بازشناسی فعالیت انسان با استفاده از مدل تعویضی ساختاری

محمدمهدی ارزانی، محمود فتحی و احمد اکبری ازیرانی

چکیده: بازشناسی خودکار فعالیت‌های انسان، بخشی جدایی‌ناپذیر از هر برنامه تعاملی با انسان است. یکی از چالش‌های عمده برای شناخت فعالیت، تنوع در نحوه فعالیت افراد است. همچنین بعضی از فعالیت‌ها ساده، سریع و کوتاه هستند، در حالی که بسیاری دیگر پیچیده و دارای جزئیات هستند و در مدت زمان طولانی انجام می‌شوند. در این مقاله، ما از داده‌های اسکلت که از توالی تصاویر RGB-D استخراج می‌شوند استفاده می‌کنیم. ما مدل گرافی را پیشنهاد می‌دهیم که قادر است فعالیت‌های پیچیده و ساده را بازشناسی کند. برای بهینه‌سازی پارامترهای مدل گرافی احتمالی از روش پیش‌بینی ساختاری توزیع‌شده استفاده می‌کنیم. این روش در سه مجموعه داده به طور گسترده مورد آزمایش (CAD-۶۰، UT-Kinect و Florence ۳D) قرار می‌گیرد که هر دو نوع فعالیت را پوشش می‌دهند. نتایج نشان می‌دهد که روش ما می‌تواند هر دو نوع فعالیت ساده و پیچیده را به طور مؤثر تشخیص دهد، در حالی که اکثر آثار قبلی تنها بر یکی از این دو نوع تمرکز می‌کنند. همچنین ما نشان می‌دهیم استفاده از روش‌های خوشه‌بندی برای مقداردهی اولیه تأثیر زیادی در نتایج دارد.

کلیدواژه: **مدل‌های گرافی احتمالی، بازشناسی فعالیت انسان، پیش‌بینی ساختاریافته توزیع‌شده، اسکلت.**

۱- مقدمه

یکی از وظایف اساسی برای روبات‌ها این است که جهان اطراف را درک کنند تا قادر به برقراری ارتباط با انسان‌ها و محیط اطراف باشند. بنابراین یکی از بخش‌های جدایی‌ناپذیر از برنامه‌های کاربردی روباتیک تعاملی، بازشناسی فعالیت‌های انسان است.

به منظور بازشناسی فعالیت انسان از ویژگی‌های مختلفی استفاده میشود که عبارت هستند از توالی تصاویر RGB و عمق. داده‌های اسکلت که از تصاویر عمق و RGB استخراج می‌شوند نیز به منظور بازشناسی فعالیت استفاده شده‌اند. اطلاعات اسکلت در برابر تغییرات روشنایی و زاویه دید مقاوم‌تر هستند. با ظهور حسگرهای عمق ارزان مانند مایکروسافت کینکت[1]، استفاده از اطلاعات عمق به طور فزاینده‌ای فراگیر شده است. بازشناسی فعالیت انسانی را می‌توان به عنوان یک مسئله برچسب‌گذاری توالی دید. مدل‌های گرافی احتمالی یک وسیله پیش‌بینی ساختاریافته هستند که می‌توانند به طور مؤثر روابط پیچیده و تودرتو بین متغیرهای تصادفی را استخراج کنند. گراف‌ها اشیای دارای ساختار هستند و مدل گرافی احتمالی، گراف‌هایی هستند که ساختار وابستگی بین متغیرهای تصادفی مختلف را نشان می‌دهند. در بسیاری از کاربردها به منظور سادگی یادگیری و استنتاج، از مدل‌های گرافی احتمالی با ساختار درختی استفاده می‌شود ولی استفاده از مدل‌های گرافی درختی محدودیت‌هایی را بر روابط کدگذاری‌شده در مدل گرافی تحمیل می‌کند. در بازشناسی فعالیت انسانی با استفاده از مدل گرافی احتمالی بر روی توالی‌های زمانی، چنین محدودیت‌هایی (در نظر گرفتن ساختار درختی در مدل گرافی) باعث محدودکردن قابلیت‌های روش می‌شود و از این رو عملکرد مدل را تضعیف می‌کند. بر این اساس، در این کار، ما از چنین فرضی اجتناب کرده و تعدادی زیرگراف حلقه‌ای[2] را به مدل اضافه می‌کنیم. بنابراین مدل ما ساختار درختی نخواهد داشت و مسلماً در مدل‌کردن روابط بین متغیرها قدرتمندتر است. زیر گراف‌های اضافه‌شده به مدل برای مدل‌سازی روابط بین مشاهدات در توالی زمانی طولانی مورد استفاده قرار می‌گیرد. با این حال، یافتن پارامترهای بهینه چنین مدلی به طور کلی چه از نظر محاسباتی و چه از نظر همگرایی پیچیده‌تر است. بنابراین در روش پیشنهادی ما پارامترهای مدل خود را مشابه ماشین‌های بردار پشتیبان ساختاری ناپیدا[3] بر اساس پیش‌بینی ساختار توزیع‌شده[4] [۱] و [۲] بهینه و از استنتاج با تکنیک‌های تقریب استفاده می‌کنیم [۳].

علاوه بر تغییرات فوق در ساختار مدل پیش‌بینی ساختاریافته، ما یک الگوی تعویضی[5] برای مدل خود پیشنهاد می‌کنیم تا عملکرد بازشناسی را بهبود ببخشیم. چنین مدل‌های تعویضی (که به عنوان تغییر رژیم[6] در برخی از کارهای قبلی نامیده می‌شود) به طور موفقیت‌آمیزی برای تحلیل سری‌های زمانی اقتصادی به کار گرفته شده است. روش کار به این صورت است که برای شرایط گوناگون چند مدل مختلف آماده می‌شود. این مدل‌ها از نظر ساختاری یکسان هستند ولی پارامترهای آنها متفاوت است. بخش تعویض رده بر اساس شرایط مسئله بین این مدل‌ها تعویض می‌کند و در هر زمان تصمیم‌گیری بر عهده یکی از مدل‌هایی است که توسط تعویض رده انتخاب شده است [۴].

در تعویض رژیم، متغیرهایی که به منظور تعویض بین مدل‌ها مورد استفاده قرار می‌گیرد، می‌تواند قابل مشاهده باشد (مشابه مدل آستانه[7] [۵]) یا می‌توانند پارامترهای پنهان یا آموخته‌شده باشند که ویژگی مارکف را دارند و اغلب به عنوان مدل تعویضی مارکوف شناخته می‌شوند [۶]. در ابتدا این رویکردهای تعویضی در اقتصاد کاربردی، برای توصیف الگوهای سری‌های زمانی مالی که اغلب با تغییرات ناگهانی روبه‌رو هستند به کار گرفته شدند. بنابراین مفهوم تغییر رژیم معمولاً برای مدل‌سازی شکاف ساختاری به دلیل گسترش و رکود اقتصادی در سری‌های زمانی اقتصادی مورد استفاده قرار می‌گیرد. هنگامی که ما بین رده‌های فعالیت تغییر می‌کنیم برای تمایز از ادبیات اقتصادی، به جای آن از اصطلاح تعویض

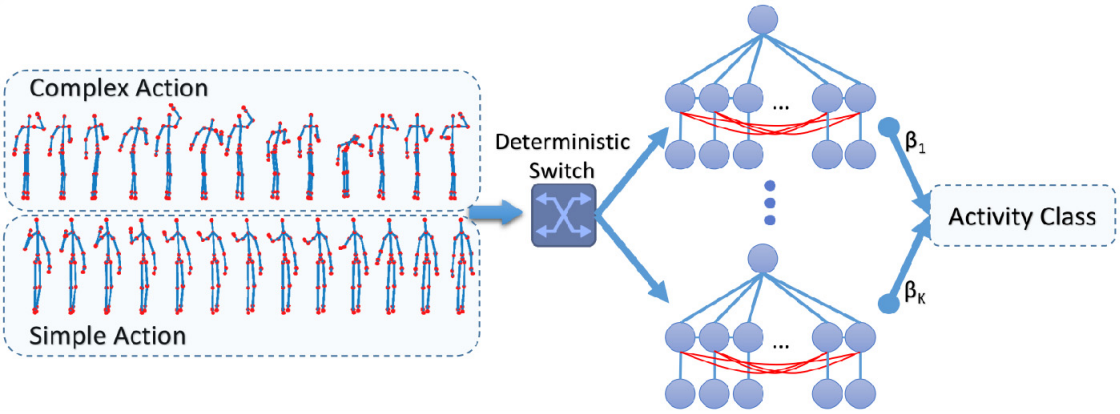
این مقاله در تاریخ ۲۷ اسفند ماه ۱۳۹۷ دریافت و در تاریخ ۲۲ آبان ماه ۱۳۹۸ بازنگری شد.

محمدمهدی ارزانی، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: marzani@iust.ac.ir).

محمود فتحی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: mahfathy@iust.ac.ir).

احمد اکبری ازیرانی، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: akbari@iust.ac.ir).

1. Kinect
2. Loopy Subgraphs
3. Latent Structural Support Vector Machines
4. Distributed Structured Prediction
5. Switching
6. Regime-Switching
7. Threshold Model



شکل ۱: مدل ما می‌تواند فعالیت‌های ساده و پیچیده را بازشناسی کند. توالی اسکلت‌ها به یک مدل تعویض رده که تصمیم می‌گیرد یک مدل گرافی احتمالی را فعال کند، داده می‌شود. زیرمدل‌های گرافی روابط کوتاه و طولانی را در طول زمان (که به ترتیب با لبه‌های آبی و قرمز مشخص می‌شود) مشخص می‌کند.

رده استفاده می‌کنیم.

در این مقاله مدل تعویضی ساختاری[1] ارائه شده است. ما یک مدل گرافی احتمالی را برای شناسایی هر دو نوع فعالیت ساده و پیچیده پیشنهاد می‌کنیم. این مدل نوعی مدل گرافی بدون جهت است که می‌تواند فعالیت‌های ساده و پیچیده را به خوبی مدل کند. ما این کار را با واردکردن روابط کوتاه و بلند بین متغیرهای مدل گرافی انجام دادیم و به جای استفاده از ساختارهای درختی مرسوم یا مدل‌های میدان تصادفی شرطی[2]، روابط و وابستگی‌های اضافی را در مدل در نظر گرفتیم. به منظور بازشناسی فعالیت انسان از داده اسکلت به عنوان ورودی استفاده می‌شود. توالی اسکلت به یک مدل تعویضی رده[3] داده می‌شود که یک مدل گرافی احتمالی را فعال می‌کند. برای تمام فعالیت‌ها، ساختار مدل گرافی احتمالی ثابت است اما پارامترها تغییر می‌کنند.

