تخمین حالت سه بعدی انسان از تک تصویر با استفاده از شبکه‌ی عصبی غیرخطی کانولوشنی مبتنی بر اطلاعات شکل

فرانک شمسافر[1] و حسین ابراهیم‌نژاد[2]

[1] دانشجوی دکتری، دانشکده‌ی مهندسی برق، دانشگاه صنعتی سهند تبریز، f_shamsafar@sut.ac.ir

[2] **نویسنده‌ی مسئول**، استاد، دانشکده‌ی مهندسی برق، دانشگاه صنعتی سهند تبریز، ebrahimnezhad@sut.ac.ir

(تاریخ دریافت مقاله: ۱۳۹۶/۱۱/۱۶ تاریخ پذیرش مقاله: ۱۳۹۷/۵/۱۷)

دانشگاه صنعتی سهند

نشریه سامانه‌های غیرخطی در مهندسی برق

دوره ۵- شماره ۱

بهار و تابستان ۹۷

صفحه ۸۳ الی ۱۰۳

ISSN: 2322-3146

http://jnsee.sut.ac.ir

چکیده

تخمین حالت سه بعدی انسان یکی از مسائل پراهمیت و پرکاربرد در بینایی ماشین است. تخمین حالت انسان، از اسکلت دو بعدی و با استفاده از اطلاعات چندگانه آغاز شده و طی یک روند تکاملی، به تخمین اسکلت سه بعدی با بهره‌گیری از حداقل اطلاعات ورودی سوق یافته است. در این مقاله، مسئله تخمین حالت سه بعدی انسان با استفاده از اطلاعات یک تصویر رنگی بررسی شده است. روش پیشنهادی جزء آن دسته از روش‌هایی محسوب می‌شود که ابتدا حالت دو بعدی را استخراج و سپس، با ارتقاء حالت دو بعدی تخمینی به فضای سه بعدی، حالت سه بعدی پیش‌بینی می‌شود. به دلیل آن‌که در این نوع روش‌ها، منشاء بیشتر خطاها ناشی از تخمین نادرست حالت دو بعدی است، در این مقاله، با ارائه روشی برای تخمین بهتری از حالت دو بعدی، حالت سه بعدی دقیق‌تری به دست آمده است. روش پیشنهادی برای تخمین حالت دو بعدی، از یادگیری عمیق و اطلاعات نقشه لبه بهره برده است. به عبارتی دیگر، در این مقاله از ویژگی لبه که یک ویژگی طراحی شده می‌باشد، برای هدایت یادگیری شبکه عصبی عمیق و یادگیری ویژگی‌ها در راستای هدف تعیین شده، استفاده شده است. آزمایش‌های انجام شده نشان می‌دهد که روش پیشنهادی به خطای کمتری در تخمین حالت دو بعدی و متعاقباً، به خطای کمتری در پیش‌بینی حالت سه بعدی تصاویر پایگاه داده Human3.6M و HumanEva-I منجر شده است.

واژه های کلیدی

تخمین حالت انسان،

یادگیری عمیق،

شبکه‌های عصبی کانولوشنی،

نقشه‌ی لبه.

Sahand University of Technology

Journal of Nonlinear Systems in Electrical Engineering

Vol.5, No.1,

Spring and Winter 1397

ISSN: 2322 – 3146

http://jnsee.sut.ac.ir

Three-Dimensional Human Pose Estimation from a Single Image Using Nonlinear Convolutional Neural Network based on Shape Information

Faranak Shamsafar[1] and Hossein Ebrahimnezhad[2]

[1] PhD candidate, Faculty of Electrical Engineering, Sahand University of Technology, f_shamsafar@sut.ac.ir

[2] **Corresponding Author**, Professor, Faculty of Electrical Engineering, Sahand University of Technology, ebrahimnezhad@sut.ac.ir

ABSTRACT

Keywords

Human pose estimation,

Deep learning,

Convolutional neural networks,

Edge map.

3D human pose estimation is one of the most significant tasks in computer vision with wide range of applications. The work for estimating human pose initialized from 2D skeletal estimation from multiple data and has proceeded toward 3D skeletal estimation from minimum input information. In this paper, 3D human pose estimation from a single RGB image is investigated. The proposed work is considered as the ones which firstly estimate 2D pose and then lift the estimated 2D configuration to 3D space. Since most of the errors in this attitude are originated by inaccurate 2D pose inference, we have proposed a method for predicting better 2D poses to obtain more accurate 3D poses. The proposed approach for estimating 2D pose has leveraged deep learning along with the information of the edge map. In other words, we have made use of edge features, which are hand-designed features, in order to guide the deep neural network in training and in learning the features in accordance with the defined objective. Experimental results have demonstrated less errors in 2D and consequently 3D pose estimation in Human3.6M and HumanEva-I benchmarks.

۱- مقدمه

تخمین حالت انسان یکی از مسائل مهم در بینایی ماشین است و به معنای به دست آوردن اطلاعات توصیف‌کننده‌ی حالت اسکلتی انسان می‌باشد. این اطلاعات، معمولاً به صورت داده‌های دو بعدی یا سه بعدی و همچنین، براساس مفاصل یا اجزاء بدن انسان ارائه می‌شوند. اطلاعات ورودی به منظور دست‌یابی به چنین هدفی، اطلاعات دیداری هستند که می‌توان این اطلاعات را به انواع ۱- تصاویر چندمنظری، ۲- تصاویر رنگی به همراه تصاویر عمق، ۳- دنباله‌ای از تصاویر (اطلاعات زمانی و حرکتی) و ۴- تک تصویر رنگی دسته‌بندی کرد. استفاده از تک تصویر رنگی، یکی از پرکاربردترین و در عین حال، پرچالش‌ترین روش‌های تخمین حالت انسان است که در سال‌های اخیر بیشتر مورد توجه واقع شده است. سایر موارد بیان شده، علی‌رغم فراهم کردن اطلاعات بیشتری برای پردازش، در بسیاری از شرایط قابل به‌کارگیری نیستند.

تخمین حالت انسان در بسیاری از مسائل کاربرد دارد که از جمله آن‌ها می‌توان موارد زیر را نام برد: نظارت ویدئویی، تعامل بین انسان و کامپیوتر، سرگرمی‌های دیجیتالی (بازی‌های کامپیوتری، انیمیشن‌های کامپیوتری و فیلم)، پزشکی و مراقبت‌های سلامتی (مانند نظارت و مراقبت از بیماران، تحلیل آسیب‌های راه رفتن و بررسی رشد ذهنی کودکان)، تحلیل صحنه‌های ورزشی، حقیقت مجازی و کاربردهای نظامی. همچنین، تخمین حالت انسان در سایر مسائل بینایی ماشین می‌تواند مفید واقع گردد، مانند بازشناسی فعالیت انسان، تحلیل و فهم فعالیت انسان، ثبت حرکت انسان، بازیابی تصویر و ویدئو بر اساس حالت انسان، ردیابی انسان و تشخیص و کنترل حرکات دست و صورت.

با توجه به شرایط تهیه تصویر، افراد موجود در تصویر و فعالیت آن‌ها، چالش‌های متفاوتی در تخمین حالت انسان وجود دارد. چالش‌های مهم عبارتند از: ساختار پیچیده اسکلت انسان و در نتیجه، حالت‌های بسیار متفاوت اسکلتی، تغییر ظاهر انسان (تفاوت در شکل و تناسب بدن، رنگ، نوع و بافت لباس افراد)، انسداد جزئی یا کلی اجزاء بدن (خودانسدادی یا انسداد با شیء دیگر)، پس‌زمینه شلوغ، تغییر در زاویه دید، تغییر مقیاس، از بین رفتن اطلاعات بعد سوم (عمق) و ایجاد ابهام، بریدگی در مرزهای تصویر و کوتاه‌نمایی. یک الگوریتم تخمین حالت انسان بایستی تا حد ممکن نسبت به این تغییرات مقاوم باشد.

در دهه گذشته، تعداد زیادی از کارهای صورت گرفته در این زمینه، در راستای تخمین حالت دو بعدی بوده، اما در چند سال اخیر، توجه محققان به سمت تخمین اسکلت سه بعدی جلب شده است. در برخی از تحقیقات انجام شده، داده‌های ورودی به گونه‌ای هستند که می‌توان اطلاعات سه بعدی را به صورت بهتری مدیریت و پردازش کرد، مانند بهره‌گیری از اطلاعات چندمنظری، تصویر عمق و اطلاعات زمانی. این نوع تحقیقات، برای دستیابی به بهترین دقت‌ها همچنان ادامه دارند. این درحالی است که تخمین حالت سه بعدی از تک تصویر، مسئله‌ای است که به دلیل حداقل بودن اطلاعات در دسترس و تشدید شدن چالش‌های تخمین حالت، به توجه بیشتری نیاز دارد؛ زیرا، بیشتر اطلاعات دیداری به دست آمده از انسان به صورت تک تصویرهای دو بعدی هستند.

از یک نگاه، می‌توان کارهای انجام شده در زمینه تخمین حالت سه بعدی از تک تصویر را به دو دسته تقسیم کرد: روش‌هایی که از مدل‌های سه بعدی بهره گرفته‌اند (روش بالا به پایین[1]) و روش‌هایی که از تخمین دو بعدی به تخمین سه بعدی دست یافته‌اند (روش پایین به بالا[2]). روش‌های نوع اول معمولاً نیازمند یک مدل‌سازی صحیح از انسان هستند و با یک تابع هدف، خطای تطابق را کمینه می‌کنند. معمولاً این مدل‌ها نمی‌توانند حالت‌های پیچیده انسان را به خوبی پوشش دهند و همچنین، نیازمند پردازش حجم بزرگی از داده‌های سه بعدی هستند. روش‌های نوع دوم، پردازش راحت‌تری دارند، اما به دلیل این‌که حالت سه بعدی از حالت دو بعدی تعمیم داده می‌شود، به دست آوردن یک حالت صحیح دو بعدی از انسان، مرحله مهم و اساسی برای به دست آوردن تخمین بهتر و دقیق‌تر حالت سه بعدی است.

