دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی
گرایش مهندسی فناوری اطلاعات

عنوان

پیاده‌سازی و مقایسه روش‌های تخمین سه‌بعدی و تک‌دوربینه حالت بدن انسان

نگارش
امرالله سیف‌الدینی بنادکوکی

استاد راهنما
دکتر حمیدرضا ربیعی

خرداد ۹۱

دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

رساله کارشناسی

پیاده‌سازی و مقایسه روش‌های تخمین سه‌بعدی و تک‌دوربینه حالت بدن انسان

نگارش: امرالله سیف‌الدینی بنادکوکی

استاد راهنما: دکتر حمیدرضا ربیعی

امضاء:

چکیده

در این پژوهش، ابتدا مروری داریم بر مسأله تخمین حالت سه بعدی و تک دوربینه انسان، سپس روش‌ها و رویکردهای اعمال شده به این مسأله را دسته‌بندی و از هر دسته، مثال‌هایی را بررسی می‌کنیم. سپس تعدادی از مشخصه‌های استفاده شده برای تخمین حالت که از سیاه‌نما به عنوان مشخصه پایه بهره می‌برند را برمی‌شماریم و مزایا و معایبشان را بررسی می‌کنیم. آنگاه برای رفع مشکل پایداری نسبت به دوران و عدم توجه به ساختار رویه‌ای داده‌ها که از مشکلات روش هیستوگرام زمینه شکل است، مشخصه‌ای جدید ارائه می‌کنیم که از کدگذاری تنک محلی و یادگیری دیکشنری برای ساخت آن استفاده می‌شود. این مشخصه در آزمایشات هم نسبت به سایر مشخصه‌ها برتری نسبی دارد و نتایج کمی و بصری نیز این ادعا را تأیید می‌کنند.

کلمات کلیدی: *تخمین حالت تک‌دوربینه انسان، بازسازی سه‌بعدی بدن انسان، سه‌بعدی، کدگذاری تنک، کدگذاری محلی، ماشین بردار وابسته*

فهرست مطالب

لیست تصاویر

لیست جداول

فصل ۱

مقدمه

از اوایل شکل‌گیری علم بینایی ماشین[۱] تاکنون، همواره پردازش تصاویر و ویدئوهای شامل انسان بسیار مورد توجه بوده است و مسائل متعددی در این باره طرح شده که برخی از آنها مانند تشخیص چهره با دقت بسیار بالایی حل شده‌اند و بسیاری نیز هنوز حل نشده باقی مانده‌اند و در حال توسعه‌اند. تخمین حالت بدن[۲] انسان یکی از این مسائل کلیدی است که دارای سابقه فعال ۱۵ ساله در پژوهش های آکادمیک بوده و این خود گویای اهمیت این حوزه است. دلیل این اهمیت، کاربردهای فراوان و متنوعی است که برای این تکنولوژی ذکر شده‌است. این کاربردها از تعامل انسان و کامپیوتر تا شناسایی حرکات انسان و سایر موضوعات مرتبط با تصویر انسان گسترده شده است. مسأله اصلی در این حوزه، تخمین زدن وضعیت قرارگیری اجزای بدن انسان در فضای سه بعدی با استفاده از یک تصویر یا ویدئو است. این وضعیت مطلوب می‌تواند با توجه به مدل مورد استفاده تعریف‌های گوناگون داشته باشد و ما امیدواریم که به کمک علم یادگیری ماشین روزی ماشین نیز بتواند مانند انسان کارهایی مثل تخمین حالت و تشخیص حرکت را با دقت و سرعت انجام دهد. لفظ تخمین از این رو برای این عمل استفاده می‌شود که ما به دلیل ماهیت دیداری مسأله و وجود چالش‌های متعدد و جدی‌ای که در این کار باید با آنها دست و پنجه نرم کنیم هرگز قادر نخواهیم بود این مسأله را بدون خطا حل کنیم. البته در اینجا منظور از این خطا، خطای خروجی با مقدارهای پیوسته جهان واقعی است نه مقدارهای گسسته حقیقت زمینه[۳] که ما فرض می‌کنیم جواب درست هستند. بنابراین همیشه به تخمینی از مقادیر واقعی بسنده می‌کنیم و بر سر به دست آوردن خطای کمتر مسابقه می‌دهیم. اکثر

---

[۱] Computer Vision

[۲] human pose estimation

[۳] groundtruth

کارهایی که در این زمینه انجام شده است فرض های محدودکننده ای را برای ساده تر کردن مسأله اعمال کرده اند. مثلاً بسیاری از تلاش ها در مورد حالات انسان در اعمال ساده ای مانند راه رفتن و دویدن انجام شده و معمولاً پس‌زمینه، ثابت و غیرپویا فرض شده است. اکثر کارها وجود فقط یک نفر را در تصویر بررسی و از انسدادهای اشیاء و افراد و حرکت خود دوربین چشم‌پوشی کرده‌اند. با توجه به کاربردهای متعدد تخمین حالت روش‌های آن نیز به دسته‌های گوناگونی تقسیم می‌شوند که هر دسته دارای مفروضات، مدل‌ها و روش‌های خاص خود است. این روش ها از نظر داده های مورد استفاده به دو دسته کلی تقسیم می‌شوند. دسته اول روش های مبتنی بر حسگر هستند که داده های خود را از حسگرهای لیزری یا مغناطیسی که روی برخی از نقاط بدن انسان نصب میشوند به دست می آورند. بنابراین در این روش ما محل دقیق برخی از اعضا یا در بدترین حالت، محدوده انسان را داریم. به دلیل همین نحوه جمع‌آوری داده و تشخیص اعضا، این روش‌ها دارای دقت بالایی هستند ولی طبعاً نیازمند تجهیزات پیشرفته‌تر و در نتیجه هزینه بالاتری هم نسبت به سایر روش‌ها هستند. دسته دوم روش های بینایی ماشین اند که داده های خود را از پردازش تصویر ورودی و تشخیص و استخراج ویژگی‌های اجزای تشکیل‌دهنده آن به دست می آورند. به دلیل ارزان تر و در دسترس‌تر بودن دوربین های فیلم برداری نسبت به حسگرهای مورد استفاده در روش‌های دسته اول و نیز افزایش قدرت پردازشی رایانه‌ها، امروزه استفاده از روش های دسته دوم برای تخمین حالت، توسعه بیشتری یافته است به طوریکه در آینده نزدیک میتوان به سادگی در گوشی‌های تلفن همراه از این فناوری بهره برد. البته هنوز هم به دلیل نیاز برخی صنایع به دقت های بالاتر، برای کاربردهای حساس بیشتر از روش های دسته اول استفاده می‌شود. در این پژوهش ما صرفاً به بررسی بخش کوچکی از روش‌های بینایی ماشین برای حل مسأله تخمین تک چشمی و ۳ بعدی حالت بدن انسان می‌پردازیم.

در ادامه این فصل ابتدا مسأله تخمین حالت انسان را به صورت اجمالی بیان و رویکردهای اعمال شده روی آن را معرفی می‌کنیم. پس از آن چالش‌های موجود برای رویکردهای بینایی ماشینی را بر می‌شماریم. سپس پایگاه‌های داده مورد استفاده این روش‌ها را معرفی خواهیم کرد و در نهایت هم ساختار ادامه این پایان‌نامه بیان خواهد شد.

۱.۱ بیان مسأله

در این بخش قصد داریم به طور جزئی‌تر این مسأله را شرح دهیم. مسأله تخمین حالت بدن انسان به طور دقیق به تخمین پیکربندی بدن انسان شامل محل، اندازه و زوایای اجزای بدن در درجه های مختلف می‌پردازد. به عبارت بهتر، ما می‌خواهیم با داشتن یک ویدئو یا تصویر، شکل و حرکت انسان

را به صورت ۳ بعدی و کامل بازسازی کنیم. در ادامه خواهیم دید که این کار بسیار سخت است و علی رغم سابقه زیادش در محیط های پژوهشی هنوز راه حل عملی و دقیقی برای آن ارائه نشده است. در واقع برای رسیدن به جواب باید تابع توزیع احتمال $p(x|z)$ را بهینه کنیم که در آن x نمایش مدل حالت، z ماتریس ویژگی‌های ورودی استفاده شده و $p(x|z)$ نیز چارچوب استنتاج این احتمال است. انتخاب‌های متفاوت برای این سه بخش، الگوریتم‌ها و روش‌های مختلف تخمین حالت را به وجود آورده است.

۱.۱.۱ کاربردها

امروزه در هر جایی که نیاز به کنترل رفتار انسان باشد میتوان به نوعی از تخمین حالت انسان برای سادگی و تسریع کار بهره برد. در واقعیت مجازی[۴]، از این تکنیک برای قراردادن انسان در محیط‌های غیرواقعی مثل استودیو، جنگل، فضا و ... استفاده میشود در حالیکه فراهم کردن و کار در چنین محیط‌هایی سخت و هزینه‌بر است. شاید جذاب‌ترین نمونه کاربردهای این فناوری را بتوان در تعامل انسان با رایانه دید. مثلاً آنجا که در ابزار کینکت[۵] مایکروسافت این فناوری در خدمت تفریح و سرگرمی آمده تا ما بتوانیم توسط حرکات بدن خود شخصیت‌های بازی را کنترل کنیم بدون اینکه نیازی به استفاده از دسته‌های بازی داشته باشیم. در اندازه‌گیری زیستی [۶] هم تخمین حالت به کمک می‌آید تا مثلاً در فیزیوتراپی بتوان آناتومی بدن انسان را تحلیل کرد و از شکل حرکات یک فرد، ناهنجاری‌های دستگاه حرکتی او را به صورت خودکار تشخیص داد. همچنین از این داده‌ها برای طراحی ربات‌های انسان‌نما یا در ربات‌های جراح برای تقلید حرکت جراح واقعی که از راه دور با حرکات خود ربات را کنترل میکند نیز استفاده می‌شود. در علم وزرش نیز از این فناوری برای بهبود حرکات ورزشکاران استفاده می‌شود. حتی اخیراً تلاش‌هایی برای شناسایی افراد از نحوه حرکات بدن (خصوصاً در راه رفتن) آنها آغاز شده است که البته به نظر، محققان برای حل این مسأله راه طولانی‌ای در پیش دارند. در بسیاری از کاربردهای مرتبط با تجسس [۷] ما نیاز داریم تا یک ویدئو یا تعدادی تصویر را برای پیدا کردن یک حرکت و حالت خاص جستجو کنیم. به طور مثال در تحلیل اطلاعات دوربین‌های مداربسته نصب شده در سطح شهر برای یافتن فرد خلافکار که حالتی مثل خم‌شدن یا رفتاری مل دویدن از خود نشان می‌دهد نمونه‌هایی از این کاربرد را میبینیم. در انیمیشن‌سازی و ساخت

---

[۴] virtual reality

[۵] Kinect

[۶] biometric

[۷] surveillance

جلوه‌های ویژه فیلم‌های سینمایی به وفور از این تکنیک برای تشخیص و ضبط حرکات واقعی انسان و سپس منتقل کردن آن به مدل اسکلتی شخصیت بازی استفاده می‌شود تا حرکات آنها کاملاً طبیعی جلوه کند. در برخی دیگر از حوزه‌های بینایی ماشین نیز از نتایج و خروجی‌های تخمین حالت استفاده می‌شود. مثلاً در تشخیص حرکات انسان این داده‌ها به عنوان ورودی مسأله به کار می‌روند و در دنبال کردن یا تشخیص انسان در ویدئو نیز از تخمین حالت برای بالا بردن دقت الگوریتم بهره می‌جویند.

۲.۱.۱ رویکردها

همانطور که در مقدمه ذکر شد، روش‌های تخمین حالت به دو دسته کلی حسگرپایه و بینایی ماشینی تقسیم می‌شوند. روش‌های بینایی ماشینی نیز خود از دیدگاه‌های مختلفی قابل دسته بندی هستند. مثلاً از نظر داده ورودی، میتوان روش‌ها را به دو دسته تقسیم کرد. دسته اول فقط از یک تصویر ثابت برای تخمین حالت فرد استفاده می‌کنند و دسته دوم از یک جریان ویدئویی که شامل چندین فریم (تصویر ثابت) در هر ثانیه است بهره میگیرند. اکثر روش‌های دسته اول بر روی ویدئو نیز قابل اعمال هستند ولی مزیت روش‌های دسته دوم در استفاده از اطلاعات زمانی و ترتیبی تصاویر است که به همواری و کاهش خطای تخمین می‌انجامد. البته از آنجا که گاهی فقط یک تصویر در اختیار داریم بهبود و استفاده از روش‌های دسته اول نیز لازم است. از دیدگاه مدل استنتاج، روش‌های بینایی ماشینی به دو دسته اصلی تمایزی[۸] و مولد[۹] منشعب میشوند. روش‌های تمایزی که اخیراً موفقیت نسبی آنها هم دراین مسأله نشان داده شده و توسعه بیشتری یافته‌اند به روش‌های پایین به بالا هم شناخته می‌شوند و از تمایز ظاهری اجزای بدن انسان مانند دست، پا، سر و ... و نیز ارتباط آنها در ساختار بدن استفاده می‌کنند تا حالت نهایی را تخمین بزنند. در مقابل، روش‌های مولد که به نام بالا به پایین هم شهرت دارند سعی می‌کنند با مقایسه ظاهر بدن در تصویر ورودی با ظاهر بدن در تصاویر آموزشی، نزدیکترین حالت را به عنوان حالت تخمینی برگردانند. در مورد این الگوریتم‌ها و تنوع آن‌ها در فصل بعد به تفصیل سخن خواهیم گفت. در میان روش‌های تخمین حالت، نمونه هایی هم وجود دارد که به طور دقیق در این دسته بندی کلی قرار نمی‌گیرند، به طور مثال [۱] از یک دوربین مادون قرمز برای گرفتن یک تصویر عمقی[۱۰] استفاده میکند و با پردازش آن، حالت بدن را با دقت خوبی تخمین میزند.[۱۱]

[۸] discriminative

[۹] generative

[۱۰] depth image

[۱۱] این مقاله اساس کار فناوری کینکت مایکروسافت است.

از نظر تعداد نقطه‌نظرهای[۱۲] ورودی، روش های تخمین حالت بینایی ماشینی به دو دسته تک‌دوربینه و چند دوربینه تقسیم میشوند. روش های چنددوربینه از چندین تصویر که با زوایای مختلفی (در بهترین حالت، مساوی و پوشاننده کل ۳۶۰ درجه) از صحنه گرفته شده‌اند استفاده می‌کنند تا با درک صحیح عمق اجزای تصویر از پیچیدگی مسأله بکاهند ولی این روش‌ها نیز پیچیدگی‌های خاص خود را دارند. بیشتر پژوهش‌های انجام شده در این زمینه بر روی یک دوربین متمرکز هستند به این دلیل که در بسیاری از کاربردهای معمولی مانند پردازش تصاویر و فیلم‌های تلویزیون ما فقط یک دوربین داریم، همچنین با وجود چند دوربین هم ما اکثر مشکلات مهمی را که در حالت تک چشمی داشتیم (مانند انسدادها) خواهیم داشت بنابراین و با توجه به اینکه انسان با دیدن یک تصویر میتواند عمل تخمین حالت را به سادگی انجام دهد، چه از نظر عملی و چه از نظر فلسفی کار بر روی حالت تک دوربینه اولویت بالاتری دارد. از دیدگاهی دیگر، تخمین حالت انسان میتواند به دو گونه ۲ بعدی و ۳ بعدی و در دستگاه مختصات مربوطه انجام شود. البته تخمین دوبعدی را معمولاً با نام تجزیه انسان هم میشناسند. امروزه به دلیل نیازهای پیچیده‌تر بر خلاف دهه اول پیدایش این حوزه، بیشتر تلاش‌ها به تخمین سه بعدی معطوف است.

با توجه به گستردگی این روش‌ها امکان بررسی همه آنها در این حجم محدود نیست بنابراین در این پژوهش ما صرفاً به بررسی برخی از روش های تخمین تک دوربینه و ۳ بعدی حالت بدن انسان که مبتنی بر بینایی ماشین هستند خواهیم پرداخت.

۳.۱.۱ ورودی‌های مسأله

همانطور که قبلاً گفته شد ما در روش تک‌دوربینه، یک تصویر یا جریان ویدئویی داریم که باید آن را در قالب ماتریس(هایی) برای پردازش رایانه‌ای بیان کنیم و با توجه به حجم زیاد داده‌ها همواره سعی داریم تا با ماتریس‌هایی کوچکتر داده‌های مفیدتری را از بخش‌های برجسته[۱۳] تصویر به الگوریتم منتقل کنیم. از طرفی چون تصاویر به دلیل اختلاف شدت نور، پوشش متفاوت افراد و پس‌زمینه‌های شلوغ[۱۴] بسیار نویزی هستند نمی‌توان به صورت مستقیم از آنها برای اعمال سطح بالا مثل تخمین حالت استفاده کرد بنابراین باید از ویژگی‌هایی سطح بالاتر مانند سیاه‌نما[۱۵]، لبه، حرکت و ... استفاده کنیم که در ادامه برخی از مهم‌ترین آنها را توضیح می‌دهیم.

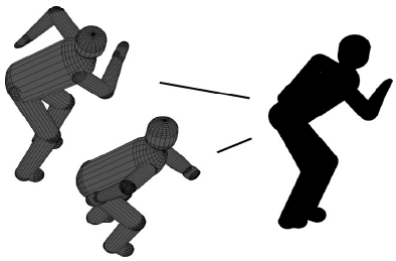
---

[۱۲] viewpoint

[۱۳] salient

[۱۴] clutter

[۱۵] silhouette

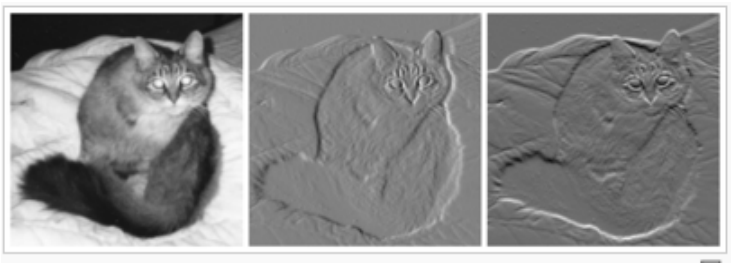


شکل ۱.۱: یک سیاه‌نما می‌تواند به دو حالت بدن نگاشت شود. (تصویر از [۵])

سیاه‌نما سیاه‌نما نمایشی سایه‌وار از انسان است که با روشی پس‌زمینه را حذف و انسان را با رنگ سیاه مشخص می‌کند. از آنجا که به طور شهودی می‌توان پذیرفت که حالت‌های مختلف بدن سیاه‌نماهای متفاوتی ایجاد می‌کنند فرض یک نگاشت از سیاه‌نما به حالت نیز در شرایط ثابت بودن پس‌زمینه قابل قبول است. در چنین شرایطی می‌توان از روش [۴] برای استخراج سیاه‌نما استفاده کرد. شاید مهمترین مشخصه استفاده شده برای تخمین حالت، سیاه‌نما باشد زیرا به تغییرات نور، رنگ، بافت و نحوه پوشش بستگی ندارد و درعین حال با فرم بدن همبستگی زیادی دارد. البته در مقابل، به دلایل متعدد مثل پس‌زمینه‌های شلوغ یا سایه‌های غیرواقعی استخراج سیاه‌نما با مشکل مواجه است و معمولاً برای کاهش این نویزها مجبوریم فیلترها و الگوریتم‌هایی را روی سیاه‌نما اجرا کنیم. در ضمن، مهمترین اشکال سیاه‌نما اینست که اطلاعات عمق را که در لبه‌ها و بافت‌ها تا حدودی وجود دارند به کلی حذف شده‌اند. نمونه‌ای از این موضوع در تصویر ۱.۱ نشان داده است.