فعالیت‌ها از طریق مدل‌های ساختاری متشکل از ژست[4]های کلیدی با روابط کوتاه و بلند در طول زمان مدل‌سازی می‌شوند. ما به تعداد K عدد از چنین مدل‌هایی می‌سازیم که هر کدام از آنها رده[5] فعالیت را مشخص می‌کند و فعالیت‌هایی که در هر رده قرار می‌گیرند، ژست و حرکات مشابهی در طول زمان دارند. ما زیر سامانه‌ای را ارائه نمودیم که بین مدل‌ها تعویض می‌کند یعنی در هر زمانی یکی از آن‌ها را انتخاب و به کار می‌گیرد. این زیرسامانه قطعی است به این معنی که از مدل‌های احتمالی استفاده نمی‌کند و بر اساس یک دسته‌بند در هر زمان به صورت قطعی خروجی آن مشخص می‌شود.

یک بردار باینری $\{\beta_i\}_{i=1}^{K}$ مشخص می‌کند که در هر زمان کدام مدل گرافی احتمالی خروجی را تعیین کند. شکل ۱ تصویر کلی روش پیشنهادی را نشان می‌دهد. باید توجه داشت که وقتی ساختار مدل گرافی پیچیده است و انواع مختلف روابط در نظر گرفته می‌شوند، یادگیری و استنتاج یک کار بسیار چالش‌برانگیز خواهد بود. بنابراین مدل با استفاده از یک چارچوب پیش‌بینی ساختاریافته، مشابه با [۱] بهینه‌سازی شده است. با این که یادگیری پارامترهای یک مدل گرافی احتمالی بسیار زمان‌گیر است، مدل پیشنهادی در یک چارچوب پردازش توزیع‌شده پیاده‌سازی می‌شود، زیرا مدل تعویضی ما می‌تواند وظایف را به بخش‌های کوچک‌تر تقسیم کند. چنین پیاده‌سازی می‌تواند فعالیت‌های انسانی را به موقع تشخیص دهد که برای رباتیک مناسب است. همچنین در این مدل ما از متغیرهای پنهان استفاده کردیم. در این کار نشان دادیم مقداردهی اولیه متغیرهای پنهان تأثیر بسزایی در نتایج دارد. برای مقداردهی اولیه ما از روش‌های مختلف خوشه‌بندی استفاده کردیم.

همچنین در این مقاله ما چند نوع موازی‌سازی را انجام دادیم. نوع اول مربوط به استفاده از روش‌های انتشار باور[6] است که بر اساس [۳] به صورت موازی پیاده‌سازی شده است. نوع دوم به علت ارائه روش تعویض رده است و از آنجایی که مسئله را به بخش‌های کوچک‌تر و مستقل افراز می‌کند می‌توان هر کدام را به صورت موازی اجرا کنیم و از آنجایی که هر کدام مسئله کوچک‌تری است در کل سرعت اجرا افزایش خواهد داشت.

به طور خلاصه نوآوری‌های ما در این مقاله عبارت است از ۱) ارائه مدل گرافی حلقه‌ای بدون جهت با ساختار جدید و فرمول‌بندی و ارائه روابط و راهکار یادگیری و استنتاج در آن، ۲) بازشناسی فعالیت‌های ساده و پیچیده با استفاده از مدل گرافی ارائه‌شده، ۳) استفاده از مفهوم تعویض رده برای افزایش دقت و سرعت در بازشناسی فعالیت انسان و ۴) بررسی تأثیر روش‌های خوشه‌بندی مختلف برای مقداردهی اولیه حالات پنهان.

۲- کارهای مرتبط

۲-۱ مدل‌های گرافی احتمالی برای بازشناسی فعالیت

چندین محقق پیشین روش‌های بازشناسی فعالیت انسان را مورد بررسی قرار داده‌اند [۷] و [۸]. در میان رویکردهای موفقیت‌آمیز برای شناسایی فعالیت انسانی، مدل‌های گرافیکی کارایی بالایی نشان داده‌اند و هر دو مدل گرافیکی جهت‌دار [۹] و غیر جهت‌دار [۱۰] تا [۱۲] برای این کار مورد استفاده قرار گرفته‌اند. در حالت عادی مدل‌های بدون جهت توالی فریم‌ها را در نظر نمی‌گیرد چرا که روابط جهتی ندارند ولی با استفاده از بررسی نحوه قرارگیری حالت‌های متغیرها در کنار یکدیگر می‌توانند کلاس مورد نظر را بازشناسی کنند. توالی فریم‌ها زمانی در مدل‌های بدون جهت در نظر گرفته می‌شود که توابع پتانسیل متغیرهای چند فریم را مدل می‌کند و به این ترتیب توالی زیرفعالیت‌ها در نظر گرفته خواهند شد که در بخش ۳-۲-۱ توابع پتانسیل مورد بحث قرار خواهند گرفت. اخیراً LSSVM[7] [۱۳] به طور مؤثر برای یادگیری پارامترهای مدل گرافیکی استفاده شده است. انواع HCRF[8] نیز توسعه یافته‌اند که در

1. Switching Structured Prediction
2. Conditional Random Fields
3. Category-Switching
4. Pose
5. Category
6. Belief Propagation
7. Least-Squares Support-Vector Machines
8. Hidden Conditional Random Fields

آنها روش بیشینه- حاشیه[1] که با نام بردار ماشین پشتیبان پنهان نیز شناخته می‌شود برای بازشناسایی فعالیت‌ها مورد استفاده قرار گرفته شده است [۱۴] و [۱۵].

هو و همکاران [۱۰] با استفاده مدل‌های گرافی برچسب‌های فازی به ویدئو اختصاص دادند که این کار منجر به خطای بخش‌بندی[2] کمتر شد. همچنین در [۹] مدل مارکوف بی‌نظمی بیشینه[3] به کار گرفته شد. برخی دیگر از کارها از مفهوم مورد استفاده شیء[4] استفاده کردند. مورد استفاده به این معنی است که هر شیء در چه فعالیتی به کار گرفته می‌شود. با استفاده از این مفهوم می‌توان در مدل بازشناسی فعالیت، وجود فعالیت‌هایی را که شیء مرتبط در آنها حضور دارد محتمل‌تر دانست. برای نمونه در [۱۱] از میدان تصادفی مارکوف[5] در تصاویر RGB-D استفاده شد، در این کار از SSVM[6] برای یادگیری پارامترهای مدل استفاده شده است. نی و همکاران در [۱۶] تصاویر سطح خاکستری[7] را با تصاویر عمق ادغام[8] کردند تا با استفاده از یک روش چندسطحی فعالیت انسان را بازشناسی کنند. الگوریتم بازشناسی در کار آنها مدل ساختاری ناپیدا[9] می‌باشد. میدان‌های تصادفی شرطی نیز از جمله مدل‌های گرافی هستند که در بازشناسی فعالیت انسانی به صورت موفقیت‌آمیزی مورد استفاده قرار گرفته‌اند. برای مثال در [۱۷] لن و سایر همکاران انواع مختلفی از میدان‌های تصادفی شرطی مختلفی را برای این کار آزمودند. آنها روابط مستقیم بین متغیرهای ورودی و خروجی و همین طور در بین متغیرهای پنهان را مورد بررسی قرار دادند. تا جایی که ما اطلاع داریم تمامی کارهای پیشین از روابط با طول کوتاه یا متوسط بین متغیرها استفاده شده است در صورتی که در بسیاری از کاربردها در نظر گرفتن روابط طولانی می‌تواند نتایج را بهبود ببخشد.

۲-۲ یادگیری منیفلد برای بازشناسی فعالیت

برخی از مقالات با استفاده از روش‌های منیفلد فعالیت انسان را بازشناسی کرده‌اند [۱۸] تا [۲۱]. با این حال باید توجه داشت که فرض منیفلد زمانی معتبر است که فعالیت انجام‌شده توسط انسان کوتاه و سریع باشد. فرض منیفلد به این معنا است که ویژگی‌های استخراج‌شده از داده‌های ورودی بر روی یک منیفلد در فضای چندگانه قرار دارند. برای نمونه با استفاده از منیفلد ریمانی ژانگ و همکاران [۱۸] نتایج خوبی را بر روی مجموعه داده UT-Kinect به دست آوردند که فعالیت‌های موجود در این مجموعه داده کوتاه است. وانگ و سایرین [۱۹] نیز نزدیک‌ترین ویدئو را در فضای منیفلد پیدا کردند و برچسب آن را به ویدئو اختصاص دادند. در [۲۰] نویسندگان منیفلد را مسطح کردند و سپس با استفاده از PCA و SVM ویژگی‌های فضای منیفلد را یاد گرفتند. در [۲۱] نویسندگان در فضای منیفلد ریمانی مفاصل را نمایش دادند و سپس با استفاده از تغییرات اسکلت (مسیرها[10]) فعالیت انسان را بازشناسی کردند.

۲-۳ یادگیری عمیق برای بازشناسی فعالیت

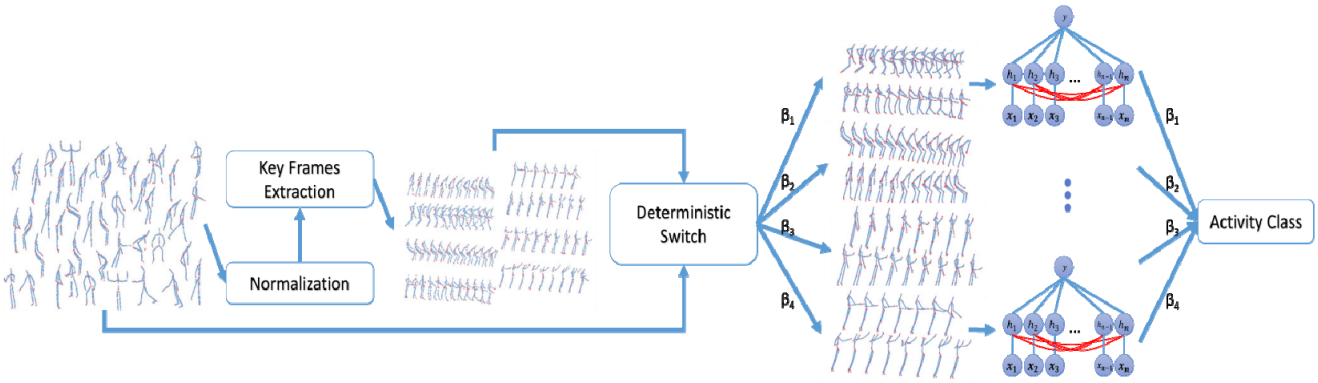
یکی از روش‌هایی که اخیراً در بازشناسی فعالیت انسانی مورد توجه قرار گرفته است، روش‌های یادگیری عمیق می‌باشد ولی در بسیاری از موارد (مانند مسئله ما) که داده کمی موجود است این روش‌ها به خوبی کار نمی‌کنند. لیو و سایرین [۲۲] از LSTM[11] به منظور بازشناسی فعالیت انسان استفاده کردند. در سال‌های اخیر برخی از مقالات مدل‌های گرافی احتمالی را با یادگیری عمیق به منظور بازشناسی فعالیت انسان ترکیب کرده‌اند [۲۳] و [۲۴]. در [۲۵] با استفاده از روشی به نام PRNN[12] که از دو شبکه CNN[13] و RNN[14] تشکیل شده‌اند تا داده‌های اسکلت و عمق را با یکدیگر ادغام کنند، فعالیت‌های انسان را بازشناسی می‌کند. آنها از اسکلت به منظور تخمین پارامترهای شبکه استفاده کردند و سپس شبکه را با استفاده از داده‌های عمق آموزش دادند. روش GCAA-LSTM[15] در [۲۶] از دو شبکه LSTM و یک حافظه سراسری هم‌بافت[16] استفاده کرد. در مرحله اول LSTM اول و حافظه سراسری هم‌بافت تصمیم می‌گیرند که باید از کدام ورودی استفاده کرد و پس از آن LSTM دوم مفصل‌های مهم را در هر فریم یاد می‌گیرد.