[1] Top-down
[2] Bottom-up

هدف از این مقاله، تخمین حالت سه بعدی انسان با استفاده از اطلاعات تک تصویر رنگی است که در آن از تعمیم حالت دو بعدی به حالت سه بعدی استفاده شده است. با توجه به اهمیت بسیار بالای تخمین حالت دو بعدی در این نوع روش‌ها، که به عنوان مرحله ابتدایی و بنیادی تلقی می‌شود، در این مقاله سعی شده است که با استفاده از ترکیب یادگیری عمیق و روش‌های سنتی استخراج ویژگی (نقشه‌ی لبه)، تخمین حالت دو بعدی انسان با دقت بالایی انجام شود. همان‌طور که می‌دانیم، ویژگی‌های یادگیری شده توسط یادگیری عمیق، در مقابل ویژگی‌های طراحی شده توسط روش‌های سنتی قرار دارند. بنابراین، می‌توان گفت روش پیشنهادی به منظور تخمین دقیق‌تر حالت دو بعدی، از ترکیب دو نوع ویژگی‌های طراحی شده و ویژگی‌های یادگیری شده بهره برده است. این امر، یعنی تخمین حالت دو بعدی دقیق‌تر، متعاقباً منجر به تخمین دقیق‌تر حالت سه بعدی می‌شود.

ساختار مقاله بدین صورت است: بخش ۲ نگاهی به کارهای مرتبط با مسئله تخمین حالت انسان دارد. در بخش ۳، شبکه‌های عصبی کانولوشنی غیرخطی به عنوان یکی از مهم‌ترین روش‌های یادگیری عمیق توصیف شده است. بخش ۴ روش پیشنهادی مقاله را شرح داده و در بخش ۵، نحوه پیاده‌سازی روش پیشنهادی بیان شده است. بخش ۶ شامل گزارشی از نتایج آزمایش‌ها می‌باشد.

۲- کارهای مرتبط

اولین کار مرتبط با تخمین سه بعدی حالت انسان از اطلاعات دو بعدی در [۱] گزارش شده است که در آن، از یک درخت تصمیم‌گیری باینری استفاده شده است. در این روش، مانند سایر روش‌های اولیه [۲-۴]، از اطلاعات آناتومیکی یا محدودیت زوایای حرکتی اسکلت انسان بهره گرفته‌اند. در برخی دیگر از روش‌ها، مانند [۵-۷]، به منظور رفع ابهام ناشی از عمق تصویر، از روش نزدیک‌ترین همسایگی استفاده شده است.

در کارهایی مانند [۸-۱۰]، برای ایجاد محدودیت آناتومیکی، یک مدل آماری به صورت مستقیم از اطلاعات ضبط حرکت[1] از پایگاه داده‌ها یادگیری شده است. نویسندگان مراجع [۸، ۱۰-۱۴] حالت انسان را ترکیب تنکی از حالت‌های پایه سه بعدی در نظر گرفته‌اند که این حالت‌های پایه را از اطلاعات ضبط حرکت به دست آورده‌اند. درحالی که محققان در [۸، ۱۱] با اعمال محدودیت در طول اجزای بدن، به ضرایب توصیف تنک و زاویه دید دوربین دست یافته‌اند، جو و همکارانش [۱۳] برای این هدف از رهاسازی محدب[2] استفاده کرده‌اند. در روشی دیگر معرفی شده توسط جو و همکارانش [۱۴] برای دستیابی به عملکرد مناسب، از اطلاعات زمانی و اطلاعات فعالیت افراد استفاده می‌شود.

در برخی دیگر از کارهای ابتدایی در تخمین حالت سه بعدی انسان، مانند [۱۵-۲۱]، از ویژگی‌های طراحی شده مانند سایه‌نما [۱۶]، متن شکل [۱۷]، توصیف‌گر SIFT [۱۸] و هیستوگرام جهت‌های لبه‌ها [۱۵] استفاده شده است. این روش‌ها بدون استفاده از حالت دو بعدی، مستقیماً از خود تصویر و توسط یک تابع نگاشت، حالت سه بعدی را استنتاج کرده‌اند. به این دسته از روش‌ها، روش‌های تمییزدهنده[3] گفته می‌شود.

با روی کارآمدن دوباره شبکه‌های عصبی کانولوشنی[4] در سال ۲۰۱۲ به عنوان یکی از مهم‌ترین روش‌های یادگیری عمیق در بحث طبقه‌بندی تصاویر [۲۲]، بسیاری از مسائل بینایی ماشین از این موفقیت بهره گرفته‌اند، مانند آشکارسازی اشیاء، بخش‌بندی تصاویر، بازشناسی چهره انسان، بازشناسی فعالیت انسان و بازیابی تصویر مبتنی بر محتوا. دلیل عمده حضور دوباره و موفقیت چشم‌گیر این نوع شبکه‌ها، وجود پایگاه داده‌های بسیار بزرگ و همچنین، روش‌های محاسباتی قوی (استفاده از پردازش موازی توسط کارت‌های گرافیکی) است. مسئله تخمین حالت دو بعدی انسان نیز از بحث یادگیری عمیق بهره‌مند شده است و روش‌های مطرح شده، بهبود عملکردی بسیار زیادی نسبت به روش‌های پیش از یادگیری عمیق نشان داده‌اند [۲۳-۲۷].

[1] Motion Capture (MoCap)
[2] Convex Relaxation
[3] Discriminative
[4] Convolutional Neural Network (CNN or ConvNet)

از کارهایی که به منظور نگاشت از فضای دو بعدی به فضای سه بعدی از یادگیری عمیق بهره برده‌اند، می‌توان به [۲۸-۳۱] اشاره کرد. دو مرجع [۳۲, ۳۳] از داده‌های سنتز شده برای آموزش با هدف تخمین سه بعدی حالت انسان استفاده کرده‌اند. پاولاکوس و همکارانش در مقاله [۳۰] با استفاده از یک شبکه از پیش تعریف شده و موفق در تخمین دو بعدی حالت انسان، به جای انجام عمل رگرسیون به نقشه‌های حرارتی دو بعدی، رگرسیون را به توزیع احتمالات در فضای سه بعدی انجام داده‌اند. در واقع، در این تحقیق، نگاشت مستقیم از فضای دو بعدی به سه بعدی (با دقت کمتری نسبت به حالتی که رگرسیون حجمی انجام می‌شود) پیشنهاد شده است. مورنو-نوگر [۲۸]، یک ماتریس فاصله‌ی دو به دو از فضای دو بعدی به فضای سه بعدی پیش‌بینی کرده‌اند که نسبت به چرخش، انتقال و انعکاس مقاوم است و بدین ترتیب توانسته است تخمین‌های غیرمحتمل را کنار بگذارد. انگیزه این دو مقاله [۲۸, ۳۰] آن است که باور دارند تخمین سه بعدی نقاط از اطلاعات دو بعدی، مسئله‌ای مشکل و پیچیده است.

اما برخلاف این دو روش، مارتینز و همکارانش در مقاله [۲۹] اثبات کرده‌اند که نگاشت مستقیم دو بعد به سه بعد، به صورت کارآمدی می‌تواند توسط شبکه‌های عصبی عمیق یادگیری شود و مسئله مهم، آنالیز مستقیم اطلاعات دیداری و تخمین حالت دو بعدی است. مرجع [۳۴] نیز یک شبکه عمیق برای نگاشت از نقاط دو بعدی به سه بعدی پیشنهاد داده است. این روش، صرفاً در حالتی ارزیابی شده است که تخمین دو بعدی کاملاً صحیح باشد و در نتیجه، عملکرد این روش به هنگام استفاده از الگوریتم‌های تخمین حالت دو بعدی نامشخص است. دو مرجع [۸, ۳۵] نقاط دو بعدی را ابتدا به صورت دستی مشخص کرده‌اند و سپس، مختصات دو بعدی به سه بعد ارتقاء داده شده است. از منظری دیگر، محققانی مانند پارک و همکارانش [۳۶] نقاط کلیدی دو بعدی را به صورت مستقیم به کار گرفته‌اند، در حالی که نویسندگان مقاله‌های [۳۷, ۳۸] به جای نقاط دو بعدی از نقشه‌های حرارتی به دست آمده بهره برده‌اند. به صورت کلی، به این دسته از روش‌ها که نتایج تخمین حالت دو بعدی به حالت سه بعدی ارتقاء داده می‌شود، اصطلاحاً روش‌های پایپ‌لاین[1] گفته می‌شود.

برخی از کارهای گزارش شده حالت انسان را بر مبنای زاویه تعریف کرده و تخمین زده‌اند؛ مانند [۱۲, ۳۹]. یکی از محاسن این روش‌ها آن است که به دلیل کاهش درجه آزادی و حرکت مفاصل، حالت‌های تخمین زده شده طوری به دست می‌آیند که ساختار معتبری برای انسان دارند. اما در [۲۹] نشان داده شده است که این نوع فرضیه، باعث یادگیری و استنتاج سخت‌تر و پیچیده‌تر به هنگام نگاشت مستقیم می‌شود و در نتیجه، منظور کردن مفاصل به صورت مکان سه بعدی، و نه بر مبنای زاویه، ترجیح داده شده است.

می‌توان مقالات مطرح شده در زمینه تخمین حالت سه بعدی را بر این مبنا نیز دسته‌بندی کرد که اطلاعات در قسمت ورودی و خروجی الگوریتم کلی چگونه منظور شده‌اند. در قسمت ورودی، تصویر خام [۱۴, ۳۰, ۳۱, ۳۶, ۳۹-۴۳] و توزیع احتمالات دو بعدی [۱۴, ۳۰] بررسی شده است. این در حالی است که در خروجی، احتمالات سه بعدی [۳۰]، پارامترهای حرکتی سه بعدی [۳۹] و ضرایب بردارهای پایه‌های حالت و پارامترهای دوربین [۸, ۱۰, ۱۲-۱۴] به دست آمده است.

از میان روش‌های بیان شده، یکی از مهم‌ترین کاستی‌هایی که روش‌های مبتنی بر رگرسیون مستقیم اطلاعات سه بعدی از ویژگی تصویر دارند، مانند [۲۴, ۳۱, ۳۹, ۴۲, ۴۴]، این است که نیازمند تصاویر بسیار زیاد به دست آمده در شرایط طبیعی و گوناگون و همراه با حاشیه‌گذاری دقیق سه بعدی هستند که چنین پایگاه داده‌ای وجود ندارد. توجه شود که فراهم کردن اطلاعات حقیقی سه بعدی برخلاف اطلاعات دو بعدی، به سادگی و به صورت دستی امکان‌پذیر نیست. بنابراین، کاربرد روش‌های مطرح شده به این سبک، صرفاً به محیطی که در آن تصاویر آموزشی به دست آمده‌اند، محدود می‌شود.

