مقاله مروری

مجله ماشین‌بینایی و پردازش‌تصویر
سال یازدهم، شماره اول، بهار ۱۴۰۳

مرور مقایسه‌ای تخمین حالت سه بعدی دست مبتنی بر پردازش تصویر

محمد مفرح[۱]، میرهادی سیدعربی[۲*]، بهزاد مظفری تازه کند[۳]، شهره کسایی[۴]

چکیده

با گسترش روزافزون فناوری و ظهور دستگاه‌های هوشمند در زندگی امروزی، برقراری ارتباط تعاملی با این ابزارهای هوشمند به امری اجتناب ناپذیر تبدیل شده است. از میان روش‌های متعدد برقراری ارتباط تعاملی که شامل استفاده از صوت و لباس‌های مجهز به سنجش‌گر حرکتی و همچنین روش‌های مبتنی بر تصویر است که هر کدام برمبنای شناسایی و نظارت بر عضو خاصی از بدن متمرکز هستند، روش‌های مبتنی بر دست انسان به دلیل دارا بودن ویژگی‌های خاص دست، از اهمیت مضاعفی در این حوزه برخوردار است و لذا تخمین حالت سه بعدی دست با استفاده از پردازش تصویر به یکی از جذاب‌ترین و چالشی‌ترین زمینه‌های پژوهشی مرتبط با ارتباط تعاملی انسان با ماشین تبدیل شده است. در این مقاله، روش‌های مختلف تخمین حالت سه بعدی دست با تاکید بر روش‌های مبتنی بر پردازش تصویر بررسی شده و نقاط قوت و ضعف آنها بیان شده و مورد مقایسه قرار گرفته است. این مقایسه بر اساس روش‌های سنتی و همچنین روش‌های مرتبط با یادگیری عمیق انجام شده است. همچنین پایگاه داده‌های مورد استفاده در تخمین حالت دست نیز معرفی شده و ویژگی‌های هرکدام از آنها در کاربردهای مختلف مورد بررسی قرار گرفته است.

کلید واژه ها

تخمین حالت سه بعدی دست، ارتباط تعاملی انسان-ماشین، مفصل‌های دست.

۱- مقدمه

با توجه به پیشرفت‌های بسیار سریع و موثر دهه‌های اخیر در حوزه الکترونیک و کامپیوتر و هوشمند شدن بسیاری از ابزارهای مورد استفاده روزانه بشر، نیاز به برقراری ارتباط بین انسان و ماشین روز به روز در حال گسترش است. روش‌های متعددی برای ارتباط تعاملی بین انسان و ماشین ارائه شده‌اند که استفاده از سیگنال‌های صوتی، بکارگیری ابزارهای پوشیدنی، استفاده از سنجش‌گرهایی که بر روی بدن کاربر سوار می‌شوند و روش‌های مبتنی بر پردازش تصویر از آن جمله هستند. در این میان محدودیت‌هایی مانند محدود کردن حرکت‌های طبیعی انسان به سبب سوار شدن ابزارهای جانبی بر روی بدن، استفاده از آن‌ها را در فعالیت‌های روزانه محدود می‌کند. دست انسان به دلیل دارا بودن ویژگی‌های خاص خود از اهمیت مضاعفی در این حوزه برخوردار است و لذا تخمین حالت سه بعدی دست با استفاده از پردازش تصویر یکی از جذاب‌ترین زمینه‌های پژوهشی این حوزه به حساب می‌آید. با توجه به افزایش هوشمندسازی ابزارهای کنونی، کاربردهای زیادی را برای تخمین حالت سه بعدی دست در زندگی امروزه می‌توان مشاهده نمود که شامل محدوده وسیعی از کاربردها، از کنترل ابزارها با حرکات دست گرفته تا واقعیت مجازی[۱]، واقعیت افزوده[۲]، واقعیت ترکیبی[۳]، بازی‌های کامپیوتری، دستیار راننده[۴] و حتی استفاده در کاربردهای امنیتی مانند تشخیص هویت با

این مقاله در دی‌ماه ۱۴۰۲ دریافت شد؛ در فروردین‌ماه ۱۴۰۳ بازنگری و در خردادماه همان سال پذیرفته گردید.

[۱] دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
m.mofarreh@tabrizu.ac.ir
[۲] استاد، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
seyedarabi@tabrizu.ac.ir
[۳] استاد، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
mozaffary@tabrizu.ac.ir
[۴] استاد، دانشکده مهندسی کامپیوتر، دانشگاه شریف، تهران، ایران
kasaei@sharif.edu

نویسنده مسئول: میرهادی سیدعربی

[۱] Virtual Reality
[۲] Augmented Reality
[۳] Mixed Reality
[۴] Drive Assistant

حرکات دست، است. کارهای ابتدایی در این حوزه با تکیه بر تصاویر و ویدیوهای دوبعدی انجام شده است که به دلیل این که رنگ دست دارای اطلاعات کافی نیست، دارای دقت لازم نبوده و حساسیت زیاد آن‌ها به شرایط محیطی استفاده از آن‌ها را در کاربردهای روزمره عملا ناممکن می‌سازد. با معرفی دوربین‌های با توانایی ثبت سه بعدی، قابلیت تخمین مختصات فضایی مفصل‌های اسکلتی بدن و به طور ویژه، مختصات سه بعدی دست فراهم شده است. در این مطالعه، روش‌های مختلفی که برای حل مساله تخمین حالت دست با بکارگیری فن[۱]های مبتنی بر پردازش تصویر مورد استفاده قرار گرفته‌اند بررسی شده و نقاط قوت و ضعف هرکدام مورد مطالعه قرار گرفته است. ساختار این مقاله به این صورت است که پس از مقدمه در بخش اول، مساله تخمین حالت سه بعدی دست بیان شده و ابزارهای مرتبط با آن و همچنین چالش‌ها و معضلات این حوزه پژوهشی بررسی می‌شود. در بخش دوم، روش‌های پرکاربرد مورد استفاده برای تخمین حالت دست با استفاده از پردازش تصویر معرفی می‌گردد و نقاط قوت و ضعف هرکدام بیان می‌شود. در بخش سوم پایگاه داده‌های استفاده شده در این حوزه مورد بررسی قرار می‌گیرند و بخش چهارم حاوی نتیجه‌گیری روش‌های بررسی شده و چشم‌اندازهای پیش روی این حوزه می‌باشد.

۲- بیان مساله

استفاده از پردازش تصویر و ویدیو برای برقراری ارتباط تعاملی انسان - ماشین به دلیل عدم افزودن محدودیت‌های اضافی به کاربر برای اجرای حالت‌ها، نسبت به سایر روش‌ها از اقبال بیشتری در کاربردهای روزمره برخوردار است. کارهای ابتدایی در این حوزه با تکیه بر تصاویر و ویدیوهای دوبعدی تهیه شده با دوربین‌های رنگی RGB معمولی[۲] انجام شده است که اطلاعات فراهم شده با این دوربین‌ها فاقد بعد سوم یعنی ژرفا[۳] هستند و دقت این روش‌ها برای استفاده در کاربردهای روزمره کافی نیست. در استفاده از اطلاعات رنگ علاوه بر بحث دقت، حساسیت این سیستم‌ها به تغییرات نورپردازی محیط و همچنین زاویه نورپردازی در کنار مشکلات مربوط به تشخیص ناحیه دست به دلیل تشابه رنگ پوست دست با سایر اعضای بدن، می‌تواند عملکرد کل سیستم را مختل کند. برای رفع این محدودیت، دوربین‌هایی که قابلیت تهیه تصاویر سه بعدی را دارا هستند، مانند دوربین‌های RGB استریو و یا دوربین‌های ToF[۴] معرفی شدند که بار محاسباتی بسیار بالا و قیمت گزاف این دوربین‌ها عملا استفاده از آن‌ها را در کاربردهای ارتباط تعاملی انسان – ماشین[۵] (HMI) محدود می‌کند. در سال ۲۰۱۰ و با معرفی دوربین کینکت مایکروسافت[۶] و متعاقب آن نسخه‌ی دوم این دوربین در سال ۲۰۱۲، پژوهش در این حوزه وارد فاز جدیدی شد و این دوربین به عنوان یک دوربین ارزان قیمت برای تهیه تصویر و ویدیوهای سه بعدی که کاربرد بسیار وسیعی در شناسایی حالت‌های بدن دارند مورد استفاده قرار گرفت [۱]. این دوربین علاوه بر استخراج اطلاعات مربوط به رنگ، اطلاعات ژرفا را نیز تولید می‌کند که به تصاویر حاصل اصطلاحا تصاویر رنگ-ژرفا RGBD[۷] گفته می‌شود. شکل ۱ نمونه‌ای از تصاویر تهیه شده با دوربین کینکت را نمایش می‌دهد.

شکل (۱) : تصویر RGBD تهیه شده با دوربین کینکت. بالا: تصویر رنگی، پایین: تصویر ژرفا [۲]

در کنار دوربین کینکت، مایکروسافت و چند شرکت دیگر نرم‌افزارهایی را نیز به منظور تهیه و استخراج اطلاعات تکمیلی مانند اطلاعات مربوط به مختصات سه بعدی مفصل‌های اسکلتی ارائه نمودند که SDK مایکروسافت[۸] و OpenNI نمونه‌هایی از این نرم‌افزارها هستند. این نرم‌افزارها قادرند اطلاعات مختصات سه بعدی ۲۰ یا ۱۵ مفصل[۹] بدن انسان را از تصاویر RGBD حاصل از کینکت استخراج نمایند که این اطلاعات در کاربردهای متعددی از جمله در سیستم‌های تعاملی انسان – ماشین، تخمین حالت بدن، تخمین حالت حرکت بدن، تخمین حالت حرکت دست، بازسازی سه بعدی انسان[۱۰] و ... نقشی اساسی ایفا می‌کنند.

---

۱ Technique
۲ Red-Green-Blue
۳ Depth
۴ Time of Flight
۵ Human Machine Interaction
۶ Microsoft Kinect v.1
۷ RGB-Depth
۸ Microsoft Software Development Kit
۹ Joint
۱۰ 3D-Human Reconstruction

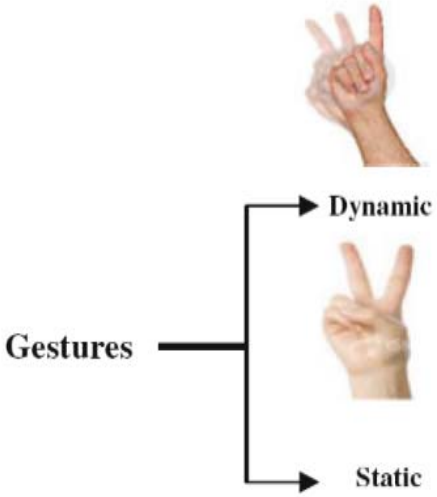
۳- تخمین حالت دست[۱] و شناسایی حالت حرکت دست[۲]

در مقایسه با سایر اعضای بدن انسان، دست‌ها به دلیل درجه آزادی حرکتی و تفکیک‌پذیری بالا، در حوزه‌ی HMI از اهمیت بسزایی برخوردارند. با توجه به وجود مفصل‌های زیاد در دست و تنوع و پیچیدگی بالا و در نتیجه امکان اجرای حالت‌های بسیار زیاد با استفاده از دست، که هم می‌تواند مفید و جذاب و در عین حال به دلیل وقوع مداوم خودانسدادی، تغییر زاویه دید، وضوح مکانی پایین، شباهت بالای انگشتان به هم، خطاهای به وقوع پیوسته در تفکیک انگشتان و نویزپذیری زیاد تصاویر دست چالش برانگیز باشد [۳] و [۴]. در هر دو کاربرد ذکر شده، با داشتن یک تصویر یا ویدیو از دست، ابتدا حالت دست تخمین زده می‌شود. این عمل معادل تخمین مختصات سه بعدی مفصل‌های ۱۶ یا ۲۱ و یا ۳۶ تایی دست است که در شکل ۲ نشان داده شده است. به عبارت دیگر، ورودی این مرحله، تصویر یا ویدیویی از دست و خروجی آن، تخمینی از مفصل‌ها خواهد بود.

پس از تعیین مختصات مفصل‌ها، حالت حرکت دست دسته‌بندی می‌شود که در آن معنای حالت حرکت دست استخراج می‌گردد. به عبارت دیگر، ورودی و خروجی این مرحله به ترتیب مختصات سه بعدی مفصل‌ها و معنای حالت حرکت دست که یک مفهوم است خواهد بود [۵].

شکل (۲) : مدل‌های ۱۶، ۲۱ و ۳۶ مفصلی دست انسان [۶] و [۷]

حالت حرکت‌های بدن و در حالت خاص، حالت حرکت دست را از لحاظ استفاده از اطلاعات توالی زمانی می‌توان در دو نوع دسته‌بندی نمود: حالت حرکت‌های ایستا[۳] و حالت حرکت‌های پویا، که در شکل ۳ نشان داده شده است.



شکل (۳) : تفاوت حالت حرکت‌های ایستا و پویا [۸]

حالت حرکت‌های ایستا حالت حرکت‌های ساده‌ای هستند که کاربر بصورت ثابت در مقابل دوربین انجام می‌دهد [۹]. حالت حرکت‌های پویا حالت حرکت‌های پیچیده‌تری هستند که می‌توان آن‌ها را بصورت یک توالی از حالت حرکت‌هایی در نظر گرفت که پشت سر هم در مقابل دوربین انجام می‌شود [۱۰]. چندین مساله در تحلیل حالت حرکت‌های پویا مطرح می‌شود از جمله شناسایی نقطه شروع و نقطه پایان حالت حرکت، استفاده از اطلاعات زمانی برای شناسایی آن و سایر عامل[۴]هایی که عمدتا بر پایه زمان استوار است.

در یک طبقه‌بندی جامع‌تر که از ترکیب دسته‌های فوق حاصل می‌شود، می‌توان حالت حرکت‌ها را در چهار گروه دسته‌بندی نمود [۱۱]:

- Gesticulation: که در ضمن صحبت و برای تاکید مورد استفاده قرار می‌گیرند.
- Emblems: که قسمتی از یک کد حالت حرکت قابل فهم مجازی مانند علامت ok را تشکیل می‌دهد.
- Pantomimes: که در غیاب صحبت استفاده می‌شود ولی کد شده نیست.

Sign Language: یا زبان اشاره که کلا به جای صحبت جایگزین می‌شود (مانند زبان ناشنوایان).

شناسایی حالت حرکت دست برای کاربردهای تعاملی عمدتا در استفاده از Gesticulation، Emblems و Pantomimes قرار می‌گیرد که برای تولید فرم جدیدی از تعامل انسان – ماشین استفاده می‌شود.

---

[۱] Hand Pose Estimation
[۲] Hand Gesture Recognition
[۳] Static Gesture
[۴] Parameter

تخمین حالت دست در حالت کلی به سه روش انجام می‌گیرد:

- روش‌های مبتنی بر شکل: در این روش‌ها، شکل مشاهده شده و ویژگی‌های آن با شکل‌های موجود در پایگاه داده مقایسه شده و حالت دست تخمینی برابر خواهد بود با نزدیکترین شکل موجود در پایگاه داده از نظر ویژگی‌های انتخابی.
- روش‌های مبتنی بر مدل سه بعدی دست: در این روش، هدف کمینه‌سازی تفاوت مدل‌های سه بعدی دست با شکل مشاهده شده است.
- روش‌های مبتنی بر مفصل‌های اسکلتی: در این روش، بخش‌های مختلف دست تشخیص داده می‌شوند و از روی آن مفصل‌های اسکلتی مشخص می‌شوند. در واقع، تخمین حالت دست با توجه به موقعیت سه بعدی این مفصل‌ها انجام می‌گیرد.

- **مدل سازی سه بعدی دست**

مدل سازی سه بعدی دست و تعقیب دست، حرکات و حالت‌های تفکیک شده‌ی دست را به خوبی بیان می‌کند. این دو، فناوری‌های پایه‌ی بسیاری از کاربردهای ارتباط تعاملی انسان و ماشین به شمار می‌روند. از یک نقطه نظر، روش‌های تخمین حالت دست را می‌توان در دو گروه روش‌های تمایزدهنده[۱] و روش‌های زایشی[۲] تقسیم بندی نمود [۱۲] و [۱۳]. علاوه بر این مدل سازی سه بعدی ترکیبی که مدل‌سازی تمایزدهنده و زایشی را با هم ترکیب می‌کند نیز مورد استفاده قرار گرفته است [۱۴].