لبه‌ها لبه‌ها در اثر تغییرات شدید و ناگهانی رنگ و نور در پیکسل‌های متوالی ایجاد می‌شوند. اطلاعات لبه‌ها را می‌توان با هزینه پردازشی کم استخراج کرد و برای تشخیص مرز بخش‌های داخلی بدن از آن بهره برد. مخصوصاً هنگامیکه از سیاه‌نما به عنوان مشخصه اصلی استفاده می‌شود می‌توان با ترکیب آن با اطلاعات لبه‌ها بخشی از ابهام عمق را رفع کرد و عضوهای جلویی را از عضوهای عقب تفکیک کرد. اگرچه لبه در مقابل رنگ، بافت و تغییرات رنگ مقاوم است و با فرم بدن همبستگی بالایی دارد اما در حضور پس‌زمینه‌های شلوغ و درهم‌ریخته یا لباس‌های با بافت ناهمگون رفتار خوبی ندارد و تعداد زیادی لبه ی اشتباه استخراج می‌شود. بنابراین معمولاً الگوریتم‌های پردازش پسینی را برای حذف این نتایج غلط و هموار کردن خروجی به کار می‌گیرند.[۸]

حرکت اگر به عنوان داده ورودی، ویدئو داشته باشیم اغلب استفاده از مشخصه حرکت به بهبود نتایج تخمین حالت کمک خواهد کرد. با فرض اینکه در هر ثانیه چندین فریم از ویدئو را در اختیار داریم میتوانیم حرکت را از تفاضل مقدار پیکسل‌های هر دو فریم متوالی به دست آوریم. البته این ساده‌ترین



شکل ۲.۱: از چپ به راست، یک تصویر سیاه‌سفید، گرادیان آن در جهت افقی، گرادیان آن در جهت عمودی (تصویر از [۱۰])

روش است و معمولاً برای تشخیص حرکت از روش‌های جریان نوری[۱۶] که در [۶] به تفصیل توضیح داده شده‌اند استفاده می‌شود.

رنگ از آنجا که پوست اعضای بدن اغلب رنگ مشخصی دارد و این رنگ با فرض شرایط نوری پایدار در طول ویدئو ثابت می‌ماند و نیز نسبت به دوران و تغییر اندازه هم مقاوم است می‌توان برای تعیین محل اعضایی مثل سر و دست و پا از مشخصه رنگ استفاده کرد. همچنین استخراج هیستوگرام رنگ[۱۷] بخش‌های تصویر هزینه پردازشی کمی دارد و می‌توان هر تصویر را به صورت مشبک به منطقه‌هایی افراز کرد و این هیستوگرام را برای هر منطقه محاسبه کرد. سپس با مقایسه توزیع رنگی هر منطقه با هیستوگرام اعضای بدن، احتمال وجود عضو در آن منطقه را مشخص کرد. البته از آنجا که ممکن است بافت لباس یا پس‌زمینه شبیه عضوی از بدن باشد معمولاً تشخیص‌های غلط بسیاری داریم و برخی از اعضا هم به دلیل انسداد دیده نمی‌شوند. برای کاهش این خطاها و پایداری نتیجه، مثلاً [۷] از محدودیت‌های هندسی پیکربندی بدن استفاده می‌کند.

گرادیان و بافت از گرادیان برای استخراج بافت[۱۸] تصویر و سپس تشخیص اجزای بدن استفاده می‌شود. در واقع گرادیان که مشتق تصویر است بیانگر تغییر شدت نور یا رنگ تصویر در یک جهت خاص (که معمولاً افقی و عمودی در نظر گرفته می‌شود) است. بافت نیز مانند رنگ در مقابل دوران و تغییر اندازه مقاوم است ولی در مواجهه با لباس‌های همگون با بدن و تغییرات نور نتایج غیرمعتبری را ارائه می‌دهد. در شکل ۲.۱ نمونه‌ای از این مشخصه نشان داده شده است.

---

[۱۶] optical flow

[۱۷] نموداری میله ای که محور افقی آن طیف رنگی و محور عمودی آن تعداد پیکسل‌های دارای یک رنگ خاص است.

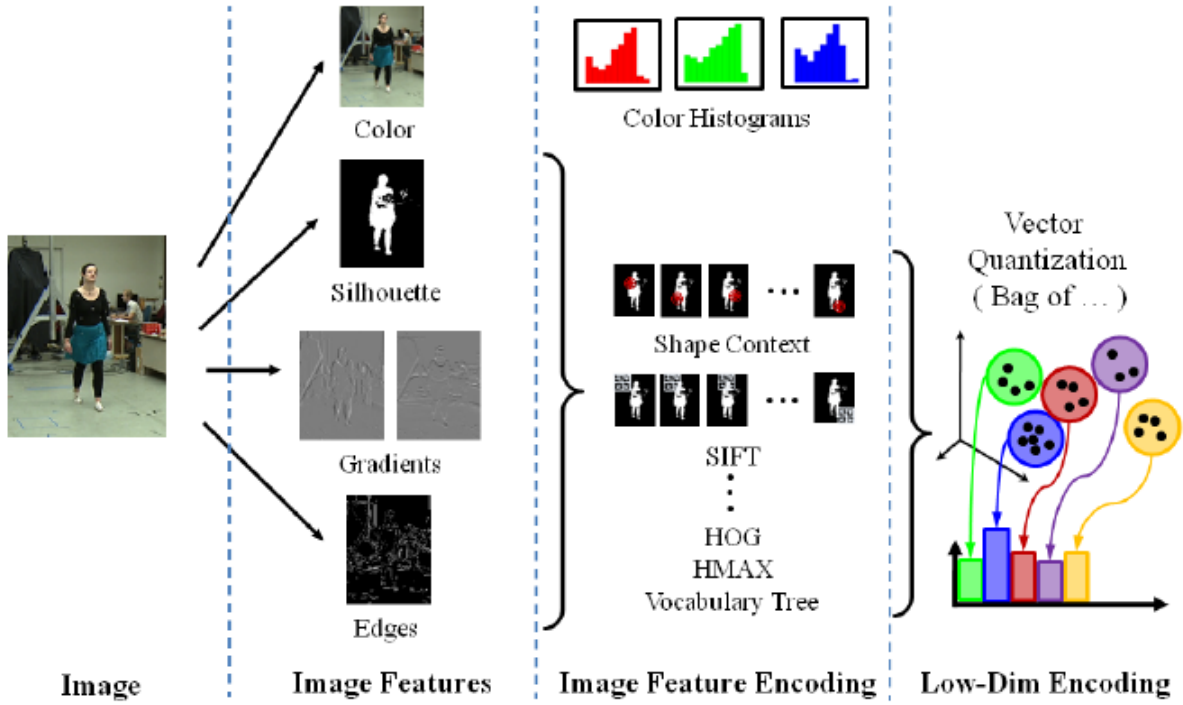
[۱۸] texture

سایر ویژگی‌ها و قالب‌ها ویژگی‌های دیگری نیز مانند سایه[۱۹] و تمرکز[۲۰] برای تخمین حالت استفاده شده‌اند ولی به اندازه موارد بالا رواج نیافته‌اند. به دلیل اینکه هر کدام از این ویژگی‌ها دارای مزایا و معایبی است معمولاً از ترکیب وزن دار آنها برای تخمین حالت استفاده می‌شود تا ضمن بهره بردن از مزایای همه در سناریوهای مختلف، نواقص همدیگر را پوشش دهند. وزن‌های مذکور می‌توانند به عنوان متغیرهای الگوریتم به صورت مکاشفه‌ای تعیین و مقداردهی شوند یا اینکه طی فرآیند استنتاج یاد گرفته شوند. اغلب برای کاهش ابعاد ماتریس‌های خروجی این مشخصه‌ها و حذف نویزهای مزاحم، این ویژگی‌های خام در قالب توصیف کننده‌های سطح بالاتر تصویر مانند زمینه شکل[۲۱]، SIFT، هیستوگرام گرادیان‌های جهت‌دار[۲۲] و حالتک‌ها [۲۳] بیان می‌شوند. همچنین می‌توان این داده‌ها را در قالب‌های سلسله مراتبی چندسطحی مانند HMAX، استوانه‌های فضایی[۲۴]، درخت لغت[۲۵] کدگذاری کرد. گاهی نیز از روش‌هایی مانند کوانتیزه کردن بردار یا سبد کلمات[۲۶] برای کاهش ابعاد داده‌ها و تبدیل آنها به هیستوگرام استفاده می‌شود ولی باعث از دست رفتن اطلاعات محلی در تصویر می‌گردد. در شکل ۳.۱ می‌توان شمایی از این چارچوب را دید.

۴.۱.۱ مدل‌های نمایش حالت

به دلیل کاربردهای مختلف تخمین حالت و نیازهای خاص هر کاربرد، مدلهای متنوعی برای نمایش حالت بدن معرفی شده‌اند که مهمترین آنها مدل درخت حرکتی[۲۷] (اسکلت حرکتی) و مدل حجمی[۲۸] است. سایر مدل‌ها را می‌توان تغییریافته این دو دانست. در مدل اسکلت حرکتی اجزای مهم و متحرک بدن با استفاده از لینک‌هایی شبیه مفاصل به هم متصل می‌شوند. هر مفصل می‌تواند حداکثر ۳ درجه آزادی[۲۹] در راستای محورهای مختصات داشته باشد که انتخاب میزان این درجه آزادی هر

[۱۹] shading
[۲۰] focus
[۲۱] shape context
[۲۲] HOG: Histogram of Oriented Gradients
[۲۳] poselet
[۲۴] spatial pyramids
[۲۵] vocabulary tree
[۲۶] bag of words
[۲۷] kinematic tree
[۲۸] volumetric
[۲۹] degree of freedom

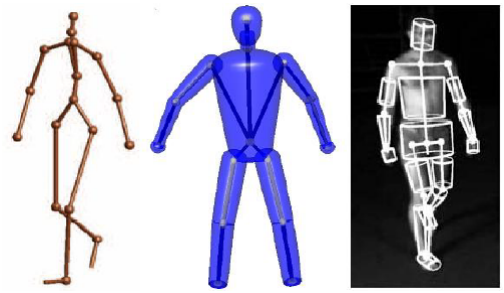


شکل ۱.۳: شمایی از ویژگی‌های تصویری رایج و قالب‌های مورد استفاده برای آنها. (تصویر از [۹])

مفصل وابسته به کاربرد است.[۲] هر چه این درجه‌ها بیشتر باشد خطای تخمین حالت کمتر و در عوض بعد داده‌ها و به تبع آن هزینه پردازشی الگوریتم بیشتر است. نحوه نمایش این مدل به صورت $x = \{\tau, \theta_\tau, \theta_1, \theta_2, \ldots, \theta_N\}$ است که در آن τ مختصات ریشه درخت است (که معمولاً لگن انتخاب می‌شود.) و θ_τ جهت این عضو را نشان می‌دهد. $\{\theta_i\}$ها نیز بیانگر زوایای نسبی اعضا نسبت به پدرشان در درخت حرکتی هستند و با توجه به نوع هر مفصل درجه آزادی آنها از ۱ تا ۳ متغیر است. این مدل می‌تواند برای سه حالت ۲ بعدی، ۳ بعدی و ۲.۵ بعدی محاسبه شود که منظور از حالت ۲.۵ بعدی، همان مدل ۲ بعدی به همراه اطلاعات عمق اعضا نسبت به هم است. مدل دیگری که به عنوان جایگزین درخت حرکتی مطرح شد، مدل مبتنی بر بخش[۳۰] است که در آن مختصات و جهت هر عضو به صورت مستقل (و نه نسبی مانند درخت حرکتی) نمایش داده شده و این بخش‌ها با محدودیت‌های آماری و حرکتی شکل بدن را می‌سازند. این مدل برخلاف اسکلت حرکتی بیشتر برای حالت ۲ بعدی به کار می‌رود و معمولاً یک پارامتر اضافه به نام s_i هم برای تغییر اندازه اعضای بدن به هر عضو اضافه می‌شود.

در مدل حجمی، اندام‌ها را توسط استوانه هایی با انتهای بیوضی شکل مدل می‌کنند. طبیعتاً این مدل می‌تواند حجم بدن را تا حد خوبی تخمین بزند ولی نیازمند تعیین طول و عرض هر کدام از این

[۳۰] part based



شکل ۴.۱: از چپ به راست، مدل اسکلت حرکتی ۳ بعدی و دو مدل حجمی (تصویر از [۲])

استوانه‌هاست. این پارامترها می‌توانند به صورت مکاشفه‌ای محاسبه و در ابتدای کار، به صورت ثابت مقداردهی شوند یا اینکه به صورت خودکار توسط الگوریتم یاد گرفته شوند. در شکل ۴.۱ نمونه‌ای از این مدل‌ها نشان داده شده‌است.

۵.۱.۱ معیارهای مقایسه

برای مقایسه روش‌های تخمین حالت به معیاری ساده و فراگیر نیاز داریم تا ضمن تطابق با اکثر پایگاه‌های داده‌ای اعداد با معنایی را نیز تولید کند. بهترین معیاری که تاکنون معرفی شده روش معروف میانگین قدرمطلق خطاها[۳۱] است. اگر $x_{i,j}$ زاویه درجه آزادی iام مفصل jام و $x'_{i,j}$ زاویه تخمین زده شده مربوطه باشد RMS طبق فرمول ۱.۱ محاسبه می‌شود.

$$RMS = \frac{۱}{mn}\sum_{i=۱}^{n}\sum_{j=۱}^{m}|((x_{i,j} - x'_{i,j} + ۱۸۰) \bmod ۳۶۰) - ۱۸۰| \qquad (۱.۱)$$

به دلیل اختلاف مدل‌های نمایش حالت یافتن یک معیار استاندارد و عمومی برای مقایسه ممکن نیست بنابراین اکثر روش‌ها کار خود را به صورت بصری یا در حالت‌های خیلی خاص با دیگران مقایسه کرده‌اند.

۲.۱ چالش ها

به دلیل پیچیدگی ظاهر بدن انسان و کمبود اطلاعات کافی برای بازسازی آن، مسأله تخمین حالت اصطلاحاً بدتعریف[۳۲] است. به عبارت دیگر، برای حل آن علاوه بر چالش‌های عام بینایی ماشین مثل پس‌زمینه‌های متحرک، حرکت دوربین، تغییرات نور و سایه باید با چالش‌های دیگری نیز که در ادامه

[۳۱] RMS: Root Mean Square

[۳۲] ill-posed

شرح داده شده‌اند دست و پنجه نرم کرد.

۱.۲.۱ چالش‌های عام بینایی ماشین

یکی از مشکلات عام بینایی ماشین تغییرات ظاهری جسم است که هرگاه بخواهیم چیزی را در یک ویدئو تشخیص دهیم یا دنبال کنیم با آن مواجه می‌شویم. در واقع تغییرات زاویه نور، رنگ، پس‌زمینه و سرعت حرکت افراد در فریم‌های متوالی باعث می‌شود تا نتوانیم به درستی مسیر حرکت افراد را دنبال کنیم.

ابهام در برچسب داده‌ها هم یکی دیگر از چالش‌های بینایی ماشین است. منظور از برچسب داده‌ها، انتساب پیکسل‌های تصویر به دسته‌های مختلف مانند پس‌زمینه، انسان یا اجزای انسان است. گاهی به دلیل کیفیت پایین عکس، الگوی شلوغ پس‌زمینه یا سناریوهای پیچیده تعامل انسان با محیط اطراف، تشخیص انسان از پس زمینه بسیار سخت می‌شود و استخراج سیاه نما و به تبع آن خروجی الگوریتم را با مشکل عدم دقت مواجه می‌کند. البته تنوع حالات بدن انسان نیز این مشکل را تشدید می‌کند.

۲.۲.۱ چالش عمق

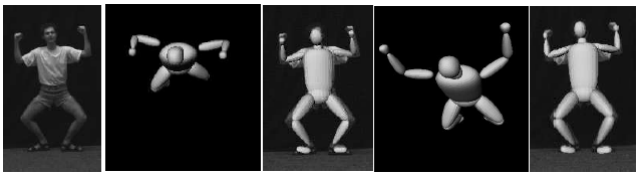
وقتی جسم را از دنیای ۳ بعدی به محیط ۲ بعدی منتقل میکنیم و با یک تصویر نمایش میدهیم در واقع اطلاعات با ارزش عمق را از دست داده ایم و این مشکل، اساسی ترین چالش تخمین حالت است. زیرا نگاشت ما از فضای حالت به فضای ظاهر، دیگر یک به یک نیست و به ازای هر تصویر می‌توان چندین حالت قرارگیری را برای بدن تخمین زد. (شکل ۵.۱) در چنین شرایطی اتفاقی که در عمل می‌افتد اینست که تابع بهینه‌سازی که قرار است حالت محتمل‌تر را برگرداند از حالت محدب[۳۳] خارج شده، چندنمایی [۳۴] می‌شود و به ازای هر اکسترمم محلی یک حالت بسیار محتمل به دست خواهیم آورد. این موضوع وقتی تشدید می‌شود که فقط از سیاه‌نما به عنوان مشخصه استفاده کنیم. بنابراین با استفاده از ترکیب مشخصه‌های با همبستگی کم[۳۵] می‌توان تا حدودی این مشکل را رفع کرد. از طرف دیگر وقتی از سیاه‌نما به عنوان ویژگی استفاده شود این موضوع تشدید خواهد شد. زیرا گاهی با مشاهده یک سیاه‌نما یا نقشه لبه‌ای[۳۶] چندین حالت قرارگیری برای شخص می‌توان متصور

[۳۳] convex

[۳۴] multimodal

[۳۵] مشخصه‌هایی که اطلاعاتشان جدید بوده و از ترکیب اطلاعات دیگر به دست نیایند.

[۳۶] edge map



شکل ۱.۵: یک نمونه از ابهام در تخمین حالت: ار چپ به راست، تصویر اصلی، نمونه‌ای از حالت ۳ بعدی متناظر آن، همان حالت از زاویه‌ای دیگر، نمونه‌ای دیگر از حالت متناظر با تصویر، حالت قبلی از زاویه‌ای دیگر که از نظر کاملاً شبیه حالت اول است. (تصویر از [۲])

شد و حتی ذهن انسان هم نمی‌تواند حالت درست را به طور قطعی تشخیص دهد. این ابهام به دلیل پیکربندی متقارن بدن و مفاصل چرخشی(مثلاً چرخیدن سر یا دست حول محور مرکزی‌اش)، پس زمینه‌های همگون با شخص، انسدادها و یا از دست رفتن اطلاعات عمق ایجاد می‌شود. این ابهام مربوط به مشاهده است و نه فرآیند استنتاج و برای حل آن باید از دانش پیشینی از جمله حرکت‌شناسی انسان استفاده کنیم.

۳.۲.۱ خودانسدادی

وقتی برخی از اجزای انسان مانع از دیده شدن بخش های دیگری از بدن او شود خودانسدادی رخ داده است و در واقع ما برخی از اطلاعات اجزای بدن را از دست داده ایم، بنابراین فرآیند استنتاج پیچیده‌تری خواهیم داشت و باید در مورد اجزای ناپیدا حدس‌هایی بزنیم. همچنین سایر اشیا یا افراد ممکن است جلوی برخی از اجزای سوژه اصلی قرار بگیرند و این چالش را تشدید کنند. این مشکل با توجه به تعداد زیاد اجزای بدن و صحنه‌های پیچیده ی ویدئوها در بسیاری از مواقع رخ می‌دهد. برای مواجهه با این چالش ابتدا باید انسداد را تشخیص بدهیم و سپس با توجه به اطلاعات قبلی که از مدل حرکتی و آناتومی بدن انسان داریم تخمینی از محل اجزای ناپیدا به دست آوریم.

۴.۲.۱ تغییرات ظاهری بدن انسان

ظاهر افراد با توجه به نوع پوشش و شکل بدنشان (مثلاً چاق یا لاغر) بسیار متغیر است و مدل کردن این پیچیدگی‌ها توسط مدل‌های ساده‌ای که در بخش ۴.۱.۱ شرح دادیم ممکن نیست و پیچیده‌تر کردن مدل‌ها نیز ابعاد داده‌ها را به صورت نمایی افزایش می‌دهد.

