دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی‌ارشد

گرایش هوش مصنوعی

تخمین ژست بدن انسان در ویدیو با مدل‌های احتمالاتی گرافی

نگارش

مینا قدیمی عتیق

استاد راهنما

دکتر احمد نیک آبادی

بهمن ماه سال ۱۳۹۷

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

اینجانب مینا قدیمی عتیق متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

مینا قدیمی عتیق

امضا

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 می‌توان 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 مربوط به عضو مربوط در فریم‌های قبلی، اطلاعات حرکتی برای تخمین حالت آن عضو در فریم‌های بعدی موجود نیست. برعکس، در این گونه موارد اطلاعات فریم‌های بعدی می‌تواند 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.

واژه‌های کلیدی:

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

فهرست مطالب

فهرست اشکال

فهرست جداول

۱

فصل اول

مقدمه

در مسائل بینایی ماشین، هدف کمک به ماشین‌ها برای دستیابی به درک بالاتری از محیط پیرامون است. یکی از مسائل چالش برانگیز در حوزه بینایی ماشین، تخمین ژست بدن انسان است. در مسئله تخمین ژست بدن انسان، هدف استخراج موقعیت مکانی نقطه‌های کلیدی بدن انسان در ورودی است. ورودی می‌تواند تصویر یا ویدیو، دو بعدی یا سه بعدی باشد. در ورودی سه بعدی، اطلاعات عمق به عنوان بعد سوم موجود است. خروجی تولید شده نیز می‌تواند با توجه به تعریف مسئله دارای دو یا سه بعد باشد و شامل مختصات نقطه‌های کلیدی بدن انسان است. نقطه‌های کلیدی می‌توانند در نسخه‌های مختلف مسئله تخمین ژست بدن انسان، دارای تعداد و تعریف‌های مختلف باشند. یکی دیگر از عوامل مهم در تعریف مسئله تخمین ژست بدن انسان، تعداد افراد موجود در ورودی است. در برخی مسئله‌ها، هدف تخمین ژست برای ورودی‌های دارای تک فرد و در برخی، هدف تخمین ژست در صورت حضور چندین فرد در ورودی است.

هدف ما در این پژوهش تخمین ژست بدن انسان در ویدیوی دو بعدی و تک فرد است و مختصات تخمین زده شده برای نقاط سر، شانه‌ی راست و چپ، بازوی راست و چپ، مچ دست راست و چپ، مفصل ران راست و چپ، زانوی راست و چپ و مچ پای راست و چپ به عنوان خروجی تولید می‌شوند. برای درک بهتر مسئله تخمین ژست بدن انسان، یک نمونه خروجی در شکل ۱-۱ نمایش داده شده است. تخمین مختصات ۱۳ نقطه کلیدی بدن با نقاط مشکلی نمایش داده شده است. همچنین روابط سینماتیکی بین نقاط مورد نظر توسط یال نشان داده شده است.

شکل ۱-۱ نمونه‌ای از خروجی تولید شده در مسئله تخمین ژست بدن انسان

مسئله تخمین ژست بدن انسان یک مسئله‌ی پایه‌ای بوده و در بسیاری از کاربردهای دنیای واقعی همچون ویدیوهای نظارتی، تعامل انسان و کامپیوتر، سرگرمی‌های دیجیتال، زمینه‌های پزشکی، صحنه‌های ورزشی و ... نقش مهمی ایفا می‌کند[1]. مثال‌هایی از کاربردهای تخمین ژست بدن انسان در شکل ۱-۲ نمایش داده شده است. به عنوان نمونه ژست بدن انسان به عنوان اطلاعات پایه‌ای در بازشناسی رفتار مورد استفاده قرار می‌گیرد. همچنین در مسئله تحلیل پوشش، ژست تخمین زده شده به قطعه‌بندی پوشش به بخش‌های تشکیل دهنده کمک می‌کند. در مسائلی همانند بازی سازی و ردیابی افراد نیز ژست بدن افراد نقش مهمی را ایفا می‌کند.

شکل ۱-۲ نمونه‌هایی از کاربردهای تخمین ژست بدن انسان[1]

مسئله تخمین ژست بدن انسان دارای چالش‌های فراوانی است. عواملی همانند تغییرات گسترده در ظاهر، زاویه‌های دید متفاوت، تغییرات در نحوه ایستادن افراد، پس‌زمینه چالش برانگیز، خود انسدادی و ... باعث افزایش پیچیدگی این مسئله می‌شوند[2]. نمونه‌هایی از ژست‌های چالش برانگیز در مسئله تخمین ژست بدن انسان در شکل ۱-۳ نمایش داده شده است. برای مثال در بخش‌های الف و ب با توجه به نحوه قرارگیری فرد، تخمین ژست دشوار است. در بخش ب، ج و د نیز با توجه به قرار گیری فرد، خود انسدادی وجود دارد. همچنین در بخش ج با توجه به اینکه فرد در حال حرکت است، تاری وجود دارد.

الف

ب

ج

د

شکل ۱-۳ ژست‌های چالش برانگیز در مسئله تخمین ژست بدن انسان[1]

در روش‌های ابتدایی تخمین ژست بدن انسان، بخش بزرگی از بار مسئله بر عهده تعریف صریح ساختار سینماتیکی بدن انسان و یا استخراج دستی ویژگی‌های سطح پایینی هم‌چون HOG، لبه‌ها، هیستوگرام رنگ و ... بود[3]. با ظهور شبکه‌های عصبی عمیق، تغییر شگرفی در راه‌حل‌های ارائه شده برای تخمین ژست بدن انسان ایجاد شد. این شبکه‌ها قابلیت تخمین ژست بدن انسان بدون نیاز به تعریف صریح ساختار بدن و یا استخراج دستی ویژگی‌های سطح پایین را دارند[5][4].

مسئله تخمین ژست بدن انسان در ویدیو را می‌توان به دو بخش تخمین ژست بدن در تصویر و استفاده از اطلاعات زمانی بین فریم‌ها تقسیم کرد.

در اکثر روش‌های ارائه شده برای تخمین ژست بدن انسان در ویدیو، علاوه بر اطلاعات موجود در هر فریم از رابطه‌ی زمانی بین فریم و فریم‌های قبل برای تخمین ژست بدن انسان در آن فریم استفاده می‌شود. استفاده از این رابطه‌ی زمانی، سازگاری زمانی بین تخمین‌ها را ایجاب می‌کند. اطلاعات زمانی به دست آمده از رابطه‌ی هر فریم و فریم‌های پیشین، اطلاعات مفیدی را فراهم می‌کند؛ اما در تمامی موارد کاربردی نیست. برای مثال در صورت رخداد انسداد در فریم‌های متوالی پیشین، رابطه‌ی زمانی با فریم‌های قبل اطلاعاتی از مکان رخداد عضو در فریم کنونی ارائه نمی‌دهند.

در این پژوهش قصد داریم تا با در نظر گرفتن ارتباط هر فریم و فریم‌های بعدی، اطلاعات زمانی جدیدی کسب کنیم. این اطلاعات در مواردی همچون انسداد کمک شایانی به بهبود تخمین به دست آمده می‌کنند. مدل‌های متفاوتی برای استخراج رابطه‌های زمانی در جهات مختلف طراحی می‌کنیم. در هنگام استفاده از اطلاعات فریم‌های پیشین برای تخمین، از پیمایش رو به جلوی فریم‌ها و در هنگام استفاده از فریم‌های پسین، از پیمایش رو به عقب استفاده می‌شود. در نتیجه دو مجموعه تخمین برای ورودی تولید می‌شود. در انتها با ارسال تخمین‌های تولید شده به مدل ترکیب، تخمین نهایی به دست می‌آید. شبکه‌ی طراحی شده با استفاده از مجموعه داده‌های Penn Action[6]و Sub-JHMDB[7] مورد آزمایش قرار گرفته و برتری خود را نسبت به روش‌های موجود نشان داده است.

در ادامه، در فصل دوم مروری بر کارهای پیشین خواهیم داشت. در فصل سوم جزئیات روش پیشنهادی شرح داده می‌شود. در فصل چهارم نیز به بررسی نتایج و ارزیابی مدل می‌پردازیم. در فصل آخر جمع‌بندی و کارهای آتی بیان می‌شود.

۲

فصل دوم

کارهای پیشین

در این بخش مروری بر کارهای پیشین مرتبط با تخمین ژست بدن انسان خواهیم داشت. ابتدا در بخش ۲-۱- روش‌های ارائه شده برای تخمین ژست بدن انسان در تصاویر ورودی مورد بررسی قرار می‌گیرد. سپس در بخش ۲-۲- روش‌های ارائه شده برای تخمین ژست بدن انسان در ویدیو شرح داده شده است. در انتها، در بخش ۲-۳- مفاهیم پایه موردنیاز برای پژوهش معرفی می‌شوند.

۲-۱- تخمین ژست بدن انسان در تصویر

تا به امروز روش‌های متعددی برای تخمین ژست بدن انسان در تصویر ارائه شده است. این روش‌ها را می‌توان در چهار دسته‌ی کلی قرار داد[9] ,[8]:

- روش‌های اولیه
- مدل پایه ساختارهای تصویری[1] و پوزلت[2]
- ساختارهای درختی
- ساختارهای گرافی
- شبکه‌های عصبی

۲-۱-۱- روش‌های اولیه

در روش‌های اولیه ارائه شده برای حل مسئله تخمین ژست بدن انسان از روش‌های کلاسیک مبتنی بر مدل استفاده شده است[14]–[10]. یکی از اولین روش‌های انجام شده در این زمینه «طرح بدن[3]»[12] است که با مشخص کردن بخش‌هایی که به هم متصل هستند به ارائه‌ی مدلی برای طرح بدن می‌پردازد. در [15] از قطعه‌بندی برای تخمین ژست استفاده شده است. در [16] ژست بدن انسان به طور مستقیم با استفاده از رنگ پوست تخمین زده می‌شود.

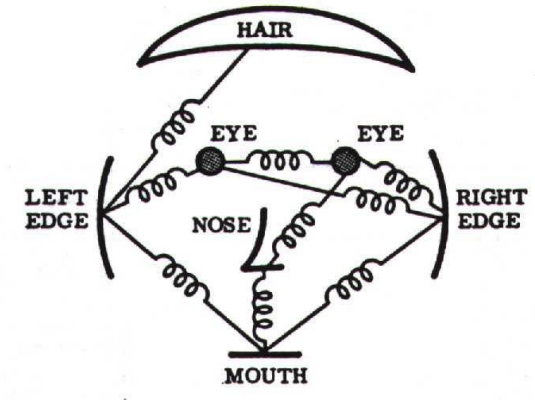
[1] Pictorial Structures

[2] poselet

[3] Body Design

۲-۱-۲- ساختارهای تصویری و پوزلت

در [17] مفهومی به عنوان ساختارهای تصویری برای اشیا ارائه می‌شود. در این کار اشیا دگردیس پذیر[4] با شکل‌های متفاوت، می‌توانند به صورت مجموعه‌ای از اجزای ایستا مدل شوند. شکل ۲-۱ این مفهوم را بر روی مدل صورت نمایش می‌دهد. در این مدل، صورت فرد به عنوان مجموعه‌ای از چشم‌ها، بینی، دهان و ... متصل به هم برای تشکیل یک پیکربندی عملی در نظر گرفته شده است.



شکل ۲-۱ مدل صورت ارائه شده در روش [17]

در مدل ارائه شده اعضای دگردیس پذیر به صورت مجموعه‌ای از اعضا در نظر گرفته می‌شود. این عضوها دارای قابلیت تغییر مکان و حالات قرارگیری مختلف هستند. در نتیجه مدل ارائه شده دارای قابلیت انعطاف‌پذیری بالایی است.

مدل پیشنهادی به صورت گراف $G = (V, E)$ تعریف می‌شود. در این گراف $V = \{v_1, v_2, \dots, v_n\}$ نشان‌دهنده‌ی مجموعه راس‌ها است که هر رأس به یک عضو تعلق دارد. هم‌چنین E نیز مجموعه یال‌های گراف G متصل کننده راس‌های گراف است. رأس $v_i \in V$ شامل اطلاعات l_i است که اطلاعات پیکربندی عضو i ام را در بردارد. l_i در بردارنده اطلاعات مکان، جهت‌گیری و مقیاس برای عضو iام است. اطلاعات پیکربندی مربوط به همه اعضا پیکربندی کل $L = \{l_1, l_2, \dots, l_n\}$ را تشکیل می‌دهند. مدل‌سازی با هدف پیدا کردن پیکربندی بهینه L^* و با استفاده از روش بیزین انجام می‌شود. با داشتن تصور I و گراف $G = (V, E)$، پیکربندی بهینه L^* با استفاده از (۲-۱) به دست می‌آید.

[4] Deformable objects

$$L^* = \boldsymbol{argmax_L} \boldsymbol{P(L|I,\theta)} \quad (۱-۲)$$

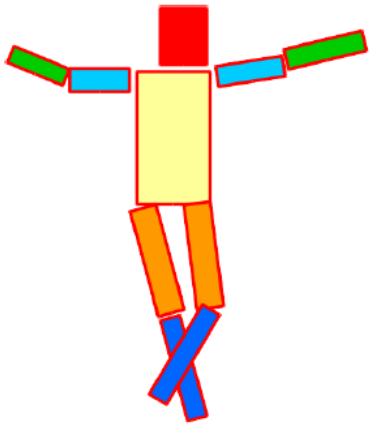
که θ پارامتر مدل است. با استفاده از مدل بیزین داریم

$$P(L|I,\theta) = \frac{P(I|L,\theta).P(L|\theta)}{P(I|\theta)} \quad (۲-۲)$$

$$P(L|I,\theta) \propto P(I|L,\theta).P(L|\theta) \quad (۳-۲)$$

عبارت $P(I|L,\theta)$ نشان‌دهنده میزان تطبیق تصویر با پیکربندی و پارامترهای مدل است. عبارت $P(L|\theta)$ نیز دربردارنده احتمال اولیه برای بررسی پیکربندی است.

همانند مدل تعریف شده برای صورت، برای کل بدن نیز می‌توان مدل مشابهی متشکل از سر، شانه، پا که اعضای دگردیس پذیر متصل بهم هستند در نظر گرفت. نمونه‌ای از مدل بدن در شکل ۲-۲ نمایش داده شده است.



شکل ۲-۲ مدل ژست بدن به صورت مجموعه‌ای از عضوهای مختلف دگردیس پذیر[8]

در [18] تعریف جدیدی از اعضا تحت عنوان پوزلت معرفی شده است. پوزلت قسمت‌هایی از بدن است که با توجه به ظاهر و نوع پیکربندی دسته‌بندی شده‌اند. در تعریف و انتخاب پوزلت دو معیار وجود دارد. اولاً، پوزلت در تصویر ورودی به راحتی قابل تشخیص باشد. دوما، پس از پیدا کردن پوزلت، چگونگی قرارگیری سه بعدی فرد به راحتی قابل تشخیص باشد. در این روش، اعضا با تشکیل خوشه‌هایی از پچ[5]های متراکم در ابعاد دو بعدی و سه بعدی یاد گرفته می‌شوند. خوشه‌بندی بر اساس ژست سه بعدی انجام می‌شود.

[5] patch

سپس برای هرکدام از پوزلت ها یک SVM آموزش داده می‌شود و با استفاده از روش پنجره لغزان برای تطبیق ژست مورد استفاده قرار می‌گیرد.

۲-۱-۳- ساختارهای درختی

در راستای بهبود روش‌های موجود در تخمین ژست بدن انسان با استفاده از ساختارهای تصویری، کارهای برجسته‌ای ارائه شده است. در ادامه به بررسی این روش‌ها می‌پردازیم.

در کار [4] در انتخاب اعضای درگردیس پذیر تغییر به وجود آمده و عضوهای کوچکتری همانند مفصل‌ها به عنوان اعضای دگردیس پذیر انتخاب شده است. نمونه‌ای از مدل در نظر گرفته برای بدن در شکل ۲-۳ نمایش داده شده است. چگونگی قرارگیری مفصل‌ها به موقعیت هندسی اعضای تشکیل دهنده‌ی آن وابسته است. با قرارگیری اعضای کوچک موجود در مدل در کنار هم اعضای بزرگتر تشکیل می‌شوند. مدل پیشنهادی به صورت ترکیبی از دو مدل ظاهری و هم‌رخدادی[6] در نظر گرفته می‌شود. در این ساختار مدل ظاهری با استفاده از ویژگی‌های HOG[7] به دست می‌آید. در حالی که مقدار مدل هم‌رخدادی میزان مشخصی به ازای هر ترکیب از اعضا است. برای گراف $G = (V, E)$ با داشتن تصویر I، تابع امتیاز مدل به صورت (۲-۴) تعریف شده است

$$S(I, z) = \sum_{i \in V} \phi_i(I, z_i) + \sum_{ij \in E} \psi_{ij}(z_i, z_j) \quad (۲-۴)$$

که

$$\phi_i(I, z_i) = w_i^{t_i} . \phi(I, l_i) + b_i^{t_i} \quad (۲-۵)$$

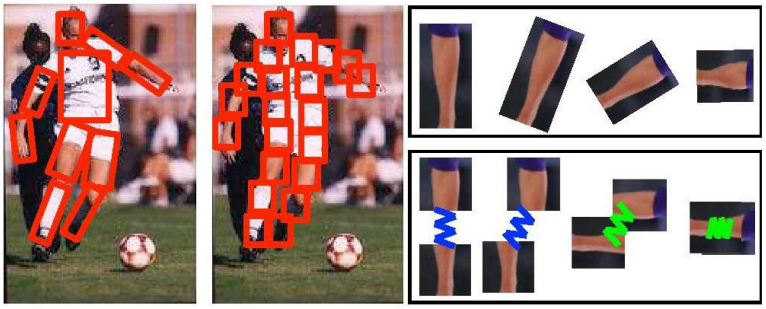
$$\psi_{ij}(z_i, z_j) = w_{ij}^{t_i, t_j} . \psi(l_i - l_j) + b_{ij}^{t_i, t_j} \quad (۲-۶)$$

که $z_i = (l_i, t_i)$ و l_i و t_i مشخص کننده مکان و نوع قرارگیری عضو i ام است. $w_i^{t_i}$، $w_{ij}^{t_i, t_j}$ ، $b_i^{t_i}$ ، $b_{ij}^{t_i, t_j}$ پارامترهای مدل هستند. $\phi_i(I, z_i)$ بردار ویژگی به دست آمده از مکان l_i در تصویر I است. این بردار ویژگی مشخص کننده مدل ظاهری بوده و با استفاده از توصیف‌گر HOG به دست می‌آید. همچنین داریم

[6] Co-occurrence

[7] Histogram Of Gradients

$\psi(l_i - l_j) = [dx\ dx^2\ dy\ dy^2]^T$ که $dx = x_i - x_j$ و $dy = y_i - y_j$ مشخص کننده مکان عضو i نسبت به عضو j هستند. یادگیری ϕ_i و ψ_{ij} با استفاده از روش یادگیری با نظارت با هدف بیشینه کردن تابع امتیاز ارائه شده در (۲-۴) انجام می‌شود. این کار با استفاده از روش برنامه‌نویسی پویا[8] قابل انجام است.



شکل ۲-۳ مدل بدن ارائه شده در [4]

در [19] توصیف‌گر مورد استفاده در مورد تغییر یافته است. در این کار از شبکه‌های عصبی کانولوشنی برای توصیف اجزای بدن انسان و استخراج نوع رابطه بین آن‌ها استفاده شده است. مدل گرافی معرفی شده در این روش از قیدهای سینماتیکی برای تعیین نوع ارتباط بین اعضا استفاده می‌کند و مدل درختی برای بدن تشکیل می‌دهد.

در این روش اطلاعات رابطه‌ی دوتایی بین اعضا و حضور اعضا در هر مکان از تصویر از تکه‌های محلی تصویر استخراج می‌شود. برای استخراج این اطلاعات یک شبکه عصبی آموزش داده می‌شود. خروجی شبکه‌ی آموزش داده شده به عنوان فاکتورهای یگانی و دوتایی در تابع امتیاز مورد استفاده می‌شود. با ترکیب قیود سینماتیکی استخراج شده از مدل در نظر گرفته شده و مقادیر به دست آمده از شبکه‌ی عصبی، تابع امتیاز $F(l,t|I)$ به صورت (۲-۷) شکل می‌گیرد.

$$F(l,t|I) = \sum_{i \in V} U(l_i|I) + \sum_{(i,j)\in E} R(l_i, l_j, t_{ij}, t_{ji}|I) + w_0 \quad (۲-۷)$$

که $U(l_i|I)$ عبارت یگانی مشخص کننده مشاهده‌های محلی حضور عضو i ام در مکان l_i بوده و بر اساس تکه تصویر $I(l_i)$ به دست می‌آید. عبارت یگانی با استفاده از (۲-۸) محاسبه می‌شود.

[8] Dynamic programming

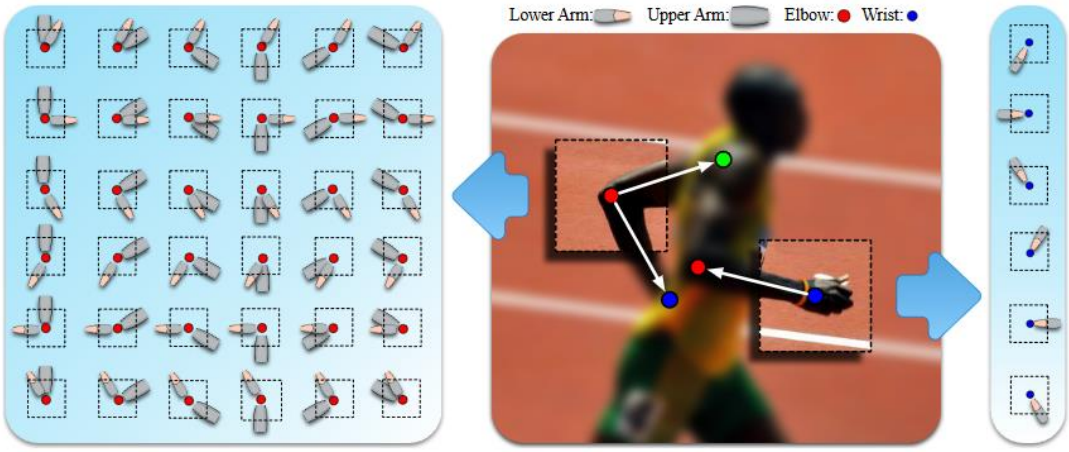
$$U(l_i|I) = w_i\phi(i|I(l_i);\theta) \quad (۸-۲)$$

که $\phi(.|.;\theta)$ عبارت ظاهری با پارامتر θ و وزن w_i است.

