مجله ماشین‌بینایی و پردازش‌تصویر
سال ششم، شماره دوم، زمستان ۱۳۹۸

تخمین حالت سه‌بعدی بدن انسان از یک تصویر بوسیله شبکه عصبی کانولوشن و کدگذاری و بازنمایی تنک با رویکرد مبتنی بر مدل

حسن علی‌کرمی[۱]، فرزین یغمایی[۲] و محمدجواد فدایی اسلام[۳]

چکیده

در زمینه تخمین و ساخت اسکلت حالت سه‌بعدی بدن انسان از طریق بندهای بدن (body joints) بوسیله یک تصویر دوبعدی، چالش‌های عمق و خودانسدادی وجود دارد که مانع از تخمین دقیق می‌گردد. در این مقاله به تخمین حالت سه‌بعدی بدن انسان با دو رویکرد مختلف پرداخته شده است. بدین منظور، رویکرد اول پیشنهادی با تمرکز بر عمق حالت دوبعدی حقیقت اصلی بوسیله کدگذاری و بازنمایی تنک و تصحیح‌گر مبتنی بر مدل، حالت سه‌بعدی بدن انسان استخراج می‌شود. در رویکرد دوم پیشنهادی به‌کمک روش مبتنی بر یادگیری شبکه‌های عصبی کانولوشن، تخمین حالت دوبعدی بدن انسان بدست می‌آید، سپس بوسیله کدگذاری و بازنمایی تنک و تصحیح‌گر مبتنی بر مدل، تخمین عمق حالت استخراج می‌شود. نتایج حاصل از این روش، برتری تخمین حالت و عمق سه‌بعدی بدن انسان را نسبت به رویکردهای پیشین نشان می‌دهد. تخمین حالت‌های سه‌بعدی انجام شده در روش پیشنهادی نشان می‌دهد میانگین خطای بازسازی نسبت به کارهای مشابه کاهش قابل توجهی داشته است.

کلیدواژه‌ها

شبکه‌های عصبی کانولوشنی، کدگذاری و بازنمایی تنک، اسکلت حالت سه‌بعدی بدن انسان، تخمین حالت سه‌بعدی

۱ مقدمه

امروزه با پیشرفت علوم مختلف از جمله هوش مصنوعی، تعامل بین انسان و ماشین امکان‌پذیر شده است. از مهم‌ترین چالش‌های موجود، ایجاد تعامل بصری بین انسان و ماشین است که نیازمند درک دقیق صحنه با تشخیص و تحلیل رفتار اجسام می‌باشد. از این رو پژوهشگران حوزه بینایی ماشین برای برقراری ارتباط یا عمل متقابل نسبت به دنیای واقعی، نیازمند تشخیص و تحلیل اجسام در تصویر و ویدیو هستند. این تشخیص و تحلیل با انگیزه‌های متفاوتی صورت گرفته‌است. برای مثال، هدف از ردیابی حرکت یک جسم، محدود کردن صحنه، فشرده‌سازی و ویرایش تصاویر است که متفاوت از تشخیص نوع حرکت جسم در صحنه هستند.

یکی از اهداف اصلی در تشخیص و تحلیل اجسام، درک صحنه با تحلیل موقعیت و جهت اجسام صلب[۱] است. در واقع پژوهشگران سعی دارند با تعریف مختصات یک جسم، قابلیت درک و توصیف آن را مقدور سازند تا از این طریق، موقعیت و جهت یک جسم در فضای دوبعدی و سه‌بعدی نسبت به مختصات جهانی را تعیین کنند. این مبحث در بینایی ماشین تحت عنوان

این مقاله در شهریورماه ۱۳۹۶ دریافت، در مردادماه ۱۳۹۷ دومین بازنگری و در آذرماه پذیرفته شد.

[۱] دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه سمنان
رایانامه: Hassan_Alikarami@semnan.ac.ir
[۲] دانشگاه سمنان، دانشکده مهندسی برق و کامپیوتر
رایانامه: F_Yaghmaee@semnan.ac.ir
[۳] دانشگاه سمنان، دانشکده مهندسی برق و کامپیوتر
رایانامه: Fadaei@semnan.ac.ir

نویسنده مسئول: فرزین یغمایی

[1] Rigid

"تخمین حالت[1]" شناخته شده است. از آنجاکه بدن انسان یک جسم صلب محسوب نمی‌شود، این عمل به‌صورت نمایش حالت بدن از طریق بندهای آن انجام می‌شود. از این رو می‌توان فرض کرد هر کدام از بندهای بدن انسان یک جسم صلب است و هدف از تشخیص حالت بدن انسان فقط حالت اسکلت است، لذا بدلیل وابستگی شکل بدن به عضلات و ماهیچه‌ها، نیازی به تشخیص شکل بدن نمی‌باشد. این مسئله در دهه اخیر به‌دلیل کاربردهای گسترده در صنعت فیلم و بازی سازی، سیستم‌های نظارتی، تجهیزات پزشکی و ... در بینایی ماشین اهمیت بسیاری یافته است.

روش‌های کاربردی موجود، مبتنی بر آنالیز حرکتی اجسام توسط سنسورهای عمق‌یاب یا تصاویر چنددیدی هستند که به‌دلیل نیاز به فضای مخصوص و هزینه بالا، لازم است روش‌هایی جهت تشخیص حالت دوبعدی و سه‌بعدی با استفاده از یک تصویر ارائه گردند. در این رویکرد، هدف تخمین حالت دو بعدی و محاسبه‌ی عمق حالت بدن انسان بر پایه استفاده از یک تصویر با کمترین خطا نسبت به حالت سه‌بعدی حقیقت اصلی[2] است.

جهت نزدیک شدن به مکانیسم مغز انسان از ایده یادگیری عمق از یک تصویر استفاده می‌شود زیرا، گرچه تحلیل تصاویر در مغز انسان براساس دو تصویر از یک مشاهده صورت می‌پذیرد، اما به‌دلیل یادگیری و تجربیات مشاهدات قبلی، انسان قادر به تشخیص عمق اجسام از طریق مشاهدات تک‌دیدی نیز می‌باشد. لذا از چالش‌های مهم در این حوزه، تخمین حالت دوبعدی جهت غلبه بر مشکل خود انسدادی[3] است. برای مثال در مقاله [۱] یک چارچوب محاسباتی مبتنی بر مدل جهت بخش‌بندی و تشخیص اشیا ارائه شده است. ایده اصلی این مقاله، نمایش اشیا توسط مجموعه‌ای از قطعات مرتب شده در یک پیکربندی انعطاف‌پذیر است. آن‌ها یک چارچوب آماری برای توصیف ظاهری اشیا در نظر گرفته و از مدل‌های ساختاری تعریف شده جهت تطابق با حالت فرد یا شی استفاده نمودند. بهترین تطابق برای مدل تصویر به‌گونه‌ای در نظر گرفته شده است که تابع انرژی را حداقل نماید. مقاله [۲] با یک نمایش جدید از توصیف بخش‌های مدل، بندها را به‌صورت بلوک‌های مخلوطی ترکیبی در نظر گرفتند. در این روش از مدل‌های مخلوط انعطاف‌پذیر برای روابط بین بخش‌ها و افزایش مدل‌هایی برای توصیف روابط فضایی استفاده کردند. بطور کلی، این رویکردها به‌دلیل دقت پایین و محاسبات زیاد، کمتر مورد توجه قرار گرفته است.

از سوی دیگر، مقاله [۳] برای کاهش بار محاسباتی و دقت تشخیص، یک چارچوب ساختاری تصویر برای حل مشکل تخمین کارآمد از بخش‌های مختلف بدن، از رگرسیون غیرخطی استفاده کرده است. آن‌ها دو لایه رگرسیون تشکیل دادند که لایه اول به عنوان یک بخش متمایزکننده برای طبقه‌بندی قطعه‌های بدن است و لایه دوم با استفاده از تخمین لایه اول، قادر به پیش‌بینی مفصل‌ها توسط مدل‌سازی از طریق وابستگی‌های مفاصل می‌باشد. این روش نشان داد عملکرد محلی‌سازی مفاصل بهتر از ساختار درختی است.

در سال‌های اخیر، متخصصین به تخمین حالت دو بعدی بدن انسان با استفاده از شبکه‌های عصبی کانولوشن پرداخته‌اند. این رویکرد برای اولین‌بار، سال ۲۰۱۳ در مقاله [۴] ارائه شد که شبکه آن مشابه Alex Net [۵] است با این تفاوت که در لایه آخر از یک لایه رگرسیون استفاده شده است. دقت بالای حاصل از این رویکرد موجب جلب توجه پژوهشگران این حوزه به شبکه‌های عصبی کانولوشن گردید. در مقاله [۶] ایده‌ای مشابه [۴] به صورت یک روش بازخوردی بالا به پایین ارائه شد. در این روش با تکرار شبکه عصبی کانولوشن خود اصلاح، خطای تخمین حالت دوبعدی انسان کاهش یافت.

در مقاله [۷] سعی شده است اطلاعات ناشی از دیدگاه کلی و بخش‌های محلی نسبت به یک تصویر، بدست آمده از دو منبع شبکه عصبی کانولوشن، یکپارچه‌سازی شود. اما مقاله [۸] روش چند مرحله‌ای از شبکه‌های عصبی کانولوشن برای تخمین حالت بدن معرفی کرد که در هر مرحله قسمتی از حالت بدن تخمین زده می‌شود. بنابراین، در تشخیص بندهای بدن انسان با خطای کمتری مواجه می‌شوند. این شبکه به‌دلیل بروز ریسک خطا در مواجهه با شیب کاهشی یا محوکنندگی در طول آموزش، از یک لایه میانی بعد از هر مرحله برای اجرا استفاده شده است.

در مقاله [۹] روشی پایان به پایان[4] برای تخمین حالت انسان پیشنهاد شده است که به‌صورت متوالی، با استخراج روابط میان ویژگی‌های نقشه مفاصل توسط شبکه عصبی کانولوشن و ترکیب مدل‌سازی ساختاری و یادگیری ویژگی بر مبنای نقشه مفصل و اصلاح آن از طریق گراف، سعی در انتخاب بهترین حالت دارد.

در موارد بالا، هدف تخمین حالت دوبعدی بدن است، اما بدن انسان یک سیستم بسیار پیچیده متشکل از اندام‌ها و مفاصل مختلف است، لذا تخمین واقعی از موقعیت مفاصل در حالت سه‌بعدی، حتی برای مغز انسان هم چالش برانگیز است. از این رو، مقاله [۱۰] سعی در حل مسائل تخمین حالت بدن انسان و بخش‌بندی آن با استفاده از دسته‌بندی فضای فرضیه داشته است. آن‌ها با تهیه یک مجموعه‌دادگان جدید حاوی تصاویر دارای حاشیه‌نویسی و اطلاعات مفاصل سه‌بعدی بدن در یک تصویر دوبعدی، سعی در پیاده‌سازی یک روش جستجوی داده‌محور به‌منظور یافتن فضای حالت داشته‌اند. از این رو به‌صورت جدا، قسمت‌های بدن را در کلاس‌هایی دسته‌بندی کرده و از طریق ماشین‌بردار پشتیبان، به تشخیص محتوا پرداخته‌اند.

[1] Pose Estimation
[2] Ground Truth
[3] Self Occlusion
[4] End To End

با وجود چالش‌های ابهام در عمق و خودانسدادی در این حوزه، می‌توان گفت تکنیک‌های خودکار، جایگزین مناسبی برای حل این مسئله هستند. چرا که روش مبتنی بر مدل، پس از ساخت مدل بدن منطبق بر بهترین مدل موجود، با اعمال محدودیت‌های اجرایی بر هر پارامتر، به‌عنوان مثال، رعایت نسبت طول بدن، اندام و زوایای مفاصل، سعی در تخمین حالت بدن انسان دارد [۱۱].