۲-۴ بازشناسی فعالیت با استفاده از داده‌های RGB-D

داده‌های به دست آمده از کینکت به طور گسترده‌ای برای برنامه‌های کاربردی بازشناسی فعالیت استفاده شده است. در این بخش، ما این روش‌ها را بررسی می‌کنیم، به ویژه آنهایی که مستقیماً بر روی سه مجموعه داده مورد استفاده در این مقاله، (UT-Kinect، CAD -۶۰ و Florence ۳D) استفاده شده‌اند. برخی از این آثار تنها از داده‌های اسکلت استفاده می‌کنند [۲۷] و [۲۸] در حالی که برخی دیگر از داده عمق [۲۹] و [۳۰] استفاده کردند و برخی از ترکیبات عمق، RGB و اطلاعات اسکلت [۳۱].

ومولاپالی و سایر همکاران [۳۲] فاصله‌های هندسی مختلفی را بین قسمت‌های مختلف اسکلت استفاده کردند. خلاصه‌سازی ویدئو به فریم‌های کلیدی هنگامی که با فعالیت‌های پیچیده سروکار داریم بسیار مهم است. در [۳۳] نویسندگان بر اساس انرژی اسکلت ویدئو را به بخش‌های استاتیک و پویا رده‌بندی کردند. آنها دو نوع فریم کلیدی را از بخش‌های پویا و ایستا استخراج کردند و بازشناسی را بر اساس آنها انجام دادند. کدینگ تنک نگهدارنده[17] ترتیب در [۳۴] برای دسته‌بندی فعالیت‌های دیتای RGB-D معرفی شد. پاریسی و سایرین [۳۵] فعالیت انسان را با استفاده از روش‌هایی که از علوم اعصاب‌شناسی ایده گرفته شده بازشناسی کردند که بر روی ژست‌ها و حرکات تمرکز می‌کند. آنها از یک شبکه SO-GWR[18] استفاده کردند. در مقاله دیگر [۳۶] از روش‌های کرنل غیر خطی برای به دست آوردن روابط بین مفاصل در فضاهای مرتبه بالاتر استفاده کردند. در [۳۷] از DBMM[19] برای ترکیب چند دسته‌بند استفاده شد.

1. Max-Margin
2. Segmentation
3. Maximum Entropy Markov Model
4. Object Affordance
5. Markov Random Field
6. Smooth Support Vector Machine
7. Grayscale
8. Fusion
9. Latent Structural Model
10. Trajectory
11. Long Short-Term Memory
12. Privileged Information Based Recurrent Neural Network
13. Convolutional Neural Networks
14. Recurrent Neural Networks
15. Global Context-Aware Attention LSTM
16. Global Context Memory
17. Order-Preserving Sparse Coding
18. Self-Organizing Growing When Required
19. Dynamic Bayesian Mixture Model



شکل ۲: نمای کلی روش پیشنهادی (توضیحات تکمیلی در مورد این شکل در متن آماده است).

۳- روش پیشنهادی

یک طرح کلی از روش ما در شکل ۲ و در جزئیات این بخش توضیح داده شده است. روش ما در اولین گام، اسکلت را به مقیاس‌ها و مشخصه‌های یکنواخت تبدیل می‌کند. در مرحله بعد از یک استخراج‌کننده فریم کلیدی برای انتخاب مناسب و مفیدترین فریم‌های کلیدی برای ساخت مدل گرافی استفاده می‌شود. سومین مرحله، یک تعویض قطعی (پیاده‌سازی شده توسط دسته‌بندی‌کننده‌های سطحی[1]) است که بین مدل‌های مدل گرافی احتمالی بر اساس پیچیدگی فعالیت مشاهده‌شده تغییر می‌کند. برخی از ویژگی‌های داده‌ها (برای مثال، جهت اسکلت) برای گام تعویضی از اسکلت نرمال‌نشده (داده‌های اصلی) گرفته می‌شوند. مرحله نهایی، فعالیت را از طریق مدل گرافی احتمالی پیشنهادی تشخیص می‌دهد. در ادامه در ابتدا ویژگی‌های مورد استفاده و نحوه استخراج فریم‌های کلیدی بیان گردیده و پس از آن مدل و معادلات مربوط به آن شرح داده خواهد شد. به منظور بیان یک مسئله در مدل‌های گرافی احتمالی بدون جهت باید فرمول کلی بیان شود و سپس توابعی که روابط بین متغیرها را تعیین می‌کنند مشخص گردد. این قسمت بسیار مهم است زیرا تفاوت مدل‌ها معمولاً در همین قسمت است. از طرفی باید روابط به صورتی تعریف شود که به خوبی بتواند کلاس‌ها را از یکدیگر تمیز دهد و از طرف دیگر مدل باید قابل استفاده باشد به این معنا که بتوان پارامترهای آن را بر اساس داده آموزش به دست آورد و همین طور بتوان در زمان آزمایش با سرعت مناسب برچسب داده ورودی را تعیین کرد. همین طور نباید مدل آن قدر پیچیده باشد که از نظر پیچیدگی محاسباتی غیر قابل پردازش باشد. در نهایت پس از تعریف مدل باید نحوه یادگیری پارامترها و استنتاج مدل نیز مشخص شود زیرا اگر نتوان پارامترهای مدل را تعیین کرد مدل قابل استفاده نمی‌باشد. هر مدل شامل تعداد زیادی پارامتر است که بر اساس داده‌های آموزشی به دست آورده می‌شوند. در نهایت در زمان آزمایش، استنتاج انجام می‌شود به این معنی که با داشتن داده‌های ورودی، محتمل‌ترین برچسب به دست می‌آید. البته از آنجایی که ما در یک چارچوب احتمالی کار می‌کنیم، برای هر ورودی احتمال تمامی برچسب‌ها محاسبه خواهد شد و برچسبی که بیشترین احتمال را دارد به عنوان کلاس فعالیت در نظر گرفته می‌شود. ذکر این نکته ضروری است که از آنجایی که ما در مدل خود از حالت‌های پنهان استفاده می‌کنیم عمل استنتاج پیچیده خواهد شد زیرا نه تنها احتمال رخداد تمامی برچسب‌ها باید محاسبه شود بلکه احتمال تمامی حالت‌های پنهان برای تمامی متغیرهای نهان نیز باید محاسبه گردد. در ادامه در بخش ۳-۲ کلیات مدل و روابط کلی بیان خواهد شد، سپس در ۳-۲-۱ روابطی که بر اساس آن مدل گرافی احتمالی ساخته خواهد شد تعریف می‌شود. در بخش ۳-۲-۲ روش یادگیری پارامترها بیان می‌شود و در ۳-۲-۳ الگوریتم و روابط محاسبه برچسب خروجی بر حسب ورودی بیان خواهد شد.

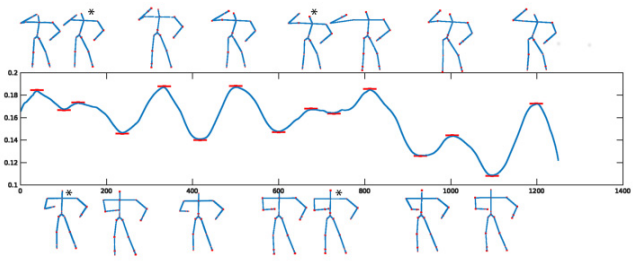
۳-۱ نمایش ویژگی و استخراج فریم‌های کلیدی

اسکلت‌ها از مجموعه‌ای از موقعیت مکانی مفاصل استخراج‌شده از اطلاعات عمق خام [۳۸] تشکیل شده‌اند. از این رو توالی‌های اسکلت‌ها به عنوان مجموعه‌ای از نمایش‌های سطح بالا برای بازشناسی فعالیت‌های انسان مفید هستند. محققان برای استفاده از این نمایش سطح بالا برای بازشناسی فعالیت انسانی از داده‌های اسکلت خام (مثلاً در [۳۹]، [۳۲] و [۴۰]) یا چندین ویژگی از آنها استفاده می‌کنند.

در این مقاله، فاصله بین مفاصل اسکلت در هر فریم کلیدی به عنوان ویژگی مورد استفاده قرار می‌گیرد (توضیح داده شده در بخش بعدی). مدل‌های گرافی احتمالی ابزار قدرتمندی برای به دست آوردن روابط زمانی هستند و بنابراین ما از ویژگی‌های اسکلتی که خصوصیات درون فریم را منعکس می‌کنند، استفاده می‌کنیم.