در عبارت دوتایی تعریف شده در این مدل اطلاعات همسایگی بین عضوهای i و j از تکه‌های کوچک تصویر استخراج می‌شود. در شکل ۲-۴ نمونه‌هایی از روابط دوتایی به دست آمده از تکه‌های تصویر نمایش داده شده است. همان‌طور که مشاهده می‌شود، با استفاده از تکه تصویر آرنج می‌توان مکان نسبی دو عضو مچ و شانه را به دست آورد. هم‌چنین با استفاده از تکه تصویر مچ می‌توان در مورد مکان نسبی آرنج اطلاعات کسب کرد. روابط نسبی بین دو عضو i و j به چندین نوع $t_{ij} \in \{1, \dots, T_{ij}\}$ با میانگین موقعیت نسبی $r_{ij}^{t_{ij}}$ دسته بندی شده است. تابع مربوط به رابطه دوتایی برای عضوهای $(i,j) \in E$ به صورت (۲-۹) تعریف می‌شود.

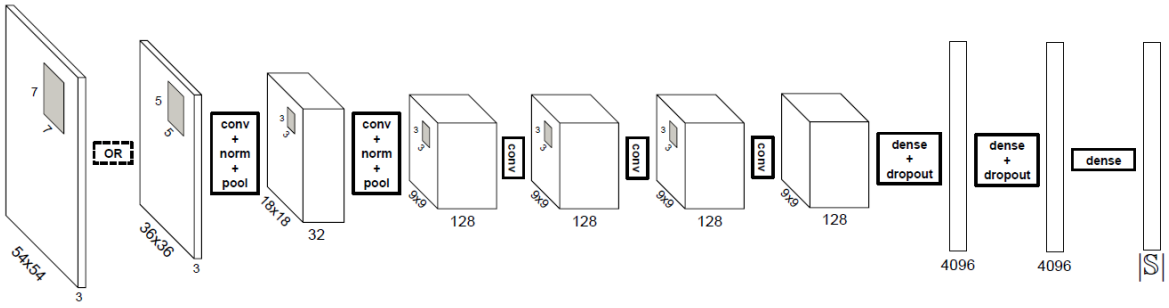
$$R(I_i, I_j, t_{ij}, t_{ji}|I) = \quad (۹-۲)$$
$$\langle w_{ij}^{t_{ij}}, \psi\left(I_j - I_i - r_{ij}^{t_{ij}}\right)\rangle +$$
$$w_{ij}\varphi(t_{ij}|I(l_i);\theta) + \langle w_{ji}^{t_{ji}}, \psi\left(I_i - I_j - r_{ji}^{t_{ji}}\right)\rangle +$$
$$w_{ji}\varphi(t_{ji}|I(l_j);\theta)$$

که $\psi(\Delta l = [\Delta x, \Delta y]) = [\Delta x\ \Delta x^2\ \Delta y\ \Delta y^2]^T$ ویژگی‌های استاندارد درجه دوم تغییر هستند. $\varphi(.|.;\theta)$ نیز عبارت دودویی وابسته به تصویر با پارامتر θاست. $w_{ij}^{t_{ij}}$، w_{ij}، $w_{ji}^{t_{ji}}$ و w_{ji} پارامترهای وزن هستند. نماد $\langle .,. \rangle$ نیز نمایش‌دهنده ضرب نقطه‌ای است.



شکل ۲-۴ روابط دوتایی به دست آمده از تکه‌های کوچک تصویر. با استفاده از نقاط اطراف یک عضو می‌توان مکان نسبی عضوهای همسایه را به دست آورد. [19]

ساختار شبکه‌ی عصبی مورد استفاده برای استخراج فاکتورهای یگانی و دوتایی در شکل ۲-۵ نمایش داده شده است. ورودی شبکه با توجه به مجموعه داده مورد استفاده دارای اندازه ۳۶*۳۶ و یا ۵۴*۵۴ است. شبکه از ۵ لایه کانولوشنی، ۲ لایه انباشت بیشینه[9] و ۳ لایه تماماً متصل[10] و خروجی با ابعاد $|\mathbb{S}|$ تشکیل شده است.



شکل ۲-۵ معماری شبکه‌ی عصبی عمیق استفاده شده در [19]

[9] Max-pooling layer

[10] Fully-connected layer

۲-۱-۴- ساختار گرافی

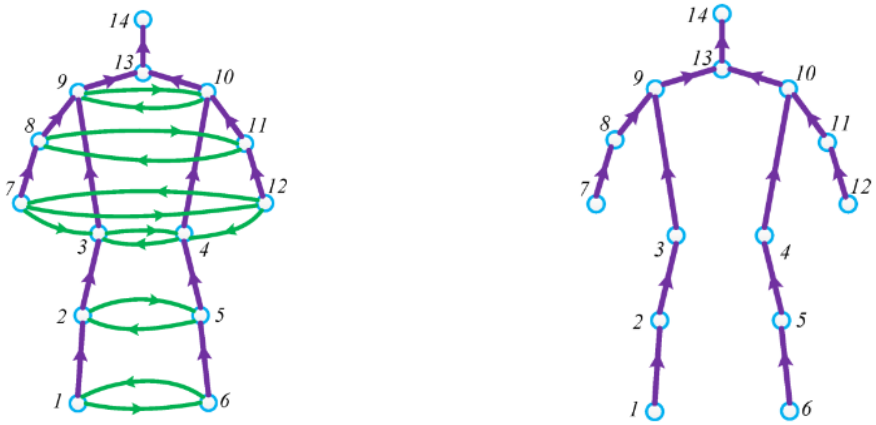
در بخش قبل به بررسی روش‌هایی پرداختیم که برای قیود سینماتیکی بدن از مدل درختی استفاده کرده‌اند. برخی از روش‌ها قیود جدیدی به مدل درختی اضافه کرده‌اند. با اضافه کردن قیود جدید، مدل به صورت دور دار خواهد بود و مدل گرافی به دست می‌آید[20]. برخلاف مدل‌های درختی، استنتاج دقیق در مدل‌های گرافی دارای پیچیدگی بیشتری است و پیچیدگی استنتاج دقیق با افزایش بزرگ‌ترین کلیک[11] موجود در گراف به صورت نمایی افزایش می‌یابد[21]. با افزایش پیچیدگی استنتاج دقیق در گراف‌ها، استفاده از استنتاج تقریبی گزینه بهتری است[22]. روش‌هایی همچون وزن‌دهی دوباره درختی[12] یا انتشار باور[13] در این زمینه پر کاربرد هستند.

در [23] یک مدل گرافی برای حل مشکل انسداد در تخمین ژست بدن انسان ارائه شده است. مدل‌های درختی علی‌رغم سادگی و استنتاج دقیق، قابلیت حل مشکل انسداد را ندارند. یکی از دلایل عدم عملکرد مناسب در این زمینه، عدم وجود اتصال در بین عضوهایی که درحالت فیزیکی اتصالی ندارند، است. مدل گرافی ارائه شده در این کار در شکل ۲-۶ دیده می‌شود. مدل گرافی ارائه شده در [23] با مدل‌های درختی رایج همانند مدل ارائه شده در [4] دارای دو تفاوت است. اولاً، در ساختار گرافی ارائه شده هر عضو دارای وضعیت نمایش است. وضعیت نمایش هر عضو می‌تواند دارای سه مقدار قابل مشاهده، دیگرانسدادی و خودانسدادی است. دوما در این مدل گرافی علاوه بر ارتباط سینماتیکی اعضا که با یال‌های بنفش مشخص شده‌اند، رابطه‌ی بین عضوهای نزدیک ولی بدون اتصال سینماتیکی نیز در نظر گرفته می‌شود. یال‌های سبز نشان‌دهنده این نوع ارتباط‌ها هستند. استفاده از این یال‌ها به حل مشکل خود انسدادی کمک می‌کند.

[11] clique

[12] Tree Reweighting

[13] Belief Propagation (BP)



ب)

الف)

شکل ۲-۶ مقایسه ساختار مدل درختی و مدل گرافی در کاربرد تخمین ژست بدن انسان. الف) مدل درختی رایج برای تخمین ژست بدن انسان استفاده شده در [4]، ب) مدل گرافی پیشنهادی در [23]

در شکل ۲-۷ نمونه‌هایی از عملکرد مدل‌های درختی و گرافی در مواجهه با انسداد نمایش داده شده است. در شکل ۲-۷ الف و ج مدل درختی قصد دارد تا ژست بدن انسان را با وجود انسداد بدن تخمین بزند. به دلیل وجود انسداد در تصویر فرایند انتشار پیام شکست می‌خورد. در صورتی که همانطور که در شکل ۲-۷ ب و د دیده می‌شود، به دلیل وجود یال‌های جدید علاوه بر روابط سینماتیکی و تشکیل مدل گرافی، استنتاج را به درستی انجام می‌شود.

الف) ب)

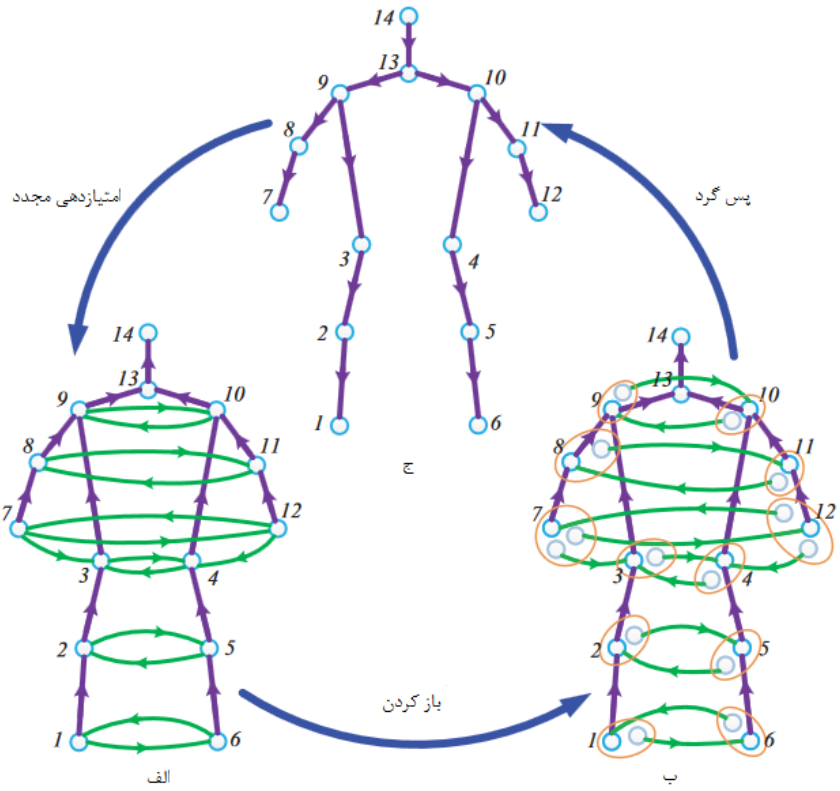
ج) د)

شکل ۲-۷ بررسی تفاوت عملکرد مدل درختی و گرافی در مواجه با خود انسدادی[23]

برای استفاده از روش‌های استنتاج دقیق در گراف‌های دارای دور نیاز به ذخیره جدول حالات و احتمال‌ها و در نتیجه حافظه داریم. در هنگام ذخیره موارد موردنیاز در گراف‌های دارای دور به حافظه بالایی نیاز است که قابل تأمین نیست[24]. در این میان روش‌های همچون [21] معرفی شده‌اند که از نظر تئوری، پیچیدگی زمانی بالایی دارند. ولی با اعمال حقه‌هایی از پیچیدگی زمانی بالای روش جلوگیری شده و در زمان قابل قبول به خروجی دست آورد.

برای استنتاج مدل گرافی ارائه شده در [23] ابتدا با استفاده از مدل درختی، کاندیدهای مختلف ژست بدن تولید می‌شوند. سپس با استفاده از مدل گرافی کاندیدهای تولید شده امتیازدهی می‌شوند. همانطور که در شکل ۲-۸ دیده می‌شود، ابتدا مدل گرافی باز می‌شود و با استفاده از مدل به دست آمده کاندیدهای

ژست بدن انسان تولید می‌شود. در مرحله باز کردن، برای هر گره که به $v_i > 1$ گره دیگر وصل است، $v_i - 1$ گره مجازی ایجاد می‌شود. عملیات بازکردن ارائه شده، معادل با انجام یک مرحله انتشار باور دارای دور[۱۴] در گره ریشه است. پس از باز کردن مدل شکل ۲-۸ ب به دست می‌آید. حال به جای انتشار پیام در مدل گرافی دارای دور، از مدل باز شده استفاده می‌شود. سپس از مدل‌های بازشده تعداد σ مدل برتر انتخاب می‌شود. در این مرحله فرض می‌کنیم که مدل بهینه نیز در بین مدل‌های انتخاب شده خواهد بود. سپس زمان پس‌گرد[۱۵] فرا می‌رسد. پس از پس‌گرد و امتیازدهی مدل با بیشترین امتیاز به عنوان ژست تخمین زده‌شده انتخاب می‌شود.



شکل ۲-۸ استنتاج ارائه شده برای مدل گرافی[23]، الف) مدل گرافی ارائه شده، ب) مدل درختی حاصل از باز کردن، ج) گره‌های به دست آمده از عملیات پس‌گرد

[14] Loopy Belief Propagation

[15] backtracking

۲-۱-۵- شبکه‌های عصبی عمیق

پس از ظهور شبکه‌های عصبی عمیق و عملکرد چشم‌گیر آن‌ها در کاربردهایی همانند دسته‌بندی تصاویر[25] و شناسایی اشیا[26]، استفاده از این شبکه‌ها به مسئله تخمین ژست بدن انسان نیز راه یافت. شبکه‌های عصبی عمیق در مسئله تخمین ژست بدن انسان، دارای قابلیت‌هایی همچون یادگیری ساختار بدن انسان بدون تعریف صریح ساختار درختی یا گرافی و استخراج خودکار ویژگی هستند.

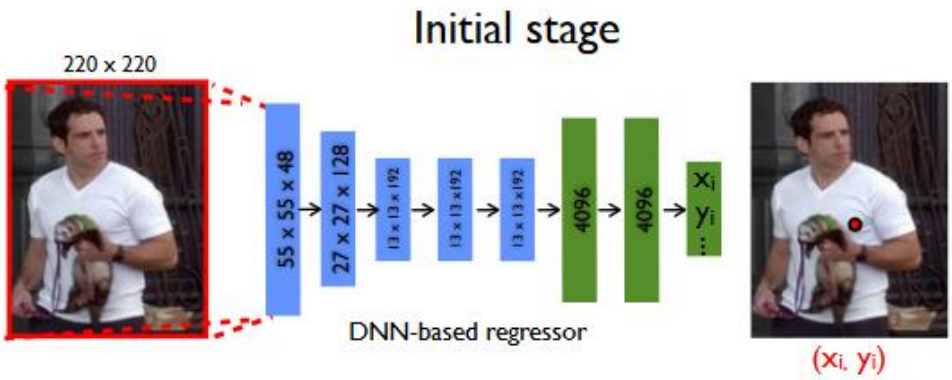
در کارهای اولیه انجام شده در مسئله تخمین ژست بدن انسان، این مسئله به صورت رگرسیون مفصل‌ها بازنمایی شده است. در [27] مسئله تخمین ژست بدن انسان به صورت تابع $\psi(x;\theta) \in \mathbb{R}^{2k}$ بتا پارامتر θ بازنمایی می‌شود. این تابع برای تصویر x خروجی بردار ژست نرمال‌شده را تولید می‌کند. در این تعریف مسئله ژست بدن انسان با استفاده از مکان k عضو کلیدی بدن مشخص می‌شود. مکان عضوهای کلیدی یک بردار ژست به صورت $y = (\dots, y_i^T, \dots)^T, i \in \{1, \dots, k\}$ که y_i شامل مختصات x و y برای عضو i است. داده‌های دارای برچسب مورد استفاده در این کار به صورت (x, y) که x نشان‌دهنده تصویر و y بردار ژست تشکیل شده است، هستند. برای به دست آوردن مختصات واقعی عضوها از (۲-۱۰) استفاده می‌شود.

$$y^* = N^{-1}\left(\psi(N(x);\theta)\right) \quad (۲\text{-}۱۰)$$

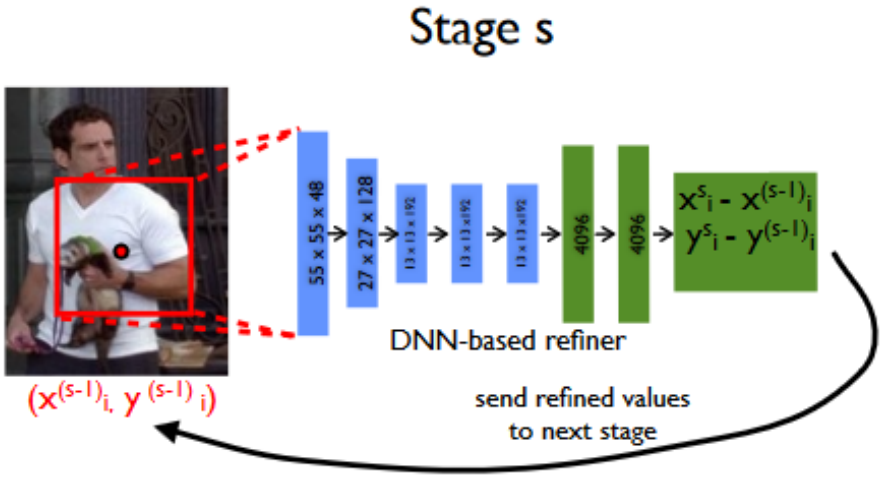
که N برای نرمال کردن ورودی و N^{-1} برای تبدیل ورودی نرمال به موقعیت واقعی استفاده شده است. قدرت و پیچیدگی رابطه تعریف شده به ψ وابسته است که این تابع بر پایه شبکه‌ی عصبی عمیق است. خطای L_2 محاسبه شده برای شبکه‌ی عصبی عمیق به صورت زیر محاسبه می‌شود.

$$argmin_\theta \sum_{(x,y)\in D_N} \sum_{i=1}^{k} \left\| y_i - \psi_i(x;\theta) \right\|_2^2 \quad (۲\text{-}۱۱)$$

شبکه‌ی مورد استفاده در شکل ۲-۹ و شکل ۲-۱۰ نمایش داده شده است. شبکه‌ی عصبی عمیق نمایش داده شده شکل ۲-۹ برای به دست آوردن تخمین اولیه و شبکه‌ی عصبی عمیق نمایش داده شده در شکل ۲-۱۰ برای دریافت نتیجه مرحله‌ی قبلی و بهبود آن استفاده می‌شوند. لایه‌های دارای رنگ آبی نمایش‌دهنده لایه‌های کانولوشنی و لایه‌های سبز نمایش‌دهنده لایه‌های تماماً متصل هستند.



شکل ۲-۹ شبکه‌ی عصبی عمیق به کار رفته برای مرحله اول تخمین ژست در [27]



شکل ۲-۱۰ شبکه‌ی عصبی عمیق به کار رفته در مراحل بعدی در تخمین ژست بدن در [27]

همان‌طور که توضیح داده شد، شبکه ارائه شده در [27] مسئله ژست بدن انسان را به صورت رگرسیون برای یافتن مختصات (x, y) برای هر عضو بدن انسان مورد بررسی قرار می‌دهد. کارهای دیگری نیز همانند [28] و [29] مسئله تخمین ژست بدن انسان را یک مسئله رگرسیون مکان نقاط در نظر می‌گیرند. اما با توجه به بررسی‌های انجام شده در [9]، استفاده از نقشه‌های اطمینان[16] به جای مختصات اعضا باعث ایجاد بهبود می‌شود. در این حالت مدل به جای رگرسیون مکان اعضا، قصد دارد تا نقشه‌های اطمینان را تولید کند. در نقشه‌های اطمینان تولید شده، نقطه‌ی دارای بیشترین اطمینان به عنوان نقطه تخمینی انتخاب می‌شود. نقشه‌های اطمینان از دو جهت باعث بهبود عملکرد مدل می‌شود.

[16] heatmap

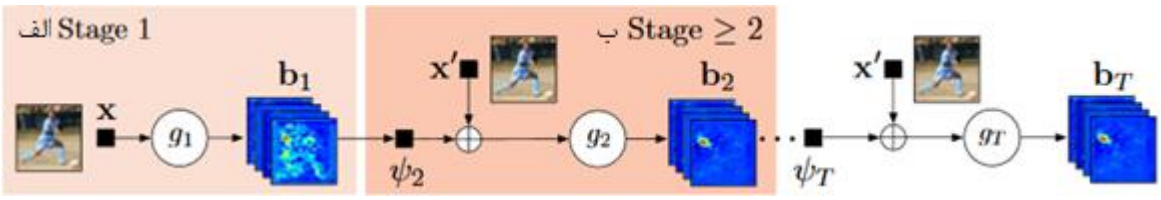
اولاً، دارای قابلیت نمایش و بررسی بهتری هستند. قابلیت نمایش بهتر نتایج منجر به درک بهتری از عملکرد شبکه و بهبود فرایند تفکر می‌شود.

دوما، از انجایی که خروجی شبکه دارای چندین مدالیته[17] خواهد بود، فرایند آموزش شبکه نیز راحت‌تر خواهد بود.

در ادامه‌ی کارهای ارائه شده در تخمین ژست بدن انسان، روش [30] به ارائه ساختار جدیدی برای تخمین ژست بدن انسان پرداخت. این ساختار از چندین مرحله تشکیل شده است که با پیش‌روی در مراحل تخمین ژست بدن بهبود پیدا می‌کند. در [30] ماشین‌های کانولوشنی ژست ارائه شد. ساختار پایه‌ی ماشین‌های کانولوشنی ژست از ماشین‌های ژست ارائه شده در [31] تشکیل شده است. در نتیجه در ادامه ابتدا به توضیح ماشین‌های ژست می‌پردازیم.

ماشین ژست ارائه شده در [31] یک الگوریتم تخمین ترتیبی است. ساختار ماشین ژست ارائه شده از ساختار انتشار پیام برای تخمین نقشه اطمینان اعضا و اصلاح و بهبود تخمین به صورت تکراری الهام گرفته است. ساختار ماشین ژست در شکل ۲-۱۱ نمایش داده شده است.

همان‌طور که در ساختار ماشین ژست در شکل ۲-۱۱ دیده می‌شود، ماشین ژست از تعدادی مرحله تشکیل شده است. در مرحله‌ی اول تنها با استفاده از اطلاعات تصویر تخمین ژست بدن انسان انجام می‌شود. در مراحل بعدی، اطلاعات تصویر و نقشه‌ی باور تولید شده در مراحل قبلی به عنوان ورودی برای تخمین ژست بدن انسان مورد استفاده قرار می‌گیرند.