۵.۲.۱ محدودیت‌های فیزیکی

بدن انسان علی رغم تنوع حرکتی زیادی که به دلیل مفاصل مختلف و متعدد دارد از نظر حالات ممکن برای قرارگیری اعضا با محدودیت‌هایی مواجه است. مثلاً هرگز سر نمی‌تواند از گردن جدا باشد یا پاها از وسط سینه رد شده باشند. این محدودیت‌ها از یک سو باعث کوچک‌تر شدن فضای جستجو و در نتیجه بهبود نتایج الگوریتم‌ها می‌شوند ولی از سوی دیگر، اعمال این محدودیت‌ها به صورت خودکار بسیار سخت است و فقط از روی مدل‌های خروجی بدیهی نیست بلکه باید از داده‌های آموزشی بسیاری استفاده شود.

۶.۲.۱ بعد بالای نمایش داده های ورودی

اکثر ویژگی های تصویری که برای تخمین حالت استفاده میشود بردارهایی بزرگ دارند که گاهی (مانند shape context) تا ۶۰ بعد میرسد، سپس تعدد اجزای بدن انسان هم این ماتریس را چند برابر می‌کند و مسئله نهایی را با بردارهایی در حدود چندصد درایه ای برای هر تصویر مواجه می‌کند که حل آن به زمان و حافظه زیادی نیاز دارد. البته معمولاً با روش‌هایی مانند Bag of words این ماتریس‌ها را به صورت هیستوگرام درآورده و به عنوان ورودی به مسأله می‌دهند. اصولاً کار کردن با چنین فضای بزرگی غیرعملی است، بنابراین از یک طرف باید به سمت توصیف‌کننده‌های بهینه‌تر حرکت کنیم و از طرف دیگر باید روشی را برای کاهش بعد داده‌ها به کار گیریم. اخیراً نشان داده شده است که داده‌های حرکتی انسان علی رغم ابعاد ظاهری زیادش، ذاتاً ابعاد بسیار کمتری دارد و می‌توان آن را توسط رویه‌ای[۳۷] چندبعدی نمایش داد. این شهود، اساس بسیاری از کارهاست که با یافتن نگاشتی از فضای ویژگی به فضای حالت و برعکس سعی در تخمین حالت در فضایی با ابعاد کم دارند. البته روش‌های نیز هستند که برای مواجهه با این چالش از همبستگی زمانی حالت در فریم‌های متوالی ویدئو یا از تقارن بدن استفاده می‌کنند.

۳.۱ پایگاه های داده

به طور کلی برای تخمین حالت می‌توان از ۳ نوع پایگاه داده‌ای استفاده کرد.

---

[۳۷] manifold

نرم‌افزارهای پویانمایی نرم‌افزارهای پویانمایی[۳۸] مانند 3dmax و Maya با استفاده از مدل‌های از پیش‌تعریف شده خود می‌توانند اشکال فیزیکی از جمله انسان را شبیه‌سازی کنند و به او حرکت ببخشند. بنابراین علاوه بر ویدئوی خروجی، ساختار پیکربندی بدن را نیز می‌توانند به ما بدهند. از این داده‌ها که به دلیل نداشتن نویزهای محیط واقعی، بدون خطا هستند به عنوان آموزش و حقیقت زمینه استفاده می‌شود. همچنین تولید و برچسب زدن داده‌ها در این نرم‌افزارها ساده تر است. یکی دیگر نرم‌افزارهای معروف در این زمینه نرم‌افزار Poser از شرکت Curious Labs است.

فیلم‌ها و تصاویر معمولی در دهه اخیر با گسترش و عمومیت ابزارهای تصویربرداری، حجم عظیمی از داده‌های ویدئویی توسط مردم و کمپانی‌ها تولید شده که از طرق مختلف مانند سایت‌های به اشتراک گذاری ویدئو (مثل ،Youtube) تلویزیون و دوربین‌های نصب شده در محل‌های خاص و ... قابل دسترسی است. تنوع موضوعات و سناریوهای این ویدئوها و همچنین کم‌هزینه بودن دستیابی به این داده‌ها آنها را برای به چالش کشیدن روش‌های تخمین حالت مناسب کرده است. البته از طرف دیگر این داده‌ها بدون برچسب هستند و برچسب‌گذاری تعداد زیادی از آنها کاری سخت و پرهزینه است بنابراین فقط روش‌های نیمه نظارتی[۳۹] که نیاز به تعداد محدودی برچسب دارند می‌توانند از این ویدئوها و تصاویر به صورت گسترده استفاده کنند.

مجموعه‌های داده‌ای پایگاه‌های داده متعددی وجود دارند که برای کار تخمین حالت میتوان از آنها بهره برد ولی تعداد بسیار کمی از آنها هستند که دارای معیارهای ارزیابی دقیق و عموماً پذیرفته شده برای مسأله تخمین حالت باشند. پایگاه داده Huamneva و CMU دو نمونه خوب از این پایگاه داده‌ها هستند که شامل چندین ویدئوی انسان در فعالیت‌های مختلف‌اند و مختصات تعدادی از نقاط بدن را به عنوان حقیقت زمینه برای تست در اختیار گذاشته‌اند.

پایگاه داده CMU که با فناوری ضبط حرکت تولید شده دارای ۲۰۶۵ ویدئو در ۶ دسته‌ی ارتباط انسان‌ها (دست دادن، ...)، ارتباط با محیط (زمین بازی، محیط ناهموار، ...)، حرکات ساده (راه رفتن، دویدن، ...)، حرکات فیزیکی و ورزش(بسکتبال، رقص، ...)، موقعیت‌ها و سناریوها (پانتومیم، ...) و در نهایت حرکات تست و ۲۳ زیردسته می‌باشد. انسان‌های هر یک از دسته‌ها در ۴۱ نقطه نشانه گذاری شده‌اند و از آنها توسط ۱۲ دوربین مادون قرمز تصویربرداری شده‌است. بدین صورت محل مفاصل و سایر نقاط مهم بدن به ازاي هر یک از ویدئوها در دست است. این مجموعه یکی از

---

[۳۸] animation

[۳۹] semi-supervised

پایگاه هاي قدیمی است که بیشتر به هدف تحقیقات بر روي تشخیص رفتار انسان ساخته شده است اما براي تخمین حالت نیز از آن می‌توان استفاده کرد.

شاید بهترین پایگاه داده براي تخمین حالت انسان پایگاه داده Humaneva باشد. این پایگاه داده که هدف اصلی آن تخمین حالت و دنبال کردن انسان است از سال ۲۰۰۶ به صورت عمومی در دسترس قرار گرفته است و شامل ۷ ویدئوي کالیبره شده از ۴ انسان است که ۶ عمل معمولی (مانند راه رفتن، آهسته دویدن و غیره) را انجام می‌دهند. این داده‌ها شامل سه بخش آموزش، اعتبار سنجی و تست (به همراه برچسب حالت واقعی) میباشد.همچنین در کنار این داده‌ها کد جداسازی پس‌زمینه نیز قرار داده شده است. اخیراً پایگاه داده Humaneva ۲ نیز در کنار Humaneva ۱ در اختیار عموم قرار گرفته است که تقریباً شبیه هم هستند.

یک پایگاه داده قدیمی و پرکاربرد دیگر در تخمین حالت، پایگاه داده راه رفتن مستقیم و روی دایره متعلق به هدویک کلسترام استاد دانشگاه KTH Royal است. ویدئوی راه رفتن روی مسیر مستقیم این پایگاه داده شاید پرکاربردترین ویدئو در مقالات تخمین حالت باشد. این پایگاه داده شامل ۴ ویدئو به طول‌های ۵۱ تا ۱۷۵ فریم و چندین مجموعه تصویر است. ویدئوها برچسب گذاری نشده‌اند ولی محل اعضای بدن شامل سر، تنه، دست و پا در مجموعه‌های تصویر تعیین شده است.

فصل ۲

کارهای پیشین

تخمین حالت انسان از اواخر دهه‌ی ۱۹۹۰ به تدریج در ادبیات بینایی ماشین ظهور پیدا کرد و رفته رفته توجه به آن افزایش یافت تا جایی که اکنون به یکی از مسائل مهم در حوزه‌ی علوم رایانه تبدیل شده است. اگر به مقالات پذیرفته‌شده در مجلات و کنفرانس‌های اصلی تشخیص الگو و بینایی ماشین در سال‌های اخیر هم نگاه کنیم تکرر پژوهش‌های در این زمینه کاملاً محسوس است. درصد قابل توجه مقالات پذیرفته شده در مجله‌هایی همچون "تحلیل الگوها و هوشمندی ماشین IEEE"[۱] و "مجله‌ی بین‌المللی بینایی ماشین Springer" [۲] و کنفرانس‌هایی مثل "کنفرانس بین‌المللی بینایی ماشین" و "بینایی ماشین و تشخیص الگو" مربوط به این حوزه هستند. همانطور که در فصل اول ذکر شد، مسأله موردنظر ما در این پژوهش، مسأله تخمین تک دوربینه و ۳ بعدی حالت بدن انسان با استفاده از ورودی تصویر است. بنابراین مسائل مربوط به وجود چند دوربین، تخمین حالت ۲ بعدی و ۲.۵ بعدی، استفاده از ویژگی‌های زمانیِ ویدئو به عنوان ورودی، تخمین جهت‌دار سر انسان و سایر زیرحوزه‌های تخمین حالت، خارج از محدوده این پژوهش قرار می‌گیرند.

۱.۲ دسته‌بندی روش‌ها

در این فصل الگوریتم‌های موجود برای استخراج حالت را همانند آنچه در بررسی [۳] صورت گرفته است به دو دسته‌ی کلی روش‌های مولد و روش‌های تمایزی تقسیم‌بندی می‌کنیم. این تقسیم‌بندی شبیه به تقسیم بندی انجام شده در برخی مقالات مانند [۱۴، ۱۳، ۱۲، ۱۱] است که روش‌ها را به دو

[۱] IEEE Pattern Analysis and Machine Intelligence (PAMI)

[۲] International Journal of Computer Vision (IJCV)

دسته‌ی روش‌های مبتنی بر مدل و روش‌های مبتنی بر یادگیری تقسیم کرده‌اند. بدیهی است که این تقسیم بندی، کلی است و احتمالاً روش‌هایی وجود دارند که به طور کامل در یک دسته نمی‌گنجند بلکه از برخی ویژگی‌های هر دو گروه استفاده می‌کنند. ولی به هر حال، ارائه چارچوب و ساختار، هر چند ناقص برای مقایسه روش‌ها همیشه سودمند بوده است.

برای یک توضیح کوتاه در مورد تقسیم‌بندی به دو دسته‌ی مبتنی بر مدل و مبتنی بر یادگیری می‌توان گفت که روش‌های دسته‌ی اول از محدودیت‌های فیزیکی حرکات انسان در مدلی که از بدن در نظر گرفته‌اند برای تخمین حالت استفاده می‌کنند در حالی که روش‌های مبتنی بر یادگیری استفاده از این محدودیت‌ها را بر عهده‌ی خود الگوریتم یادگیری می‌گذارند. در بخش روش‌های مولد شرح کاملی از ایده‌ی مطرح در روش‌های مولد و نحوه‌ی حل مسئله با کمک گرفتن از این ایده خواهیم داد. سپس برای روشن شدن بحث یکی از روش‌های مولد مطرح را با جزئیات کافی توضیح خواهیم داد. از آنجا که پایه‌ی این پژوهش بیشتر بر روی روش‌های تمایزی بنا شده است آنها را به صورت مفصل در فصل بعد بررسی می‌کنیم.

۲.۲ روش‌های مبتنی بر مدل

مسأله تخمین حالت، از آنرو که باید از نمونه تصاویر ورودی، حالت احتمالی بدن انسان به ازای یک الگوی تصویری را یاد بگیرد، یک مسأله یادگیری ماشین است و مانند بسیاری دیگر از مسائل یادگیری ماشین، آن را می‌توان به عنوان مسئله‌ی پیدا کردن احتمال رخ دادن یک واقعه به شرط دانستن مقداری اطلاعات (که همان زوج‌های تصاویر ورودی و حالت متناظر آن است) در نظر گرفت. به عبارتی اگر دانسته‌های ما از مسئله که شامل داده‌های آموزشی و دانسته‌های پیشین ما از مسئله می‌شود را با x و رخداد واقعه‌ای که مجهول مسئله‌ی ما است را با yنشان دهیم، مسائل یادگیری به دنبال پیدا کردن $p(y|x)$ هستند. روش‌های یادگیری ماشین بر اساس رویکردی که به پیدا کردن این احتمال دارند به دو دسته‌ی روش‌های مبتنی بر مدل و مبتنی بر یادگیری تقسیم می‌شوند. نام رایج‌تر این روش‌ها به ترتیب، روش‌های مولد و تمایزی است. تلاش اصلی روش‌های مولد پیدا کردن $p(x|y)$و سپس تخمین زدن $p(y|x)$ از روی آن و به کمک قانون بیز است. از طرف دیگر هدف روش‌های تمایزی پیدا کردن مستقیم $p(y|x)$ است.

وجه تسمیه روش‌های مولد این است که این روش‌ها با استفاده از پارامترهای مجهول مسئله و یک مدل از پیش‌تعریف شده انسان، قابلیت تولید (یا به بیان دقیق‌تر پیدا کردن توزیع احتمال) داده‌های قابل مشاهده که همان مشخصه‌های ورودی مسئله هستند را دارند. روش‌های مولدی که از سیاه‌نما به عنوان

ورودی استفاده می‌کنند معمولاً قابلیت این را دارند که با داشتن زوایای مفصل‌های بدن، سیاه‌نمای متناظرش را تولید کنند. این روش‌ها با استفاده از تکنیک‌های بهینه سازی یا ساخت نمونه‌های تصادفی، دسته پارامترهایی را پیدا می‌کنند که تصویر نزدیکی به ورودی مسئله تولید کند. سپس با استفاده از قانون بیز به هر یک از این حالت‌ها احتمالی برای درست بودن آن نسبت داده می‌شود. مرحله استنتاج با جستجوی فضای حالت[۳] برای اکسترمم‌های بیشینه تابع درست‌نمایی ($p(x|y)$) در دو بخش انجام می‌شود. در بخش اول نحوه‌ی مقدار دهی اولیه به y و نحوه‌ی تغییر آن برای حرکت بر روی فضای حالت تعیین شود. ساده ترین راه برای حل این مسئله مقداردهی اولیه به صورت دستی(که با تجربه بهبود می‌یابد) و جستجوی فضا به صورت پیاده‌روی تصادفی[۴] است. بخش بعد شامل تولید نمونه‌ای از مدل بدن، متناظر با پارامترهای حالت و سنجش میزان شباهت آن با ورودی مسأله است. جست و جوی این فضای حالت که ابعاد بالایی دارد مسئله‌ای پیچیده است و روش‌های متنوعی برای مواجهه با آن مطرح شده است. پس از پیدا کردن حالت‌هایی که نقاط بیشینه‌ی $p(x|y)$ را می‌سازند، با استفاده از قانون بیز و با در دست داشتن مدل ورودی‌های مسئله، $p(x)$ و احتمال پیشین حالت‌ها، $p(y)$ احتمال حالت به شرط ورودی را به صورت زیر بدست می‌آوریم:

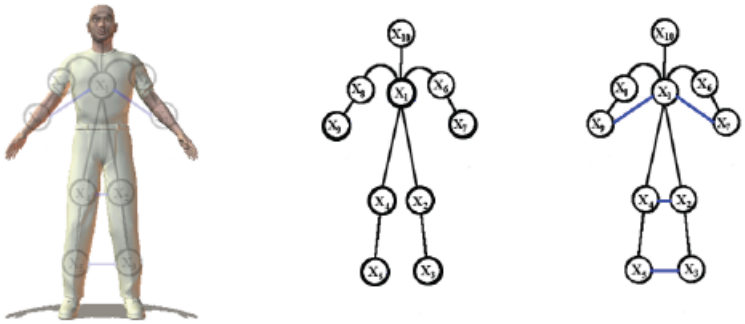
$$p(y|x) = \frac{p(x|y)p(x)}{p(x)} \qquad (۱.۲)$$

رویکرد مولد در حوزه تخمین حالت، از قدمت بالاتری نسبت به روش‌های تمایزی برخوردار است اما هنوز چالش‌های اساسی آن پابرجاست. از جمله اینکه بالا بودن بعد فضای حالت خروجی جستجو برای پیدا کردن بیشینه‌های این فضا را دشوار کرده و نیازمند داشتن محدودیت‌های قوی روی فضا است. یکی از مهم‌ترین این محدودیت‌ها که با استفاده از فریم‌های قبل و بعد فریم جاری به وجود می‌آید اینست که حالت انسان در دو فریم متوالی نمی‌تواند تغییر زیادی بکند. بر همین اساس، می‌توانیم فضای جستجو را به میزان قابل توجهی هرس کنیم. سایر محدودیت‌ها اکثراً ناشی از محدودیت‌های فیزیکی بدن انسان هستند. مثلاً اینکه زانو و آرنج نمی‌توانند به داخل خم شوند. به دلیل نیاز به این محدودیت‌ها اکثر روش‌های مولد برای کاربرد دنبال کردن انسان که معادل پیدا کردن حالت (و در حالت عمومی‌تر، محل) انسان در تمام فریم‌های یک ویدئو است طراحی شده‌اند و اکثراً روش‌هایی مبتنی بر مدل هستند.۱.۲ در ادامه‌ی این بخش دو نمونه از روش‌هایی که به تخمین حالت انسان با رویکرد مولد پرداخته‌است را مختصراً معرفی می‌کنیم.

---

[۳] Pose Space

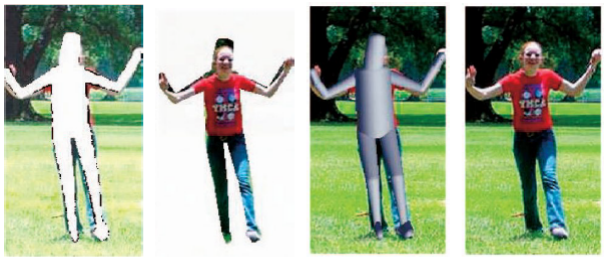
[۴] Random Walk



شکل ۱.۲ : مدل ساختار تصویری

یکی از روش‌های مولد مبتنی بر مدل مطرح که در سال ۲۰۰۶ ارائه شد[۱۴] ابتدا با استخراج محل تقریبی اعضای بزرگ بدن (دست، پا، سر و ...) اسکلت دوبعدی‌ای از حالت انسان می‌سازد. سپس با روش مخصوصی که معرفی کرده، این اسکلت دوبعدی را به حالت سه‌بعدی اولیه تبدیل می‌کند. واضح است که این کار، خطای زیادی دارد و چالش اصلی تخمین حالت هم یافتن همین نگاشت است. بنابراین این اسکلت سه‌بعدی فقط تخمینی اولیه از حالت نهایی انسان است و در ادامه مقاله سعی شده است که آن را بهبود داده و به تخمین قابل قبول نزدیک کند. این کار به صورت گام به گام و توسط زنجیره‌ی مارکوف مونت کارلو[۵] که روشی تکراری برای بهبود احتمالی است انجام می‌شود. مشکل جستجوی سریع و کامل فضای حالت که به عنوان چالش عمومی روش‌های مولد مطرح کردیم در این مقاله هم وجود دارد. به عنوان یک نوآوری دیگر، این مقاله برای مواجهه با این چالش، عملگری ارائه کرده است که شانس جستجوی کامل فضا را در پیاده‌روی تصادفی به کمک پرش‌هایی به سایر مکان‌های عموماً چگال فضای حالت افزایش می‌دهد. اساس کار این عملگر که از شهود عمومی نسبت به نحوه قرارگیری دست و پاها نسبت به تنه الهام گرفته، بدین صورت است که با داشتن یک حالت به عنوان نقطه‌ای از فضای حالت، نقطه‌ی بعد با قرینه کردن مختصات یکی از دست‌ها یا پاها نسبت به محور عمودی بدن به دست می‌آید. علاوه بر روش جستجوی فضای حالت نحوه‌ی ارزش‌گذاری نقاط این فضا نیز باید مشخص باشد تا برای تعیین کاندیدای بهتر به کار گرفته شود. در این مقاله با توجه به مدلی که از بدن انسان در نظر گرفته شده است و توجه به حالت خروجی الگوریتم، تصویری شماتیک از بدن انسان ساخته شده و با تصویر واقعی مقایسه می‌شود و میزان شباهت آین دو به عنوان معیاری برای دقت تخمین به کار گرفته می‌شود. این مقایسه تحت سه فاکتور صورت می‌پذیرد که عبارتند از: توزیع رنگ پیش‌زمینه و پس‌زمینه، همپوشانی انسان ساختگی و اصلی

[۵] Markov Chain Monte Carlo (MCMC)



(د) (ج) (ب) (الف)

شکل ۲.۲ : نمونه‌ای از کارکرد تابع درست‌نمایی الف) تصویر اصلی ب) تصویر ساختگی ج)تخمین محل انسان د) تخمین پس‌زمینه [۱۴]

در تصویر و در نهایت، رنگ پوست انسان. در واقع همانطور که در شکل ۲.۲ مشخص است، پس از انداختن تصویر آدمک ساخته شده از حالت خروجی، روی آدم اصلی که محل آن را از قبل داریم، هرچه که انطباق محل بدن دو آدم بیشتر باشد این حالت امتیاز بهتری دریافت می‌کند.