این نکته مهم است که فریم‌های متوالی در ویدئو اغلب بسیار هم‌بسته هستند و مقدار زیادی تکرار دارند. بنابراین برای بهبود راندمان محاسباتی الگوریتم از یک انتخابگر فریم کلیدی استفاده می‌کنیم. این انتخابگر فریم کلیدی یک مجموعه مختصر از فریم‌ها را شناسایی می‌کند که می‌توانند فعالیت را مشخص کنند. برای این منظور، ما رویکرد استفاده‌شده در [۴۱] را با برخی معیارهای اضافی برای انتخاب فریم کلیدی بسط می‌دهیم. استفاده از انرژی جنبشی در [۴۱] به عنوان یک معیار برای تعیین فریم کلیدی استفاده شده است. در این کار ویژگی استخراج‌شده از اسکلت عبارت است از مربع تفاضل بین مکان مفاصل تقسیم بر بازه زمانی. با این تعریف، آنها فریم‌هایی را انتخاب می‌کنند که مقدار صفر یا مقدار بسیار کمی انرژی جنبشی داشته باشد. رویکرد آنها دو عیب عمده دارد: ۱) در برخی موارد، فعالیت ممکن است به حالت ساکن و یا نزدیک به ساکن برسد، در حالی که جابه‌جایی‌های کوچک در اندام‌ها وجود دارد (مانند دو نقطه در شکل ۳ با علامت *). حالات ساکن نماینده خوبی برای فریم‌ها هستند زیرا نقاط کرانی سیگنال انرژی جنبشی را نشان می‌دهند. با این حال، انرژی جنبشی ژست برای این حالت‌های نشان داده شده در شکل نسبتاً مقدار زیادی دارد. روش [۴۱] این موارد را نادیده می‌گیرد و همه فریم‌های کلیدی در این شرایط را انتخاب نمی‌کند. ۲) در برخی از موارد ژست‌هایی وجود دارد که در بین دو حالت ژست با انرژی صفر (یا کم) قرار دارند ولی به منظور بازشناسی فعالیت ژست‌های مناسبی می‌باشند (برای مثال قله‌های انرژی در شکل ۳ که در قسمت بالایی شکل نشان

1. Shallow Classifier



شکل ۳: انرژی حرکت اسکلت برای فعالیت نوشیدن آب در وسط شکل به همراه اسکلت‌های قله‌ها در بالا و دره‌ها در پایین.

داده شده است). این اطلاعات آموزنده و فریم‌های مربوط نیز توسط [۴۱] نادیده گرفته می‌شود.

برای حل این مسایل، ما تغییرات در تابع انرژی جنبشی را در نظر می‌گیریم. به طور خاص، ما هر دو دره و قله را به ترتیب در انرژی جنبشی در نظر می‌گیریم، به جای آن که تنها به نقاط با انرژی تقریباً صفر نگاه کنیم. از آنجا که این نقاط کران انرژی هستند انتظار می‌رود که آنها حاوی اطلاعات ژست مهمی باشند. شکل ۳ انرژی جنبشی ژست یک توالی نمونه از اسکلت را نشان می‌دهد (فعالیت "آشامیدن آب"). همان طور که دیده می‌شود، فریم‌های کلیدی انتخاب‌شده (ژست‌های نشان داده شده در قسمت بالا و پایین نمودار) در کمینه یا بیشینه محلی (به شدت بزرگ‌تر یا کوچک‌تر در همسایگی آنها) از تابع انرژی قرار می‌گیرند. جالب است که در طول مدت زمانی که "آشامیدن آب" در حال رخدادن است، دست چندین بار بالا می‌رود و پایین می‌آید، بالاترین و پایین‌ترین موقعیت دست دو ژست متمایز هستند که توالی و تغییرات می‌تواند نمایانگر فعالیت مؤثر باشد. استراتژی انتخاب فریم کلیدی ما تضمین می‌کند که حداقل یک فریم کلیدی برای هر کدام از آنها انتخاب شود. علاوه بر این در برخی موارد ویدئو بسیار کوتاه بوده و شامل تنها چند دره و قله مانند مجموعه داده UT-Kinect است (بعداً مورد بحث قرار خواهد گرفت). در چنین مواردی، ما به طور یکنواخت چندین فریم از کل ویدئو را انتخاب می‌کنیم تا از پوشش محدوده زمانی ویدئو اطمینان حاصل کنیم.

همان طور که بیان شد ما در این مقاله از ویژگی‌های اسکلت استفاده می‌کنیم ولی یکی از چالش‌های اصلی در استفاده از داده اسکلت قابل اعتماد و ثبات آنها می‌باشد. در اینجا ما سه سناریو را که در آنها داده اسکلت مخدوش می‌باشد بررسی می‌کنیم: ۱) زمانی که یک فریم از ویدئو و بالطبع تمامی اسکلت‌های مربوط به آن از دست برود، ۲) زمانی که تعدادی از مفاصل گم شود و ۳) زمانی که دستگاه کینکت به درستی اسکلت را استخراج نکند. ادعای ما این است که در هر سه حالت الگوریتم ما عملکرد درستی خواهد داشت. در حالت‌های اول و سوم قسمت انتخاب فریم کلیدی ما فریم‌های مناسب را که به خوبی بتوانند فعالیت را نمایش دهند، انتخاب خواهد کرد. فریم‌های مخدوش یا گم‌شده معمولاً در این قسمت نادیده گرفته خواهند شد زیرا آنها شامل حرکات سریع و تصادفی هستند. در حالت دوم در صورتی که اطلاعات مفاصل ضروری وجود داشته باشد مشکلی ایجاد نمی‌شود. مفاصل از دست رفته را با مقدار صفر جایگزین می‌کنیم و بردار ویژگی را از روی داده اسکلت با طول ثابت ایجاد می‌کنیم. پس بنابراین روش ما در برابر تمامی این مشکلات داده اسکلت می‌تواند عملکرد قابل قبول و مناسبی داشته باشد.

۳-۲ مدل

به منظور آموزش مدل، K نمونه $(x^i \in X, y^i \in Y)_{i=1}^{K}$ از مجموعه داده $\mathcal{D}$ استفاده می‌شود که هر کدام شامل یک ویدئو است. به تعداد n_i فریم کلیدی (در قسمت قبلی در مورد آن بحث شد) از iامین ویدئو انتخاب خواهد شد $x^i, (\forall 1 \le i \le K)$. jامین فریم، با استفاده از بردار x_j^i نشان داده خواهد شد و بنابراین هر ویدئو به صورت $\{x_1^i, x_2^i, \ldots, x_{n_i}^i\} \in x^i$ نمایش داده می‌شود. هدف ما طراحی یک مدل گرافی احتمالی متمایزکننده است تا بتواند برچسب خروجی y^i را با فرض ورودی x^i پیش‌بینی کند. بدین منظور هر بردار x_j^i به یک حالت پنهان h_j^i مرتبط خواهد بود که این کار یک سری بردار پنهان را نتیجه می‌دهد $h^i = \{h_1^i, h_2^i, \ldots, h_{n_i}^i\}$. از این به بعد برای سادگی ما اندیس‌های i و j را نمایش نخواهیم داد.

یک تابع ارزیابی $g(x,y,h,\theta)$ بر اساس تابع پیش‌بینی $f_\theta(x) = \arg\max_{y,h} g(x,y,h,\theta)$ پیکربندی پارامترها را ارزیابی می‌کند. تابع خروجی f_θ پارامترهای مدل θ را برای هر ویدئوی داده‌شده x تعریف می‌کند. از این به بعد از $f_\theta(x)$ به عنوان تابع پیش‌بینی نام برده خواهد شد. با تعریف یک تابع خسارت[1] $\Delta(y, f_\theta(x))$ که هزینه پیش‌بینی اشتباه بین برچسب پیش‌بینی شده $f_\theta(x)$ و برچسب درست y را محاسبه می‌کند، می‌توانیم خسارت کلی پیش‌بینی ویدئوی $l(x,y,h,\theta) = \Delta(y, f_\theta(x))$ را محاسبه کنیم. پس بنابراین احتمال هر برچسب با داشتن مجموعه‌ای از حالت‌های پنهان و داده ورودی متناسب با رابطه زیر است

$$p_\theta(x,y,h,\theta) \propto \theta^T \phi(x,y,h) + l(x,y,h,\theta) \qquad (۱)$$

که در آن $\phi(.)$ تابع پتانسیل است که $\psi(.)$ را در فرم لگاریتم خطی تعریف می‌کند

$$\psi(x,y,h) = \exp(\theta^T \phi(x,y,h) + l(x,y,h,\theta)) \qquad (۲)$$

خانواده توزیع احتمال توأم[2] در یک مدل گرافی را می‌توان بر حسب ضرب توابع پتانسیل بیان کرد [۴۲]. تابع خسارت یک تنبیه[3] را زمانی تعریف می‌کند که مقادیر متغیر پیش‌بینی‌شده متفاوت از مقادیر حقیقی باشند. در اینجا این نکته باید مورد توجه قرار گیرد که مدل ارائه‌شده توسط ما از نظر ساختار تعمیمی از ساختار HCRF [۴۳] است با این تفاوت که ما وابستگی‌های طولانی بین متغیرهای پنهان اضافه کرده‌ایم. این امر ما را قادر می‌سازد تا ویژگی‌های فعالیت‌های ساده را مدل کنیم. در مدل ما تابع ارزیابی $g(x,y,h,\theta)$ به صورت زیر تعریف می‌شود

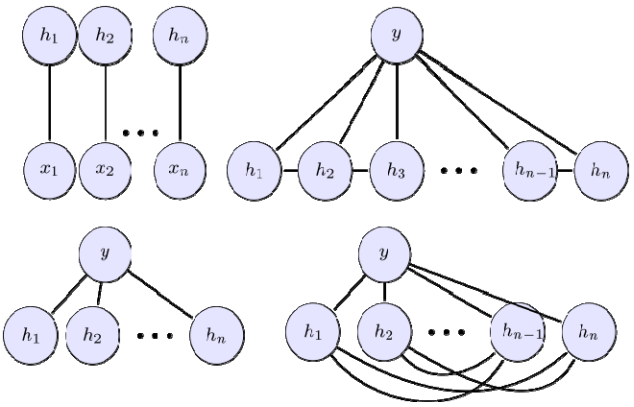
$$g(x,y,h,\theta) = \theta^T \phi(x,y,h) \qquad (۳)$$

که پیش‌بینی شبکه را با توجه به ورودی اندازه‌گیری می‌کند. همان طور که مشخص است، به حداقل رساندن تابع هدف در (۱) غیر عملی است چون تابع خسارت به صورت تکه‌ای ثابت[4] است و این موضوع روش‌های بهینه‌سازی مبتنی بر شیب را بلااستفاده می‌کند. بر اساس نظریه یادگیری آماری، برای به دست آوردن یک راه‌حل بهینه، کمینه‌کردن کران بالای محدب (۱) کافی است [۴۴]. بنابراین تابع خسارت با کران بالای آن جایگزین خواهد شد

$$l(x,y,h,\theta) = \max_{y \in Y} \Delta(y, f_\theta(x)) - g(x,y,h,\theta) + g(x, f_\theta(x), h, \theta) \qquad (۴)$$

این فرمول حد بالای تابع خسارت است زیرا برای هر $f_\theta(x)$ داریم

1. Loss
2. Joint Probability Distributions
3. Penalty
4. Piece-Wise Constant



شکل ۴: چهار نوع ارتباط برای توابع پتانسیل در مدل گرافی احتمالی پیشنهادی که به صورت جداگانه در ابتدا آموزش داده می‌شوند. بالا چپ: تابع پتانسیل یکتا، پایین چپ: تابع پتانسیل دوتایی، بالا راست: تابع پتانسیل سه‌تایی نوع ۱ و پایین راست: تابع پتانسیل سه‌تایی نوع ۲.

$$\begin{aligned}&\Delta(y,f_\theta(x))\le\\&\Delta(y,f_\theta(x))-g(x,y,h,\theta)+g(x,f_\theta(x),h,\theta)\le\\&\max_{y\in Y}\Delta(y,f_\theta(x))-g(x,y,h,\theta)+g(x,f_\theta(x),h,\theta)=\\&l(x,y,h,\theta)\end{aligned}\quad(۵)$$

بنابراین (۱)، مشابه LSSVM [۱۳] از طریق به حداقل رساندن خسارت لولای تنظیم‌شده محدب[1] بهینه می‌شود.