- **مدل‌سازی سه بعدی تمایزدهنده**

روش‌های تمایزدهنده، مدل سه بعدی دست را محدود به درجه آزادی مشخصی نمی‌کنند ولی به جای آن دسته‌بندها را آموزش می‌دهند تا بطور معکوس ویژگی‌های ظاهری تصویردانه[۳]های مربوط به دست را به عامل‌های دست ناشناخته (مانند برچسب بخش، عامل حالت و ...) نگاشت کنند. دسته‌بندها معمولا بصورت برون‌خط[۴] و با استفاده از نمونه‌های آموزشی بسیار زیاد آموزش داده می‌شوند. همچنین بسیاری از روش‌های فوق از دسته‌بندهای مبتنی بر درخت تصمیم‌گیری استفاده می‌کنند تا سرعت تخمین را در هر قاب[۵] بصورت مجزا افزایش دهند.

- **مدل‌سازی سه بعدی زایشی**

روش‌های زایشی در کارهای اخیر تعقیب و مدل‌سازی سه بعدی دست بسیار عمومیت پیدا کرده‌اند. این روش‌ها، تعقیب دست مبتنی بر مدل نیز نامیده می‌شوند که در آن‌ها تلاش می‌شود یک مدل با درجه آزادی سه بعدی با یک داده‌ی دست مشاهده شده، تطابق یابد. توجه شود که تصاویر دو بعدی و سه بعدی دست را می‌توان برای تخمین مدل سه بعدی دست استفاده کرد.

- **اجزای سیستم تشخیص حالت دست**

یک سیستم تخمین حالت حرکت دست و تخمین حالت دست مبتنی بر ویدیو یا تصاویر ژرفا، از قسمت‌ها و اجزای زیر تشکیل می‌شود.

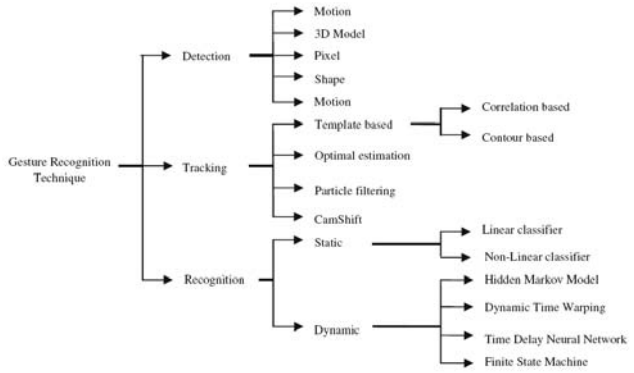
الف- بخش جمع آوری داده

ب- مکان‌یابی دست

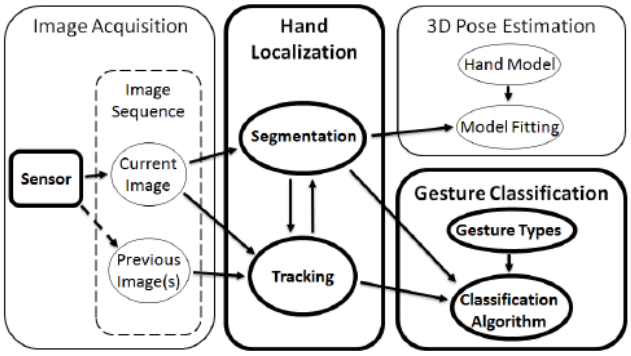
ج — تخمین حالت سه بعدی دست

د — دسته‌بندی حالت حرکت

شکل ۴ بخش‌های مختلف اجزای سیستم تشخیص حالت دست و شکل ۵ نحوه ارتباط این قسمت‌ها را نشان می‌دهند.



شکل (۴) : بخش‌های مختلف یک سیستم تخمین حالت حرکت دست[۸]



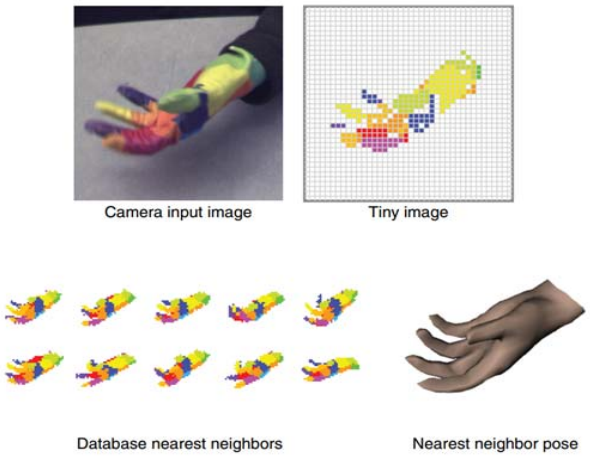
شکل (۵) : ارتباط بخش‌های مختلف یک سیستم تخمین حالت حرکت دست [۱۱]

تا قبل از پیدایش و توسعه فن یادگیری عمیق، روش‌های مرسوم یادگیری ماشین و بینایی ماشین برای حل مساله تخمین حالت دست استفاده می‌شد. با معرفی روش‌های مبتنی بر یادگیری عمیق، پژوهش‌ها در این حوزه وارد فاز جدیدی گردید. در ادامه مقاله، بررسی روش‌های استفاده شده در دو فاز قبل و بعد از پیدایش فن‌های یادگیری عمیق صورت می‌گیرد. لازم به ذکر است که برخی

---

۱ Discriminative Approaches
۲ Generative Approaches
۳ Pixel
۴ Offline
۵ Frame

پژوهش‌ها با الهام گرفتن از تخمین حالت بدن و سپس تعمیم آن به تخمین حالت دست انجام شده‌اند.
در [۱۵] Wang و Popovic از یک دستکش رنگی استفاده کرده و سپس با استفاده از روش نزدیک‌ترین همسایه، مکان هر رنگ در تصویر دست را تعیین نموده و از روی آن مکان هر بخش خاص از دست را مشخص می‌کنند. شکل ۶ مراحل اجرای این روش را نشان می‌دهد.



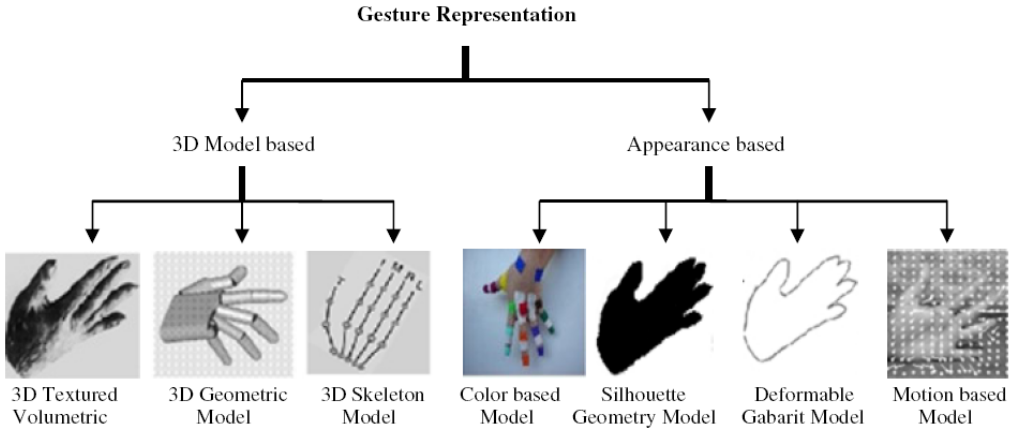
شکل (۶) : مراحل اجرای روش استفاده شده در [۱۵]

- **مدل‌سازی مبتنی بر تکه**[۱]

در مقاله‌ی مشهوری که Shotton و همکارانش در سال ۲۰۱۱ منتشر کرده‌اند [۱۶] یک فن شناسایی حالت بدن انسان را ارائه کرده‌اند که بخش‌های اسکلتی بدن انسان را با استفاده از یک سنجش‌گر کینکت استخراج می‌کند. آن‌ها جنگل‌های تصمیم‌گیری تصادفی با ۳۰۰،۰۰۰ تصویر آموزشی ساختگی[۲] برای آموزش هر درخت تصمیم‌گیری را تشکیل دادند تا هر تصویردانه از تصویر ژرفای بدن را برای بخش‌های مختلف بدن برچسب گذاری نمایند. با الهام از این کار، Keskin, Furkan, Yunus و Lale در [۱۷] ایده‌ی Shotton و همکارانش را برای تخمین حالت سه بعدی دست با استفاده از تصاویر ژرفا مورد استفاده قرار دادند. آن‌ها از یک مدل مش پوسته سه بعدی[۳] به عنوان تولید کننده‌ی حالت دست برای تولید ۲۰۰،۰۰۰ تصویر ساخته شده‌ی آموزشی استفاده کردند تا درخت‌های تصمیم‌گیری را برای بخش‌های مختلف دست آموزش دهند. در طی مرحله‌ی آموزش، ویژگی‌های نمونه‌های آموزشی به درخت‌های تصادفی تصمیم‌گیری[۴] وارد می‌شوند تا آفست‌های بهینه برای هر گره برگ[۵] آموزش داده شود. امتیاز نهایی دسته‌بندی با میانگین‌گیری از همه‌ی توزیع‌ها در جنگل هر بخش به دست می‌آید. تصویردانه‌ای به عنوان قسمت خاص دست برچسب‌گذاری می‌شود که دارای بالاترین امتیاز باشد. برای به دست آوردن اسکلت دست، از mean shift برای به دست آوردن مرکز هر بخش استفاده می‌شود [۱۶] و [۱۷]. با این حال، جنگل‌های بخش‌ها با یک پایگاه داده عظیم از تصاویر ساخته شده و با نیاز به یک حافظه بسیار بزرگ آموزش داده می‌شوند.
یک مدل مرز برچسب گذاری شده[۶] برای تخمین حالت حرکت توسط Yao و Fu معرفی شده است [۱۸]. در این روش ابتدا تصویر ژرفا با استفاده از ویژگی مکان، به بخش‌های مختلف دست دسته‌بندی می‌شود و سپس این نتایج برای برچسب گذاری نقاط برای فراهم کردن اطلاعات بیشتر برای تطابق، مورد استفاده قرار می‌گیرد. به عنوان یک روش تکمیلی، Liang, Yuan, Thalmann و Zhang استفاده از ICP[۷] را برای ساختن یک ارتباط زمانی مابین قاب قبلی و قاب کنونی برای جداسازی دست پیشنهاد دادند و لبه‌های بخش‌های مختلف دست را به عنوان قیدهای اضافه برای بهبود نتایج مورد استفاده قرار دادند [۱۹]. از آنجایی که این روش برای جداسازی دست وابستگی زیادی به مرجع زمانی دارد لذا به طور ذاتی به خطاهای تعقیب بسیار حساس خواهد بود بخصوص اینکه در این روش فرض می‌شود که تغییرات در قاب‌های متوالی بسیار جزئی هستند.
به عنوان یک روش غیر مبتنی بر مدل، Keskin, Kirac, Kara و Akarun یک روش چندلایه مبتنی بر جنگل‌های تصمیم‌گیری تصادفی را برای جداسازی بخش‌های مختلف دست پیشنهاد کرده‌اند تا مفاهیمی همچون تغییرات حالت را نیز دربرگیرد [۲۰]. آن‌ها حالت‌های دست را به دسته‌های الگوی مرتبط[۸] تخصیص داده‌اند و از تخمین‌گر حالتی که منحصرا برای آن دسته آموزش داده شده است استفاده کرده‌اند. ابتدا عامل‌های پیکربندی دست در الگوهای موجود در پایگاه داده آموزشی با استفاده از خوشه‌بندی طیفی[۹] خوشه‌بندی شده و به k دسته مختلف تقسیم‌بندی می‌شوند. در طی فرآیند یادگیری، یک دسته‌بند RDF برای دسته‌بندی شکل دست SCF[۱۰] آموزش داده می‌شود که کل تصویر ژرفای دست را به k دسته دسته‌بندی می‌کند. یک دسته‌بند RDF منحصر به فرد GEN[۱۱] بر روی هرکدام از k زیر بخش‌های پایگاه داده اصلی برای جداسازی دست به انگشت‌ها و کف دست و با استفاده از ویژگی مورد استفاده در [۱۶] آموزش داده می‌شود. در مرحله آزمون ابتدا دسته‌بند SCF به تصویر ژرفای دست اعمال می‌شود و نتیجه آن برای انتخاب دسته‌بندهای GEN به منظور تفکیک دست مورد استفاده قرار می‌گیرد.

---

۱ Part based modelling
۲ Synthesized
۳ 3D Skinned Mesh Model
۴ Randomized Decision Tree
۵ Leaf Node
۶ Labeled Contour Model
۷ Iterative Closest Point
۸ Shape Classes
۹ Spectral Clustering
۱۰ Shape Classification Forest
۱۱ Global Expert Network



شکل (۷) : انواع روش‌های توصیف حالت حرکت [۸]

اگرچه گزارش‌ها از بهبود قابل توجهی نسبت به روش Shotton حکایت دارد اما چهارچوب چندلایه‌ی RDFها به وضوح فرض را بر این گرفته‌اند که پایگاه داده مورد استفاده برای آموزش، حاوی الگوهای بسیاری است که دارای عامل‌های پیکربندی دست مشابهی هستند که با واقعیت مسائل تفکیک دست در سناریوهای حرکت دست با درجه آزادی کامل، فاصله زیادی دارد. Hernandez و همکارانش در [۲۱] نتایج دسته‌بندی تصویردانه‌ای Shotton را با اعمال بهینه‌سازی برش گراف[۱] بهبود داده‌اند که در آن 5.96% بهبود در مقایسه با نتایج Shotton گزارش شده است. با توجه به اینکه در آن مقاله هیچ اشاره‌ای به زمان مورد نیاز محاسبات نشده است لذا کارآیی این روش از نقطه نظر زمان اجرای الگوریتم می‌تواند مورد شک واقع شود.

- **برچسب گذاری تکه بهینه[۲]**

Keskin, Furkan, Yunus و Lale بر مبنای کارهای قبلی خود در [۱۷]، جنگل دسته‌بندی شکل SCF را برای دسته‌بندی شکل دست و نه بخش‌های دست پیشنهاد دادند [۲۰]. با توجه به اینکه روش‌های تمایزدهنده بر مبنای تنها یک قاب استوار هستند، موضوعات مربوط به زمان برای آن‌ها موضوعیت ندارد و می‌توان این روش‌ها را برای کاربردهای بلادرنگ مورد استفاده قرار داد ولی نیازمند حجم عظیمی از داده‌های آموزشی با کیفیت بالا هستند تا شناسایی کننده‌های بخش‌های مختلف دست را آموزش دهند. عدم وجود محدودیت‌های حرکتی آن‌ها را در برابر برخی موارد مانند خودانسدادی نامقاوم می‌سازد.

- **مدل‌سازی با تصاویر دو بعدی**

در برخی کارهای ابتدایی، پیشنهاد استفاده از روش‌های زایشی برای تخمین حالت دست سه بعدی با استفاده از تصاویر تک رنگ دو بعدی ارائه شده است [۲۲] و [۲۳]. در [۲۳] Gorce و Paragios دست را بصورت یک درخت حرکتی بند بند[۳] AKT با ۲۸ درجه آزادی مدل‌سازی کرده‌اند. از آنجایی که مدل‌سازی سه بعدی دست با استفاده از تصاویر تک رنگ دوبعدی بار محاسباتی بسیار زیادی دارد لذا کاربردهای آن محدود است.