شکل ۲-۱۱ ساختار ماشین ژست ارائه شده در[31] متشکل از مرحله اولیه الف و مرحله‌های متوالی ب

حال به بررسی دقیق‌تر راه حل ارائه شده در [31] می‌پردازیم. در روش پیشنهاد شده، مکان p امین عضو بدن به صورت $Y_p \in \mathbb{Z} \subset \mathbb{R}^2$ نمایش داده می‌شود. که در این نمایش $\mathbb{Z}$ نشان‌دهنده مجموعه تمامی نقاط

[17] Multi modal

مختصات (u, v) در تصویر است. هدف ما پیش‌بینی $Y = (Y_1, \dots, Y_P)$ نشان‌دهنده مکان P عضو بدن انسان، است. یک ماشین ژست از تعدادی پیش‌بینی کننده‌ی چندکلاسه $g_t(.)$ تشکیل شده است. این پیش‌بینی کننده‌ها برای پیش‌بینی مکان هر عضو در هر مرحله از ساختار سلسله مراتبی آموزش دیده‌اند.در هر مرحله $t \in \{1 \dots T\}$ ، دسته‌بندی کننده‌های g_t، میزان باور رخداد هر عضو در مکان‌های مختلف ورودی را تخمین می‌زنند. در واقع دسته‌بند میزان باور رخداد هر عضو را در مکان‌های $Y_p = z, \forall z \in \mathbb{Z}$ بر اساس ویژگی‌های استخراج‌شده از تصویر در مکان z که با $x_z \in \mathbb{R}^d$ نمایش داده می‌شود و اطلاعات زمینه‌ای حاصل از دسته‌بند قبلی در اطراف هر Y_p در مرحله t، پیش‌بینی می‌کند. در نتیجه دسته‌بند در مرحله $t = 1$ مقادیر باور زیر را تولید می‌کند.

$$g_1(x_z) \rightarrow \{b_1^p(Y_p - z)\}_{p \in \{0,\dots,P\}} \quad (۱۲-۲)$$

که $b_1^p(Y_p = z)$ امتیاز پیش‌بینی شده توسط دسته‌بند g_1 برای رخداد عضو p ام در مرحله اول شبکه در مکان z است. همه باورهای متعلق به عضو p محاسبه‌شده در هر مکان $z = (u, v)^T$ در تصویر به صورت $b_t^p \in \mathbb{R}^{w \times h}$ نمایش داده می‌شود. w و h طول و عرض تصویر هستند. در نتیجه داریم

$$b_t^p[u, v] = b_t^p(Y_p = z) \quad (۱۳-۲)$$

برای راحتی، نقشه‌های باور متعلق به همه‌ی اعضا به صورت $b_t \in \mathbb{R}^{w \times h \times (P+1)}$ (p تعداد اعضای بدن به اضافه‌ی پس‌زمینه) نمایش داده می‌شود.

در مرحله‌های بعدی، دسته‌بند بر اساس ویژگی‌های $x_z^t \in \mathbb{R}^d$ به دست آمده از تصویر و اطلاعات زمینه‌ای حاصل از دسته‌بند قبلی در اطرافهر Y_p، به انتساب باور به هر مکان برایعضو $Y_p = z, \forall z \in \mathbb{Z}$ می‌پردازد.

$$g_t(x'_z, \psi_t(z, b_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p \in \{0,\dots,P+1\}} \quad (۱۴-۲)$$

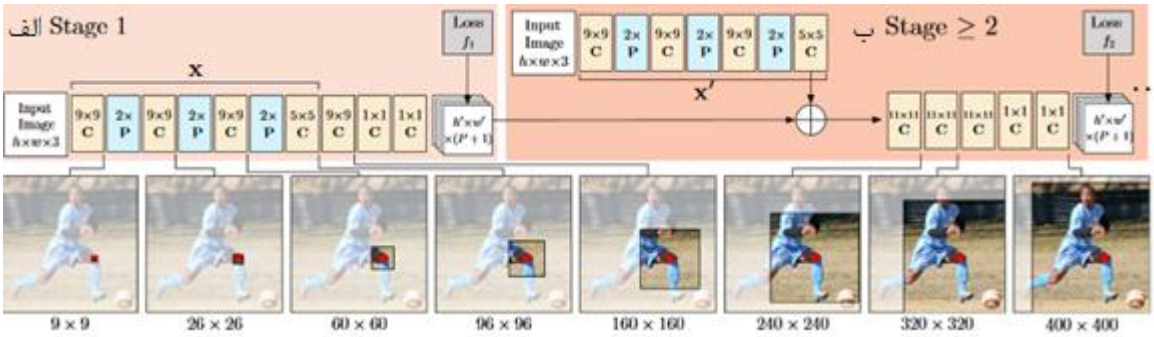
که $\psi_{t>1}(.)$ تابعی برای انجام نگاشت از باور b_{t-1} به ویژگی‌های زمینه‌ای است. باورهای محاسبه شده در هر مرحله، تخمین‌های بهبود یافته برای احتمال رخداد اعضا در مختصات مختلف ورودی را فراهم می‌کنند. در ساختار طراحی شده ویژگی‌های x'_z به کار رفته در هر مرحله الزاماً با ویژگی‌های مرحله اول x یکسان نیست. ماشین ژست ارائه شده از جنگل‌های تصادفی تقویت شده[18] به عنوان پیش‌بینی کننده $(\{g_t\})$

[18] Boosted Random Forests

استفاده می‌کند. هم‌چنین ویژگی‌های استخراج شده از تصویر در همه مراحل یکسان در نظر گرفته می‌شود. در نتیجه داریم $x' = x$. برای ویژگی‌های زمینه‌ای نیز از ویژگی‌های دستی زمینه‌ای استخراج شده $(\psi_t(.))$ استفاده شده و به استخراج محتوای مکانی در تمامی مراحل می‌پردازد.

از آنجایی که ساختار پایه‌ی ماشین کانولوشنی ژست ماشین ژست است، ویژگی‌های مؤثر ماشین ژست شامل یادگیری غیر صریح وابستگی‌های دوربرد[19]، استفاده از سرنخ‌های متعدد، ادغام یادگیری و استنتاج و طرحی چند مرحله‌ای را به ارث می‌برد. با استفاده از ساختار کانولوشنی مزیت‌هایی هم‌چون یادگیری مستقیم ویژگی‌های تصویر و ویژگی‌های محلی زمینه‌ای از تصویر ورودی، قابلیت یادگیری انتها به انتها و قابلیت استفاده برای تخمین ژست بدن انسان در مجموعه داده‌های بزرگ به قابلیت‌های ماشین ژست اضافه می‌شود.

همان‌طور که در شکل ۲-۱۲ دیده می‌شود، با اضافه کردن ساختار کانولوشنی عمیق به ماشین ژست برای استخراج ویژگی و تخمین ژست، ماشین کانولوشنی ژست به دست می‌آید. ماشین‌های کانولوشنی ژست نیز همانند ماشین‌های ژست از چندین مرحله تشکیل شده‌اند که شکل ۲-۱۲ الف نشان‌دهنده ساختار مرحله اول و شکل ۲-۱۲ ب نشان‌دهنده مراحل بعدی است. در این ساختار لایه‌هایی که با حرف C نمایش داده شده‌اند، لایه‌های کانولوشنی و لایه‌هایی که با حرف P نمایش داده شده‌اند، لایه‌های انباشت هستند.

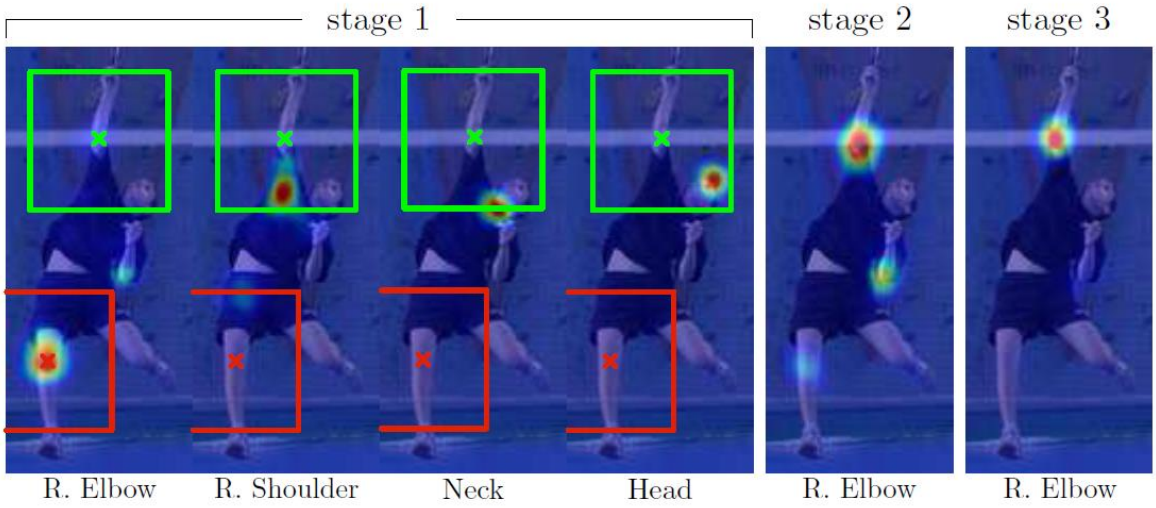


شکل ۲-۱۲ مدل پایه ارائه شده در [30] متشکل از چندین مرحله، بخش الف مرحله اول و بخش ب مراحل بعدی

[19] Long-range

همان‌طور که درشکل ۲-۱۲ الف دیده می‌شود، شبکه به کار رفته در مرحله اول از ۵ لایه کانولوشنی با اندازه کرنل‌های بزرگتر از یک به همراه ۲ لایه کانولوشنی با اندازه کرنل ۱ × ۱ تشکیل شده است. در این مرحله، ماشین کانولوشنی ژست تنها با استفاده از مشاهدات محلی تصویر به پیش‌بینی نقشه‌های باور اعضای بدن می‌پردازد. برای دست یافتن به دقت بهینه، تصاویر به اندازه ۳۶۸ × ۳۶۸ تبدیل شده و به عنوان ورودی به شبکه ارسال می‌شوند. در این صورت میدان تأثیر شبکه دارای اندازه ۱۶۰ × ۱۶۰ پیکسل خواهد بود. با توجه به اینکه در مرحله اول میدان تأثیر شبکه کوچک بوده و تنها به اطراف مکان پیکسل خروجی محدود است، اطلاعات به دست آمده در این بخش اطلاعات محلی هستند. ساختار ارائه شده به صورت شبکه عمیق لغزنده بر روی تصویر حرکت می‌کند. با حرکت بر روی تصویر، اطلاعات محلی در میدان تأثیر با اندازه ذکر شده استخراج می‌شود. سپس بردار خروجی به اندازه $P + 1$ شامل امتیاز برای حضور P عضو و پس‌زمینه در هر مکان از تصویر تولید می‌شود.

در تخمین مکان عضوهای بدن انسان، دقت تخمین ژست برای عضوهایی همانند سر و شانه مقدار قابل قبول و بالایی است. دلیل عملکرد خوب مدل‌ها در مواجهه با این عضوها، ظاهر ثابت آن‌ها در شرایط مختلف است. اما در بخش‌های پایینی اسکلت بدن به دلیل تغییرات گسترده در شکل و نحوه قرارگیری، دقت به دست آمده کمتر است. اطلاعات موجود در نقشه باور در اطراف نقاط مورد نظر، حتی در صورت وجود نویز، حاوی اطلاعات مفیدی هستند. همانطور که در شکل ۲-۱۳ مشاهده می‌شود، در هنگام تشخیص مکان اعضای چالش برانگیز همانند آرنج راست، نقشه باور شانه راست که دارای اطمینان بالایی است، به عنوان نشانه‌ی قوی عمل کرده و به تخمین آرنج راست کمک می‌کند. برای مثال، در مرحله اول مکان آرنج راست اشتباه تشخیص داده شده است. سپس در مرحله‌های بعد با استفاده از اطلاعات مکانی شانه، گردن و سر مکان پیش‌بینی شده تصحیح شده است.



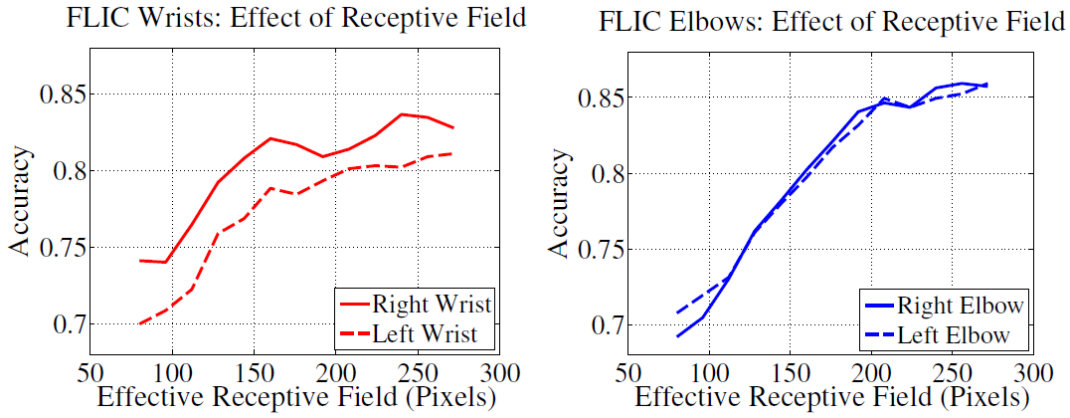
شکل ۲-۱۳ استفاده از اطلاعات مکانی نقشه باور نقاط آسان به عنوان نشانه برای تخمین مکان نقاط سخت [30]

پیش‌بینی کننده در مرحله‌های بعد از مرحله اول $(g_t > 1)$ می‌تواند از اطلاعات مکانی نقشه باور در اطراف آدرس z، $(\psi_{t>1}(.))$ استفاده کرده و با توجه به سازگاری مکانی اعضای بدن تخمین‌های موجود را بهبود دهد. در مرحله دوم، ویژگی‌های تصویر x_z^2 و ویژگی‌های به دست‌آمده از نقشه باور اعضای بدن که توسط تابع ویژگی ψ برای هرکدام از اعضا با استفاده از نقشه باور مرحله قبل محاسبه می‌شود، به عنوان ورودی به دسته‌بند g_2 ارسال می‌شوند. تابع ویژگی ψ، به کدگذاری نقشه باورهای مرحله قبلی در حول مکان z می‌پردازد. در ماشین کانولوشنی ژست نیازی به تعریف صریح تابع ψ نبوده و به عنوان میدان تأثیر بر روی میدان تأثیر مرحله قبلی در نظر گرفته می‌شود.

ساختار شبکه به گونه‌ای طراحی شده است که در لایه خروجی مرحله دوم میدان تأثیر دارای اندازه بزرگی باشد. در نتیجه شبکه دارای قابلیت یادگیری ارتباطات پیچیده و دارای فاصله طولانی بین اجزا است. در هر مرحله ویژگی‌های استخراج شده از خروجی مرحله قبل به عنوان ورودی دریافت می‌شود. با در اختیار داشتن ویژگی‌های خروجی مرحله قبل، لایه‌های کانولوشنی موجود در ساختار مرحله متعاقب با انتخاب موثرترین ویژگی‌ها به ترکیب اطلاعات زمینه‌ای می‌پردازند. نقشه‌های باور در اولین مرحله توسط شبکه‌ای با بررسی محلی تصویر و میدان تأثیر کوچک تولید می‌شوند. در مرحله‌ی دوم شبکه به گونه‌ای طراحی شده است که میدان تأثیر به میزان قابل ملاحظه‌ای افزایش می‌یابد. برای دستیابی به میدان تأثیر بزرگ و دقت بالاتر، دو رویکرد افزایش اندازه کرنل به کاررفته و یا افزایش تعداد لایه‌های کانولوشنی موجود در

ساختار وجود دارد. در حالت اول تعداد پارامترهای شبکه بسیار زیاد خواهد بود. در حالت دوم نیز به دلیل افزایش تعداد لایه‌های کانولوشنی باپیش‌روی در شبکه مشکل گرادیان محوشونده رخ می‌دهد. در ساختار ماشین کانولوشنی ژست افزایش تعداد لایه‌های کانولوشنی به عنوان راهکار برای افزایش اندازه میدان تأثیر استفاده شده است. در نتیجه در مقایسه با راه‌حل افزایش اندازه کرنل‌های موجود، تعداد پارامترها کمتر خواهد بود. با تکرار شبکه در مرحله‌های مختلف و دریافت نقشه باور به عنوان ورودی، از اطلاعات زمینه‌ای تصویر استفاده شده و تخمین‌های اشتباه موجود نیز تصحیح می‌شود.

همان‌طور که در شکل ۲-۱۴ دیده می‌شود، بر اساس آزمایش‌های انجام شده بر روی مجموعه داده‌ی FLIC[32] با افزایش اندازه میدان تأثیر، دقت حاصل نیز افزایش می‌یابد. در این آزمایش، تأثیر اندازه میدان تأثیر با افزایش تعداد لایه‌های کانولوشنی، بدون تغییر تعداد پارامترها و با ورودی به اندازه ۳۰۴ × ۳۰۴ مورد بررسی قرار می‌گیرد. همان‌طور که مشاهده می‌شود، با افزایش اندازه میدان تأثیر دقت نیز افزایش یافته و در اندازه ۲۵۰ به حالت اشباع می‌رسد. با توجه به بهبود دقت با افزایش اندازه میدان تأثیر، می‌توان نتیجه گرفت که میدان تأثیر بزرگ باعث پوشش و کسب اطلاعات بهتری در مورد رابطه‌های با فاصله زیاد بین اعضا می‌شود. در بهترین تنظیمات با توجه به شکل ۲-۱۴ تصویر برش داده شده به اندازه ۳۶۸ × ۳۶۸ تغییر اندازه داده می‌شود. همچنین اندازه میدان تأثیر در مرحله دوم ۳۱ × ۳۱ خواهد بود. این میدان تأثیر در تصویر اصلی محدوده‌ای به اندازه ۴۰۰ × ۴۰۰ را پوشش می‌دهد که قابلیت پوشش همه‌ی جفت اعضای بدن را دارد. با پیش‌روی در مرحله‌های بعد اندازه میدان تأثیر نیز بزرگ‌تر می‌شود.



شکل ۲-۱۴ اثر میدان تأثیر بزرگ در دریافت اطلاعات زمینه‌ای [30]

ساختار طراحی شده برای ماشین کانولوشنی ژست یک ساختار عمیق با تعداد لایه‌های زیاد است. آموزش شبکه با تعداد لایه‌های زیاد مستعد رخداد مشکلی همچون گرادیان محو شونده است. در واقع اندازه گرادیان برگشت داده‌شده[20] پس از گذشت از تعداد لایه‌های میانی زیاد بین لایه ورودی و خروجی کاهش پیدا می‌کند. اما شبکه‌ی طراحی شده در ماشین کانولوشنی ژست از مشکل گرادیان محو شونده جلوگیری می‌کند.

هر مرحله از ماشین ژست کانولوشنی، نقشه باور برای اعضای بدن انسان تولید می‌شود. در هر مرحله تابع خطا بر اساس فاصله l_2 بین نقشه باور پیش‌بینی شده و نقشه باور درست و ایده‌آل محاسبه می‌شود. با مشخص شدن فاصله به عنوان تابع هدف و تلاش در راستای کاهش آن، شبکه در راستای نزدیک شدن به نقشه باور ایده‌آل پیش می‌رود. نقشه باور ایده‌آل برای یک عضو p به صورت $b_*^p(Y_p = z)$ مشخص می‌شود. این نقشه توسط ایجاد قله گاوسی به مرکزیت مکان درست هر عضو p به دست می‌آید. تابع هزینه‌ای که در خروجی هر مرحله محاسبه شده و سعی در کاهش آن داریم به صورت (۲-۱۵) تعریف می‌شود.

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathbb{Z}} \left\| b_t^p(z) - b_*^p(z) \right\|_2^2 \quad (۲\text{-}۱۵)$$

تابع هدف f_t برای هر مرحله در ساختار ماشین‌های کانولوشنی ژست تعریف می‌شود. هدف نهایی در ساختار کلی شبکه توسط جمع هزینه‌های f_t محاسبه شده در هر مرحله به دست می‌آید و به صورت (۲-۱۶) نمایش داده می‌شود.

$$\mathbb{F} = \sum_{t=1}^{T} f_t \quad (۲\text{-}۱۶)$$

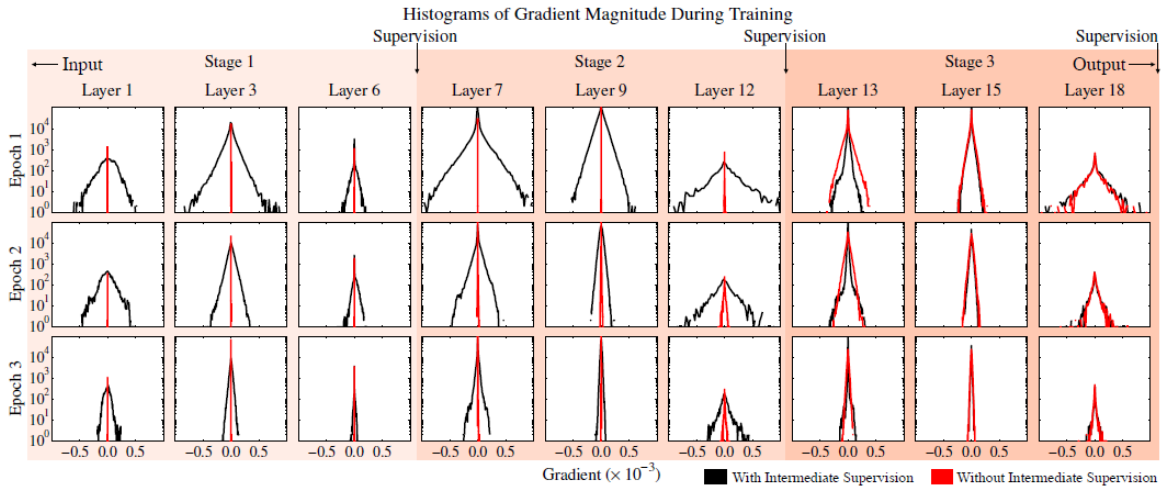
با استفاده از گرادیان نزولی تصادفی[21] آموزش تمامی T مرحله در شبکه انجام می‌شود. برای اینکه ویژگی تصویر x' در تمامی مراحل بزرگتر از یک به اشتراک گذاشته شود، بایستی وزن‌های لایه‌های کانولوشنی در لایه‌های موجود در مراحل مذکور نیز به اشتراک گذاشته شوند.

[20] Back-propagated

[21] Stochastic Gradient Descent

تابع هزینه ارائه شده در (۲-۱۶) دارای قابلیت محاسبه در مراحل مختلف شبکه است. در نتیجه در انتهای هر مرحله از شبکه هزینه محاسبه می‌شود. محاسبه تابع هزینه در انتهای هر مرحله نقش نظارت میانی[۲۲] را ایفا می‌کند. استفاده از نظارت میانی در ساختار شبکه باعث می‌شود تا علی‌رغم وجود تعداد لایه‌های زیاد در ساختار شبکه، مشکل گرادیان محو شونده رخ ندهد. زیرا تابع هزینه محاسبه شده در هر مرحله به تقویت گرادیان محاسبه شده کمک می‌کند.