۳-۲-۱ توابع پتانسیل

همان طور که بیان شد، روابط بین متغیرها که از طریق توابع پتانسیل تعریف می‌شوند، مدل گرافی احتمالی را تعریف می‌کنند. تابع پتانسیل سازگاری بین متغیرها را اندازه‌گیری می‌کند. در این مقاله، چهار نوع مختلف از توابع پتانسیل برای به دست آوردن تمام جنبه‌های فعالیت‌های پیچیده در نظر گرفته شده‌اند. این چهار نوع در شکل ۴ نشان داده شده‌اند که شامل یک تابع یکتایی، یک تابع باینری و دو تابع سه‌تایی است. تابع یکتایی رابطه بین داده‌های ورودی و متغیرهای پنهان را مشخص می‌کند در حالی که تابع پتانسیل دوگانه (یا جفتی) روابط بین متغیرهای پنهان و برچسب‌ها را مدل‌سازی می‌کند. می‌بایست توجه شود که یک بخش از تابع پتانسیل دوتایی (یعنی برچسب y) مشاهده[2] می‌شود. مشاهده‌شدن یک متغیر به این معنی است که در هنگام آموزش مقدار یا حالت آن متغیر مشخص است و از آنجایی که متغیر y کلاس یا همان شماره فعالیت در حال انجام در ویدئو است، در دسته متغیرهای مشاهده‌شده قرار می‌گیرد. ولی متغیرهای پنهان (h_i) ما به ازای خارجی ندارند پس مشاهده‌شده نیستند.

به علاوه، ما دو نوع از توابع پتانسیل سه‌تایی را تعریف می‌کنیم که نوع اول بین متغیرهای پنهان متوالی مانند HCRF استاندارد تعریف می‌شود در حالی که نوع دوم بین متغیرهای ابتدایی[3] و پایانه[4] تعریف می‌شود. شهود تعریف این نوع از پتانسیل این است که در برخی موارد، فعالیت‌ها بسیار صاف و سریع هستند و در نتیجه، ژست‌های میانی اغلب متعدی هستند که باعث می‌شود که ابتدا و انتهای دنباله توالی کاملاً تعیین‌کننده باشد.

1. Convex Regularized Hinge Loss
2. Observed
3. Initial
4. Terminal

اگر ویژگی‌های استخراج‌شده از فریم x توسط $\zeta(x)$ نشان داده شود، تابع پتانسیل یکتا توسط رابطه زیر تعریف می‌شود

$$\psi_1(x,h_i)=\exp(\theta_1^i(h_i)\zeta(x)+l(x,.,h_i,\theta_1^i))\quad(۶)$$

که در آن در صورتی که ارزش واقعی و پیش‌بینی شده برای x داده‌شده متفاوت باشد، خسارت باعث هزینه می‌شود. در اینجا . مکانی برای متغیرهای ورودی است که هیچ تأثیری در عملکرد فعلی ندارند. علاوه بر این، تابع پتانسیل باینری رابطه بین حالت‌های پنهان و برچسب‌ها را اندازه‌گیری می‌کند

$$\psi_2(h_i,y)=\exp(\theta_2^i(y,h_i)+l(.,y,h_i,\theta_2^i))\quad(۷)$$

از آنجا که برخی از حالت‌های میانی در برخی فعالیت‌ها رخ می‌دهند (و مؤثر هستند)، در حالی که برای برخی دیگر آنها اتفاق نمی‌افتند (یا به ندرت اتفاق می‌افتد)، (۷) در تعیین کلاس فعالیت مؤثر است زیرا سازگاری بین هر حالت پنهان و برچسب مربوط را اندازه‌گیری می‌کند. با این حال، برای برخی فعالیت‌ها، ما باید روابط بین دنباله‌های متغیر پنهان و برچسب را در نظر بگیریم. آنها در تعاریف پتانسیل سه‌تایی ما پوشش داده می‌شوند. به منظور در نظر گرفتن الگوی تغییرات در حالت‌های پنهان متوالی با توجه به هر برچسب فعالیت، ما پتانسیل سه‌تایی نوع ۱ را بین برچسب‌ها و هر جفت از متغیرهای پنهان متوالی تعریف می‌کنیم

$$\psi_3(h_{i-1},h_i,y)=\exp(\theta_3^i(y,h_{i-1},h_i)+l(.,y,h_{i-1},h_i,\theta_3^i))\quad(۸)$$

به علاوه، ما یک تابع پتانسیل سه‌تایی نوع ۲ را تعریف می‌کنیم که برای تشخیص نوع خاصی از فعالیت‌های ساده مفید است که در آن آغاز و پایان فعالیت بسیار مهم هستند. برای مثال فعالیت، "هل‌دادن" را در نظر بگیرید. این فعالیت می‌تواند با شروع و پایان توالی (یعنی حالت‌های ابتدایی و انتهایی) به خوبی نمایش داده شود، در حالی که حالت‌های میانی اهمیت کمتری دارند. از آنجا که این فعالیت کوتاه است و افراد مختلف ممکن است با سرعت‌های متفاوت عمل کنند، حتی تعداد حالات میانی هم ممکن است برابر یا مشابه نباشند، در حالی که حالت اولیه (یعنی h_1 و h_2) و پایانه (یعنی h_{n-1} و h_n) بسیار مهم هستند. از این رو ما رابطه زیر را تعریف می‌کنیم

$$\begin{aligned}&\psi_4(h_j,h_k,y)=\\&\exp(\theta_4^{j,k}(y,h_j,h_k)+l(.,y,h_j,h_k,\theta_4^{j,k}))\end{aligned}\quad(۹)$$

که در آن j و k به ترتیب اندیس فریم‌های ابتدایی و انتهایی است. از آنجایی که حالت‌های پنهان متغیرهایی هستند که مشاهده نمی‌شوند، ما اطلاعاتی در مورد آنها نداریم پس ما حالت‌های پنهان را با عمل جمع روی متغیرها حذف می‌کنیم

$$p_\theta(y|x)=\sum_h p_\theta(y,h|x)\quad(۱۰)$$

با توجه به توابع پتانسیل تعریف‌شده، احتمال شرطی به صورت زیر تعریف می‌شود

$$\begin{aligned}&p_\theta(y|x)=\sum_h[\frac{1}{z_\theta(x)}\prod_{i,j,k}(\psi_1(x,h_i)\times\psi_2(h_i,y)\times\\&\psi_3(h_{i-1},h_i,y)\times\psi_4(h_j,h_k,y))]\end{aligned}\quad(۱۱)$$

که در آن

$$\begin{aligned}&z_\theta(x)=\sum_{y,h}\prod_{i,j,k}(\psi_1(x,h_i)\times\psi_2(h_i,y)\times\\&\psi_3(h_{i-1},h_i,y)\times\psi_4(h_j,h_k,y))\end{aligned}\quad(۱۲)$$

نیازی به تعریف صریح حالات پنهان وجود ندارد. مقادیر و پارامترهای آنها به طور خودکار در طول آموزش یاد گرفته می‌شوند، با این حال مقداردهی اولیه می‌تواند مفید باشد. بر اساس توابع پتانسیل تعریف‌شده در بالا، ما می‌توانیم به طور جمعی فعالیت‌های ساده و پیچیده را توأمان تشخیص دهیم.

۳-۲-۲ یادگیری پارامترها

برای یادگیری پارامترها ما از روش مشابه [۱] پیروی می‌کنیم. فرض می‌کنیم داده‌ها i.i.d[1] هستند و ما N نمونه $(x,y)_1^N$ داریم و می‌خواهیم لگاریتم درست‌نمایی منفی[2] ورودی را به حداقل برسانیم، علاوه بر آن از یک پیش‌فرض $p(\theta) \propto (\theta)_p^p$ به عنوان عبارت قاعده‌مندی[3] برای جلوگیری از بیش‌برازش[4] استفاده می‌کنیم. بنابراین عبارت زیر باید کمینه شود

$$-\ln(p(\theta))\prod_{(x,y)\in D} p(y|\theta) \qquad (۱۳)$$

سپس خواهیم داشت

$$\frac{c}{p}\|\theta\|_p^p + \sum_{(x,y)\in D}(\ln Z_\theta(x) - \ln\sum_{h\in H}\exp(\sum_{i,j,k} E_{i,j,k}(x,h,y))) \qquad (۱۴)$$

که در آن H مجموعه‌ای از همه حالت‌های پنهان است و E به صورت زیر تعریف می‌شود

$$\begin{aligned} E_{i,j,k}(x,h,y) = {} & \theta_1^i(h_i)\zeta(x) + l(x,.,h_i,\theta_1^i) + \\ & \theta_2^i(h_i,y) + l(.,y,h_i,\theta_2^i) + \theta_3^i(h_{i-1},h_i,y) + \\ & l(.,y,h_{i-1},h_i,\theta_3^i) + \theta_4^{j,k}(h_j,h_k,y) + l(.,y,h_j,h_k,\theta_4^i) \end{aligned} \qquad (۱۵)$$

