DOR :20.1001.1.2538161.1402.8.2.2.6

تابستان ۱۴۰۲، دوره ۸، شماره ۲، پیاپی ۲۹

دریافت مقاله: ۱۴۰۲/۰۱/۲۸

پذیرش مقاله: ۱۴۰۲/۰۵/۱۷

نوع مقاله: پژوهشی

تشخیص تعاملات انسان و شیء برمبنای ویژگی‌های استخراج شده از داده‌های عمق با استفاده از شبکه عصبی سیامی

منصوره رضائی

دانشجوی دکتری دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

پست الکترونیکی: mansooreh.rezaei@stu.yazd.ac.ir

مهدی رضائیان*

دانشیاردانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

پست الکترونیکی: mrezaeian@yazd.ac.ir

چکیده

تشخیص تعامل انسان و شیء (HOI) مجموعه‌ای از سه تایی‌های (انسان، شیء، تعامل) از یک تصویر را استخراج می‌کند. این حوزه، یکی از زمینه‌های تحقیقاتی رو به رشد در بینایی کامپیوتر است. علاوه بر اطلاعات دو بعدی مانند ظاهر انسان و اشیاء و موقعیت مکانی آن‌ها، وضعیت سه بعدی به خصوص در پیکربندی بدن انسان می‌تواند نقش مهمی در یادگیری تعاملات میان انسان و شیء داشته باشد. در این مقاله، مشخصه‌های بصری انسان، زمینه و شیء که به ترتیب از وضعیت‌های انسانی و پیکربندی‌های فاصله‌ای انسان و شیء به دست می‌آید، استخراج می‌شود. علاوه بر این، یک شبکه سیامی برای یادگیری ویژگی‌های ساختاری جفت‌های انسان- شیء استفاده می‌شود. شبکه سیامی بهبود یافته برای استخراج ویژگی‌های مشابه هدف، از ابرهای نقطه پیشنهاد می‌شود. شبکه سیامی بهبود یافته برای یافتن شباهت دو ورودی و استخراج ویژگی‌های مربوط به انسان و شیء پیشنهاد شده است این شبکه، توانایی تمایز را برای مرحله دوم که مرحله تشخیص تعاملات است، بهبود می‌بخشد. ما از یک بازنمایی داده بر اساس نگاشت سه بعد به دو بعد به صورت چند دیدی استفاده می‌کنیم، سپس از شبکه سیامی برای استخراج ویژگی‌های مربوط به این نگاشت به همراه جعبه شیء در یک توصیف‌گر محلی ۳۲ بعدی استفاده می‌کنیم. نتایج آزمایش‌ها روی مجموعه داده HICO-DET نشان می‌دهد که روش ما توانسته است نتیجه تعاملات را بهبود بخشد و معیار mAP را نسبت به روش DJ-RN به میزان ۲/۷۸ افزایش دهد.

واژه‌های کلیدی: تعاملات انسان و شیء، مشخصه‌های سه بعدی، شبکه سیامی، توانایی تمایز، نگاشت سه بعدی به دو بعدی.

مقدمه

تشخیص تعامل انسان و شیء[1] (HOI)، وظیفه‌ای برای پیش‌بینی تعاملات بین انسان و شیء است. این زمینه سعی می‌کند ابتدا، انسان‌ها و اشیاء را تعیین محل کند و سپس تعاملات بین آن‌ها را تشخیص دهد. اخیراً مطالعات زیادی در مورد درک صحنه بصری مانند تشخیص اشیاء [۱] و

1- Human-Object Interaction

* نویسندهٔ مسئول

تشخیص عمل [۲] انجام شده است. بسیاری از تحقیقات به تشخیص تعامل انسان و شیء پرداخته اند [۹-۳]. روش‌های سنتی تشخیص تعاملات انسان- شیء ویژگی‌های دو بعدی مانند ویژگی ظاهری انسان و شیء و روابط فاصله‌ای آن‌ها را بر اساس اطلاعات دو بعدی استخراج می‌کنند [۵، ۶، ۱۰]. در این روش‌ها ابتدا پیشنهادهای انسان- شیء مشخص می‌شود. برای تعیین محل مکان انسان و اشیاء از تکنیک‌های تشخیص اشیاء استفاده می‌شود. در مرحله بعد، پیشنهادها طبقه‌بندی شده و تعاملات بین انسان و اشیاء پیش‌بینی می‌شود. مجموعه داده‌هایی که اغلب در این زمینه استفاده می‌شوند V-COCO [۱۳] و HICO-DET [۸] هستند. در حالت دوبعدی، وضعیت انسان و همچنین روابط فاصله‌ای بین انسان و شیء نمی‌تواند به طور کامل ویژگی‌های انسان و شیء متعاملش را بازنمایی کند و به عنوان یک عیب محسوب می‌گردد. مشخصه‌های سه بعدی می‌توانند حاوی اطلاعات ارزشمندی باشند که روند تشخیص تعاملات را بهبود بخشند. بنابراین با استفاده از ویژگی‌های مربوط به داده عمق، فاصله بین انسان و شیء به طور کامل بازنمایی می‌شود. ابر نقطه[۳] و مش‌های[۴] مربوط به آن فضای زیادی را به‌خصوص در مجموعه داده‌های بزرگ مانند HICO-DET اشغال می‌کنند و از نظر زمان پردازش و پیچیدگی مشکلات زیادی را ایجاد می‌کنند. بنابراین، ما برای غلبه به مشکلات کمبود حافظه و پیچیدگی اطلاعات ابر نقطه‌ای از نگاشت ابرنقطه به صورت چند دیدی[۵] استفاده می‌کنیم. در این مطالعه، از مزایای داده‌های سه بعدی و مشخصه‌های عمق برای بهبود پیش‌بینی تعاملات بین انسان و شیء استفاده شده است. هدف این مقاله، استخراج پیکربندی سه بعدی بدن انسان بر اساس یک تصویر دو بعدی به منظور بازیابی اطلاعات انسان به جامع‌ترین شکل است که از HICO-DET به عنوان یک مجموعه داده معیار استفاده می‌شود[۸].

اخیرا شبکه‌های عصبی سیامی در حوزه یادگیری ماشین به طور گسترده مورد استفاده قرار گرفته‌اند[۳۰-۳۲]. شبکه‌های عصبی هم‌آمیختی سیامی در مواردی که تعداد رده‌ها زیاد است و تعداد نمونه‌های آموزشی در هر رده کم است به خوبی کار می‌کنند[۳۳]. شبکه‌های عصبی سیامی[۶] می‌توانند اندازه‌گیری شباهت بین جفت تصویر را بیاموزند و بر اساس آن شباهت تصاویر را تعیین کنند. یک شبکه عصبی سیامی از دو یا چند زیرشبکه یکسان استفاده می‌کند که ویژگی‌های یکسانی دارند. شبکه‌های فرعی می‌توانند پرسپترون‌های چند لایه[۷] (MLP)، شبکه‌های عصبی هم‌آمیختی[۸] (CNN) و غیره باشند. شبکه سیامی می‌تواند ویژگی‌های پایدار و دقیق رمزگذاری کند. با مقایسه کارایی روش پیشنهادی با روش‌های دیگر، مزایای زیر را می‌توان برشمرد:

- تعداد ابعاد در قسمت سه بعدی شبکه در شاخه ویژگی‌های انسان- شیء، ۳۲ است که نشان از ابعاد کم ویژگی‌های استخراج شده دارد.
- مش بدن انسان و ویژگی‌های سه بعدی مشکلاتی مانند پیچیدگی و حجم زیاد دارند که مشکلات زیادی را هم از نظر پیاده‌سازی و زمان اجرا و هم از نظر فضای اشغال شده ایجاد می‌کنند. به دلیل استفاده از نگاشت چنددیدی و نگاشت سه بعد به دو بعد، ورودی شبکه دو بعدی است که مشکلات کار با تصاویر سه بعدی را حل کرده و در مقایسه با آنها پیچیدگی مناسبی دارد.
- روش‌های موجود بر بازنمایی‌های ظاهر و وضعیت انسان در حالت دوبعدی تکیه می‌کنند، که مشخص کردن و بازیابی وضعیت بدن به طور کامل غیرممکن است. علاوه بر این، اطلاعات عمق نقش اساسی در تعیین وضعیت و ویژگی‌های ظاهری انسان ایفا می‌کند. در حالی که روش پیشنهادی از مزایای بازیابی سه‌بعدی بدن انسان استفاده می‌کند، اما نقطه ضعف پیچیدگی و فضای زیاد اشغال شده توسط داده‌های مش را ندارد.