برای اثبات ادعای موجود تأثیر محاسبه تابع هزینه و اضافه کردن آن به ساختار شبکه در شکل ۲-۱۵ نمایش داده شده است. در این شکل اندازه گرادیان در عمق‌های مختلف شبکه در اپوک‌های[۲۳] آموزش با اعمال تابع هزینه و بدون اعمال آن مورد بررسی قرار گرفته است.



شکل ۲-۱۵ تأثیر افزودن تابع هزینه به عنوان سرپرست میانی در رفع مشکل گرادیان محو شونده[30]

ماشین کانولوشنی ژست را بدون وجود نظارت میانی مورد بررسی قرار می‌دهیم. در صورتی که در اپوک‌های اولیه از لایه‌های خروجی به سمت لایه‌های ورودی حرکت کنیم، اندازه گرادیان به صورت قله‌ی تیز در حول نقطه صفر خواهد بود. اگر اندازه گرادیان را در شرایط ذکر شده و در ماشین کانولوشنی ژست در حضور نظارت میانی به دست آوریم، شاهد واریانس بیشتری خواهیم بود. واریانس بزرگتر گرادیان حاکی از

[22] Intermediate Supervision

[23] epoch

رخداد یادگیری در تمامی لایه‌ها است. البته شایان توجه است که با پیشروی فرایند آموزش، واریانس موجود در توزیع گرادیان کاهش می‌یابد که این امر نشان‌دهنده همگرایی مدل است.

ماشین کانولوشنی ژست ساختار جدید و ارزشمندی برای تخمین ژست بدن انسان در تصویر ارائه داد. در ادامه کارهایی هم‌چون [33]، [5]،[34]، [35] راهکارهایی برای تخمین ژست بدن انسان با استفاده از شبکه‌های عصبی عمیق ارائه داده‌اند. برای مثال، به بررسی مختصر [33] می‌پردازیم. ماهیت روش ارائه شده در [33] مشابه ماشین‌های کانولوشنی ژست است. با این تفاوت که دارای واحدهای ساختاری متفاوتی است و از ماژول آورگلس[۲۴] که دارای ساختاری مشابه ساعت شنی است استفاده می‌کند.. هم‌چنین در ساختار ماشین‌های کانولوشنی ژست برای محاسبه تخمین ژست در مقیاس‌های مختلف، تمامی شبکه به طور جداگانه برای هر مقیاس اجرا می‌شود. در واقع ماشین کانولوشنی ژست از خط لوله[۲۵]های متعدد برای به دست آوردن تخمین در مقیاس‌های مختلف استفاده می‌کند. در صورتی [33] از تنها یک خط لوله استفاده می‌کند.

۲-۲- تخمین ژست بدن انسان در ویدیو

در بخش ۲-۱- به بررسی روش‌های موجود برای تخمین ژست بدن انسان در تصاویر ثابت پرداختیم. حال قصد داریم تا ژست بدن انسان را در دنباله‌ای از فریم‌های ورودی یا همان ویدیوی ورودی تخمین بزنیم. در چالش تخمین ژست بدن انسان در ویدیو علاوه بر اطلاعات استخراج شده از تک تک فریم‌ها به صورت جداگانه، اطلاعات بین فریم‌ها را نیز در دست داریم. حال برای تخمین ژست بدن انسان در ویدیو می‌توان با صرف نظر از ارتباط زمانی بین فریم‌ها، از روش‌های ارائه شده در بخش قبل برای تخمین ژست بدن انسان در تصویر استفاده کرد. حتی در صورت استفاده از روش‌های مرز علم هم‌چون شبکه‌های عصبی کانولوشنی عمیق ارائه شده در بخش قبل، به علت وجود چالش‌هایی در ورودی ویدیو، نتایج مناسبی به دست نمی‌آید. چالش‌های موجود در ورودی ویدیوی ورودی شامل مواردی هم‌چون ژست‌های نامعمول، تغییر شکل چالش برانگیز، زاویه‌ی دید، خودانسدادی و یا انسداد با حضور اجسامی بر روی بدن است. در

[24] hourglass

[25] pipeline

شبکه‌های عصبی، داده‌ی آموزش نشانه‌گذاری شده دارای اهمیت بالایی است، در حالی که با رخداد چالش‌های ذکر شده، داده‌ی آموزش مناسب در دسترس وجود ندارد. از این رو شبکه‌های عصبی در تخمین ژست بدن انسان تنها با در نظر گرفتن فریم‌ها به صورت جداگانه عملکرد خوبی از خود نشان نمی‌دهند[36]، [37]، [38].

برای بررسی روش‌های ارائه شده برای تخمین ژست بدن انسان در ویدیو، آن‌ها را در دو دسته‌ی کلی زیر قرار می‌دهیم.

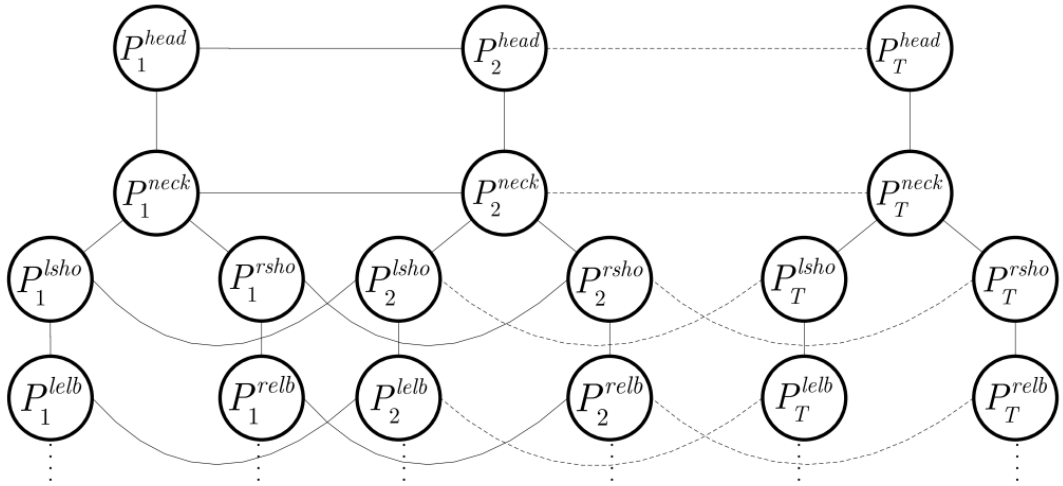
- مدل‌های گرافی مکانی-زمانی
- شبکه‌های عصبی عمیق

۲-۲-۱- مدل‌های گرافی زمانی مکانی

در این دسته از روش‌ها برای مدل‌سازی ژست بدن انسان در دنباله‌ای فریم‌ها از مدل‌های گرافی استفاده می‌شود. مدل گرافی به نحوی تعریف می‌شود که رابطه مکانی بین اعضای بدن در یک فریم و رابطه زمانی بین مکان یک عضو در فریم‌های مختلف را پوشش دهد. با توجه به پیچیدگی داده‌های ویدیو، در صورتی که مدل گرافی تعریف شده نیز پیچیده باشد، استنتاج دقیق و دستیابی به تخمین مسئله NP-hard خواهد بود. در این راستا می‌توان با ساده کردن مدل تعریفی برای ژست بدن انسان در ویدیو به استنتاج بهتر کمک کرد. همچنین می‌توان به جای استفاده از استنتاج دقیق از روش‌های تقریبی استنتاج استفاده کرد[2]، [36].

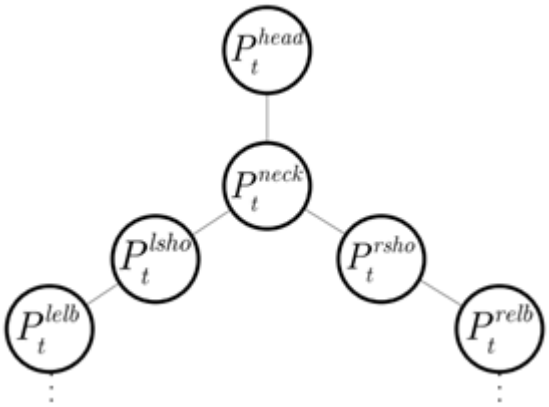
روش ارائه شده در [39] یک نمونه‌ی مناسب از چگونگی عملکرد مدل گرافی در مواجهه با مسئله تخمین ژست بدن انسان در ویدیو است. از این رو به توضیح دقیق‌تر آن می‌پردازیم.

مدل کلی در نظر گرفته شده برای ژست بدن انسان در توالی فریم‌های ورودی در شکل ۲-۱۶ نمیش داده شده است. گره‌های موجود، نقاط کلیدی موردنظر بدن هستند. این نقاط شامل سر، گردن، شانه راست و چپ و ... هستند. بین گره‌های موجود در مدل دو نوع ارتباط تعریف می‌شود: ارتباط مکانی بین اجزای مختلف بدن در یک فریم (مثل ارتباط بین P_1^{head}و P_1^{neck}) و ارتباط بین یک جز از بدن در فریم‌های مختلف (مثل ارتباطبین P_1^{head} و P_2^{head}).



شکل ۲-۱۶ مدل مکانی-زمانی ژست بدن انسان[39]

حال با استفاده از روش تقسیم و غلبه، مسئله به دو مسئله مختلف و مدل به دو مدل تقسیم شده و حل می‌شود. در یک مدل، از رابطه‌ی زمانی بین فریم‌ها چشم‌پوشی کرده و صرفاً به بررسی رابطه‌ی بین اجزای بدن در هر فریم پرداخته می‌شود که مدل به دست آمده در شکل ۲-۱۷ نمایش داده شده است.



شکل ۲-۱۷ مدل ساختار تصویری در عدم حضور زمان و بررسی ژست بدن در تصویر [39]

در این مدل احتمال رخداد ژست بدن P با داشتن فریم موردنظر I محاسبه می‌شود، که $P = \{P^i\}_{i=1}^{K}$ متشکل از اجزای بدن است. در نهایت P توسط گراف G نمایش داده می‌شود. در این گراف، گره‌ها بیانگر اجزای بدن و یال بین دو گره نشان‌دهنده ارتباط فیزیکی دو جزء است. در نتیجه قصد داریم تا$p(P|I)$ را به دست آوریم. این احتمال با استفاده از رابطه‌ی (۲-۱۷) به دست می‌آید.

$$p(P|I) \propto \exp(\sum_{(i,j)\in E} \Psi(P^i, P^j) + \sum_i \Phi(P^i, I)) \quad (۲-۱۷)$$

در واقع از دو فاکتور $\Psi(P^i, P^j)$ و $\Phi(P^i, I)$ برای محاسبه احتمال موردنظر مورد نیاز است. فاکتور $\Psi(P^i, P^j)$ به محاسبه سازگاری هر دو جزء برای قرار گرفتن در کنار هم می‌پردازد. برای محاسبه این فاکتور از رابطه (۲-۱۸) استفاده می‌شود.

$$\Psi(P^i, P^j) = \theta_{ij}^{z^i, z^j} + \alpha_{ij}^{z^i, z^j} \psi(P^i, P^j) \quad (۱۸-۲)$$

که $\theta_{ij}^{z^i, z^j}$ عبارت نشان‌دهنده سازگاری و رخداد دو نوع عضو بدن در کنار هم است. با درنظر داشتن $dx = x^i - x^j$ و $dy = y^i - y^j$ داریم $\psi(P^i, P^j) = [dx, dx^2, dy, dy^2]$ که نشان‌دهنده مکان نسبی عضو i نسبت به عضو j است. $\alpha_{ij}^{z^i, z^j}$ پارامتر مدل است که در صورت انتخاب نوع z^i و z^j برای بخش‌های i و j ، به عنوان فاصله نسبی در نظر گرفته می‌شوند.

فاکتور $\Phi(P^i, I)$ نیز به بررسی امکان قرار گیری جزء موردنظر i در مکان (x^i, y^i) می‌پردازد. این فاکتور نیز توسط رابطه (۲-۱۹) محاسبه می‌شود.

$$\Phi(P^i, I) = \theta_i^{z^i} + \beta_i^{z^i} \phi(I, P^i) \quad (۱۹-۲)$$

که $\theta_i^{z^i}$ نشان‌دهنده سازگاری نوع اختصاص داده شده برای بخش i است. $\beta_i^{z^i}$ الگوی وابسته به نوع در بخش i است و $\phi(I, P^i)$ بردار ویژگی تولید شده توسط هیستوگرام گرادیان‌ها در بخش p^i در فریم I است. برای یادگیری پارامترهای موجود در روابط بالا شامل $\theta_{ij}^{z^i, z^j}$، $\alpha_{ij}^{z^i, z^j}$، $\theta_i^{z^i}$ و $\beta_i^{z^i}$ ، از روش بیشینه‌سازی حاشیه[۲۶] و یا بردار پشتیبان ساخت یافته استفاده می‌شود.

پس از معرفی چگونگی محاسبه فاکتورهای موردنیاز برای محاسبه احتمال، نوبت به بخش استنتاج می‌رشد. در بخش استنتاج قصد داریم تا با ژست بهینه p^* برای تصویر I را بیابیم. از آنجایی که مدل در نظر گرفته‌شده دارای ساختار درختی است، استنتاج دقیق قابل انجام بوده و می‌توان از بیشینه‌سازی پسین[۲۷] با استفاده از بیشینه ضرب[۲۸] برای یافتن پیکربندی بهینه جهانی استفاده کرد. پیغام منتشر شده از بخش i به بخش j با استفاده از رابطه‌های (۲-۲۰) و (۲-۲۱) محاسبه می‌شود:

[26] Max Margin

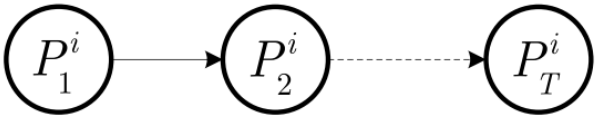
[27] Maximum a Posterior

[28] Max product

$$m_i(P^j) \propto \sum_{Pi} \Psi(P^i, P^j) a_i(P^i) \quad (۲۰-۲)$$

$$a_i(P^i) \propto \Phi(P^i, I) \prod_{k \epsilon kids_i} m_k(P^i) \quad (۲۱-۲)$$

حال به بررسی مدل دیگر می‌پردازیم. در مدل دوم که در شکل ۲-۱۸ نمایش داده شده است، رابطه زمانی بین فریم‌ها درنظر گرفته می‌شود. در این مدل هدف بررسی ارتباط بین اعضا در فریم‌های متوالی است. در این راستا از فیلتر نمونه‌ها[۲۹] استفاده شده است.



شکل ۲-۱۸ مدل پنهان مارکوفی در صورت بررسی ارتباط زمانی اجزا

P_i^t نشان‌دهنده متغیر حالت مشخص کننده پارمتر وابسته حرکت در بخش i در زمان t است. با در دست داشتن دنباله‌ای از تصاویر مشاهده شده به صورت $I_{1:t} = \{I_1, \dots, I_t\}$ می‌توان به محاسبه احتمال پسین P_t^i به صورت زیر پرداخت.

$$p(P_t^i | I_{1:t}) \propto p(I_t | P_t^i) \int p(P_t^i | P_{t-1}^i) p(P_{t-1}^i | I_{1:t-1}) dP_{t-1}^i \quad (۲۲-۲)$$

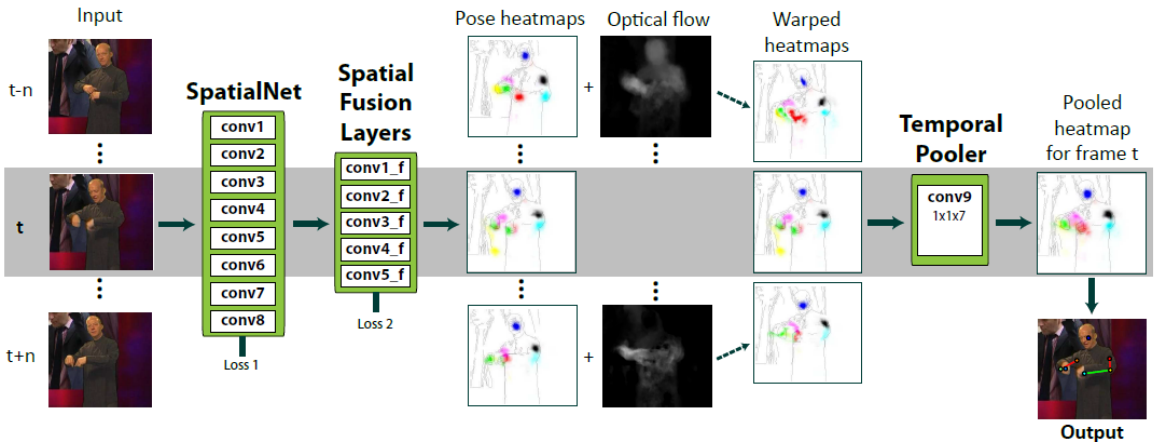
در واقع می‌توان محاسبه احتمال را به دوبخش $p(P_t^i | P_{t-1}^i)$ که مدل دینامیک بین دو حالت است و $p(I_t | P_t^i)$ که مدل مشاهده است، تقسیم کرد.

۲-۲-۲- شبکه‌های عصبی عمیق

همان‌طور که در بخش تخمین ژست بدن انسان در تصویر نیز بررسی کردیم، شبکه‌های عصبی عمیق دارای مزیت‌هایی همچون عدم نیاز به تعریف صریح مدل گرافی برای مدل‌سازی روابط بین اعضا و عدم نیاز به استخراج دستی ویژگی‌ها و ... هستند. از این رو پس از ظهور شبکه‌های عصبی عمیق، در چالش تخمین ژست بدن انسان در ویدیو نیز جایگاه ارزشمندی برای خود کسب کرده‌اند.

[29] Particle Filter

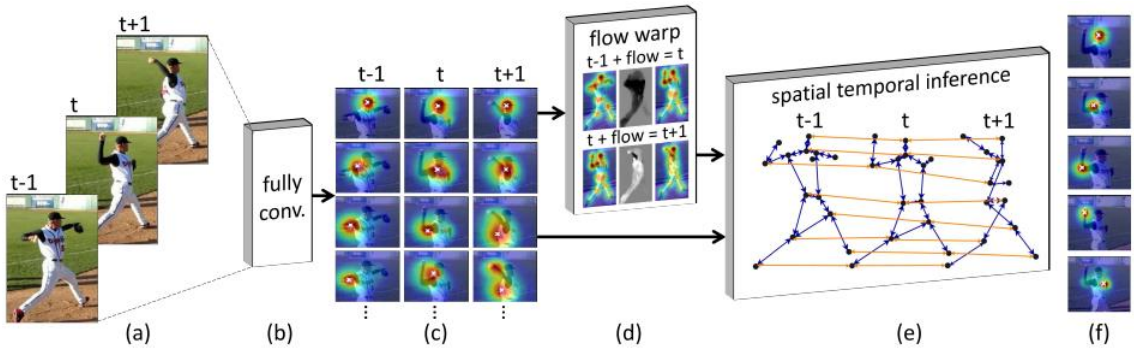
در [40] شبکه‌ی عصبی عمیق شامل سه بخش طراحی شده است. ساختار این شبکه در شکل ۲-۱۹ نمایش داده شده است. همان‌طور که دیده می‌شود بخش اول شبکه SpatialNet نامیده شده است. این بخش مسئولیت بررسی رخداد هر عضو را در مکان از تصویر با توجه به ویژگی‌های ظاهری بر عهده دارد. بخش دوم شبکه که Spatial Fusion Layers نام دارد، دارای وظیفه ارتباط بین زوج عضوهای بدن را بر عهده دارند. در واقع این دو بخش اول به نوعی مسئولیت مدل گرافی در نظر گرفته شده برای تخمین ژست بدن انسان در تصویر را انجام می‌دهند. حال در بخش سوم، پس از محاسبه آپتیکال فلو[۳۰] و پیچش هرکدام از نقشه‌های اطمینان به دست آمده برای همسایه‌های فریم هدف با آپتیکال فلو بین فریم هدف و فریم همسایه، کاندیدی برای نقشه اطمینان عضو مورد نظر به دست می‌آید. حال با ترکیب این کاندیدها با استفاده از یک لایه کانولوشنی با نام Temporal Pooler تخمین نهایی برای هر فریم به دست می‌آید.



شکل ۲-۱۹ ساختار شبکه‌ی عصبی عمیق ارائه شده در [40]

در [40] برای مدل‌سازی رابطه‌ی زمانی بین فریم‌ها از آپتیکال فلو استفاده شده است. در [41] نیز شبکه‌ای با ساختار متفاوت ولی ماهیت مشابه پیشنهاد شده است. در این روش نیز از آپتیکال فلو برای استخراج رابطه‌ی زمانی استفاده می‌کند. ساختار کلی این شبکه در شکل ۲-۲۰ نمایش داده شده است. به بررسی این روش می‌پردازیم.

[30] Optical Flow



شکل ۲-۲۰ ساختار شبکه ارائه شده در [41]

شبکه‌ی ارائه شده در [41] ابتدا تعداد کمی فریم مرتبط را به عنوان ورودی دریافت می‌کند. در مرحله‌ی اول شبکه برای رخداد هر عضو در هر کدام از مکان‌های فریم‌های ورودی احتمالی را نسبت داده و نقشه اطمینان تولید می‌کند. برای تولید نقشه‌ی اطمینان شبکه‌ی عصبی عمیق استفاده می‌شود که در بخش (b) شکل ۲-۲۰ نمایش داده شده است. نقشه اطمینان تولید شده در بخش (c) درشکل ۲-۲۰ نمایش داده شده است. نقشه اطمینان حاصل به لایه پیچش جریان[۳۱] و لایه استنتاج مکانی زمانی ارسال می‌شود. مدل در لایه پیچش جریان قصد دارد تا از اطلاعات زمانی بین فریم‌ها استفاده کند. برای برقراری سازگاری زمانی بین مکان‌های پیش‌بینی شده در فریم‌های همسایه از مدل حرکتی صریحی استفاده نشده است. زیرا همان‌طور که می‌دانید مدل حرکت انسان غیرقابل پیش‌بینی است. لایه پیچش جریان که در بخش (d) در شکل ۲-۲۰ نمایش داده شده است، با در نظر گرفتن شار نوری متراکم[۳۲] به تغییر نقشه اطمینان اعضا پرداخته و سازگاری مکان اعضا در فریم‌های همسایه را در نظر می‌گیرد. در نهایت فریم کنونی و نقشه اطمینان حاصل از لایه پیچش جریان به عنوان ورودی به لایه استنتاج مکانی و زمانی داده می‌شوند. این بخش از مدل در بخش (e) در شکل ۲-۲۰ نمایش داده شده است. این لایه با توجه به ارتباط مکانی زمانی بین اعضا به استنتاج پرداخته و برای مکان هرکدام از اعضای بدن تخمین‌های نهایی خود را ارائه می‌دهد. نقشه‌های اطمینان حاصل از تخمین‌های نهایی در بخش (f) شکل ۲-۲۰ نمایش داده شده است.