یکی دیگر از روش‌های مولد که در کنفرانس CVPR۲۰۱۱ ارائه شد، استفاده از حالتک[۶]های سلسله مراتبی برای تخمین حالت دو بعدی[۳۱] است. البته این روش می‌تواند با تغییرات اندکی برای تخمین سه‌بعدی هم استفاده شود. حالتک[۳۰] برای اولین بار توسط Bourdev از دانشگاه برکلی، در کنفرانس ICCV۲۰۰۹ ارائه شد و پس از آن در کاربردهای مختلف مرتبط با انسان از آن استفاده شد. الگوریتم حالتک تلاشی است برای معرفی یک تشخیص‌دهنده بخش [۷] جدید که بیشتر برای موجودیت‌هایی که فضای حالت گسترده‌ای دارند مفید واقع می‌شود. در این چارچوب، هر بخش تعریف دیگری دارد و به صورت بخش‌هایی از تصویر که هم در فضای ظاهر و هم در فضای پیکربندی، خوشه‌های فشرده‌ای تشکیل می‌دهند تعریف می‌شود. ایده اولیه این تعریف از آنجا ناشی شد که در مسأله تشخیص بخش‌های بدن انسان، به دلیل تنوع فیزیک بدن انسان و نیز شکل ظاهری پوشش، رنگ و صحنه، ابهام زیادی در یافتن یک بخش وجود داشت. معمولاً موارد تشخیص اشتباه، مواردی بود که الگوریتم فریب ظاهر یکسان یا نمونه‌های آموزشی خورده بود در حالیکه خروجی تصاویر در فضای پیکربندی کاملاً متفاوت بود. بنابراین ایده یک تعریف جدید برای بخش که در هر دو فضا شباهت بالایی داشته باشد شکل گرفت و نمونه‌های آموزشی برای تصاویر انسان ساخته شد. این نمونه‌ها در پایگاه داده‌ی H۳D ارائه شد. الگوریتم آموزش، ابتدا خوشه‌هایی که در هر دو فضا فشرده باشند را می‌یابد و هر کدام را یک حالتک فرض می‌کند. البته برای کم کردن این تعداد، جزئیات دیگری هم

---

[۶] Poselet

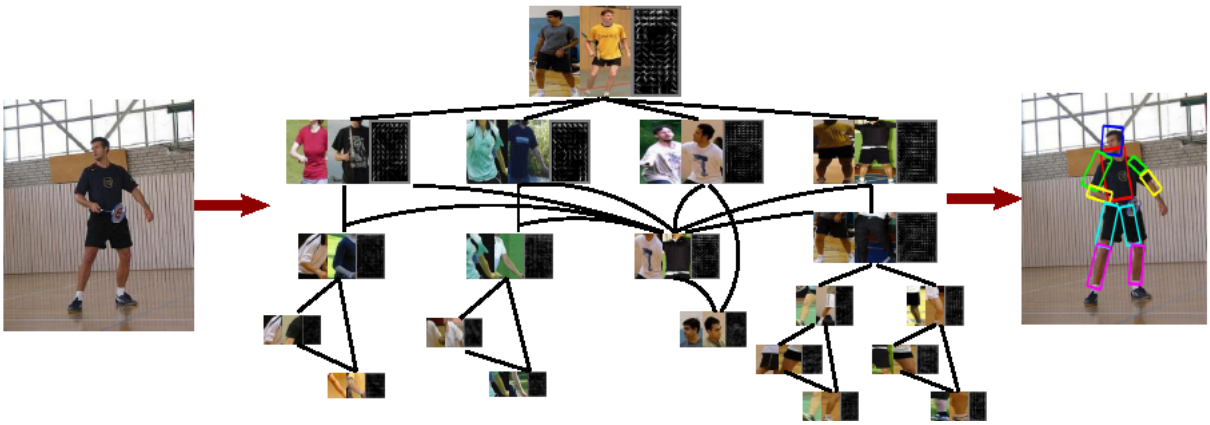
[۷] part detector

شکل ۳.۲: نمونه‌ای از نتایج تشخیص حالتک‌ها [۳۰]

در الگوریتم گنجانده شده که خارج از بحث این پژوهش است. نتیجه این الگوریتم برای تشخیص انسان کاملاً خوب بوده و در مسابقات Pascal سال‌های اخیر، مقام اول را به دست آورده. در شکل نمونه‌ای از نتایج تشخیص این روش را مشاهده می‌کنید که توانسته است بخش‌های بزرگتری از یک بخش را هم به صورت دقیق تشخیص دهد.

در مقاله [۳۱] که به آن اشاره شد، هدف، شکستن انسان به بخش‌های بدنش و تعیین حالت دقیق هر بخش است ولی برای این کار از نمایش سنتی انسان که آن را به عنوان مجموعه‌ای از بخش‌های کوچک به هم وصل‌شده در نظر می‌گرفتند، اجتناب کرده و مفهوم جدیدی به نام حالتک سلسله مراتبی را جایگزین این نحوه نمایش کرده است. در این شیوه نمایش، یک بخش تمام خاصیت‌های حالتک را به ارث برده است یعنی می‌تواند از یک بخش کوچک واقعی شروع شده و در حالت حدی، تا اندازه کل بدن بزرگ فرض شود. این تعریف عمومی‌تر به ما امکان می‌دهد تا الگوهای تمایزی‌تری برای تشخیص بخش‌ها داشته باشیم؛ چراکه معمولاً الگوی بخش‌های کوچک و ثابت مانند دست، پا و ... از قابلیت تمایز کافی برخوردار نیست. مثلاً اکثر الگوریتم‌های تشخیص بخش، مجبورند از ویژگی‌های پایه مانند خطوط موازی، دایره‌ها و ... برای تمایز الگوی بخش موردنظر خود استفاده کنند در حالیکه وجود تعداد زیادی از نمونه‌های این الگوها در پس‌زمینه و سایر اشیای تصویر باعث کاهش قدرت تشخیص این الگوریتم‌ها می‌شود. در عوض، بخش‌های بزرگ‌تر بدن مانند حالت پایین تنه، دست راست و سر، بالاتنه و ... الگوهای قوی‌تری برای تمایز در اختیار ما می‌گذارند که در پس‌زمینه یافت نمی‌شود.

این الگوریتم، بدن را متشکل از ۲۰ حالتک فرض کرده که در ساختار سلسله‌مراتبی با هم ارتباط دارند. بخش ریشه، کل بدن است و برگ‌ها همان بخش‌های پایه در روش سنتی هستند. هر بخش، اطلاعاتی مانند محل دقیقش، شناسه حالتک یافت‌شده و درصد قطعیت را برای سایر بخش‌هایی که در درخت ذکرشده با آن ارتباط دارند مبادله می‌کند. یعنی این انتقال اطلاعات به صورت دوطرفه است. یکبار از بالا به پایین اطلاعات توزیع می‌شود و سپس از پایین به بالا تجمیع اطلاعات صورت می‌گیرد



شکل ۲.۴: ساختار سلسله مراتبی و ارتباط بین لایه‌ای حالتک‌ها برای شکل‌دهی بدن انسان [۳۱]

تا محل دقیق اعضای کوچک در قالب اعضای بزرگ مشخص شود. در داخل هر گره، هر بخش با مقایسه تصویر فضای بخش‌های پدرش با الگویی که الگوریتم آموزش برای آن بخش یاد گرفته، سعی در یافتن نمونه‌هایی از آن الگو در آن فضاها می‌کند و بین نمونه‌های یافت شده، محتمل‌ترین را به عنوان حالتک یافت شده برای این بخش به سایر بخش‌های زیرمجموعه‌اش معرفی می‌کند. در برگشت هم، اطلاعات بخش‌های کوچک‌تر با هم مقایسه می‌شود و محتمل‌ترین ترکیب به عنوان وضعیت نهایی این بخش به پدرانش معرفی می‌شود.

این الگوریتم در کاربرد تجزیه اجزای انسان بسیار موفق بوده و روی پایگاه داده UIUC people توانسته درصد موفقیت ۶۷ بگیرد در حالیکه بهترین الگوریتم بعد از آن نتیجه ۵۱ درصد را داشته. به صورت خلاصه می‌توان استفاده از بخش‌های تمایزی‌تر، ساختار سلسله مراتبی و انتقال اطلاعات به تمام بخش‌های مرتبط را دلایل موفقیت این الگوریتم دانست. در عوض، این رفت و برگشت اطلاعات بسیار زمانگیر خواهد بود و این الگوریتم را در کاربردهای بلادرنگ دچار تاخیر می‌کند. همچنین برای آموزش تشخیص‌دهنده‌های حالتک‌ها به حجم زیادی داده آموزشی با برچسب‌های دقیق نیاز است که کار جمعآوری چنین پایگاه داده‌ای را زمانبر می‌کند. از طرفی اگر قرار به استفاده از این روش برای تخمین سه بعدی باشد، باید این برچسب‌ها به صورت مختصات سه‌بعدی باشند که برای تصویرهای واقعی و بدون داشتن سنسور، امکان پذیر نیست.

۳.۲ نتیجه‌گیری

در این فصل به بررسی مقالات مرتبط با مسئله‌ی تخمین حالت انسان که رویکرد مولد داشتند پرداختیم. در ابتدا روش‌های تخمین سه بعدی حالت انسان را به دو دسته‌ی کلی روش‌های مولد و روش‌های

تمایزی تقسیم کردیم. روش‌های مولد روش‌هایی هستند که با جستجوی فضای حالت به دنبال نقاطی که نزدیک‌ترین ورودی به ورودی داده‌ی آزمون را تولید می‌کنند و سپس استفاده از قانون بیز برای محاسبه‌ی احتمال درستی این نقاط و برگرداندن محتمل‌ترین نقطه، تخمینی از حالت انسان می‌زنند. این روش‌ها به دلیل پیچیدگی و اغلب غیر محدب بودن جستجو در فضای حالت و نیاز آنها به قدرت پردازشی بالا و مشکلاتی مانند انتشار خطا در این پژوهش چندان مورد تاکید نیستند. دسته‌ی دوم روش‌ها،روش‌های تمایزی هستند که به دنبال پیدا کردن تابعی مستقیم بین داده‌های ورودی و حالت خروجی هستند. درباره این روش‌ها در فصل بعد به طور مفصل صحبت خواهیم کرد.

فصل ۳

روش‌های مبتنی بر یادگیری

در فصل پیش، کلیت روش‌های مولد را به عنوان یکی از روش‌های آماری به کار رفته برای تخمین حالت به صورت مختصر توضیح دادیم. در کاربردهای بینایی ماشین، همواره دو بخش اصلی است که تعیین‌کننده عملکرد رویکردهاست: مشخصه استفاده شده و روش احتمالی استفاده شده. در تخمین حالت هم، این موضوع قابل مشاهده است. برخی کارها، تمرکز خود را روی انتخاب یا تولید یک مشخصه خوب گذاشته‌اند در حالیکه بقیه، بهبود روش‌های یادگیری الگوی به کار رفته در تخمین حالت را هدف اصلی کار خود قرار داده‌اند. در این فصل قصد داریم نمونه‌هایی از روش‌ها و رویکردهای مبتنی بر یادگیری را بررسی کنیم. همانطور که قبلاً ذکر شد، این روش‌ها به نام روش‌های تمایزی نیز شناخته می‌شوند. فصل بعد را نیز به انتخاب مشخصه اختصاص داده‌ایم.

۱.۳ پایه‌ی عملکرد روش‌های تمایزی

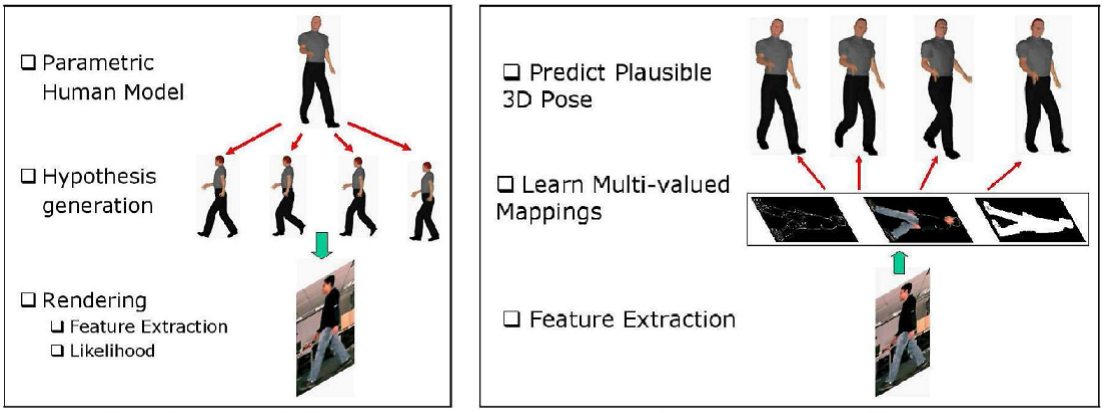
در فصل دوم گفتیم که روش‌های تمایزی مستقیماً $p(y|x)$ را تخمین می‌زنند. این روش‌ها یک فرم پارامتری یا غیرپارامتری برای توزیع شرطی $p(x|y)$ در نظر گرفته و پارامترهای آن را با استفاده از نمونه‌های آموزشی یاد می‌گیرند. در رویکرد تمایزی از این واقعیت که فضای حالتی که در عمل اتفاق می‌افتد بسیار کوچکتر از فضای حالات ممکن است استفاده می‌کنند. این روش‌ها مشکلات نیاز به مدل‌سازی دقیق سه‌بعدی و مقداردهی اولیه حالات در این رویکرد تا حد خوبی رفع شده است اما این روش‌ها نیز مشکلاتی دارند که تلاش مقاله‌های ارائه شده رفع یا کاهش چنین مشکل‌هایی است. بزرگترین مشکل این روش‌ها تناظر یک به چند بین فضای تصاویر ورودی و فضای حالت است. این موضوع باعث می‌شود که از نظر تئوری نتوان تابعی پیدا کرد که داده‌های ورودی را به

حالت‌های خروجی نگاشت کند. در عمل نیز به دلیل نگاشته شدن داده‌هایی با ورودی بسیار نزدیک به خروجی‌هایی کاملاً متفاوت، چنین تابعی هر چند وجود دارد اما پیدا کردن آن کار بسیار دشواری است. برای رفع این مشکل معمولاً سعی می‌شود تا با استفاده از اطلاعاتی که از طریق داده‌های برچسب‌دار در مورد فضای خروجی می‌توان کسب کرد، به نوعی ابهام موجود در نگاشت بین ورودی و خروجی را کاهش داد . یکی دیگر از مشکلات روش‌های تمایزی بعد بالای فضای خروجی است. این در حالیست که بین ابعاد مختلف خروجی داده‌ها در این فضای با بعد زیاد همبستگی قابل توجهی وجود دارد. این همبستگی همان چیزی است که باعث می‌شود ما تنها با بخش کوچکی از فضای خروجی سروکار داشته باشیم. اما پیدا کردن آن و استفاده از آن در تخمین حالت به دلیل پیچیدگی همبستگی بین اعضای مختلف بدن کار بسیار دشواری است. روش‌های کاهش بعد برای محدود کردن فضای جستجو و استفاده از ارتباط بین نقاط این فضا به کار آمده‌اند[۲۴، ۲۵، ۲۶]. در نهایت یکی دیگر از مشکلات روش‌های تمایزی نیاز به داده‌های آموزش زیاد است، چراکه هرچند حالت‌هایی که بدن انسان به خود می‌گیرد بخش کوچکی نسبت به کل فضای حالت را در بر می‌گیرد اما همین بخش حالت‌های بسیار زیادی را شامل می‌شود.

با وجود مشکلات گفته شده، تجربه نشان داده است که روش‌های تمایزی بهتر از روش‌های مولد عمل می‌کنند. البته این روش‌ها در مسئله‌ی تخمین حالت انسان رویکری نسبتاً نوپا به حساب می‌آیند و مقایسه‌ی مستقیم چندانی هم بین این دو دسته روش صورت نگرفته است. در ادامه‌ی این قسمت مقالات مهم ارائه شده که رویکردی تمایزی به مسئله‌ی تخمین حالت انسان داشته‌اند را به دو بخش تقسیم کرده‌ایم. در بخش اول به بررسی مقالاتی می‌پردازیم که به یادگیری یک تابع رگرسیون خوب می‌پردازند. بخش دوم که به طور مفصل در فصل بعد بررسی خواهند شد فعالیت‌هایی را شامل می‌شود که به یادگیری مشخصه‌های خوب توجه داشته‌اند. هرچند بخش‌های ذکر شده با هم اشتراك دارند اما سعی شده است مقالاتی در هر بخش قرار گیرند که جنبه‌ی مورد بحث در آن بخش در آنها پررنگ‌تر باشد. درشکل ۱.۳ کلیت مراحل استنتاج روش‌های مولد و تمایزی نشان داده شده است.

۲.۳ روش‌های مبتنی بر یادگیری نظارتی

در علم یادگیری ماشین، اصولاً سه رویکرد کلی برای یادگیری وجود دارد: یادگیری نظارتی، نیمه نظارتی و بدون نظارت. روش‌های نظارتی بر اساس داده‌های برچسب‌دار آموزشی، تابع رگرسوری را برای تعیین برچسب داده‌های تست یاد می‌گیرند در حالیکه روش‌های بدون نظارت، فقط از ساختار کلی داده‌ها برای یادگیری این تابع استفاده می‌کنند. روش‌های بدون نظارت زمانی کاربرد دارند که



Top-down modeling Bottom-up modeling

شکل ۱.۳ : مقایسه روش‌های مولد و تمایزی

داده برچسب دار وجود ندارد یا به صرفه نیست. در این میان، روش‌های نیمه نظارتی هم هستند که سعی می‌کنند مزایای هر دو روش را ترکیب کنند، یعنی هم از داده‌های برچسب‌دار استفاده می‌کنند و هم از ویژگی‌های ساختاری داده‌های بدون برچسب. بررسی جزئیات بیشتر نحوه کار هر روش از حوزه این پژوهش خارج است.