2- http://www-personal.umich.edu/~ywchao/hico/
3- Point-cloud
4- Mesh
5- Multi-view projection
6- Siamese Neural Network
7- Multi-Layer Perceptron
8- Convolutional Neural Network

در این مطالعه برای اولین بار شبکه هم‌آمیختی سیامی برای استخراج ویژگی در حوزه تعاملات انسان- شیء پیشنهاد شده است.

در این مقاله برای اولین بار از داده‌های عمق به صورت کارا در حوزه تعاملات انسان- شیء استفاده شده است.

مجموعه داده HICO-DET شامل ٤٧/٧٧٦ تصویر (٣٨/١١٨ در مجموعه آموزشی و ٩/٦٥٨ در مجموعه آزمایشی)، ٦٠٠ دسته HOI ساخته شده توسط ٨٠ دسته شیء و ١١٧ رده فعل است. روش پیشنهادی توانسته است بر روی مجموعه داده معیار در حالت پیش‌فرض کامل، به mAP =٢٤/١٢ برسد. روش DJ-RN -که از ابرنقطه شیء و انسان در حوزه تعاملات انسان- شیء استفاده می‌کند- دارای mAP =٢١/٣٤ است که روش پیشنهادی نسبت به این روش معیار mAP را به میزان ٢/٧٨ افزایش داده است. نتایج نشان می‌دهد روش پیشنهادی در همه حالات نسبت به روش DJ-RN بهتر عمل می‌کند. علاوه بر این، پیچیدگی و بار محاسباتی ناشیء از ابر نقطه و مش انسان و شیء را کاهش داده است و در نتیجه حجم اشغال شده توسط ابر نقطه کاهش یافته و به دنبال آن زمان اجرای برنامه پایین آمده است. یک انسان قادر به تشخیص روابط بین اشیاء در یک تصویر هست. این روابط می‌تواند به یک ماشین هوشمند کمک کند تا معنای اصلی تصویر یا صحنه را تفسیر کند و بنابراین، یک قدم به درک دنیای واقعی و فهم عمیق صحنه نزدیک‌تر شود. تحقیقات به دست آمده در مورد تشخیص HOI می‌تواند به وظایف مختلفی مانند شرح تصویر[9] و بازیابی آن، بازیابی داده‌های بصری[10]، پاسخ به سؤالات بصری[11] و درک صحنه[12] در حوزه صنعتی (به عنوان مثال، ماشین‌های خودران) کمک کند [٢٧، ١٧]. علاوه بر این روش پیشنهادی در حوزه نظارت برای کمک به سالمندان و بیماران نقش موثری دارد. این مقاله با استفاده از داده عمق، گام نوینی به سمت بهبود پیش‌بینی تعاملات برداشته است و با استفاده از داده‌های سه‌بعدی وضعیت انسان و پیش‌بینی عمل او در رابطه با شیء را تسهیل می‌بخشند. در ادامه بخش‌های مختلف به این صورت سازمان‌دهی گردیده است: در بخش دوم مروری بر تعدادی از روش‌های مرتبط را خواهیم داشت و در بخش سوم روش پیشنهادی مطرح خواهد شد. سپس در بخش چهارم نتایج حاصل از روش پیشنهادی و همچنین مقایسه با روش‌های دیگر نشان داده شده است و نهایتاً در بخش پنجم نتیجه گیری آورده شده است.

۱. کارهای مرتبط

در سال‌های اخیر، تشخیص تعاملات انسان و شیء به یک زمینه مورد علاقه برای محققان تبدیل شده است. تکنیک‌های تشخیص شیء [١، ١١، ١٢، ١٥] یک جعبه محدودکننده ردهٔ خاص را در اطراف هر شیء و ردهٔ مربوطه تعیین می‌کنند. اما این تکنیک‌ها تعامل بین اشیاء را مشخص نمی‌کنند. برای تشخیص تعاملات انسان- شیء، اولین قدم تعیین محل اشیاء و در نتیجه استفاده از آشکارسازهای شیء است. گوپتا و همکاران برای نخستین بار، مسئله برچسب‌گذاری نقش معنایی بصری[13] را مطرح کردند[١٣]. تحقیقات آن‌ها به تعیین مکان شیء و انسان منجر شد. اخیراً، به لطف یادگیری عمیق، الگوریتم‌های تشخیص اشیاء و سپس مسئله تشخیص HOI به پیشرفت‌های قابل توجهی دست یافته‌اند[١١، ١٢]. چائو و همکاران، برای اولین بار از شبکه‌های عمیق برای مسئله تشخیص HOI استفاده کردند[٨]. آنها یک معماری چند جریانه که شامل جریان انسانی، جریان شیء و جریان روابط فاصله‌ای بین انسان و شیء است را پیشنهاد کردند که جفت جعبه‌ها را با استفاده از یک الگوی تعامل رمزگذاری می‌کند. در هر جریان، امتیازات برای هر دسته تعامل محاسبه می‌شود و رتبه نهایی با استفاده از ترکیب این رتبه‌ها استخراج می‌شود. علاوه بر این، چائو و همکاران، مجموعه داده HICO-DET را به عنوان یک

9- Image Captioning
10- visual data retrieval
11- visual question answering
12- scene understanding
13- visual semantic role labeling