- **مدل‌سازی دست سه بعدی تکی**

Oikonomidis, Kyriazis و Argyros یک روش برای بازیابی حالت سه بعدی دست با استفاده از تطبیق مدل دست ۲۶ درجه آزادی با دیدهای دوربین‌های چندگانه را در [۲۴] ارائه داده‌اند. اجرا کردن[۴] سطح سه بعدی دست با بکارگیری یک کره و یک استوانه‌ی بریده شده به عنوان الگوهای پایه انجام شده است. در مقاله آنها، روش‌هایی برای تعقیب دست سه بعدی زایشی بر اساس سنجش‌گر کینکت ارائه شده و استفاده از اطلاعات ژرفا و سایه‌نمای دست موجب کارآیی بهتر تابع هدف شده و درنتیجه ساده‌تر شدن سیستم گردیده است. همچنین با استفاده از قابلیت‌های سازگاری پذیری سنجش‌گر کینکت، مقاومت در برابر تغییرات نور بیشتر می‌شود. در [۲۵] با استفاده از هیستوگرام شیب زاویه[۵] HOG و هیستوگرام جریان[۶] HOF، روشی جهت تخمین حالت سه بعدی دست مستقل از دید و مقاوم به انسداد در یک جریان از فریم‌های تصویر حاوی شکل دست، معرفی می شود. برخلاف کارهای پیشین که از اطلاعات کلی تصویر استفاده می‌کنند، آنها از روش بخش‌بندی محلی تصویر استفاده کرده و ویژگی ظاهری و جنبشی هر بخش را توسط دو فن HOG و HOF استخراج می‌کنند و در ادامه، با استفاده از طبقه‌بند غیرخطی SVM، نشان داده‌اند که روش آنها در برابر انسداد از روش‌های پیشین مقاوم‌تر است. شکل ۷ انواع روش‌های توصیف حالت حرکت دست را که در دو گروه و هفت زیرگروه تقسیم‌بندی شده است، نشان می‌دهد.

---

۱ Graph-cut Optimization

۲ Efficient part labeling

۳ Articulated kinematic tree

۴ Render

۵ Histograms of Oriented Gradients

۶ Histogram of Optical Flow

۴- جمع‌آوری داده[۱]

برای اخذ داده‌های سه بعدی، دوربین‌های متعددی معرفی شده است که سه مورد از پرکاربردترین دوربین‌ها عبارتند از:

- **Microsoft Kinect**

نسخه‌های ۱ و ۲ دوربین کینکت توسط مایکروسافت به ترتیب در سال‌های ۲۰۱۰ و ۲۰۱۲ معرفی شدند که شامل یک دوربین رنگ RGB با وضوح VGA[۲] (۶۴۰×۴۸۰) و یک دوربین ژرفا با وضوح QVGA[۳] (۳۲۰×۲۴۰) است. هر دوی این دوربین‌ها قادر هستند که تصاویر ویدیویی را با نرخ ۳۰ قاب در ثانیه ثبت کنند. کینکت قادر است با کمک نرم افزارهای جانبی خود مختصات سه بعدی مفصل‌های اسکلتی کل بدن را اخذ و محاسبه نماید. این دوربین برای تخمین ژرفا از امواج مادون قرمز ساخت یافته استفاده می‌کند، به این صورت که یک پروژکتور مادون قرمز، آرایه‌ای از نقاط غیر یکنواخت حجیم را به روی صحنه می‌تاباند و یک گیرنده‌ی مادون قرمز انعکاس این پرتوها را از صحنه ثبت می‌کند. از آنجایی که الگو و فاصله‌ی نقاط تابانده شده معلوم است، پردازش‌گرهای داخلی دوربین قادر خواهند بود با مقایسه فاصله و ساختار نقاط بازتابانیده شده با پرتوهای تابانده شده، ژرفای اجسام در صحنه را مشخص نمایند. این دوربین برای عملکرد مطلوب، دارای محدودیت فاصله‌ی حداقل ۱۲۰ سانتی متر و حداکثر ۳۵۰ سانتی متری است.

- **Leap Motion**

برخلاف کینکت که اطلاعات ژرفای کل بدن را ثبت می‌کند، Leap Motion بر موقعیت‌یابی دقیق سه بعدی دست تمرکز کرده است. این سنجش‌گر قادر است با دقت 0.01 mm دست و انگشتان را آشکارسازی کند. پس از معرفی آن در سال ۲۰۱۳، چندین کار بر اساس این سنجش‌گر انجام شده است که بیشتر بر روی زبان اشاره متمرکز شده‌اند [۲۶] تا [۲۸].

- **سنجش‌گر ToF**

قبل از پیدایش کینکت و Leap Motion، سنجش‌گرهای ToF برای اندازه‌گیری ژرفا مورد استفاده قرار می‌گرفتند. در اصل نسخه دوم کینکت نیز نوعی سنجش‌گر ToF به شمار می‌رود. برای حالت حرکت‌های ایستا، این سنجش‌گرها در برابر پس‌زمینه[۴]های شلوغ و تغییرات اندازه و راستای دست بسیار مقاوم هستند. از این نوع سنجش‌گر در کارهای [۲۹] تا [۳۲] و به خصوص به منظور تخمین زبان‌های اشاره‌ی پیچیده استفاده شده است.

در قسمت جمع‌آوری داده بر اساس نوع سنجش‌گر مورد استفاده، یک سری از تصاویر ژرفای RGBD تهیه می‌شود. بسته به اینکه سیستم شناسایی حالت حرکت برای تخمین حالت حرکت‌های پویا یا ایستان طراحی شده باشد، یک یا چند تصویر پشت سر هم، مورد استفاده‌ی بخش‌های آتی قرار خواهد گرفت.

برای مکان‌یابی دست، باید قسمت‌های مختلف تصویر از هم تفکیک شوند. یکی از مزایای اصلی تصاویر ژرفا نسبت به تصاویر رنگی در مرحله بخش‌بندی آشکار می‌شود. در کاربردها وقتی از کاربر خواسته می‌شود در برابر دوربین ایستاده و حالت حرکت خاصی را اجرا کند، عمل بخش بندی دست به سادگی با اعمال یک آستانه‌گذاری ساده در محور z (ژرفا) قابل انجام است. از این فن در کارهای [۳۳] تا [۳۸] استفاده شده است. آستانه‌گذاری دست به این معنی است که دست، نزدیکترین قسمت تصویر به دوربین است و این آستانه‌گذاری را می‌توان به دو صورت از پیش تعیین شده توسط کاربر و یا به صورت نزدیکترین نقاط به دوربین تعیین کرد. به منظور کاهش حساسیت به نویز می‌توان محدوده‌ای برای سطح در نظر گرفته شده برای دست (یا تعداد تصویردانه‌هایی که به عنوان تصویردانه‌های دست در مرحله آستانه‌گذاری در نظر گرفته می‌شود) را در نظر گرفت. این بهینه‌سازی در کارهای [۳۹] و [۴۰] اعمال شده است. بهینه‌سازی بعدی را می‌توان در تعیین مکان دست بر اساس مکان سایر اعضای بدن و نه لزوما نزدیکترین تصویردانه‌ها به دوربین دانست. Van den Bergh و Van Gool در [۴۱] و Droeschel, Stückler و Behnke در [۴۲] از ابزار موجود در OpenCV برای شناسایی موقعیت مکانی صورت استفاده کرده و بر اساس آن محتمل‌ترین قسمت برای "دست بودن" را تعیین کرده‌اند. Biswas و Basu از آستانه‌گذاری برای جداسازی بدن از پس‌زمینه استفاده کرده و بجای تصمیم‌گیری‌های سخت برای تعیین بخش‌های مختلف بدن، از بافت‌نگاشت[۵] ژرفا استفاده نموده است تا نویز کاهش یافته و پیوستگی نواحی مربوط به اعضای بدن حفظ شود [۴۳].

بدون استفاده از اطلاعات ژرفا برای بخش بندی دست‌ها، روش معمول استفاده از رنگ پوست [۴۴] و [۴۵] و استفاده از دسته‌بندهای پشت سر هم روی ویژگی‌های شبه هار[۶] [۴۶] و [۴۷] است. روش مبتنی بر رنگ پوست به شدت در برابر شرایط نورپردازی ضعیف و یا تغییرات زاویه نورپردازی آسیب پذیر است، حتی وقتی که از روش‌های تغییرناپذیر در برابر تغییر نورپردازی استفاده شود. در برخی کارها نیز از ترکیب آستانه‌گذاری ژرفا و اطلاعات رنگ برای بهبود دقت بخش بندی دست استفاده شده است [۴۸].

در حالت‌های خاص می‌توان از روش‌های اختصاصی مخصوص آن شرایط خاص برای بخش بندی دست بهره برد. Jojic با معلوم بودن تصویر پس‌زمینه، با یک تفریق ساده، عمل بخش بندی را انجام داده است [۴۹]. در دو مطالعه دیگر، Park [۵۰] و Raheja

---

[۱] Data Acquisition
[۲] Video Graphics Array
[۳] Quarter Video Graphics Array
[۴] Background
[۵] Histogram
[۶] Haar-like

[۵۱] در ابتدای اجرای حالت حرکت از کاربر می‌خواهند تا دست خود را تکان دهد و آستانه‌گذاری ژرفا در محدوده‌ی تصویردانه‌هایی با بیشترین حرکت اعمال می‌شود.

تعقیب دست: تعقیب دست اطلاعات زمانی را علاوه بر اطلاعات مکانی اخذ می‌کند و لذا برای حالت حرکت‌های پویا مورد استفاده قرار می‌گیرد [۵۲] و [۵۳]. با معرفی کینکت و SDK ارائه شده با آن و همچنین نرم‌افزار OpenNI، قابلیت تعقیب اجزای بدن با این بسته‌های نرم‌افزاری فراهم شده است و با توجه به مختصات مفصل‌های ۲۰ نقطه‌ای که این نرم‌افزارها ارائه می‌دهند، تعقیب قسمت‌های مختلف بدن از جمله دست فراهم شده است. کارهای [۵۲] و [۵۴] تا [۵۷] مستقیما از اطلاعات فراهم شده با این نرم‌افزارها برای تعقیب دست استفاده کرده‌اند.

یکی از روش‌های معمول برای تخمین حالت حرکت‌های پویا استفاده از مدل مارکوف پنهان یا HMM[۱] است [۵۸] که بر مبنای یادگیری احتمالات گذار بین حالت‌های مختلف بدن انسان (دست) است. با توجه به اینکه حالت حرکت‌های پویا شامل مجموعه‌ی متوالی از حالت حرکت‌های ایستا است، مدل مارکوف پنهان ابزارهای مفیدی را برای بهره‌گیری از اطلاعات زمانی بین قاب‌ها فراهم می‌سازد. استفاده از روش‌های جنگل‌های تصادفی مشروط CRF[۲] یکی دیگر از ابزارهایی است که برای تحلیل دنباله‌های زمانی پشت سر هم استفاده می‌شود ولی همه این روش‌ها فرض را بر این می‌گیرند که تعداد حالت‌های هر حرکتی به روشنی مشخص است [۵۹]. به منظور دوری از چنین محدودیتی، روش پیچش زمانی پویا DTW[۳] برای استفاده در تخمین حالت حرکت توسط Reyes پیشنهاد شده است [۶۰]. روش DTW اجازه می‌دهد که دو دنباله زمانی مفروض که ممکن است در زمان باهم اختلاف داشته باشند با هم هم‌تراز شوند. این اختلاف زمانی به نحوه اجرای حالت حرکت توسط افراد مختلف بر می‌گردد. هزینه این هم‌ترازی می‌تواند به عنوان مشخصه ظاهری حالت حرکت مورد استفاده قرار گیرد. مهمترین کاری که توسط Reyes ارائه شده است معرفی یک معیار وزن دهی ویژگی در چهارچوب DTW است که کارآیی تخمین حالت حرکت دست را بهبود می‌بخشد. مقاوم بودن این روش برای تخمین حالت حرکت‌های مختلف و همچنین عملکرد بسیار خوب آن در شناسایی ابتدا و انتهای یک حالت حرکت نشانگر بهبود قابل ملاحظه کارآیی در استفاده از این روش دارد.

یکی از روش‌های بسیار کاربردی و مهم برای استخراج حالت سه بعدی دست با استفاده از تصاویر ژرفا، استفاده از جنگل‌های تصمیم تصادفی یا RDF است و استفاده از RDF و نسخه‌های مختلف آن بطور وسیعی در کاربردهای تخمین حالت بدن معمول شده است. این ایده بعدها توسط پژوهشگران پخته‌تر شد تا در تخمین چالشی‌تر حالت دست مورد استفاده قرار گیرد. الگوریتم تخمین حالت بدن به تحقیقات Shotton و همکارانش نسبت داده می‌شود [۱۶]. کار آن‌ها توسط Girshick توسعه بیشتری داده شد [۶۱] و در کنار استفاده از RDF، هر تصویردانه برچسب دار یک بردار جابجایی را برای هر مفصل پیشنهاد می‌کند و بردار جابجایی میانگین، مکان نهایی هر مفصل را تعیین می‌کند. این روش بطور موثری دقت مکان مفصل‌ها را بخصوص برای مفصل‌های داخلی بدن افزایش می‌دهد حتی در موارد وقوع خودانسدادی.

برای تخمین حالت دست "یک دست تنها"، روش کلی بالا به پایین مبتنی بر مدل که از یک مدل سه بعدی دست برای تطبیق داده‌های تست استفاده می‌کند مورد استفاده قرار می‌گیرد. این روش‌ها با تطبیق دادن مدل، مسائلی همچون خودانسدادی، تغییر زاویه دید و محدودیت‌های حرکتی را مدیریت می‌کنند. با توجه به ویژگی‌های فوق این روش‌ها معمولا در مواقعی که دست با یک یا چند شیء دیگر در ارتباط باشند نیز مورد استفاده قرار می‌گیرند. با این حال این روش‌ها به یک مرحله راه‌اندازی بسیار دقیق نیازمندند [۶۲] و [۶۳]، هم برای مکان دست و هم برای سایر اندازه‌گیری‌ها. اگر خطایی در مرحله راه‌اندازی اولیه صورت گیرد بازیابی این روش‌ها بسیار سخت خواهد بود.

از کارهای دیگری که در این زمینه انجام شده است می‌توان به کار Tang، Chang، Tejani و Kim اشاره کرد که در آن از ساختار LRF[۴] استفاده کرده‌اند که حاوی مجموعه‌ای از درخت‌های تصمیم تصادفی دودویی برای تخمین حالت دست است [۶۴]. الگوریتم بکار رفته در این کار بهینه است و بجای اختصاص برچسب به تصویردانه‌ها، یک فرآیند جستجوی تقسیم-چیرگی دوبخشی را اجرا می‌کند و در آن فرآیند جستجوی درشت به ریز، در نهایت به موقعیت‌های دقیق مفصل‌های اسکلتی همگرا می‌شود. اغلب روش‌های اولیه که از RDF استفاده می‌کنند، برچسب‌هایی را به تمام نقاط موجود در ابر نقطه‌های تصویر در مراحل اولیه اختصاص می‌دهند و سپس درباره مختصات مکانی مفصل‌ها رأی‌گیری انجام می‌دهند. با توجه به اینکه این برچسب‌گذاری بر روی تمام نقاط انجام می‌شود به نظر می‌رسد می‌توان با روش‌هایی که برچسب‌گذاری را برای تمام نقاط انجام نمی‌دهند کارآیی این روش را افزایش داد. یکی از این کارها، پیشنهاد X. Liang، P. Li و Li Liao است که در آن قبل از اینکه تصمیم نهایی در نقاط گره برگ اتخاذ شود، تنها یک مجموعه انعطاف‌پذیر از نقاط با اهمیت مجازی تحت تعقیب قرار می‌گیرند که نقاط SIP[۵] نامیده می‌شوند [۶۵]. استراتژی رشد جنگل در روش SIP به گونه‌ای طراحی شده است که تصمیم‌های آن بر مبنای ویژگی‌های تصادفی‌ای است که