در این بخش قصد داریم یکی از روش‌های تخمین حالت انسان، مبتنی بر یادگیری نظارتی را معرفی کنیم. این روش [۱۱] در سال ۲۰۰۶ در کنفرانس CVPR معرفی شد و سپس نسخه کامل‌تر آن در سال ۲۰۰۸ در مجله PAMI به چاپ رسید. در این روش برای اولین بار از مشخصه‌ای به نام هیستوگرام زمینه شکل[۱] برای نمایش سیاه‌نمای یک تصویر استفاده شده و عملیات استنتاج و تخمین حالت نیز با استفاده از یک رگرسیون غیرخطی به نام ماشین بردار وابسته[۲] انجام گرفته است. از آنجا که به دست آوردن حقیقت زمینه برای تخمین حالت بسیار سخت است و اغلب هم به دلیل خطای دید، دقیق نیست، اکثر الگوریتم‌ها به مقایسه بصری کفایت کرده‌اند اما در این پژوهش، از نرم‌افزار Poser برای ساخت مدل بدن انسان استفاده شده، در نتیجه، مختصات و زوایای دقیق تمام اعضای بدن در اختیار است. البته برای ارزیابی روش در محیط‌های واقعی، الگوریتم با تصاویر و ویدئوهای واقعی نیز تست شده و نتایج قابل قبولی ارائه شده است. مدلی که ما استفاده می‌کنیم با توجه به امکاناتی که نرم‌افزار Poser در اختیار گذاشته، از ۱۸ مفصل بدن تشکیل شده که برای هر کدام نیز ۳ درجه آزادی زاویه‌ای در راستای محورهای مختصات سه بعدی در نظر گرفته شده است. پس خروجی الگوریتم تخمین حالت، برداری ۵۴ درایه ای است که با تعیین یک محل فرضی برای ریشه بدن، میتوانیم

[۱] Histogram of Shape Context

[۲] Relevance Vector Machine

مختصات سایر بخش‌ها به صورت نسبی تعیین کنیم. نکته دیگری که در مورد این روش حائز اهمیت است، تمایزی و ساده بودن آن است به صورتیکه نه تنها نیازی به تعریف یک مدل ساختاری از بدن انسان ندارد بلکه تعداد پارامترهای الگوریتم نیز بسیار محدود است. در واقع نگاشتی مستقیم بین فضای ورودی و خروجی حالت، توسط رگرسیون یاد و برای استنتاج، از آن بهره گرفته می‌شود.

در این بخش کارکرد دو نوع رگرسیون را برای تخمین حالت بررسی می‌کنیم. در معادلاتی که در این بخش استفاده شده فضای حقیقی ۱۰۰ بعدی ورودی با x و فضای ۵۴ بعدی خروجی با y نمایش داده شده. نکته قابل ذکر دیگر اینکه، اگرچه با توجه ارتباط x و y به دلیل ابهام مسأله تخمین حالت به صورت رابطه‌ای[۳] و نه تابعی[۴] است ولی در اینجا برای استفاده از رگرسیون، فرض کردیم که این رابطه میتواند به صورت یک تابع که همان تابع رگرسور است هم تخمین زده شود:

$$y = \sum_{k=1}^{p} a_k \phi_k(X) + \epsilon \qquad (۱.۳)$$

در اینجا $\{\phi_k(X)|k = 1...p\}$ توابع پایه، a_k ضرایب پایه‌ها و ϵ هم بردار خطای باقیمانده است. برای فشرده‌تر شدن معادله، بردار ضرایب در یک بردار به نام A که قرار است یاد گرفته شود تجمیع شده است و بردار پایه‌ها را نیز با $f(X) = (\phi_1, \phi_2, ..., \phi_p)^T$ نمایش داده می‌شود. با توجه به داده‌هایی آموزشی که جفت‌هایی به صورت $\{(y_i, x_i)|i = 1...n\}$ با n تصویر آموزش هستند، و اینکه در این مقاله از نرم دوم (فاصله اقلیدسی) بردار yها استفاده شده، مسأله تخمین حالت به مسأله بهینه‌سازی ۲.۳ تبدیل می‌شود:

$$A := arg\,min_A \Big\{ \sum_{i=1}^{n} \|Af(x_i) - y_i\|^2 + R(A) \Big\} \qquad (۲.۳)$$

که $R(-)$یک تابع هموارساز روی A است. با تجمیع تمام دوتایی‌های آموزشی در قالب ماتریس $m \times n$ خروجی به صورت $Y \equiv \{y_1 y_2 ... y_n\}$ و ماتریس مشخصه $p \times n$ به صورت $F \equiv \{f(x_1), f(x_2) ... f(x_n)\}$ مسأله تخمین به صورت زیر درمی‌آید:

$$A := arg\,min_A \{\|AF - Y\|^2 + R(A)\} \qquad (۳.۳)$$

۱.۲.۳ رگرسور Ridge

از آنجا که مسأله تخمین حالت، بدتعریف و دارای ابعاد بالاست، بدون داشتن عبارت هموارساز $R(A)$، تابع رگرسور دچار Overfitting می‌شود و نمی‌تواند برای داده‌های تست، به خوبی داده‌های آموزش کار

---

[۳] relational

[۴] functional

کند. راحت‌ترین راه برای تأمین این همواری، تابع $\lambda\|A\|^2$ است که در آن باید پارامتر λ را به اندازه کافی و لازم بزرگ انتخاب کنیم تا مقادیری از ماتریس ضرایب A که به صورت ناهمگون با بقیه بزرگ هستند را متناسب با اندازه‌شان جریمه کنیم. با این کار، ضرایب هموارتر می‌شوند و عمومیت الگوی یافت‌شده توسط رگرسور افزایش می‌یابد. البته باید در انتخاب λ دقت کرد، چون مقادیر بیش از حد بزرگ آن باعث می‌شود ماتریس A به سمت صفر میل کند و مجدداً دقت خروجی بسیار کم شود. به این نوع رگرسور، Ridge می‌گویند و معادله ۳.۳ را به صورت زیر برای آن بازنویسی می‌کنند:

$$\|A\tilde{F} - \tilde{Y}\|^2 := \|AF - Y\|^2 + \lambda\|A\|^2 \qquad (۴.۳)$$

که در آن $\tilde{F} := (F \quad \lambda I)$ و $\tilde{Y} := (Y \quad ۰)$ هستند. جواب این مسأله با حل معادله خطی $A\tilde{F} = \tilde{Y}$ به وسیله تجزیه برداری یا معادلات نرمال به دست می‌آید. از آنجا که تغییر اندازه (ضرب ورودی در یک مقدار ثابت) در جواب رگرسور ridge موثر است باید قبل از حل معادله، ورودی را به صورت استاندارد و با واریانس یکسان در بیاوریم.

۲.۲.۳ ماشین بردار وابسته

ماشین بردار وابسته[۵] [۱۸، ۱۹] یک رویکرد مبتنی بر قانون احتمالاتی بیز و مفهوم بردار تنک به مسأله رگرسیون و دسته‌بندی کردن است. در این روش، شرط‌های پیشین برای پارامترها تعیین می‌شود که با فراپارامترها کنترل می‌شوند. اساس کار این ماشین هم مانند ماشین بردار پایه است که با یافتن بردارهای مهم‌تر فضای ورودی، سعی در کاهش حجم داده مورد استفاده و نیز کاهش تأثیر نویز دارد. انتگرال‌گیری روی فراورودی‌ها متغیرهای پیشینی را به صورت $p(a) \sim ||a||^{-v}$ برای هر پارامتری رگرسیون ارائه می‌دهد که بسیار غیرمحدب هستند و توسط v به عنوان فراپارامتر کنترل می‌شوند. با لگاریتم‌گیری از این عبارت، تابع جریمه $R(a) = v\log||a||$ برای هموارسازی به دست می‌آید که با مشتق‌گیری از آن نسبت به a، نیروی هموارسازی را به صورت $\frac{\partial R}{\partial a} \sim \frac{v}{||a||}$ خواهیم داشت. از این تابع برای جریمه پارامترهایی که تغییرات خروجی‌ها را بیش از اندازه ناهموار می‌کنند استفاده می‌شود. همانطور که مشخص است، این نیروی هموارسازی با a نسبت عکس دارد، یعنی به ازای مقادیر بزرگ a جریمه کمتر و تبع آن، هموارسازی کمتر را داریم. در حالیکه به ازای مقادیر کوچک a جریمه بالایی به چنین پارامترهایی تحمیل میک‌نیم و آنها نمیتوانند در رقابت با سایر پارامترها به

[۵] Relevance Vector Machine

سطوح بالای بهینگی دست پیدا کنند، در نتیجه پارامترهایی انتخاب می‌شوند که هموارسازی بهتری روی خروجی‌ها دارند. اتفاقی که در ماشین بردار وابسته می‌افتد اینست که به ازای مقادیر به اندازه کافی کوچک a، در تکرارهای الگوریتم، داده‌ها دیگر برای غیرصفر نگه‌داشتن پارامتر رگرسیون در مقابل نیروی هموارسازی کافی نیستند و پارامتر به سرعت به سمت صفر همگرا می‌شود. از همین روست که دسته پارامترهای رگرسیون حاصل از RVM تنک است. این الگوریتم به صورت خودکار، زیرمجموعه‌ای از مرتبطترین بردارهای پایه برای بازسازی تابع رگرسیون را انتخاب می‌کند که تنک نیز باشند. از آنجا که ضرب ورودی‌ها در یک مقدار ثابت، فقط مخرج نیروی هموارسازی را به نسبت آن مقدار، تغییر می‌دهد، به صورت نسبی، تاثیری در خروجی الگوریتم نخواهد داشت. به عبارت دیگر، RVM نسبت به تغییر اندازه مقاوم است. این الگوریتم در حالت کلی، غیرمحدب است و اکسترمم‌های محلی زیادی در آن وجود دارد که فرآیند بهینه‌سازی را با مشکل مواجه می‌کند، زیرا ممکن است بردارهای وابسته، به صورت اتفاقی در دام این اکسترمم‌ها گیر بیفتند و تبدیل به صفر شوند. اما در عمل می‌بینیم که با توجه به نوع داده‌ها، این اتفاق کم رخ می‌دهد و برآیند کلی الگوریتم مثبت است. قابل ذکر است که خروجی رگرسیون ،RVM دقتی برابر با بهترین روش‌های موجود از جمله SVM و گواسین دارد در حالیکه تعداد پایه های استفاده شده در آن بسیار کمتر است و این یک موفقیت قابل توجه در ایجاد مدل ساده و کارا به شمار می‌آید.

برای آموزش RVM با استفاده از داده‌های آموزش، از یک روش پیوسته برای تخمین متوالی هموارسازهای $v\log||a||$ استفاده می‌کنیم که از نمودار پل‌های درجه دوم $v(||a||/a_{scale})^2$ بهره میبرد. این کار باعث می‌شود تا حدی از به دام افتادن زودهنگام پارامترها در تله صفر جلوگیری شود. زیرا پارامترها با حرکت روی سهمی که به عنوان پل عمل می‌کند از محور صفر عبور می‌کنند و با ریسک کمتری نسبت به تابع اکیداً نزولی لگاریتم مواجه هستند. ما برای تأمین تنک بودن، برای تابع هموارساز از ترکیب ستونی متغیرها استفاده کردیم به صورتی که یک مجموعه مشترک بردار پایه مرتبط برای تمام ستون‌های A انتخاب می‌شود. پس داریم: $R(A) = v\sum_k \log||a_k||$ که در آن a_k، kامین ستون A است. پس عبارتی که باید ماکزیمم شود به این صورت است:

$$||AF - Y||^2 + v\sum_k \log||a_k|| \qquad (۵.۳)$$

برای انتخاب پایه های رگوسور، ما دو روش را تست کردیم. ۱) پایه‌های خطی ۲) پایه‌های هسته‌ای[۶]. در روش اول، تابع، همان بردار ورودی را برمی‌گرداند و در نتیجه، RVM بردارهای مرتبط

[۶] kernel

RVM Training Algorithm

1) Initialize A with ridge regression. Initialize the running scale estimates $a_{\text{scale}} = \|\mathrm{a}\|$ for the components or vectors a.
2) Approximate the $\nu \log \|\mathrm{a}\|$ penalty terms with "quadratic bridges", the gradients of which match at a_{scale}. *I.e.* the penalty terms take the form $\frac{\nu}{2}(\mathrm{a}/a_{\text{scale}})^2 + \text{const}$.
(One can set const $= \nu(\log\|a_{\text{scale}}\| - \frac{1}{2})$ to match the function values at a_{scale}, but this value is irrelevant for the least squares minimization.)
3) Solve the resulting linear least squares problem in A.
4) Remove any components a that have become zero, update the scale estimates $a_{\text{scale}} = \|\mathrm{a}\|$, and continue from 2 until convergence.

شکل ۲.۳: الگوریتم ماشین بردار وابسته (تصویر از [۱۱])

را که در واقع اجزای ف٠غ ورودی هستند انتخاب میکند. در روش هسته، پایه‌ها توسط توابع هسته $K(x, x_i)$ که با ورودی‌ها تغذیه شده‌اند ساخته می‌شوند. آزمایش‌ها نشان داد که پایه‌های هسته‌ای نسبت به حالت خطی، بهبود بیشتری در نتایج دارند ضمن اینکه انتخاب نوع هسته هم تأثیر قابل توجهی در این بهبود ندارد. ما در این پژوهش از هسته گواسین به صورت $K(x, x_i) = e^{-\beta||x-x_i||^2}$ استفاده می‌کنیم که پارامتر βی آن از انحراف معیار داده‌های آموزش یاد گرفته می‌شود. شمای کلی الگوریتم RVM در تصویر ۲.۳ نشان داده شده است.

۳.۳ روش‌های مبتنی بر یادگیری نیمه‌نظارتی

پژوهش‌های نیمه نظارتی انجام شده برای تخمین حالت انسان را می‌توان به سه دسته کلی تقسیم کرد: روش‌های مبتنی بر فرآیندهای گاوسی [۳۳، ۳۴]، روش‌های مبتنی بر رویه[۳۵، ۲۱] و روش‌های مکاشفه‌ای[۳۶]. در این بخش به بررسی دو دسته‌ی اول روش‌های مبتنی بر یادگیری نیمه نظارتی می‌پردازیم.

۱.۳.۳ روش‌های مبتنی بر فرآیند گاوسی

روش‌های مبتنی بر فرآیندهای گاوسی برای توابع خروجی ممکن بر روی اجتماع داده‌های آموزش و تست یک توزیع گاوسین در نظر می‌گیرند. هدف این روش‌ها محدود کردن این توزیع تا حد ممکن و سپس اعلام محتمل‌ترین تابع این توزیع به عنوان تخمینی از تابع خروجی است. در این روش‌ها، ابتدا توسط داده‌های برچسب‌دار، توزیع اولیه‌ی توابع ممکن تولید می‌شود. سپس داده‌های بدون برچسب به کمک می‌آیند تا توزیع توابع ممکن را محدودتر کنیم. در نهایت هم، امید ریاضی این توزیع توابع، محتمل ترین تابعی که نگاشت بین ورودی و خروجی را مدل می‌کند به دست می‌دهد.

این دسته از روش‌ها بر یک اصل اساسی استوارند که عبارتست از : توابع مختلفی که می‌توانند عمل رگرسیون را بین فضای ورودی و خروجی انجام دهند دارای یک توزیع نرمال هستند. در یکی از این روش‌ها که TGP [۳۷] نام دارد، داده‌های برچسب‌دار توزیع اولیه‌ی توابع ممکن را محدود می‌کنند و توزیع احتمال درست‌نمایی اولیه را به توزیعی پسین بر روی توابع تبدیل می‌کنند. سپس داده‌های بدون برچسب به کمک می‌آیند تا توزیع توابع ممکن را با در دست داشتن داده‌های برچسب‌دار، برچسب آنها و داده‌های بدون برچسب به دست آوریم. امیدریاضی این توزیع توابع، محتمل ترین تابعی که نگاشت بین ورودی و خروجی را مدل می‌کند به دست می‌دهد. مشکل اصلی فرآیندهای گاوسی این است که از ساختار داده‌ها استفادهی مستقیمی نمی‌کنند بلکه این داده‌ها تنها در محدود کردن توزیع توابع برچسب‌دهی ممکن استفاده می‌شوند.

۲.۳.۳ روش‌های مبتنی بر رویه

کار کردن با داده‌های با بعد بالا یکی از اساسی ترین مشکلات اکثر مسائل تشخیص الگو است. این مشکل هرچند در کارکردن با داده‌های ورودی در اکثر مسائل از جمله تخمین حالت انسان وجود دارد، اما هنگامی که داده‌های خروجی نیز بعد بالایی داشته باشند(مثل تخمین حالت)، این مسئله بسیار شدیدتر می‌شود. در برخورد با داده‌های ورودی با بعد بالا، از آنجا که داده‌ها به ما داده می‌شوند می‌توان با استفاده از روش‌های کاهش بعد، تعداد ابعاد را کاهش داد به گونه‌ای که کمترین حجم اطلاعات از دست برود. یکی از مهم‌ترین روش‌های کاهش بعد، استخراج رویه با بعد کمتر از داده‌هاست. در اصلاح هندسی، رویه را فضایی مسطح و توپولوژیک تعریف می‌کنند که با گام‌های به اندازه کوچک می‌تواند فضای اقلیدسی متناظر خود را پوشش دهد. برای مثال، خط یا دایره، رویه‌های یک بعدی و سطح، یک رویه دو بعدی است. هر رویه قابل تخمین زدن به صورت یک گراف معمولاً وزن‌دار است. مثلاً می‌توانیم کره زمین را که سه‌بعدی است با رویه‌ای دو بعدی نشان دهیم و تمام نقاط آن را

نیز بپوشانیم. فرض منطقی‌ای که روشهای مبتنی بر رویه در نظر گرفته می‌شود چنین است: اگر گرافی وزندار داده‌های ورودی را به یکدیگر متصل کند به گونه‌ای که وزن یال بین دو داده ارتباط مستقیمی با میزان شباهت آنها داشته باشد، نمونه‌هایی که فاصله‌ی کمی بر روی گراف دارند احتمالا برچسب مشابهی خواهند داشت. بنابراین، این روش‌ها معمولا گرافی بر روی داده‌های ورودی می‌سازند که در آن وزن یال‌ها متناسب با شباهت نقاط است. سپس بر روی این گراف پیش‌فرض رویه در قالب یک عبارت منظم‌ساز اعمال می‌شود. در اعمال پیش‌فرض رویه، این روش‌ها معمولا از عملگر لاپلاسین روی گراف برای محاسبه‌ی تغییرات تابع برچسب‌دهی بر روی رویه استفاده می‌کنند.

برای معرفی هر روش یادگیری نیمه‌نظارتی مبتنی بر گراف باید نحوه‌ی ساخت گراف همسایگی، روش اعمال فرض خمینه و روش تخمین برچسب‌ها به کمک فرض خمینه را تعیین کنیم که برای هر یک هم روش‌های متنوعی وجود دارد. در اینجا دو روش ساخت گراف را معرفی می‌کنیم.

روش آستانه‌گذاری اولین راهی که برای مدل کردن چگالی داده‌های یک منطقه به نظر می‌رسد، متصل کردن هر گره به تمام گره‌های همسایه‌ی آن در یک شعاع همسایگی است. بدین ترتیب گره‌هایی که در مناطق چگال قرار دارند دارای درجه‌ی بالایی خواهند شد. هر چند از نظر تئوری این روش قابلیت مدل کردن توزیع حاشیه‌ای داده‌ها را دارد، اما مشکل اصلی این روش تعیین مقدار پارامتر است. مقادیر کوچک این پارامتر ممکن است منجر به گراف‌های بسیار تنک یا حتی ناهمبند شود و مقادیر زیاد این پارامتر باعث می‌شود گراف بسیار چگال یا حتی کامل شود.