مجموعه داده معیار جمع‌آوری کردند[۸]. به منظور بازیابی کامل اطلاعات از انسان، وضعیت بدن انسان نیز استخراج شد و به عنوان ورودی شبکه در نظر گرفته شد[۱۸-۱۴، ۱۶]. گائو و همکاران، یک شبکه توجه پیشنهاد کردند که در آن اطلاعات زمینه تصویر و ویژگی‌های ظاهری برای تعیین ویژگی نمونه استخراج شد[۷]. ویژگی نمونه شامل نشانه‌هایی در تصویر است که قسمت‌های مهم را مشخص می‌کند. در برخی از مطالعات، تصویر به عنوان یک گراف در نظر گرفته می‌شود که گره‌های آن انسان و اشیاء هستند و یال‌ها تعاملات را مدل‌سازی می‌کنند. یانگ و همکاران یک مدل تشخیص HOI مبتنی بر گراف را معرفی کردند. در شبکه آن‌ها، یک گره انسانی به عنوان یک گره مرکزی و بقیه گره‌ها به عنوان گره‌های معنایی مدل می‌شوند. این کار برای همه انسان‌های حاضر در صحنه انجام می‌شود و برای هر انسان یک گراف کاملاً متصل ساخته می‌شود. برای هر گره مرکزی و معنایی، بازنمایی ویژگی‌ها و روابط لبه‌ها طراحی شده است. در هر گره معنایی، مکان نسبی و اطلاعات دسته شیء نیز تعبیه شده است. سپس، یک شبکه هم‌آمیختی گراف[۱۴] برای به روزرسانی ویژگی‌ها و استخراج تعاملات انسان-شیء استفاده می‌شود[۲۰]. در مطالعات دیگری نیز از گراف برای مدل‌سازی تشخیص HOI استفاده کردند[۲۱، ۲۲]. اولوتان و همکاران استدلال فاصله‌ای نسبی و ارتباطات ساختاری بین اشیاء را برای تشخیص تعاملات انسان- شیء اتخاذ کردند[۲۳]. مطالعه آن‌ها ویژگی‌های بصری را از پیکربندی‌های انسان و شیء و پیکربندی فاصله‌ای بین جفت‌های انسان- شیء بازیابی می‌کند و از ارتباطات ساختاری بین این جفت‌ها با استفاده از پیچش‌های گراف استفاده می‌کند. لیو و همکاران یک تعامل دو وجهی را معرفی کردند[۲۴]. در این کار، با توجه به یک شیء مشخص در تصویر، پس از جفت شدن یک شیء با انسان، جفت‌های تولید شده یا تعاملی هستند یا غیر تعاملی. به طورکلی، جفت‌های تولید شده عمدتا غیرتعاملی هستند. بر اساس یک تعامل دو وجهی پیشین، یک حوزه تعامل[۱۵] برای مدل سازی توزیع تعاملی جهانی تصاویر HOI پیشنهاد شد. حوزه آموزش دیده باید با ملاحظات تصویر واقعی HOI سازگار باشد، بنابراین بر اساس تفاوت در حوزه تعاملی ذاتی در زمینه جفت‌های تعاملی در مقابل غیرتعاملی، ارتباط یا عدم ارتباط جفت‌های انسان و شیء مشخص می‌شود. با توجه به محدودیت نمونه‌ها در برخی از رده‌ها که به عنوان نادر شناخته می‌شوند، یادگیری بدون نمونه[۱۶] اتخاذ شد[۲۵، ۲۶]. اخیراً مبدل‌های[۱۷] تشخیص تعاملات انسان و شیء به دلیل سرعت اجرای بالا و یادگیری پایان به پایان[۱۸] مورد استفاده قرار می‌گیرند. مبدل‌هایی مانند QPIC [۲۷] و HOITrans [۲۸]، چندین سه‌گانه <انسان، شیء، تعامل> را در تصاویر با یک رمزگشای[۱۹] واحد مشخص می‌کنند. مبدل‌های دارای شاخه‌های موازی مانند HOTR [۴] و ASNet [۲۹]، اشیاء و تعاملات را به صورت موازی و هم‌زمان با استفاده از رمزگشاها پیش‌بینی می‌کنند و تطبیق بین اهداف مربوطه را برای تعیین پیش‌بینی‌ها انجام می‌دهند. اطلاعات دوبعدی مانند ظاهر انسان/اشیاء و مکان‌ها، به تنهایی نمی‌تواند به طور کامل ویژگی‌های انسان و شیء متعاملش را بازنمایی کند. بنابراین اطلاعات سه بعدی و وضعیت انسان در حالت سه بعدی به دلیل استقلال در دید بسیار مفید هستند. یانگ و همکاران، وضعیت انسان در حالت سه بعدی در حوزه تعاملات انسان- شیء را مورد استفاده قرار دادند[۱۴]. آن‌ها یک روش یادگیری نمایش مشترک دوبعدی- سه بعدی را پیشنهاد دادند. ابتدا، از روش تصویربرداری از بدن انسان به صورت تک دید برای به دست آوردن اشکال سه بعدی بدن، صورت و دست استفاده می‌شود. در مرحله بعد، مکان و اندازه شیء سه بعدی با توجه به پیکربندی فاصله‌ای دوبعدی انسان و شیء و گروه‌های مختلف اشیاء تخمین زده می‌شود. در نهایت، یک چارچوب یادگیری مشترک دوبعدی- سه بعدی برای یادگیری

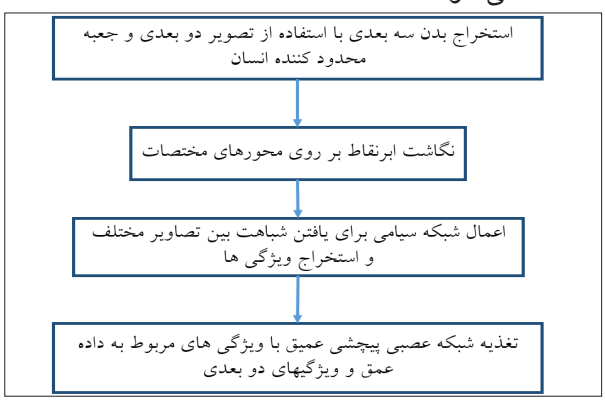
14- Graph Convolutional Network

15- Interactiveness Field
16- Zero-Shot Learning
17- Transformer
18- End-to-end
19- Decoder

نمایش HOI در نظر گرفته می‌شود. روش آن‌ها با استفاده از اطلاعات سه بعدی بازیابی شده پیش‌بینی تعاملات را به صورت چشمگیری بهبود می‌بخشد، اما به دلیل پیچیدگی بالای داده‌های سه بعدی و فضایی که اشغال می‌کنند دارای محدودیت‌هایی از جمله بالا بودن زمان اجرا است. همچنین برای پیاده‌سازی برنامه نیاز به سیستمی با فضای ذخیره‌سازی زیاد است. روش پیشنهادی توانسته است علاوه بر بهبود پیش‌بینی تعاملات، حجم اشغال شده توسط ابر نقطه و بار محاسباتی ناشی از آن را کاهش دهد و به دنبال آن برنامه با سرعت بالایی اجرا شود.

۲. روش پیشنهادی

در این مقاله، از شبکه‌های سیامی برای استخراج مشخصه‌های محلی مؤثر و قوی استفاده شده است. در این بخش، با استفاده از تصاویر دو بعدی و جعبه محدودکننده انسان، مولفه‌های سه بعدی استخراج شده و بازسازی سه بعدی انسان انجام می‌شود. در مرحله بعد، با توجه به نگاشت مش بدن انسان از سه محور مختصات، تصاویر دو بعدی استخراج می‌شود. سپس این نگاشت سه دیدی به همراه جعبه شیء متعامل با انسان مذکور به شبکه عصبی سیامی داده می‌شود تا ویژگی‌ها را استخراج کند. مراحل انجام شده در روش پیشنهادی به صورت زیر خلاصه می‌شود:



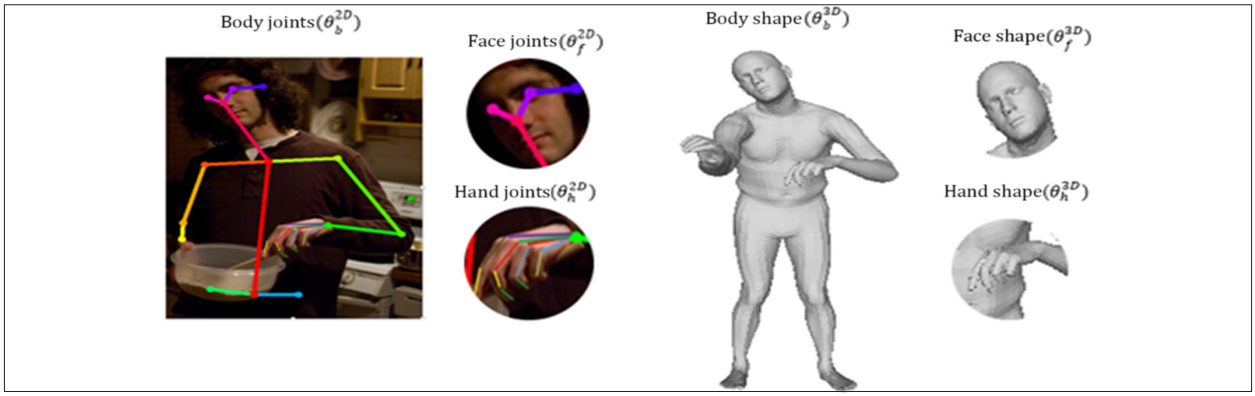
۲-۱. بازسازی انسان سه بعدی

برای اولین بار، یانگ و همکاران، وضعیت انسان در حالت سه بعدی را در حوزه تعاملات انسان- شیء مورد استفاده قرار دادند[۱۴]. برای استخراج ویژگی‌های سه‌بعدی انسان، بر اساس یک تصویر مشخص، در نخستین گام با استفاده از روش تشخیص شیء[20] [۱۲] و تخمین وضعیت انسان[21] [۳۴]، جعبه محدود کننده دوبعدی انسان b_h و شیء b_o و همچنین وضعیت دو بعدی بدن انسان $\theta^{2D} = \{\theta_b^{2D}, \theta_f^{2D}, \theta_h^{2D}\}$-که مفاصل بدن، مفاصل صورت و مفاصل دست هستند- استخراج می‌شود[۱۴]. سپس بر اساس b_h، b_o و θ^{2D}، پیکربندی سه بعدی استخراج می‌شود که به عنوان ورودی SMPLify-X [۳۵] برای استخراج انسان سه‌بعدی، یعنی پارامترهای شکل $\{\theta_b^{3D}, \theta_f^{3D}, \theta_h^{3D}, \beta, \varphi\}$ در نظر گرفته می‌شوند. روش SMPLify-X بر اساس وضعیت دوبعدی انسان و و جعبه‌های محدودکننده، بدن کامل بخصوص مفاصل دست و نشانه‌های چهره را بازسازی می‌کند. در این روش از هزاران پویش سه بعدی برای آموزش یک مدل جدید، یکپارچه و سه بعدی از بدن انسان استفاده می‌شود. در این روش، یک مدل سه بعدی از بدن انسان به نام SMPL-X آموزش داده می‌شود که به طور مشترک بدن، صورت و دست انسان را مدل‌سازی می‌کند. شکل ۱، تشخیص دوبعدی و سپس بازسازی سه بعدی بدن انسان را در یک تصویر نشان می‌دهد.

۲-۲. نگاشت مولفه‌های سه بعدی بر روی صفحات دو بعدی

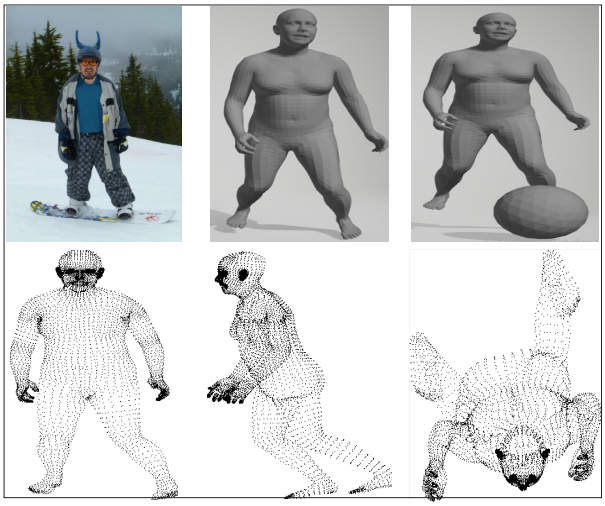
در این بخش نگاشتی از مش به صفحات دو بعدی صورت می‌گیرد که به کامل‌ترین شکل اطلاعات سه بعدی را با کمترین میزان پیچیدگی و حجم اشغال شده بازنمایی می‌کند. یکی از جنبه‌های اصلی داده‌های تصویر سه بعدی، نقشه عمق است. داده‌های سه بعدی حاوی اطلاعات ارزشمندی هستند، اما در عین حال معایبی نیز دارند. از جمله معایب آن‌ها می‌توان به فضای زیاد مورد نیاز برای ذخیره اطلاعات و پیچیدگی زیاد این داده‌ها که منجر به ناکارآمدی آن‌ها می‌شود اشاره کرد. راه

20- Object Detection
21- Pose Estimation



شکل (۱): وضعیت بدن انسان استخراج شده و پارامترهای شکل در حالت دو بعدی و سه بعدی توسط Open-Pose[۳۴] و SMPLify-X [۳۵]

غلبه بر این مشکلات استفاده از داده‌های سه‌بعدی در قالب بسیار کارآمدتر در برنامه‌ها است. یک مش شامل رئوسی است که توسط لبه‌هایی به هم متصل شده و صورت‌هایی به شکل چند ضلعی می‌سازد. بر اساس این اتصالات، یک ساختار داده، نحوه ذخیره هر عنصر و ویژگی‌های ارجاعی به همسایگان خود را تعریف می‌کند. نگاشت داده‌های سه‌بعدی در فضای دوبعدی جایگزین، نمایش دیگری از داده‌های مش است که در آن داده‌های تصویر شده بسیاری از ویژگی‌های کلیدی مش اصلی را کپسوله می‌کنند. نگاشت‌های متعددی وجود دارد که هر یک از آن‌ها داده‌های سه بعدی را با اطلاعات مخصوص به خود به فضای جایگزین منتقل می‌کند. نگاشت سه‌دیدی پیشنهادی، یک بازنمایی از مولفه‌های سه‌بعدی است که اطلاعات آن با ترکیب سه تصویر دوبعدی شیء از نماهای مختلف بازیابی می‌شود. بازنمایی داده‌های سه‌بعدی در این حالت منجر به یادگیری چندین مجموعه از ویژگی‌های مختلف برای کاهش نوفه، انسداد[۲۲] و مشکلات روشنایی[۲۳] می‌شود. در اینجا، ابتدا مش بدن انسان بر روی هر یک از محورهای مختصات y، x و z نگاشت می‌شود و سه تصویر دو بعدی استخراج می‌شود. شکل (۲)، نگاشت سه- دیدی از مش بدن انسان را به تصاویر دو بعدی نشان می‌دهد. سپس این سه تصویر که اطلاعات بدن انسان هستند به همراه جعبه محدودکننده شیء به عنوان ورودی شبکه عصبی هم‌آمیختی سیامی در نظر گرفته



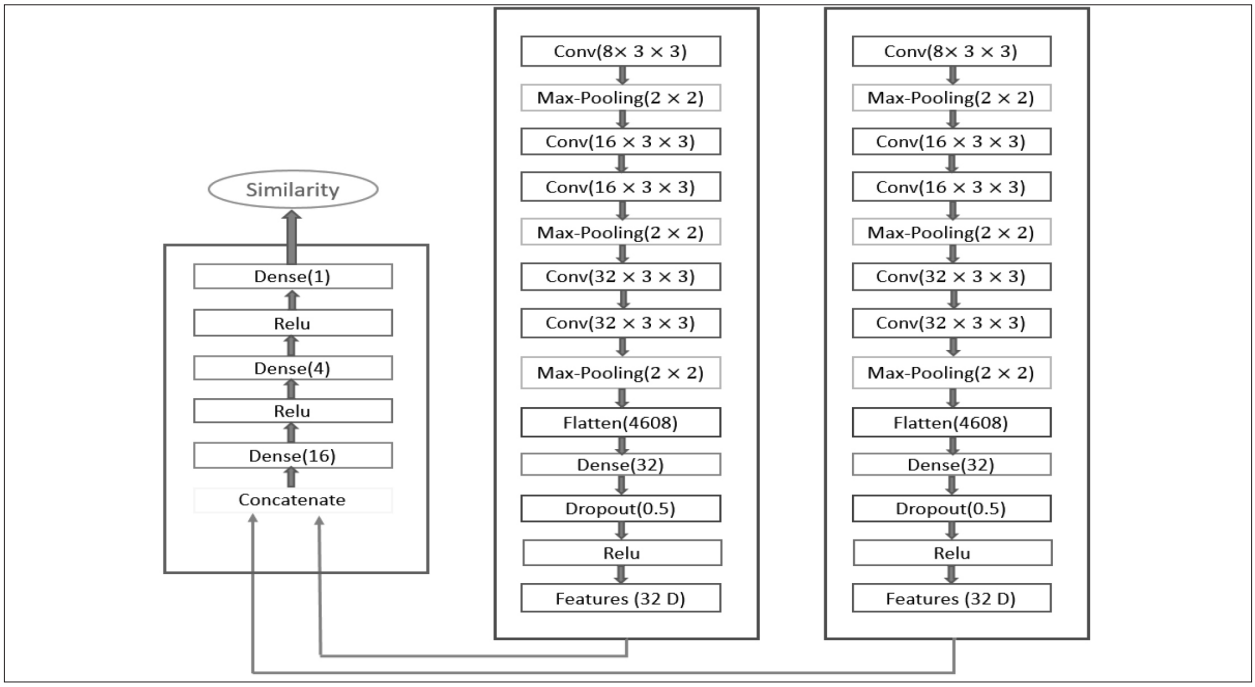
شکل (۲): نگاشت سه- دیدی مش بدن انسان به تصاویر دوبعدی. هر تصویر دوبعدی نشان دهنده یک انسان خاص است که از زاویه متفاوتی مشاهده می‌شود. ردیف اول به ترتیب تصویر اصلی، مش بدن انسان و پیکربندی فاصله‌ای انسان و شیء را مشخص می‌کند. ردیف دوم تصاویر انسان مرتبط را نشان می‌دهد که به سه محور نگاشت شده است.