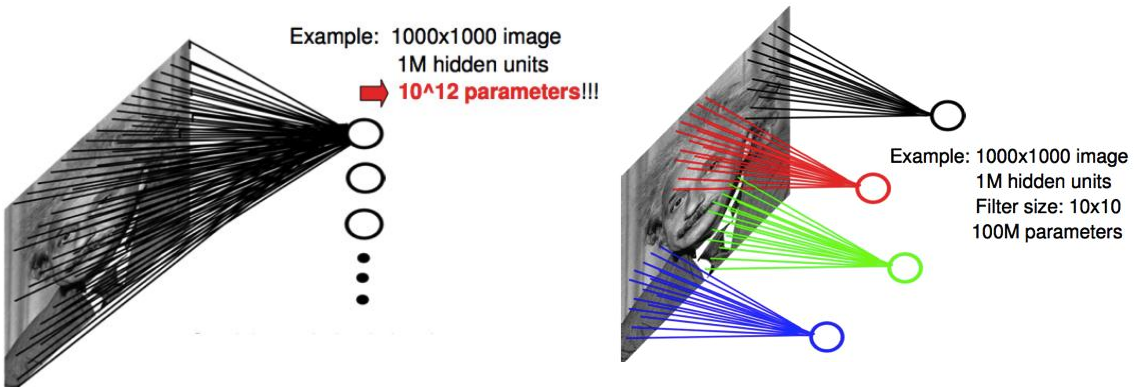
مرحله اولیه روش‌هایی که ابتدا حالت دو بعدی و سپس حالت سه بعدی را به دست آورده‌اند (یعنی مرحله‌ی تخمین حالت دو بعدی) بسیار حائز اهمیت است؛ چراکه بسیاری از خطاهای تخمین سه بعدی نهایی، ناشی از ادراک اشتباه از حالت دو بعدی انسان است. این در حالی است که بسیاری از روش‌ها توجه خود را صرفاً به مرحله نگاشت حالت دو بعدی به حالت سه بعدی معطوف کرده‌اند. مارتینز و همکارانش [۲۹] با بررسی خطاهای به دست آمده از تخمین سه بعدی، به این نتیجه رسیده‌اند که نگاشت حالت از فضای دو بعد به فضای

[1] Pipeline

سه بعد، خطای بسیار ناچیزی دارد و مقدار بیشتر خطاها، ناشی از تحلیل نادرست تصاویر خام و تخمین غیردقیق یا نادرست حالت دو بعدی است. در نتیجه، در این مقاله، برخلاف روش‌هایی که از اطلاعات حقیقی دو بعدی برای تخمین اطلاعات سه بعدی استفاده می‌شود، با درنظر گرفتن روشی نوین برای به دست آوردن تخمین دقیق حالت دو بعدی انسان، بهبود زیادی در تخمین سه بعدی حاصل شده است.

۳- شبکه‌های عصبی کانولوشنی غیرخطی عمیق

شبکه‌های عصبی کانولوشنی که با تأثیر از سیستم بینایی انسان طراحی شده‌اند، یکی از انوع مهم روش‌های یادگیری عمیق می‌باشند. این شبکه‌ها شامل تعدادی لایه‌های غیرخطی سلسله‌مراتبی هستند که توصیف فشرده‌تری از داده ورودی، پس از عبور از هر لایه، به دست می‌دهند. عموماً، این شبکه‌ها از سه نوع لایه اصلی تشکیل یافته‌اند: لایه‌های کانولوشنی، لایه‌های ادغام و لایه‌های تماماً متصل[1]. لایه‌های کانولوشنی شامل تعدادی پارامتر (وزن‌های کرنل‌ها یا فیلترها) هستند که بین داده ورودی آن لایه و پارامترهای همان لایه، عمل کانولوشن اعمال می‌کنند. داده پس از عبور از لایه کانولوشنی، به صورت یک نقشه ویژگی، به دست آمده و سپس، توسط لایه‌ی ادغام، نمونه‌برداری می‌شود. ادغام به سبک ماکزیمم‌گیری[2]، یکی از رایج‌ترین روش‌های ادغام است. عمل ادغام، علاوه بر کاهش ابعاد نقشه ویژگی به دست آمده و کاهش تعداد وزن‌ها در لایه‌های بعدی، باعث می‌شود که ویژگی به دست آمده، تا حدی نسبت به تغییرات انتقال مقاوم باشد. عملکرد لایه تماماً متصل نیز مانند لایه‌ی کانولوشنی است؛ با این تفاوت که برخلاف لایه‌ی کانولوشنی که به صورت محلی به فیلترها اعمال می‌شود، در لایه‌ی تماماً متصل، فیلتر به صورت کلی و یک‌جا در داده ورودی عمل کانولوشن را انجام می‌دهد. همین امر باعث می‌شود که تعداد پارامترهای لایه تماماً متصل، بسیار بیشتر از تعداد پارامترهای لایه کانولوشنی باشد. تفاوت عملکرد این دو لایه در شکل (۱) نشان داده شده است.



(ب) (الف)

شکل (۱). نحوه عملکرد فیلترهای (الف) لایه کانولوشنی و (ب) لایه تماماً متصل. درلایه کانولوشنی تعداد پارامترها بسیار کمتر از لایه تماماً متصل است.

در شبکه عصبی کانولوشنی، اگر Y_L^k، k–امین نقشه ویژگی لایه L–ام باشد، با توجه به وزن‌های لایه کانولوشنی L–ام (W_L^k) و مقادیر بایاس همان لایه (b_L^k)، Y_L^k از طریق محاسبات غیرخطی به صورت زیر حاصل می‌گردد:

$$Y_L^k = f\left(W_L^k * Y_{L-1} + b_L^k\right) \quad (۱)$$

در فرمول (۱)، علامت *، بیانگر عمل کانولوشنی و $f(.)$، بیانگر یک تابع غیرخطی است. در شبکه‌های عصبی کانولوشنی، برخلاف شبکه‌های عصبی متداول، این تابع به صورت زیر تعریف می‌شود که «واحدهای خطی تصحیح شده[3]»، نام دارد:

$$f(x) = \max(0, x) \quad (۲)$$

[1] Fully-connected
[2] Max Pooling
[3] Rectified Linear Units (ReLU)

در حقیقت، هدف نهایی از آموزش شبکه‌های عصبی کانولوشنی، یافتن مقادیر پارامترهایی مانند Θ است که با استفاده از تخمین همانند بیشینه، با داده‌های آموزشی سازگار شوند:

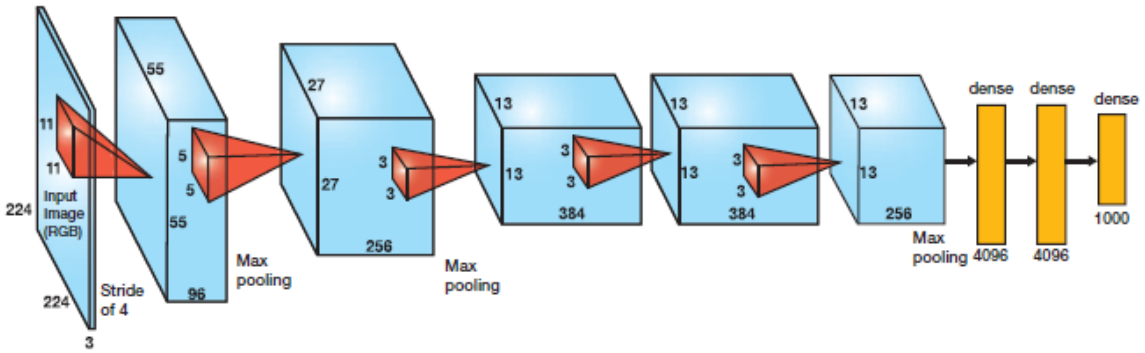
$$\arg\max_{\Theta} L(\Theta, D) = \arg\max_{\Theta} \prod_{n=1}^{N} h(X \mid \Theta) \qquad (۳)$$

در این رابطه، $h(X \mid \Theta)$ احتمال پسین نمونه‌ی X است. عبارت (۳) را می‌توان به صورت اتلاف آنتروپی متقابل[1] به شکل زیر نوشت:

$$-\ln[P(D \mid \Theta)] = -\sum_{n=1}^{N} [yh(X;\Theta) + (1-y)(1-h(X;\Theta))] \qquad (۴)$$

پارامتر y بیانگر برچسب کلاس است. به منظور بهینه‌سازی این عبارت، از روش کاهش تصادفی گرادیان[2] و برای تنظیم وزن‌ها، از روش انتشار پسرو[3] استفاده می‌شود.

مهم‌ترین شبکه‌ای که به عنوان نقطه عطفی برای به کارگیری یادگیری عمیق در داده‌های تصویری مطرح شده است، شبکه AlexNet [۲۲] می‌باشد. این شبکه هفت لایه‌ای در سال ۲۰۱۲ در مسابقه طبقه‌بندی تصاویر ImageNet [۴۵] به ۱۰۰۰ نوع کلاس، به خطای کمتری دست یافته است. ساختار این شبکه، در شکل ۲ به نمایش در آمده است.



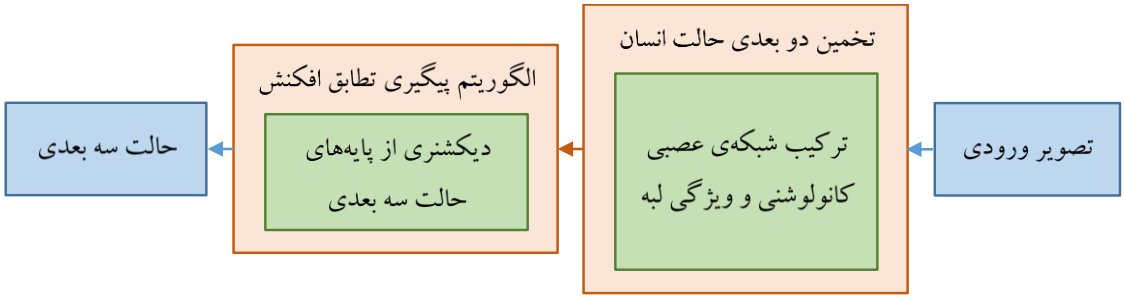
شکل (۲). ساختار شبکه AlexNet برای طبقه‌بندی تصویر ورودی به ابعاد ۳×۲۲۴×۲۲۴ به ۱۰۰۰ عدد کلاس

٤- روش پیشنهادی

در این مقاله، تخمین سه بعدی حالت انسان به صورت غیرمستقیم و با استفاده از تخمین حالت دو بعدی انجام می‌شود. هدف، به دست آوردن یک توصیفی از حالت انسان به صورت $p = (x_1, y_1, z_1, ..., x_n, y_n, z_n)$ با استفاده از اطلاعات خام تصاویر منفرد است. p نشانگر بردار حالت و x_i ، y_i و z_i نشانگر مختصات مفصل i–ام و n تعداد مفاصل مفروض است. تعداد مفاصل مفروض در این مقاله چهارده و به صورت استاندارد پایگاه داده‌ی LSP [۴۶] در نظر گرفته شده‌اند (شکل (۳)). بلوک‌دیاگرام روش پیشنهادی در شکل (۴) مشاهده می‌شود. در ادامه، به شرح قسمت‌های مختلف این بلوک‌دیاگرام خواهیم پرداخت.