گام اصلی در مطالعه حرکت سه‌بعدی انسان، استخراج ویژگی دقیق از سیگنال‌های ورودی است. روش‌های اولیه از ویژگی‌های سطح پایینی استفاده می‌کردند که جهت کاهش ابعاد بردار ویژگی‌ها، از توصیف‌گرهای تصاویر بهره می‌بردند.

در مقاله [۱۲] پس از نرمال سازی مجموعه‌دادگان حالت‌های سه‌بعدی، در مرحله اول، براساس نوع حالت سه‌بعدی، حالات دوبعدی مربوطه از دیدهای مختلف، استخراج شده است. در مرحله دوم، با استفاده از حالت دوبعدی استخراج شده از تصویر ورودی، بوسیله ساختاری درختی، سعی در اصلاح آن از طریق مقایسه با حالات استخراجی مرحله اول براساس روش k نزدیک‌ترین همسایه، صورت گرفته است و در نهایت تصمیم‌گیری با به حداقل رساندن خطای طرح نهایی انجام شد.

در این حوزه جهت دستیابی به تخمین دقیق‌تر، روش‌های مختلفی بر مبنای استخراج عمق دقیق پیشنهاد شده است. اما نتایج تحقیقات پژوهشگران نشادن داده است که روش‌های یادگیری عمیق و کدگذاری و بازنمایی تنک، از دقت و کارایی بالاتری برخوردارند. به‌عبارتی، پس از موفقیت روش‌های مبتنی بر تکنیک‌های یادگیری عمیق جهت تخمین حالت دو بعدی، از این تکنیک در حوزه تخمین حالت سه‌بعدی نیز استفاده شد. از این رو در مقاله [۱۳] تخمین حالت، براساس رگرسیون یک فریم انجام شد. اما در مقاله [۱۴] آموزش شبکه های عصبی کانولوشن با دو نگرش صورت گرفته است. در نگرش اول، بصورت همزمان شبکه جهت تخمین حالت بدن بوسیله‌ی رگرسیون و نیز جهت تشخیص اشیا، آموزش انجام شده است. در نگرش دوم، ابتدا شبکه بصورت مستقل جهت تشخیص اشیا آموزش دیده و سپس در مرحله‌ی بعد بوسیله‌ی رگرسیون، تخمینی از حالت بدن انسان بدست می‌آید. آن‌ها نشان دادند که شبکه در آخرین لایه‌های خود، شامل نمایش داخلی مناسبی از ساختار اسکلت و همبستگی بین متغیرهای خروجی جهت تعیین حالت بدن است.

مقاله [۱۵] از یک چارچوب شامل یک تصویر و حالت سه‌بعدی به همراه ارزش همخوانی مربوطه به عنوان ورودی استفاده می کند. در این رویکرد، شبکه عصبی کانولوشن همزمان وظیفه‌ی انجام تبدیلات غیرخطی از تصویر و ترکیب‌بندهای تخمین‌زده شده و ساخت حالت بدن را بر عهده دارد.

از طرفی مقاله [۱۶] با کمک یک چارچوب بیزی و براساس کدگذاری و بازنمایی تنک، واژه‌نامه‌ای از نمونه‌های حالت ایجاد کرده است و به تشخیص شبیه‌ترین حالت نسبت به تصویر ورودی می‌پردازد. مقاله [۱۷] یک روش نیمه نظارتی کدگذاری تنک برای تخمین حالت سه‌بعدی ارائه داد. در این روش با استفاده از الگوریتم محلی کوانتوم خطی، داده‌های بدون برچسب به مدل پیشنهادی وارد می‌شود. از این رو در مقاله [۱۸] با روش شبکه های عصبی کانولوشن، تخمین حالت دوبعدی انسان انجام شد و از طریق واژه‌نامه ساخته شده شامل تمام حالت‌های سه‌بعدی، سعی در تخمین حالت سه‌بعدی یک تصویر بوسیله روش KNN دارد. مقاله [۱۹] بر اساس چارچوبی بر پایه شبکه عصبی کانولوشن DeepCut [۲۰] جهت تخمین حالت دوبعدی از رویکردی بالا به پایین استفاده کرده است. در این مقاله پس از تخمین حالت دو بعدی، از میان مجموعه‌دادگان سه‌بعدی شکل[1]، نمونه‌ای با کمترین اختلاف شباهت با این تخمین، انتخاب شد. این روش برای مجموعه‌دادگان سه‌بعدی کوچک نیز مناسب است. اما مقاله [۲۱] شبکه‌ای عصبی کانولوشنی با هدف تخمین حالت دوبعدی و سه‌بعدی بصورت مشترک ارائه داد. در این شبکه وزن-های آخرین لایه تمام اتصال تخمین حالت دوبعدی به‌عنوان بخشی از لایه تمام اتصال تخمین حالت سه‌بعدی قرار گرفته است.

مقاله [۲۲] با شبکه عصبی کانولوشن، اطلاعات حرکت را در دنباله ای کوتاه بررسی و تخمین حالت سه‌بعدی انجام داده است. اما مقاله [۲۳] نقشه ارتفاع[2] را از تصویر دوبعدی استخراج کرده است. سپس بوسیله شبکه‌های عصبی کانولوشنی، تخمینی از حالت دوبعدی بدن انسان استخراج شده است و توسط نقشه عمق اصلاح می‌گردد. در ادامه با ساخت واژه‌نامه از حالت‌های سه‌بعدی، با الگوریتم PCA به استخراج حالت سه‌بعدی می‌پردازد. از این رو مقاله [۲۴] توسط شبکه‌های عصبی کانولوشن و کدگذاری تنک به تخمین حالت سه‌بعدی می‌پردازد. در این مقاله تخمین حالت دوبعدی توسط ترکیب نتایج دوشبکه عصبی کانولوشن انجام می‌گیرد سپس در ادامه توسط کدگذاری و بازنمایی تنک به تخصیص بهترین حالت سه‌بعدی نسبت به حالت دوبعدی تخمین زده شده می‌پردازد. از طرفی مقاله [۲۵] برای بازسازی حالت سه‌بعدی از یک شبکه عصبی کانولوشن استفاده کرده است. در این روش، شبکه عصبی کانولوشن برای تخمین حالت دوبعدی بدن انسان و زاویه بین بندها در مرحله لایه تمام اتصال و در نهایت استخراج حالت سه‌بعدی بدن انسان، آموزش انجام شده است.

استخراج حالت سه‌بعدی انسان از یک تصویر دوبعدی، با چالش‌های مختلفی روبرو است. از این رو در برخی مقالات، هدف تخمین حالت سه‌بعدی انسان با فرض وجود حقیقت اصلی حالت دو بعدی انجام گرفته‌است. در مقاله [۲۶]، منطبق بر حقیقت اصلی حالت دوبعدی، مشابه‌ترین حالات سه‌بعدی موجود در مجموعه‌دادگان تعیین می‌شود. در ادامه، با اعمال شروط حرکت‌شناسی حالت انسان، دقیق‌ترین حالت سه‌بعدی تخمین زده شد. از طرفی نویسندگان مقاله [۲۷] با دو واژه‌نامه مختص به

[1] Shape
[2] Height Map

داده‌های ورودی سایه‌نما و حالت‌های سه‌بعدی حقیقت اصلی، تخمینی از اطلاعات ورودی برحسب ترکیب خطی مولفه‌های واژه‌نامه اول استخراج شد، سپس با اعمال این ترکیب خطی در واژه‌نامه دوم، تخمین حالت سه‌بعدی انجام شد.

در مقاله[۲۸] با استفاده از کدگذاری و بازنمایی تنک، از رابطه‌های محدب به‌منظور ساخت نزدیک‌ترین مدل از ترکیب مولفه های واژه‌نامه استفاده شد. در این روش از الگوریتم ADMM با متغیر کمکی استفاده شده است و بصورت بازسازی مستقیم از خروجی الگوریتم ADMM، طرح نهایی تشکیل شد. لذا با توجه به خطای بازسازی مستقیم طرح سه‌بعدی که ناشی از تخمین نادرست وزن‌های مولفه‌های ترکیبی و ماتریس چرخش[1] است، مقاله [۲۹] سعی در بهبود این روش داشتند. آن‌ها وزن و ماتریس چرخش را بر اساس یک معادله نرم Frobenius، جایگزین مرحله بازسازی مستقیم کرده و با حل و بروز رسانی آن از طریق الگوریتم Altern، سعی در بازسازی مناسبی برای مدل بودند. هدف این مقاله استفاده از کدگذاری و بازنمایی تنک به‌منظور بهبود دقت تخمین بوده است اما متاسفانه نتایج تخمین حالت سه‌بعدی، بهبود ملموسی را نشان نمی‌دهد. با این حال، این روش در تخمین عمق از دقت بالاتری برخوردار است.

در این مقاله، روشی جدید مبتنی بر شبکه های عصبی کانولوشن و کدگذاری و بازنمایی تنک با تصحیح‌گر مبتنی بر مدل، جهت تخمین حالت سه بعدی انسان از یک تصویر ارائه شده است. بخش های مختلف به این صورت سازماندهی می شود: در بخش دوم، مفاهیم اولیه بیان شده است. در بخش سوم، روش پیشنهادی بصورت چند مرحله ای بیان شده است. در بخش چهارم نتایج تجربی و مقایسه با سایر مقالات بیان شده است. بخش پنجم نیز حاوی نتیجه‌گیری می‌باشد.

۲ مفاهیم اولیه

۲-۱ ارتباط حالت دوبعدی و سه‌بعدی

تبدیل تصویر سه‌بعدی به دوبعدی، موجب از دست رفتن ابعاد و اطلاعات عمق فضا طی یک نگاشت یک به یک می شود. گرچه استفاده از تصاویر دوبعدی چندیدی امکان بازسازی تصویر و تخمین مناسب با استفاده از عمق استخراجی را ممکن می‌سازد، اما تخمین حالت سه‌بعدی بدن انسان از تنها یک تصویر، بصورت بازسازی یک پیکربندی نقاط دلخواه سه‌بعدی بوسیله‌ی یک تصویر تک‌چشمی RGB است. این موضوع باعث تطابق حالت‌های سه‌بعدی متفاوت با حالت دوبعدی می‌شود. لذا خطای جزئی در حالت دوبعدی می‌تواند تأثیر بالایی بر تخمین حالت سه‌بعدی داشته باشد[۳۰]. در شکل ۱ نمونه‌ای از تصویر دوبعدی با حالت دوبعدی آن نمایش داده شده است.

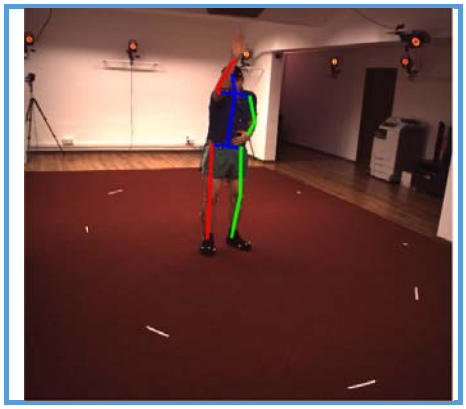
به‌طور کلی تخمین حالت سه‌بعدی با رویکردهای مبتنی بر مدل، مبتنی بر یادگیری و مبتنی بر نمونه قابل انجام است. در روش مبتنی بر مدل، هدف تطابق مدل اولیه سه‌بعدی در نظر گرفته شده با تصویر دوبعدی شخص است، به‌گونه‌ای که با ایجاد تغییرات متناسب با تصویر دوبعدی، مدل سه‌بعدی با هدف به حداقل رساندن مقدار خطا در تطابق دوبعدی نسبت به تصویر تغییر داده شود. پس از اعمال تغییرات می‌توان مدل سه‌بعدی را تخمینی از حالت دوبعدی بدن انسان در نظر گرفت. اما روش مبتنی بر مدل به‌دلیل زمان‌بر بودن در مراحل مختلف، کمتر مورد توجه قرار می‌گیرد. روش مبتنی بر یادگیری، سعی در یادگیری حالت‌های ورودی در مرحله آموزش نسبت به تصاویر ورودی دارد تا با توجه به ویژگی‌ها و الگوهای موجود در تصویر، حالت‌های مفاصل را تشخیص دهد. روش مبتنی بر یادگیری دارای سرعت بالایی می‌باشد اما با توجه به گسترده بودن فضای حالت بدن، نمی‌توان تمام حالت‌ها را با تخمینی مناسب بازسازی کرد. بر همین اساس روش مبتنی بر نمونه، حالت‌های متفاوت را در حافظه ذخیره کرده و با جستجو در بین آن‌ها، مشابه‌ترین حالت‌ها را انتخاب می‌کند. اما این روش به‌دلیل نیاز به حجم بالای حافظه و تامین انواع حالت‌ها برای تخمین حالت مناسب، هزینه‌بر خواهد بود.