می‌شوند.

۲-۳. استخراج ویژگی با استفاده از شبکه عصبی سیامی

در این مقاله شبکه‌های سیامی برای اندازه‌گیری شباهت و توصیف ویژگی‌های کارآمد و قوی پیشنهاد شده است. یک شبکه عصبی هم‌آمیختی سیامی از دو یا چند شبکه فرعی مشابه تشکیل شده است که شبکه‌های عصبی هم‌آمیختی (CNN) هستند. پارامترها و وزن‌ها در زیرشبکه‌ها یکسان بوده و در هر دو شبکه به روزرسانی می‌شوند. شبکه سیامی مشخصه‌های دو ورودی را مقایسه می‌کند. اگر جفت‌های ورودی برچسب ردهٔ یکسانی

22- Occlusion
23- illumination



شکل (۳): معماری شبکه هم‌آمیختی سیامی مورد استفاده در روش پیشنهادی

داشته باشند، شبکه سعی می‌کند فاصله آن‌ها را به حداقل برساند و اگر برچسب‌های متفاوتی داشته باشند، شبکه آن‌ها را از یکدیگر دور نگه می‌دارد. شبکه سیامی در مواردی که فقط تعداد محدودی داده از هر رده وجود دارد، به خوبی کار می‌کند [۳۶]. در روش پیشنهادی، به ازای هر جفت انسان و شیء، ماتریس تلفیقی از مولفه‌های انسان و شیء ساخته می‌شود که به عنوان ورودی شبکه سیامی در نظر گرفته می‌شود. این شبکه برای در نظر گرفتن اطلاعات بدن انسان و شیء در یک توصیفگر محلی[۲۴] ۳۲ بعدی اتخاذ شده است. در اینجا شبکه سیامی را در نظر گرفته می‌شود که شاخه‌های شبکه ویژگی‌ها را استخراج می‌کند و بالای شبکه یک تابع شباهت را مشخص می‌کند. شکل (۳)، معماری شبکه سیامی را برای استخراج مشخصه‌ها نشان می‌دهد. ورودی شبکه دو بردار با اندازه ۳۶۰۰ و خروجی آن، احتمالی است که میزان شباهت بین دو تصویر را نشان می‌دهد. معماری هر زیر شبکه که با مستطیل قرمز در شکل نشان داده شده است به شرح زیر است: لایه هم‌آمیختی[۲۵] (۳×۳×۸) یعنی ۸ هسته[۲۶] با اندازه ۳×۳، لایه Relu، لایه هم‌آمیختی (۳×۳×۱۶)، لایه Relu، لایه تجمیع حداکثری۲۷ (۲×۲)، لایه هم‌آمیختی (۳×۳×۱۶)، لایه Relu، لایه هم‌آمیختی (۳×۳×۳۲)، لایه Relu، لایه تجمیع حداکثری (۲×۲). سپس یک لایه مسطح[۲۸]، یک لایه متراکم[۲۹] با اندازه ۳۲، یک حذف[۳۰] (۰/۵) و تابع فعال سازی[۳۱] Relu اتخاذ می‌شود. در نهایت یک بردار ویژگی ۳۲ بعدی به عنوان خروجی زیر شبکه خواهیم داشت. در آخرین لایه متراکم از تابع فعال‌سازی سیگموئید[۳۲] استفاده می‌شود. در بقیه لایه‌های هم‌آمیختی و متراکم، تابع فعال‌سازی خطی اتخاذ می‌شود. تابع ضرر آنتروپی متقاطع دودویی[۳۳] و تابع معیار میانگین خطای مطلق[۳۴] استفاده شده است.

۲-٤. شبکه HOI چند جریانی

شبکه پیشنهادی از بلوک‌های دوبعدی و سه‌بعدی تشکیل شده است. در بخش دوبعدی شبکه، برای استخراج

24- Local Descriptor
25- Convolutional Layer
26- Kernel

27- Max-pooling Layer
28- Flatten Layer
29- Dense Layer
30- Dropout
31- Activation Layer
32- Sigmoid
33- Binary Cross Entropy
34- Mean absolute error

مشخصه‌ها، از ویژگی‌های ظاهری انسان و شیء و ویژگی‌های فاصله‌ای مربوط به آن‌ها استفاده می‌شود. ویژگی‌های ظاهری انسان و شیء به ترتیب به صورت f_h^{2D} و f_o^{2D} درنظر گرفته می‌شوند. برای تعیین جعبه‌های محدودکننده انسان و شیء، از الگوریتم Faster R-CNN با ستون فقرات ویژگی ResNet-50-FPN استفاده می‌شود. جعبه‌های انسانی با رتبه s_h بالاتر از ۰/۸ و جعبه‌های شیء با رتبه s_o بالاتر از ۰/٤ در نظر گرفته می‌شوند. ما از COCO[۳۷] که بر روی Faster RCNN از پیش آموزش‌دیده [۳۵۱۱]، برای استخراج مشخصه‌های تجمیع ROI[۳۶] از جعبه‌های شناسایی شده استفاده می‌کنیم[۳۷]. در بلوک دوبعدی، برای مشخصه‌های ظاهری انسان و شیء، از بلوک iCAN استفاده شده است[۷]. همان‌طور که ذکر شد، جریان فاصله‌ای نیز رابطه بین انسان و شیء را از نظر فاصله مشخص می‌کند. جریان فاصله‌ای مکان‌های نسبی دوبعدی انسان و شیء را در نظر می‌گیرد. نقشه فاصله‌ای، دو کانال دارد که شامل نقشه‌های انسان و شیء است. بر اساس جعبه‌های محدودکننده انسان و شیء، الگوی تعامل به صورت یک تصویر دودویی با دو کانال تولید می‌شود. کانال اول دارای مقدار ۱ در پیکسل‌های مشخص شده در جعبه محدودکننده انسان و مقدار ۰ در جای دیگر است. کانال دوم دارای مقدار ۱ در پیکسل‌هایی که جعبه محدود کننده شیء را مشخص می‌کند و مقدار ۰ در مکان‌های دیگر است. با توجه به جعبه محدود کننده انسان، روش OpenPose برای استخراج ۱۷ نقطه کلیدی بدن انسان به کار گرفته شده است[۳٤]. سپس، نقاط کلیدی با خطوطی –که مقادیر ۰/۱۵ تا ۰/۹۵ را دارند– برای مشخص کردن اعضای بدن مرتبط می‌شوند. به بقیه مناطق مقدار ۰ تخصیص داده شده است. در نهایت نقشه وضعیت به شکل جعبه‌ای با اندازه ٦٤×٦٤ استخراج می‌شود.در بلوک سه بعدی همان‌طور که در قسمت قبل توضیح داده شد، f_H^{3D}، تخمینی از سه‌بعدی بدن انسان و f_H^{3D-2D}، مشخصه‌هایی هستند که با استفاده از شبکه سیامی استخراج می‌شوند. معماری کلی روش پیشنهادی در شکل (٤) نشان داده شده است. طبقه‌بندی‌کننده‌ها در هر جریان ابتدا از دو لایه کاملاً متصل به اندازه ۱۰۲٤ و سپس یک لایه سیگموئید عبور می‌کنند. برای هر جریان دو بعدی و سه بعدی، رتبه‌ها به طور جداگانه محاسبه می‌شود. در جریان دوبعدی، رتبه نهایی به صورت ترکیبی از رتبه‌های جریان انسان، شیء و جریان فاصله‌ای محاسبه می‌شود.

$$S^{2D} = (S_H^{2D} + S_O^{2D}) \times S_{sp}^{2D} \quad (۱)$$

در جریان سه بعدی رتبه کلی به صورت زیر محاسبه می‌شود.

$$S^{3D} = S_H^{3D} + S_H^{3D-2D} \quad (۲)$$

رتبه نهایی، از جمع رتبه‌های بلوک‌های دو بعدی و سه بعدی به‌دست می‌آید.