برای درک عملکرد لایه مکانی زمانی توضیح مختصری ارائه می‌دهیم. گراف نشان‌داده شده در بخش (e) در شکل ۲-۲۰ را به $G = (V, E)$ نشان می‌دهیم که V نشان‌دهنده راس‌های این گراف و $E \subseteq V \times V$

[31] Flow warping layer

[32] Dense optical flow

نشان‌دهنده یال‌های این گراف برای مدل‌سازی رابطه‌های مکانی زمانی است که یال‌های معرفی‌کننده ارتباط مکانی بین اعضا در یک فریم توسط E_s و یال‌های معرفی کننده ارتباط زمانی بین اعضا در فریم‌های متوالی توسط E_f نمایش داده می‌شوند.

با در دست داشتن تصویر I، ژست p دارای ساختار گرافی G، به صورت مختصات دوبعدی در فضای تصویر تعریف می‌شود که داریم $p = \{p_i = (x_i, y_i) \in \mathbb{R}^2 : \forall i \in V\}$. ورودی شبکه برای تخمین ژست، تعدادی از فریم‌های ویدیو $\mathbb{I} = (I_1, I_2, \dots, I_T)$ است که در هر فریم ژست $\mathbb{P} = (p^1, p^2, \dots, p^T)$ تخمین زده می‌شود. با در نظر داشتن اینکه ژست تخمین زده شده برای فری‌های متوالی بایستی دارای سازگاری زمانی باشند، تابع هزینه به صورت (۲-۲۳) تعریف می‌شود.

$$S\left(\mathbb{I}, \mathbb{P}\right)_{slice} = \sum_{t=1}^{T} S(I^t, P^t) + \sum_{(i,i^*)\in E_f} \psi_{i,i^*}(p_i, p'_{i^*}) \quad (۲-۲۳)$$

که $S(I^t, P^t)$ تابع هزینه برای هر فریم است که همانند روش‌های ذکر شده برای تخمین ژست بدن انسان در تصویر از حاصل جمع فاکتور یگانی و دوتایی به دست می‌آید. عبارت دوتایی $\psi_{i,i^*}(p_i, p'_{i^*})$ به برقراری سازگاری زمانی بین عضو i در یک فریم و فریم‌های همسایه می‌پردازد. در این روش $p'_{i^*} = p_{i^*} + f_{i,i^*}(p_{i^*})$ که $f_{i,i^*}(p_{i^*})$ شار نوری محاسبه شده در p_{i^*} است. این بخش نماینده فرایند پیچش جریان است که با اعمال بردار جریان به ازای هر پیکسل، نقشه اطمینان را تغییر داده و نقشه‌ی اطمینان همسایه حاصل می‌شود.

در دو روش بررسی شده از محاسبه‌ی آپتیکال فلو برای استخراج رابطه‌ی زمانی بین فریم‌ها می‌پردازیم. راهکار دیگر در این زمینه استفاده از حافظه کوتاه مدت طولانی کانولوشنی [۳۳] است. به عنوان نمونه به بررسی آخرین کار انجام شده در این زمینه می‌پردازیم.

در [42] از ایده‌ی ماشین‌های کانولوشنی ژست[30] و ماشین‌های ژست[31] برای تخمین ژست بدن انسان در ویدیو استفاده شده است. در نتیجه ابتدا روابط به کار رفته در ماشین کانولوشنی ژست را مجدداً بررسی می‌کنیم.

[33] LSTM

فرض کنیم نقشه اطمینان یا باور به دست آمده برای P عضو بدن و پس‌زمینه در مرحله $s \in \{1,2,\dots,S\}$ به صورت $b_s \in \mathbb{R}^{W \times H \times (P+1)}$ نمایش داده می‌شود و داریم.

$$b_s = g_s(X),\ s = 1, \quad (۲۴-۲)$$

$$b_s = g_s(F_s(X) \oplus b_{s-1}),\ s = 2,3,\dots,S. \quad (۲۵-۲)$$

که $X \in \mathbb{R}^{W \times H \times C}$ تصویر اصلی ورودی ارسال شده به هر مرحله است. $F_s(.)$ یک شبکه عصبی کانولوشنی برای استخراج ویژگی از تصویر است. ویژگی‌های استخراج شده به همراه باورهای به دست آمده از مرحله قبل به یک شبکه عصبی دیگر $g_s(.)$ برای تولید نقشه باور ارسال می‌شوند. در ماشین کانولوشنی ژست تابع‌های $F_s(.)$ و $g_s(.)$ در مراحل مختلف شبکه با وجود یکسان بودن معماری یکسان نیستند. البته $g_s(.)$ در مرحله اول و مراحل بعدی دارای ساختار متفاوتی است. با توجه به اینکه مرحله اول تنها تصویر را به عنوان ورودی دریافت می‌کند، ساختار تعریف شده برای مرحله اول عمیق‌تر از ساختارهای تعریفی برای بقیه مراحل است تا بتواند به دقت خوبی دست یابد. با اعمال نظارت میانی در هر مرحله و پیش‌روی در مراحل تمین به دست آمده بهبود می‌یابد. اما این شبکه به دلیل عدم استفاده از رابطه‌ی زمانی بین فریم‌ها گزینه‌ی مناسبی برای تخمین ژست بدن انسان در ویدیو نیست. از این رو تغییراتی در ساختار شبکه ایجاد شده است. وزن تابع‌های $F_s(.)$ و $g_s(.)$ در تمامی مراحل به اشتراک گذاشته می‌شود. در نتیجه یک ماشین بازگشتی ژست با فرمول‌های (۲-۲۶) و (۲-۲۷) به دست می‌آید.

$$b_t = g_0(X_t),\ t = 1, \quad (۲۶-۲)$$

$$b_t = g(F(X_t) \oplus b_{t-1}),\ t = 2,3,\dots,T. \quad (۲۷-۲)$$

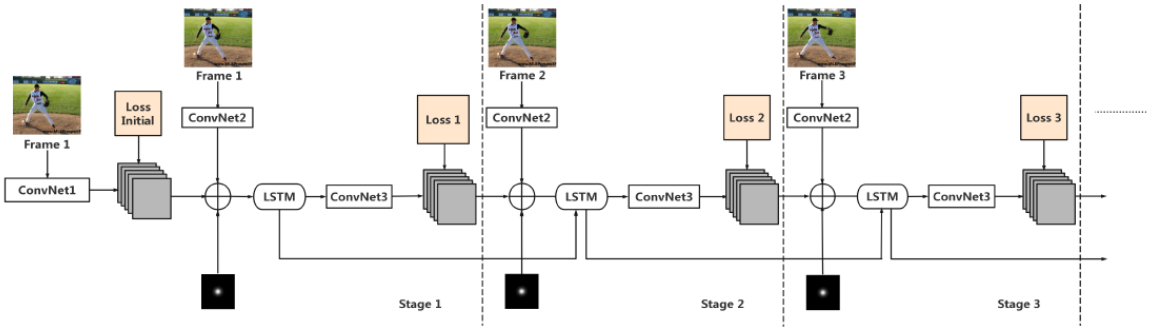
در این تعریف b_t نشان‌دهنده باور به دست آمده برای فریم‌های ویدیوی ورودی است. $X_{t(1\le t\le T)}$ نشان‌دهنده فریم‌های متوالی ویدیو است. تابع $g_0(.)$ دارای ساختار مستقل و متفاوتی با بقیه مرحله‌ها است. اما همه‌ی بقیه‌ی مراحل دارای $g(.)$ یکسانی و وزن‌های مشترکی هستند. با تغییرات ایجاد شده، ماشین بازگشتی ژست با قابلیت دریافت ورودی با تعداد فریم‌های متغیر تولید می‌شود. البته شبکه بازگشتی ایجاد شده عملکرد خوبی نشان نمی‌دهد. از این رو برای بهبود عملکرد شبکه حافظه کوتاه مدت طولانی کانولوشنی به ساختار آن اضافه شده است.

$$b_t = g\left(\tilde{L}\left(F'(X_t)\right)\right),\ t = 1, \quad (۲۸-۲)$$

$$b_t = g\left(\tilde{L}(F(X_t) \oplus b_{t-1})\right),\ t = 2, 3, \dots, T. \quad (۲۹-۲)$$

که $\tilde{L}$ تابعی برای کنترل ورودی و خروجی حافظه است. در رابطه جدید تعریف شده، بخش‌های استخراج ویژگی و تولید خروجی در $g_0(.)$ جدا شده و حافظه کوتاه مدت طولانی کانولوشنی در بین آن‌ها قرار می‌گیرد، زیرا حافظه در نظر گرفته شده تنها برای دریافت و ذخیره و ارسال ویژگی‌ها در نظر گرفته شده است. با تغییر ایجاد شده تابع $g(.)$ در تمامی مراحل یکسان بوده ولی تابع استخراج ویژگی $F'(.)$ در مرحله اول عمیق‌تر از سایر مراحل خواهد بود.

ساختار شبکه که معادل با (۲-۲۸) و (۲-۲۹) است، در شکل ۲-۲۱ نمایش داده شده است.



شکل ۲-۲۱ ماشین ژست با حافظه کوتاه مدت طولانی ارائه شده در [42]

همان‌طور که در شکل ۲-۲۱ دیده می‌شود، فریم‌های همسایه در ویدیو به عنوان ورودی به مرحله‌های مختلف شبکه‌ی طراحی شده ارسال می‌شوند و ژست متناظر با هر فریم به عنوان خروجی آن مرحله تولید می‌شود.

۲-۳- مفاهیم پایه – حافظه کوتاه مدت طولانی کانولوشنی

شبکه‌های عصبی بازگشتی[۳۴] 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[43].

[34] Recurrent Neural Networks

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.

$$g_t = tanh\big(W_{xg} * X_t + W_{hg} * h_{t-1} + \epsilon_c\big) \quad (۳۰\text{-}۲)$$

$$f_t = sigmoid\big(W_{xf} * X_t + W_{hf} * h_{t-1} + \epsilon_f\big) \quad (۳۱\text{-}۲)$$

$$i_t = sigmoid(W_{xi} * X_t + W_{hi} * h_{t-1} + \epsilon_i) \quad (۳۲\text{-}۲)$$

$$o_t = sigmoid(W_{xo} * X_t + W_{ho} * h_{t-1} + \epsilon_o) \quad (۳۳\text{-}۲)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \quad (۳۴\text{-}۲)$$

$$h_t = o_t \odot tanh(C_t) \quad (۳۵\text{-}۲)$$

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.

۲-۴- جمع‌بندی

در این فصل به بررسی کارهای پیشین موجود در تخمین ژست بدن انسان پرداختیم. از آنجایی که هر ویدیو از دنباله‌ای از فریم‌ها تشکیل شده است، در ساده‌ترین حالت می‌توان مسئله تخمین ژست بدن انسان

در ویدیو را به صورت تخمین ژست بدن انسان در مجموعه‌ای از تصاویر مدل‌سازی کرد. از این رو ابتدا روش‌های موجود برای تخمین ژست بدن انسان در تصویر معرفی شدند. سپس در راستای استفاده از اطلاعات زمانی بین فریم‌ها، روش‌های موجود برای تخمین ژست بدن انسان در ویدیو مورد بررسی قرار گرفتند. در ادامه به معرفی مفاهیم پایه موردنیاز در روش پیشنهادی می‌پردازیم.

۳

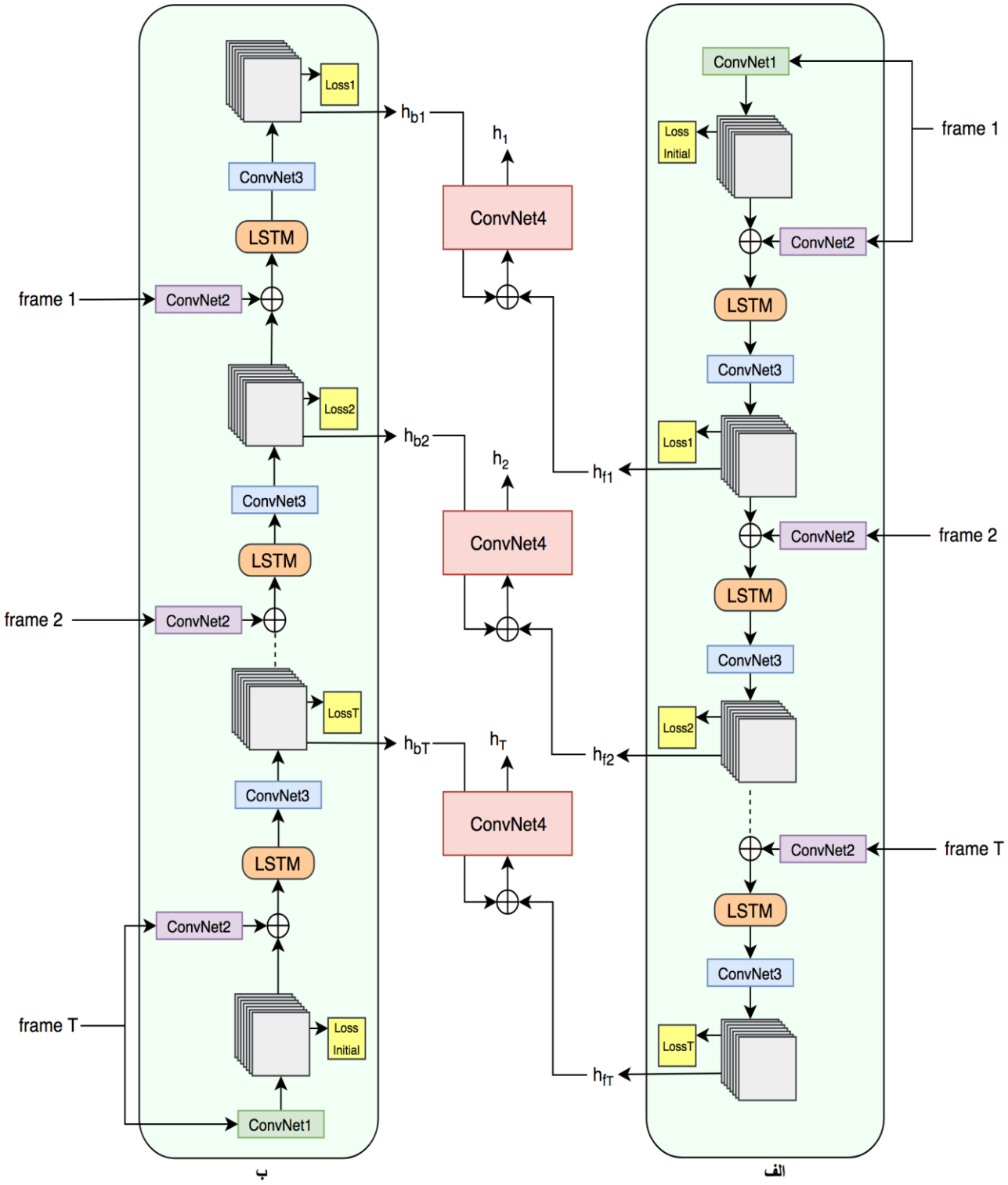
فصل سوم

روش پیشنهادی

روش پیشنهادی ارائه شده در این پژوهش مبتنی بر روش ارائه شده در [42]، [30] و [31] است. ساختار کلی روش پیشنهادی شکل ۳-۱ نمایش داده شده است. شبکه‌ی طراحی شده از سه بخش شبکه رو به جلو، شبکه رو به عقب و شبکه ترکیب تشکیل شده است. در این روش اطلاعات حرکتی رو به جلوی ویدیوی ورودی توسط شبکه رو به جلو و اطلاعات حرکتی رو به عقب ویدیوی ورودی، توسط شبکه رو به عقب استخراج می‌شود. در هر دو شبکه، ژست بدن انسان با اطلاعات استخراج‌شده از ویدیوهای ورودی تخمین زده می‌شود. پس از تولید دو مجموعه تخمین توسط دو شبکه‌ی طراحی شده، تخمین نهایی با استفاده از شبکه‌ی ترکیب به دست می‌آید. خروجی شبکه‌ی ترکیب ژست نهایی تولید شده توسط مدل کلی است. هدف این پژوهش بهبود تخمین ژست بدن انسان 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- شبکه رو به جلو

ساختار شبکه رو به جلو در شکل ۳-۱ بخش الف نمایش داده شده است. این شبکه، دنباله‌ای از T فریم متوالی ویدیو را به عنوان ورودی دریافت می‌کند. هر کدام از فریم‌های ویدیو به عنوان ورودی به یکی از مرحله‌های مدل ارسال می‌شوند و ژست متناظر با آن فریم، به عنوان خروجی مرحله متناظر تولید می‌شود. شبکه با دریافت ورودی $\{frame_t, 1 \le t \le T\}$، خروجی $\{h_{ft}, 1 \le t \le T\}$ را تولید می‌کند. ورودی شبکه دارای ابعاد $h \times w \times 3 \times T$ است که h و w ارتفاع و پهنای فریم ورودی هستند. نقشه اطمینان تولید شده به ازای هر فریم h_{ft} با ابعاد $h' \times w' \times (P+1)$ است. h' و w' ارتفاع و پهنای نقشه اطمینان خروجی هستند. $(P+1)$ نیز تعداد عضوهای کلیدی بدن انسان به همراه پس‌زمینه است. در نقشه‌های اطمینان تولید شده به عنوان خروجی، 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.



شکل ۳-۱ ساختار مدل پیشنهادی ارائه شده در پژوهش

پس از دریافت مجموعه فریم‌های ورودی، پردازش و تخمین ژست از فریم اول شروع شده و به سمت فریم آخر پیش می‌رود. با ارسال فریم‌های ورودی به شبکه، فریم اول از دنباله ورودی ($t = 1$)، به شبکه $ConvNet1$ ارسال می‌شود. این شبکه به طور خاص برای پردازش فریم اول در دنباله‌ی ورودی طراحی شده است. با دریافت این فریم، ژست اولیه توسط $ConvNet1$ تخمین زده می‌شود. تخمین اولیه فریم اول، تنها با استفاده از فریم ورودی تولید می‌شود. از این رو، این تخمین دارای دقت بالایی نیست. بنابراین، برای بهبود دقت نقشه اطمینان تخمین زده شده برای فریم اول، مرحله دیگری طراحی شده است. در راستای بهبود تخمین به دست آمده و تولید تخمین ثانویه، ابتدا فریم $t = 1$ به شبکه $ConvNet2$ ارسال می‌شود. وظیفه‌ی شبکه $ConvNet2$، استخرج ویژگی از فریم ورودی است. با اضافه کردن ویژگی‌های استخراج شده از فریم اول به نقشه‌های اطمینان اولیه‌ی به دست آمده و تخمین مجدد ژست بدن انسان، تخمین ثانویه‌ای با دقت بالاتری به دست می‌آید. ساختار طراحی شده برای بهبود دقت در ادامه دقیق‌تر توضیح داده شده است.

نقشه‌های اطمینان اولیه‌ی به دست آمده از شبکه‌ی $ConvNet1$ و ویژگی‌های استخراج شده از شبکه $ConvNet2$ در کنار هم قرار داده می‌شود. از این دست اطلاعات برای تخمین ژست با دقت بالاتر استفاده می‌کنیم. نکته‌ای مهم در تخمین ژست بدن در ویدیو، بحث رعایت سازگاری زمانی بین تخمین‌های به دست آمده از فریم‌های متوالی است. از این رو، ژست تخمین زده شده برای فریم اول در ژست فریم دوم و همچنین فریم‌های بعدی تاثیرگذار است. برای دخیل کردن اطلاعات حاصل از فریم اول در تخمین ژست فریم‌های بعدی، از حافظه کوتاه مدت طولانی کانولوشنی استفاده شده است. خروجی‌های شبکه‌های $conveNet1$ و $ConvNet2$ که به ترتیب نقشه اطمینان و ویژگی‌های فریم اول هستند، به حافظه ارسال می‌شوند. حافظه کوتاه مدت طولانی کانولوشنی دارای قابلیت به یاد سپاری اطلاعات قدیم، دریافت اطلاعات جدید و فراموش کردن اطلاعات قدیمی بدون فایده است. در نتیجه با پیش‌روی در طول شبکه و دریافت اطلاعات فریم‌های متوالی، فریم‌هایی که دارای فاصله‌ی زیادی با فریم کنونی هستند، دارای اثر کم‌تری نسبت به فریم‌های نزدیک هستند.

پس از ارسال خروجی‌های شبکه‌های $ConvNet1$ و $ConvNet2$ به حافظه کوتاه مدت طولانی کانولوشنی، خروجی حاصل از این حافظه به $ConvNet3$ ارسال می‌شود. شبکه $ConvNet3$ و نقشه‌های اطمینان متناظر با فریم اول را تولید می‌کند. نقشه‌های اطمینان تولید شده توسط این شبکه دارای دقت

بالاتری نسبت به نقشه‌های اولیه تولید شده توسط $ConvNet1$ هستند. عملکرد شبکه برای تولید نقشه‌های اطمینان در فریم اول به صورت (۳-۱) بیان می‌شود.

$$h_{ft} = g_f\left(\tilde{L}_f\left(F'_f(X_t)\right)\right),\ t = 1 \quad (۳\text{-}۱)$$

که h_{ft} نقشه‌های اطیمنان به دست آمده برای فریم اول است.$F'_f(X_t)$ برابر $F_{f0}(X_t) + F_f(X_t)$ است. $F_{f0}(X_t)$ تخمین اولیه به دست آمده برای فریم اول(X_t) توسط شبکه $ConvNet1$ و $F_f(X_t)$ نیز ویژگی‌های استخراج شده برای فریم اول توسط شبکه $ConvNet2$ است. $\tilde{L}_f$ حافظه‌ی کوتاه مدت طولانی کانولوشنی به کار رفته در مدل رو به جلو و g_f نیز شبکه‌ی $ConvNet3$ که تخمین نهایی فریم اول را تولید می‌کند، است.

پس از تخمین ژست و تولید نقشه‌های اطمینان برای فریم اول به سراغ فریم‌های دیگر می‌رویم. با دریافت هر فریم، نقشه‌های اطمینان به دست آمده برای فریم قبلی 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. از آنجایی که این نقشه اطمینان با استفاده از خروجی حافظه که حاوی ترکیبی از اطلاعات نقشه‌های اطمینان پیشین و ویژگی‌های استخراج شده از فریم‌ها است تولید می‌شود، دارای دقت بالاتری است. عملکرد مدل رو به جلو برای فریم‌های غیر از فریم اولیه در (۳-۲) نمایش داده شده است.

$$h_{ft} = g_f\left(\tilde{L}_f\left(F_f(X_t) \oplus h_{f(t-1)}\right)\right),\ t = 2, 3, \dots, T. \quad (۳\text{-}۲)$$

که h_{ft} نقشه‌های اطمینان به دست آمده برای فریم‌های به غیر از فریم اول است. بدین ترتیب، نقشه‌های اطمینان برای فریم‌های ورودی با پیمایش رو به جلوی فریم‌ها تولید می‌شوند.