روش نزدیک‌ترین$\mathbf{k}$- همسایه در این روش هر گره به نزدیک‌ترینk همسایه‌اش متصل می‌شود. برای اینکه ماتریس لاپلاسین به دست آمده از چنین گرافی مثبت نیمه معین باشد، یال‌های این گراف بدون جهت در نظر گرفته می‌شوند و ماتریس همسایگی متقارن خواهد شد. این امر باعث می‌شود در عمل درجه‌ی گره‌ها در مناطق چگال بسیار بالاتر از مناطق تنک باشد.

برای اعمال فرض خمینه هم روش‌های زیادی وجود دارد ولی یکی از قدیمی‌ترین روش‌ها استفاده از ماتریس لاپلاسین است. اگر $f = f_1, f_2, ..., f_{l+u}$ تابع برچسب‌دهی و W ماتریس مجاورت گراف همسایگی رویه در روش kنزدیک‌ترین همسایه باشد ماتریس لاپلاسین این گراف به صورت $L = D - W$ تعریف می‌شود که در آن، D ماتریس قطری درجات رئوس گراف است. با مشتق‌گیری و محاسبات جبری به دست میآید که برای کمینه کردن خطای بازسازی گراف از رویه باید عبارت $f^T L f$ کمینه شود. به عبارت دیگر، رأس‌هایی که در گراف نزدیک هم هستند برچسب شبیه هم خواهند داشت. اما از آنجا که اکثر روش‌های مطرح کاهش بعد برگشت پذیر نیستند کار کردن در

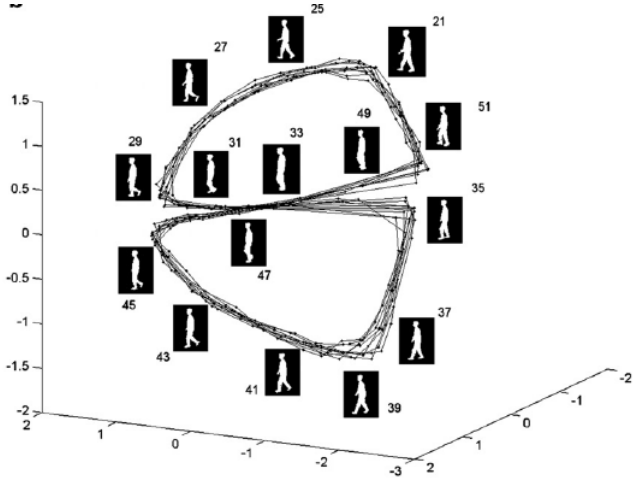
یک فضای خروجی کاهش بعد یافته و سپس پیدا کردن خروجی‌های اصلی از روی خروجی‌های تخمین‌زده شده در این فضا کار آسانی نیست. بدین ترتیب باید روشی برای پیدا کردن نگاشتی از فضای کاهش بعد یافته‌ی خروجی به فضای اصلی خروجی پیدا کنیم[۲۵، ۲۷]. روش‌های مختلفی برای این کار معرفی شده که از بحث این پژوهش خارج است. در ادامه به بررسی یکی از مقالاتی که در این دسته قرار میگیرد می‌پردازیم.

خلاصه‌ی این کار[۲۴] که در سال ۲۰۰۴ ارائه شده بدین صورت است که با استفاده از تعبیه‌ی محلی خطی[۷] و با کمک برچسب‌هایی که از داده‌ها در اختیار است، رویه‌ای که حالت‌های مختلف راه رفتن بر روی آن قرار می‌گیرند را تخمین زده‌است. سپس با در دست داشتن داده‌های آموزش و توسط توابع پایه‌ی شعاعی تعمیم‌یافته، نگاشتی از این رویه به فضای ورودی یافته و سپس با پیدا کردن معکوس این نگاشت، نگاشت ورودی به این رویه را پیدا کرده است. از طرف دیگر، نگاشت معکوس بین رویه و داده‌های خروجی نیز توسط توابع پایه‌ی شعاعی تعمیم یافته تخمین زده شده است. بدین ترتیب ورودی هر داده‌ی تست به رویه‌ی فعالیت نگاشته شده و سپس از آنجا به فضای خروجی نگاشته می‌شود. این پژوهش علاوه بر اینکه روشی خوب و تنها مبتنی بر کاهش ابعاد برای تخمین حالت انسان ارائه کرده است، با نمایش داده‌های خروجی در فضای کاهش بعد یافته به شکل تجربی ثابت کرده است که حالت‌های راه رفتن انسان بر روی رویه‌ای با بعد حداکثر سه قرار می‌گیرند۳.۳. همچنین از آنجا که نگاشتی از داده‌های ورودی به این رویه‌ی حداکثر سه بعدی پیدا شده است، این مقاله به شکل ضمنی نشان داده‌است که داده‌های ورودی حرکت راه رفتن نیز بر روی رویه‌ای با بعد حداکثر سه قرار می‌گیرند.

اگرچه در ابتدا این پژوهش فقط روی فعالیت راه رفتن انجام شد ولی برای تعمیم آن، در سال ۲۰۰۷ نویسنده همین مقاله[۲۴]، آزمایش خود را روی برخی فعالیت‌های پیچیده‌تر انسان نیز تکرار کرد[۲۶] و نشان داد که در همه آن‌ها با تقریب خوبی، داده‌های خروجی فضای حالت روی یک رویه با بعد کم قرار می‌گیرند. از طرف دیگر نشان داد که فرض همواری این رویه نیز برقرار است. بنابراین، فرض غیرموجهی نیست اگر فضای حالت همه فعالیت‌های انسان را روی رویه‌های با ابعاد کم تصور کنیم. ما نیز از این شهود در کار خود بهره برده‌ایم.

---

[۷] Local Linear Embedding



شکل ۳.۳: رویه‌های فعالیت راه رفتن [۵]

۴.۳ نتیجه‌گیری

در این فصل به بررسی مقالات مرتبط با مسئله‌ی تخمین حالت انسان که رویکرد تمایزی داشتند پرداختیم. این روش‌ها را در دو دسته کلی یادگیری نظارتی و نیمه‌نظارتی طبقه‌بندی کردیم و از هر دسته، مثال‌هایی را معرفی کردیم. روش‌های رگرسیون Ridge و ماشین بردار وابسته را از دسته نظارتی و روش‌های مبتنی بر گواسین و رویه را هم از دسته نظارتی برشمردیم. همچنین، روش‌های TGP و رویه‌ی فعالیت را به ترتیب به عنوان مثالی از روش‌های مبتنی بر گواسین و رویه مطرح کردیم.

فصل ۴

یادگیری مشخصه

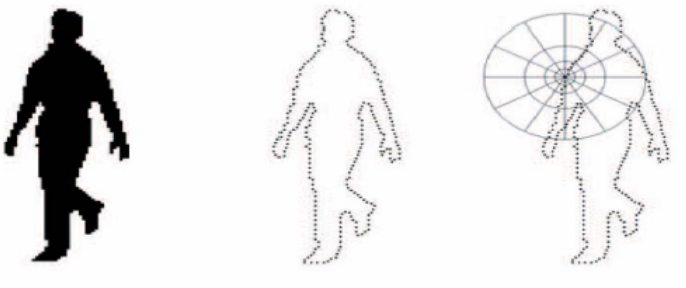
در فصل قبل، روش‌های تمایزی به کار رفته برای تخمین حالت را مختصراً معرفی کردیم. اما برای رسیدن به نتایج مطلوب، تنها استفاده از روش مناسب کارساز نیست و باید مشخصه‌های خوبی نیز به کار گرفته شود. مشخصه‌ای خوب است که ضمن دربرداشتن اطلاعات کافی از ظاهر انسان موجود در تصویر، ابعاد کمی هم داشته باشد تا سرعت و حافظه پردازشی را به وضعیت بلادرنگ نزدیک‌تر کند. در این فصل قصد داریم مشخصه‌های به کار رفته برای تخمین حالت را بررسی و مشخصه معرفی‌شده در این پژوهش را شرح دهیم. از آنجا که بسیاری از روش‌ها از سیاه‌نما به عنوان مشخصه‌ی ورودی اصلی خود استفاده می‌کنند ما هم در این فصل، آن را مبنای ساختن مشخصه‌های سطح بالاتر در نظر می‌گیریم. سیاه‌نما سه مزیت عمده نسبت به سایر توصیف‌کننده‌ها دارد: ۱) تقریبا به صورت قابل اطمینان از تصویر استخراج می‌شود، درصورتیکه تصویر پس‌زمینه ثابت باشد یا به صورت یکنواخت حرکت کند و از مشکلات سایه‌ها هم اجتناب شده باشد. ۲) نسبت به مشخصه‌های سطح مانند نحوه پوشش و بافت، مستقل است. ۳) حجم قابل‌توجهی از اطلاعات حالت بدن انسان را در خود دارد بدون اینکه نیاز به هرگونه اطلاعات برچسب‌زنی وجود داشته باشد. اما در عوض، سیاه‌نماها مشکلاتی نیز دارند از این قبیل:۱) در تصاویر دنیای واقعی، به دلیل عمود نبودن زاویه تابش نور، معمولاً سایه وجود دارد، همچنین پس‌زمینه نیز غالباً ناهمگون است، این موارد، کار الگوریتم‌های قطعه بندی را که هنوز هم به کارایی ایده‌آل نرسیده‌اند دشوار می‌کند و باعث خطا در نتایج تخمین حالت می‌شود. ۲) سیاه‌نماها چندین درجه آزادی را از تصویر حذف یا مشاهده آنها را بسیار سخت می‌کنند. این مورد اغلب در مواردی که بخشی از بدن توسط بخش دیگر پوشیده می‌شود(انسداد) رخ می‌دهد. اگرچه ترکیب سیاه‌نما با سایر مشخصه‌ها مثل نقشه لبه داخلی بدن می‌تواند در ابهام‌زدایی از این موارد مؤثر باشد، ولی به دلیل سختی این کار و تاثیر کم آن، در اکثر پژوهش‌ها از این ترکیب خودداری شده است.

۱.۴ استخراج سیاه‌نما

اولین مرحله برای تولید مشخصه لازم برای تخمین حالت، استخراج سیاه‌نماست. این کار (تشخیص و جدا سازی انسان از تصاویر) که از زیرشاخه‌های قعطه‌بندی تصویر محسوب می‌شود به واسطه پیچیده بودن آن، علی‌رغم نتایج خوب بدست آمده برای آن، هنوز جای کار زیادی دارد. گام اول، جداسازی انسان از بقیه‌ی تصویر است که بدون فرض‌های محدود کننده بسیار سخت است. اما با فرض ثابت بودن پس‌زمینه که با توجه به ثابت بودن دوربین در اکثر کاربردهای تخمین حالت انسان از جمله وارسی ویدئو، بازی‌های رایانه‌ای، ورزشی و غیره فرض معقولی است، روش‌های کارآمدی برای جداسازی پس‌زمینه از انسان وجود دارد. اکثر این روش‌ها بر پایه‌ی تفاضل تصویر ثابت پس‌زمینه از تصویر موجود که شامل پیش‌زمینه می‌شود و اعمال یک آستانه برای جداسازی این دو قسمت از هم بنا شده‌اند [۲۰، ۲۸]. تفاوت در مدل نمایش پس‌زمینه و نحوه‌ی به روز رسانی این مدل روش‌های مختلف را به وجود آورده است. در سال ۹۹ روشی موثر برای جداسازی پس‌زمینه با فرض ثابت بودن آن ارائه شد [۲۰]. این روش حتی در مقابل تغییرات معمول پس‌زمینه مانند تغییرات جزئی نور به خوبی مقاوم است. سادگی و قدرت بالای این روش باعث شده است که از زمان ارائه‌ی آن تا کنون بسیاری از مقاله‌های تخمین حالت که از سیاه‌نما به عنوان ورودی استفاده می‌کنند با تغییراتی جزئی از این روش برای جداسازی سیاه‌نمای انسان از محیط استفاده کنند. در این مقاله ابتدا با استفاده از تصاویر آموزشی، یک الگوی گواسین برای پس‌زمینه یاد گرفته شده. سپس هر پیکسل تصویر ورودی به صورت ترکیبی از گاوسین‌ها مدل شده است. به روز رسانی پارامترها و وزن هر یک از گاوسین‌ها بدین صورت است که در فریم جدید هر پیکسل در صورت شباهت کافی به نزدیک‌ترین گاوسین مکان خود تخصیص داده می‌شود و وزن این گاوسین زیاد و وزن سایر گاوسین‌ها کم می‌شود. اگر شباهت پیکسل جدید به هیچ یک از گاوسین‌ها به اندازه‌ی کافی نباشد کم‌وزن‌ترین گاوسین کنار گذاشته شده و یک گاوسین به مرکز پیکسل جدید و با واریانس به اندازه کافی زیاد ایجاد خواهد شد. جداسازی پیکسل‌های پیش‌زمینه با مقایسه محتمل‌ترین گاوسین هر پیکسل با مدل پس‌زمینه و تشخیص شباهت به اندازه کافی انجام می‌شود.

۲.۴ نمایش سیاه‌نما

سیاه‌نما را می‌توان به صورت تصاویری خام مورد استفاده قرار داد و یا با نمونه‌برداری از لبه‌های آنها مختصات نقاط نمونه را به عنوان مشخصه‌ی ورودی به الگوریتم‌ها داد اما با استخراج مشخصه‌های



شکل ۱.۴: از چپ به راست، سیاه‌نمای استخراج شده توسط حذف پس‌زمینه، نقاط لبه‌ی نمونه‌برداری‌شده، استخراج زمینه‌ی شکل (تصویر از [۱۱])

معنی‌دار و با بعد کمتر برای نمایش سیاه‌نما می‌توان کار الگوریتم‌های تشخیص الگو را آسان‌تر کرد. سیاه‌نما به دو صورت کلی محلی و سراسری می‌تواند نمایش داده شود. برای تخمین حالت اجزای بدن به ویژگی‌های محلی نیاز داریم چراکه بسیار محتمل است که ترکیب‌های متفاوت محلی اجزا، مشخصه‌های سراسری یکسانی تولید کنند. از جمله روشهایی سراسری می‌توان Distance Transform را نام برد که هر پیکسل تصویر را با فاصله‌اش از نزدیک‌ترین مرز سیاه‌نما نمایش می‌دهد. همچنین، روش‌هایی مثل Hu Moments یا پواسون[۱] هم مشخصه‌هایی آماری از سرتاسر سیاه‌نما استخراج می‌کنند. در تخمین حالت، مشخصه‌ای مناسب است که نسبت به انتقال و تغییر اندازه غیروابسته در حالیکه نسبت به دوران و انعکاس باید وابسته باشد. از طرف دیگر، روش‌هایی مثل Contour Signature هم که پیکسل‌های روی مرز سیاه‌نما را به صورت چندتادرمیان، نمونه‌برداری و به صورت ترتیبی استفاده می‌کنند در برابر تغییرات سایه‌های سیاه‌نما و عدم پیوستگی آن پایدار نیستند.

در میان روش‌های محلی روش Shape Context (SC) که مقبولیت خوبی به دست آورده است را در ادامه بررسی می‌کنیم.

۳.۴ مشخصه هیستوگرام زمینه شکل

مشخصه‌ای که در اکثر مقالات از آن برای نمایش سیاه‌نماها استفاده شده است زمینه‌ی شکل[۲] است که توسط Mori در سال ۲۰۰۰ معرفی شد. ایده‌ی اصلی استخراج این مشخصه در شکل ۱.۴ قابل مشاهده است.

اگر فرض کنیم نمونه‌های $x_1, x_2, ..., x_n$ از مرز خارجی سیاه‌نما برداشته شده‌اند و x_i یک نقطه‌ی

[۱] Poisson

[۲] Shape Context

نمونه‌ی است، زمینه‌ی شکل در آن نقطه را با دنباله‌ی h_i نشان می‌دهیم و آن را بدین صورت تعریف می‌کنیم:

$$h_i(k) = |\{x_i : j \neq i, x_i \in bin_{xj}(k)\}| \qquad (۱.۴)$$

که در آن $bin_{xj}(k)$ قطاع $k-$ام حول نقطه‌ی x_jرا نشان می‌دهد. همان طور که در شکل ۱.۴ مشخص است زاویه‌ی این قطاع‌ها مقداری ثابت است و شعاع آنها به صورت نمایی زیاد می‌شود. به عبارت ساده‌تر فضای اطراف هر نقطه‌ی نمونه را به صورت قطبی-لگاریتمی در نظر می‌گیریم و تعداد نقاط نمونه که در هر یک از قطاع‌ها قرار می‌گیرند را شمرده و به شکل یک هیستوگرام نمایش می‌دهیم.

بدین ترتیب مجموعه‌ی زمینه‌های شکل هر سیاه‌نما را خواهیم داشت اما برای نمایش این مجموعه باید روش ساده‌تری نسبت به نمایش دنباله‌ای پیدا کرد چراکه ابعاد مشخصه بسیار بالا خواهد بود. همچنین، ترتیب آنها در دنباله هم دردسرساز است. برای حل این مشکل، [۱۱] استفاده از هیستوگرام زمینه‌ی شکل را توصیه کرده است. این مشخصه در واقع توزیعی از زمینه‌های شکل هر سیاه‌نما است. ایده‌ی اصلی هیستوگرام زمینه‌ی شکل خوشه بندی فضای زمینه‌ی شکل به تعدادی خوشه و نمایش هر نقطه‌ی فضا توسط برداری است که درایه‌ی $i-$ام آن میزان تعلق آن نقطه به خوشه‌ی $i-$ام را با وزنی گاوسی نشان می‌دهد. این روش وزندهی که اصطلاحاً به آن وزن‌دهی نرم گفته می‌شود باعث می‌شود که تغییرات اندك زمینه‌های شکل باعث تغییرات زیاد در هیستوگرام زمینه‌ی شکل نشوند. این مقاله نشان داد که این روش نمایش سیاه‌نماها با وجود بعد کم حجم زیادی از اطلاعات را حفظ می‌کند. مزیت عمده زمینه شکل نسبت به سایر مشخصه‌های مشتق شده ز سیاه‌نما، عدم وابستگی آن به دوران، تغییر اندازه و برگردان شدن است.

الگوریتم ذکرشده برای محاسبه این مشخصه دارای پارامترها و جزئیات زیادی هست. ما در این پژوهش، برای قابل مقایسه بود نتایج، روشی که در [۱۷] استفاده شده را برای تولید مشخصه‌مان به کار بردیم که به شرح زیر است. برای محاسبه زمینه شکل، ابتدا باید محدوده سیاه‌نما را مشخص کرد، سپس به تعداد مورد نیاز که متناسب با پایگاه داده، اندازه و کیفیت عکس و فعالیت انسان موجود در عکس است از لبه‌های آن، نقطه‌هایی را به عنوان نمونه برمی‌داریم. از آنجا که کدکردن تمام نقاط لبه، حجم زیادی داده تولید خواهد کرد و منظم بودن ویژگی را تضعیف می‌کند، لذا با نمونه‌برداری، ضمن کاهش قابل توجه حجم داده، ویژگی منظم‌تری تولید می‌کنیم. این تعداد معمولاً برای دقت‌های بالا، ۳۰۰ الی ۴۰۰ نمونه است. در مرحله بعد برای اعمال خاصیت محلی، یک دایره به قطر اندازه یک عضو بدن در نظر میگیریم و با تقسیم کردن آن در هر دو راستای شعاعی و زاویه‌ای، فضایی

لگاریتمی-قطبی[۳] ایجاد می‌کنیم. همانطور که در شکل ۱.۴ مشخص است، محیط دایره با ۱۲ خط و سطح آن با ۶ خط تقسیم‌بندی شده است و در مجموع، ۶۰ خانه را ایجاد کرده‌اند. با شروع از یک نقطه تصادفی لبه، دایره را به مرکز آن درنظر میگیریم و تعداد نقاط دیگر سیاه‌نما را که در ۶۰ خانه سطح دایره قرار گرفته‌اند می‌شماریم و به عنوان هیستوگرام آن نقطه درنظر می‌گیریم. با حرکت دادن این دایره روی تمام نقاط نمونه‌برداری شده و محاسبه هیستوگرام آنها، میتوانیم هر سیاه‌نما را با یک ماتریس n در ۶۰ که n ، تعداد نمونه‌هاست نمایش دهیم.