[1] Cross-Entropy Loss
[2] Stochastic Gradient Descent (SGD)
[3] Back Propagation

شکل (۳). چهارده مفصل مفروض در اسکلت انسان که حالت انسان را مشخص می‌کنند (مطابق استاندارد پایگاه داده‌ی LSP)



شکل (۴). بلوک‌دیاگرام کلی روش پیشنهادی به منظور تخمین حالت سه بعدی

٤-١- تخمین حالت دو بعدی

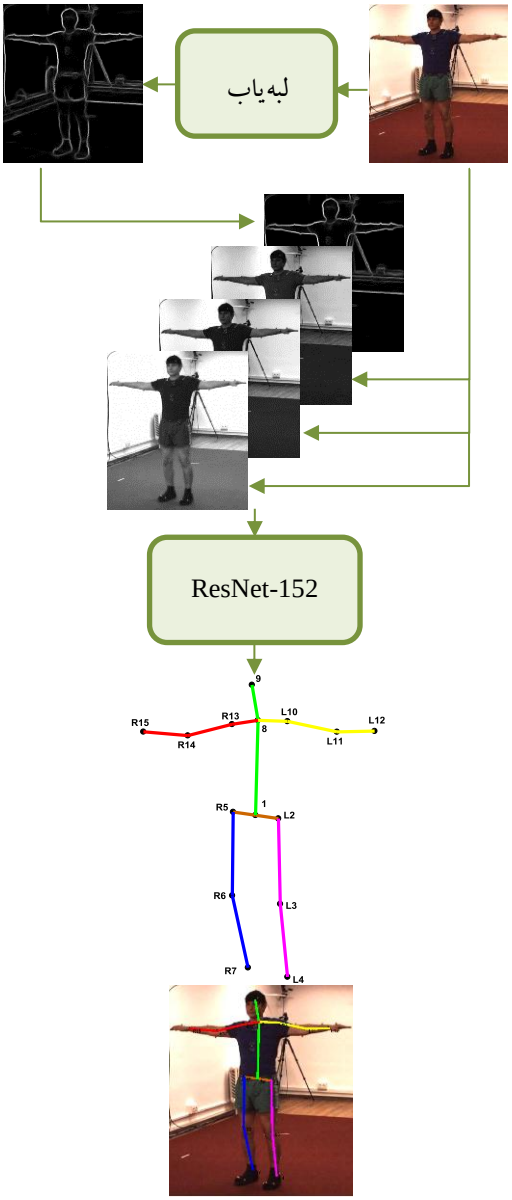
در این مقاله، به منظور استخراج حالت دو بعدی از یک تصویر رنگی، شبکه عصبی کانولوشنی به گونه‌ای آموزش داده می‌شود که یک خروجی به صورت برداری مانند $p=(x_1,y_1,...,x_n,y_n)$ به دست می‌دهد. به عبارتی دیگر، شبکه عصبی کانولوشنی، نقش یک تابع نگاشت غیرخطی را خواهد داشت که عمل رگرسیون انجام می‌دهد. در روش پیشنهادی، علاوه بر این‌که خود تصویر به صورت سه کانال RGB به عنوان ورودی در نظر گرفته می‌شود، کانال دیگری به همراه سه کانال R، G و B وارد شبکه شده و آن تصویر احتمال لبه (نقشه‌ی لبه) است. بنابراین، داده ورودی شبکه‌ی عصبی کانولوشنی، به ابعاد $h\times w\times 4$ و داده خروجی آن، به ابعاد $n\times 2$ می‌باشد که h و w ابعاد تصویر و n تعداد مفاصل مفروض است. بدین ترتیب، روش پیشنهادی به منظور تخمین حالت دو بعدی انسان، روش ترکیبی است که از ویژگی‌های دستی و طراحی شده (نقشه لبه) و ویژگی‌های یادگیری شده توسط شبکه‌های عصبی کانولوشنی استفاده می‌کند. شکل (۵) بلوک‌دیاگرام روش پیشنهادی را برای تخمین حالت دو بعدی با استفاده از ترکیب ویژگی نقشه لبه و شبکه عصبی کانولوشنی نمایش می‌دهد.

در حقیقت، به علت این‌که اساس عملکرد شبکه‌های عصبی کانولوشنی بر عمل کانولوشن (ضرب مقادیر فیلترها در مقادیر داده ورودی، و جمع حاصل‌ضرب‌ها) است، در صورتی که داده‌های ورودی شامل اطلاعات ناخواسته یا غیرمرتبط با هدف آموزشی باشند، با اعمال تأکید بیش‌تر در نواحی مطلوب و ایجاد توجه کم‌تر به داده‌های نامطلوب، به شبکه کمک می‌شود که هدف مورد نظر را بهتر آموزش ببینند. اعمال تأکید بیش‌تر و ایجاد توجه کم‌تر، با قرار دادن مقادیر بزرگ‌تر و کم‌تر، به ترتیب در قسمت‌های مرتبط با داده‌های مطلوب و غیرمطلوب، عملی می‌شود.

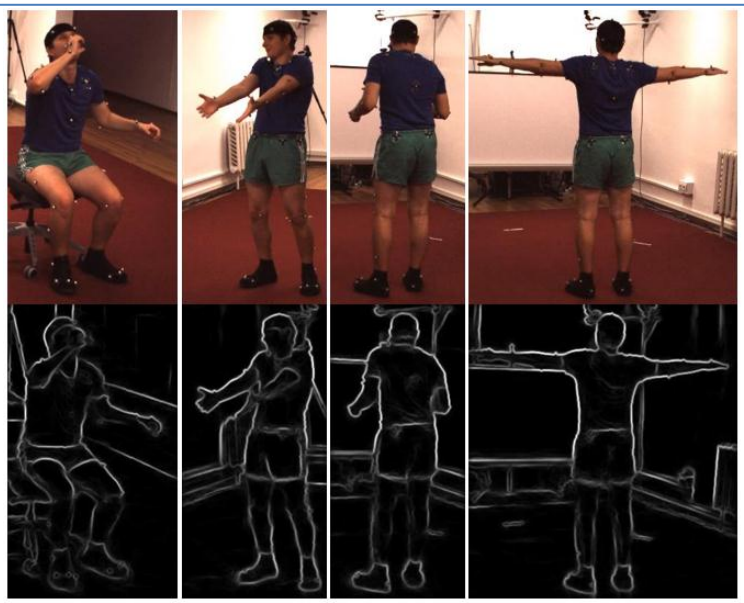
چنین مقادیری در مسئله مورد بررسی، یعنی تخمین حالت انسان، در نقشه لبه ایجاد شده است. در نقشه لبه، شکل انسان برای شبکه فراهم می‌شود و همان‌طور که می‌دانیم، شکل انسان، یکی از مؤثرترین اطلاعات برای تخمین حالت انسان است. از طرفی دیگر، نقشه لبه باعث می‌شود اطلاعاتی که در تخمین حالت انسان اثری ندارند (مانند اطلاعات رنگ و بافت لباس، و اطلاعات پس‌زمینه) به هنگام

آموزش نقش کمتری داشته باشند. این نوع اطلاعات گاهی باعث بایاس شدن یا گمراه شدن شبکه به هنگام یادگیری می‌شوند. به عنوان مثال، اگر در تصاویر آموزشی، یک رنگ مشخصی از لباس بیشتر وجود داشته باشد، در صورت ورودی RGB و بدون استفاده از نقشه‌ی لبه، فیلترها برای آن رنگ، وزن‌های بیشتری به خود می‌گیرند و در واقع، شبکه به آن رنگ بایاس می‌شود. حال اگر به هنگام آزمایش، فردی در حالت مشابه و در یک لباس با رنگ متفاوتی باشد، شبکه دچار خطا خواهد شد. اما در تصویر لبه، اطلاعات غیرمرتبط با حالت انسان، مانند رنگ وجود ندارد و در صورت استفاده از نقشه لبه در کنار ورودی شبکه، در نواحی مطلوب که کانتور، لبه و شکل انسان (مقادیر بیش‌تر و روشن‌تر) وجود دارد، بر داده‌ها تأکید بیشتری خواهد شد و در قسمت‌های با مقادیر کم‌تر (تیره‌تر) توجه کم‌تری به داده‌های آموزشی می‌شود.

در شکل (۶)، نمونه‌ای از تصاویر احتمال لبه مشاهده می‌گردد. ملاحظه می‌شود که قسمت‌هایی از تصویر که مرتبط با شکل انسان است، با درجه روشنایی بیشتری به دست آمده است و اغلب اطلاعات نامرتبط یا کمرنگ شده‌اند و یا از بین رفته‌اند. برای به دست آوردن اطلاعات لبه از الگوریتم پشنهادی در مرجع [۴۷] استفاده شده است.