---

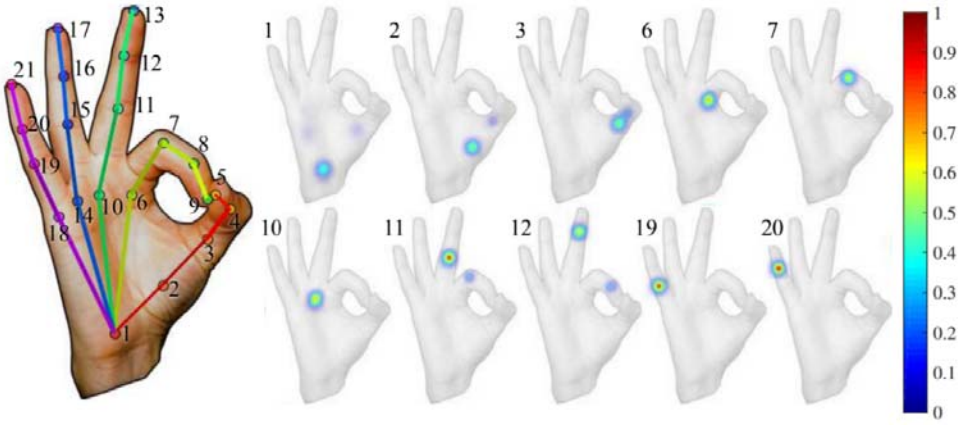
[۱] Hidden Markov Model

[۲] Conditional Random Forest

[۳] Dynamic Time Warping

[۴] Latent Regression Forest

[۵] Segmentation Index Point



شکل (۸) : خروجی روش مبتنی بر آشکارسازی [۶۶]. سمت چپ، تصویر ورودی با ۲۱ مفصل آشکارسازی شده

توسط SIP ها پیشنهاد می‌شود. سرعت آموزش نیز افزایش یافته است زیرا فقط SIP ها و نه تمام مفصل‌های اسکلتی در گره‌های غیر برگ تخمین زده می‌شوند. نتایج شبیه‌سازی این روش بر روی پایگاه داده‌های معمول، بدون موازی کردن و با استفاده از یک پردازنده معمولی نشان از عملکرد آن با نرخ 55.5 قاب در ثانیه دارد که کاملا برای کاربردهای معمول قابل قبول است. البته در کنار این بهبودها باید توجه داشت که SIP زمان بیشتری را بخصوص در تفسیرهای سه بعدی نیازمند بوده و همچنین دارای بار محاسباتی زیادی هم برای آموزش و هم برای اجرا دارد که می‌توان آن را معایب SIP در نظر گرفت. پس از ظهور و فراگیر شدن فن‌های یادگیری عمیق، و بخصوص توانایی بسیار بالای این فن در حل مشکلات موجود، بسیاری از پژوهش‌ها را به سمت استفاده از این فن‌ها به عنوان ابزاری مناسب برای حل مسائل مربوطه سرازیر کرد که حوزه تخمین حالت دست نیز خارج از این قاعده نیست. در این بخش، یادگیری عمیق و روش استفاده از آن در حوزه تخمین حالت دست توصیف شده و مبانی مرتبط با آن تشریح می‌گردد.

۵- یادگیری عمیق در حوزه تخمین حالت دست

روش‌هایی که از یادگیری عمیق برای تخمین حالت دست استفاده می‌کنند را می‌توان بر اساس معیارهای مختلفی از جمله تخمین اسکلت بندی دو بعدی یا سه بعدی، الگوریتم مبتنی بر آشکارسازی یا مبتنی بر وایازش[۱]، استفاده از شبکه‌های CNN[۲] دو بعدی یا سه بعدی، استفاده از تصاویر ژرفا یا رنگی و ... دسته‌بندی نمود. در روش‌های مبتنی بر آشکارسازی، مدل طراحی شده، برای هر مفصل یک نقشه چگالی احتمال تولید می‌کند که مکان دقیق هر مفصل با اعمال یک تابع argmax بر روی نقشه آن مفصل تعیین می‌گردد. شکل ۸ نحوه انجام این روش را توصیف می‌کند که در [۶۶] استفاده شده است.

در مقابل، روش‌های مبتنی بر وایازش مستقیما سعی در تخمین و به دست آوردن مکان هر مفصل را دارند. بعنوان مثال اگر مدلی از ۲۱ مفصل استفاده کند، به تعداد 3×21 نورون در لایه آخر نیازمند است تا بتواند مختصات سه بعدی (x, y, z) هر مفصل را تخمین بزند. به دلیل غیر خطی بودن روابط در این روش، برای استفاده از روش مبتنی بر وایازش، به مقدار زیادی داده آموزشی و حلقه تکرار نیاز خواهد بود ولی از آنجایی که تولید نقشه چگالی احتمال برای هر مفصل از لحاظ بار محاسباتی بسیار سنگین است لذا روش‌های مبتنی بر وایازش برای تخمین حالت سه بعدی دست بیشتر مورد استفاده قرار می‌گیرند [۶۷].

۵-۱- روش‌های مبتنی بر تصاویر ژرفا

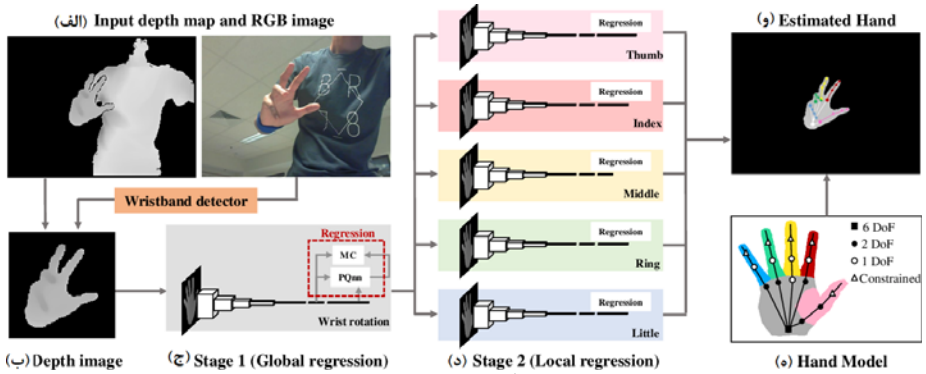
عموما روش‌های مبتنی بر تصاویر ژرفا روش غالب در تخمین حالت دست و تخمین حالت بدن هستند. در [۶۸] از یک روش مبتنی بر وایازش با استفاده از تصاویر ژرفا برای تخمین مفصل‌های ۲۱ تایی دست استفاده شده است. آن‌ها تلاش کرده‌اند تا مکان هر مفصل روی انگشتان را بطور مستقل تخمین بزنند و به همین منظور برای هر انگشت یک شبکه مجزا آموزش داده‌اند تا مفصل‌های سه گانه واقع در آن را تخمین بزنند. باید توجه کرد که با اینکه روش آن‌ها مبتنی بر تصاویر ژرفا است ولی برای شناسایی مکان دست و حذف سایر قسمت‌ها از تصاویر رنگ هم بهره گرفته‌اند ولی بر خلاف اکثر کارهای مشابه، آن‌ها شبکه‌ای مستقل را برای جداسازی ناحیه دست مورد استفاده قرار نداده‌اند که احتمالا به دلیل کاهش بار محاسباتی بوده است. آن‌ها برای مکان‌یابی دست و حذف تصویردانه‌های اضافه، از رنگ پوست استفاده کرده و رنگ‌های خارج از محدوده رنگ پوست را حذف می‌کنند. شکل ۹ روش استفاده شده در [۶۸] را نشان می‌دهد.

Ge، Liang، Yuan و Thalmann در [۶۹] از روشی جدید برای تخمین حالت دست سه بعدی با استفاده از شبکه‌های دو بعدی بهره بردند. برای این منظور آن‌ها تصویر ژرفای دست را بر روی سه صفحه متعامد نگاشت کرده و برای هر کدام از تصاویر نگاشت

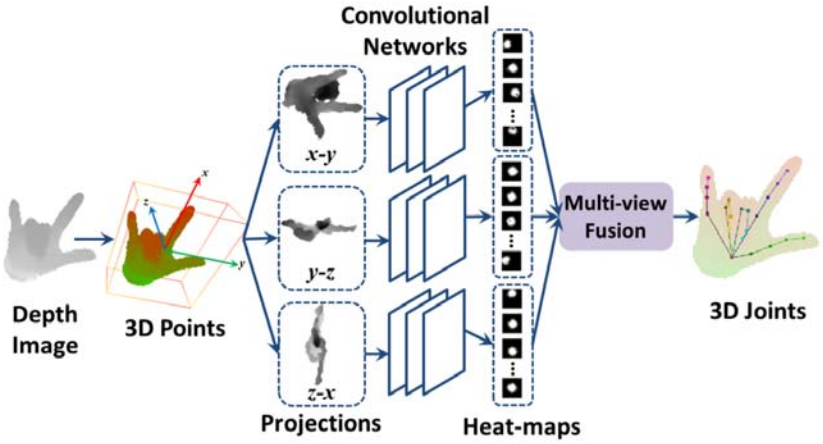
---

[۱] Regression

[۲] Convolutional Neural Network



شکل (۹) : روش استفاده شده در [۶۸]. تصاویر ورودی ژرفا و رنگ (الف)، استخراج ناحیه دست و برش تصویر ژرفا با کمک گیری از مچ بند رنگی (ب)، استخراج ویژگی‌ها و عامل‌های کلی حالت دست (ج)، استفاده از خروجی مرحله قبل و استخراج عامل‌های مفصل‌ها برای هر انگشت (د)، حالت دست نهایی استخراج شده و نمایش آن در صفحه نمایش (ه) و (و).



شکل (۱۰) : روش استفاده شده در [۶۹]

شده بر روی این صفحات، یک شبکه CNN دو بعدی آموزش دادند. سپس در نهایت نتایج حاصل از این سه شبکه را با هم تلفیق کرده و حالت سه بعدی دست از روی آن به دست می‌آید.شکل ۱۰ روش مورد استفاده آن‌ها در [۶۹] را نشان می‌دهد. اگرچه آن‌ها در این روش به روشنی بیان نکرده‌اند که مدل سه بعدی چگونه با استفاده از نتایج حاصل از سه شبکه دو بعدی تولید می‌شود ولی از مقالاتی که به این کار استناد کرده‌اند می‌توان فهمید که تولید مدل سه بعدی با استفاده از روش‌های یادگیری ماشین انجام نشده است [۶۷].

روش استفاده شده در [۷۰] و [۷۱] در دسته روش‌های مبتنی بر آشکارسازی طبقه بندی می‌شود. در [۷۰] که توسعه ای از کارهای [۷۲] و [۷۳] به شمار می‌رود نویسندگان از یک شبکه ۱۷ لایه با ۶۴ خروجی ویژگی برای ۱۶ لایه اول استفاده کرده‌اند که خروجی آن نقشه‌های چگالی احتمال برای تمام مفصل‌های ۲۱ تایی دست است. با الهام از روش نگاشت سه بعدی به دو بعدی و تلفیق نتایج، Moon، Chang و Lee در قالب یک روش مبتنی بر آشکارسازی، یک شبکه حجم‌دانه[۱] به حجم‌دانه V2V برای تخمین مستقیم محل هر قسمت دست بر اساس شکل تخمینی سه بعدی دست طراحی کردند [۷۱]. از آنجایی که ورودی و خروجی هر دو سه بعدی هستند لذا تمام عملیات کانولوشن و دی کانولوشن در حوزه سه بعدی انجام می‌شود. در واقع آن‌ها از یک CNN سه بعدی برای تخمین احتمال حجم‌دانه[۲] هر مفصل و در ادامه از یک CNN برای تخمین مرکز جرم تصویر ژرفای برش خورده در مرحله قبل استفاده کردند. شکل ۱۱ روش پیشنهادی استفاده شده در [۷۱] را نشان می‌دهد. از نکات جالب روش فوق، این است که به راحتی می‌توان آن را به سایر تخمین حالت‌ها از جمله تخمین حالت بدن تعمیم داد.

از میان روش‌های مبتنی بر وایازش می‌توان به کارهای انجام شده در [۶۸] و [۷۴] تا [۷۹] اشاره نمود که در این روش بصورت مستقیم تصویر ژرفا یا به مکان و یا به زاویه مفصل‌های مدل دست[۳] [۶۸] و [۸۰] نگاشت می‌شوند. برخی از پژوهش‌ها نگاشتی از تصویر ژرفا به مفصل‌ها را با ترکیبی از شبکه‌های دیگر مانند ResNet[۴] [۸۱] و شبکه REN [۸۲] برای تخمین مستقیم

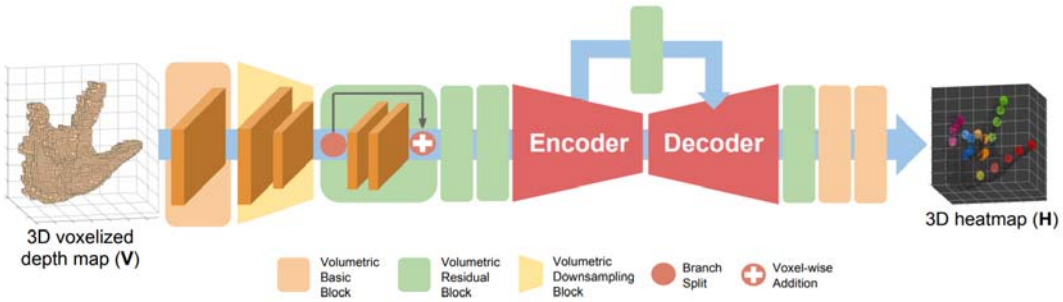
---

[۱] Voxel

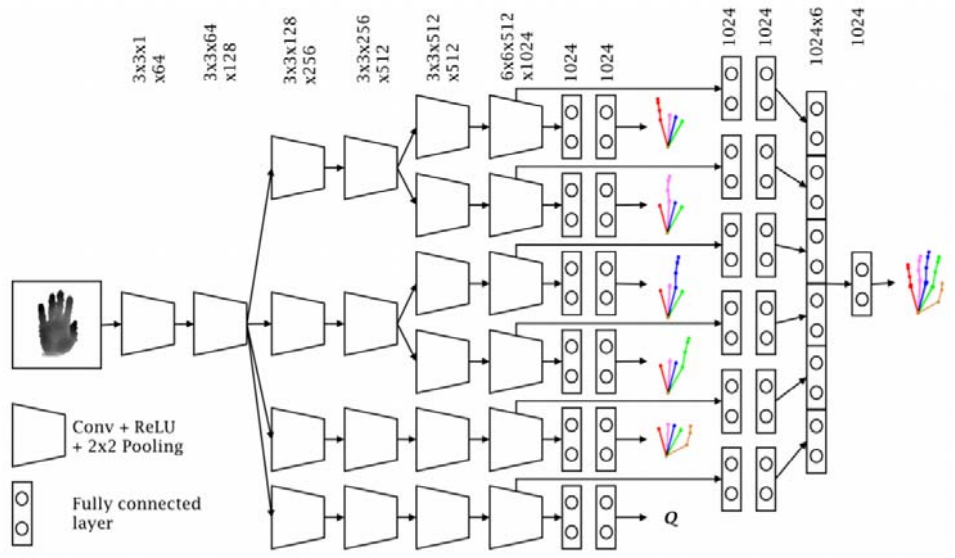
[۲] Per-voxel likelihood

[۳] Joint angels of hand model

[۴] Residual Network



شکل (۱۱) : معماری شبکه V2V-PoseNet [۷۱]



شکل (۱۲) : ساختار شبکه سلسله مراتبی طراحی شده در [۷۷]

مفصل‌ها انجام داده‌اند. در [۸۳] نویسندگان با توجه به کاهش اطلاعات مفید در گذار از لایه‌های متعدد شبکه عمیق، با افزودن ایده شبکه RESNet [۸۱] اثر محوشدگی اطلاعات را کاهش داده و موفق به بهبود دقت تخمین مفصل‌ها شده‌اند.