که در آن تابع افراز z به صورت زیر تعریف می‌شود

$$z_\theta(x) = \sum_{h,y}\exp(\sum_{i,j,k} E_{i,j,k}(x,h,y)) \qquad (۱۶)$$

LSSVM و HCRF هر دو می‌توانند به طور موفقیت‌آمیزی برای یادگیری پارامترها به کار گرفته شوند و با این حال در [۱] و [۴۵] نشان داده شده است که یک چارچوب کلی با عنوان پیش‌بینی ساختاریافته توزیع‌شده (DSP)[5] وجود دارد که هر دو روش را پوشش می‌دهد و می‌تواند در یک محیط توزیع‌شده بهینه شود

$$\frac{c}{p}\|\theta\|_p^p + \sum_{(x,y)\in D}(\varepsilon\ln\sum_{h,y'}\exp(\frac{\sum_{i,j,k} E_{i,j,k}(x,h,y')}{\varepsilon}) - \varepsilon\ln\sum_{h\in H}\exp\frac{\sum_{i,j,k} E_{i,j,k}(x,h,y)}{\varepsilon}) \qquad (۱۷)$$

این مدل دارای یک فراپارمتر (ε) است. اگر $\varepsilon = 1$، فرمول HCRF را نشان می‌دهد و اگر $\varepsilon \to 0$ معادله به (LSSVM) متمایل می‌شود. ما از [۱] برای حل این مشکل پیروی می‌کنیم. معادله (۱۷)، یک جمع بر روی عبارات محدب و مقعر است. برای سادگی، بخش مقعر آن با کران بالایی که محدب است و یک عبارت غیر محدب دوخطی[6]، جایگزین می‌شود

$$\forall q_{x,y}(h) \in \Lambda, \varepsilon\ln\sum_{h\in H}\exp\frac{\sum_{i,j,k} E_{i,j,k}(x,h,y)}{\varepsilon} = \varepsilon\mathbb{H}(q_{x,y}(h)) + \mathbb{E}_{q_{x,y}(h)}[\sum_{i,j,k} E_{i,j,k}(x,h,y)] \qquad (۱۸)$$

Λ یک احتمال ساده است، $\mathbb{H}(.)$ آنتروپی را اندازه‌گیری می‌کند و $\mathbb{E}$ انتظار است. اگر ما (۱۸) را در (۱۷) قرار دهیم، جمعی از یک محدب، یعنی $\mathbb{H}(q_{x,y}(h))$ و عبارات دوخطی $\mathbb{E}_{q_{x,y}(h)}[\sum_{i,j,k} E_{i,j,k}(x,h,y)]$، یعنی که قابل حل است [۱] خواهیم داشت.

۳-۲-۳ استنتاج

استنتاج در مورد مدل‌های گرافیکی با ساختار درختی می‌تواند به طور دقیق و مؤثر از طریق انتشار باور انجام شود، با این حال چنین روشی نمی‌تواند زمانی اعمال شود که گراف‌ها دارای ساختارهای غیر درختی (مانند مدل ما) باشند. در چنین مواردی، اغلب روش‌های تقریبی استفاده می‌شوند که از نظر محاسباتی بسیار سنگین هستند. در این مقاله از یک رویکرد تجزیه دوگانه[7] استفاده می‌شود که از محاسبات کمتری برخوردار است و پیاده‌سازی آن می‌تواند به سادگی موازی‌سازی شود. در مرحله آزمایش، ما به دنبال بهترین پیکربندی متغیرها هستیم، یعنی ما داده‌های ورودی $\{x_1,x_2,\ldots,x_n\}$ را داریم و می‌خواهیم $\{h_1,h_2,\ldots,h_n\}$ و y را به دست آوریم، در این صورت بر اساس (۱۱) خواهیم داشت

$$Y,H = \arg\max_{Y,H}\prod_{x,y,h}\psi(x,y,h) \qquad (۱۹)$$

در مدل لگاریتم خطی (۱۱)، ما تابع افراز (z) را هم داریم، اما همان طور که در (۱۲) همه متغیرها در توزیع احتمالی شرطی حاشیه[8] شده‌اند، تابع افراز در (۱۹) نادیده گرفته می‌شود. اگر ما از لگاریتم (۱۹) استفاده کنیم و با استفاده از یک تابع نشانگر[9] $b(x,y,h)$ (۱۹) را به یک برنامه عددی خطی[10] تبدیل کنیم، مسئله زیر به دست می‌آید

$$\max_{y,h}\sum_{y,h,x} b(x,y,h)\ln\psi(x,y,h) \qquad (۲۰)$$

که در آن

$$b(.)\in\{0,1\}\ ,\ \sum_{x,y,h} b(x,y,h) = 1 \qquad (۲۱)$$

معادله (۲۰) یک مسئله برنامه‌نویسی خطی عدد صحیح[11] را معرفی می‌کند که NP-hard است. ما می‌توانیم آن را با استفاده از یک ساده‌سازی برنامه‌نویسی عدد صحیح با جایگزین‌کردن محدودیت‌های مساوی با نابرابری غیر منفی حل کنیم و فقط راه‌حل‌های صحیح را قبول کنیم. حل این مسئله نیز از نظر محاسباتی سنگین است، با این حال معمولاً برای حل این نوع از مسایل، محدودیت‌های ساده[12] استفاده می‌شوند. محدودیت‌ها با استفاده از تابع آنتروپی تقریب زده می‌شوند

$$\max_{y,h}\sum_{y,h,x} b(x,y,h)\ln\psi(x,y,h) + \varepsilon(\sum_{x,y,h}\mathbb{H}(b)) \qquad (۲۲)$$

که این مسئله باعث می‌شود مسئله به یک مسئله اکیداً مقعر تبدیل شود و شکل دوگانه صاف[13] داشته باشد. ما از چارچوب انتشار باور توزیع‌شده[1] [۳]

1. Independent and Identically Distributed
2. Negative Log-Likelihood
3. Regularization Term
4. Overfitting
5. Distributed Structured Prediction
6. Non-Convex Bi-Linear
7. Dual Decomposition
8. Marginalized
9. Indicator
10. Linear Integer Program
11. Integer Linear Programming
12. Simplex Constraints
13. Smooth Dual Form

برای حل مسئله (۲۲) استفاده کردیم و این استنتاج را انجام دادیم.

۳-۳ تعویض- رده

همان طور که قبلاً طیف گسترده‌ای از انواع فعالیت که ما هدف قرار می‌دهیم مورد بحث قرار گرفت، پارامترهای مدل بر روی ساختار باید متفاوت باشد. یک طرح تعویضی رده می‌تواند این کار را انجام دهد.

ذکر این نکته مهم است که فعالیت‌ها در مجموعه داده‌های مختلف دارای ویژگی‌های متفاوتی هستند و یک معیار ساده می‌تواند آنها را به چند رده کلی[2] تقسیم کند. به منظور دستیابی به پیچیدگی محاسباتی بهتر (از طریق موازی‌سازی) و به طور هم‌زمان عملکرد بهتر، ما به طور خودکار کلاس‌های فعالیت‌ها را در مجموعه داده‌ها به برخی از رده‌های کلی تقسیم می‌کنیم. سپس برای هر گروه یک مدل گرافی احتمالی (با همان ساختار توصیف‌شده در بخش ۳-۲) آموزش می‌دهیم. این نکته باید مورد توجه قرار گیرد که روش تعویضی ارائه‌شده توسط ما، هیچ گونه محدودیتی را بر روی کاربرد کلی مدل تحمیل نمی‌کند زیرا عمل تعویض برای تقسیم کل فضا به افرازهای کوچک، یک عمل قطعی (پیاده‌سازی شده با استفاده از آستانه) برای گروه‌بندی فعالیت‌های مشابه است. به طور کلی ما در آزمایش‌های خود، مشاهده کردیم که تقسیم خودکار داده‌ها به زیرمجموعه‌های کوچک‌تر مفید است و پارامترهای مدل را در بخش‌های کوچک‌تر یاد می‌گیریم. علاوه بر این، این فرایند چندین فرایند مجزا را در این مدل ایجاد می‌کند که می‌تواند به راحتی قابل موازی‌سازی باشد و روی پلتفرم‌های محاسبات توزیعی اجرا شود. همان طور که در آزمایش‌ها گزارش می‌شود، زمانی که روش خود را بدون گام تعویضی- رده اجرا کردیم هم نتایج خوبی به دست آوردیم. این تعویض به سادگی به یک مدل دقیق‌تر و سریع‌تر منجر می‌شود. ما از یک طرح تعویضی قطعی (شبیه به [۵] و [۴۶]) استفاده می‌کنیم تا این گام تعویضی را طراحی کرده و آن را در کنار مدل احتمالاتی خود اجرا کنیم. برای یک مدل گرافی احتمالی $p_w(y,h,x)$ که برای یکی از دسته‌های رده کلی تعریف شده است، تعویض به صورت زیر فرموله شده است

$$\beta_i p_w(y,h,x)\ ,\ \forall i \in \{1,\dots,K\} \qquad (۲۳)$$

که در آن K تعداد مدل‌هایی است که می‌توان بین آنها تعویض کرد که هر کدام به یک متغیر β_i مرتبط است. رده‌ها فضای برچسب‌ها را به K زیرمجموعه افراز می‌کند و بنابراین در صورتی که Y فضای تمامی برچسب‌ها باشد خواهیم داشت $\cup_{i=1}^{K} G_i = y$.

باید توجه داشت که β_i برای تنها یک $i \in \{1,\dots,K\}$ می‌تواند برابر با یک باشد و بقیه باید صفر باشند. بنابراین بردار β، تعویض رده را برای مدل پیش‌بینی ساختار تعویضی تعریف می‌کند. در هر زمان تنها نیاز به انجام محاسبات بر روی یکی از مدل‌ها می‌باشد و بنابراین حجم محاسبات کاهش می‌یابد.

ما یک دسته‌بند ساده را آموزش دادیم تا مشخص کند کدام یک از β_iها مساوی یک است. برای این منظور ما دیتاست را به K رده کلی $G_1, G_2, \dots, G_K$ تقسیم می‌کنیم که هر کدام از این رده‌ها شامل زیرمجموعه‌ای از برچسب‌ها می‌باشد. سپس به طور موقت هر رده G_i را یک برچسب جدید در نظر می‌گیریم و تمامی نمونه‌های آموزشی مربوط به برچسب‌های زیرمجموعه آن را به برچسب جدید اختصاص می‌دهیم. از آنجایی که این کلاس‌ها بسیار گسترده هستند و داده‌های درون هر یک تفاوت زیادی با هم دارند، دسته‌بند، کار چندان مشکلی برای جداسازی آنها نخواهد داشت. پس بنابراین هر نمونه جدید بر اساس این دسته‌بند، رده‌بندی و β مربوط به آن برابر یک خواهد شد.