شکل (۵). بلوک‌دیاگرام روش پیشنهادی برای تخمین حالت دو بعدی



شکل (۶). نمونه‌هایی از تصاویر پایگاه داده‌ی Human3.6M [۴۸] و تصاویر احتمال لبه متناظر

٤-۲- تخمین حالت سه بعدی از حالت دو بعدی

به منظور ارتقاء حالت پیش‌بینی شده از فضای دو بعدی به فضای سه بعدی، از روش پیگیری تطابق افکنش در [۸] استفاده می‌شود. در این الگوریتم، حالت سه بعدی به کمک افکنش آن در فضای دو بعدی و نیز یک دیکشنری به دست می‌آید. به عبارتی دیگر، هدف از این روش، کمینه کردن تفاوت بین حالت دو بعدی ورودی و افکنش حالت سه بعدی پیش‌بینی شده از زاویه دوربین است. حالت دو بعدی ورودی، حالت دو بعدی به دست آمده از روش پیشنهادی این مقاله منظور می‌شود. حالت سه بعدی انسان را می‌توان به صورت ترکیب خطی از یک حالت میانگین و تعدادی بردارهای پایه و به صورت توصیف تنک نشان داد:

$$X = m + \sum_i w_i B_i \qquad (۵)$$

در معادله بالا، m حالت میانگین است که از میانگین‌گیری در بین حالت‌های داده‌های آموزشی به دست می‌آید، B_i ها بردارهای پایه هستند که به عنوان عناصر دیکشنری منظور می‌شوند و w_i ها وزن‌های متناظر با بردارهای پایه می‌باشند.

همان‌طور که می‌دانیم، با در دست داشتن مشخصات ذاتی دوربین که توسط ماتریس چرخش و بردار انتقال بیان می‌شود، افکنش نقاط سه بعدی به فضای دو بعدی از زاویه دید دوربین، به صورت زیر به دست می‌آید که در آن علامت $\otimes$ بیانگر ضرب کرونکر، T ماتریس انتقال و M ماتریس توأم چرخش و مقیاس است:

$$x = (I_n \otimes M)X + 1_{n\times 1} \otimes T \qquad (۶)$$

با توجه به توضیحات ذکر شده، تابع هدف در روش پیگیری تطابق افکنش به صورت زیر تعریف می‌شود:

$$\underset{M,T,W,\beta}{\arg\min} \| (x_{ES} - (I_n \otimes M)(m + \sum_i w_i B_i) + 1_{n\times 1} \otimes T) \|_2 \qquad (۷)$$

که در آن، M ماتریس توأم چرخش و مقیاس، T ماتریس انتقال، β زیرمجموعه‌ای از پایه‌های دیکشنری، W بردار وزن‌های پایه‌های موجود در β و x_{ES} حالت دو بعدی تخمینی است. در این روش، در تکرارهای مختلف، بردارهای پایه دیکشنری، به حالت میانگین اضافه می‌شوند و نهایتاً مجموعه‌ای از بردارهای پایه انتخاب می‌شود که باعث ایجاد کم‌ترین تفاوت بین افکنش حالت سه بعدی تخمینی از زاویه دوربین تخمینی و مختصات دو بعدی پیش‌بینی شده به روش پیشنهادی (x_{ES}) شود.

به طور دقیق‌تر، در اولین تکرار، تخمین اولیه برای حالت سه بعدی، میانگین حالت‌های سه بعدی آموزشی در نظر گرفته می‌شود. سپس، پارامترهای دوربین به صورت زیر به دست می‌آیند که اگر ماتریس x، حالت دو بعدی و ماتریس X، حالت سه بعدی باشد، به

کمک تجزیه مقدارهای منفرد می‌توان نوشت:

$$Z = xX^T (XX^T)^{-1} = UDV^T \quad (۸)$$

در این صورت، ماتریس مقیاس دوربین، اولین قسمت به اندازه ۲×۲ از ماتریس D خواهد شد و ماتریس چرخش دوربین، برابر است با DV^T. سپس، تفاوت حالت دو بعدی با تفاوت افکنش حالت سه بعدی از زاویه دوربین تخمینی به دست می‌آید. بردار پایه‌ای از دیکشنری که با این بردار تفاوت، هم‌ترازی بیشتری داشته باشد (ضرب داخلی‌شان بیشتر شود)، انتخاب شده و با اضافه شدن به حالت سه بعدی، حالت سه بعدی جدید تخمین زده می‌شود. سپس، تکرار بعدی، مشابه این مرحله، با تخمین پارامترهای دوربین با حالت سه بعدی جدید آغاز می‌شود. به همین ترتیب، این روال تکرار شده و تا جایی ادامه پیدا می‌کند که خطای بین حالت دو بعدی با افکنش حالت سه بعدی، کمتر از مقدار معینی شود.

۵- پیاده‌سازی

۵-۱- آموزش شبکه به منظور تخمین حالت دو بعدی

برای تخمین حالت دو بعدی انسان، از شبکه عصبی کانولوشنی ResNet-152 [۴۹] استفاده شده است. این شبکه ۱۵۲ لایه‌ای بر اساس ایده شبکه‌های مبتنی بر یادگیری باقیمانده[1] طراحی شده و در مسابقه ILSVRC2015 و در بخش طبقه‌بندی تصاویر پایگاه داده‌ی ImageNet [۴۵] با خطای ٪۳/۵۷ جایگاه اول را به خود اختصاص داده است. ورودی این شبکه، با توجه به روش پیشنهادی این مقاله به ابعاد ۴×۲۲۴×۲۲۴ و خروجی، با استفاده از تغییر لایه‌های آخر به ابعاد ۱۴×۲ تغییر داده می‌شود. در جدول (۱) مشخصات شبکه‌ی مورد استفاده آمده است. مشخصات این شبکه، مشابه مشخصات شبکه ResNet-152 است، با این تفاوت که: ۱- در لایه اول، ابعاد فیلتر از ۶۴×۳×۷×۷، به ۶۴×۴×۷×۷ تغییر کرده است؛ ۲- در آخر شبکه، لایه تماماً متصل از ابعاد ۱۰۰۰×۱×۱×۲۰۴۸ به ۲۸×۲۰۴۸×۱×۱ تغییر یافته است؛ ۳- به جای استفاده از تابع هدف رگرسیون منطقی[2] (طبقه‌بندی) با استفاده از تابع SoftMax، از رگرسیون حقیقی با استفاده از تابع MSE[3] استفاده شده است.

جدول (۱). مشخصات اصلی شبکه‌ی مورد استفاده

layer name	conv1	conv2_x		conv3_x	conv4_x	conv5_x	FC
output size	112×112	56×56		28×28	14×14	7×7	1×1
layer info	7×7, 64, stride	3×3 max pool, stride 2	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	FC1 (2048×2048) FC2 (2048×28) MSE loss

برای مقادیر اولیه وزن‌های تمام لایه‌ها از وزن‌های از پیش آموزش داده شده در پایگاه ImageNet، استفاده می‌شود و وزن‌های مرتبط با کانال چهارم در لایه اول، به صورت نویز تصادفی با توزیع گوسین و با واریانس ۰/۱ منظور شده‌اند. از تصاویر پایگاه داده MPII [۵۰] و LSP [۴۶, ۵۱] برای آموزش شبکه بهره گرفته شده است. جمعاً به تعداد ۲۴۵۶۸ تصویر آموزشی تعیین شده است که با افزودن داده

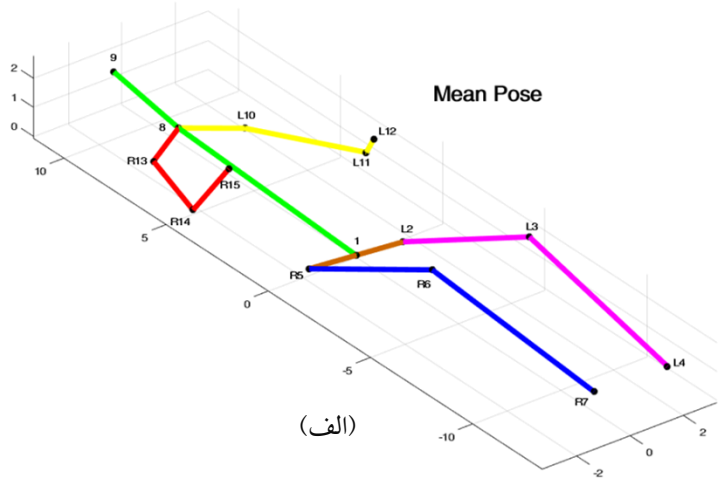
[1] Residual Learning Networks
[2] Logistic Regression
[3] Mean Square Error

با استفاده از روش‌های قرینه کردن و چرخش بین ۴۰- و ۴۰ درجه با گام‌های ۱۰ درجه‌ای، تعداد تصاویر به ۴۴۲۲۲۴ افزایش می‌یابد. افزایش داده[1] یکی از مراحل مهم برای آماده کردن داده‌های آموزشی شبکه‌های عصبی کانولوشنی است و هدف اصلی آن جلوگیری از بیش‌برازش شدن[2] شبکه می‌باشد.

شبکه با استفاده از ۴۴۲۲۲۴ تصویرآموزشی به تعداد ۱۰ اپک آموزش داده می‌شود. یادگیری براساس کاهش تصادفی گرادیان همراه با ممنتوم با تعداد دسته‌های ۱۶ تایی، با استفاده از روش انتشار پسرو و با نرخ یادگیری ثابت به مقدار ۰/۰۰۱ انجام شده است. به منظور پیاده‌سازی، نرم‌افزار Matlab وچهارچوب Caffe که در تولباکس MatConvNet فراهم آورده شده، به کارگرفته شده است.