در رویکرد مبتنی بر نمونه، با در نظر گرفتن پارامتر P،بعنوان تعداد موقعیت‌های مدل، B_i حالت پایه سه‌بعدی و C_i وزن تاثیرگذاری برای حالت شماره i، می‌توان حالت سه‌بعدی بدن را به صورت رابطه ۱ تعریف کرد[۲۸].

$$S = \sum_{i=1}^{k} C_i B_i \qquad (۱)$$

در این رابطه $S \in R^{3\times P}$ ، حالت استخراجی براساس رویکرد مبتنی بر نمونه می‌باشد. از طرفی، ارتباط بین تصاویر دوبعدی و سه‌بعدی با فرض $W \in R^{2\times P}$ حالت دوبعدی تبدیل یافته، $R \in R^{2\times 3}$ ماتریس چرخش متعامد و T بردار جابجایی، بصورت رابطه ۲ در نظر گرفته می‌شود.

$$W = RS + T1^T \qquad (۲)$$



شکل ۱: نمونه تصویر دوبعدی بدن انسان و حالت دوبعدی با ۱۵ بند.

[1] Rotation Matrix

اما به‌دلیل نرمال‌سازی اسکلت بدن انسان، از بردار جابجایی صرف نظر کرده و رابطه ۱ و ۲ بصورت رابطه ۳ بازنویسی می‌شود.

$$W = R\sum_{i=1}^{k} C_i B_i \qquad (۳)$$

در این مدل به دلیل نزدیک بودن بدن انسان نسبت به صحنه، ماتریس کالیبراسیون نادیده گرفته شده است[۲۸].

۲-۲ مقدمه‌ای بر شبکه عصبی کانولوشن

یادگیری عمیق[۳۱] به عنوان نسخه الهام گرفته شده از پرسپترون چند لایه (MLPS)، در زمینه‌های مختلف بینایی‌ماشین معرفی شده است. این مدل از یادگیری، تکنیکی مبتنی بر شبکه‌های عصبی است و روشی برای ارائه‌دادن یادگیری فرضیه‌ها [۳۲] می‌باشد که از تبدیلات غیرخطی چند منبعی تشکیل شده است. سلسه مراتب ویژگی‌ها در این مدل، از ویژگی‌های سطح پایین و تجمیع برای ویژگی‌های سطح بالاتر تشکیل شده است. در بکارگیری یادگیری عمیق، با توجه به روش و نوع معماری مورد استفاده، می‌توان از یادگیری بدون ناظر یا با ناظر و یا ترکیب این دو استفاده کرد [۳۳].

یکی از معروف‌ترین و موفق‌ترین معماری‌های یادگیری عمیق در حوزه آنالیز تصویر، شبکه عصبی کانولوشن است. علیرغم معرفی این شبکه‌ها در دهه ۸۰ [۱۰]، عمده گسترش آن‌ها در وظایف بینایی ماشین به‌دلیل در دسترس بودن منابع محاسباتی، در سال‌های اخیر صورت گرفته است. از اولین کاربردهای پیاده‌سازی شده در این زمینه می‌توان به شبکه Le Net [۱۱] اشاره کرد که جهت دسته‌بندی ارقام دست‌نویس در سال ۱۹۹۸ ارائه شد.

در سال‌های اخیر، معماری‌های متفاوتی متناسب با کاربردهای مختلف بر اساس شبکه‌های عصبی کانولوشنی ارائه شده است که غالبا متشکل از سه نوع لایه اصلی شامل لایه کانولوشن[۱]، لایه ادغام[۲] و لایه تماما متصل[۳] با ترکیب‌های متفاوت هستند. لایه کانولوشن برای محاسبه نقشه ویژگی، محاسبات را با فعال‌سازی یک نقشه با قابلیت تولید پاسخ فیلتر در هر موقعیت مکانی انجام می‌دهد. از لایه ادغام جهت کاهش بار محاسباتی لایه‌های کانولوشنی استفاده می‌شود. هدف از قرار دادن این لایه، دستیابی به فضایی تغییرناپذیر با کاهش رزولوشن نقشه ویژگی‌های مورد نظر است. از طرف دیگر لایه تماما متصل مسئول استدلال سطح بالا در شبکه‌های عصبی کانولوشن است. این لایه معمولا آخرین لایه از شبکه است و تمام نودهای لایه‌های قبلی را به‌عنوان ورودی پذیرفته و به یک نرون واحد متصل می‌کند. لذا پس از این لایه، تجسم و تحلیل اطلاعات مکانی داده‌ها به‌صورت یک‌بعدی امکان‌پذیر است.

۲-۳ مقدمه‌ای بر کدگذاری تنک

سیستم بینایی به‌عنوان یک سیستم کدگذاری به دنبال حذف وابستگی‌های آماری جهت توصیف تصاویر در قالب رویدادهای مستقل است. در بینایی ماشین نیز به‌منظور نمایش و ذخیره‌سازی داده‌های بزرگ از مکانیسم کدگذاری جهت تحلیل بهتر داده‌ها با حذف افزونگی‌های ناشی از وابستگی‌های آماری پیچیده استفاده می‌شود [۳۴]. از این رو محققین با استفاده از این سیستم‌های کدگذاری در سال‌های اخیر چالش‌های زیادی مانند تشخیص چهره [۳۵]، تشخیص حالت بدن انسان [۲۴ و ۳۶] و ... را حل کرده‌اند.

یکی از سیستم‌های کدگذاری معروف، کدگذاری تنک است. این کدگذار به‌دنبال بازسازی سیگنال‌ها در قالب یک ساختار تنک می‌باشد و داده ورودی را به‌صورت ترکیب خطی از مؤلفه‌های واژه‌نامه بیان می‌کند[۳۷]. بطور کلی هدف از این سیستم، حل چالش‌هایی با ابعاد بالای داده‌ها با مدل‌کردن مساله به کمک معادلات ریاضی می‌باشد. تنکی[۴] یک بردار با تعداد مؤلفه‌های صفر آن بردار تعیین می‌شود. در این کدگذاری، اگر بازنمایی بهینه در یک مسئله به اندازه کافی تنک باشد، آن مسئله با بهینه‌سازی محدب قابل حل خواهد بود. الگوریتم‌هایی که برای بدست آوردن جواب‌های تنک ارائه شده‌اند، همگی به دنبال توصیف تصویر ورودی y با ترکیب چند مولفه اولیه، از مجموعه مولفه‌های موجود در واژه‌نامه D با ضرایب A بصورت رابطه ۴ می‌باشد.

$$y = DA \qquad (۴)$$

با فرض $D=[D_1,D_2,...,D_n]$ و، $A=[A_1,A_2,...,A_n]$ این رابطه را نشان می‌دهد. کد گذاری تنک می‌تواند بصورت خطی رابطه (۴) را با نرم صفر کمینه کند[۳۸]. بنابراین مسئله بهینه‌سازی به‌صورت رابطه ۵ بازنویسی می‌شود.

$$\hat{A}\sum_{i=1}^{n} D_i^2 = \arg\min \| A \|_0 \quad s.t. \quad y = DA \qquad (۵)$$

که در آن $\|.\|_0$ نشان‌دهنده تعداد عناصر غیرصفر در بردار مربوطه می‌باشد. از این رو، اگر k مولفه از واژه‌نامه D، نمونه آزمایشی را توصیف کند رابطه بهینه‌سازی بصورت رابطه ۶ تبدیل می‌شود.

$$y = DA \quad s.t. \quad \| A \|_0 \le k \qquad (۶)$$

از طرفی برای حذف نویز و طبق قضیه ضرب‌کننده لاگرانژ[۵] با ضریب ثابت α که مقدار انقباض A را مشخص می‌کند، رابطه ۶ بصورت رابطه ۷ بازنویسی می‌شود:

$$\hat{A} = L(A,\alpha) = \arg\min \| y - DA\|_2^2 + \alpha\|A\|_0 \qquad (۷)$$

از الگوریتم‌های متداول در حل اینگونه مسائل بهینه‌سازی، MP[۶] است. این الگوریتم به صورت حریصانه و با در نظر گرفتن نرم صفر، سعی در پیدا کردن بهترین مولفه در واژه‌نامه را دارد. در

[1] Convolutional Layer
[2] Pooling Layer
[3] Fully Connected Layer
[4] Sparsity
[5] Lagrange Multiplier
[6] Matching Pursuit

ادامه، الگوریتم OMP[1] با متعامد[2] ساختن برای رسیدن به بهترین تخمین در هر تکرار، الگوریتم MP را بهبود داد[۳۷]. اما بسیاری از مسائل را می‌توان در چارچوب بهینه سازی محدب قرار داد. لذا الگوریتم های بهینه‌سازی موازی به‌عنوان مکانیسمی برای حل مسائل آماری در مقیاس بزرگ معرفی و روش‌های تقسیم پذیر را به فرم زیر تبدیل می‌کند.

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } Ax + Bz = c \end{aligned} \qquad (۷)$$

که $A \in \mathrm{R}^{p\times n}$ و $B \in \mathrm{R}^{p\times m}$ و با فرض $x \in \mathrm{R}^{n}$ ، $z \in \mathrm{R}^{m}$ و توابع f و g بصورت محدب می‌باشد. از الگوریتم‌هایی که این روش را شامل می‌شود ADMM[3] [۳۴] و relaxed ADMM [۳۹] می‌باشد که دارای بهترین تخمین در مسائل درجه دوم و مخروطی است. در نهایت باید گفت که هر یک از این روش‌های بهینه‌سازی، نقاط قوت و ضعفی دارند که با توجه به نوع و ماهیت دادگان و نیز شرایط مساله، مورد استفاده قرار می‌گیرند.

۳ روش پیشنهادی

مراحل کلی روش پیشنهادی جهت تخمین حالت سه بعدی بدن انسان از یک تصویر در شکل ۳ نمایش داده شده است. در ادامه با توجه به خلاصه گفته شده از شبکه عصبی کانولوشن و کدگذاری و بازنمایی تنک، روش پیشنهادی شرح داده خواهد شد.

۳-۱ تخمین حالت دوبعدی از تصویر بوسیله شبکه عصبی کانولوشن

یکی از معماری‌های با عملکرد مناسب در حوزه پردازش تصویر و استخراج ویژگی، معماری VGGNet [۴۰] می‌باشد. معماری اولیه این شبکه با ۱۶ لایه جهت طبقه‌بندی تصاویر، با یادگیری مجموعه‌دادگان ImageNet ارائه شده است [۴۰]. با این حال این معماری دارای پیکربندی‌های دیگری نیز می‌باشد که تنها تفاوت آن‌ها در عمق شبکه است و حداکثر ۱۹ لایه می‌باشند. این شبکه به‌دلیل توالی مناسب در لایه‌های کانولوشنی جهت برقراری همبستگی بین ویژگی‌ها و عمق بالا، موجب ارائه عملکرد مناسب شده است. مقاله [۴۱] نشان داد که این شبکه با کاهش تعداد پارامترها از طریق تغییر اندازه تصویر ورودی و حذف لایه ادغام در آخرین لایه، تخمین حالت دوبعدی بدن انسان را به خوبی انجام می‌دهد. از این رو در این مقاله با رویکرد مبتنی بر یادگیری، شبکه‌ای مشابه این معماری، برای تخمین حالت دوبعدی بدن انسان در نظر گرفته شده است. معماری پیشنهادی، شامل ۱۳ لایه کانولوشنی و ۳ لایه تماما متصل به‌صورت پیشرو منظم می‌باشد (شکل ۲).