٤- نتایج و آزمایش‌ها

در این بخش سه مجموعه داده مورد استفاده برای آزمایش‌ها را معرفی می‌کنیم و بعد از آن تنظیمات آزمایش‌ها و نتایج روی این مجموعه داده‌های مختلف ارائه خواهد شد. پیاده‌سازی‌ها با استفاده از ابزارهایی که توسط [۱] تا [۳] ارائه گردیده انجام شده که به صورت دستورات فرمان خط[3] در محیط ترمینال لینوکس اجرا می‌شوند ولی هر فرمان نیاز به فایل‌های کمکی دارد که در آنها مدل و همین طور دیتای ورودی باید تعریف شده باشد که این فایل‌ها توسط نرم‌افزار Matlab تولید شده‌اند.

٤-١ دیتاست‌ها

ما از سه مجموعه داده برای ارزیابی مدل پیشنهادی استفاده می‌کنیم: CAD-۶۰ [۹]، UT-Kinect [۴۷] و Florence ۳D [۴۸]. این مجموعه داده‌ها گستره وسیعی از فعالیت‌های ساده و پیچیده را پوشش می‌دهند. برای مثال UT-Kinect و Florence ۳D دارای فعالیت‌هایی از حرکات ساده، هسته‌ای[4] و سریع هستند در حالی که CAD-۶۰ شامل تعدادی از فعالیت‌های پیچیده با چندین عمل هسته‌ای است. تمام این دیتاست‌ها شامل داده اسکلت هستند که توسط دستگاه کینکت جمع‌آوری شده‌اند.

٤-٢ پیکربندی

در ابتدا داده‌های اسکلت به طور کامل با بیشترین مقدار برای فاصله بین مفصل‌ها در مجموعه داده‌ها نرمال شده‌اند. همه اسکلت‌ها در هر فریم به گونه‌ای می‌چرخند که همه به سمت یک صفحه از پیش تعیین شده روبه‌رو شوند. ما $\varepsilon = 0$ را در (۱۸) قرار دادیم و از چارچوب LSSVM [۱۳] برای حل مسئله استفاده کردیم. لازم به ذکر است که ما همچنین مقادیر دیگر ε را (در دامنه [۰،۱]) مورد بررسی قرار داده‌ایم اما این مقادیر زمان همگرایی را افزایش داد و در عین حال دقت کلی کاهش پیدا کرد. برای مقداردهی اولیه حالت‌های پنهان، ما داده‌های آموزشی را با استفاده از سه روش kmeans، k-medoids و Gaussian Mixture Model خوشه‌بندی می‌کنیم، سپس توابع پتانسیل یکتا را به صورت کاملاً مشاهده‌شده آموزش می‌دهیم، هر خوشه توسط یک حالت پنهان نشان داده می‌شود. وزن به دست آمده به عنوان وزن‌های مقداردهی اولیه برای مدل به کار می‌رود.

اگرچه ما از یک چارچوب پردازش توزیع‌شده استفاده کردیم که سرعت آموزش را افزایش می‌دهد، ولی به علت پارامترهای بسیار زیاد مدل، فرایند آموزش طولانی خواهد بود. بنابراین برای یادگیری سریع‌تر پارامترها، یک رویکرد تقسیم و ادغام برای تسریع روند آموزشی پیشنهاد می‌کنیم. همان طور که در بخش ۳-۲-۱ اشاره شد، مدل گرافی احتمالی چهار نوع توابع پتانسیل (که در شکل ۴ نشان داده شده است) دارد. ما چهار مدل گرافی احتمالی مختلف را می‌سازیم که هر یک شامل یک نوع توابع پتانسیل جداگانه هستند. سپس وزن هر یک از آنها را ترکیب می‌کنیم و در نهایت مدل گرافی احتمالی کامل را برای چند تکرار دیگر آموزش می‌دهیم. ذکر این نکته ضروری است که ما نمی‌توانیم

1. Distributed Convex Belief Propagation
2. Broad Categories
3. Command Line
4. Atomic

جدول ۱: رده‌های کلی برای مجموعه داده CAD-۶۰.

شماره ماژول	به سمت دیوار	ایستادن	فعالیت
۱	×	✓	مسواک‌زدن، زدن لنز به چشم، حرف‌زدن با تلفن، نوشیدن آب، بازکردن در قوطی
۲	✓	×	کار با کامپیوتر
۳	✓	✓	آب‌کشیدن دهان، پخت‌وپز (خرد کردن)، پخت‌وپز (سرخ‌کردن)، نوشتن روی تخته وایت‌برد
۴	×	×	صحبت‌کردن روی کاناپه، استراحت روی کاناپه

جدول ۲: رده‌های کلی برای مجموعه داده UT-Kinect.

شماره ماژول	حرکت	ایستادن	استفاده از دو دست	فعالیت
۱	✓	×	N/A	راه‌رفتن، حمل‌کردن
۲	×	✓	N/A	نشستن، ایستادن، بلندکردن
۳	×	×	×	پرتاب‌کردن، هل‌دادن، کشیدن
۴	×	×	✓	دست‌تکان‌دادن و کف‌زدن

پتانسیل‌های دوگانه و سه‌گانه را آموزش دهیم، مگر این که حالات متغیرهای پنهان را تعیین کنیم. بنابراین در ابتدا ما پتانسیل یکتا را آموزش می‌دهیم، سپس بر روی آن استنتاج می‌کنیم تا حالت‌های پنهان را تعیین کنیم، بعد از آن ما از خروجی مدل یکتا برای آموزش پارامترهای مدل دوگانه و سه‌گانه استفاده می‌کنیم. ما پارامترها را ۶۰ بار سریع‌تر با استفاده از روش تقسیم و ادغام یاد می‌گیریم. الگوریتم‌ها با دو پردازنده اینتل Xeon E۵-۲۶۲۰ v۳ بر روی یک ماشین اجرا شده‌اند.

۴-۳ رده‌های فعالیت‌ها مورد استفاده برای تعویضی رده

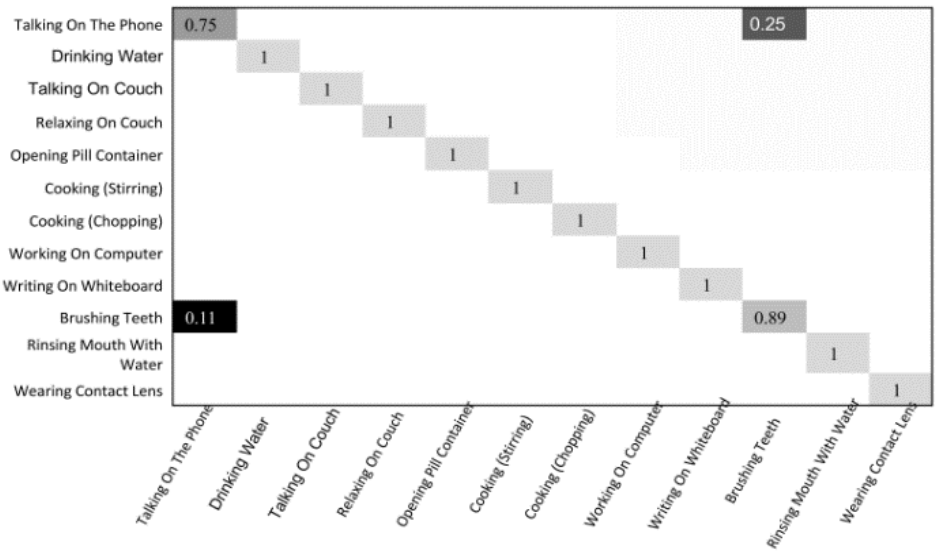
همان طور که قبلاً بحث شد، در چارچوب پیشنهادی ما ابتدا فعالیت‌ها بر اساس ویژگی‌های به دست آمده از اسکلت، به رده‌های کلی رده‌بندی می‌شوند (سطح اول). سپس برای هر رده، ما یک مدل گرافی احتمالی (سطح دوم) را تشکیل می‌دهیم و فعالیت‌ها را بازشناسی می‌کنیم. ساختارهای هر مدل برای دسته‌بندی یکسان هستند اما پارامترهای آنها متفاوت خواهد بود. این موضوع باید مورد توجه قرار گیرد که هر دو سطح اول و دوم به طور خودکار برای تمام مجموعه‌های داده انجام می‌شوند. حالاتی که ما برای اولین سطح عمومی برای تقسیم مجموعه داده CAD-۶۰ استفاده کردیم، ۱) ایستاده یا نشسته‌بودن و ۲) آزادانه به سمت هدفی مانند دیوار یا میز ایستادن است. این دو وضعیت برای CAD-۶۰ رده‌بندی می‌شوند و فعالیت‌ها را به چهار رده کلی طبقه‌بندی می‌کند. آنها به سادگی توسط آستانه‌های ساده از روی داده‌ها شناسایی می‌شوند. همان طور که در جدول ۱ مشاهده می‌کنید، اگر از این دو ویژگی ساده استفاده کنیم، "کارکردن با کامپیوتر" را می‌توان از فعالیت‌های دیگر جدا کرد. رده‌بندی به سمت دیوار بر اساس زاویه اصلی بدن انسان با توجه به محورهای افقی و عمودی انجام می‌شود. همچنین ما از فاصله بین مفاصل پایین‌تنه و مفاصل بالاتنه برای جداسازی وضعیت نشستن از ایستادن استفاده می‌کنیم.

ما از سه شرط برای تقسیم داده‌های UT-Kinect به رده‌های کلی استفاده کردیم. شروط حرکت، ایستادن و استفاده از یک یا دو دست مورد استفاده قرار گرفتند. همان طور که در جدول ۲ نشان داده شده است، برای این مجموعه داده ما ۴ رده در نظر گرفتیم، چرا که برخی شرایط برای برخی از گروه‌ها قابل اجرا نیستند و ما آنها با 'N/A' مشخص کردیم. در اینجا "نشستن" نشان می‌دهد که فرد حداقل برای بخشی از زمانی که فعالیت را انجام می‌دهد، ایستاده نیست. حرکات در جهت عمودی در طول فعالیت "حرکت" ناچیز است در نتیجه ما از تغییر موقعیت مفاصل در مسیرهای افقی برای تشخیص این حرکات استفاده کردیم. از طرف دیگر، حرکت در جهت عمودی در طول فعالیت‌های "نشستن" قابل توجه است. مانند CAD-۶۰، ما از فواصل بین محل اتصالات در بالاتنه و پایین‌تنه استفاده کردیم. دسته سوم مشخص می‌کند که کدام فعالیت از طریق یک یا دو دست انجام می‌شود. اگر فعالیت با یک دست انجام شود، یک دست حرکت می‌کند و دست دیگر حرکت ناچیزی دارد با این وجود، هنگام استفاده از دو دست، موقعیت مفاصل دو دست به مقدار قابل توجه و تقریباً یکسان است.