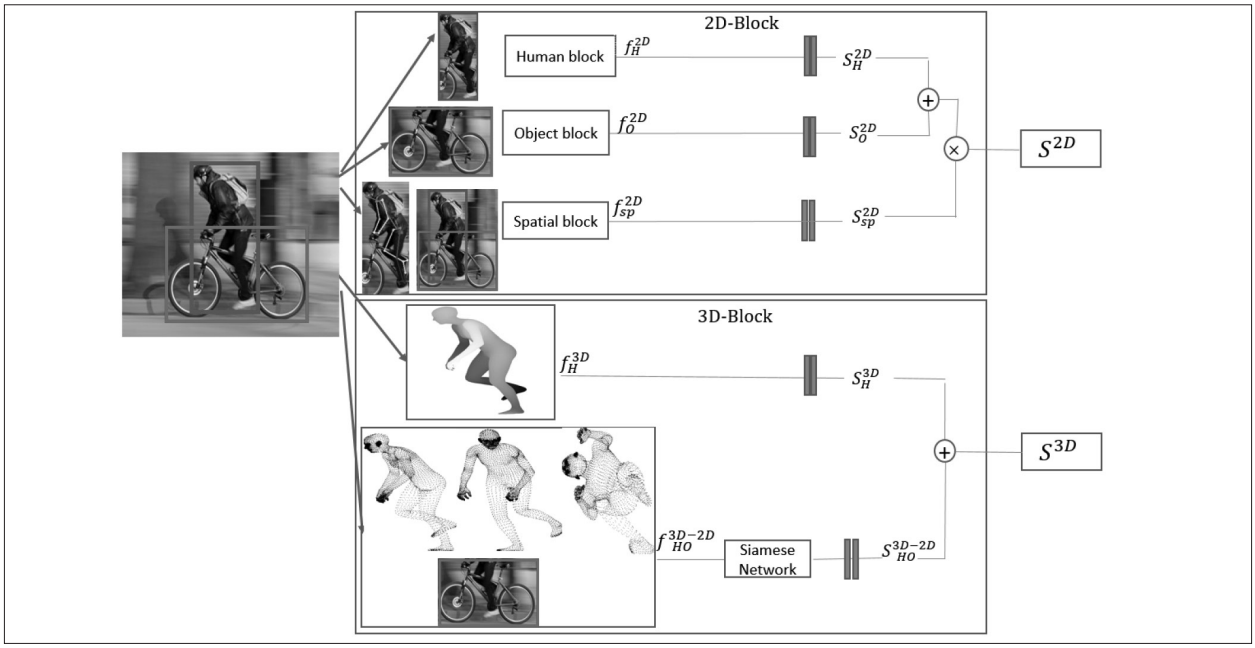
$$S = S^{3D} + S^{2D} \quad (۳)$$

۳. نتایج و بحث

برای ارزیابی روش پیشنهادی، از مجموعه داده HICO-DET[۸] استفاده شده است. این مجموعه داده، گسترده‌ترین مجموعه داده معیار در زمینه تشخیص تعاملات انسان–شیء است. مجموعه داده HICO-DET برای مشخص کردن داده HICO با حاشیه‌نویسی‌های جعبه محدود کننده برای وظیفه تشخیص تعاملات انسان– شیء جمع‌آوری شده است. مجموعه داده HICO-DET دارای تنظیمات حالت پیش‌فرض[۳۷] و حالت شناخته شده[۳۸] است. هر تنظیم دارای سه حالت مختلف است – حالت کامل[۳۹] با ٦۰۰ دسته تعاملات انسان–شیء، حالت نادر[٤۰] با ۱۳۸ دسته تعاملات انسان–شیء که کمتر از ۱۰ نمونه آموزشی دارند و حالت غیرنادر[٤۱] با مابقی ٤٦۲ دسته تعاملات انسان– شیء. این داده شامل ٤۷,۷۷٦ تصویر (۳۸/۱۱۸ در مجموعه آموزشی و ۹/٦۵۸ در مجموعه آزمایشی)، ٦۰۰ دسته

35- Pre-trained
36- Region of Interest

37- Default
38- Known
39- Full mode
40- Rare mode
41- Non-rare mode



شکل (۴): معماری روش پیشنهادی

HOI ساخته شده توسط ۸۰ دسته شیء و ۱۱۷ ردهٔ فعل است. ورودی f_H^{3D}، شامل مشخصه‌های استخراج شده از روش SMPLify-X هستند[۳۵]. خروجی f_H^{3D} یک بردار با اندازه ۸۵ می‌باشد. ورودی f_H^{3D-2D} نیز یک بردار با اندازه ۳۶۰۰ می‌باشد. این بردار، الحاقی از سه ماتریس نگاشت با اندازه ۳۰×۳۰ که نگاشت مش انسان بر روی سه محور مختصات هستند و همچنین جعبه محدود کننده شیء با تغییر اندازه ۳۰×۳۰ می‌باشد. تغییر شکل این چهار ماتریس به بردار و الحاق آن‌ها منجر به تولید یک بردار با اندازه ۳۶۰۰ می‌باشد که به عنوان ورودی شبکه سیامی در نظر گرفته می‌شود. خروجی حاصل از شبکه سیامی یک بردار مشخصه با اندازه ۳۲ می‌باشد. در برخی از تصاویر که در آن‌ها انسداد مشهود است، بازیابی سه بعدی بدن انسان ممکن است با شکست روبه‌رو شود، زیرا در هر تصویر، روش SMPLify-X قادر به بازسازی بدن انسان در حالتی است که سر، لگن، یک شانه و یک مفصل ران برای او تشخیص داده شود. در بلوک دو بعدی، برای استخراج ویژگی، شبکه بر اساس R-CNN Faster [۱۱] با ResNet-50 [۳۵] پیاده‌سازی می‌شود. بلوک دو بعدی دارای ساختار چند جریانی است که برای استخراج ویژگی‌های ظاهری انسان و شیء از روش iCAN [۷] و برای استخراج مشخصه‌های روابط فاصله‌ای انسان و شیء از تنظیمات روش [۱۶] استفاده شده است. برای استخراج بدن سه بعدی، در ابتدا از روش OpenPose [۳۴] برای تشخیص حالت دو بعدی بدن، صورت و دست‌ها استفاده می‌شود. خروجی OpenPose به همراه داده تصویر به الگوریتم SMPLify-X[۳۵] داده می‌شود تا بدن سه‌بعدی انسان را ثبت کند. در بلوک سه‌بعدی، مشخصه‌ها بر اساس پارامترهای SMPLify-X استخراج می‌شوند. این پارامترها عبارتند از مفاصل بدن، شکل صورت و دست‌ها، بیان صورت و بدن متشکل از فک، انگشتان و مفاصل بدن. برای شکل و بیان بدن که f_H^{3D} نامگذاری شده است، پارامترهای آنها به طور مستقیم استفاده می‌شود[۱۴]. در مرحله آموزش، شبکه پیشنهادی با تکرار ۴۰۰۰۰۰ با نرخ یادگیری[۴۲] ۰/۰۰۱، کاهش‌دهنده وزن[۴۳] ۰/۰۰۰۵ و مقدار تکانه[۴۴] ۰/۹ آموزش داده شد. پارامترهای تشخیص شیء طبق پیشنهاد چن و همکاران تنظیم شده است[۷]. میانگین دقت متوسط[۴۵] (mAP) به عنوان معیار ارزیابی در آزمایش‌ها استفاده شده