اکنون با در اختیار داشتن زمینه شکل هر سیاه‌نما، هنوز هم مقایسه دو سیاه‌نما کار زمانبری است و ابعاد بالای هر سیاه‌نما، الگوریتم رگرسیون را با مشکل مواجه می‌کند. بنابراین یه سطح دیگر از هیستوگرام کردن را نیز به کار می‌بریم تا هر سیاه‌نما را به یک بردار ۱۰۰ بعدی تقلیل دهیم. برای این کار، تمام بردارهای زمینه از تمام تصاویر آموزشی را بوسیله یک الگوریتم خوشه‌بندی مانند k-means به ۱۰۰ خوشه افراز می‌کنیم و مراکز این خوشه‌ها را به عنوان کتاب کد[۴] در نظر می‌گیریم. حال، برای کدکردن هر سیاه‌نمای ورودی، فاصله هر بردار زمینه آن را تا چند مرکز خوشه‌های نزدیکش محاسبه کرده و به عنوان امتیاز این بردار زمینه به آن خوشه‌ها در نظر می‌گیریم. در نهایت با جمع کردن امتیازهایی که به هر خوشه داده شده، یک بردار ۱۰۰ بعدی به نام هیستوگرام زمینه شکل حاصل خواهد شد که در الگوریتم رگرسیون از آن به عنوان نماینده یک سیاه‌نما استفاده می‌کنیم. این نحوه امتیازدهی را soft voting می‌نامند زیرا باعث می‌شود تأثیر مخرب کوانتیزه کردن فاصله‌ای که در فرآیند تولید هیستوگرام زمینه شکل رخ میدهد کمتر شود. آزمایش‌ها نشان داده که این مشخصه در برابر انسداد و خطاهای جزئی در قطعه‌بندی مقاوم است، و علی‌رغم ابعاد کم آن، حجم قابل قبولی از اطلاعات سیاه‌نما را در خود دارد.

اگرچه ما در اینجا، مدلی خاص شامل بردارهایی با اندازه مشخص برای ورودی و خروجی ارائه کردیم ولی روش ارائه شده در این مقاله، به هیچ وجه وابسته به این مدل‌ها نیست و در واقع هیچ حسی از معنی این داده‌ها ندارد؛ بلکه فقط به دنبال تابعی هموار می‌گردد که با جفت‌های ورودی و خروجی هماهنگی بیشتری داشته باشد. داده‌هایی که ما برای آموزش و تست الگوریتم استفاده می‌کنیم از پایگاه داده ضبط حرکت دانشگاه CMU اخذ شده. از آنجا که سیاه‌نمای این داده‌ها در آن پایگاه داده موجود نیست به ناچار برای حذف خطای استخراج سیاه‌نما از نرم‌افزار Poser بهره می‌گیریم. در واقع با دادن خروجی حالت به این برنامه، تصویر متناظرش و سیاه‌نمای آن تصویر را ایجاد می‌کنیم.

بعد از معرفی هیستوگرام زمینه‌ی شکل اکثر مقالات از این مشخصه برای نمایش سیاه‌نماها بهره

---

[۳] log-polar

[۴] codebook

می‌برند. مشخصه‌های دیگری نیز مانند [۱۵، ۱۶] برای نمایش سیاه‌نما استفاده شده‌اند اما به دلیل اینکه کاربرد آنها در تخمین حالت انسان بسیار محدود است، از معرفی آنها خودداری می‌کنیم.

۴.۴ کدگذاری تنک

مشخصه‌های ذکرشده در بالا هر کدام دارای مشکلاتی هستند. مشخصه‌های سراسری که اصولاً با ذات مسأله تخمین حالت سازگاری ندارند. روش HoSC هم فقط به ارتباط فاصله‌ای بین کلمات دیکشنری و مشخصه توجه می‌کند در حالیکه می‌دانیم این ارتباط باید به جهت هم وابسته باشد. یعنی با دوران تصویر، مشخصه‌ها یکسان نباشد. بنابراین با توجه به ماهیت داده‌های تخمین حالت، به نظر می‌رسد که استفاده از کدگذاری تنک برای حل این مشکل مفید باشد. چراکه در کدگذاری، هر عنصر با ترکیب خطی از کلمات دیکشنری توصیف می‌شود و در نتیجه، اطلاعات جهت هم در آن دخیل می‌گردد. طبق همین شهود، ما دو روش کدگذاری تنک را روی آن اعمال کردیم و مقایسه نتایج حاکی از بهبود نسبی تخمین حالت داشت. لازم به ذکر است که تاکنون پژوهشی در زمینه کاربرد کدگذاری تنک در تخمین حالت انجام نشده است و این پژوهش می‌تواند شروعی برای کارهای آتی در این زمینه باشد، به خصوص اینکه نتایج هم از لحاظ بهبود دقت تخمین، امیدوارکننده است. برای معرفی روش‌های استفاده شده، لازم است ابتدا کمی در مورد کدگذاری تنک صحبت کنیم.

همانطور که از عبارت تنک برمی‌آید، هدف کدگذاری تنک اینست که یک داده را که می‌تواند تصویر یا هر نوع داده ماتریسی دیگری باشد، بر حسب ترکیبی از تعداد کمی بردار پایه کدگذاری کند به طوریکه تعداد ضرایب غیرصفر در بردار ضرایب حاصل، کم باشد. معمول‌ترین نوع ترکیب مورد استفاده، ترکیب خطی است. طبیعتاً در مواقعی که ما بردارهای پایه مناسب فضا را از پیش نمی‌شناسیم مجبوریم با استفاده از یافتن الگوی داده‌های آموزش، سعی کنیم بردارهای پایه‌ای را برای فضا تخمین بزنیم. هر کدام از این دسته بردارهای کاندیدا، هنگام توصیف داده های آموزشی به صورت کدهای تنک، مقداری هم خطای بازسازی دارند که عبارتست از اختلاف مقدار اصلی داده با داده ای که از ضرب بردار ضرایب خروجی در بردار پایه‌ها به دست می‌آید. طبیعتاً از بین کاندیداها، مجموعه‌ای که دارای کمترین خطای بازسازی باشد بهترین گزینه برای مجموعه بردار پایه است. پس باید با یک الگوریتم بهینه سازی، کاندیدای بهینه را پیدا کنیم. روش‌های متنوعی برای کدگذاری خطی معرفی شده است که در تعریف و یافتن الگوی پنهان داده‌ها و نیز در تعریف عبارات بهینه‌سازی با هم تفاوت دارند اما اساس کار همه، حداقل کردن خطای بازسازی است.

۱.۴.۴ کدگذاری تنک سریع

اولین روشی که انتخاب کردیم برگرفته از مقاله [۳۲] است که در کنفرانس NIPS ۲۰۰۶ به چاپ رسیده. این روش مسئله کدگذاری را که یک مسئله بهینه‌سازی پیچیده و زمانگیر است و راه‌حل تحلیلی هم ندارد به صورت دو مسئله بهینه‌سازی محدب درآورده و آنها را با سرعت و کارایی بالا به صورت تخمینی حل کرده است. مسئله اول، کمینه مربعات نرم L۱ هموارشده[۵] است که توسط الگوریتم جستجوی feature sign حل شده. مسئله بعدی هم شبیه همین بهینه‌سازی ولی برای نرم L۲ و به صورت محدودیت‌دار است، نه هموار، که توسط دوگان لاگرانژ حل شده است. مزیت عمده این روش، که به صورت تکراری کار می‌کند سرعت بالا و گیر نیفتادن در بهینه‌های محلی است. با توجه به سرعت بالای این الگوریتم، می‌توانیم حجم داده آموزشی زیادی را هم به آن بدهیم و کدگذاری دقیق‌تری داشته باشیم. ما این الگوریتم را روی تمام گروه داده‌های آموزش اعمال کردیم و کدهای استخراج شده را به همراه بردار حالت مربوطش به عنوان مشخصه ورودی به الگوریتم رگرسیون دادیم تا تابع نگاشت یاد گرفته شود. نتایج در برخی فعالیت‌ها نسبت به حالت پایه بهبود اندکی داشت ولی در اکثر گروه فعالیت‌ها نتایج، یکسان یا ضعیف‌تر بود. مهمترین دلیلی که برای این نتیجه می‌توان یافت، عدم دخیل کردن ساختار رویه‌ای داده‌ها در کدگذاری است، بدین معنی که کار کردن با فاصله اقلیدسی در بهینه‌سازی، برای مسائلی که ویژگی‌ها ساختار رویه‌ای دارند - از جمله تخمین حالت انسان- مناسب نیست و باعث گمراهی تابع شباهت می‌شود.

۲.۴.۴ کدگذاری خطی محلی

روش دیگری که از آن برای کدگذاری تنک استفاده کرده‌ایم برگرفته از مقاله [۲۹] است که در کنفرانس NIPS ۲۰۱۰ ارائه شده. در ادامه، جزئیات این روش ر ا توضیح می‌دهیم. این الگوریتم، کدگذاری خطی محلی[۶] یا به اختصار LLC نام دارد. LLC نشان می‌دهد که در برخی مسائل که پارامترها روی یک رویه با بعد کم قرار می‌گیرند، مبتنی بر محلی بودن کدگذاری، مهمتر از تنک بودن آن است. چون محلی بودن، به تنک بودن منجر می‌شود ولی لزوماً برعکس آن صحیح نیست. بنابراین LLC بجای تنک بودن، روی محلی بودن پارامترها محدودیت می‌گذارد و عبارت بهینه سازی را به صورت زیر درمی‌آورد:

$$min_C \sum_{i=1}^{N}(||x_i - Bc_i||^2 + \lambda||d_i \odot c_i||^2).s.t.1^\top c_i = 1, \forall i \qquad (۲.۴)$$

[۵] L1-regularized Least Squares

[۶] Locality-constrained Linear Coding

که در آن $\odot$ بیانگر ضرب مولفه‌ای است و d_i اعمال کننده شرط محلیت است بدین صورت که به هر بردار پایه، به نسبت میزان شباهتش به ورودی x_i، آزادی حرکت برای دور شدن از محلیت را می‌دهد. این متغیر اینگونه تعریف می‌شود:

$$d_i = \exp(\frac{dist(x_i, B}{\sigma}) \qquad (۳.۴)$$

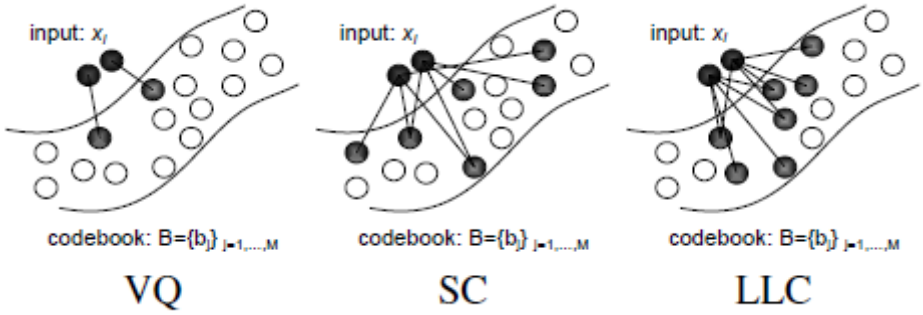
که در آن $dist(x_i, B) = [dist(x_i, b_۱), dist(x_i, b_۲), ..., dist(x_i, b_M)]^T$ و $dist(x_i, b_j)$ هم فاصله اقلیدسی بین x_i و b_j است. پارامتر σ هم سرعت کاهش وزن این شرط را تنظیم می‌کند. معمولاً با کم کردن ماکزیمم $dist(x_i, B)$ها از همه، عبارت d_i را نرمال می‌کنیم به صورتیکه بین صفر و یک قرار بگیرد. محدودیت $۱^\top c_i = ۱$ هم باعث می‌شود کد حاصل از LLC، نسبت به جابجایی مقاوم باشد. اگرچه کد حاصل از LLC در تعریف نرم صفر، تنک نیست (یعنی تعداد ضرایب غیرصفر آن کم نیست) ولی تعداد ضرایب بزرگ آن کم است و به راحتی می‌توان با مشخص کردن یک حد آستانه، تعداد زیادی از ضرایب را صفر کرد.

ویژگی‌های LLC

برای اینکه الگوریتم دسته‌بندی، کارایی خوبی داشته باشد باید برای ورودی‌های یکسان، کدهای بسیار شبیه به هم تولید کند. اگرچه این یک معیار پذیرفته شده است ولی بسیاری از روش‌ها، معیارهای دیگر را فدای این می‌کنند. در فرمول‌بندی LLC، جمله $||d_i \odot c_i||^۲$ وظیفه هموارسازی محلی را بر عهده دارد که ویژگی‌هایی را به کدگذاری ما اضافه می‌کند:

بازسازی بهتر در کوانیزاسیون برداری، هر بردار فقط توسط یکی از بردارهای پایه تخمین زده میشود، در نتیجه بازسازی داده‌ها خطای زیادی دارد و برای جبران آن معمولاً هسته‌های غیرخطی به کار برده می‌شود. همچنین در چنین حالتی، از ارتباط بین پایه‌ها هم کاملاً غفلت شده است در حالیکه در روش LLC، ما هر بردار را با ترکیبی از چندین بردار پایه نزدیک آن می‌سازیم. در نتیجه، بازسازی بهتری از داده‌ها داریم و از ارتباط بین پایه‌ها هم برای بهبود این بازسازی و کم کردن تعداد پایه‌ها بهره برده‌ایم.

تنک بودن هموار محلی در روش صرفاً کدگذاری تنک، هر بردار وروردی توسط چند بردار پایه که بازسازی بهتری برای آن داشته باشند ساخته می‌شود. مشکل این روش اینست که ممکن است کدهای خروجی منتسب به ورودی‌های نزدیک به هم، از هم دور باشند، یعنی پایه‌های انتخاب شده برای داده‌های نزدیک به هم، لزوماً نزدیک نیستند چراکه صرفاً به تنک بودن و خطای بازسازی کمتر



شکل ۲.۴ : تعیین پایه‌ها در روش‌های (از چپ به راست) کوانتیزه کردن برداری، کدگذاری تنک و محلی تنک. (تصویر از [۲۹])

توجه شده است و فاصله پایه از ورودی تأثیری ندارد. اما در روش LLC، موضوع محلیت پایه‌ها هم به عنوان یک معیار مهم در تعیین ضرایب دخیل شده است. در اینجا هم هر بردار ورودی توسط ترکیبی از چند پایه ساخته می‌شود ولی این پایه‌ها باید حتماً به اندازه کافی نزدیک به ورودی باشند تا تأثیرشان در ضرایب، در مقایسه با جریمه فاصله‌شان، قابل توجه شود. در نتیجه، در روش LLC، به داده‌های نزدیک، ضرایب شبیه به هم داده می‌شود و این همان چیزی است که با شهود ما از کدگذاری منصفانه هم سازگاری دارد.

راه‌حل تحلیلی روش کدگذاری تنک بار محاسباتی بالایی برای حل الگوریتم بهینه‌سازی دارد، چون راه حل تحلیلی برای آن وجود ندارد و با جستجوی فضای ورودی باید بهترین مجموعه پایه را پیدا کرد. ولی روش LLC، دارای راه‌حل تحلیلی به صورت زیر است:

$$\tilde{c}_i = (C_i + \lambda diag(d))\backslash 1 \qquad (۴.۴)$$

$$c_i = \tilde{c}_i / 1^\top \tilde{c}_i \qquad (۵.۴)$$

که در آن، $C_i = (B - 1x_i^\top)(B - 1x_i^\top)^\top$ کواریانس ماتریس داده‌هاست. این راه حل، حتی می‌تواند به صورت تقریبی تخمین زده شود و در نتیجه در عمل بسیار سریع است. در کل، زمان اجرای الگوریتم آموزش و تست در LLC، به نسبت یک چهارم از روش HoSC کمتر است و این موضوع برای تخمین حالت در ویدئو که هدف، رسیدن به نتایج نزدیک به لحظه‌ای است بسیار حائز اهمیت است.

در تصویر ۲.۴ نحوه تعیین کد در سه روش کدگذاری و مزایای روش LLC دیده می‌شود.

۵.۴ نتیجه گیری

در این فصل ابتدا با توجه به اهمیت سیاه‌نما به عنوان اصلی‌ترین مشخصه پایه که در این کاربرد استفاده شده است، روش‌های استخراج و نمایش سیاه‌نما را شرح دادیم. سپس سعی کردیم تعدادی از مشخصه‌های استفاده شده برای تخمین حالت که از سیاه‌نما به عنوان عنصر پایه استفاده می‌کنند را مقایسه کنیم. این مشخصه‌ها عبارتند از: هیستوگرام زمینه شکل، Distance Transform، پواسون، Hu Moments و Contour Signature. سپس با توجه به مشکلات این روش‌ها و شهودی که از روش‌های کدگذاری تنک و ذات داده‌های مسأله تخمین حالت داشتیم استفاده از روش‌های کدگذاری تنک را برای این کاربرد، مفید دانستیم. لازم به ذکر است که استفاده از کدگذاری تنک در تخمین حالت، تاکنون انجام نشده بود و این پژوهش آغاز این مسیر است. سپس دو نمونه از روش‌های کدگذاری تنک شامل کدگذاری تنک سریع و کدگذاری خطی محلی را شرح دادیم و نحوه استفاده از آنها در این کاربرد را بررسی کردیم.

فصل ۵

نتایج تجربی

برای بررسی کارایی ویژگی‌ای که معرفی کردیم، آزمایش‌هایی را بر مبنای داده‌های سنتز شده توسط نرم افزار Poser طراحی و اجرا کرده‌ایم. دلایل استفاده از داده‌های مصنوعی را در انتهای فصل اول شرح داده‌ایم که یکی از آنها، سخت بودن تولید داده دقیق و طبیعی برای تخمین حالت است. مهمترین مزیت این داده‌ها، بدون خطا بودن حقیقت زمینه و دقت سیاه‌نمای استخراج شده است. چرا که این دو کار را خود نرم افزار Poser برای ما انجام می‌دهد. مدل خروجی ای که ما استفاده کرده‌ایم، ۵۴ بعد دارد که حاصل از ۳ درجه آزادی برای ۱۸ نقطه بدن است.