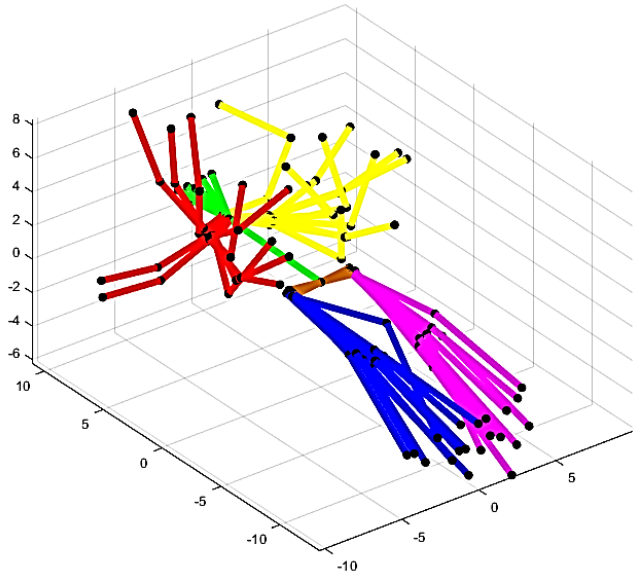
۵-۲- ایجاد دیکشنری پایه‌ها

به منظور ایجاد دیکشنری، از پایگاه داده CMU [۵۲] استفاده شده است. CMU یک پایگاه داده ضبط حرکت است که حالت‌های موجود در این پایگاه داده نسبت به پایگاه داده Human3.6M [۴۸] دارای تنوع بیشتری است. این پایگاه شامل تقریباً ۲۵۰۰ دنباله ویدئویی و بیش از سه میلیون حالت سه بعدی است که از انجام ۳۰ نوع فعالیت توسط ۱۴۴ فرد جمع‌آوری شده است. در این پایگاه داده، مشابه مقاله [۸]، از فعالیت‌های زیر استفاده شده است: running، waving، arm movement، walking، jumping، jumping jacks، run، sit، boxing و bend. برای هر حالت سه بعدی، یک دستگاه مختصات محلی منظور شده است و حالت‌ها در فضای سه بعدی به گونه‌ای تغییر می‌کنند (انتقال، چرخش و تغییر مقیاس) که دستگاه مختصات محلی آن‌ها، هم‌تراز با دستگاه مختصات محلی یک حالت مرجع (حالت میانگین) شود. دستگاه مختصات محلی بدین صورت تعریف می‌شود: ۱- محور ستون فقرات به عنوان محور y، ۲- ضرب خارجی برداری که دو مفصل ران را مشخص می‌کند در محور y، به عنوان محور x و ۳- ضرب خارجی دو محور x و y به دست آمده، محور z در نظر گرفته می‌شود. در شکل (۷) تعدادی از حالت‌ها پس از تغییر مقیاس، انتقال و چرخش و هم‌تراز شدن به یک دستگاه مختصات محلی و همچنین حالت میانگین، به نمایش درآمده است. در گام بعدی، با اعمال تحلیل اجزای اصلی[3] در هر کدام از فعالیت‌های ذکر شده و در نظر گرفتن اجزای اصلی که ۹۹٪ انرژی را دربردارند، بردارهای پایه‌های دیکشنری ایجاد شده است. مرجع [۵۳] در روشی مشابه (توصیف تنکی از حالت سه بعدی)، خود حالت‌ها را که به صورت تصادفی انتخاب می‌شوند، به حالت میانگین اضافه می‌کند. در مقاله‌ی [۵۴]، اهمیت نحوه‌ی تشکیل بردارهای پایه در دیکشنری بررسی شده است.



(الف)

[1] Data Augmentation
[2] Overfitting
[3] Principal Component Analysis (PCA)



شکل (۷). (الف) حالت میانگین، (ب) نمونه‌هایی از حالت‌های سه بعدی آموزشی پس از تغییر مقیاس، انتقال، چرخش و درنتیجه، همتراز شدن با حالت میانگین

٦- نتایج آزمایش‌ها

٦-١- نتایج تخمین حالت دو بعدی

پایگاه داده‌ی Human3.6M [۴۸] بزرگ‌ترین پایگاه داده برای تخمین سه بعدی حالت انسان است. اطلاعات ضبط حرکت پایگاه داده‌ی Human3.6M شامل ۳/۶ میلیون حالت سه بعدی است. تصاویر این پایگاه از انجام ۱۵ فعالیت روزمره (مانند قدم زدن، صحبت کردن با تلفن، نشستن، خوردن و ...) توسط ۱۱ فرد (۵ زن و ۶ مرد) گردآوری شده است. پروتکل‌های مختلفی برای ارزیابی در این پایگاه داده پیشنهاد شده‌اند. در پروتکل استاندارد خود پایگاه داده، اطلاعات جعبه محاط مشخص است و ارزیابی در داده‌های آموزشی و آزمایشی یک نوع فعالیت مشخص صورت می‌گیرد. در [۱۲, ۵۵] نیز پروتکل‌های دیگری معرفی شده‌اند. در پروتکل معرفی شده در [۵۵] یا اصطلاحاً پروتکل شماره ۲، اطلاعاتی در مورد فعالیت شخص وجود ندارد و داده‌های آموزشی شامل تمام کلاس‌های مرتبط با فعالیت‌های مختلف است. در این پروتکل از پایگاه داده Human3.6M، از تصاویر افراد S1، S5، S6، S7، S8 و S9 به منظور آموزش و از تصاویر مربوط به فرد S11 برای ارزیابی کمی روش تخمین استفاده می‌شود. به منظور آزمایش از هر ۶۴ فریم دنباله ویدئویی S11، یک فریم انتخاب و همچنین، تمام زوایای دید (۴ دوربین) و همه نمایش‌ها (دو نمایش) منظور شده‌اند. با توجه به این بیانات، به تعداد ۳۶۸۳ تصویر آزمایشی وجود دارد.

در جدول‌های (۲) و (۳)، خطای تخمین حالت دو بعدی به ترتیب در تصاویر آزمایشی پروتکل ۲ پایگاه داده Human3.6M و ۱۰۰۰ تصویر آزمایش پایگاه داده LSP ملاحظه می‌گردد. روش به دست آوردن چنین خطایی مانند مرجع [۱۴] منظور شده است. در هر دو پایگاه، روش پیشنهادی در مقایسه با سایر روش‌ها، خطای کمتری نتیجه داده است. به منظور بررسی میزان بهبود عملکرد به هنگام استفاده از تصویر لبه، نتایج تخمین حالت دو بعدی به هنگام عدم استفاده از تصویر لبه به عنوان کانال چهارم، در جدول (۴) آورده شده است. ژو و همکارانش در [۱۴] از شبکه‌های عصبی کانولوشنی برای به دست آوردن نقشه حرارتی مفاصل دو بعدی بهره گرفته‌اند و تخمین حالت سه بعدی با استفاده از اطلاعات ویدئویی انجام می‌شود. این در حالی است که اطلاعات ورودی روش پیشنهادی مقاله حاضر تنها یک تصویر دو بعدی است و به هنگام آزمایش، از اطلاعات زمانی به عنوان ورودی استفاده نمی‌شود.

در جدول (۳) نیز، روژز و اشمیت در مقاله‌ی [۳۲] از شبکه‌ی عصبی کانولوشنی به عنوان یک طبقه‌بند با دو منبع اطلاعات ورودی

(تصاویر دو بعدی و حالت‌های سه بعدی) بهره گرفته‌اند. از منظری دیگر، در حالی که در روش [۳۲] تخمین حالت سه بعدی هدف اصلی روش پیشنهاد شده است، دو مرجع [۵۶, ۵۷] صرفاً با هدف تخمین حالت دوبعدی شبکه‌های عصبی کانولوشنی را به کار گرفته‌اند. در این جدول، ملاحظه می‌شود که در مفاصل مچ پاها و مچ دست‌ها که مشکل‌ترین مفاصل برای مکان‌یابی در تخمین حالت انسان هستند، روش پیشنهادی عملکرد بهتری دارد. بنابراین، با توجه به نتایج به دست آمده، استفاده از ویژگی‌های طراحی شده (نقشه لبه) به منظور هدایت و به کارگیری در یادگیری عمیق، اهمیت شایان توجهی دارد.

جدول (۲). خطای تخمین حالت دو بعدی تحت پروتکل ۲ پایگاه داده Human3.6M. بردارهای حالت قبل از محاسبه به مقدار ۲۲۰ نرمالیزه شده‌اند.

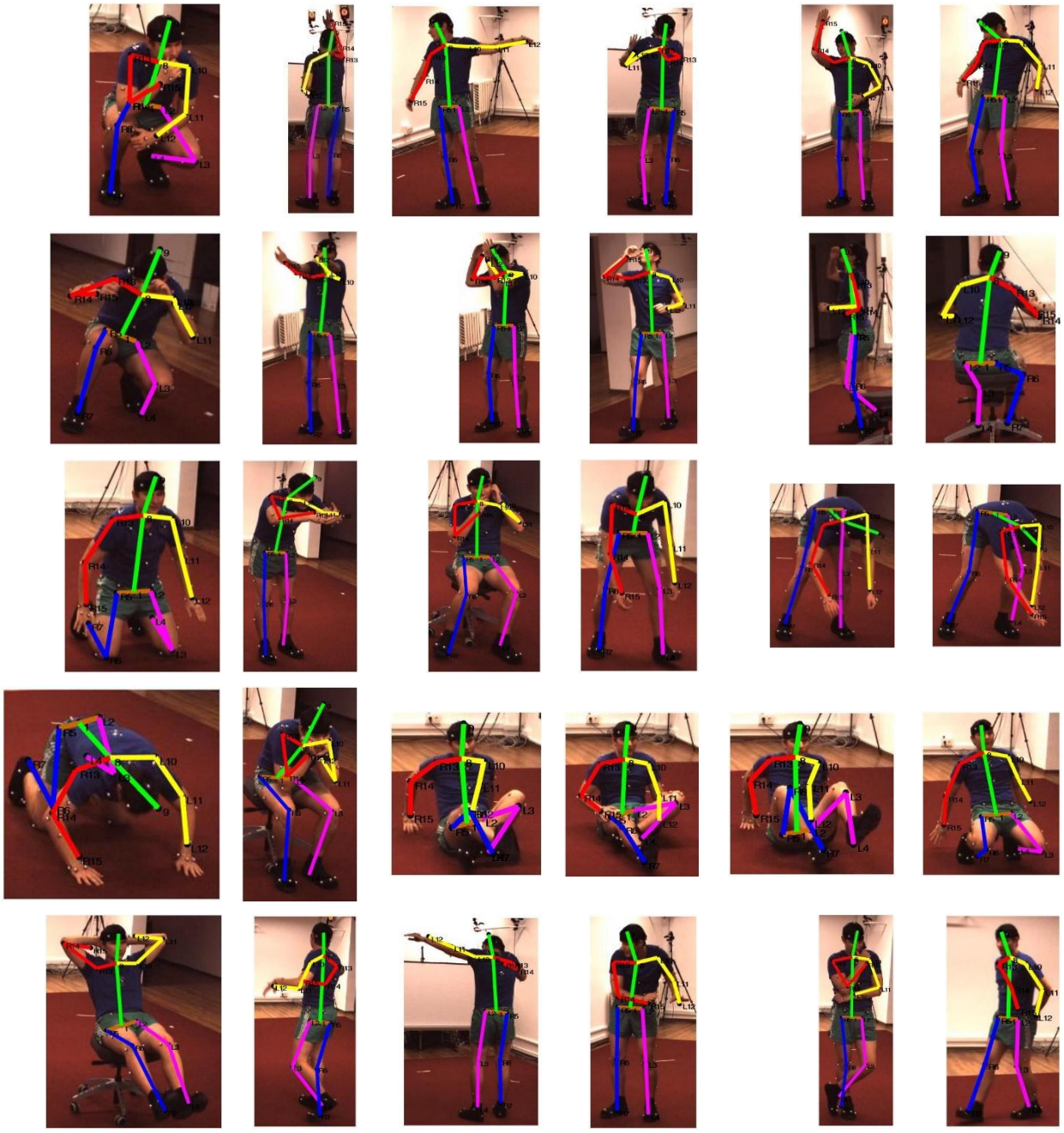
روش	خطای اقلیدسی (برحسب پیکسل)
بدون استفاده از تصویر لبه	۱۱/۲
[۱۴]	۱۰/۸
روش پیشنهادی	۱۰/۵

جدول (۳). خطای تخمین حالت دو بعدی در تصاویر آزمایشی پایگاه داده LSP در مفاصل مختلف اسکلت (خطای اقلیدسی برحسب پیکسل). بردارهای حالت قبل از محاسبه به مقدار ۲۲۰ نرمالیزه شده‌اند.