۳-۲- شبکه رو به عقب

طراحی شبکه رو عقب مشابه شبکه رو به جلو بوده و در شکل ۳-۱ بخش ب نمایش داده شده است. این مدل فریم‌های ورودی را با ترتیب آخر به اول پیماپیش می‌کند. در این شبکه نیز دنباله‌ای از T فریم ویدیو به عنوان ورودی مرحله‌های مختلف به شبکه ارسال می‌شود. شبکه با دریافت ورودی با اندازه‌های $h \times$

$w \times 3 \times T$خروجی‌های$h_{bt}, 1 \leq t \leq T$ را تولید می‌کند. برای سازگاری تخمین‌های تولید شده توسط شبکه رو به جلو و رو به عقب، تنظیمات شبکه‌ها به نحوی انتخاب می‌شود که خروجی‌ها دارای ابعاد یکسانی باشند.

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 در دنباله‌ی ورودی، یا همان فریم آخر ویدیو 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. چگونگی تخمین ژست در این فریم در (۳-۳) نمایش داده شده است.

$$h_{bt} = g_b\left(\tilde{L}_b\left(F'_b(X_t)\right)\right),\ t = T \quad (۳\text{-}۳)$$

که h_{bt} نقشه‌ی اطمینان به دست آمده برای فریم آخر $t = T$ با استفاده از مدل رو به عقب است. $F'_b(X_t)$ برابر $F_{b0}(X_t) + F_b(X_t)$ است که $ConvNet1$ نقش $F_{b0}(X_t)$ را ایفا کرده و تخمین اولیه‌ای برای فریم اول در دنباله‌ی ورودی یعنی فریم آخر ویدیو تولید می‌کند. هم‌چنین $ConvNet2$ نیز نقش $F_b(X_t)$ را ایفا کرده و ویژگی‌های فریم را استخراج می‌کند. $\tilde{L}_b$ حافظه‌ی کوتاه مدت طولانی کانولوشنی به کار رفته در مدل رو به عقب و g_b نیز شبکه‌ی $ConvNet3$ که تخمین نهایی فریم آخر را تولید می‌کند، است.

سپس شبکه به تخمین ژست برای فریم $T - 1$ می‌پردازد و با پیش‌روی برعکس در دنباله فریم‌های ورودی، برای هر فریم از دنباله، ژست تخمین زده می‌شود. عملکرد مدل در مواجهه با این فریم‌ها در (۳-۴) نمایش داده شده است.

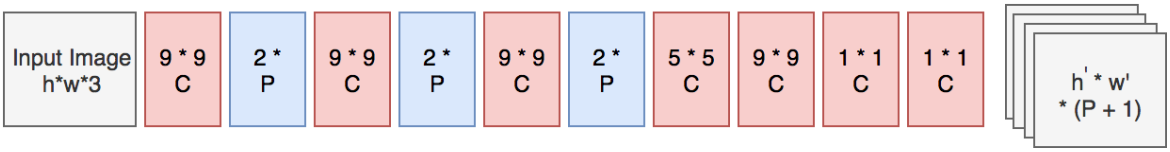
$$h_{bt} = g_b\left(\tilde{L}_b\left(F_b(X_t) \oplus h_{b(t+1)}\right)\right),\ t = 2, 3, \dots, T. \quad (۴\text{-}۳)$$

که h_{bt} نقشه اطمینان به دست آمده برای فریم ورودی به جز فریم آخر توسط مدل رو به عقب است. همان‌طور که دیده می‌شود، خروجی شبکه $ConvNet2$ که ویژگی‌های استخراج شده از فریم ورودی است، به همراه $h_{b(t+1)}$ که نقشه اطمینان تولید شده برای فریم بعدی در پیمایش رو به جلو (فریم قبلی در پیمایش رو به عقب) است، به عنوان ورودی به حافظه‌ی کوتاه مدت طولانی کانولوشنی که با $\tilde{L}_b$ نمایش داده شده است، ارسال می‌شوند.

۳-۳- ساختار شبکه‌های به کار رفته در مدل‌های رو به جلو و رو به عقب

در توضیحات ارائه شده برای شبکه رو به جلو در ۳-۱- و شبکه رو به عقب در ۳-۲- کاربرد شبکه‌های $ConvNet1$، $ConvNet2$ و $ConvNet3$ در تولید نقشه اطمینان مورد بررسی قرار گرفت. اما ساختار درونی این شبکه‌ها هم‌چنان نامعلوم است. از این رو در ادامه به بررسی دقیق‌تر ساختار شبکه‌های $ConvNet1$، $ConvNet2$ و $ConvNet3$ می‌پردازیم.

ساختار شبکه $ConvNet1$ در شکل ۳-۲ نمایش داده شده است. در شبکه‌های نمایش داده شده لایه‌هایی که با C مشخص‌شده‌اند، لایه‌های کانولوشنی و لایه‌هایی که با P مشخص شده‌اند، لایه‌های انباشت هستند. ورودی شبکه دارای ابعاد $h \times w \times 3$ است که h و w ارتفاع و عرض فریم ورودی هستند. خروجی این شبکه نیز دارای ابعاد $h' \times w' \times (P+1)$ است که هم‌بعد با نقشه‌ی اطمینان‌های نهایی تولید شده توسط شبکه برای هر فریم است. شبکه‌ی $ConvNet1$ با هدف تولید نقشه اطمینان اولیه برای فریم ورودی که فریم اول در دنباله ورودی است، طراحی شده است. در هنگام استفاده از مدل رو به جلو، فریم اول ویدیو به عنوان ورودی به این شبکه ارسال می‌شود. در هنگام استفاده از مدل رو به عقب نیز فریم آخر ویدیو که فریم اول دنباله ورودی شبکه است، به شبکه ارسال می‌شود. از آنجایی که شبکه $ConvNet1$ قصد دارد تنها با استفاده از فریم ورودی، تخمینی از نقشه اطمینان ارائه دهد، دارای تعداد لایه‌های بیشتری است. البته تخمین به دست آمده از این شبکه میزان اطمینان بالایی نداشته و به عنوان تخمین اولیه برای شروع فرایند تولید نقشه اطمینان مورد استفاده قرار می‌گیرد.

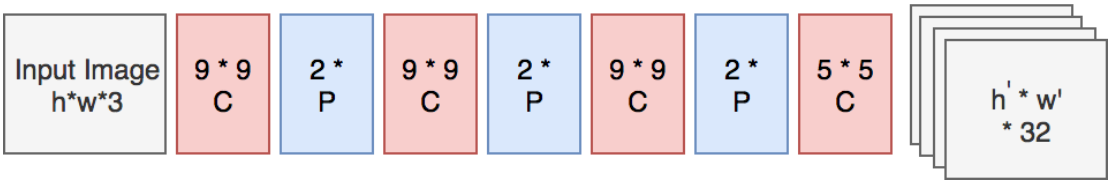


شکل ۳-۲ ساختار شبکه ConvNet1

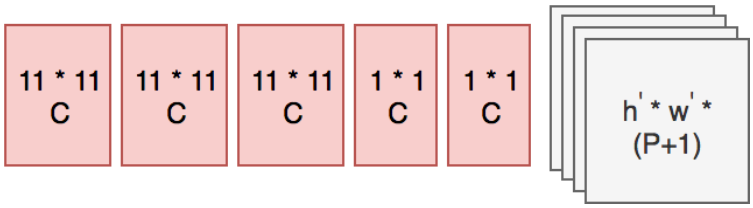
ساختار شبکه $ConvNet2$ در شکل ۳-۳ نمایش داده شده است. ورودی این شبکه فریم‌های ویدیو است که دارای ابعاد $h \times w \times 3$ هستند. h و w ارتفاع و عرض فریم است. هر کدام از فریم‌های ورودی به این شبکه ارسال شده و بردار ویژگی متناظر به صورت خودکار به دست می‌آید.

در شکل ۳-۴ نیز ساختار شبکه $ConvNet3$ نمایش داده شده است. این شبکه دارای وظیفه تولید تخمین نهایی برای هر فریم است. خروجی این شبکه دارای ابعاد $h' \times w' \times (P+1)$ است که نقشه‌های

اطمینان تولید شده برای P عضو به همراه پس‌زمینه است. همان‌طور که دیده می‌شود ابعاد خروجی دو شبکه $ConvNet1$ و $ConvNet3$ یکسان است. زیرا هر دو شبکه دارای هدف تولید نقشه‌ی اطمینان برای اعضا و پس‌زمینه هستند.



شکل ۳-۳ ساختار شبکه ConvNet2



شکل ۳-۴ ساختار شبکه ConvNet3

در مدل‌های رو به جلو و رو به عقب طراحی شده شبکه‌ی $ConvNet3$ با دریافت خروجی حافظه کوتاه مدت طولانی کانولوشنی به تولید نقشه‌های اطمینان می‌پردازد. حافظه کوتاه مدت طولانی کانولوشنی نیز در هر مرحله، تخمین ژست مرحله‌ی قبل یا همان فریم قبل و ویژگی‌های استخراج شده از فریم مدنظر با استفاده از $ConvNet2$ را دریافت می‌کند. در نتیجه ویژگی‌های استخراج شده از فریم نقش مهمی در کسب اطلاعات در زمینه‌ی فریم کنونی و تخمین ژست فریم کنونی دارد. تخمین ژست مرحله قبل که با استفاده از $ConvNet1$ یا $ConvNet3$ تولید شده و به صورت نقشه اطمینان در اختیار قرار دارد، سازگاری زمانی بین تخمین‌های تولید شده را الزام می‌کند.

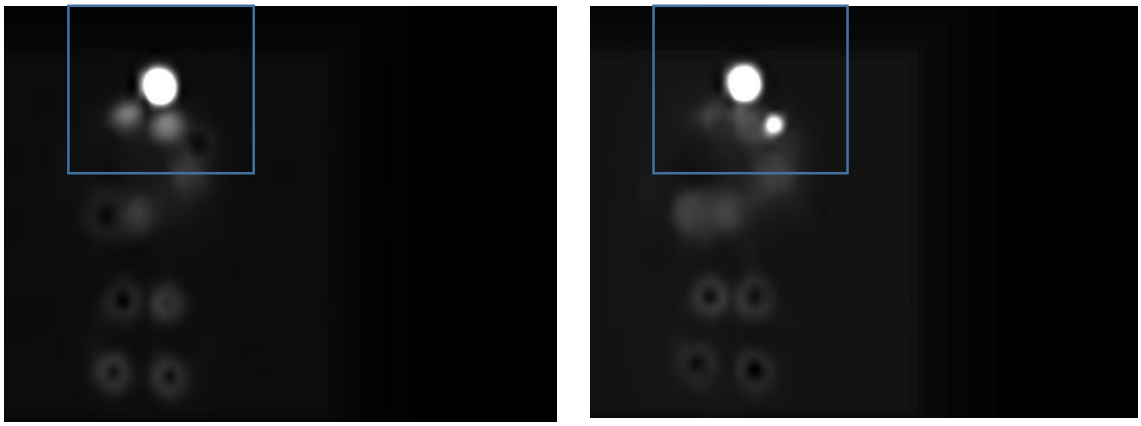
همان‌طور که در شکل ۳-۱ دیدیم، به ازای هر فریم ورودی به مدل‌های رو به جلو و رو به عقب، یک شبکه $ConvNet2$، یک شبکه $ConvNet3$ و یک حافظه کوتاه مدت طولانی کانولوشنی وجود دارد. توجه به بررسی‌های انجام شده در فرمول‌های (۳-۱)، (۳-۲)، (۳-۳) و (۳-۴) این شبکه‌ها که با توابع F (F_f در شبکه رو به جلو و F_b در شبکه رو به عقب)، g (g_f در شبکه رو به جلو و g_b در شبکه رو به عقب) و $\tilde{L}$ ($\tilde{L}_f$ در شبکه رو به جلو و $\tilde{L}_b$ در شبکه رو به عقب) مشخص شده‌اند، در طول مدل‌ها یکسان هستند.

یکسان بودن این توابع در طول شبکه‌های رو به جلو و رو به عقب، با وزن‌های مشترک در طول مدل پیاده سازی شده است. البته وزن‌های توابع در شبکه‌های رو به جلو و رو به عقب یکسان نیستند. متفاوت بودن این پارامترها به مدل‌ها کمک می‌کند تا به طور جداگانه ویژگی‌های حرکتی و ژست را در پیمایش رو به جلو و رو به عقب استخراج کند. در صورتی که این شبکه‌ها در مدل‌های رو به جلو و رو به عقب یکسان در نظر گرفته شود، مدل در یادگیری ویژگی‌های حرکتی موفق نمی‌شود.

۳-۴- شبکه ترکیب

پس از ارسال فریم‌های ویدیو به مدل‌های رو به جلو و رو به عقب، دو مجموعه نقشه اطمینان توسط شبکه‌های رو به جلو و رو به عقب تولید می‌شود. هدف در این بخش تولید نقشه اطمینان نهایی با استفاده از دو مجموعه نقشه اطمینان تولید شده است.

در شکل ۳-۵ نمونه‌ای از نقشه‌های اطمینان به دست آمده از شبکه رو به جلو و رو به عقب برای عضو سر نمایش داده شده است، که نقشه اطمینان سمت راست از شبکه رو به جلو و نقشه اطمینان سمت چپ از شبکه رو به عقب به دست آمده است. بالا بودن مقدار متناظر با هر نقطه فریم ورودی در نقشه اطمینان، نشان‌دهنده بالا بودن احتمال رخداد عضو موردنظر در آن مکان است. برای دست یافتن به تخمین نهایی و نقشه اطمینان نهایی، هر دو نقشه اطمینان به مدل ترکیب طراحی شده ارسال می‌شوند.



شکل ۳-۵ نقشه اطمینان به دست آمده برای سر از مدل‌های رو به جلو و رو به عقب

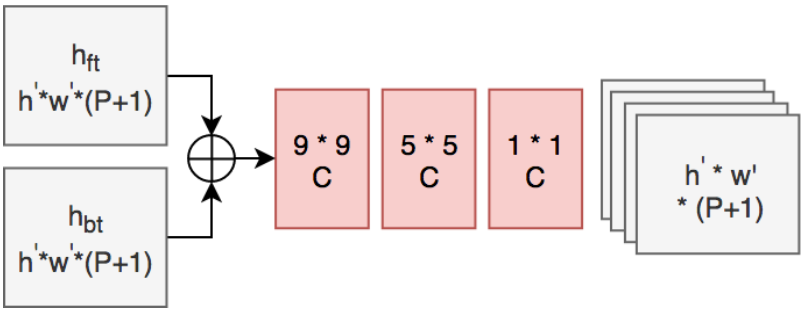
در این راستا شبکه ترکیب $ConvNet4$ طراحی شده است. ساختار این شبکه در شکل ۳-۶ نمایش داده شده است. شبکه ترکیب $ConvNet4$ دو نقشه اطمینان h_{ft} و h_{bt} را که $1 \leq t \leq T$، به عنوان

ورودی دریافت می‌کند. این نقشه‌های اطمینان دارای اندازه $h' \times w' \times (P+1)$ بوده و توسط شبکه‌های رو به جلو و رو به عقب تولید شده‌اند. شبکه $ConvNet4$ دارای وظیفه‌ی تولید نقشه اطمینان نهایی با توجه به اطلاعات موجود در نقشه اطمینان‌های h_{ft} و h_{bt} است. برای مدل‌سازی عملکرد شبکه ترکیب داریم:

$$h_t = Cmb(h_{ft} \oplus h_{bt}), t = 1,2, \dots, T. \quad (۵-۳)$$

که h_t نقشه اطمینان نهایی به دست آمده است. تابع Cmb که نقشه‌های اطمینان به دست آمده از شبکه‌های رو به جلو و رو به عقب را به عنوان ورودی دریافت می‌کند، همان شبکه $ConvNet4$ طراحی شده است.

شبکه ترکیب تخمین نهایی هر مکان را با دریافت پنجره‌ای از نقشه‌های اطمینان رو به جلو و رو به عقب تولید می‌کند. مثالی از پنجره‌های دریافت شده در شکل ۳-۵ نمایش داده شده است (اندازه نواحی دقیق نیست.). شبکه با دریافت اطلاعات نقشه اطمینان در کادرهای مشخص شده، مقدار اطمینان یا باور نهایی را تولید می‌کند. بنابراین در تولید میزان اطمینان نهایی هر نقطه، میزان اطمینان نقطه در نقشه‌های اطمینان رو به جلو و رو به عقب و میزان اطمینان همسایگی این نقاط در هر دو نقشه اطمینان تاثیرگذار هستند.



شکل ۳-۶ ساختار شبکه ترکیب ConvNet4

با توجه به ساختار تعریف شده برای شبکه رو به جلو، اطلاعات فریم‌های قبل در تخمین ژست بدن انسان در هر فریم اثرگذار هستند. همچنین در تخمین ژست در شبکه رو به عقب، اطلاعات فریم‌های بعد به ایفای نقش می‌پردازند. در نتیجه در هنگام استفاده همزمان از شبکه‌های رو به جلو و رو به عقب و نقشه‌های اطمینان تولید شده توسط این شبکه‌ها، از اطلاعات فریم‌های قبل و فریم‌های بعد در تخمین استفاده می‌شود.

۳-۵- تابع هزینه

با ارسال دنباله فریم‌های ورودی به شبکه طراحی شده، خروجی شبکه‌های رو به جلو، رو به عقب و ترکیب نقشه‌های اطمینان متناظر با ورودی است. نقشه‌ی اطمینان تولید شده توسط این شبکه‌ها را به ترتیب برابر با h_{ft}، h_{bt} و h_t هستند. آموزش شبکه بایستی در راستای یادگیری نقشه‌های اطمینان و تولید نقشه‌های h_{ft}، h_{bt} و h_t مشابه مقادیر درست باشد. از این رو هدف در هنگام آموزش، کمینه کردن تفاوت نقشه‌های اطمینان تولید شده و نقشه‌های اطمینان واقعی و درست است. با داشتن مختصات واقعی و درست هر کدام از عضوهای بدن، نقشه اطمینان درست را تولید می‌کنیم. برای تولید نقشه‌ی اطمینان درست، تابع گاوسی در مرکزیت مختصات نقاط قرار داده می‌شود. در راستای کمینه کردن تفاوت نقشه‌های اطمینان تولید شده و نقشه‌های اطمینان درست، از فاصله l_2 استفاده می‌کنیم. با کمینه کردن میزان فاصله‌ی حاصل شده، شبکه به سمت تولید نقشه‌های اطمینانی مشابه با نقشه اطمینان درست پیش می‌رود. در راستای بهینه‌سازی شبکه برای تولید نقشه‌های اطمینان با شرایط ذکر شده، تابع‌های هزینه بخش‌های مختلف به صورت (۳-۶)، (۳-۷) و (۳-۸) تعریف می‌شود.

$$F_{forward} = \sum_{t=1}^{T}\sum_{p=1}^{P+1}\left\|h_{ft}(p) - g.t._t(p)\right\|^2 \quad (۳-۶)$$

$$F_{backward} = \sum_{t=1}^{T}\sum_{p=1}^{P+1}\|h_{bt}(p) - g.t._t(p)\|^2 \quad (۳-۷)$$

$$F_{total} = \sum_{t=1}^{T}\sum_{p=1}^{P+1}\|h_{tt}(p) - g.t._t(p)\|^2 \quad (۳-۸)$$

که $F_{forward}$، $F_{backward}$ و F_{total} به ترتیب تابع هزینه برای مدل رو به جلو، تابع هزینه برای مدل رو به عقب و تابع هزینه کلی است. $h_{ft}(p)$، $h_{bt}(p)$ و $h_{tt}(p)$ نقشه‌های اطمینان تولید شده توسط مدل‌های رو به جلو، رو به عقب و مدل نهایی است. $g.t._t(p)$ نیز میزان نقشه اطمینان واقعی برای فریم‌های ورودی است که در هر سه مدل یکسال است. خطای محاسبه شده برای مدل‌های رو به جلو و رو به عقب در انتهای هر مرحله محاسبه شده و در آموزش شبکه دخالت داده می‌شود. این کار باعث نظارت میانی شده و عملکرد شبکه را بهبود می‌بخشد. همچنین خطا برای نقشه‌ی اطمینان نهایی تولید شده نیز محاسبه شده و در آموزش شبکه دخالت داده می‌شود.

۳-۶- شبکه مورد استفاده در آزمایش

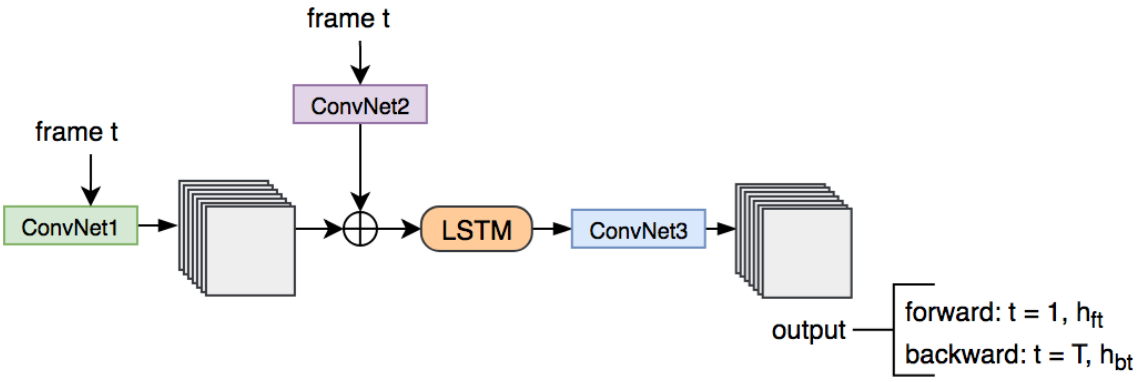
ساختارهایی که تا به این بخش توضیح دادیم، در زمان آموزش مورد استفاده قرار می‌گیرند. در زمان آموزش T فریم به عنوان ورودی به شبکه ارسال می‌شوند. در زمان آزمایش به ازای هر فریم، خروجی شبکه رو به جلو و شبکه رو به عقب تولید می‌شود. پس از تولید خروجی شبکه‌های رو به جلو و رو به عقب، خروجی نهایی با استفاده از شبکه ترکیب به دست می‌آید.