این معماری با توجه به تعداد بالای تصاویر آموزشی، جهت کاهش تعداد پارامترها، ابعاد لایه ورودی از ۲۲۴×۲۲۴ به ۱۱۲×۱۱۲ تغییر یافته و به همین نسبت ابعاد تمام لایه‌های کانولوشنی کاهش پیدا کرده‌است. از سوی دیگر، بدلیل تفاوت تخمین حالت دوبعدی بدن انسان با طبقه‌بندی تصاویر، جهت ترکیب وزن‌ها به‌عنوان ورودی لایه‌های تماما متصل و استفاده از اطلاعات فضایی محلی، از پنجره ۷ در ۷ (بدون قاب گذاری) برای شبکه تماما متصل استفاده و به‌منظور تشخیص بندها، لایه ادغام آخر شبکه طبقه‌بندی VGGNet حذف می‌گردد.

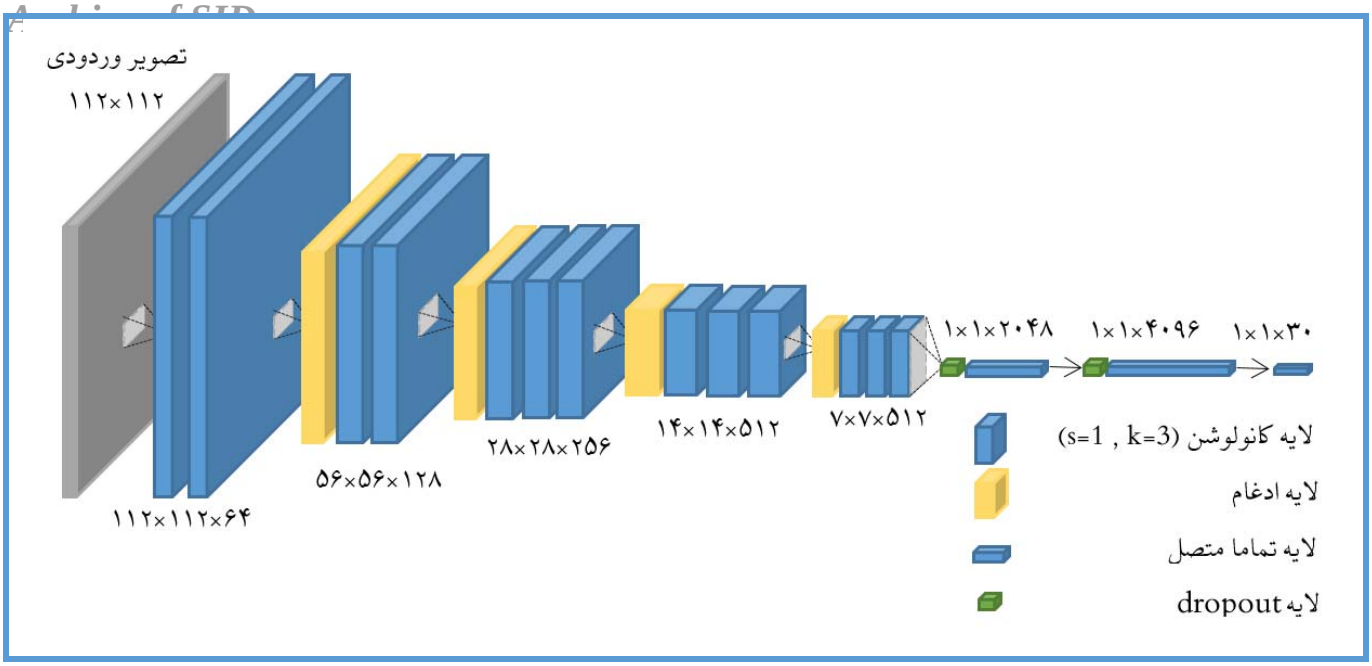
در معماری‌های شبکه عصبی کانولوشن، به‌دلیل تعداد بسیار زیاد پارامترهای استفاده شده، ممکن است بیش‌برازش[4] ایجاد شود. از این رو روش Dropout [۴۲ و ۴۳] به‌منظور جلوگیری از بروز این پدیده معرفی شده است. در هر مرحله از آموزش این روش، هر نورون با احتمال p-۱، از شبکه حذف[5] می‌شود بطوریکه بطوریکه نهایتا یک شبکه کاهش داده شده باقی بماند. در صورت حذف یک نود، یال‌های ورودی و خروجی آن نیز حذف می‌شود.

بدین ترتیب در هر مرحله داده‌ها فقط بر روی شبکه کاهش یافته آموزش خواهد دید. پس از هر مرحله، نودهای حذف شده به همراه وزن‌های سابق آن‌ها (قبل از حذف شدن) دوباره به درون شبکه وارد می‌شوند. از این رو در معماری به‌منظور جلوگیری از ایجاد بیش‌برازش شبکه، قبل از هر لایه تماما متصل، یک لایه dropout با نرخ ۰.۲۵ قرار گرفته است.

با اعمال تصاویر آموزشی و حالت‌های دوبعدی اسکلت‌بندی بدن انسان به‌عنوان ورودی شبکه، وزن‌ها جهت تشخیص مناسب بندهای بدن، اصلاح شده است. همچنین در لایه‌های تماما متصل، لایه اول از ابعاد ۴۰۹۶ به ۲۰۴۸ و بردار ویژگی با ابعاد ۴۰۹۶×۱ به ۳۰×۱ جهت مشخص نمودن ۱۵ بند اسکلت‌بندی و در انتها از کاهشی آنتروپی[6] برای شبکه استفاده شده است.

در روش پیشنهادی، به‌دلیل یادگیری و تخمین مناسب شبکه عصبی کانولوشن با ابعاد پایین‌تر، همانند پژوهش‌های پیشین از تصاویر برش داده شده براساس قاب محتوای اصلی حالت دوبعدی بدن انسان در تصویر استفاده شده است و محتوای فاقد اطلاعات حذف گردیده است. به بیان دیگر، به‌دلیل تجهیزات پایین و تعداد زیاد تصاویر مجموعه‌دادگان جهت آموزش شبکه‌های عصبی کانولوشنی با پارامتر بالا، از رویکرد برش محتوای اصلی استفاده شده است و منجر به یادگیری و تخمین حالت دقیق‌تری از حالت دوبعدی بدن انسان در شبکه‌های عصبی کانولوشن با پارامترهای بسیار پایین‌تر می‌شود.

[1] Orthogonal Matching Pursuit
[2] Orthogonalization
[3] Alternating Direction Method of Multipliers
[4] Overfitting
[5] Dropped Out
[6] Cross Entropy Loss



شکل ۲: معماری شبکه عصبی کانولوشن روش پیشنهادی

بطورکلی این شبکه با توجه به کاهش ابعاد لایه‌ها به نصف شبکه اولیه ۱۶ لایه VGG، تقلیل ابعاد بردار ویژگی اول از ۴۰۹۶ به ۲۰۴۸ در قسمت تماما متصل و کاهش ابعاد دسته‌بند انتهایی از ۱۰۰۰ به ۳۰، تعداد پارامتر شبکه نسبت به شبکه VGG[۴۰]، ۷۴ میلیون پارامتر و نسبت به شبکه مقاله [۴۱]، ۵۹ میلیون پارامتر کاهش داشته است و دارای ۶۴ میلیون پارامتر است. آموزش این شبکه عصبی کانولوشنی با نرخ یادگیری ۰/۰۰۰۱، بر روی یک دستگاه سیستم توسط پردازشگر گرافیکی GTX 1080 با ۸ گیگابایت حافظه پردازش در مدت زمان ۱۲ روز انجام گرفته است.

۳-۲ تخمین حالت سه‌بعدی بدن از حالت دوبعدی با روش کدگذاری تنک

همانگونه که بیان شد، در روش پیشنهادی به‌منظور تخمین حالت سه‌بعدی براساس حالت‌های دو بعدی از نرم ۲ الگوریتم ADMM relaxed استفاده می‌شود. به گونه‌ای که با ترکیبی از مدل‌های سه‌بعدی واژه‌نامه، سعی در ساخت بهترین حالت سه‌بعدی بدن انسان نسبت به حالت دوبعدی اولیه می‌باشد. در این الگوریتم با استفاده از یک متغییر کمکی و بصورت رویکردی تکراری، روش مستقیم را با تبدیل به دو زیرمسئله حل می‌کند. از این رو براساس مقاله [۲۸] با فرض $M_i \in R^{2\times 3}$ تشکیل شده از ماتریس ضرایب C_i و دو ردیف اول ماتریس چرخش R_i ، رابطه ۳ به فرم رابطه ۸ تبدیل می‌شود.

$$min_{M_1,\dots,M_k} \sum_{i=1}^{k} M_{i2}, \quad \text{s.t.} \quad W = \sum_{i=1}^{k} M_i B_i \qquad (۸)$$

با فرض $\tilde{M} = [M_1 \dots M_k]$ و $\tilde{B} = [B_1 \dots B_k]^T$ ، رابطه ۸ را می‌توان بصورت رابطه ۹ بازنویسی کرد.

$$\min_{\tilde{M},Z} \frac{1}{2}\|W - Z\tilde{B}\|_F^2 + \alpha\sum_{i=1}^{k}\|M_i\|_2 \quad \text{s.t } \tilde{M} = Z \qquad (۹)$$

در این معادله Z یک متغیر کمکی و ضریب α براساس آزمایشات تجربی صورت گرفته برابر ۰,۱ در نظر گرفته شده است.
معادله بالا با استفاده از ضریب لاگرانژ بصورت رابطه ۱۰ بازنویسی می‌شود.

$$L_\mu(\tilde{M},Z,Y) = \frac{1}{2}\|W - Z\tilde{B}\|_F^2 + \alpha\sum_{i=1}^{k}\|M_i\|_2 + < Y, \tilde{M} - Z> + \frac{\mu}{2}\|\tilde{M} - Z\|_F^2 \qquad (۱۰)$$

که در این رابطه، Y متغیر دوگانه و μ پارامتری برای کنترل اندازه گام‌ها در بهینه‌سازی می‌باشد[۳۴]. در نتیجه با روش الگوریتم relaxed ADMM [۳۹] در هر مرحله با ضریبی از $\tilde{M}^{t+1}$ و $\tilde{M}^t$ ، گام‌های زیر بروزرسانی می‌شوند.

$$\tilde{M}^{t+1} = \arg\min_{\tilde{M}} L_\mu(\tilde{M}, Z^t, Y^t) \qquad (۱۱)$$

$$Z^{t+1} = \arg\min_{Z} L_\mu(\beta\tilde{M}^{t+1} + (1-\beta)\tilde{M}^t, Z^t, Y^t) \qquad (۱۲)$$

$$Y^{t+1} = Y^t + \mu\left(\left(\beta\tilde{M}^{t+1} + (1-\beta)\tilde{M}^t\right) - Z^{t+1}\right) \qquad (۱۳)$$

در رابطه‌های ۱۲ و ۱۳، $\beta \in (0,2)$ می‌باشد که با آزمایشات محدود انجام شده، بهترین نتایج در این روش با در نظر گرفتن مقدار ۰.۹ برای β استخراج گردید. لازم به ذکر است اگر ضریب β برابر ۱ قرار گیرد این روش همانند روش ADMM [۳۵] خواهد بود. در

ادامه به استفاده از روش relaxed ADMM [۴۰] و با محاسبه لاگرانژ، اگر هر ستون i از $Z^t-\frac{1}{\mu}Y^t$ برابر Q_i^t در نظر گرفته شود[۲۴]، رابطه ۱۱ به صورت رابطه ۱۴ بازنویسی می‌شود.

$$M^{t+1}=D_{\frac{\alpha}{\mu}}\left(Q_i^t\right) \quad (۱۴)$$

این عملیات بصورت مجزا و تکراری، برای هر ستون براساس مسئله پروکسیمال انجام می‌گیرد[۲۴].