روش	مچ پاها	زانوها	ران‌ها	مچ دست‌ها	آرنج‌ها	شانه‌ها	سر	میانگین
(Alexnet) [۳۲]	۱۹/۱	۱۳/۰	۴/۹	۲۱/۴	۱۶/۶	۱۰/۵	۱۰/۳	۱۳/۸
(VGG) [۳۲]	۱۶/۲	۱۰/۶	**٤/۱**	۱۷/۷	۱۳/۰	**۸/٤**	۹/۸	۱۱/۴
[۵۶]	۱۵/۷	۱۱/۵	۸/۱	۱۵/۶	۱۲/۱	۸/۶	**٦/۸**	۱۱/۲
[۵۷]	۱۵/۵	۱۱/۵	۸/۰	۱۴/۷	۱۲/۲	۸/۹	۷/۴	۱۱/۱
روش پیشنهادی	**۱۰/۷**	**۹/۳**	۵/۳	**۱٤/۲**	**۱۱/۳**	۹/۲	۸/۵	**۹/۸**

نمونه‌هایی از تخمین حالت دو بعدی در تصاویر پایگاه داده Human3.6M در شکل (۸) مشاهده می‌گردد. همان‌طور که ملاحظه می‌شود، روش پیشنهادی به هنگام وجود چالش‌هایی مانند انسداد مفصلی از بدن که بایستی مکان‌یابی شود و حالت‌های پیچیده‌تر، تخمین مناسبی به دست می‌دهد. همچنین، این روش، اعضای متقارن بدن انسان را به درستی تشخیص می‌دهد؛ به عبارتی دیگر، در تشخیص اعضای سمت چپ و راست بدن دقت بالایی دارد. مهمترین چالشی که در تخمین حالت سه بعدی از یک تک تصویر دو بعدی وجود دارد، از بین رفتن اطلاعات بعد سوم (اطلاعات عمق) در تصویر دو بعدی است. این چالش هنگامی که حالت فرد به گونه‌ای باشد که اصطلاحاً کوتاه‌نمایی[1] اتفاق افتاده باشد، تشدید می‌شود. به طور دقیق‌تر، کوتاه‌نمایی به حالتی گفته می‌شود که از زاویه دید مورد بررسی، طول یک قسمت بدن (حد فاصل بین دو مفصل) کوتاه شده باشد و بیشتر اطلاعات آن قسمت در بعد سوم (که در تصویر دو بعدی حذف شده است) وجود داشته باشد. روش پیشنهادی در چنین حالت‌هایی نیز تخمین دو بعدی مناسبی به دست می‌دهد. علاوه بر موارد بیان شده، مشاهده می‌شود در قسمت‌هایی از تصویر که رنگ پس‌زمینه مشابه رنگ قسمتی از بدن است، اضافه کردن تصویر لبه به مکان‌یابی بهتر مفصل مورد نظر کمک می‌کند.

[1] Foreshortening



شکل (۸). نتایج تخمین حالت دو بعدی روش پیشنهادی در برخی تصاویر Human3.6M

۶-۲- نتایج تخمین حالت سه بعدی

معیار ارزیابی کمی تخمین حالت سه بعدی در پروتکل ۲، دقت معرفی شده در [۵۸] است که به «خطای حالت سه بعدی[1]» مشهور است. به منظور به دست آوردن این مقدار خطا، اسکلت پیش‌بینی شده، توسط یک تابع تبدیل صلب (تحلیل پروکروستس[2])، به اسکلت حقیقی هم‌تراز شده و سپس، مقدار متوسط فاصله‌ی اقلیدسی بین مفاصل مرتبط محاسبه می‌شود.

در جدول (۴)، خطای حالت سه بعدی با استفاده از روش پیشنهادی مقاله حاضر و در مقایسه با سایر روش‌ها آمده است. ملاحظه می‌شود که خطای به دست آمده از روش پیشنهادی ما کمتر می‌باشد. دقت شود که در پیاده‌سازی روش پیشنهادی، برخلاف سایر روش‌های ذکر شده در جدول، به هنگام آموزش از داده‌های آموزشی Human3.6M استفاده نشده است. مسلماً با منظور کردن داده‌های

[1] 3D Pose Error
[2] Procrustes Analysis

آموزشی Human3.6M، یعنی ساخت دیکشنری با استفاده از داده‌های سه بعدی آن پایگاه داده، خطای به دست آمده کمتر خواهد شد. اساس دو روش [۵۹] و [۳۲] به منظور تخمین سه بعدی حالت انسان تقریباً یکسان است؛ هر دوی آن‌ها، به علت کمبود داده‌های آموزشی به صورت تصاویر دو بعدی واقعی (و نه در محیط کنترل شده‌ی آزمایشگاه) همراه با حاشیه‌گذاری سه بعدی، از دو مسئله بازیابی حالت سه بعدی و تخمین حالت دو بعدی برای به دست آوردن حالت سه بعدی بهره برده‌اند. تفاوت این دو روش در آن است که یاسین و همکارانش در [۵۹] از جنگل‌های تصادفی برای مدل‌سازی استفاده کرده‌اند، در حالی که روژز و اشمیت در [۳۲]، از شبکه‌های عصبی کانولوشنی برای یادگیری مدل بهره برده‌اند. روش [۵۹]، برای تخمین حالت دو بعدی، مدل ساختارهای تصویری را به کار گرفته است.

همان‌طور که ملاحظه شد، در روش پیشنهادی، پایگاه داده‌های آموزشی دو بعدی و سه بعدی کاملاً مستقل از هم هستند. برای آموزش تخمین دو بعدی، از پایگاه داده‌های MPII و LSP، برای ایجاد دیکشنری پایه‌ها از پایگاه داده CMU و برای آزمایش از پایگاه داده Human3.6M استفاده شده است. در نتیجه، ملاحظه می‌شود که روش پیشنهادی، قابلیت تعمیم به حالت‌ها و تصاویر جدید را دارد.

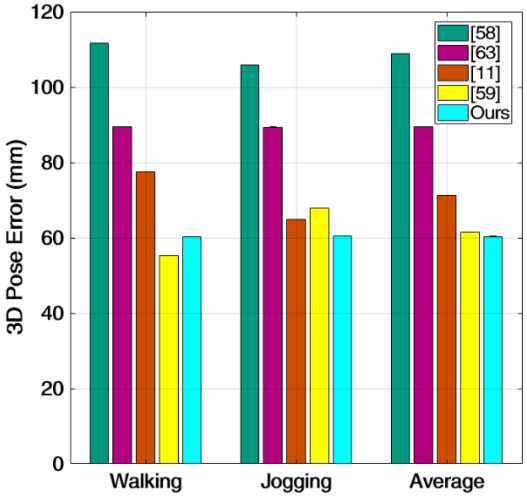
جدول (۴). خطای حالت سه بعدی در تصاویر آزمایشی پروتکل ۲ پایگاه داده‌ی Human3.6M

روش	خطای حالت سه بعدی (mm)
[۶۰]	۱۱۷/۹
[۵۵]	۱۱۵/۷
[۵۹]	۱۰۸/۳
[۳۲]	۸۸/۱
روش پیشنهادی	۷۸/۶

جدول (۵) نتایج ارزیابی بر روی پایگاه داده HumanEve [۶۱] را نشان می‌دهد. این پایگاه داده شامل ویدئوهایی از انجام ۶ فعالیت، توسط ۴ فرد و در سه مرتبه اجرای نمایش می‌باشد. روش ارزیابی مانند پروتکل مورد استفاده در مراجع [۵۵, ۵۹, ۶۲] است. به عبارتی دیگر، روش پیشنهادی در هر ۵ فریم از فریم‌های اعتبارسنجی در دنباله‌های ویدئویی walking (A1) و jogging (A2) برای هر سه فرد S1، S2 و S3 از دوربین C1 بررسی شده است. در شکل (۹)، مقدار خطا به صورت نمودار برای مقایسه بهتر به نمایش درآمده است. روش پیشنهادی به صورت میانگین به خطای کمتری در تخمین حالت سه بعدی دست یافته است.