در هنگام آزمایش، ابتدا برای فریم‌های $t = 1$ و $t = T$ که T در این بخش برابر با تعداد فریم‌های ویدیو است، نقشه‌های اطمینان اولیه تخمین زده می‌شود. برای تخمین نقشه‌های اطمینان فریم‌های شروع کننده دنباله از شبکه $deploy1$ که در شکل ۳-۷ 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. در (۳-۹) تخمین اولیه برای فریم $t = 1$ و در (۳-۱۰) تخمین اولیه برای فریم $t = T$ تولید می‌شود. توابع به کار رفته در این بخش برای تولید نقشه‌های اطمینان اولیه، مشابه توابع تعریف شده در (۳-۱) و (۳-۳) هستند.

$$h_{ft} = g_f\left(\tilde{L}_f\left(F'_f(X_t)\right)\right),\ t = 1 \quad (۳-۹)$$

$$h_{bt} = g_b\left(\tilde{L}_b\left(F'_b(X_t)\right)\right),\ t = T \quad (۳-۱۰)$$

پارامترهای توابع $F'_f(X_t)$، $\tilde{L}_f$، g_f متعلق به شبکه رو به جلو بوده و در هنگام آموزش شبکه یاد گرفته می‌شوند. همچنین پارامترهای توابع $F'_b(X_t)$، $\tilde{L}_b$ و g_b نیز متعلق به شبکه رو به عقب هستند. توابع استفاده شده برای شبکه‌های رو به جلو و رو به عقب، دارای مقادیر متفاوتی بوده و در مرحله آموزش به صورت جداگانه یاد گرفته می‌شوند. همان‌طور که در شکل ۳-۷ دیده می‌شود، تخمین‌های به دست آمده در این مرحله تخمین‌های اولیه فریم‌های اول $(h_{ft}, t = 1)$ و آخر $(h_{bt}, t = T)$ دنباله ورودی بوده و شروع کننده تخمین ژست برای فریم‌های متوالی هستند.



شکل ۳-۷ ساختار شبکه deploy1

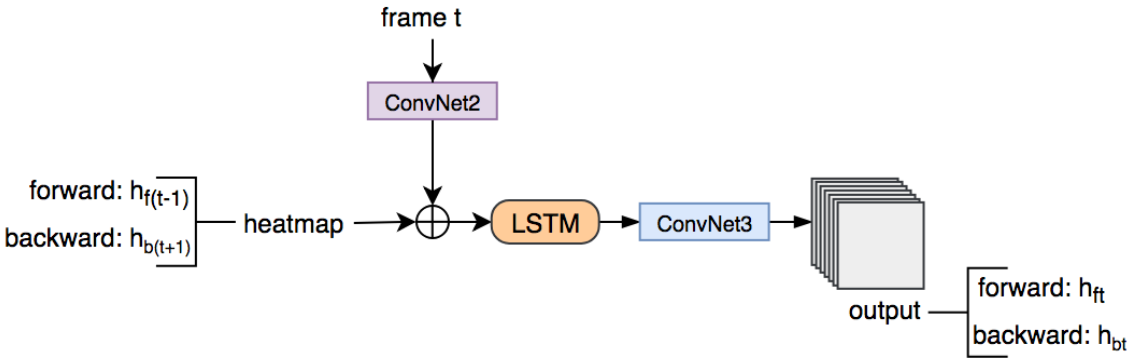
پس از به دست آوردن خروجی برای فریم‌های اول و آخر ویدیو، به سراغ فریم‌های بعدی یعنی $t = 2$ و $t = T - 1$ می‌رویم. برای تخمین ژست فریم‌های غیر از فریم اول و آخر توسط مدل‌های رو به جلو و رو به عقب، از شبکه $deploy2$ که در شکل ۳-۸ نمایش داده شده است، استفاده می‌کنیم.

در هنگام تخمین ژست برای فریم‌های $t = 2$ و $t = T - 1$، نقشه اطمینان تخمین زده شده برای فریم‌های $t = 1$ و $t = T$ به عنوان ورودی به شبکه ارسال می‌شود و نقشه‌ی اطمینان متناظر با فریم‌های ورودی تولید می‌شود. به همین ترتیب با پیش‌روی در دنباله فریم‌ها و دریافت ژست تخمین زده شده برای فریم قبل و بعد به عنوان ورودی، کل ویدیو پیمایش می‌شود. همان‌طور که در شکل ۳-۸ نمایش داده شده است، در صورت استفاده از $deploy2$ با وزن‌های یادگرفته شده در شبکه رو به جلو، شبکه فریم t و نقشه اطمینان تخمین زده شده برای فریم قبل $h_{f(t-1)}$ را به عنوان ورودی دریافت کرده و h_{ft} را تولید می‌کند. در صورت استفاده از وزن‌های یادگرفته شده در شبکه رو به عقب، از نقشه اطمینان تخمین زده شده برای فریم بعد $h_{b(t+1)}$ استفاده شده و h_{bt} تولید می‌شود.

عملکرد این بخش نیز در (۳-۱۱) و (۳-۱۲) نمایش داده شده است. تمامی توابع به کار رفته در این بخش نیز مشابه (۳-۲) و (۳-۴) است.

$$h_{ft} = g_f\left(\tilde{L}_f\left(F_f(X_t) \oplus h_{f(t-1)}\right)\right),\ t = 2, 3, \dots, T. \quad (۱۱-۳)$$

$$h_{bt} = g_b\left(\tilde{L}_b\left(F_b(X_t) \oplus h_{b(t+1)}\right)\right),\ t = 2, 3, \dots, T. \quad (۱۲-۳)$$



شکل ۳-۸ ساختار شبکه deploy2

پس از پیمایش تمامی فریم‌های ویدیو و تخمین ژست‌های $\{h_{ft}, 1 \leq t \leq T\}$ توسط شبکه رو به جلو و ژست‌های $\{h_{bt}, 1 \leq t \leq T\}$ توسط شبکه رو به عقب، زمان تخمین ژست‌های نهایی فرا می‌رسد. در این مرحله دو تخمین به دست آمده از شبکه‌های رو به جلو و رو به عقب را در دست داریم. در این مرحله از شبکه $deply3$ استفاده می‌کنیم که دارای ساختاری مشابه با $ConvNet4$ نمایش داده شده در شکل ۳-۶ است. با ارسال تخمین‌های به دست آمده به شبکه $ConvNet4$، این شبکه به تولید نقشه اطمینان نهایی می‌پردازد.

۳-۷- جمع‌بندی

در این فصل به بررسی روش پیشنهادی برای تخمین ژست بدن انسان پرداختیم. شبکه طراحی شده برای تخمین ژست بدن انسان از بخش‌های شبکه رو به جلو، شبکه رو به عقب و شبکه ترکیب تشکیل شده است. در بخش‌های مجزا ساختار این شبکه‌ها مورد بررسی قرار گرفت. پس از آموزش شبکه‌های طراحی شده، شبکه‌های رو به جلو و رو به عقب ژست بدن انسان را در فریم ورودی تخمین می‌زنند. پس از تخمین ژست با استفاده از شبکه‌های رو به جلو و رو به عقب، تخمین نهایی با استفاده از شبکه ترکیب تولید می‌شود.

۴

فصل چهارم

نتایج و ارزیابی

پس از معرفی روش پیشنهادی در فصل سوم، به ارزیابی و بررسی عملکرد روش پیشنهادی می‌پردازیم. ابتدا به معرفی مجموعه داده‌های مورد استفاده می‌پردازیم. برای استفاده از مجموعه داده مورد استفاده در آموزش و آزمایش از پیش‌پردازش‌هایی استفاده می‌شود، که به معرفی پیش‌پردازش‌ها می‌پردازیم. هم‌چنین در ادامه معیارهای ارزیابی، نحوه تنظیم پارامترها، ارزیابی مدل پیشنهادی و مقایسه با سایر روش‌ها بررسی شده است.

۴-۱- معرفی مجموعه داده

در این پژوهش از دو مجموعه داده Penn Action[6] و Sub-JHMDB[7] استفاده شده است. در ادامه به معرفی این دو می‌پردازیم.

مجموعه داده Penn Action [6] یک مجموعه داده ویدیویی شامل ۲۳۲۶ ویدیو در حوزه ورزشی 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 هر عضو تعریف شده است، که این برچسب مشخص کننده دیده شدن یا انسداد عضو مربوط است.

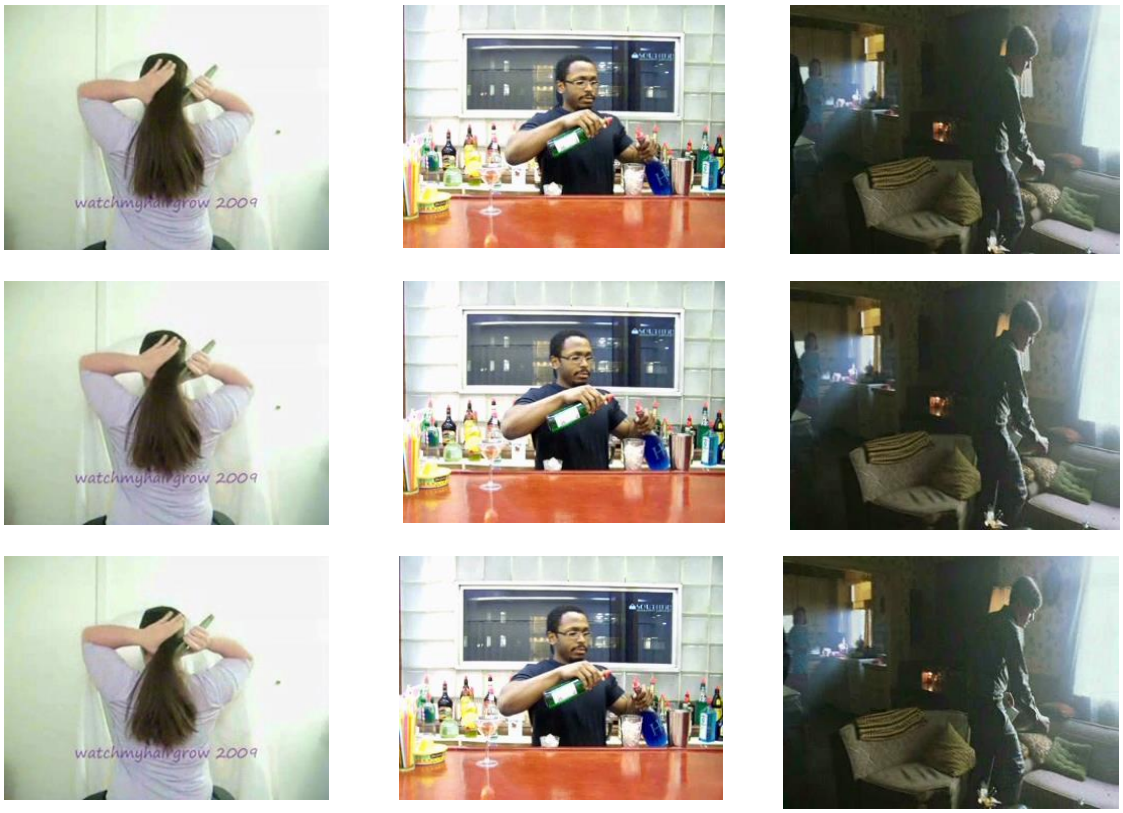
نمونه‌هایی از این مجموعه داده در شکل ۴-۱ نمایش داده شده است. نمونه‌ها شامل ۳ فریم متوالی از ۳ ویدیو هستند. همان‌طور که دیده می‌شود، بر خلاف مجموعه داده‌های تولید شده در آزمایشگاه، این مجموعه داده بدون هیچ محدودیتی جمع آوری شده است. داده‌های موجود در محیط‌های مختلف، با زاویه‌های دید متفاوت و نورپردازی‌های متغیر جمع‌آوری شده است.

شکل ۴-۱ نمونه‌هایی از مجموعه داده Penn Action

مجموعه داده Sub-JHMDB [7] مجموعه داده JHMDB یک مجموعه داده ویدیویی برای تخمین ژست بدن انسان در ویدیو است. این مجموعه داده دارای ویدیوهای مربوط به فعالیت‌های مختلف همانند شانه کردن، ریختن، برداشتن است که نمونه‌هایی از داده‌ها در شکل ۴-۲ نمایش داده شده است، که ۳ فریم متوالی از ۳ ویدیو از مجموعه داده نمایش داده شده است.

برای مقایسه نتایج به دست آمده بر روی این مجموعه داده با سایر روش‌های موجود در تخمین ژست بدن انسان، تنها از بخشی از این 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 زیرمجموعه از سه بخش مختلف تشکیل شده 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 به دست آمده بر روی سه زیر مجموعه به دست می‌آید.



شکل ۴-۲ نمونه‌هایی از مجموعه داده JHMDB

در فرایند پیش‌پردازش ویدیوهای ورودی، از تکنیک افزودگی داده[۳۵] استفاده می‌شود. استفاده از افزودگی داده، باعث بالا رفتن تنوع مجموعه داده ورودی می‌شود. از این رو تا حدودی از بیش‌برازش نیز جلوگیری خواهد شد. از آنجایی که فریم‌های ویدیوی ورودی به صورت همزمان به شبکه ارسال می‌شوند، پیش‌پردازش‌های انجام شده بر روی همه فریم‌هایی که به صورت همزمان به شبکه ارسال می‌شوند، یکسان خواهد بود. به عنوان افزودگی داده از عملیات مقیاس کردن[۳۶]، چرخاندن[۳۷]، بریدن[۳۸] و وارونه کردن[۳۹] تصویر استفاده می‌شود.

[35] Data Augmentation

[36] Scale

[37] Rotate

[38] Crop

[39] Flip

برای مقیاس کردن تصاویر، یک عدد تصادفی انتخاب شده و با اعمال آن بر روی تصویر، تصویر مقیاس شده به دست می‌آید. برای مجموعه داده Penn Action مقدار تصادفی در بازه ۰,۸ تا ۱,۴ انتخاب می‌شود. برای مجموعه داده sub-JHMDB به دلیل کوچکتر بودن اندازه‌ی بدن از عدد بزرگتری برای مقیاس کردن انتخاب شده و عدد تصادفی در بازه ۱,۲ تا ۱,۸ انتخاب می‌شود. برای چرخش تصویر، زاویه‌ای بین ۴۰− درجه تا ۴۰ درجه به صورت تصادفی انتخاب می‌شود. وارونه کردن یا نکردن تصویر نیز به صورت تصادفی مشخص می‌شود. در مرحله بریدن، خروجی با اندازه ثابت ۳۶۸ × ۳۶۸ تولید می‌شود. انجام برش به نحوی صورت می‌گیرد که بدن فرد در مرکز تصویر قرار بگیرد.

۴-۲- معیارهای ارزیابی

برای ارزیابی تخمین‌های حاصل در هر فریم از ویدیو از معیار ارزیابی Percentage Correct Keypoint (PCK) 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.

۴-۴- تنظیم پارامترها

با توجه به اینکه در بخشی از مدل پایه از [42] استفاده شده است، برای مقداردهی اولیه شبکه‌ی طراحی شده از وزن‌های این شبکه استفاده کردیم. البته همان‌طور که در بخش روش پیشنهادی بررسی شد، مدل از دو بخش رو به جلو و رو به عقب تشکیل شده است، که دارای ساختار مشابه هستند. ولی در بین این مدل اشتراک وزنی وجود ندارد. در نتیجه هر دو مدل رو به جلو و رو به عقب به صورت جداگانه با استفاده از وزن‌های حاصل از [42] مقدار دهی اولیه می‌شوند. شبکه‌ی ترکیب که توسط خودمان برای ترکیب نقشه‌های اطمینان دریافتی و تولید نقشه اطمینان واحد خروجی طراحی شده است، به صورت تصادفی وزن دهی اولیه می‌شود.

همان‌طور که در بخش‌های مدل رو به جلو و رو به عقب توضیح دادیم، شبکه در هنگام آموزش دنباله‌ای از فریم‌ها با طول T را به عنوان ورودی دریافت می‌کند. برای کنترل اندازه مدل تولیدی و پیچیدگی مدل، تمامی فریم‌های ویدیوی ورودی به شبکه ارسال نمی‌شوند. در نتیجه مقدار T با طول ویدیوی ورودی برابر نخواهد بود. با بررسی‌های انجام شده، $T = ۵$ برای طول ویدیوی ورودی انتخاب می‌شود. در نتیجه هر ویدیویی که برای آموزش به شبکه ارسال می‌شود، تنها ۵ فریم از آن انتخاب می‌شود. بر اساس بررسی‌های انجام شده، تعداد ۵ فریم مقدار مناسبی برای استخراج ویژگی‌های حرکتی از ویدیو است. حال در هنگام آموزش برای وجود تنوع در داده‌ها، از نقطه‌ای تصادفی در ویدیوی ورودی شروع کرده و T فریم انتخاب کرده و به مدل‌ها ارسال می‌کنیم.

نرخ یادگیری اولیه در آموزش برابر با ۸×۱۰^{-۵} انتخاب شد. برای یافتن بهینه در فرایند آموزش از الگوریتم گرادیان نزولی تصادفی[40] با میزان مومنتوم[41] ۰/۹ و کاهش وزن[42] ۵×۱۰^{-۵} استفاده کردیم. اندازه دسته‌های استفاده شده در آموزش نیز برابر با ۲ است و آموزش به اندازه ۳۰۰۰۰ قدم پیش رفته و متوقف می‌شود.

این مدل با استفاده از کتابخانه Caffe[44] و زبان برنامه نویسی متلب پیاده سازی شده است و با استفاده از Geforce GTX 1080 Ti NVIDIA GPU اجرا شده است.

۵-۴- نتایج آزمایش‌ها

در این بخش عملکرد قسمت‌های مختلف روش پیشنهادی بر روی مجموعه داده‌های Penn Action و Sub-JHMDB مورد ارزیابی قرار می‌گیرد. سپس عملکرد مدل به دست آمده با سایر روش‌های موجود مقایسه می‌شود.

[40] Stochastic Gradient Descent

[41] momentum

[42] Weight decay

۴-۵-۱- ارزیابی مدل ترکیب بر روی مجموعه داده Penn Action

در ارزیابی مدل ابتدا عملکرد مدل‌های رو به جلو و رو به عقب به صورت جداگانه بررسی می‌شود. پس از بررسی عملکرد جداگانه مدل‌های رو به جلو و رو به عقب، عملکرد مدل ترکیبی در مقایسه با این دو بررسی می‌شود.

مدل رو به جلو در ۳-۱- معرفی شد و ساختار این مدل در شکل ۳-۲ نمایش داده شده است. همان‌طور که در ۴-۴- گفتیم، وزن‌دهی اولیه مدل رو به جلو با استفاده از وزن‌های به دست آمده از مدل آموزش دیده شده در [42] انجام می‌شود. دقت به دست آمده از آزمایش این مدل در جدول ۴-۱ نمایش داده شده است. در این جدول دقت جداگانه بر روی اعضای سر، شانه، بازو، مچ، ران، زانو و مچ پا نیز گزارش شده است.

جدول ۴-۱ دقت به دست آمده برای مدل رو به جلو بر روی مجموعه داده Penn Action

روش	Head	Sho	Elb	Wri	Hip	Knee	Ank	Mean
روبه‌جلو	۹۸/۹	۹۸/۵	۹۶/۶	۹۶/۵	۹۸/۲	۹۸/۲	۹۷/۵	۹۷/۷

نمونه‌هایی از ژست‌های تخمین زده شده توسط این مدل در شکل ۴-۳ نمایش داده شده است. در این شکل، ۳ فریم متوالی از سه ویدیوی مختلف نمایش داده شده است.

همان‌طور که در خروجی‌های حاصل در شکل ۴-۳ دیده می‌شود، مکان ۱۳ عضو بدن توسط مدل تخمین زده شده است. سپس ارتباط‌های سینماتیکی بین این اعضا که یال‌های مدل گرافی هستند نیز رسم شده است. هر کدام از یال‌های موجود با رنگ منحصر به فردی نمایش داده شده است.

شکل ۴-۳ نمونه‌هایی از تخمین ژست‌های به دست آمده از مدل رو به جلو بر روی مجموعه داده Penn Action

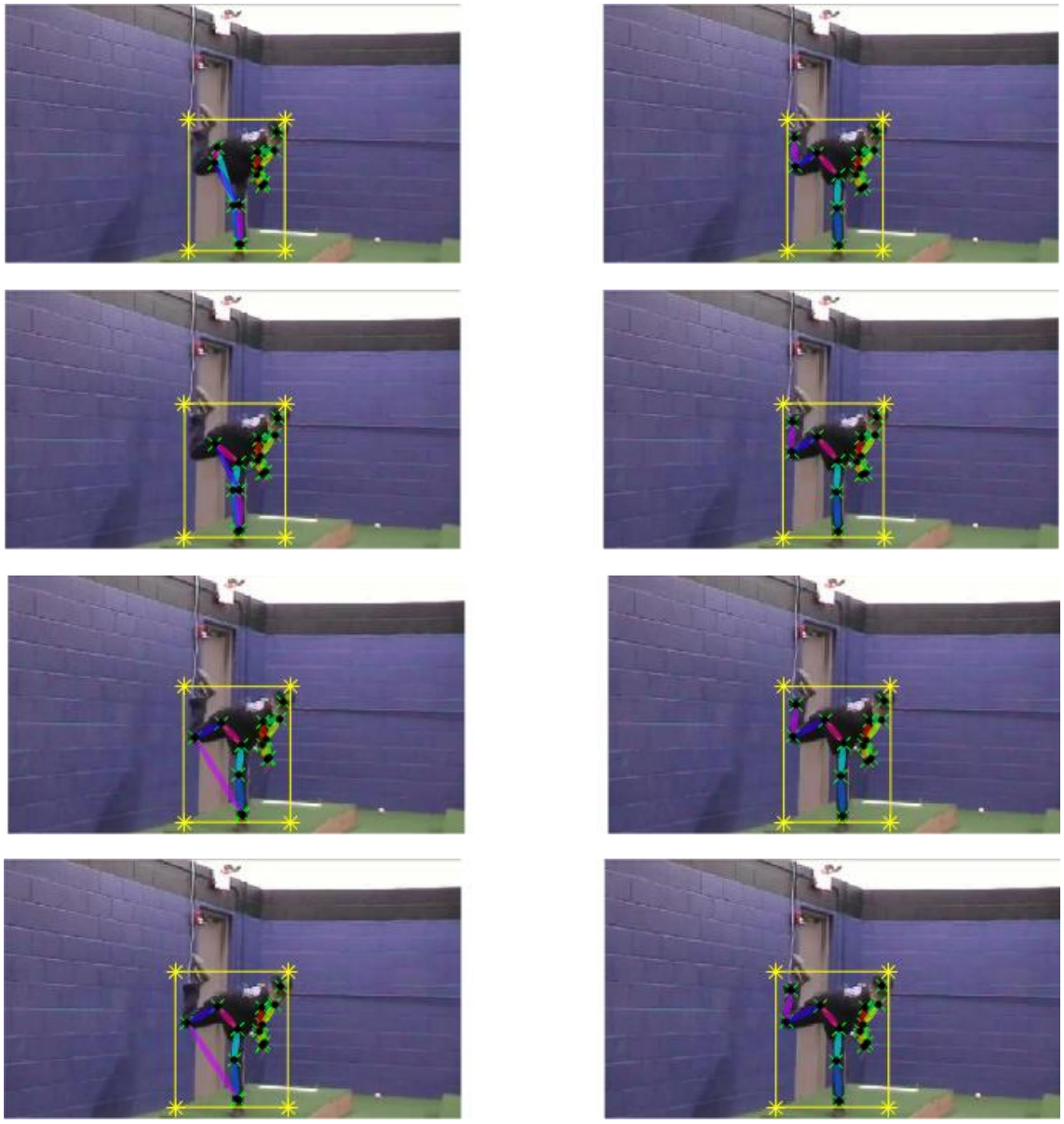
به طور مشابه نتایج مدل رو به عقب نیز در جدول ۴-۲ نمایش داده شده است.