از طرفی در معادله ۱۲، $L_\mu\left(\beta\tilde{M}^{t+1}+(1-\beta)\tilde{M}^t,Z^t,Y^t\right)$ به فرم درجه دوم می‌باشد و بصورت رابطه ۱۵ بازنویسی می‌شود.

$$Z^{t+1}=(W\tilde{B}^T+\mu\left(\left(\beta\tilde{M}^{t+1}+(1-\beta)\tilde{M}^t\right)+Y^t\right)\left(\tilde{B}\tilde{B}^T+\mu I\right)^{-1} \quad (۱۵)$$

با توجه به آنچه گفته شد، شبه‌کد روش پیشنهادی بصورت الگوریتم ۱ می‌باشد. همانند حالت‌های سه‌بعدی اعمال شده به واژه‌نامه، حالت‌های دوبعدی آزمایشی نیز نیاز به نرمال‌سازی دارند. نرمال‌سازی به منظور ایجاد میانگین برابر با صفر و انحراف معیار یک برای مقادیر هر سطر می‌باشد.

روش پیشنهادی ارائه شده در این مقاله سعی در ساختن بهترین تطابق از حالت دوبعدی با حالت‌های سه‌بعدی از طریق مولفه‌های واژه‌نامه دارد.

Input: W,α

Output: $M_1 \dots M_k$

Initialize $Z=Y=0,\ \mu>0,\ \beta\in(0,2)$

While not converged do

Parallel for i=1 to k do

$$Q_i^t=\text{the i-th column-triplet of } Z^t-\frac{1}{\mu}Y^t$$

$$M^{t+1}=D_{\frac{\alpha}{\mu}}\left(Q_i^t\right)$$

End

$$Z^{t+1}=(W\tilde{B}^T+\mu\left(\left(\beta\tilde{M}^{t+1}+(1-\beta)\tilde{M}^t\right)+Y^t\right)\left(\tilde{B}\tilde{B}^T+\mu I\right)^{-1}$$

$$Y^{t+1}=Y^t+\mu\left(\left(\beta\tilde{M}^{t+1}+(1-\beta)\tilde{M}^t\right)-Z^{t+1}\right)$$

End

الگوریتم ۱: حل رابطه ۹ با الگوریتم relaxed ADMM.

۳-۳ بازسازی مدل سه‌بعدی

بازسازی حالت سه‌بعدی تخمین زده شده از حالت‌های استخراجی از الگوریتم ۱ به‌صورت مبتنی بر مدل انجام می‌گیرد. شرح مراحل مختلف این بخش از روش پیشنهادی به این صورت می‌باشد:

- مرحله اول: تخمین حالت سه‌بعدی توسط الگوریتم ۲ و خروجی‌های الگوریتم ۱ متشکل از ماتریس‌های M مطابق مولفه‌های واژه‌نامه، ساخته می‌شود.

Input: $M_1 \dots M_k$

Output: S

For i=1 to k do

$$c_i=\|M_i\|_2$$

$$r_i^{(1)}=m_i^{(1)}/c_i$$

$$r_i^{(2)}=m_i^{(2)}/c_i$$

$$r_i^{(3)}=r_i^{(1)}\times r_i^{(2)}$$

$$R_i=\left[r_i^{(1)},r_i^{(2)},r_i^{(3)}\right]^T$$

End

$$S=\sum_{i=1}^{k}c_iR_iB_i$$

الگوریتم ۲: بازسازی مستقیم سه‌بعدی [۲۹].

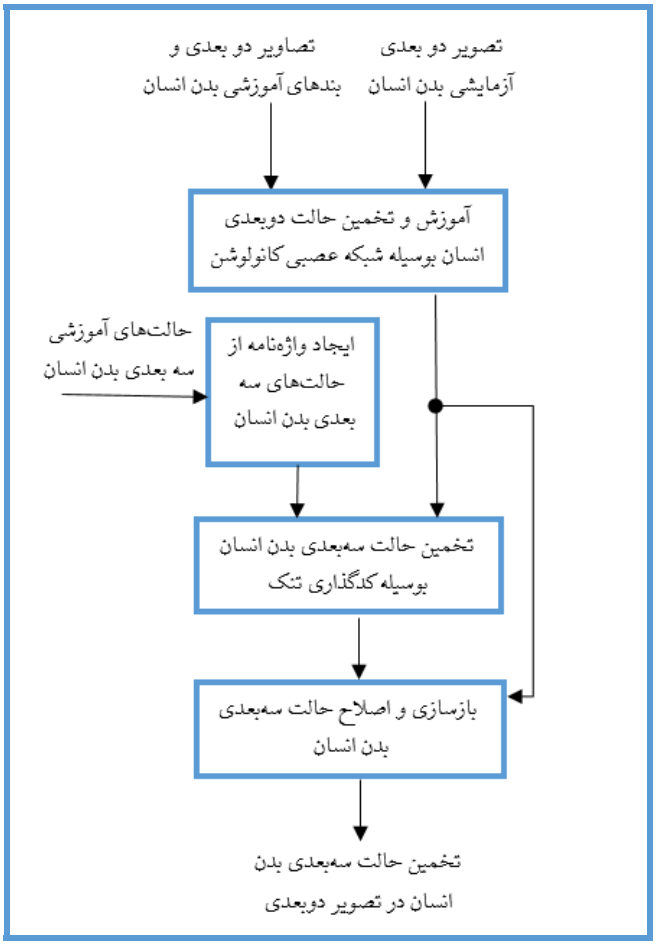
- مرحله دوم: تخمین عمق انجام شده در مرحله اول به نرمال تخمین حالت دوبعدی استخراج شده توسط شبکه عصبی کانولوشن اضافه می‌شود.
- مرحله سوم: ساختار درختی ۲ مدل برای حالت سه‌بعدی بدن انسان، براساس نزدیک‌ترین حالت‌های دوبعدی همتراز شده در واژه‌نامه با تخمین حالت دوبعدی، در نظر گرفته می‌شود.
- مرحله چهارم: با توجه به اینکه مدل‌های استخراجی در مرحله سوم از نظر نسبت اندازه، با حالت تخمین زده شده متفاوت می‌باشد، از این رو، ضریب نسبت بین ۲ مدل سه‌بعدی و مدل تخمین زده شده از طریق محاسبه جمع فاصله بندها با معیار فاصله اقلیدوسی در ساختار درختی، انجام می‌گیرد. در ادامه لازم است این ضرایب در مدل‌های مرحله سوم ضرب شوند تا نسبت بین اعضا برقرار گردد.
- مرحله پنجم: برای ۲ مدل بروزرسانی شده و مدل تخمین زده شده، فاصله اقلیدوسی بندها با اولویت فاصله از ریشه در فضای سه‌بعدی محاسبه می‌شود. سپس اندازه فاصله اقلیدوسی هر خط فاصل بین دو بند مدل تخمین زده شده و بازه اندازه فاصله اقلیدوسی دو مدل، مقایسه می‌شود. اگر اندازه فاصله دو بند در بین بازه تعیین شده قرار نگیرد، میانگین اندازه فاصله اقلیدوسی دو مدل ساختاری، به‌عنوان اندازه فاصله دو بند تخمین زده شده، در نظر گرفته می‌شود. در ادامه با استفاده از حالت دوبعدی تخمین زده و اندازه فاصله، بر اساس هم جهت بودن با عمق تخمینی اصلی، عمق بند انتهایی بروز رسانی می‌شود.

۳-۴ ساخت واژه‌نامه

سیستم کدگذاری تنک به‌منظور ساخت واژه‌نامه، با هدف کاهش حجم داده‌ها و پوشش دادن تمام حالات ممکن آموزشی صورت می‌گیرد. از این رو در بعضی مسائل، از تمام داده‌های آموزشی جهت ذخیره در واژه‌نامه استفاده می‌شود، اما در رویکردی دیگر، بدلیل حجم بالای اطلاعات آموزشی، واژه‌نامه بصورت یادگیر، بگونه‌ای که تمام حالات را بر اساس حالات پایه پوشش دهد ساخته می‌شود. لذا در این پژوهش همانند مقاله [۲۹] جهت ساخت واژه‌نامه از بهینه‌سازی رابطه ۱۶ استفاده شده است و تعداد حالت‌های سه‌بعدی واژه نامه ۱۲۸ در نظر گرفته شده است.

$$\begin{aligned} &\min_{B,C} \sum_{j=1}^{n} \frac{1}{2} \|S_j - \sum_{i=1}^{k} C_{i,j} B_i\|_F^2 + \lambda \|C\|_1 \\ &\text{s.t. } C_{i,j} \geq 0, B_{iF} \leq 1, \\ &\forall i \in [i,k],\ j \in [1,n], \end{aligned} \quad (۱۶)$$

در این معادله S_j حالت سه‌بعدی آموزشی، B_i حالت پایه‌ی یادگیری شده، C_{ij} ضریب iامین ضریب برای jامین حالت‌های آموزش می‌باشد[۲۹].



شکل ۳: مراحل روش پیشنهادی تخمین حالت سه‌بعدی بدن انسان

۴ نتایج تجربی

جهت تحلیل و ارزیابی دقیق‌تر نتایج حاصل از روش‌های مختلف، در ابتدای این بخش مجموعه‌دادگان استفاده شده معرفی خواهد شد. در بخش ۴-۲ آشنایی مختصری با معیارهای ارزیابی بکار گرفته شده صورت می‌گیرد. نهایتا در بخش ۴-۳ نتایج حاصل از پیاده‌سازی روش پیشنهادی در مقایسه با کارهای پیشین ارائه خواهد شد.

۴-۱ مجموعه‌دادگان مورد استفاده

مجموعه‌دادگان Human 3.6M [۱۳] یک مجموعه‌دادگان بزرگ برای تخمین حالت سه‌بعدی انسان است و در این مقاله مورد استفاده قرار گرفته است. این مجموعه‌دادگان شامل میلیون‌ها حالت سه‌بعدی بدست آمده از تصاویر دوربین‌های کالیبره[1] شده است. فیلم‌های موجود در این مجموعه‌دادگان از چند دید[2] به صورت هماهنگ شده می‌باشند. همچنین این مجموعه داده‌های دوبعدی و سه‌بعدی از حالت‌های بدن انسان جهت ارزیابی با تعداد بالا قرار دارد. از مدل‌های انجام آزمایش بر روی این مجموعه‌دادگان، استفاده از مجموعه‌ای شامل ۱۱ بازیگر، ۷ مدل و ۱۵ موضوع می‌باشد و از ۵ مدل (S1, S5, S6, S7,S8) در ۱۵ موضوع به‌منظور آموزش و از ۲ مدل (S9, S11) در ۱۵ موضوع جهت آزمایش استفاده می‌شود. نتایج این مقاله نیز براساس این مدل ارزیابی شده است. ویدیوهای این مجموعه‌دادگان برای هر موضوع، با دوبار تکرار از ۴ دید متفاوت تهیه شده است و هر موضوع با نرخ اصلی ۵۰ فریم بر ثانیه است. در اکثر کارهای پیشین و روش پیشنهادی این مقاله، این نرخ به ۱۰ فریم بر ثانیه تبدیل شده است. استخراج اسکلت‌بندی بدن انسان توسط ۳۲ حسگر انجام گرفته است که در این پژوهش همانند سایر مقالات از ۱۵ بند اصلی استفاده شده است(شکل ۱).