42- Learning Rate
43- Weight Decay
44- Momentum
45- Mean Average Precision

جدول (۱): نتایج به‌دست آمده از روش پیشنهادی و سایر روش‌ها بر روی مجموعه داده HICO-DET

روش	پیش‌فرض			شناخته شده		
	کامل	نادر	غیر نادر	کامل	نادر	غیر نادر
iCAN [۷]	۱۴/۸۴	۱۰/۴۵	۱۶/۱۵	۱۶/۲۶	۱۱/۳۳	۱۷/۷۳
Transferable [۱۶]	۱۷/۰۳	۱۳/۴۲	۱۸/۱۱	۱۹/۱۷	۱۵/۵۱	۲۰/۲۶
DJ-RN[۱۴]	۲۱/۳۴	۱۸/۵۳	۲۲/۱۸	۲۳/۶۹	۲۰/۶۴	۲۴/۶۰
VSGNet[۲۳]	۱۹/۸	۱۶/۰۵	۲۰/۹۱	–	–	–
۳D۲-DSiameseNet	۲۴/۱۲	۲۰/۹۱	۲۴/۸۹	۲۵/۰۲	۲۲/۹۸	۲۶/۹۱

است. در بینایی ماشین، میانگین دقت متوسط یک معیار ارزیابی رایج است که برای تشخیص شیء (به‌عنوان‌مثال بومی‌سازی[۴۶] و طبقه‌بندی[۴۷]) استفاده می‌شود. یک نتیجه سه‌گانه <انسان، عمل، شیء> به عنوان یک مثبت واقعی[۴۸] در نظر گرفته می‌شود اگر تعامل به درستی پیش بینی شود و معیار اشتراک بر اجتماع[۴۹] (IoU) بین جعبه انسان/شیء شناسایی شده و داده مرجع[۵۰] مربوطه بزرگتر از ۰/۵ باشد. برای محاسبه میانگین دقت متوسط باید ابتدا دقت متوسط[۵۱] (AP) محاسبه می‌شود. دقت متوسط، منطقه زیر منحنی دقت- فراخوان[۵۲] است. برای یافتن میانگین دقت متوسط باید دقت متوسط درون‌یابی‌شده[۵۳] محاسبه گردد. اگر IoU ≤ ۰,۵ باشد، پیش‌بینی مثبت[۵۴] است. همچنین، اگر چندین کشف از یک جسم شناسایی شود، اولین مورد را مثبت حساب می‌کند درحالی‌که بقیه منفی[۵۵] است. برای محاسبه دقت متوسط درون‌یابی‌شده، یک میانگین AP برای ۱۱ نقطه محاسبه می‌شود. ابتدا مقدار دقت از ۰ تا ۱ به ۱۱ نقطه تقسیم می‌شود. سپس، میانگین حداکثر مقدار دقت برای این ۱۱ مقدار صحت محاسبه می‌شود.

$$AP = \frac{1}{11} \times (AP_r(0) + AP_r(0.1) + \cdots + AP_r(1.0)) \quad (٤)$$

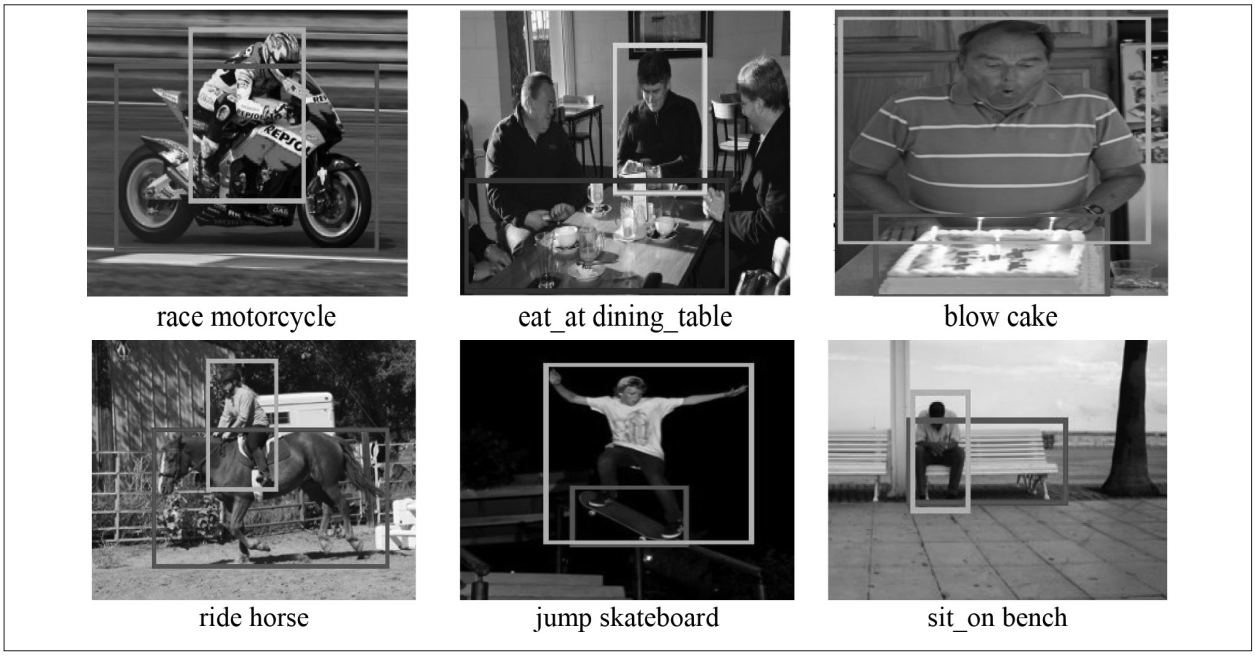
و در نهایت، mAP برای تشخیص اشیاء، میانگین AP محاسبه‌شده برای همه طبقات است. جدول (۱) نتایج به‌دست آمده در روش پیشنهادی و روش‌های دیگر روی مجموعه داده HICO-DET را نشان می‌دهد.

روش پیشنهادی که به عنوان 3D-2DSiameseNet نامگذاری شده است، توانسته است مشخصه‌های سه بعدی و داده‌های عمق را به شبکه تزریق کند تا اطلاعات کامل‌تری در رابطه با نوع عمل به‌دست آید. در روش DJ-RN از وضعیت سه بعدی بدن انسان استفاده می‌شود[۱۴]. روش آن‌ها به دلیل استفاده از مش‌های سه بعدی انسان و شیء پیچیدگی بالایی دارد و علاوه بر آن، ابر نقطه استخراج شده فضای زیادی را اشغال می‌کند که اجرای آن با مشکلات زیادی روبه‌رو است. روش پیشنهادی علاوه بر آن‌که توانسته است این مشکلات را حل کند، پیش‌بینی تعاملات را نیز بهبود بخشیده است. همان‌طور که از جدول (۱) مشخص است روش پیشنهادی و مشخصه‌های استخراج شده در روش پیشنهادی که در ردیف آخر پررنگ شده است، توانسته تا حد قابل توجهی مقدار mAP را بهبود بخشد. نتایج کیفی روش پیشنهادی در شکل (۵) نشان داده شده است. این شکل نشان می‌دهد که مدل پیشنهادی می‌تواند تعاملات انسان و شیء را در انواع مختلف پیش‌بینی کند. در پایین هر تصویر، تعامل و شیء متعامل با انسان را نشان می‌دهد. جعبه‌های انسان و شیء به ترتیب در کاردرهای مستطیلی سبز و آبی مشخص شده اند.