جدول ۴-۲ دقت به دست آمده برای مدل رو به عقب بر روی مجموعه داده Penn Action

روش	Head	Sho	Elb	Wri	Hip	Knee	Ank	Mean
روبه‌عقب	۹۸/۹	۹۸/۵	۹۵/۹۵	۹۶	۹۸/۲۵	۹۷/۹۵	۹۶/۹۵	۹۷/۴۳

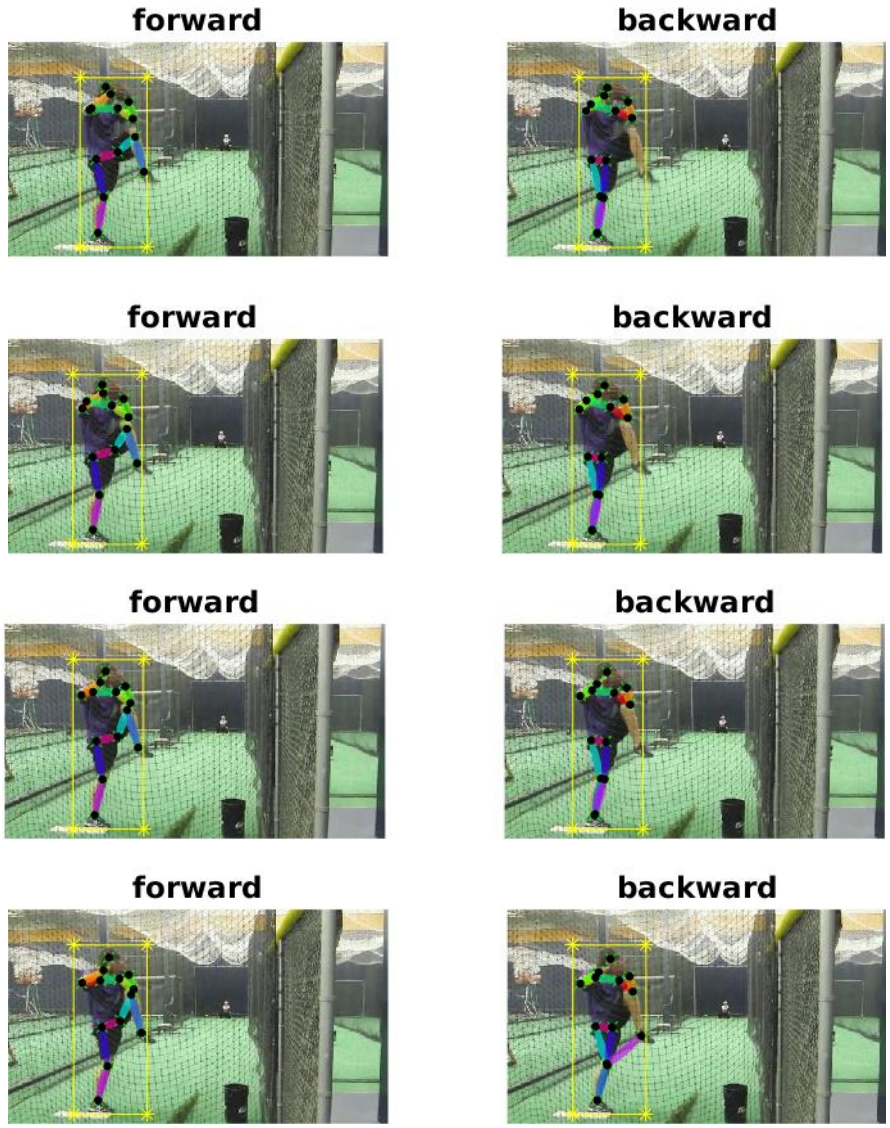
با داشتن دو مدل رو به جلو و رو به عقب، نتایج این دو مدل به دست می‌آید. پس از استخراج تخمین‌های به دست آمده برای مدل رو به جلو و رو به عقب، فریم‌هایی که دارای تخمین‌های رو به جلو و رو به عقب متفاوتی هستند بررسی می‌شوند. برای مثال در شکل ۴-۴ مدل رو به جلو نتوانسته است تخمین درستی ارائه دهد. در مجموعه فریم‌های نمایش داده شده، پای فرد در حال حرکت است. در حرکت رو به جلو، تاری و انسداد مدل رو به جلو دچار مشکل شده و مکان نادرستی برای پا تخمین می‌زند. در صورتی که

مدل رو به عقب با پیمایش فریم‌های ورودی در جهت برعکس، حرکت پا را از جهت برعکس دریافت کرده و می‌تواند با وجود تاری و یا انسداد تخمین درستی ارائه دهد.



شکل ۴-۴ نمونه‌هایی از تخمین‌های به دست آمده از مدل‌های رو به جلو و رو به عقب – عملکرد درست مدل رو به عقب در مقابل عملکرد غلط مدل رو به جلو

حال به طور مشابه در ادامه در شکل ۴-۵، نمونه‌ای از فریم‌هایی که مدل رو به جلو موفق به تخمین ژست صحیح شده است اما مدل رو به عقب شکست خورده است، نمایش داده می‌شود.



شکل ۴-۵ نمونه‌ای از تخمین به دست آمده از مدل‌های رو به جلو و رو به عقب – عملکرد درست مدل رو به جلو در مقابل عملکرد غلط مدل رو به عقب

همان‌طور که در دو مثال ذکر شده نیز دیدیم، در برخی از فریم‌ها مدل رو به جلو عملکرد درستی از خود نشان داده و مدل رو به عقب دچار خطا شده است. همچنین در برخی موارد نیز با وجود شکست مدل رو به جلو، مدل رو به عقب قادر به حصول تخمین درست شده است. حال به دنبال به دست آوردن مدل ترکیبی برای رفع خطاهای دو مدل رو به جلو و رو به عقب هستیم. مدل ترکیب در صورتی می‌تواند به بهبود نتایج کمک کند که، مدل‌های رو به جلو و رو به عقب بیشتر دارای خطا در موقعیت‌های متفاوتی باشند. از این رو به بررسی این مورد در جدول ۴-۳ می‌پردازیم. در جدول ۴-۳ تعداد ویدیوهای دارای اشتباه‌های غیرهمزمان مدل‌های رو به جلو و رو به عقب با توجه به معیار درستی‌های مختلف ذکر شده است. معیار درستی را برابر با تعداد عضوهای دارای تخمین درست در نظر می‌گیریم. برای مثال در حالتی که این مقدار را ۹ در نظر بگیریم، یک فریم دارای تخمین درست خواهد بود، در صورتی که ۹ عضو در آن فریم دارای تخمین درستی باشند. حال با معیارهای درستی متفاوت، به شمارش تعداد ویدیوهایی می‌پردازیم که تعداد اشتباه‌های غیر همزمان مدل‌های رو به جلو و رو به عقب از تعداد اشتباه‌های همزمان این دو مدل بیشتر باشد.

جدول ۴-۳ میزان خطاهای غیرهمزمان مدل‌های رو به جلو و رو به عقب

تعداد ویدیوهای دارای اشتباه‌های غیرهمزمان مدل‌های رو به جلو و رو به عقب	معیار درستی فریم
۸۵۵	۹
۷۷۲	۱۰
۶۷۰	۱۱

همان‌طور که در جدول ۴-۳ مشاهده می‌شود، با معیارهای درستی انتخاب شده تعداد ویدیوهایی که اشتباه‌های غیرهمزمان بیشتری دارند، اکثریت مجموعه داده تست را پوشش می‌دهند. از این رو به دنبال استفاده از مدل ترکیبی برای پیش‌بینی ژست پیش می‌رویم. در جدول ۴-۴ نتیجه‌ی به دست آمده از مدل ترکیب نمایش داده شده است.

جدول ۴-۴ مقایسه نتایج به دست آمده در مدل‌های رو به جلو، رو به عقب و مدل ترکیب پیشنهادی بر روی مجموعه داده Penn Action

روش	Head	Sho	Elb	Wri	Hip	Knee	Ank	Mean
روبه‌جلو	۹۸/۹	۹۸/۵	۹۶/۶	۹۶/۵	۹۸/۲	۹۸/۲	۹۷/۵	۹۷/۷
روبه‌عقب	۹۸/۹	۹۸/۵	۹۵/۹۵	۹۶	۹۸/۲۵	۹۷/۹۵	۹۶/۹۵	۹۷/۴۳
ترکیب	**۹۸/۹**	**۹۸/۵**	**۹۷**	**۹۷**	**۹۸/۴**	**۹۸/۵**	**۹۷/۷**	**۹۷/۹۴**

عضوهایی همانند سر و شانه که دارای ساختار و ظاهر تقریباً مشابهی در شرایط متفاوت هستند، دارای دقت بالایی در مدل‌های رو به جلو و رو به عقب و یکسان در مدل ترکیب هستند. اما در بقیه‌ی عضوها همانند بازو، مچ و ... که دارای جهت‌گیری‌های مختلف و شکل‌های ظاهری بسیار متفاوتی در ژست‌های مختلف هستند، مدل ترکیب توانسته است به خوبی دقت تخمین را بهبود دهد.

با توجه به اینکه مدل ترکیب توانسته است نسبت به مدل رو به عقب و رو به جلو که مدل‌های پایه‌ی تشکیل دهنده‌ی مدل ترکیب هستند، نتیجه‌ی بهتری ارائه دهد ادعا می‌کنیم که مدل ترکیب ارائه شده موفقیت آمیز عمل کرده است.

۴-۵-۲- مقایسه عملکرد مدل به دست آمده بر روی مجموعه داده Penn Action

پس از مشاهده بهبود به دست آمده در مدل ترکیب نسبت به مدل رو به جلو و رو به عقب، نتیجه به دست آمده از مدل ترکیب با روش‌های مختلف موجود در تخمین ژست در جدول ۴-۵ مقایسه می‌شود.

روش‌های [45]، [46]،[47] و [48] دارای دقت به نسبت پایین‌تری بوده و قدرت رقابت با مدل پیشنهادی را ندارند. نکته جالب برتری نتایج روش [30] که روشی ارائه شده برای تخمین ژست بدن انسان در تصویر است، بر روش [41] است. همان‌طور که در جدول ۴-۵ دیده می‌شود، روش پیشنهادی دقت بالاتری نسبت به روش‌های پیشین ارائه داده است. برای بررسی چگونگی حصول به دقت بالاتر، دقت به دست آمده برای هرکدام از عضوها که در جدول ۴-۶ گزارش شده است، را بررسی می‌کنیم. روش پیشنهادی توانسته است با بهبود تخمین عضوهای دارای تغییر زیاد، تخمین ژست بهتری نسبت به روش‌های پیشین ارائه دهد.

جدول ۴-۵ مقایسه‌ی دقت به دست آمده در روش‌های تخمین ژست بدن انسان بر روی مجموعه داده Penn Action

روش	دقت
[45]	۴۵/۳
[46]	۴۸/۰
[47]	۸۱/۱
[48]	۹۱/۸
[41]	۹۶/۵
[30]	۹۷/۱
[42]	۹۷/۷
روش ارائه شده	**۹۷/۹۴**

جدول ۴-۶ مقایسه نتایج به دست آمده در تخمین اعضای بدن انسان بر روی مجموعه داده Penn Action با پارامتر $\alpha = 0.2$

روش	Head	Sho	Elb	Wri	Hip	Knee	Ank
[45]	۶۲/۸	۵۲/۰	۳۲/۳	۲۳/۳	۵۳/۳	۵۰/۲	۴۳/۰
[46]	۶۴/۲	۵۵/۴	۳۳/۸	۲۴/۴	۵۶/۴	۵۴/۱	۴۸/۰
[47]	۸۹/۱	۸۶/۴	۷۳/۹	۷۳/۰	۸۵/۳	۷۹/۹	۸۰/۳
[48]	۹۵/۶	۹۳/۸	۹۰/۴	۹۰/۷	۹۱/۸	۹۰/۸	۹۱/۵
[41]	۹۸/۰	۹۷/۳	۹۵/۱	۹۴/۷	۹۷/۱	۹۷/۱	۹۶/۹
[30]	۹۸/۶	۹۷/۹	۹۵/۹	۹۵/۸	۹۸/۱	۹۷/۳	۹۶/۶
[42]	۹۸/۹	**۹۸/۶**	۹۶/۶	۹۶/۶	۹۸/۲	۹۸/۲	۹۷/۵
ارائه شده	**۹۸/۹**	۹۸/۵۵	**۹۷/۰**	**۹۷/۰**	**۹۸/۴**	**۹۸/۵**	**۹۷/۷**

با توجه به مقایسه‌ی نتایج به دست آمده از مدل پیشنهادی و روش‌های پیشین، مدل ارائه شده توانسته است بر روی مجموعه داده Penn Action عملکرد بهتری ارائه دهد.

۴-۵-۳- ارزیابی مدل ترکیب بر روی مجموعه داده Sub-JHMDB

در این بخش نیز همانند بخش ۴-۵-۱- عملکرد مدل ترکیب با توجه به مدل‌های پایه‌ی استفاده شده بررسی می‌شود.

برای بررسی عملکرد روش پیشنهادی بر روی مجموعه داده JHMDB، ابتدا عملکرد این روش با بخش‌های تشکیل دهنده روش پیشنهادی مقایسه شده است. همان‌طور که در جدول ۴-۷ مشاهده می‌شود، روش پیشنهادی توانسته است دقت بالاتری نسبت به مدل رو به جلو و مدل رو به عقب کسب کند. در نتیجه می‌توان ادعا کرد که روش پیشنهادی از نقاط قوت مدل‌های رو به جلو و رو به عقب استفاده کرده و توانسته است با تصحیح خطاهای موجود نتیجه نهایی را بهبود دهد.

جدول ۴-۷ مقایسه نتایج به دست آمده بر روی مدل‌های رو به جلو، رو به عقب و روش پیشنهادی بر روی مجموعه داده JHMDB

روش	دقت
شبکه رو به جلو	۹۳/۶
شبکه رو به عقب	۹۳/۵
روش پیشنهادی	**۹۴/۰**

همچنین علاوه بر مقایسه نتیجه به دست آمده بر روی مدل پیشنهادی و مدل‌های رو به جلو و رو به عقب، نیاز به مقایسه روش پیشنهادی با سایر روش‌های موجود است.

همان‌طور که در جدول ۴-۸ مدل ترکیبی ارائه شده بر روی مجموعه داده Sub-JHMDB نیز نسبت به روش‌های پیشین عملکرد بهتری نمایش داده است.

جدول ۴-۸ مقایسه‌ی دقت به دست آمده در روش‌های تخمین ژست بدن انسان بر روی مجموعه داده Sub-JHMDB

دقت	روش
۵۲/۵	[45]
۵۵/۷	[46]
۷۳/۸	[47]
۹۲/۱	[41]
۹۱/۹	[30]
۹۲/۲	[42] بدون LSTM
۹۳/۶	[42]
۹۴/۰	روش ارائه شده

با توجه به مقایسه‌ی نتایج به دست آمده از مدل پیشنهادی و روش‌های پیشین، مدل ارائه شده توانسته است بر روی مجموعه داده JHMDB نیز عملکرد بهتری ارائه دهد.

۴-۶- جمع‌بندی

در این بخش ابتدا به معرفی مجموعه داده، پیش‌پردازش‌های موردنیاز برای تبدیل داده‌های موجود به شکل مناسب برای شبکه طراحی شده پرداختیم. در ادامه معیار ارزیابی مورد استفاده برای محاسبه دقت و چگونگی تنظیم پارامترها توضیح داده شد.

سپس نتایج به دست آمده بر روی مجموعه داده‌های استفاده شده گزارش شد. همان‌طور که در نتایج مشاهده شد، شبکه طراحی شده با استفاده از اطلاعات فریم‌های قبل و بعد، به خوبی توانسته است که دقت تخمین ژست بدن انسان را نسبت به مدل‌های پایه تشکیل دهنده (مدل رو به جلو و رو به عقب) بهبود دهد. همچنین با مقایسه روش ارائه شده با سایر روش‌های موجود در تخمین ژست بدن انسان در ویدیو، برتری روش ارائه شده مشاهده شد.

۵

فصل پنجم

جمع‌بندی و نتیجه‌گیری

در این مقاله، مسئله تخمین ژست بدن انسان مورد بررسی قرار گرفت. ابتدا روش‌های موجود در تخمین ژست بدن انسان در تصویر و سپس روش‌های موجود برای تخمین ژست بدن انسان در ویدیو توضیح داده شد.

در تخمین ژست بدن انسان در ویدیو اطلاعات زمانی موجود در بین فریم‌ها و لزوم سازگاری زمانی بین تخمین‌های فریم‌ها، اطلاعات اضافی در اختیار مدل تخمین ژست قرار می‌دهد. با توجه به تأثیر مثبت اطلاعات فریم‌های بعد در تخمین ژست فریم موردنظر، علاوه بر تأثیر مثبت فریم‌های قبل در این تخمین، مدلی برای استفاده از اطلاعات تمامی فریم‌های قبل و بعد فریم موردنظر طراحی کردیم.

در مدل طراحی شده ابتدا T فریم از فریم‌های ورودی به صورت پیمایش رو به جلو و پیمایش رو به عقب به مدل‌های رو به جلو و رو به عقب ارسال می‌شوند. هرکدام از مدل‌های رو به جلو و رو به عقب تخمینی از ژست دنباله فریم‌های ورودی تولید می‌کنند. پس از پیمایش فریم‌ها و تولید دو مجموعه نقشه اطمینان برای هرکدام از فریم‌ها، از شبکه ترکیب برای تولید تخمین نهایی استفاده می‌کنیم. شبکه‌ی ترکیب طراحی شده با دریافت دو مجموعه نقشه اطمینان، نقشه اطمینان نهایی را تولید می‌کند. با توجه به آزمایش‌های انجام شده بر روی مجموعه داده‌های Penn Action و Sub-JHMDB توانایی مدل در تخمین بهتر نقشه‌های اطمینان نشان داده شد.

در ادامه برای بهبود بیشتر نتایج تخمین ژست بدن انسان در ویدیو قصد داریم تا از نقشه‌های اطمینان تولید شده توسط شبکه‌های رو به جلو و رو به عقب برای تولید مجموعه‌ای از ژست‌های کاندید استفاده کنیم. پس از تولید مجموعه‌ی ژست‌های کاندید، با استفاده از تابع امتیاز طراحی شده ژست بهینه انتخاب خواهد بود. با این کار عوامل متعدد مؤثر که در مدل طراحی شده لحاظ نشده‌اند، در تابع امتیاز دخیل شده و به تولید تخمین نهایی بهتر تولید می‌کنند.

۶ منابع و مراجع

[1] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: the body parts parsing based methods," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 10–19, 2015.

[2] D. Zhang and M. Shah, "Human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2012–2020.

[3] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[4] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, 2013.

[5] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 468–475.

[6] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.

[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[8] C. Science and D. Singh, "Human Pose Estimation : Extension and Application," no. September, 2016.

[9] T. Pfister, "Advancing Human Pose and Gesture Recognition," no. April, 2015.

[10] K. Rohr, "Towards model-based recognition of human movements in image sequences," *CVGIP-Image Underst.*, vol. 59, no. 1, pp. 94–115, 1994.

[11] N. Hogg, "Human monocytes have prothrombin cleaving activity.," *Clin. Exp. Immunol.*, vol. 53, no. 3, p. 725, 1983.

[12] D. A. Forsyth and M. M. Fleck, "Body plans," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 678–683.

[13] J. O'rourke and N. I. Badler, "Model-based image analysis of human motion

using constraint propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 522–536, 1980.

[14] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, 1998, pp. 8–15.

[15] X. Ren, A. C. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2005, vol. 1, pp. 824–831.

[16] G. Hua, M.-H. Yang, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 2, pp. 747–754.

[17] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Comput.*, vol. 100, no. 1, pp. 67–92, 1973.

[18] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1365–1372.

[19] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in neural information processing systems*, 2014, pp. 1736–1744.

[20] D. Tran and D. Forsyth, "Improved human parsing with a full relational model," in *European Conference on Computer Vision*, 2010, pp. 227–240.

[21] T.-P. Tian and S. Sclaroff, "Fast globally optimal 2d human detection with loopy graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 81–88.

[22] M. Wainwright, T. Jaakkola, and A. Willsky, "MAP estimation via agreement on (hyper) trees: Message-passing and linear programming approaches," in *Proceedings of the annual allerton conference on communication control and computing*, 2002, vol. 40, no. 3, pp. 1565–1575.

[23] L. Fu, J. Zhang, and K. Huang, "Beyond tree structure models: A new occlusion aware graphical model for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, no. c, pp. 1976–1984.

[24] D. Koller, N. Friedman, and F. Bach, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[28] L. Pishchulin *et al.*, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[29] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, 2016, pp. 34–50.

[30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[31] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*, 2014, pp. 33–47.

[32] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

[33] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.

[34] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *European Conference on Computer Vision*, 2016, pp. 246–260.

[35] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1290–1299.

[36] L. Zhao, X. Gao, D. Tao, and X. Li, "Tracking human pose using max-margin markov models," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5274–5287, 2015.

[37] F. Zhou and F. la Torre, "Spatio-Temporal Matching for Human Pose Estimation in Video.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1492–1504, 2016.

[38] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3063–3072.

[39] L. Zhao, X. Gao, D. Tao, and X. Li, “Learning a tracking and estimation integrated graphical model for human pose tracking,” *IEEE Trans. neural networks Learn. Syst.*, vol. 26, no. 12, pp. 3176–3186, 2015.

[40] T. Pfister, J. Charles, and A. Zisserman, “Flowing ConvNets for Human Pose Estimation in Videos,” 1913.

[41] J. Song and L. Wang, “Thin-Slicing Network : A Deep Structured Model for Pose Estimation in Videos.”

[42] Y. Luo *et al.*, “Lstm pose machines,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5207–5215.

[43] K. Greff, R. K. Srivastava, J. Koutn\’\ik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Trans. neural networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017.

[44] Y. Jia *et al.*, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.

[45] D. Park and D. Ramanan, “N-best maximal decoders for part models,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2627–2634.

[46] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, “Joint action recognition and pose estimation from video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.

[47] U. Iqbal, M. Garbade, and J. Gall, “Pose for action-action for pose,” in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, 2017, pp. 438–445.

[48] G. Gkioxari, A. Toshev, and N. Jaitly, “Chained predictions using convolutional neural networks,” in *European Conference on Computer Vision*, 2016, pp. 728–743.

Amirkabir University of Technology
(Tehran Polytechnic)

Computer Engineering Department

MSc Thesis

Human Pose Estimation in Video

By
Mina Ghadimi Atigh

Supervisor
Dr. Ahmad Nickabadi

Feburary 2019