۴-۲ معیارهای ارزیابی

از معیارهای متداول جهت سنجش نتایج روش‌های مختلف، محاسبه خطای هر بند، از میانگین فاصله اقلیدوسی در تمام مفاصل می‌باشد. در این معیار اگر موقعیت بندهای تخمین زده شده $\hat{x}_1,\ldots,\hat{x}_n$ و موقعیت حقیقت اصلی $x_1,\ldots,x_n$ فرض شود، بصورت رابطه ۱۷ است.

$$e = \frac{1}{n} \sum_{i=1}^{n} \| \hat{x}_i - x_i \|_2 \quad (۱۷)$$

مقدار خطا در رابطه ۱۷ وابسته به مقیاس و چرخش تخمین صورت گرفته است. از طرفی ابهام در عمق و مقیاس در بازسازی حالت سه‌بعدی از تصویر دوبعدی به‌صورت کامل قابل حل نمی‌باشد. از این رو برای مقایسه و سنجش دقت بازسازی، از

[1] Calibration
[2] Multi View

میانگین حالت‌های آموزشی استفاده و حالت تخمین زده شده نسبت به مکان‌های ریشه اسکلت بدن هم‌تراز می‌شود.

لذا ترازسازی به صورت دلخواه قابل قبول نمی‌باشد. بر این اساس، خطای بازسازی[1] برای حالت سه‌بعدی به صورت رابطه ۱۸ تعریف می‌شود[۴۴].

$$r = min_T \frac{1}{n}\sum_{i=1}^{n} \| \hat{x}_i - T(x_i) \|_2 \qquad (۱۸)$$

که T نشان‌دهنده ماتریس تبدیل است. عمدتا خطای بازسازی در ساختارها به منظور ارزیابی دقت ساختار بهبود یافته بدون در نظر گرفتن مقیاس و صلب بودن حالت است.

۴-۳ ارائه و تحلیل نتایج تجربی

در این پژوهش تخمین حالت سه‌بعدی بدن انسان می‌تواند از طریق حالت دوبعدی یا از طریق استخراج ویژگی در تصویر و تخمین حالت دوبعدی انجام شود. این دو حالت به‌دلیل ساختاری با هم متفاوت هستند، زیرا در تخمین حالت دوبعدی از یک تصویر، سعی در تخمین حالت دوبعدی است و هدف نسبت دادن بهترین عمق به آن می‌باشد. اما در تخمین حالت سه‌بعدی از حالت دوبعدی فرض بر اینست که تخمین حالت دوبعدی انجام گرفته و هدف تخمین یک حالت سه‌بعدی بر اساس واژه‌نامه تشکیل شده بگونه‌ای است که از نظر خطای دوبعدی کمترین خطا را بدون استفاده از حالت دوبعدی اولیه داشته باشد. در نتیجه در این مقاله، با دو رویکرد متفاوت سعی در حل این چالش داریم. رویکرد اول شامل تخمین حالت سه‌بعدی بدن انسان براساس حالت دوبعدی مجموعه‌دادگان می‌باشد. این رویکرد براساس الگوریتم ۱ ضرایب مورد نیاز استخراج شده و با استفاده از مراحل بازسازی و بدون استفاده از شبکه عصبی کانولوشن، بصورت مبتنی بر مدل بهبود داده می‌شود. در رویکرد دوم پس از تخمین حالت دوبعدی بدن انسان از تصویر بوسیله شبکه عصبی کانولوشن، حالت‌های تخمین زده نرمال شده و در ادامه بوسیله الگوریتم ۱ ضرایب استخراج می‌شوند. سپس بازسازی توسط ۵ مرحله مطابق آنچه پیش‌تر در بخش ۳-۳ گفته شد انجام می‌گیرد.

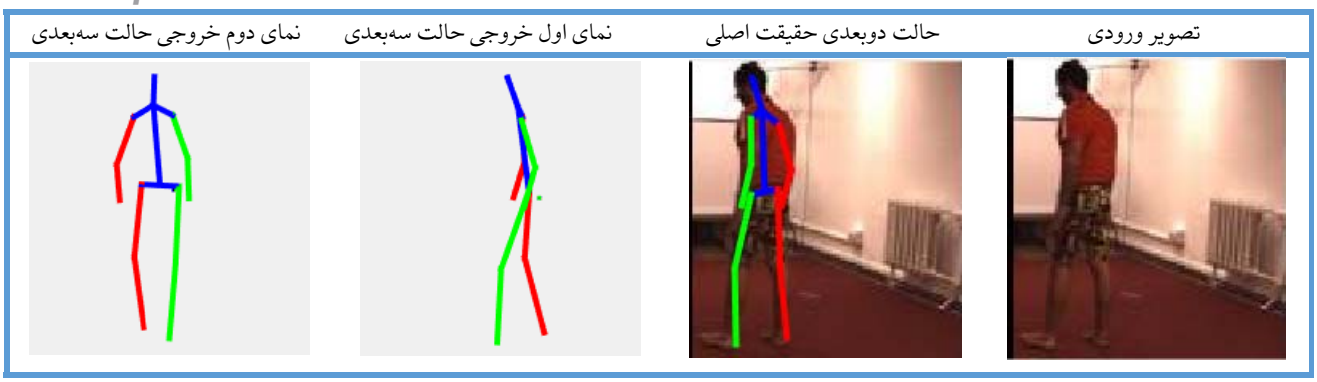
نتایج حاصل از آزمایش‌های صورت گرفته رویکرد اول بر روی مجموعه دادگان Human 3.6M با ۱۵ بند اصلی در جدول ۱ با عنوان روش پیشنهادی ارائه می‌گردد. همانطور که مشخص است، نتایج حاصل از پیاده سازی رویکرد اول با بهره‌گیری از کدگذاری تنک منجر به میانگین خطای بازسازی ۴۷.۵ می‌شود. این بهبود نشان از اهمیت بکارگیری روش‌های ترکیبی مبتنی بر مدل و مبتنی بر نمونه می‌باشد.

با توجه به جدول ۱، رویکرد حریصانه الگوریتم OMP در مقاله PMP [۳۶] نشان از ضعف این مدل در ترکیب و ساخت مدل با کمترین خطا، بدلیل عملیات حریصانه و استفاده از نزدیک‌ترین حالت‌ها است. به بیان دیگر، مقاله [۳۶] از محاسبه نزدیک‌ترین حالت دوبعدی و ترکیب با حالت‌های استخراجی قبلی استفاده می‌کند و قابلیت جلوگیری از انتخاب نزدیک‌ترین حالت‌های دوبعدی که منجر به خطای بالاتر در ترکیب حالت‌ها را ندارد. از طرفی با توجه به ثابت بودن دوربین در هنگام فیلم‌برداری (اگرچه بازیگر بطور محدود در داخل صحنه جابجا می‌شود)، در مقاله [۴۵] محدودیت هندسی در بازسازی سه‌بعدی ایجاد می‌کند وخطای تخمین بالا می‌رود. اما در روش پیشنهادی، به‌دلیل استفاده از ساختارهای درختی از بندها به‌عنوان مدل، ساختار حالت بدن انسان حفظ می‌شود و مانع از ایجاد عمق نامناسب نسبت به اعضای بدن می‌گردد. در این نگرش برای حفظ تعادل بین خطای تخمین حالت دوبعدی و عمق تخمین زده شده، بازه تعیین شده در مرحله بازسازی، به‌عنوان یک تصحیح‌گر عمق‌های نامناسب عمل می‌کند. همچنین می‌توان چنین استنباط کرد تخمین‌های دقیق‌تر بیشتر در حوزه تخمین عمق می‌باشد و نسبت به حالت دوبعدی بهبود قابل توجهی بدست نیامده است.

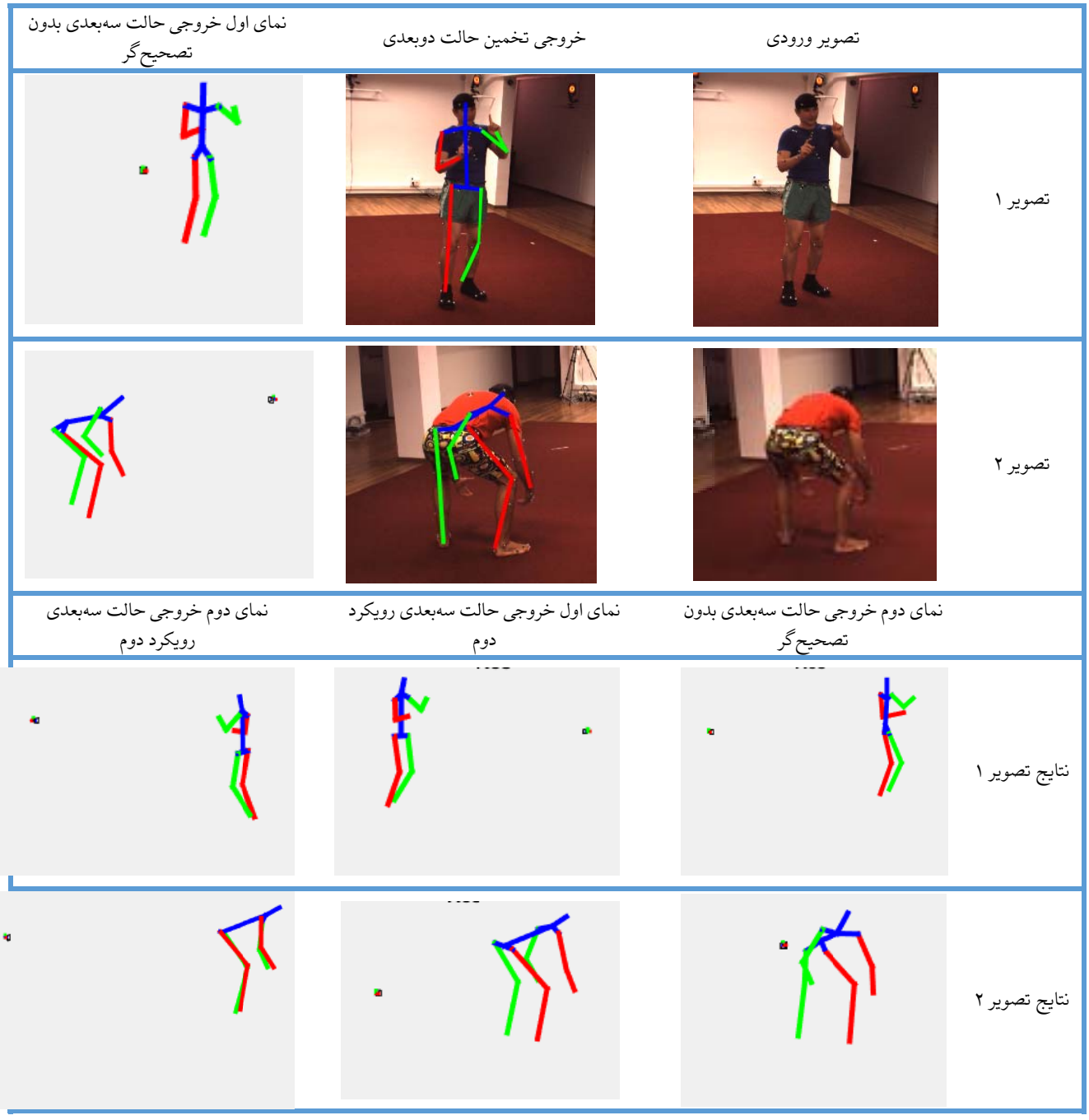
با توجه به جدول ۱، بعضی حالات بدلیل تفاوت در مقیاس اندازه‌های ساختار درختی استخراج شده، خطای بازسازی عمق افزایش یافته و از حالت اصلی عمق حالت بدن، فاصله گرفته است. از این رو استفاده از ساختار مبتنی بر مدل، وابستگی بالایی به مدل ساخته شده را نشان می‌دهد. این رویکرد با توجه به محدودسازی واژه‌نامه و قابلیت تعمیم آن جهت ساخت حالت‌های جدید سه‌بعدی با رویکرد مقالات [۱۶، ۱۷ و ۲۷] متفاوت می‌باشد و بجای ذخیره سازی ۱۵۸۰۰۰ تصویر آموزشی، به ۱۲۸ حالت که کمترین خطا نسبت به ساخت تمام حالات را داشته باشد اکتفا می‌کند. در این مقالات علاوه بر اطلاعات فضایی حالت، از نقشه سیاه نما نیز استفاده می‌کند و نقطه ضعف این روش‌ها، عدم امکان نادیده گرفتن نزدیک‌ترین حالت دوبعدی است زیرا این حالات همواره نزدیک‌ترین حالت سه‌بعدی نسبت به هم را ندارند و همین امر باعث افزایش خطا می‌گردد. لذا با توجه به روابط مقاله [۱۷ و ۲۷] قیدهای نرم یک و نرم دو هستند و موجب می‌شود جواب منحصربفردی داشته باشد؛ در عین حال که تنک بودن پاسخ‌ها نیز تحقق یابد، بدین ترتیب با توجه به ساختار روابط می‌توان روابط را به شکل ساختار روابط رگرسیون در نظر گرفت. در شکل ۴ نمونه‌ی خروجی مراحل رویکرد اول نشان داده شده است.