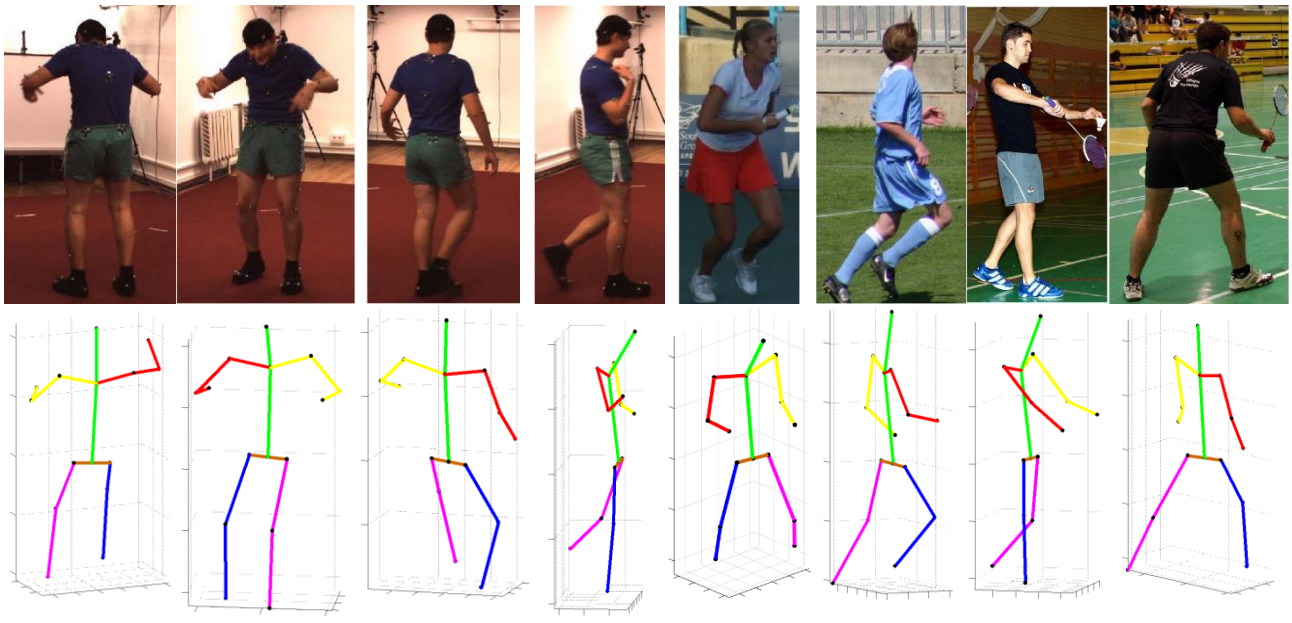
جدول (۵). خطای حالت سه بعدی در تصاویر آزمایشی پایگاه داده‌ی HumanEva-I

روش	Walking (A1, C1)			Jogging (A2, C1)			میانگین (mm)
	S1	S2	S3	S1	S2	S3	
[۵۸]	۹۹/۶	۱۰۸/۳	۱۲۷/۴	۱۰۹/۲	۹۳/۱	۱۱۵/۸	۱۰۸/۹
[۶۳]	۷۵/۱	۹۹/۸	۹۳/۸	۷۹/۲	۸۹/۸	۹۹/۴	۸۹/۵
[۱۱]	۷۱/۹	۷۵/۷	۸۵/۳	۶۲/۶	۷۷/۷	۵۴/۴	۷۱/۳
[۵۹]	۵۲/۲	۵۱/۰	۶۲/۸	۷۴/۵	۷۲/۴	۵۶/۸	۶۱/۶
روش پیشنهادی	۵۶/۲	۵۸/۰	۶۶/۸	۶۱/۲	۵۹/۶	۶۱/۰	۶۰/۴



شکل (۹). خطای سه بعدی در پایگاه داده HumanEva-I

شکل (۱۰) نمونه‌هایی از تخمین حالت سه بعدی را به منظور ارزیابی کیفی در برخی تصاویر آزمایشی پایگاه داده Human3.6M و نمونه‌هایی از تصاویر پایگاه داده‌ی LSP نمایش می‌دهد. به دلیل این‌که در پایگاه داده‌ی LSP، صرفاً حاشیه‌گذاری مفاصل دو بعدی وجود دارد، به دست آوردن خطای تخمین حالت سه بعدی ممکن نیست. روش پیشنهادی، الزامی برای وجود تصویر در شرایط آزمایشگاهی (مانند پایگاه داده Human3.6M) ندارد و می‌تواند در شرایط غیرآزمایشگاهی (مانند پایگاه داده‌ی LSP) که در آن افراد با لباس‌های متنوع در پس‌زمینه متفاوت و شلوغ و در زوایای دید گوناگون هستند، پاسخ مناسبی به دست دهد.



شکل (۱۰). نمونه‌هایی از تخمین حالت سه بعدی. چهار تصویر سمت راست، مربوط به پایگاه داده‌ی LSP و چهار تصویر سمت چپ، از تصاویر فرد S11 پایگاه داده Human3.6M است.

۷- نتیجه‌گیری

در این مقاله، روشی برای تخمین حالت سه بعدی انسان از تک تصویر پیشنهاد شده است. با توجه به این‌که منشاء اصلی خطا به هنگام ارتقاء حالت، از فضای دو بعدی به فضای سه بعدی، تخمین غیردقیق مکان‌های دو بعدی مفاصل است، این مقاله، با بهبود عملکرد تخمین حالت دو بعدی به تخمین دقیق‌تر حالت سه بعدی دست یافته است. روش پیشنهادی به منظور تخمین حالت دو بعدی با الهام گرفتن

از این‌که شکل انسان سرنخ‌های اصلی را برای پیش‌بینی حالت دو بعدی انسان به دست می‌دهد، از اطلاعات شکل انسان (نقشه احتمال لبه) برای هدایت شبکه‌ی عصبی عمیق بهره برده است. بنابراین، می‌توان این‌گونه بیان کرد که روش پیشنهادی ترکیبی از روش‌های سنتی (طراحی شده) و روش‌های نوینِ (یادگیری شده) استخراج ویژگی است.

٦- مراجع

[1] H. J. Lee and Z. Chen, "Determination of 3D human body postures from a single view," *Computer Vision, Graphics and Image Processing*, vol. 30, pp. 148-168, 1985.

[2] C. Barr´on and I. A. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image," *Computer Vision and Image Understanding*, vol. 81, pp. 269-284, 2001.

[3] V. Parameswaran and R. Chellappa, "View independent human body pose estimation from a single perspective image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. II–16.

[4] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000, pp. 677–684.

[5] H. Jiang, "3D human pose reconstruction using millions of exemplars," presented at the International Conference on Pattern Recognition (ICPR), 2010.

[6] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[7] C.-H. Chen and D. Ramanan, "3D human pose estimation = 2D pose estimation + matching," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," *European Conference on Computer Vision (ECCV)*, 2012.

[9] X. Fan, K. Zheng, Y. Zhou, and S. Wang, "Pose locality constrained representation for 3D human pose reconstruction," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 174–188.

[10] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3D human pose and shape from a single image," *European Conference on Computer Vision (ECCV)*, 2016.

[13] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1648-1661, 2016.

[14] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4966-4975, 2016.

[15] G. Shakhnarovich, P. A. Viola, and T. J. Darrell, "Fast pose estimation with parameter-sensitive hashing," presented at the IEEE International Conference on Computer Vision (ICCV), 2003.

[16] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression,"

presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004.

[17] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1052-1062, 2006.

[18] L. F. Bo, C. Sminchisescu, A. Kanaujia, and D. N. Metaxas, "Fast algorithms for large scale conditional 3D prediction," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[19] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 44–58, 2006.

[20] C. H. Ek, P. H. S. Torr, and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," in *International Workshop on Machine Learning for Multimodal Interaction*, 2007.

[21] L. Sigal, R. Memisevic, and D. Fleet, "Shared kernel information embedding for discriminative inference," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[22] A. Krizhevsky ,I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114.

[23] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Neural Information Processing Systems (NIPS)*, 2014.

[24] A. Toshev and C. Szegedy, "DeepPose: human pose estimation via deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[25] 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 (ECCV)* Cham, 2016, pp. 34-50.

[26] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] F. Shamsafar and H. Ebrahimnezhad, "Understanding holistic human pose using class-specific convolutional neural network," *Multimedia Tools and Applications,* 2018.

[28] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," *arXiv preprint arXiv:1611.09010*, 2016.

[29] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[30] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[31] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3D human pose with deep neural networks," presented at the British Machine Vision Conference (BMVC), 2016.

[32] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3D pose estimation in the wild," presented at the Advances in Neural Information Processing Systems (NIPS), 2016.

[33] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, *et al.*, "Learning from synthetic humans," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[34] R. Zhao, Y. Wang, and A. M. Martinez" ,A Simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[35] Hashim Yasin, Björn Krüger, and A. Weber, "Model based full body human motion reconstruction from video data," in *International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications*, 2013.

[36] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," presented at the European Conference on Computer Vision Workshops, 2016.

[37] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Fusing 2D uncertainty and 3D cues for monocular body pose estimation," *arXiv preprint arXiv:1611.05708*, 2016.

[38] A. Popa, M. Zanfir, and C. Sminchisescu, "Deep multitask architecture for integrated 2D and 3D human sensing," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei" ,Deep kinematic pose regression," presented at the Computer Vision–ECCV 2016 Workshops, 2016.

[40] Y. a. W. Du, Yongkang and Liu, Yonghao and Han, Feilin and Gui, Yilin and Wang, Zhen and Kankanhalli, Mohan and Geng, Weidong, "Marker-less 3D human motion capture with monocular image sequence and height-maps," in *European Conference on Computer Vision*, 2016, pp. 20-36.

[41] M. F. Ghezelghieh, R. Kasturi, and S. A Sarkar, "Learning camera viewpoint using CNN to improve 3D body pose estimation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 685-693.

[42] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3D human pose estimation.," *International Journal of Computer Vision*, vol. 122, pp. 149-168, 2017.

[43] B. T. A. R. V. L. P. Fua, "Direct prediction of 3D body poses from motion compensated sequences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 991-1000.

[44] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision (ACCV)*, 2014, pp. 332–347.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, pp. 211-252, 2015.

[46] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," presented at the British Machine Vision Conference (BMVC), 2010.

[47] P. Doll´ar and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 37, pp. 1558-1570, 2015.

[48] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, 2014.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[50] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[51] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[52] "CMU MoCap: Carnegie Mellon University Graphics Lab Motion Capture Database," C. M. U. G. Lab, Ed., ed, 2014.

[53] M. Zolfaghari, A. Jourabloo, S. G. Gozlou, B. Pedrood, and M. T. Manzuri-Shalmani, "3D human pose estimation from image using couple sparse coding," *Machine Vision and Applications*, vol. 25, pp. 1489-1499, 2014.

[54]F. Shamsafar and H. Ebrahimnezhad, "3D Human pose estimation from 2D images using fuzzy c-means," in *Pattern Recognition and Image Analysis (IPRIA2015)*, 2015.

[55] I. Kostrikov and J. Gall, "Depth sweep regression forests for estimating 3D human pose from images," presented at the British Machine Vision Conference (BMVC), 2014.

[56] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," presented at the Advances in Neural Information Processing Systems (NIPS), 2014.

[57]W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single image 3D human pose estimation from noisy observations," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[59] U. I. H. Yasin, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[60] L. Bo and C .Sminchisescu, "Twin gaussian processes for structured prediction," *International Journal of Computer Vision*, 2010.

[61] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, 2010.

[62] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer., "A joint model for 2D and 3D pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[63] I. Radwan, A. Dhall, and R. Goecke, "Monocular image 3D human pose estimation under self-occlusion," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.