در نتایج قابل مشاهده است، مدل ارائه شده به خوبی توانسته است دقت تخمین ژست انسان را در مجموعه داده Penn Action بهبود دهد و در زمینه عضوهای دارای حرکت زیاد و شکل‌های مشاهده متفاوت همانند دست، آرنج و ... به خوبی عمل کرده است. دلیل این امر در استفاده از اطلاعات رو به عقب فریم‌های ویدیو و شیوه ترکیب آن است. مدل‌های رو به عقب و رو به جلو در بیشتر موارد در شرایط متفاوتی دچار خطا می‌شوند. در نتیجه ترکیب این دو مدل می‌تواند به بهبود نتایج کمک کند. برای مثال پس از وقوع انسداد و رفع آن، مدل رو به جلو نمی‌تواند تخمین درستی ارائه دهد. اما مدل رو به عقب با دریافت فریم‌ها با ترتیب برعکس از مکان عضو پس از رخداد انسداد در مدل رو به جلو آگاه است.

نتایج به دست آمده بر روی مجموعه داده Sub-JHMDB نیز در جدول جدول (۳) قابل مشاهده است. همان‌طور که در بخش مجموعه داده‌ها توضیح داده شد، مدل به صورت جداگانه بر روی هر کدام از سه زیر مجموعه آموزش داده شده است. سپس با میانگین‌گیری از دقت‌های به دست آمده از این مجموعه‌ها دقت نهایی به دست آمده است.

جدول (۳): دقت به دست آمده بر روی مجموعه داده Sub-JHMDB

دقت	روش
۵۲/۵	[۱۳]
۵۵/۷	[۱۴]
۷۳/۸	[۱۵]
۹۲/۱	[۸]
۹۱/۹	[۷]
۹۲/۲	[۹] بدون LSTM
۹۳/۶	[۹]
۹٤/۰	روش ارائه شده

۵- نتیجه گیری

در این مقاله، مسئله تخمین ژست بدن انسان در ورودی ویدیو مورد بررسی قرار گرفت. با توجه به عدم استفاده اکثر روش‌ها از اطلاعات فریم‌های بعد و تأثیر مثبت این اطلاعات در تخمین ژست در هر فریم، یک مدل دو طرفه برای تخمین ژست بدن پیشنهاد شد. مدل طراحی شده ابتدا T فریم از فریم‌های ورودی را به صورت رو به جلو و رو به عقب به عنوان ورودی دریافت می‌کند. با تولید دو تخمین برای هر فریم توسط دو مدل موجود، نیاز به مدل ترکیب برای تولید تخمین نهایی وجود دارد. در نتیجه یک شبکه ترکیب نیز برای ترکیب نقشه‌های اطمینان به دست آمده از دو مدل طراحی شد. مدل طراحی شده به خوبی توانست دقت تخمین ژست در مجموعه داده‌ی Penn Action و sub-JHMDB را بهبود دهد.

مراجع

[1] Li u, Z., Zhu, J., Bu, J., Chen, Ch., *A Survey of Human Pose Estimation: The Body Parts Parsing based methods*, Journal of Visual Communication and Image Representation, Vol. 32, pp. 10-19, 2015.

[2] Zhang, D., Shah, M., Human Pose Estimation in Videos, The IEEE International Conference on Computer Vision (ICCV), pp. 2012-2020, Santiago Chile, 2015.

[3] Tompson, J.J., Jain, A., LeCunn, Y., Bregler, Ch., *Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation*, Advances in neural information processing systems, pp. 1799 – 1807, 2014

[4] Belagiannis, V., Zisserman, A., *Recurrent Human Pose Estimation*, IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, 2017.

[5] Yang, Y., Ramanan, D., *Articulated Pose Estimation with Flexible Mixture-of-Parts*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, pp. 2878-2890, IEEE, 2012.

[6] Chen, X., Yullie, Alan L., *Articulated Human Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations*, Advances in Neural Information Processing Systems, pp. 1736-1744, 2014.

[7] Wei, Sh., Ramakrishna, V., Kanade, T., Sheikh, Y., *Convolutional Pose Machines*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724 – 4732. Las Vegas, 2016.

[8] Song, J., Wang, L., Van Gool, L., Hilligies, O., *Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2017.

[9] Luo, Y., Ren, J., Wang, Zh., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L., *LSTM Pose Machines*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[10] Greff, K., Srivastava, Rupesh J., Koutnik, J., Steunebrink, Bas R., Schmidhuber, J., *LSTM: A search Space Odyssey*, IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, pp. 2222-2232, 2016.

[11] Zhang, W., Zhu, M., Derpanis, K. G., *From actemes to action: A strongly-supervised representation for detailed accuracy understanding*, The International Conference on Computer Vision (ICCV), 2013.

[12] Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J., *Towards understanding action recognition*, The International Conference on Computer Vision (ICCV), 2013.

[13] Park, D., Ramanan, D., *N-best Maximal Decoders for part models*, The International Conference on Computer Vision (ICCV), 2011.

[14] Nie, B. X., Xiong, C., Zhu, S. C., *Joint action recognition and pose estimation from video*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] Iqbal, U., Garbade, M., Gall, J., *Pose for Action-Action for Pose*, arxiv, 2016.

پانویس‌ها

[1] Flow warping

[2] Recurrent Neural Networks

[3] Pooling Layer

[4] Percentage Correct Keypoints