در رویکردی دیگر، زمانی که حالت دوبعدی بدن انسان به‌عنوان ورودی اعمال نشود از رویکرد دوم استفاده می‌شود. این رویکرد در جدول ۲ با عنوان روش پیشنهادی با دیگر روش‌ها مقایسه شده است و بهبود عملکرد ۴ و ۲۰ درصدی نسبت به بهترین تخمین مقالات بررسی شده و میانگین خطای مقالات گذشته را نشان می‌دهد. همانطور که پیش‌تر بیان شد در حوزه‌های با فضای حالت بالا، قابلیت تعمیم تمام حالت‌ها در شبکه‌های عصبی کانولوشنی وجود ندارد.

[1] Reconstruction Error



شکل ۴: نمونه‌ای از خروجی مراحل رویکرد اول روش پیشنهادی



شکل ۵: نمونه‌ای از خروجی مراحل رویکرد دوم پیشنهادی

از طرفی شبکه‌های عصبی کانولوشن در برخورد با فضای حالت بالا، خطای بالاتری را نشان می دهد. با توجه به نتایج جدول ۲، بدلیل فضای حالت بالای حرکات، نیاز به روش‌های ترکیبی هرچه بیشتر احساس می‌شود. لذا نتایج روش پیشنهادی بدون تصحیح‌گر، قدرت عملکرد تجمیع نتایج شبکه‌های عصبی کانولوشنی و مراحل مبتنی بر مدل هنگام بازسازی را نشان می‌دهد و از طرفی با توجه به نتایج مقاله [۲۴] می‌توان چنین استنباط کرد کاهش پارامتر در بخش شبکه عصبی کانولوشن منجر به خطای بالاتر در تخمین حالت دوبعدی می‌گردد اما استفاده دوباره از نتایج خروجی شبکه کانولوشنی، خطای دوبعدی را نسبت به بازسازی مستقیم کاهش می‌دهد و با اعمال ساختار درختی در بازسازی، امکان ایجاد عملکرد مناسبی در تخمین حالت سه‌بعدی گردد. لذا این جدول نشان می‌دهد تخمین حالت سه‌بعدی وابستگی بالایی به تخمین مناسب عمق دارد. اما بخش بزرگی از خطا، وابسته به حالت دوبعدی ورودی می‌باشد. از این رو جهت تخمین مناسب حالت دوبعدی براساس روش‌های یادگیر، شبکه‌های عصبی کانولوشن بدلیل برخورداری از دقت بالا، مورد استفاده قرار گرفته است. در شکل ۵ نمونه‌ی خروجی‌های مراحل رویکرد دوم بصورت تغییر ابعاد تخمین حالت دوبعدی جهت نمایش بر روی تصویر با ابعاد بزرگتر، نشان داده شده است. لذا بهبود عملکرد روش پیشنهادی در تخمین حالت سه‌بعدی بدن انسان براساس محل قرارگیری دوربین، قابل مشاهده می‌باشد.

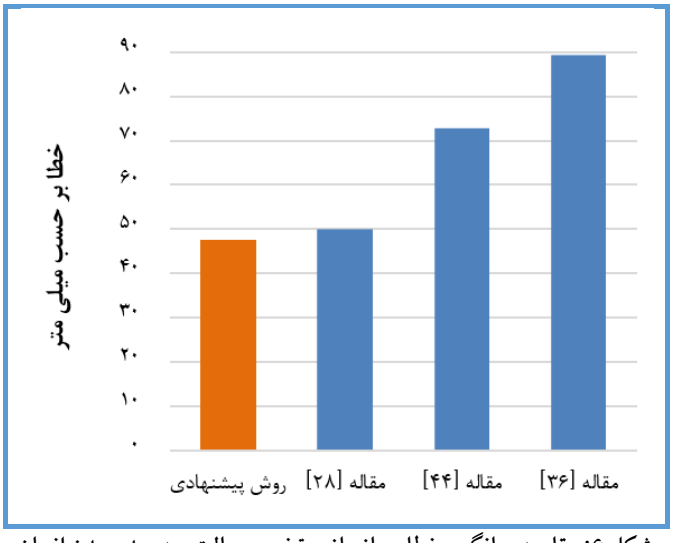
میانگین نتایج حاصل از آزمایش‌های صورت گرفته روش پیشنهادی نسبت به مقالات پیشین مطابق جدول ۱ و ۲ در شکل ۶ و ۷ بصورت نمودار نمایش داده شده است.

جدول ۱: دقت تخمین حالت سه بعدی بدن انسان از حالت دوبعدی حقیقت اصلی براساس میانگین خطای بازسازی (اعداد ذکر شده خطا بر حسب میلی متر می باشند)

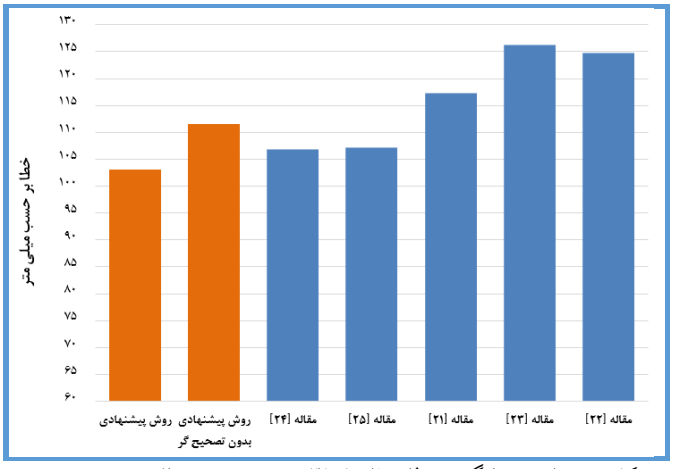
	روش پیشنهادی	مقاله [۲۸]	مقاله [۴۵]	مقاله [۳۶]
Directions	**۳۶.۶**	۳۸.۲	۷۹.۲۶	۶۸.۵
Discussion	**۴۰.۸**	۴۵.۰	۶۰.۷۵	۷۷.۵
Eating	**۴۳.۱**	۴۷.۲	۱۲۵.۷۹	۹۵.۷
Greeting	۴۹.۲	**۴۸.۵**	۷۴.۹۷	۸۶.۴
Phoning	**۳۹.۵**	۴۵.۹	۳۷.۱۴	۷۳.۹
Photo	**۵۸.۲**	۶۳.۷	۵۸.۷۴	۹۵.۲
Posing	**۴۳.۰**	۴۷.۰	۶۱.۳۸	۷۸.۳
Purchases	**۳۸.۹**	۴۳.۸	۸۸.۴۷	۹۷.۸
Sitting	۵۸.۲	**۵۳.۵**	۱۱۲.۷۴	۱۰۳.۶
SittingDown	۷۵.۰	**۷۰.۱**	۱۱۴.۶۸	۱۲۳.۴
Smoking	**۳۶.۴**	۴۱.۲	۴۵.۰۹	۷۲.۸
Waiting	**۴۶.۱**	۴۸.۴	۷۸.۰۰	۹۵.۱
WalkDog	**۴۷.۹**	۵۳.۷	۶۲.۷۹	۸۳.۴
Walking	۴۸.۴	**۴۷.۳**	۶۳.۲۰	۹۵.۱
WalkTogether	**۵۱.۸**	۵۶.۵	۳۱.۶۷	۹۴.۹
Average	**۴۷.۵**	۵۰	۷۲.۹۸	۸۹.۴

جدول ۲: دقت تخمین حالت سه بعدی بدن انسان از یک تصویر دوبعدی براساس خطای میانگین فاصله اقلیدسی (اعداد ذکر شده خطا بر حسب میلی متر می باشند).

	روش پیشنهادی	روش پیشنهادی بدون تصحیح-گر	مقاله [۲۴]	مقاله [۲۵]	مقاله [۲۱]	مقاله [۲۳]	مقاله [۲۲]
Directions	**۷۹.۲**	۸۶.۷	۸۳.۳	۹۱.۸	۱۰۰.۳	۸۵.۰	۱۰۲.۴
Discussion	**۹۳.۷**	۱۰۵.۲	۹۸.۴	۱۰۲.۴	۱۱۶.۱	۱۱۲.۶	۱۴۷.۷
Eating	**۸۳.۶**	۸۶.۸	۸۴.۸	۹۶.۶	۸۹.۹	۱۰۴.۹	۸۸.۸
Greeting	۹۹.۴	۱۰۳.۹	۱۰۴.۲	**۹۸.۷**	۱۱۶.۴	۱۲۲.۰	۱۲۵.۲
Phoning	**۹۶.۸**	۱۱۵.۳	۱۱۲.۷	۱۱۳.۳	۱۱۵.۳	۱۳۹.۰	۱۱۸.۰
Photo	**۱۱۲.۷**	۱۳۸.۴	۱۳۲.۴	۱۲۵.۲	۱۴۹.۵	۱۳۵.۹	۱۸۲.۷
Posing	۹۳.۷	۱۰۵.۴	۱۰۳.۲	**۹۰.۰**	۱۱۷.۵	۱۰۵.۹	۱۱۲.۳
Purchases	**۸۹.۶**	۹۹.۶	۹۷.۲	۹۳.۸	۱۰۶.۹	۱۶۶.۱	۱۲۹.۱
Sitting	۱۲۱.۲	۱۲۴.۸	**۱۱۳.۳**	۱۳۲.۱	۱۳۷.۲	۱۱۷.۴	۱۳۸.۸
Sitting Down	۱۹۱.۴	۱۹۶.۹	۱۸۳.۶	**۱۵۸.۹**	۱۹۰.۸	۲۲۶.۹	۲۲۴.۹
Smoking	**۹۸.۶**	۱۰۶.۲	۱۰۴.۳	۱۰۶.۹	۱۰۵.۷	۱۲۰.۰	۱۱۸.۴
Waiting	۱۰۳.۳	۱۱۵.۵	۱۰۹.۱	**۹۴.۴**	۱۲۵.۱	۱۱۷.۶	۱۳۸.۷
WalkDog	۱۱۳.۱	۱۱۴.۶	**۱۰۶.۴**	۱۲۶.۰	۱۳۱.۹	۱۳۷.۳	۱۲۶.۲
Walking	۷۸.۵	۷۹.۷	۷۶.۹	۷۹.۰	۶۲.۶	۹۹.۲	**۵۵.۰**
Walk Together	۹۱.۶	۹۵.۱	۹۳.۲	۹۸.۹	۹۶.۱	۱۰۶.۵	**۶۵.۷**
Average	**۱۰۳.۱**	۱۱۱.۶	۱۰۶.۹	۱۰۷.۲	۱۱۷.۴	۱۲۶.۴	۱۲۴.۹



شکل ۶: مقایسه میانگین خطای بازسازی تخمین حالت سه بعدی بدن انسان از حالت دوبعدی حقیقت اصلی.