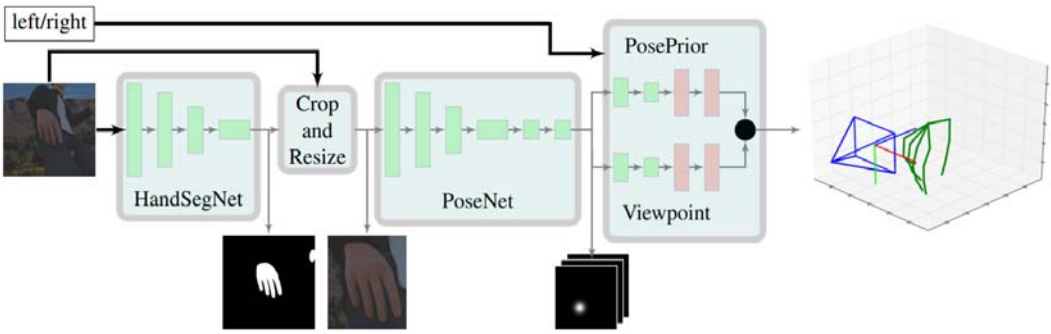
- **مدل‌های سلسله مراتبی**[۱]

این روش‌ها مساله تخمین را به چند زیر مساله تقسیم کرده و با انجام تک تک آن مسایل و تلفیق خروجی‌ها، نتیجه اصلی به دست می‌آید [۷۷] و [۸۲] و [۸۴] تا [۸۶]. روش‌های بررسی شده مفصل‌های دست را یا بر اساس انگشت [۷۷] و [۸۶] و یا بر اساس نوع مفصل [۷۷] و [۸۴] و [۸۵] تقسیم بندی کرده‌اند. در [۷۷] Madadi، Escalera، Baro و Gonzalez یک ساختار CNN سلسله مراتبی با تقسیم بلوک‌های (pooling+relu+conv) به شش شاخه (هر شاخه برای هر انگشت و یک شاخه برای کف دست) طراحی نمودند که هر کدام از این شاخه‌ها به یک لایه تمام متصل وصل می‌شود. لایه آخری هر شاخه باهم درهم آمیخته می‌شود تا تمام مفصل‌ها تخمین زده شود. شکل ۱۲ ساختار شبکه فوق را نشان می‌دهد.

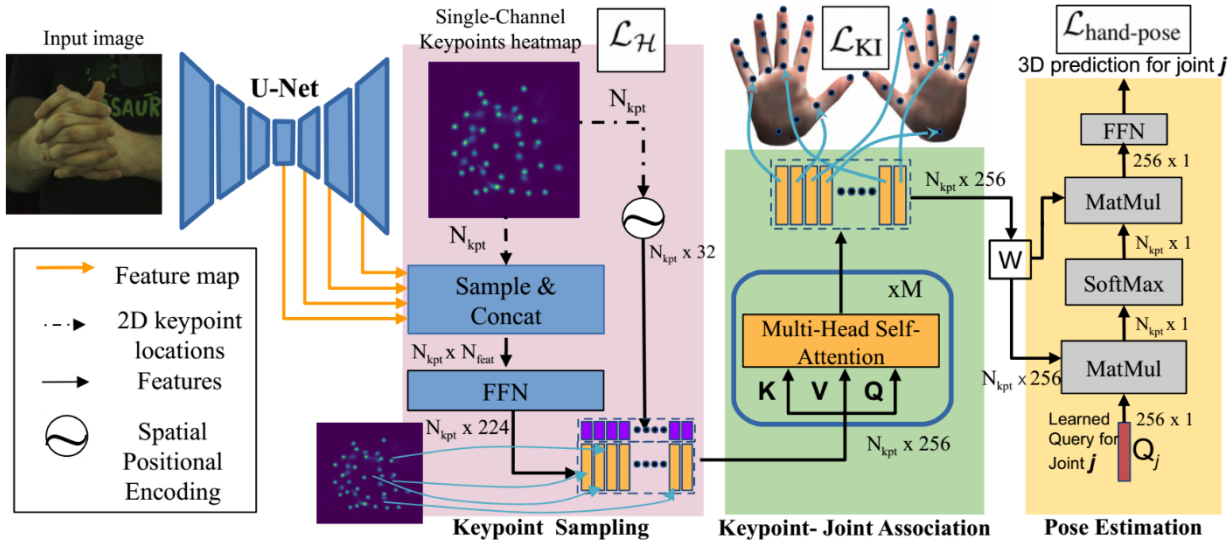
۵-۲- روش‌های مبتنی بر تصویر رنگی

همانگونه که پیشتر بیان شد، با اینکه استفاده از تصویرهای رنگی به دلیل زیاد بودن دوربین‌های با قابلیت ثبت تصاویر رنگی نسبت به دوربین‌های ژرفا، می‌تواند بسیار فراگیرتر باشد و در کاربردهای بسیار بیشتری می‌توان آن را مورد استفاده قرار داد ولی در سمت مقابل، تخمین حالت دست با تصاویر رنگی هم از لحاظ محاسباتی و هم از لحاظ نیاز به پایگاه داده نسبت به تصاویر ژرفا بسیار مشکل‌تر است. همچنین تصاویر رنگی اطلاعات بسیار کمتری نسبت به تصاویر ژرفا برای مساله تخمین حالت دست در اختیار کاربر قرار می‌دهند. از این رو در مثال‌هایی که در ادامه برای این روش ذکر شده است معمولا هر روش، پایگاه داده مختص خود را تولید کرده است که همین موضوع مقایسه کارآیی این دسته از روش‌ها را سخت‌تر می‌کند. همچنین باید توجه داشت که در روش‌های مبتنی بر تصویر رنگی، باید ابتدا ناحیه دست تشخیص

[۱] Hierarchical Models



شکل (۱۳) : ساختار معماری شبکه استفاده شده در [۸۸] برای تخمین حالت دست با استفاده از تصاویر رنگی



شکل (۱۴) : ساختار معماری شبکه مورد استفاده در KeyPoint Transformer [۸۹]، در کل ۴۲ نقطه مفصل برای دو دست تخمین زده می‌شود

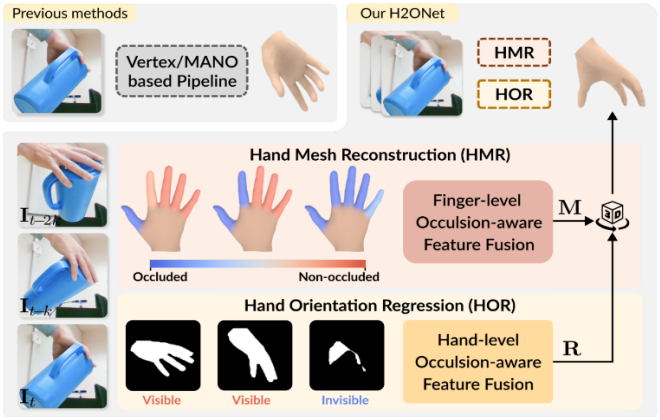
داده شده و برش داده شود و سپس تصویر حاصل به شبکه اعمال شود تا حالت آن تخمین زده شود. بنابراین بسیاری از شبکه‌های مبتنی بر تصویر رنگی از نسخه‌های مختلفی از SegNet [۸۷] استفاده می‌کنند که مخصوصا برای بخش بندی تصاویر طراحی شده است. به دلیل دسته‌بندی باینری در مساله تشخیص ناحیه دست (دسته بندی به دو دسته ی دست یا پس‌زمینه) و همچنین عدم تغییرات زیاد در تصاویر ورودی، این قسمت از مساله بار پردازشی زیادی به سیستم تحمیل نمی‌کند.

یکی از کارهای شاخص که حالت دست را با استفاده از تصاویر رنگی تخمین می‌زند، توسط Zimmermann و Brox در [۸۸] انجام شده است. آن‌ها از چهار جریان یادگیری عمیق برای تخمین حالت سه بعدی دست با استفاده از یک تصویر رنگی استفاده کرده‌اند. ابتدا از یک CNN با نام HandSegNet برای یافتن و برش ناحیه دست استفاده کرده‌اند. مطابق شکل ۱۳، خروجی HandSegNet نقابی است که تصویردانه‌های دست را نشان می‌دهد.

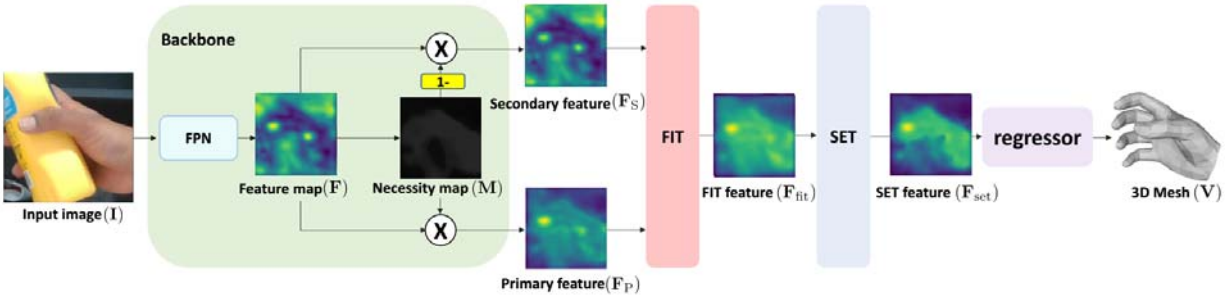
پس از برش تصویر دست و تغییر اندازه و متناسب سازی آن برای ورود به شبکه تخمین حالت، نتیجه حاصل شده وارد شبکه‌ای مبتنی بر آشکارسازی به نام PoseNet می‌شود که این شبکه برای هر مفصل یک تابع چگالی احتمال تولید می‌کند. برای تبدیل فضای دو بعدی به تخمین سه بعدی، از یک شبکه مبتنی بر وایازش بنام PosePrior استفاده شده است که تخمینی از مختصات سه بعدی نرمالیزه شده مفصل‌ها را بعنوان خروجی تولید می‌کند.

در مواردی که خود انسدادی و انسداد توسط یک شی به وجود می‌آید، تخمین حالت دست به حل همزمان مسئله مکان‌یابی نقاط مفصل و شناسایی نقاط مفصل منجر می‌شود که از لحاظ پردازش به خصوص در زمینه‌های شلوغ بسیار سنگین است. در [۸۹] برای جداسازی این دو مساله، از ساختاری که ابتدا با استفاده از یک شبکه CNN نقاط مفصل به عنوان نقاط کلیدی دو بعدی تخمین زده می‌شود و سپس ویژگی‌های استخراج شده توسط CNN برای شناسایی این نقاط مفصل به کار گرفته می‌شود، استفاده شده است. این ساختار در شکل ۱۴ نشان داده شده است که در آن، پس از شناسایی نقاط کلیدی، پیمانه خود توجه[۱] ویژگی‌های آگاه به زمینه[۲] نقاط کلیدی را که برای شناسایی نقاط مفصل حیاتی هستند تولید می‌کند. در ادامه با توجه به اطلاعات به دست آمده، حالت دست تخمین زده می‌شود. توجه شود که در این روش الزاماً همه

[۱] Self-Attention
[۲] Context-Aware



شکل (۱۵) : استفاده ازقاب‌های زمانی و اطلاعات کلی راستای دست برای استخراج اطلاعات نواحی مسدود شده در H2ONet [۹۰]



شکل (۱۶) : ساختار شبکه HandOccNet و نحوه ارتباط بخش‌های FIT و SET [۹۱]

نقاط کلیدی نقاط مفصل نیستند و همه نقاط مفصل، الزاماً روی نقاط کلیدی تخمینی قرار نمی‌گیرند. برای ارزیابی این روش، از پایگاه داده InterHand 2.6M استفاده شده است که در استفاده از یک دست، میانگین خطای ۱۰/۹۹ و برای دو دست، میانگین خطای ۱۴/۳۴ به دست آمده است.

بسیاری از روش‌ها اطلاعات مربوط به قاب‌های دارای انسداد را در تصاویر برای استفاده در تخمین حالت دست نادیده می‌گیرند ولی در [۹۰] نشان داده شده است که چنین قاب‌هایی می‌تواند حاوی اطلاعات مفیدی برای بهبود دقت تخمین باشند. در این مقاله حل مساله تخمین حالت دست به دو مرحله‌ی استفاده از اطلاعات بدون انسداد انگشتان دست در میان قاب‌های زمانی مختلف و در ادامه، استفاده از اطلاعات کلی جهت‌گیری دست برای بهبود تخمین مرحله قبل تقسیم شده است. ساختار این شبکه موسوم به H2ONet در شکل ۱۵ نشان داده شده است.

میانگین خطای تخمین در استفاده از تک قاب با سرعت ۴۳ قاب در ثانیه، ۹/۰ و در استفاده از چند قاب با سرعت ۳۵ قاب در ثانیه، ۸/۵ می‌باشد که کاملاً برای کاربردهای برخط مناسب است.

در کاری مشابه، Park و همکارانش در [۹۱] اطلاعات موجود در نواحی انسدادی رابطور موثری برای بهبود دقت تخمین مورد استفاده قرار داده‌اند. برای این منظور از دو پیمانه پشت سر هم مبتنی بر مبدل استفاده شده است که به ترتیب مبدل تزریق ویژگی[۱] FIT و مبدل خود افزاینده[۲] SET نامیده می‌شوند. مبدل FIT اطلاعات دست را با توجه به مقادیر همبستگی اجزای دست به ناحیه انسداد تزریق می‌کند و SET این اطلاعات را در راستای بهبود تخمین مورد استفاده قرار می‌دهد. ساختار این شبکه در شکل ۱۶ نشان داده شده است.

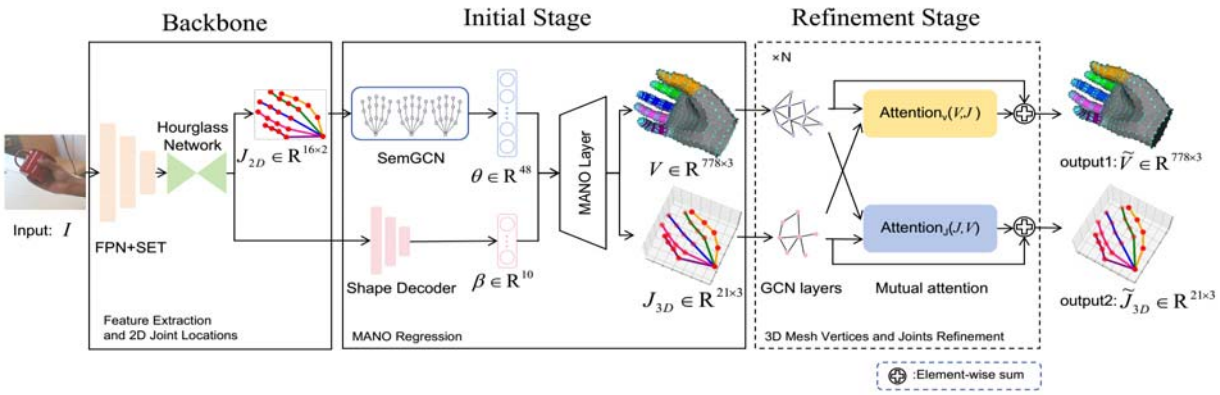
در مقاله دیگری که توسط Ziwei و همکارانش در [۹۲] ارائه شده است، از برازاندن مستقیم تور[۳] برای بهبود دقت هم راستاسازی و هدایت مدل MANO برای قابل قبول بودن حالت دست از نقطه نظر فیزیک بدن، استفاده شده است. در واقع این روش از محدودیت‌های فیزیولوژیکی دست برای بهبود دقت تخمین استفاده می‌کند.