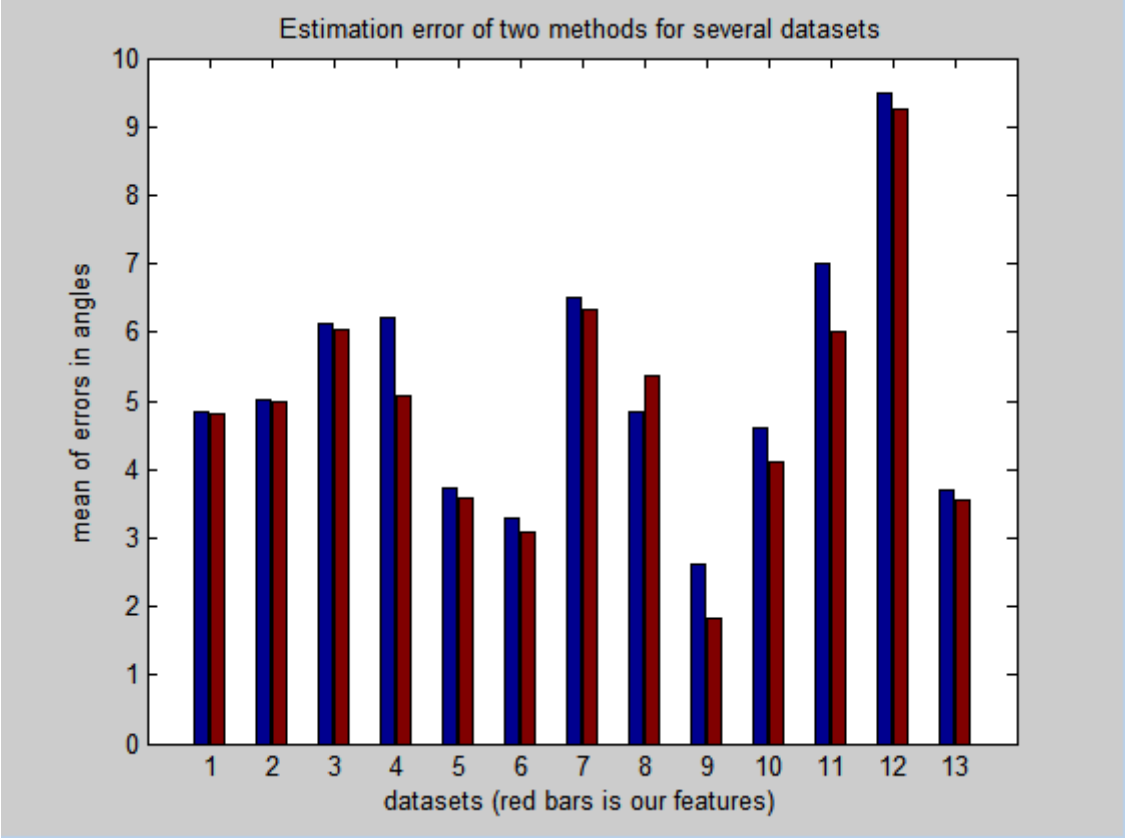
الگوریتم روش‌های HoSC، LLC TGP و Lap روی داده‌های آموزش و تست اجرا شده‌اند و نتایج زیر به دست آمده است. لازم به ذکر است که تعداد داده آزمون در هر مجموعه داده ما متفاوت است ولی به طور میانگین، هر مجموعه ۶۰۰ فریم دارد. میانگین این نتایج هم در جدول ۱.۵ آمده است.

	LLC	Lap	TGP	HoSC
Average error in RMS	۴.۸	۵.۴	۵.۶	۵.۳

جدول ۱.۵ : نتایج عددی حاصل از اجرای الگوریتم‌ها روی مجموعه داده

همانطور که میبینیم نتایج نسبت به خط پایه[۱] بهبود مناسبی دارد که به صورت میانگین، ۸ درصد است. این نتایج به ازای تغییرات پارامتر الگوریتم خوشه‌بندی kmeans هم بررسی شده‌اند و نتایج به ازای مقادیر بزرگتر این پارامتر اندکی بهتر است ولی تفاوت، چشمگیر نیست. برای کسب این نتایج، ما این پارامتر را ۱۰ در نظر گرفتیم.

[۱] baseline



شکل ۱.۵: مقایسه نتایج خطای تخمین حالت در روش های HoSC (آبی-سمت چپ) و LLC (قرمز-سمت راست)

از آنجا که مقایسه اصلی این پژوهش با روش HoSC صورت گرفته، میانگین خطای خروجی در اجراهای متوالی این دو الگوریتم که روی تک تک مجموعه‌های داده انجام شده در تصویر ۱.۵ نشان داده شده است. محور عمودی میزان خطا در فرمول مربع اختلاف‌هاست و محور افقی، نتایج دو الگوریتم روی هر مجموعه داده را نشان می‌دهد. هر مجموعه داده مربوط به یک فعالیت خاص است و با توجه به آن می‌توان به صورت تقریبی نتیجه گرفت که این روش در مورد کدام فعالیت و کاربرد موفقیت بیشتری خواهد داشت.

۱.۵ نتایج بصری

برای مشاهده مدل انسان تخمین زده شده، باید از نرم افزار Poser استفاده کنیم. این نرم افزار با گرفتن فایل مخصوصی که از روی زوایای استخراج شده و به کمک فایل اولیه مدل ساخته می‌شود، بدن انسان را سنتز کرده و به صورت گرافیکی نمایش می‌دهد. در تصاویر ۲.۵ و ۳.۵ نمونه‌هایی از این مدل‌ها را مشاهده می‌کنیم که در هر ستون، سطر اول، مدل حاصل از اجرای الگوریتم LLC و

سطر دوم، داده حقیقت‌زمینه است.

همانطور که مشاهده می‌شود، خروجی الگوریتم پیشنهادی، بسیار به برچسب واقعی نزدیک است.

۲.۵ بحث و تحلیل

برای تحلیل این بهبود لازم است به فلسفه کدگذاری تنک و یادگیری دیکشنری دقت کنیم. در روش HoSC، برای ساختن بردار ویژگی‌ها از مفهومی به نام سبد کلمات[۲] استفاده شده. بدین معنی که فضای ورودی را خوشه‌بندی کرده و هر خوشه را به مثابه یک کلمه در نظر می‌گیرد. ورودی‌های جدید مثل سبدی هستند که تعدادی از این مجموعه کلمات را با ضریبی در خود دارند مثلاً ۰.۲ تا از لغت اول، ۰.۰۳ تا ار لغت دوم و همینطور تا آخر. برای تعیین این ضرایب هم از فاصله اقلیدسی هر بردار ورودی با تمام کلمات استفاده می‌شود و فقط فاصله چند نزدیک‌ترین کلمه به عنوان ضرایب در نظر گرفته می‌شود. پس هر بردار زمینه شکل،می‌تواند به صورت هیستوگرامی که از کلمات نمایش داده شود و در نهایت هم مجموع این هیستوگرام‌ها، بردار ویژگی نهایی را برای یک سیاه‌نما تشکیل می‌دهد.

اما در روش LLC، ما به جای استفاده تنها از فاصله اقلیدسی (تحت مفهوم هیستوگرام) که در یک شعاع ثابت اطراف کلمه دیکشنری، ثابت است، سعی می‌کنیم ترکیبی از کلمات که خطای بازسازی بردار ورودی را به صورت محلی کمینه می‌کند به عنوان ضرایب برگردانیم و این مهمترین دلیل بهبود نتایج در این روش است. چراکه جهت (محل قرارگیری بردار ویژگی نسبت به کلمه دیکشنری) هم در ضرایب حاصل شده تأثیر دارد. در اینجا هم مانند HoSC، به محلی بودن پایه‌های انتخاب‌شده اهمیت زیادی داده می‌شود بدین صورت که هر بردار ورودی فقط می‌تواند با ترکیب چند کلمه نزدیک خود بازسازی شود. این موضوع علاوه بر تأمین تنک بودن بردار خروجی باعث می‌شود بردارهای دور از ورودی که نمی‌توانند تخمین‌زننده خوبی برای ورودی باشند اثر کمتری در ضرایب خروجی داشته باشند. دلیل عمده این موضوع هم به ساختار رویه‌ای فضای ورودی برمی‌گردد. همانطور که قبلاً در فصل سوم ذکر شد، در تخمین حالت انسان، فضای ورودی را که در نگاه اول، دارای ابعاد زیاد است در واقع می‌توان با رویه‌ای هموار با ابعاد بسیار کمتر مدل کرد به صورتیکه با حرکت روی این رویه، تغییر هموار حالت بدن را داشته باشیم. این خاصیت در بسیاری از فعالیت‌های بدن انسان مشاهده شده و استفاده از آن به عنوان یک فرض اولیه برای تخمین حالت، منطقی به نظر می‌رسد. بنابراین در این پژوهش هم ما این فرض را در فرآیند تولید ویژگی دخیل کردیم و یکی از دلایل انتخاب LLC، به

[۲] bag of words

شکل ۲.۵: مدل خروجی الگوریتم کدگذاری تنک محلی در مقایسه با حقیقت زمینه (از بالا به پایین)

شکل ۳.۵: مدل خروجی الگوریتم کدگذاری تنک محلی در مقایسه با حقیقت زمینه (از بالا به پایین)

عنوان روش کدگذاری، همین بوده است.

اخیراً نیز در بسیاری کارها، نشان داده شده که روش‌های کدگذاری و یادگیری دیکشنری در مقایسه با سبد کلمات، ویژگی‌های بهتری به دست می‌دهند. یکی دیگر از دلایل این بهبودها، اینست که در سبد کلمات، ما کلمات را با توجه به همه ورودی‌ها یاد نمی‌گیریم بلکه به صورت محلی سعی در یافتن یک کلمه خوب برای مناطق مختلف داریم و از سایر کلمات هم در تولید کلمات بهتر بهره‌ای نمی‌بریم. در حالیکه در یادگیری دیکشنری، ما سعی داریم بهترین مجموعه لغاتی را پیدا کنیم که خطای بازسازی ورودی‌ها با آنها کمینه و بردار ضرایب نیز تا حد قابل قبولی تنک باشد. همین امر سبب می‌شود تا کلمات عمومی‌تر و ویژگی‌های تنک بهتری به دست آید. البته در LLC، ما از یادگیری دیکشنری استفاده نکردیم و از همان مراکز خوشه‌های Kmeans به عنوان کلمات دیکشنری بهره بردیم. دلیل این کار هم تأثیر ناچیز یادگیری دیکشنری در بهبود نتایج و نیز زمانبر بودن بهینه‌سازی آن در این کاربرد خاص بود.

فصل ۶

نتیجه‌گیری

در این پژوهش مروری بر روش‌های تخمین حالت و مشخصه‌هایی که در این روش‌ها به کار گرفته شده‌اند داشتیم و مشخصه‌ای جدید رو معرفی کردیم. ابتدا رویکردهایی که نسبت به مسأله تخمین حالت وجودد دارند را به دو دسته کلی مبتنی بر مدل و مبتنی بر یادگیری تقسیم کردیم. در فصل دوم، روش‌های مبتنی بر مدل را بررسی کردیم و مثال‌هایی از این روش‌ها را مختصراً شرح دادیم. سپس، روش‌های مبتنی بر یادگیری را در فصل بعدی به دو زیردسته کلی نظارتی و نیمه‌نظارتی تقسیم کردیم. از روش‌های نیمه نظارتی که از داده‌های بدون برچسب هم استفاده می‌کنند، دو روش فرآیندهای گواسی دوقلو و روش رویه‌ی فعالیت که از ساختار رویه‌ای داده‌ها و عملگر لاپلاسین بهره می‌برد، را معرفی کردیم. از روش‌های نظارتی هم دو روش رگرسور Ridge و ماشین بردار وابسته را بررسی کردیم. در فصل چهارم دیدیم که طراحی مشخصه خوب هم به اندازه الگوریتم خوب برای حصول نتایج بهتر مهم است. بنابراین مشخصه پایه سیاه‌نما را به عنوان اصلی‌ترین مشخصه استفاده شده برای تخمین حالت معرفی کردیم و نحوه استخراج آن را شرح دادیم. برای نمایش این مشخصه هم تعدادی روش را که به مشخصه‌های سطح بالاتر و کاراتر منجر می‌شوند معرفی کردیم که عبارتند از هیستوگرام زمینه شکل، Distance Transform و Contour Signature. اشکالات رایج این روش‌ها عبارتند از: عدم نگاه محلی به اجزای سیاه‌نما، پایداری نسبت به دوران و عدم پایداری نسبت به انتقال یا تغییر اندازه. بنابراین برای رفع مشکل پایداری نسبت به دوران، با استفاده از کدگذاری تنک، جهت را هم در بردارهای هیستوگرام‌های دخیل کردیم. از بین روش‌های کدگذاری تنک و یادگیری دیکشنری، دو روش کدگذاری تنک سریع و کدگذاری تنک محلی را معرفی کردیم. در روش تنک سریع، دو مسأله بهینه‌سازی محدب به صورت سریع و افزایشی با دقت دلخواه حل می‌شوند و برای کارکردهای بلادرنگ مناسب‌تر است در حالیکه روش کدگذاری محلی، از ساختار رویه‌ای داده‌ها برای ایجاد کدگذاری بهتر

بهره می‌برد. در این روش از مراکز الگوریتم خوشه‌بندی kmeans به عنوان دیکشنری استفاده شده و هر بردار زمینه شکل تصویر به صورت ترکیبی خطی از چندهمسایه نزدیک آن نمایش می‌شود. سپس از مجموع این بردارها، بردار نهایی حاصل می‌گردد که نسبت به دوران پایدار نیست و نسبت به انتقال یا تغییر اندازه پایدار است. انتخاب چند همسایه نزدیک باعث می‌شود که به جای فاصله اقلیدسی، فاصله رویه‌ای دخیل شود و تخمین مناسب‌تری از میزان شباهت دو بردار در دست باشد. شهودهایی برای بهبود نتایج وجود داشت ولی برای مقایسه کمّی، آزمایش‌هایی را برای بررسی کارایی الگوریتم اجرا کردیم و نتایج نشان از بهبود مناسب روش پیشنهادی داشت به طوریکه در میانگین، بهبود ۱ درجه‌ای را شاهد بودیم. در انتها هم تحلیل مختصر از نتایج سایر الگوریتم‌های پیاده‌سازی شده ارائه کردیم و برتری روش LLC را توضیح دادیم.

کتاب‌نامه

[1] J. Shotton, A. Fitzgibbon, M. Cook, A. Blake, ***Real-time Human Pose Recognition in Parts from Single Depth Images***, CVPR, 2011.

[2] Hen, Y. and Paramesran, R., ***Single camera 3d human pose estimation: A review of current techniques,*** In Technical Postgraduates (TECHPOS), 2009 International Conference for, pp.1–8, IEEE.

[3] Sminchisescu, C., ***3d human motion analysis in monocular video techniques and challenges,*** In IEEE International Conference on Advanced Video and Signal based Surveillance, 2006.

[4] C. Stauffer and W. Grimson, ***Adaptive Background Mixture Models for Real-time Tracking***, in Proc IEEE Conf. Computer Vision and Pattern Recognition, 1999, pp. 23-25.

[5] R. Poppe, ***Vision-based human motion analysis: An overview***, Computer Vision and Image Understanding(CVIU), vol. 108, pp. 4-18, 2007.

[6] S. S. Beauchemin and J. L. Barron, ***The Computation of Optical Flow,*** ACM Computing Survey, vol. 27(3), pp. 433-466, 1995.

[7] C. Sminchisescu, A. Kanaujia and D. N. Metaxas, ***BMA^{3}E : Discriminative Density Propagation for Visual Tracking,*** IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI), vol. 29(11), pp. 2030-2044, 2007.

[8] D. Ramanan, D. A. Forsyth, and A. Zisserman, ***Tracking People by Learning Their Appearance,*** IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI), vol. 29, pp. 65-81, 2007.

[9] Leonid Sigal, ***Human pose estimation,*** Disney Research, Pittsburgh.

[10] `http://en.wikipedia.org/wiki/Image_gradient`

[11] Agarwal, A., Triggs, B., I., A. R., and Montbonnot, F., ***Recovering 3d human pose from monocular images,*** IEEE transactions on pattern analysis and machine intelligence, Vol.28, No.1, pp.44–58, 2006.

[12] Lee, M. and Nevatia, R., ***Human pose tracking in monocular sequence using multilevel structured models,*** IEEE transactions on pattern analysis and machine intelligence, Vol.31, No.1, pp.27–38, 2009.

[13] Moeslund, T. and Granum, E.,***A survey of computer vision-based human motion capture,*** Computer Vision and Image Understanding, Vol.81, No.3, pp.231–268, 2001.

[14] Lee, M. and Cohen, I., ***A model-based approach for estimating human 3d poses in static images,*** IEEE transactions on pattern analysis and machine intelligence, Vol.28, No.6, pp.905–916, 2006.

[15] Ling, H. and Jacobs, D., ***Shape classification using the inner-distance,*** IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.286–299, 2007.

[16] Guocheng, A., Fengjun, Z., Hongan, W., and Guozhong, D., ***Shape filling rate for silhouette representation and recognition,*** In 2010 International Conference on Pattern Recognition, pp.507–510, IEEE, 2010.

[17] S. Belongie, J. Malik, and J. Puzicha. ***Shape Matching and Object Recognition using Shape Contexts***. IEEE Trans. Pattern Analysis & Machine Intelligence, 24(4):509–522, 2002.

[18] M. Tipping. ***The Relevance Vector Machine***. In Neural Information Processing Systems, 2000.

[19] M. Tipping. ***Sparse Bayesian Learning and the Relevance Vector Machine***. J. Machine Learning Research, 1:211–244, 2001.

[20] D. Lowe. ***Object Recognition from Local Scale-invariant Features***. In Int. Conf. Computer Vision, pages 1150–1157, 1999.

[21] Kanaujia, A., Sminchisescu, C., and Metaxas, D., ***Semi-supervised hierarchical models for 3d human pose reconstruction,*** In IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[22] Agarwal, A. and Triggs, B., ***Monocular human motion capture with a mixture of regressors,*** In IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[23] Rosales, R. and Sclaroff, S., ***Learning body pose via specialized maps,*** In Neural Information Processing Systems, 2002.

[24] Elgammal, A., ***Inferring 3d body pose from silhouettes using activity manifold learning,*** In IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[25] Navaratnam, R., Fitzgibbon, A. W., and Cipolla, R., ***The joint manifold model for semi-supervised multi-valued regression,*** In IEEE International Conference on Computer Vision, 2007.

[26] Lee, C. and Elgmmal, A., ***Modeling view and posture manifolds for tracking,*** In IEEE International Conference on Computer Vision, 2007.

[27] Stauffer, C. and Grimson, W., ***Adaptive background mixture models for realtime tracking,*** In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., Vol.2, IEEE, 1999.

[28] Ridder, C., Munkelt, O., and Kirchner, H., ***Adaptive background estimation and foreground detection using kalman-filtering,*** In Proceedings of International Conference on recent Advances in Mechatronics, pp.193–199, Citeseer, 1995.

[29] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, ***Locality-constrained Linear Coding for Image Classification,*** NIPS, 2010.

[30] Lubomir Bourdev, Jitendra Malik, ***Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations,*** ICCV 2009

[31] Yang Wang, Duan Tran, Zicheng Liao, David Forsyth, ***Learning Hierarchical Poselets for Human Parsing*** CVPR 2011

[32] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng., ***Efficient sparse coding algorithms*** NIPS 2006

[33] Brand, M., ***Shadow puppetry,*** In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, Vol.2, pp.1237–1244, IEEE, 1999.

[34] Ramanan, D. and Sminchisescu, C., ***Training deformable models for localization,*** In Computer Vision and Pattern Recognition (CVPR). IEEE Conference on, IEEE Computer Society, 2006.

[35] Li, Y., Jia, K., and Zhang, G., ***Semi-supervised human pose estimation piloted by manifold structure,*** In International Conference on Information Engineering and Computer Science, 2009.

[36] Ramanan, D. and Forsyth, D., ***Finding and tracking people from the bottom up,*** In IEEE Conference on Computer Vision and Pattern Recognition, 2003

[37] Bo, L. and Sminchisescu, C., ***Twin gaussian processes for structured prediction,*** International Journal of Computer Vision, Vol.87, No.1, pp.28–52, 2010.

[38] Navaratnam, R., Fitzgibbon, A., and Cipolla, R., ***Semisupervised learning of joint density models for human pose estimation,*** In Proc. BMVC, pp.679–688, Citeseer, 2006.

Abstract

In this research we do a quick survey on Monocular 3D Human Pose Estimation problem. Then we clasify methods and aproches toward this problem and describe some examples from each category. Then we count several features used for human pose estimation which exploit silhouette as base feature. We see pros and cons of thes features. Then for rectifing the rotation invariance problem and unconsideration of manifold nature of data that are problems of HoSC feature, we introduce a new feature which is generated by Locality-constrained Linear Coding as a sparse coding method and Kmeans as a Dictionary Learning method. This feature gets better results in experiments too.

Keywords: ***monocular human pose estimation, 3D reconstruction of human body, 3D, sparse coding, Locality-constrained Linear Coding, relevance vector machin***

Sharif University of Technology
Department of Computer Engineering

Bachelor Thesis

Major in Information Technology Engineering

Topic

Implementation and comparison of 3D human pose estimation methods

By
Amrollah Seifoddini Banadkooki

Supervisor
Prof. Hamid Reza Rabiee

May 2012