شکل ۷: مقایسه میانگین خطای فاصله اقلیدسی تخمین حالت سه بعدی بدن انسان از یک تصویر دوبعدی.

۵ نتیجه‌گیری و پیشنهادهای آتی

در این مقاله، یک چارچوب تخمین حالت سه‌بعدی بدن انسان از تنها یک تصویر دوبعدی ارائه شده است. این موضوع علاوه بر کمک به تشخیص حالت بدن انسان، می‌تواند در مراکز حساس اداری و نظامی به‌عنوان سیستم هشدار جهت تشخیص حالت‌های مشکوک، صنعت پویانمایی و فیلم‌سازی در جهت طبیعی شدن حرکات موجودات ساخته‌شده با جلوه‌های ویژه و ... مورد استفاده قرار گیرد. از این رو نیاز به مدل‌کردن مکانیسمی مشابه عملکرد قشر بینایی و مغز انسان می‌باشد، لذا استخراج ویژگی‌های با رویکرد تشخیص حالت بدن انسان، جز با تحلیل دقیق مکانیسم تشخیص و تصمیم‌گیری در مغز انسان میسر نخواهد بود. از این رو، در روش پیشنهادی سعی گردید که این موارد لحاظ شود.

استخراج ویژگی‌های تصویر، مرحله اصلی و مهم تخمین حالت دوبعدی می‌باشد. از این رو، از قدرت بالای شبکه عصبی کانولوشن جهت تخمین و آشکارسازی بندهای بدن انسان استفاده شده است. در این مقاله از شبکه‌ای ۱۶ لایه‌ای با ابعاد لایه پایین استفاده شده است تا علاوه بر رویکرد کاهش پارامتر، دقت کافی جهت تخمین ۱۵ بند حالت دوبعدی در نظر گرفته شده را داشته باشد.

از سوی دیگر، در چالش تخمین حالت سه‌بعدی، با توجه به ماهیت حالت بدن انسان، بر پایه سیستم کدگذاری وبازنمایی تنک به تخمین و بازسازی توسط واژه‌نامه حالت سه‌بعدی پرداخته شده است و عمق را به خروجی نرمال شده شبکه عصبی کانولشنی اضافه می‌کنیم، سپس عمق حالت تخمین زده شده را بصورت مبتنی بر مدل تصحیح شده است تا معضل عمق نامناسب را نسبت به مقالات گذشته کاهش دهیم و از تاثیر خطای ایجاد شده در تخمین حالت سه‌بعدی در کدگذاری و بازنمایی تنک بکاهیم. نتایج نشان می‌دهد این روش از نظر دقت با معیار متداول میانگین خطای بازسازی و میانگین خطای بازسازی هر اتصال نسبت به کارهای پیشین برتری دارد.

علیرغم نقاط قوت ذکر شده، متاسفانه روش تصحیح‌گر حالت سه‌بعدی در مواردی که تخمین حالت دوبعدی نامناسب انجام گرفته است عمق را نسبت به حالت اصلی کاهش می‌دهد و علاوه بر خطای بالای حالت دوبعدی، خطای عمق نیز افزایش می‌یابد. لذا در مرحله تخمین حالت دوبعدی از تصویر، نیاز به شبکه عصبی کانولوشنی دقیق‌تر می‌باشد. از طرفی بازسازی حالت سه‌بعدی به‌صورت مستقیم نیز باعث خطای بالایی می‌شود و نیاز به روش‌های دقیق‌تر با کدگذاری و بازنمایی تنک احساس می‌شود.

مراجع

[1] Felzenszwalb, P.F., Huttenlocher, D.P., “Pictorial structures for object recognition”, International Journal of Computer Vision, Vol. 61(1), pp. 55–79, 2005.

[2] Yang, Y., Ramanan, D., “Articulated pose estimation with flexible mixtures-ofparts”, Computer Vision and Pattern Recognition (CVPR), pp. 1385–1392, 2011.

[3] Dantone, M., Gall, J., Leistner, C., Van Gool, L., “Human pose estimation using body parts dependent joint regressors”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3041–3048, 2013.

[4] Toshev, A., Szegedy, C., “Deeppose: Human pose estimation via deep neural networks”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660, 2014.

[5] Krizhevsky A., Sutskever I., Hinton, G.E. “Imagenet classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105, 2012.

[6] Carreira J., Agrawal P., Fragkiadaki K., Malik J., “Human pose estimation with iterative error feedback”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] Fan, X., Zheng, K., Lin, Y., Wang, S., “Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1347–1355, 2015.

[8] Wei S.E., Ramakrishna V., Kanade T., Sheikh Y., “Convolutional pose machines”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[9] Chu, X., Ouyang, W., Li, H., Wang, X., “Structured feature learning for pose estimation”, 2016.

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

[11] Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I. A., “3D human pose estimation: A review of the

[26] Simo-Serra, E., Ramisa, A., Aleny`a, G., Torras, C., MorenoNoguer, F., "Single Image 3D Human Pose Estimation from Noisy Observations," Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2673-2680, 2012.

[27] Zolfaghari M., Jourabloo A., Ghareh Gozlou S., Pedrood B., Manzuri-Shalmani M. T., "3D Human Pose Estimation from Image Using Couple Sparse Coding," Machine Vision and Applications, 25(6), pp. 1489-1499, 2014.

[28] Zhou, X., Leonardos, S., Hu, X., Daniilidis, K., "3D Shape Reconstruction from 2D Landmarks: A Convex Formulation", IEEE International Conference on Computer Vision and Pattern Recognition, 2015.

[29] Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K., "3D shape estimation from 2D landmarks: A convex relaxation approach", IEEE transactions on pattern analysis and machine intelligence, Vol: 39 (8), pp: 1648-1661, 2017.

[30] Agarwal A., Triggs B., "Recovering 3D human pose from monocular images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol: 28 (1), pp. 44–58, 2006.

[31] Bengio, Y., "Learning deep architectures for AI," Found. trends Mach. Learn., vol. 2, no. 1, pp. 1–127, 2009.

[32] Bengio, Y., Courville, A., Vincent, P., "Representation learning: A review and new perspectives", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol: 35 (8), pp. 1798–1828, 2013.

[33] Deng, L., Yu, D., "Deep learning: Methods and applications, Foundations and Trends in Signal Processing",Vol: 7, pp. 197–387, 2014.

[34] Boyd, S., "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2010.

[35] Xu, Y., Zhang, Z., Lu, G., Yang, J., "Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification," Pattern Recognition, 54, pp. 68-82, 2016.

[36] Ramakrishna, V., Kanade, T., Sheikh, Y., "Reconstructing 3D human pose from 2D image landmarks," ECCV, 2012.

[37] Zhang, Z., Xu, Y., Yang, J., Li, X., Zhang, D.,"Asurvey of sparse representation: algorithms and applications", IEEE Access3, pp. 490–530, 2015.

[38] Donoho, D., Elad, M., "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," Proceedings of the National Academy of Sciences, vol. 100, no. 5, pp. 2197–2202, 2003.

[39] Banjac, G., Goulart, P., Stellato, B., Boyd, S., "Infeasibility detection in the alternating direction literature and analysis of covariates", CVIU, 152, pp. 1–20, 2016.

[12] Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall J., "A dual-source approach for 3D pose estimation from a single image", Proc IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[13] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," PAMI, vol. 36, no. 7, pp. 1325–1339, 2014.

[14] Li, S., Chan, A. B., "3D human pose estimation from monocular images with deep convolutional neural network", Proc. 12th Asian Conference on Computer Vision, pp. 332–347, 2014.

[15] Li, S., Zhang, W., Chan, A. B., "Maximum-margin structured learning with deep networks for 3D human pose estimation", Proc. IEEE International Conference on Computer Vision, pp. 2848–2856, 2015.

[16] Babagholami-Mohamadabadi B., Jourabloo A., Zarghami A., Kasaei S., "A Bayesian framework for sparse representation-based 3-D human pose estimation". IEEE Signal Proc Lett. 21(3):297–300, 2014.

[17] Andalib, A., Babamir, S. M., Faraji, A., "A NEW SPARSE REPRESENTATION ALGORITHM FOR 3D HUMAN POSE ESTIMATION", Computing and Informatics, Vol. 35, pp. 1338-1355, 2016.

[18] Chen, C. H., Ramanan, D., "3d human pose estimation= 2d pose estimation+ matching", 2016.

[19] Bogo F., Kanazawa A., Lassner C., Gehler P., Romero J., Black M. J., "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image". ECCV, pp. 561–578, 2016.

[20] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B., "DeepCut: Joint subset partition and labeling for multi person pose estimation". IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 4929-4937, 2016.

[21] Park, S., Hwang, J., Kwak, N., "3D human pose estimation using convolutional neural networks with 2D pose information", ECCVW, 2016.

[22] Tekin, B., Rozantsev, A., Lepetit, V., Fua, P., "Direct prediction of 3D body poses from motion compensated sequences,", CVPR, 2016.

[23] Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., Kankanhalli, M., Geng, W., "Marker-less 3D human motion capture with monocular image sequence and height-maps," ECCV, 2016.

[24] Alikarami, H., Yaghmaee, F., Fasaeieslam, M. j., "Sparse representation and convolutional neural networks for 3D human pose estimation", 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), pp. 188 – 192, 2017.

[25] Zhou, X., Sun X., Zhang, W., Liang, S., Wei, Y., "Deep kinematic pose regression", ECCVW, 2016.

method of multipliers for convex optimization", Foundations and Trends in Machine Learning, eprints for the optimization community, 2017.

[40] Simonyan, K., Zisserman, A., "Very deep convolutional networks for large-scale image recognition,", 2014.

[41] Yang, W., Ouyang, W., Li, H., Wang, X., "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation", CVPR, 2016.

[42] Dai, Q., D. Hoiem. "Learning to localize detected objects", Computer Vision and Pattern Recognition (CVPR), 2012.

[43] Dai, J., He K., Sun J.. "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation", Proceedings of the IEEE International Conference on Computer Vision. 2015.

[44] Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K. G., Daniilidis, K. "Monocap: Monocular human motion capture using a cnn coupled with a geometric prior", 2017.

[45] Dai, Y., Li, H., and He, M., "A simple prior-free method for nonrigid Structure from motion factorization," IJCV, vol. 107, no. 2, pp. 101–122, 2014.

حسن علی‌کرمی تحصیلات خود را در دوره کارشناسی و کارشناسی ارشد خود را به ترتیب در سال‌های ۱۳۹۳ و ۱۳۹۶ در رشته مهندسی کامپیوتر- نرم‌افزار و مهندسی کامپیوتر-هوش مصنوعی و رباتیکز از دانشگاه های شهید چمران اهواز و دانشگاه سمنان به اتمام رسانده است. ایشان هم اکنون دانشجوی دکترای مهندسی کامپیوتر- هوش مصنوعی و رباتیکز در دانشگاه سمنان می‌باشد. زمینه‌های مورد علاقه ایشان پردازش تصویر، بازشناسی الگو و یادگیری عمیق می‌باشد.

فرزین یغمایی دوره کارشناسی کامپیوتر را در دانشگاه صنعتی امیرکبیر، کارشناسی‌ارشد و دکترای خود را در رشته مهندسی کامپیوتر- هوش مصنوعی در دانشگاه صنعتی شریف به پایان رسانده است و در حال حاضر دانشیار گروه مهندسی کامپیوتر دانشگاه سمنان می‌باشد. زمینه‌های مورد علاقه ایشان پردازش تصویر و ویدئو، بازشناسی الگو و متن‌کاوی می‌باشد.

محمدجواد فدایی‌اسلام استادیار گروه مهندسی کامپیوتر دانشگاه سمنان می‌باشد. ایشان مدارک کارشناسی مهندسی کامپیوتر- سخت‌افزار و کارشناسی‌ارشد و دکترای مهندسی کامپیوتر- هوش مصنوعی را از دانشگاه علم و صنعت تهران دریافت کرده است. زمینه‌های مورد علاقه ایشان پردازش تصویر و ویدیو، داده‌کاوی است.