در یک دسته‌بندی کلی که در بخش ۳ این مقاله نیز ذکر گردید، روش‌های تخمین حالت سه بعدی دست در دو گروه مبتنی بر مدل و غیر مبتنی بر مدل دسته‌بندی می‌شوند. روش‌های مبتنی بر مدل شامل نگاشت‌های غیر خطی و محاسبات پیچیده‌ای بوده و اغلب نیازمند اطلاعات پیشین در مورد ساختار و حالت دست می‌باشند و در مقابل، روش‌های غیرمبتنی بر مدل در برابر تغییرات محیط بسیار آسیب‌پذیرند [۹۳] و [۹۴].در [۹۳] مقاله شبکه‌ای موسوم به SemGCN توسعه داده شده است که از مزایای روش‌های مبتنی

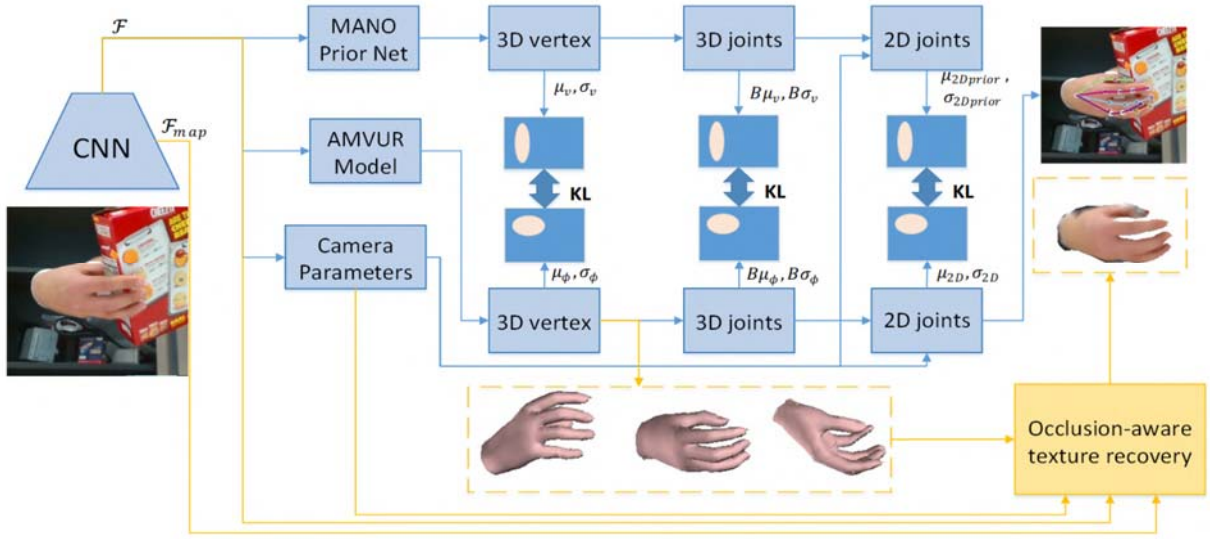
[۱] Feature Injecting Transformer

[۲] Self-Enhancing Transformer

[۳] Direct Mesh Fitting



شکل (۱۷) : اجزای مختلف شبکه SemGCN، از چپ به راست تصویر ورودی، استخراج ویژگی و نقاط مفصل دو بعدی، مرحله راه‌اندازی اولیه مبتنی بر مدل MANO و مرحله اصلاح غیرمبتنی بر مدل [۹۳]

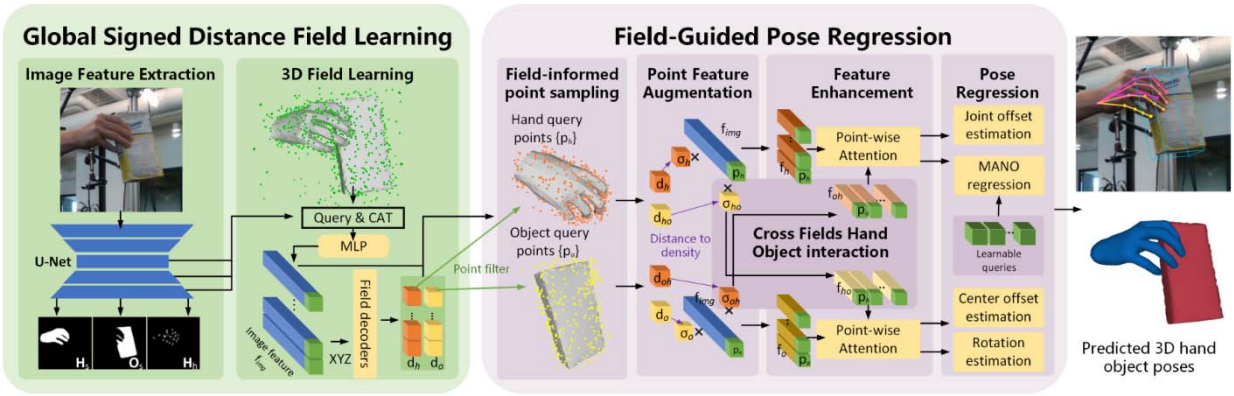


شکل (۱۸) : ساختار شبکه [۹۴] شامل بخش مبتنی بر مدل (MANO)، بخش غیرمبتنی بر مدل (AMVUR) و بخش وایازش بافت دست

بر مدل و غیرمبتنی بر مدل برای ایجاد تعادل بین دقت و امکان‌پذیری فیزیکی حالت دست بهره می‌برد. در این شبکه، ابتدا در مرحله راه‌اندازی اولیه از یک پیمانه[۱] برای وایازش پارامترهای حالت دست که مستقیماً از نقاط مفصل دوبعدی استفاده می‌کند، استفاده می‌شود که از پردازش‌های سنگین نگاشت‌های غیرخطی بی‌نیاز است و به مختصات دقیق نقاط بند وابسته نیست. خروجی این مرحله یک تخمین نه چندان دقیق از نقاط بند مبتنی بر مدل MANO است که در مرحله قبل تولید شده است. در ادامه و در مرحله موسوم به اصلاح[۲]، با استفاده از یک روش غیر مبتنی بر مدل برای اصلاح نتایج مرحله قبل تحت هدایت MANO دقت تخمین بهبود می‌یابد. شکل ۱۷ ساختار شبکه استفاده شده در این مقاله را نشان می‌دهد. برای ارزیابی کارآیی این روش از دو پایگاه داده DexYCB و HO3Dv2 استفاده شده است که میانگین خطای MPJPE و PA-MPJPE برای آنها به ترتیب ۱۲,۷، ۵/۵، ۱۹,۲ و ۸,۳ به دست آمده است.

در کاری مشابه [۹۴]، از ترکیب روش‌های مبتنی بر مدل و غیر مبتنی بر مدل استفاده شده که با ترکیب مزایای هرکدام از این روش‌ها و دوری از نقاط ضعف آنها، مدلی توسعه داده شده است که عملکرد خوبی را در مقایسه با سایر روش‌ها که از یکی از این فن‌ها استفاده می‌کنند، به نمایش می‌گذارد. این مدل AMVUR نامیده می‌شود که شکل ۱۸ ساختار این مدل را نشان می‌دهد. مدل AMVUR توزیع احتمالات شرطی نقاط مفصل و راس‌ها را تخمین زده وبرای بهبود دقت تخمین، از یک مدل توجه متقابل[۳] برای بهره‌گیری از همبستگی[۴] بین نقاط مفصل سه بعدی و رئوس

[۱] Module
[۲] Fitting
[۳] Cross-Attention
[۴] Correlation



شکل (۱۹) : ساختار شبکه HOISDF [۹۵] که از دو بخش SDF و وایازش تشکیل شده است

تور[۱] محتمل استفاده می‌کند. همچنین این شبکه حاوی مدل وایازش بافت دست است که در موارد انسداد شدید، در بازشناسی دست بسیار موثر است.

از جدیدترین کارها در تخمین حالت دست می توان به [۹۵] اشاره کرد که در آن نویسندگان برای تخمین حالت سه بعدی دست شبکه‌ای را به نام HOISDF توسعه داده‌اند که از SDF[۲]ها برای هدایت تخمین‌گر سه بعدی دست استفاده می‌کند. این شبکه دارای دو بخش است. بخش اول که برای تخمین SDFها آموزش داده می‌شود و بخش دیگر برای وایازش حالت دست-شی، مورد استفاده قرار می‌گیرد که توسط SDFها هدایت می‌گردد. SDF ها محدودیت‌های سختگیرانه‌ای را برای نقاط مفصل دست در نظر می‌گیرند که باعث می‌شود در مرحله بعدی، تخمین نقاط مفصل دچار خطا نشود. ساختار شبکه HOISDF در شکل ۱۹ نشان داده شده است. نتایج شبیه‌سازی روی دو پایگاه داده DexYCB و HO3Dv2 به ترتیب به میانگین خطای ۱۰/۱، ۱۹/۰، ۵/۳۱ و ۹/۲ منجر می‌شود که دو عدد اول معیار MPJPE و دو عدد دوم PA-MPJPE را نشان می‌دهد. سرعت اجرای این روش ۳۰/۷ قاب در ثانیه[۳] است که برای کاربردهای برخط کاملاً مناسب است. البته باید توجه داشت در استفاده از پایگاه داده DexYCB به صورت کامل و غیرانتخابی، میانگین خطای ۱۰/۱ و ۵/۱۳ به دست می‌آید.

خلاصه‌ای از روش‌های توصیف شده در قسمت‌های پیشین در جدول‌های ۱ و ۲ ارائه شده است. جدول ۱ روش‌های کلاسیک و عمدتا قبل از یادگیری عمیق را شامل می‌شود و جدول ۲ عمدتا شامل روش‌های مبتنی بر شبکه‌های عصبی و یادگیری عمیق می‌باشد. توجه شود که معیارهای ارزیابی مورد استفاده برای مقایسه، معیارهای موجود در مقالات است که شامل MPJPE[۴]، PA-MPJPE[۵]، AUC[۶]، Average per class accuracy، MAE[۷] و Mean 3D error می‌باشد.

پایگاه داده‌های مورد استفاده برای پیاده‌سازی روش مورد نظر برای هر مرجع ذکر شده است. در مواردی که پایگاه داده نوشته نشده است نویسندگان مقاله از داده‌های تولیدی خود برای پیاده‌سازی روش خود استفاده کرده‌اند.

---

[۱] Mesh
[۲] Signed-Distance Field
[۳] Frame per Second
[۴] Mean Per Joint Position Error
[۵] Procrustes Alignment MPJPE
[۶] Area Under Curve
[۷] Moving Average Error

جدول (۱): مقایسه روش‌های کلاسیک مورد استفاده برخی مراجع پیشرو و ویژگی خاص آنها

مقاله مرجع	مدل مورد استفاده	ویژگی خاص روش	نوع داده	پایگاه داده	معیار ارزیابی		سال انتشار
					MAE (mm)	میانگین تخمین صحیح (%)	
[۲۱]	**Graph-Cuts**	بهینه سازی برش گراف بهبود نتایج Shotton [۱۶]	ژرفا واقعی	–	–	۸۷/۹۱	۲۰۱۲
[۶۴]	**LRF**	جلوگیری از انتشار خطا	ژرفا واقعی	–	۲۶ (در استفاده از ۱۰۰۰ قاب تصویر)	–	۲۰۱۴
[۲۰]	**GEN**	معماری چند لایه	ژرفا واقعی-سنتز	–	–	۹۱/۲	۲۰۱۲
[۲۰]	**LEN**	معماری چند لایه	ژرفا واقعی-سنتز	–	–	۹۰/۹	۲۰۱۲
[۶۰]	**DTW**	وزن‌دهی ویژگی‌ها و بهبود تشخیص ابتدا و انتهای حرکت	ژرفا	–	–	۶۸	۲۰۱۱
[۶۱]		تخمین حالت حرکت بدن	ژرفا	MSRC-5000	–	۸۰	۲۰۱۱
[۶۵]	**SIP**	کاهش فضای جستجو با محدود کردن نقاط نامزد مفصل	ژرفا واقعی	–	–	۷۷ (با در نظر گرفتن بیشترین فاصله مجاز ۵۰ میلیمتر)	۲۰۱۶

جدول (۲): مقایسه روش‌های مبتنی بر شبکه‌های عصبی و یادگیری عمیق مورد استفاده برخی مراجع پیشرو و ویژگی خاص آنها

مقاله مرجع	مدل مورد استفاده	ویژگی خاص روش	نوع داده	پایگاه داده	معیار ارزیابی				سال انتشار
					Mean 3D Error (mm)	AUC (%)	PA-MPJPE (mm)	MPJPE (mm)	
[۷۸]	**DeepPrior++**	سادگی شبکه	ژرفا واقعی	ICVL	۸/۱	–	–	–	۲۰۱۷
				NYU	۱۲/۳				
				MSRA	۹/۵				
[۷۱]	**V2V-PoseNet**	قابل تعمیم برای سایر اعضای بدن	ژرفا واقعی	ICVL	۶/۲۸	–	–	–	۲۰۱۸
				NYU	۸/۴۲				
				MSRA	۷/۴۹				
[۸۵]	**Pose-Ren**	–	ژرفا واقعی	ICVL	۷	–	–	–	۲۰۲۰
[۱۰۴]	**GAN**	مقاوم در برابر انسداد	رنگی واقعی	NYU	–	۹۶/۵	–	–	۲۰۱۸
[۸۲]	**Ren**	–	ژرفا واقعی	NYU	۱۳/۴	–	–	–	۲۰۱۷
[۷۷]	**Madadi et. al.**	استفاده از محدودیت‌های فیزیولوژیکی دست‌ها نسبت به هم برای بهبود تخمین	ژرفا واقعی	NYU	۱۱	–	–	–	۲۰۱۷
				MSRA	۹/۷				
[۸۹]	**Keypoint Transformer v2**	–	رنگی واقعی	InterHand 2.6M	–	–	یک دست ۱۰/۸	–	۲۰۲۲
							دو دست ۱۴/۳۴		
[۹۱]	**HandOccNet**	استفاده از اطلاعات قسمت‌های زیر انسداد	رنگی واقعی	HO3D	–	–	۹/۱	–	۲۰۲۲
[۹۰]	**H2ONet**	استفاده از اطلاعات قسمت‌های زیر انسداد	رنگی	HO3Dv2	–	–	SF: ۹/.	MPJPE SF*: –	۲۰۲۳
							MF: ۸/۵	MF**: ۲۳	
				DexYCB			SF: ۵/۷	SF*: ۱۴	
							MF: ۵/۳	MF**: ۱۳/۷	
[۹۲]	**Trans.**	–	رنگی	FreiHand	–	–	–	۷/۲۸	۲۰۲۳
				InterHand 2.6M				۱۰/۰۸	
				DexYCB				۹/۱۳	
[۹۴]	**AMVUR**	ترکیب روش‌های مبتنی و غیرمبتنی بر مدل	رنگی	HO3Dv2	–	–	۸/۳	–	۲۰۲۳
[۹۳]	**SemGCN**	ترکیب روش‌های مبتنی و غیرمبتنی بر مدل	رنگی	HO3Dv2	–	–	۸/۳	۱۹/۲	۲۰۲۴
				DexYCB			۵/۵	۱۲/۷	
[۹۵]	**HOISDF**	–	رنگی	HO3Dv2	–	–	۹/۲	۱۹/۰	۲۰۲۴
				DexYCB full			۵/۱۳	۱۰/۱	
				DexYCB split			۵/۳۱	۱۰/۱	

* Single Frame
** Multi Frame

۶- پایگاه داده‌های موجود در زمینه تخمین حالت سه بعدی دست

از میان پایگاه داده‌های متعددی که برای استفاده در روش‌های مبتنی بر بینایی ماشین فراهم شده است هر دو پایگاه داده‌های دو بعدی و سه بعدی با استفاده از فعالیت‌های انسان برای تحقیق در زمینه تشخیص حالت حرکت تهیه شده‌اند [۹۶]. از آنجایی که این پایگاه داده‌ها حالت حرکت‌های دست بسیار محدودی مانند تکان دادن دست را فراهم می‌کنند پایگاه داده‌های سه بعدی از جذابیت بیشتری برای تحقیق در زمینه‌ی چالش‌های واقعی برخوردار هستند. پایگاه داده‌های حالت حرکت ایستا معمولا حالت حرکت‌های کف دست و انگشتان را در حوزه RGBD ثبت می‌کنند که می‌توانند نمادهای پایه (مانند اعداد یا یک سری نماد مانند علامت پیروزی) را بیان کنند. در ادامه مهم‌ترین پایگاه داده‌های این حوزه معرفی شده و ویژگی‌های هرکدام از آن‌ها بیان می‌شود.

۶-۱- پایگاه داده FreiHand

این پایگاه داده مجموعه‌ای از حالت‌های مختلف سه بعدی دست است که توسط ۳۲ نفر اجرا شده‌اند. برای هر تصویر دست، تفسیر مبتنی بر مدل MANO نیز در این پایگاه داده در دسترس است. در حال حاضر این پایگاه داده حاوی ۳۲۵۶۰ تصویر منحصر برای آموزش و ۳۹۶۰ تصویر برای آزمون است. تصاویری که برای آموزش تهیه شده است با استفاده از یک پس‌زمینه سبز رنگ تهیه شده‌اند که در تصاویر براحتی قابل جداسازی است. همچنین برای اهداف داده‌افزایی، سه مرحله پیش پردازش روی آن‌ها اعمال شده است ولی برای داده‌های آزمون چنین پردازش‌هایی انجام نگرفته است.