۴. نتیجه‌گیری

در این مقاله، مجموعه‌ای از مشخصه‌های نوین برای تشخیص HOI استخراج شد. مجموعه‌ای از مشخصه‌ها با پیچیدگی و ابعاد کم بر اساس شبکه سیامی معرفی شد که

46- Localization
47- Classification
48- True Positive
49- Intersection over Union
50- Ground Truth
51- Average Precision
52- Precision-Recall
53- Interpolated AP
54- Positive
55- Negative



شکل (۵): پیش‌بینی تعاملات در روش پیشنهادی در مجموعه داده HICO-DET. در پایین هر تصویر، تعامل و شیء متعامل با انسان را نشان می‌دهد. جعبه‌های انسان و شیء به ترتیب در کادرهای مستطیلی سبز و آبی مشخص شده اند.

از مشخصه‌های سه‌بعدی برای تولید ویژگی‌های قوی و کارآمد استفاده می‌کند. این شبکه سعی می‌کند از مش بدن انسان و جعبه محدودکننده شیء برای تولید مشخصه‌ها آموزش ببیند. در تنظیمات سه‌بعدی، حرکات و موقعیت بدن انسان به وضوح قابل مشاهده است. با این حال، مش بدن انسان و ویژگی‌های سه بعدی مشکلاتی مانند پیچیدگی و حجم زیاد دارند که مشکلات زیادی را هم از نظر پیاده‌سازی و زمان اجرا و هم از نظر فضای اشغال شده ایجاد می‌کنند. مجموعه‌ای از ویژگی‌های جدید بر اساس نگاشت چنددیدی سه بعد به دو بعد ارائه شد که هم حاوی اطلاعات کامل و ارزشمندی هستند و هم پیچیدگی کمی دارند. استفاده از این نگاشت نه تنها باعث رمزگذاری اطلاعات هندسی غنی و دقیق می‌شود، بلکه به مشکل بازنمایی سه بعدی بدن انسان و پیچیدگی نیز پرداخته می‌شود. با استفاده از روش پیشنهادی، ورودی بلوک سه‌بعدی شبکه پیشنهادی، دو بعدی است، بنابراین این معماری بر بسیاری از چالش‌های ناشی از داده‌های سه بعدی غلبه کرده است. روش‌های موجود بر بازنمایی‌های ظاهر و وضعیت انسان در حالت دوبعدی تکیه می‌کنند، که مشخص کردن و بازیابی وضعیت بدن به طور کامل غیرممکن است. علاوه بر این، اطلاعات عمق نقش اساسی در تعیین وضعیت و ویژگی‌های ظاهری انسان ایفا می‌کند. در حالی که روش پیشنهادی از مزایای بازیابی سه‌بعدی بدن انسان استفاده می‌کند، اما نقطه ضعف پیچیدگی و فضای زیاد اشغال شده توسط داده‌های مش را ندارد.

مراجع

1. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie , Feature pyramid networks for object detection, in Proc. IEEE Conf. Comput. Vision and Pattern Recognit, pp. 2117–2125 ,2017.
2. Gupta, Saurabh, and Jitendra Malik, Visual semantic role labeling., arXiv preprint arXiv:1505.04474, 2015.
3. Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C. and Sang, N., Mar: Masked autoencoders for efficient action recognition. arXiv preprint arXiv:2207.11660,2022.
4. Kim, Bumsoo, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim, Hotr: End-to-end human-object interaction detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 74-83. 2021.
5. Wang, Tiancai, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun, Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4116-4125. 2020.
6. G. Gkioxari et al., Detecting and recognizing human-object interactions, in Proc. IEEE Conf. Comput. Vision and Pattern Recognit., pp. 8359–8367, 2018.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20113-20122, 2022.
25. L. Shen et al., Scaling human-object interaction recognition through zero-shot learning, in IEEE Winter Conf. Appl. Comput. Vision (WACV), IEEE, pp. 1568–1576, 2018.
26. Sarullo, Alessio, and Tingting Mu. Zero-shot human-object interaction recognition via affordance graphs. arXiv preprint arXiv:2009.01039, 2020.
27. Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10410–10419, 2021.
28. Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen-Wei, End-to-end human object interaction detection with hoi transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11825– 11834, 2021.
29. Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9004– 9013, 2021.
30. Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou, ABCNN: Attention-Based Convolutional Neural Network for Modeling sentence pairs. TACL, pp. 259–272, 2016.
31. Huang, Kai, Peixuan Qin, Xuji Tu, Lu Leng, and Jun Chu. SiamCAM: A Real-Time Siamese Network for Object Tracking with Compensating Attention Mechanism. Applied Sciences, Vol. 12, no. 8, 3931, 2022.
32. He, Anfeng, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4834-4843. 2018.
33. S. Chopra, R. Hadsell, and Y. LeCun, Learning a similarity metric discriminatively, with application to face verification. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 539–546, 2005.
34. Z. Cao et al., Realtime multi-person 2D pose estimation using part affinity fields, in Proc. IEEE Conf. Comput. Vision and Pattern Recognit., pp. 7291–7299, 2017.
35. G. Pavlakos et al., Expressive body capture: 3D hands, face, and body from a single image, in Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit., pp. 10975–10985, 2019.
36. Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, vol. 2, No. 1, 2015.
37. T.-Y. Lin et al., Microsoft COCO: common objects in context, Lect. Notes Comput. Sci. 8693, pp. 740–755, 2014.
38. 40. K. He et al., Deep residual learning for image recognition, in Proc. IEEE Conf. Comput. Vision and Pattern Recognit., pp. 770–778, 2016.
7. C. Gao, Y. Zou, and J.-B. Huang, ICAN: instance-centric attention network for human-object interaction detection, https://arxiv.org/abs/1808.10437, 2018.
8. Y.-W. Chao et al., Learning to detect human-object interactions, in IEEE Winter Conf. Appl. Comput. Vision (WACV), IEEE, pp. 381–389, 2018.
9. Su, Zhan, Yuting Wang, Qing Xie, and Ruiyun Yu, Pose graph parsing network for human-object interaction detection. Neurocomputing 476,pp. 53-62, 2022.
10. Fang, Hao-Shu, Jinkun Cao, Yu-Wing Tai, and Cewu Lu, Pairwise body-part attention for recognizing human-object interactions. In Proceedings of the European conference on computer vision (ECCV), pp. 51-67. 2018.
11. S. Ren et al., Faster R-CNN: towards real-time object detection with region proposal networks, in Proc. 28th Int. Conf. Adv. Neural Inf. Process. Syst., pp. 91–99, 2015.
12. He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.
13. S. Gupta and J. Malik, Visual semantic role labeling, https://arxiv.org/abs/1505.04474, 2015.
14. Y.-L. Li et al., Detailed 2D-3D joint representation for human-object interaction, in Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit., pp. 10166–10175 (2020).
15. W. Liu et al., SSD: single shot multibox detector, Lect. Notes Comput. Sci. 9905, 21–37, 2016.
16. Y.-L. Li et al., Transferable interactiveness knowledge for human-object interaction detection, in Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit., pp. 3585–3594, 2019.
17. B. Wan et al., Pose-aware multi-level feature network for human object interaction detection, in Proc. IEEE/CVF Int. Conf. Comput. Vision, pp. 9469–9478, 2019.
18. Naveed, Humza, Fareed Jafri, Kashif Javed, and Haroon Atique Babri. Driver activity recognition by learning spatiotemporal features of pose and human object interaction. Journal of Visual Communication and Image Representation, Vol. 77, 2021.
19. Su, Z., Wang, Y., Xie, Q. and Yu, R., Pose graph parsing network for human-object interaction detection. Neurocomputing, 476, pp.53-62, 2022.
20. Yang, Wenhao, Guanyu Chen, Zhicheng Zhao, Fei Su, and Hongying Meng. iCGPN: Interaction-centric graph parsing network for human-object interaction detection. Neurocomputing 502, pp. 98-109, 2022.
21. Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In ECCV, 2018.
22. Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction. arXiv preprint arXiv:2010.10001, 2020.
23. O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, VSG-Net: spatial attention network for detecting human object interactions using graph convolutions, in Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit., pp. 13617–13626, 2020.
24. Liu, X., Li, Y.L., Wu, X., Tai, Y.W., Lu, C. and Tang, C.K., Interactiveness Field in Human-Object Interactions. In