۶-۲- HandNet

پایگاه داده HandNet یکی از بزرگ‌ترین پایگاه داده‌های حاوی تصاویر ژرفا به شمار می‌رود که با استفاده از دوربین کینکت و توسط ۱۰ نفر شامل ۵ مرد و ۵ زن تهیه شده است تا اندازه‌های مختلف دست را شامل شود. این پایگاه داده حاوی ۲۰۲۰۰۰ تصویر برای مجموعه آموزشی و ۱۰۰۰۰ تصویر برای مجموعه آزمون است. در این پایگاه داده از مدل ۶ مفصل برای دست استفاده شده است.

۶-۳- پایگاه داده HO-3D

این مجموعه یک پایگاه داده حاوی تفاسیر سه بعدی از تعاملات دست-شیء است و دارای ۶۶۰۳۴ تصویر آموزشی و ۱۱۵۲۴ تصویر تست از ۶۸ دنباله تصاویر رنگی است. دنباله تصاویر با استفاده از ساختارهای چند دوربینه و تک دوربینه تهیه شده‌اند و در آن ۱۰ فرد مختلف هرکدام با ۱۰ شیء مجزا از پایگاه داده YCB حرکات تعاملی را اجرا می‌کنند. تفاسیر حالت اشیاء هم برای هردو مجموعه آموزش و آزمون در پایگاه داده در دسترس است.

۶-۴- MSRA14

پایگاه داده MSRA Hand Tracking database یا اصطلاحا MSRA14 [۹۷]در سال ۲۰۱۴ ارائه شده و با استفاده از دوربین اینتل و با مشارکت ۶ فرد راست دست تولید شده است و برای هر فرد ۴۰۰ قاب و در مجموع ۲۴۰۰ قاب تهیه شده است. هر فایل bin حاوی یک تصویر ژرفای ۳۲۰×۲۴۰ تصویردانه است که مقادیر تصویردانه به میلیمتر بیان شده و مقدار ژرفا را نشان می‌دهد. فایل تصویر رنگی مرتبط با آن تنها به منظور نمایش ارائه شده است. فایل متن موجود در پایگاه داده، اطلاعات ۴۰۰×۲۱ مفصل را ذخیره کرده است و هر خط دارای ۶۳=۲۱×۳ مقدار برای ۲۱ نقطه سه بعدی در مختصات (x,y,z) است.

۶-۵- BigHand 2.2M

پایگاه داده BigHand 2.2M [۹۸] یکی از بزرگترین پایگاه داده‌های دست تا امروز به شمار می‌رود که همانگونه که از نام آن پیداست، شامل ۲/۲ میلیون تصویر ژرفا است که توسط ۱۰ نفر مختلف اجرا شده و داده‌های قاب‌ها با استفاده از سنجش‌گرهای الکترومغناطیسی جمع آوری شده است. این پایگاه داده از مدل ۲۱ نقطه ای استفاده می‌کند.

۶-۶- پایگاه داده DexYCB

این پایگاه داده حاوی دنباله‌هایی از گرفتن اشیاء با دست است که در آن از ۲۰ شیء مختلف از پایگاه داده YCB-Video استفاده شده و حاوی اجراهای متعددی از ۱۰ فرد است. برای هر اجرا، از فرد خواسته می‌شود که یک شیء هدف را از بین ۲ تا ۴ شیء دیگر بردارد. هر اجرا ۵ بار تکرار شده و در کل برای هر فرد ۱۰۰ اجرا و در کل ۱۰۰۰ اجرای کل در پایگاه داده وجود دارد.

۶-۷- پایگاه داده دانشگاه نانیانگ سنگاپور

این پایگاه داده شامل ۱۰ حالت حرکت‌های سه بعدی ایستای دست هستند که با استفاده از یک دوربین کینکت جمع آوری شده‌اند [۳۳]. این پایگاه داده از ۱۰ فرد که ۱۰ حالت حرکت مختلف را در حالی که یک مچبند سیاه به دست آن‌ها بسته شده بود را انجام می‌دهند تشکیل شده است و بنابراین حاوی ۱۰۰

نمونه حالت حرکت است. هر نمونه از حالت حرکت‌ها حاوی یک تصویر رنگی و یک تصویر ژرفا است. این پایگاه داده تقریبا یک پایگاه داده چالش برانگیز است چرا که نمونه‌ها در یک محیط کنترل نشده و در پس‌زمینه‌های نسبتا شلوغ تهیه شده‌اند. همچنین راستای قرار گیری افراد اجرا کننده حالت حرکت، اندازه دست‌ها و مقیاس تصاویر متغیر می‌باشند. پایگاه داده‌های مشابه بعدی، پایگاه داده‌های موسوم به ASL Finger Spelling و UESTC-ASL می‌باشند که اولی حاوی علائم الفبایی و شامل ۴۸۰۰۰ حالت حرکت مختلف از ۴ فرد مختلف بوده و دومی شامل ۱۰۰ حالت حرکت از ۱۰ فرد مختلف است که در راستاها و اندازه‌ها و مقیاس‌های مختلف و بدون استفاده از مچ بند تهیه شده است [۱۶].

۶-۸- پایگاه داده Chalearn

معروف‌ترین پایگاه داده حالت حرکت‌های پویا، پایگاه داده Chalearn [۹۹] است که در سری چالش‌های Chalearn مورد استفاده قرار گرفته است. این پایگاه داده عمدتا بر روی حالت حرکت‌های دست و بازو به منظور ارتباط بین انسان و ماشین تمرکز کرده است. برخی از آن‌ها توسط کل بدن ارائه می‌شوند ولی اکثر آنها مربوط به قسمت بالاتنه می‌شود. حالت حرکت‌ها توسط بازگشت به یک موقعیت استراحت از هم تفکیک می‌شوند

۶-۹- پایگاه داده دانشگاه Sheffield

این پایگاه داده در دانشگاه Sheffield تولید شده است [۱۰۰] که در تهیه آن از سنجش‌گر کینکت استفاده شده و شامل ۱۰ نوع فعالیت[۱]، ۳ نوع پس‌زمینه[۲]، ۳ نوع حالت[۳] و تحت ۲ نوع نورپردازی[۴] است. حالت حرکت‌ها توسط ۶ فرد انجام شده و حالت‌ها بصورت با انگشت‌های بسته، با یک انگشت و با ۵ انگشت بوده و نورپردازی‌های کم نور و پر نور دو حالت آن را تشکیل می‌دهد. در مجموع این پایگاه داده حاوی ۱۰۸۰ ویدیوی ژرفا و ۱۰۸۰ ویدیوی رنگی بوده و در مجموع ۲۱۶۰ دنباله را در خود جای داده است.

۶-۱۰- پایگاه داده SynthHands

این پایگاه داده شامل تصاویر رنگی و ژرفای ساختگی است و در آن تصاویر دست زن و مرد، با و بدون وجود اشیاء تهیه شده است [۱۰۱]. تصاویر دست و اشیاء موجود بصورت ساختگی تهیه شده‌اند ولی حرکت آن‌ها از اجراهای واقعی برداشت شده است. این پایگاه داده حاوی ۶۳۵۳۰ قاب از حرکات دست انسان است و در نمایش بافت اشیاء و تصاویر پس‌زمینه، از داده‌های واقعی استفاده شده است.

۶-۱۱- پایگاه داده NYU

پایگاه داده NYU شامل سه دسته تصویر رنگی، ژرفا و تصاویر ساختگی است [۷]. برای تولید این پایگاه داده از ۳ دوربین کینکت استفاده شده است که پیکربندی دوربین‌ها بصورت یک دوربین در روبرو و دو تا در کناره‌ها بوده و مجموعه‌های آموزش و آزمون در آن بصورت مجزا ارائه شده است. در مجموع ۸۲۵۲ قاب تهیه شده توسط دو نفر برای آزمون و ۷۲۷۵۷ ویدیو تهیه شده توسط یک نفر برای آموزش تهیه شده و اطلاعات واقعی مکان مفصل‌های ۳۶ گانه در قالب یک فایل mat. به همراه پایگاه داده ارائه شده است. برای هر ۳۶ مفصل، مختصات سه گانه فوق بصورت یک سه تایی uvd است که u و v نشانگر تصویردانه و d نشانگر ژرفا به میلیمتر است.

۶-۱۲- پایگاه داده 3D Hand Pose

این پایگاه داده در سال ۲۰۱۷ تهیه شده و یک پایگاه داده چند دیدی از حالت‌های دست است که بصورت تصاویر رنگی تهیه شده و تفاسیر مختلفی از حالت‌های دست را شامل می‌شود [۱۰۲]. از جمله آن‌ها می‌توان به محدوده مکان دست و یا مکان دو بعدی و سه بعدی مفصل‌های دست اشاره نمود.

۶-۱۳- پایگاه داده InterHand2.6M

این پایگاه داده یک مجموعه بسیار بزرگ از تصاویر رنگی دست است که حاوی ۲,۶ میلیون تصویر از حالت‌های مختلف دست و ترکیب حالت‌های دو دست باهم است که همگی آن‌ها برچسب‌گذاری شده‌اند [۱۰۳].

۶-۱۴- پایگاه داده GANerated

این پایگاه داده حاوی ۳۳۰۰۰۰ تصویر رنگی ساختگی[۵] از دست انسان به همراه برخی اشیاء است که با هدف تامین نیازمندی‌های مرتبط با وقوع انسداد، به آن افزوده شده است [۱۰۴]. این تصویرها با استفاده از تصویرهای پس‌زمینه تصادفی، بازتولید شده‌اند تا بیشتر به تصویر واقعی شبیه شوند.

جدول ۳ مشخصات و ویژگی‌های اصلی پایگاه‌های داده مورد استفاده در تخمین حالت دست را به اختصار نشان می‌دهد.

---

[۱] Action
[۲] Background
[۳] Pose
[۴] Illumination
[۵] Synthetic

جدول (۳): مشخصات برخی از پایگاه داده‌های مورد استفاده در تخمین حالت دست

نام پایگاه داده	تعداد تصویر آموزشی	تعداد تصویر آزمون	واقعی-ساختگی	مفصل‌ها	تعداد افراد	نوع تصویر
FreiHand	۳۲۵۶۰	۳۹۶۰	واقعی	۲۱	۳۲	رنگی
HandNet	۲۰۲۰۰۰	۱۰۰۰۰	واقعی	۶	۱۰	ژرفا
HO-3D	۶۶۰۳۴	۱۱۵۲۴	واقعی	۲۰	۱۰	رنگی-ژرفا
MSRA14	۲۴۰۰		واقعی	۲۱	۶	ژرفا
BigHand 2.2M	۲۲۰۰۰۰۰		واقعی	۲۱	۱۰	ژرفا
DexYCB	۵۸۲۰۰۰		واقعی	۲۱	۱۰	رنگی-ژرفا
نانیانگ سنگاپور	۱۰۰۰		واقعی	–	۱۰	رنگی-ژرفا
Sheffield	۲۱۶۰ دنباله ویدیویی		واقعی	۱۰	۶	رنگی-ژرفا
SynthHands	۶۳۵۳۰		ساختگی	۲۱	–	رنگی-ژرفا
NYU	۷۲۷۵۷	۸۲۵۲	واقعی	۳۶	۲	ژرفا
3D HandPose	۸۰۰۰۰		واقعی	۲۰	–	رنگی
InterHand2.6M	۱۳۶۰۰۰۰	۸۴۹۰۰۰	واقعی	۲۱	–	رنگی
GANerated	۳۳۰۰۰۰		ساختگی	۲۱	–	رنگی

شایان ذکر است، معادل‌های فارسی مورد استفاده در این مقاله از معادل‌های تهیه شده در آزمایشگاه پردازش تصویر IPL[۱] دانشگاه صنعتی شریف [۱۰۵] انتخاب شده است که آخرین نسخه این واژه‌نامه در وبگاه این آزمایشگاه در دسترس است.

۷- جمع‌بندی و نتیجه‌گیری

با توجه به استفاده بسیار زیاد از دست‌ها در برقراری ارتباط تعاملی بین انسان و ماشین، مساله تخمین حالت سه بعدی دست در سال‌های اخیر پیشرفت‌های بسیاری را شاهد بوده است. روش‌های ابتدایی با تکیه بر ساختارهای مکانیکی از جمله سنجش‌گرهای شتاب‌سنج و سنجش‌گرهای نصب شده بر مفصل‌ها عمل می‌کردند که موجب محدودیت‌های حرکتی متعددی برای کاربر می‌شوند. روش‌های مبتنی بر پردازش تصویر از لحاظ این محدودیت‌ها عملکرد بسیار بهتری دارند ولی خود این روش‌ها از موارد متعددی آسیب‌پذیر هستند. این مقاله به بررسی انواع این روش‌ها پرداخته است که آن‌ها را از چند بعد مطالعه و دسته‌بندی نموده است. روش‌هایی که برای تخمین حالت تنها از یک تصویر رنگی استفاده می‌کنند به دلیل فراگیری دوربین‌های رنگی در زندگی روزمره از رغبت بیشتری برخوردارند ولی آسیب‌پذیری پردازش مبتنی بر رنگ به دلیل تشابه بالای رنگ پوست دست به سایر قسمت‌های بدن و همچنین وابستگی بسیار بالای این روش‌ها به نورپردازی، زاویه تابش نور، میزان شدت نور محیط و انسداد و خوداسدادی دست‌ها موجب خطای بالای این سیستم‌ها در تشخیص حالت دست می‌گردد. در مقابل روش‌های مبتنی بر تصاویر ژرفا از لحاظ دقت عملکرد بهتر هستند ولی عدم فراگیری بالای دوربین‌هایی که قادر به ثبت اطلاعات بعد سوم یعنی ژرفا هم باشند، استفاده از این روش‌ها را در کاربردهای روزمره محدود می‌کند. روش‌های مبتنی بر تصاویر رنگ-ژرفا می‌توانند عملکرد بهتری را ارائه دهند ولی پیچیدگی بالای محاسبات مشکل دیگری است که در این سیستم‌ها نمود پیدا می‌کند. کما اینکه مشکل عدم فراگیری دوربین‌های با قابلیت ثبت اطلاعات بعد سوم برای این سیستم‌ها نیز هنوز پابرجا است. ولی با تمام این موارد، با توجه به پیشرفت‌های شگرف در زمینه هوشمندسازی ابزارهای روزمره بشر در کنار توسعه روزافزون روش‌های یادگیری عمیق، انتظار می‌رود محدودیت‌های ذکر شده به زودی از میان برداشته شوند.

مراجع

[1] H. Jungong, S. Ling, X. Dong, and S. Jamie, "Enhanced Computer Vision With Microsoft Kinect Sensor: A Review," (en), IEEE Transactions on Cybernetics, vol. 43, no. 5, pp. 1318–1334 , 2013.

[2] Available online at: http://grouplab.cpsc.ucalgary.ca/cookbook/index.php/Technologies/Kinect

[3] Y. Chen, B. Luo, Y. L. Chen, G. Liang, and X. Wu, "A real-time dynamic hand gesture recognition system using kinect sensor," in *2015 IEEE International Conference on Robotics and Biomimetics, IEEE-ROBIO 2015*, pp. 2026–2030, 2016.

[۱] Image Processing Lab

[4] D.-L. Dinh, J. T. Kim, and T.-S. Kim, "Hand Gesture Recognition and Interface via a Depth Imaging Sensor for Smart Home Appliances," (en), *Energy Procedia*, pp. 576–582, 2014.

[5] M. Ludovico, M. Giulio, and Z. Pietro, "3D hand shape analysis for palm and fingers identification," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pp. 1–6, 2015.

[6] P. Li, H. Ling, X. Li, and C. Liao, "3D hand pose estimation using randomized decision forest with segmentation index points," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 819–827, 2016.

[7] NYU Hand Dataset, Available online at: http://cims.nyu.edu/~tompson/NYU_Hand_Pose_Dataset.htm

[8] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," (da), *Artificial Intelligence Review*, vol. 43, no. 1, 2012.

[9] M. K. Ahuja and A. Singh, "Static vision based Hand Gesture recognition using principal component analysis," in Proceedings of the 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education, MITE 2015, pp. 402–406, 2016.

[10] Y. Zhou, G. Jiang, and Y. Lin, "A novel finger and hand pose estimation technique for real-time hand gesture recognition," (en), Pattern Recognition, vol. 49, pp. 102–114, 2016.

[11] S. Jesus and R. M. Robin, "Hand gesture recognition with depth images: A review," in 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 411–417, 2012.

[12] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," (ca), *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[13] C. Zhang and Y. Tian, "Histogram of 3D Facets: A depth descriptor for human action and hand gesture recognition," (en), *Computer Vision and Image Understanding*, vol. 139, pp. 29–39, 2015.

[14] C. Hong, Y. Lu, and L. Zicheng, "Survey on 3D Hand Gesture Recognition," (ca), *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1659–1673, 2016.

[15] Wang, R.Y., Popovic, J.: "Real-time hand-tracking with a color glove". *ACM transactions on graphics (TOG), vol.* 28, pp.1-8, 2009.

[16] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. "Real-time human pose recognition in parts from single depth images". *Int. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1297-1304. IEEE, 2011.

[17] K. Cem, K. r. Furkan, E. K. Yunus, and A. Lale, "Real time hand pose estimation using depth sensors," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1228–1234, 2011.

[18] Y. Yao and Y. Fu, "Real-time hand pose estimation from RGB-D sensor," (en), *Proceedings - IEEE International Conference on Multimedia and Expo*, pp. 705–710, 2012.

[19] H. Liang, J. Yuan, D. Thalmann, and Z. Zhang, "Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization," (nl), *User Modeling and User-Adapted Interaction*, vol. 29, no. 6-8, pp. 837–848, 2013.

[20] C. Keskin, F. Kiraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, pp. 852–863), 2012.

[21] A. Hernandez-Vela *et al.,* "Graph cuts optimization for multi-limb human segmentation in depth maps," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 726–732, 2012.

[22] M. De La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3D hand pose estimation from monocular video," (nl), *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[23] M. De La Gorce and N. Paragios, "A variational approach to monocular hand-pose estimation," (da), *Computer Vision and Image Understanding*, vol. 114, no. 3, pp. 363–372, 2010.

[24] Oikonomidis, N. Kyriazis, and A. A. Argyros, "Marker less and efficient 26-DOF hand pose recovery," series Lecture Notes in Computer Science, Vol. 6494, pp. 744-757, 2010.

[۲۵] نصیری محمدآبادی، محمد، محمد، عرب نیا، حمیدرضا و ندیمی محمدحسین، "**تخمین حالت سه بعدی دست، مستقل از دید و مقاوم به انسداد با استفاده ازدوربین monocular**"، اولین همایش تخصصی علوم، فناوری و سامانه های مهندسی برق، تهران، ۱۳۹۲.

[26] K. Sabir, C. Stolte, B. Tabor, and S. I. O'Donoghue, "The molecular control toolkit: Controlling 3D molecular graphics via gesture and voice," in *BioVis 2013 - IEEE Symposium on Biological Data Visualization 2013, Proceedings*, pp. 49–56, 2013.

[27] A. J. Porfirio, K. L. Wiggers, L. E. Oliveira, and D. Weingaertner, "LIBRAS sign language hand configuration recognition based on 3D meshes," in *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 1588–1593, 2013.

depth camera," in *Proceedings of the 6th International Conference on Tangible, Embedded and Embodied Interaction, TEI 2012*, pp. 217–224, 2012.

[40] Z. Li and R. Jarvis, "Real time hand gesture recognition using a range camera," in *Proceedings of the 2009 Australasian Conference on Robotics and Automation, ACRA 2009*, 2009.

[41] M. Van Den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *2011 IEEE Workshop on Applications of Computer Vision, WACV 2011*, pp. 66–72, 2011.

[42] D. Droeschel, J. Stückler, and S. Behnke, "Learning to interpret pointing gestures with a time-of-flight camera," in *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 481–488, 2011.

[43] K. K. Biswas and S. K. Basu, "Gesture recognition using Microsoft Kinect," in ICARA 2011 - Proceedings of the 5th International Conference on Automation, Robotics and Applications, pp. 100–103, 2011.

[44] M. Sushmita and A. Tinku, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[45] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," (nl), *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, 2007.

[46] M. Kölsch and M. Turk, "Robust hand detection," in Proceedings - Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 614–619, 2004.

[47] E. J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Proceedings - Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 889–894, 2004.

[48] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using kinect," (nl), *BMVC 2011 - Proceedings of the British Machine Vision Conference 2011*, 2011.

[49] N. Jojic, B. Brumitt, B. Meyers, S. Harris, and T. Huang, "Detection and estimation of pointing gestures in dense disparity maps," in *Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, pp. 468–475, 2000.

[50] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee, "3D hand tracking using Kalman filter in depth space," (af), *Eurasip Journal on Advances in Signal Processing*, vol. 2012, no. 1, 2012.

[51] J. L. Raheja, A. Chaudhary, and K. Singal, "Tracking of fingertips and centers of palm using KINECT," in *Proceedings - CIMSim 2011: 3rd International Conference on Computational Intelligence, Modelling and Simulation*, pp. 248–252, 2011.

[28] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the Leap Motion Controller," (en), *Sensors (Switzerland)*, vol. 13, no. 5, pp. 6380–6393, 2013.

[29] A. Kolb, E. Barth, and R. Koch, "ToF-sensors: New dimensions for realism and interactivity," in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, 2008.

[30] D. Droeschel, J. Stückler, and S. Behnke, "Learning to interpret pointing gestures with a time-of-flight camera," in *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 481–488, 2011.

[31] S. Oprisescu, C. Rasche, and B. Su, "Automatic static hand gesture recognition using ToF cameras," in *European Signal Processing Conference*, pp. 2748–2751, 2012.

[32] T. Kapuscinski, M. Oszust, and M. Wysocki, "Recognition of signed dynamic expressions observed by ToF camera," in *Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, SPA*, pp. 291–296, 2013.

[33] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops*, pp. 1093–1096, 2011.

[34] V. Frati and D. Prattichizzo, "Using Kinect for hand tracking and rendering in wearable haptics," in *2011 IEEE World Haptics Conference, WHC 2011*, pp. 317–321, 2011.

[35] Z. Mo and U. Neumann, "Real-time hand pose recognition using low-resolution depth images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1499–1505, 2006.

[36] P. Breuer, C. Eckes, and S. Müller, "Hand gesture recognition with a novel IR time-of-flight range camera-A pilot study," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 247–260, 2007.

[37] D. Uebersax, J. Gall, M. Van Den Bergh, and L. Van Gool, "Real-time sign language letter and word recognition from depth data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 383–390, 2011.

[38] B. Yoo *et al.*, "3D user interface combining gaze and hand gestures for large-scale display," in *Conference on Human Factors in Computing Systems - Proceedings*, pp. 3709–3714, 2010.

[39] F. Klompmaker, K. Nebe, and A. Fast, "DSensingNI - A framework for advanced tangible interaction using a

interacting hands," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1862–1869, 2012.

[64] D. Tang, H. J. Chang, A. Tejani, and T. K. Kim, "Latent regression forest: Structured estimation of 3D articulated hand posture," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3786–3793, 2014.

[65] P. Li, H. Ling, X. Li, and C. Liao, "3D hand pose estimation using randomized decision forest with segmentation index points," in Proceedings of the IEEE International Conference on Computer Vision, pp. 819–827, 2016.

[66] Tomas Simon, Hanbyul Joo, Iain A Matthews, and Yaser Sheikh. "Hand keypoint detection in single images using multiview bootstrapping". In CVPR, volume 1, page 2, 2017.

[67] Doosti, B.: "Hand pose estimation: A survey". arXiv preprint arXiv:1903.01013, 2019.

[68] Ayan Sinha, Chiho Choi, and Karthik Ramani. "Deephand: Robust hand pose estimation by completing a matrix imputed with deep features". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4150-4158, 2016.

[69] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. "Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3593-3601, 2016.

[70] P. Molchanov, J. Kautz, and S. Honari. 2017 hand challenge nvresearch and umontreal team. Hands 2017.

[71] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. "V2V-posenet: Voxel-tovoxel prediction network for accurate 3d hand and human pose estimation from a single depth map". In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[72] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. "Improving landmark localization with semisupervised learning". In CVPR, 2018.

[73] S. Honari, J. Yosinski, P. Vincent, and C. Pal. "Recombinator networks: Learning coarse-to-fine feature aggregation". In CVPR, 2016.

[74] L. Ge, Y. Cai, J. Weng, and J. Yuan. "Hand PointNet: 3d hand pose estimation using point sets". In CVPR, 2018.

[75] L. Ge and J. Yuan. 2017 hand challenge imi ntu team. Hands 2017.

[76] S. Li and D. Lee. 2017 hand challenge/frame-based/team hcr: Method description. Hands 2017.

[77] M. Madadi, S. Escalera, X. Baro, and J. Gonzalez. "End-toend global to local cnn learning for hand pose recovery in depth data". arXiv preprint arXiv:1705.09606, 2017.

[52] K. A., Z. Z., and L. Z., "A real time system for dynamic hand gesture recognition with a depth sensor," in Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pp. 1975–1979, 2012.

[53] P. Guillaume and M. C. Ana, "Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping," (en), IEEE Transactions on Instrumentation and Measurement, vol. 65, no. 2, pp. 305–316, 2016.

[54] C. Yang, Y. Jang, J. Beh, D. Han, and H. Ko, "Gesture recognition using depth-based hand tracking for contactless controller application," in Digest of Technical Papers - IEEE International Conference on Consumer Electronics, pp. 297–298, 2012.

[55] C. Bellmore, R. Ptucha, and A. Savakis, "Interactive display using depth and RGB sensors for face and gesture control," in 2011 Western New York Image Processing Workshop, WNYIPW 2011 - Proceedings, pp. 5–8, 2011.

[56] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," (da), Research in Developmental Disabilities, vol. 32, no. 6, pp. 2566–2570, 2011.

[57] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in ICMI'11 - Proceedings of the 2011 ACM International Conference on Multimodal Interaction, pp. 279–286, 2011.

[58] Baum, L. E.; Petrie, T. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics. 37 (6): 1554–1563. Retrieved 28 November 2011.

[59] D. Tang, T. H. Yu, and T. K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3224–3231, 2013.

[60] M. Reyes, G. Dominguez, S. Scalera, "Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data," in IEEE International Conference in Computer Vision, Barcelona, 2011.

[61] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in Proceedings of the IEEE International Conference on Computer Vision, pp. 415–422, 2011.

[62] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2088–2095, 2011.

[63] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Tracking the articulated motion of two strongly

Computer Vision and Pattern Recognition, pp. 1496-1505. 2022.

[92] Yu, Ziwei, Chen Li, Linlin Yang, Xiaoxu Zheng, Michael Bi Mi, Gim Hee Lee, and Angela Yao. "Overcoming the trade-off between accuracy and plausibility in 3d hand shape reconstruction." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 544-553. 2023.

[93] Shuang, Feng, Wenbo He, and Shaodong Li. "3D hand reconstruction via aggregating intra and inter graphs guided by prior knowledge for hand-object interaction scenario." Journal of Visual Communication and Image Representation 2024.

[94] Jiang, Zheheng, Hossein Rahmani, Sue Black, and Bryan M. Williams. "A Probabilistic Attention Model with Occlusion-aware Texture Regression for 3D Hand Reconstruction from a Single RGB Image." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 758-767. 2023.

[95] Qi, Haozhe, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. "HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields." arXiv preprint arXiv:2402.17062, 2024.

[96] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," (en), Computer Vision and Image Understanding, vol. 141, pp. 152–165, 2015.

[97] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. "Realtime and robust hand tracking from depth". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1106-1113, 2014.

[98] Shanxin Yuan, Qi Ye, Bj¨orn Stenger, Siddhant Jain, and Tae-Kyun Kim. "Bighand2.2m benchmark: Hand pose dataset and state of the art analysis". In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pp. 2605-2613, 2017.

[99] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, "ChaLearn gesture challenge: Design and first results," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6, 2012.

[100] Available online at: http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm.

[101] Mueller. F, Mehta. D, Sotnychenko. O, Sridhar. S, Casas. D, and Theobalt. C, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor". In *Proceedings of the IEEE International Conference on Computer Vision*, p. 1154-1163, 2017.

[102] Gomez-Donoso, Francisco, Sergio Orts-Escolano, and Miguel Cazorla. "Large-scale multiview 3d hand pose

[78] Oberweger, Markus, and Vincent Lepetit. "Deepprior++: Improving fast and accurate 3d hand pose estimation." In Proceedings of the IEEE international conference on computer vision Workshops, pp. 585-594. 2017.

[79] F. Yang, K. Akiyama, and Y. Wu. Naist rv's solution for 2017 hand challenge. Hands 2017.

[80] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. "Model-based deep hand pose estimation". In IJCAI, 2016.

[81] He, K., Zhang, X., Ren, S., Sun, J.: "Deep residual learning for image recognition". In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[82] Guo, H., Wang, G., Chen, X., Zhang, C.: "Towards good practices for deep 3d hand pose estimation". arXiv preprint arXiv:1707.07248, 2017.

[83] Mofarreh-Bonab, M., Hadi Seyedarabi, Behzad Mozaffari Tazehkand, and Shohreh Kasaei. "3D hand pose estimation using RGBD images and hybrid deep learning networks." The Visual Computer (2022): pp. 1-10, 2022.

[84] K. Akiyama, F. Yang, and Y. Wu. Naist rvlab g2's solution for 2017 hand challenge. Hands 2017.

[85] Chen, X., Wang, G., Guo, H., Zhang, C.: "Pose guided structured region ensemble network for cascaded hand pose estimation". *Neurocomputing 395*, pp.138-149, 2020.

[86] F. Yang, K. Akiyama, and Y. Wu. Naist rv's solution for 2017 hand challenge. Hands 2017.

[87] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". *arXiv preprint arXiv:1511.00561*, 2015.

[88] Zimmermann, C., Brox, T.: "Learning to estimate 3d hand pose from single rgb images". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4903-4911, 2017.

[89] Hampali, Shreyas, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11090-11100. 2022.

[90] Xu, Hao, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. "H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17048-17058. 2023.

[91] Park, JoonKyu, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. "Handoccnet: Occlusion-robust 3d hand mesh estimation network." In Proceedings of the IEEE/CVF Conference on

dataset." *Image and Vision Computing 81, pp.* 25-33, 2019.

[103] Moon, Gyeongsik, Shoou-I. Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. "Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK*, August 23–28, 2020, *Proceedings*, Part XX 16, pp. 548-564. Springer International Publishing, 2020.

[104] Mueller, Franziska, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. "Ganerated hands for real-time 3d hand tracking from monocular rgb." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49-59. 2018.

[105] Available online at: http://ipl.ce.sharif.edu/glossary.html

محمد مفرح دارای مدرک دکترای مهندسی برق - مخابرات سیستم از دانشگاه تبریز بوده و در حال حاضر عضو هیات علمی دانشگاه بناب هستند. زمینه پژوهشی ایشان شامل پردازش تصویر، یادگیری عمیق، FPGA و پیاده‌سازی سخت افزاری است.

میرهادی سید عربی دارای مدرک دکترای تخصصی مهندسی برق - مخابرات سیستم از دانشگاه تبریز بوده و در حال حاضر استاد دانشکده مهندسی برق - کامپیوتر دانشگاه تبریز هستند. زمینه پژوهشی ایشان شامل پردازش تصویر، بینایی ماشین، یادگیری عمیق، یادگیری ماشین و پردازش تصاویر پزشکی است.

بهزاد مظفری تازه‌کند دارای مدرک دکترای تخصصی مهندسی برق - مخابرات سیستم از دانشگاه تبریز بوده و در حال حاضر استاد دانشکده مهندسی برق - کامپیوتر دانشگاه تبریز هستند. زمینه پژوهشی ایشان شامل تئوری مخابرات، مخابرات بی‌سیم، طیف گسترده و پردازش سیگنال است.

شهره کسایی دارای مدرک دکترای تخصصی مهندسی برق از دانشگاه Queensland استرالیا بوده و در حال حاضر استاد دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف هستند. زمینه پژوهشی ایشان شامل پردازش تصویر و ویدیو، بینایی ماشین، یادگیری عمیق، یادگیری ماشین، بازشناسی چهره و واقعیت مجازی است.