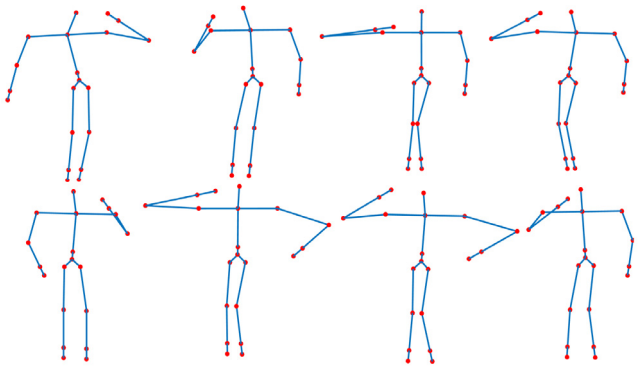
رده‌بندی داده برای مجموعه داده Florence ۳D کمی پیچیده‌تر است. برای جداکردن نمونه‌ها در این مجموعه داده برخلاف دو مجموعه داده دیگر، ما از چندین ویژگی استفاده کردیم و از یک دسته‌بندی کننده SVM ساده برای انجام رده‌بندی سطح اول استفاده می‌کنیم. در ابتدا ما به راحتی داده‌ها را به دو دسته بر اساس موقعیت ایستاده‌بودن یا نبودن (با استفاده از یک آستانه ساده) تقسیم می‌کنیم. از این رو مقوله اول شامل تکان‌دادن دست، کف‌زدن، نوشیدن از یک بطری، پاسخ به تلفن و خواندن است در حالی که بقیه در دسته دوم قرار دارند. هر یک از این دسته‌ها مجدداً تقسیم می‌شوند تا چهار دسته به دست آید (نشان داده شده در جدول ۳). برای دو ماژول اول جدول (یعنی ۱ و ۲) با استفاده از یک SVM با فواصل مشترک درون و بین فریم‌های بعدی به عنوان ویژگی‌های ورودی آن انجام می‌شود. ماژول‌های ۳ و ۴ با استفاده از SVM با استفاده از جهت حرکت مفاصل بین فریم‌ها به عنوان ویژگی ورودی آن جدا شده‌اند.

۴-۴ نتایج

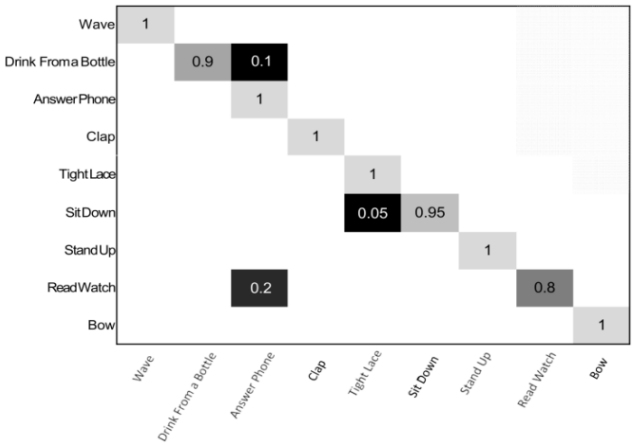
شکل ۵ ماتریس درهم‌ریختگی برای آزمایش روی مجموعه داده CAD-۶۰ را نشان می‌دهد. همان طور که دیده می‌شود، رویکرد ما برای اغلب کلاس‌ها عالی عمل می‌کند. اغلب، فعالیت‌های "صحبت‌کردن با تلفن" و "مسواک‌زدن" مستعد خطا هستند، توجه شود که ما تنها از داده‌های اسکلت برای تفکیک فعالیت‌ها استفاده می‌کنیم و حرکات اسکلت این دو فعالیت بسیار شبیه هم است. اگرچه در دنیای واقعی، "مسواک‌زدن" حرکات جزئی دارد که می‌تواند این کلاس را از "صحبت‌کردن با تلفن" جدا کند، با این حال از آنجا که اسکلت (که با استفاده از کینکت به دست می‌آید) دارای برخی سطوح نویز است، این تفاوت در داده‌ها مشخص نیست و این روش نمی‌تواند این کلاس‌ها را به راحتی از هم جدا کند. در شکل ۶ اسکلت این دو فعالیت به تصویر کشیده می‌شود. در اینجا باید دقت شود که وقتی ما از تغییر رده استفاده نمی‌کنیم دقت ۸۹٫۱۶٪ را به دست می‌آوریم که هنوز قابل مقایسه با دقت به‌دست‌آمده هنگامی که رده‌بندی را در نظر می‌گیریم است (یعنی ۹۷٫۶٪). با استفاده از استراتژی تعویضی، سرعت آموزش حدود ۴ برابر



شکل ۵: ماتریس درهم‌ریختگی برای مجموعه داده ۶۰- CAD.



شکل ۶: ژست برای دو فعالیت، در بالا صحبت‌کردن با تلفن و در پایین مسواک‌زدن.



شکل ۷: ماتریس درهم‌ریختگی بر روی مجموعه داده Florence ۳D.

جدول ۳: رده‌های کلی برای مجموعه داده‌های Florence ۳D.

شماره ماژول	جهت حرکت	ایستادن	فاصله بین مفاصل	فعالیت
۱	N/A	✓	✓	تکان‌دادن دست، کف‌زدن
۲	N/A	✓	×	نوشیدن از یک بطری، پاسخ‌دادن به تلفن، خواندن ساعت مچی
۳	×	×	N/A	بستن بند کفش، نشستن
۴	✓	×	N/A	ایستادن و تعظیم‌کردن، دست‌تکان‌دادن و کف زدن

سریع‌تر می‌شود، در حالی که زمان آزمایش تقریباً یکسان باقی می‌ماند.

شکل ۵ ماتریس درهم‌ریختگی[1] ۶۰- CAD را در روش پیشنهادشده نشان می‌دهد وقتی که تعداد حالت‌های پنهان ۲۴ است، با این حال از آنجایی که ما به طور خودکار داده‌ها را به چهار گروه تقسیم می‌کنیم، می‌توانیم تعداد مختلفی از حالات پنهان برای هر دسته را انتخاب کنیم.

شکل ۷ ماتریس درهم‌ریختگی برای آزمایش روی مجموعه داده ۳D Florence را نشان می‌دهد. همان طور که در ماتریس درهم‌ریختگی دیده می‌شود، عمل "نوشیدن از بطری" با "پاسخ‌دادن به تلفن" اشتباه شده است، زیرا در هر دو فعالیت دست بالا رفته و نزدیک صورت قرار می‌گیرد. گاهی اوقات نیز "خواندن ساعت مچی" به عنوان "پاسخ‌دادن به تلفن" شناخته می‌شود. این مسئله زمانی اتفاق می‌افتد که دست شخص برای خواندن ساعت خیلی به صورتش نزدیک می‌شود. این امر به این دلیل اتفاق می‌افتد که روش ما مبتنی بر توالی از ژست هستند و اگر ژست و توالی آنها مشابه باشد، چنین خطاهایی رخ می‌دهند. باید اضافه کنیم که اسکلت‌هایی که به دست می‌آیند بر اساس داده‌های عمق هستند در نتیجه آنها خیلی دقیق نیستند. همچنین گروه‌های فعالیت "پاسخ‌دادن به تلفن"، "مسواک‌زدن" و "آشپزی (سرخ‌کردن)"، "آشپزی (خردکردن)" گاهی در مجموعه داده‌های ۶۰- CAD به دلیل مشابه اشتباه می‌شوند، همان طور که در شکل ۵ نشان داده شده است.

صحت[2] (Acc)، دقت[3] (Pre) و بازخوانی[4] (Rec) تشخیص آزمایش‌ها روی ۶۰- CAD در مقایسه با چندین روش دیگر در جدول ۴ نشان داده شده است. نتایج مربوط به ۶۰- CAD میانگین نتایج در ۴ دور LOOCV

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall

جدول ۴: نتایج روش پیشنهادی بر روی مجموعه داده CAD −۶۰ و مقایسه آن با سایر کارها.

خلاصه روش	بازخوانی	دقت	صحت	مقاله
DBN	۵۵٫۵	۶۷٫۹	–	[۹] Sung ۲۰۱۲
MRF	۷۱٫۴	۸۰٫۸	–	[۱۱] Koppula ۲۰۱۲
BOW + SVM	۸۴٫۰	۸۶٫۰	–	[۵۱] Zhang ۲۰۱۲
Eigen joints	۶۶٫۶	۷۱٫۹	–	[۵۲] Yang ۲۰۱۳
GMM + HMM	۷۸٫۰	۷۰٫۰	–	[۵۳] Piyathilaka ۲۰۱۳
Multilevel depth & image fusion	۶۹٫۵	۷۵٫۹	–	[۱۶] Ni ۲۰۱۳
Codewords + Ensemble	۷۵٫۴	۷۸٫۱	–	[۵۴] Gupta ۲۰۱۳
Fourier Temporal Pyramid	–	–	۷۴٫۷	[۵۵] Wang ۲۰۱۴
STIP + skeleton	۸۴٫۶	۹۳٫۲	–	[۳۰] Zhu ۲۰۱۴
Dynamic Bayesian Mixture	۹۱٫۹	۹۱٫۱	–	[۳۷] Faria ۲۰۱۴
Keypose, HMM, Random Forest	۹۴٫۵	۹۳٫۸	–	[۴۱] Shan ۲۰۱۴
SVM, HMM	۷۶٫۶	۷۷٫۳	–	[۵۶] Gaglio ۲۰۱۴
Self-Organizing Neural	۹۰٫۲	۹۱٫۹	–	[۳۵] Parisi ۲۰۱۵
Atomic Motion, Naïve Bayes	۹۴٫۸	۹۴٫۶	–	[۳۳] Zhu ۲۰۱۶
PRNN	–	–	۸۷٫۶	[۲۵] Shi ۲۰۱۷
SSP	۹۷٫۷	۹۷٫۶	۹۷٫۶	روش پیشنهادی

جدول ۵: نتایج روش پیشنهادی بر روی مجموعه داده UT-Kinect و مقایسه آن با سایر کارها.

خلاصه روش	صحت	مقاله
Gram Matrix on Riemannian Manifolds	۱۰۰	[۵۰] Zhang ۲۰۱۶
Multimodal Feature Fusion	۱۰۰	[۴۹] Ye ۲۰۱۵
Graph kernel	۹۷٫۴۴	[۴۰] Wang ۲۰۱۶
Dictionary of soft-quantization	۹۳٫۴۷	[۵۷] Wang ۲۰۱۶
Body parts model	۹۷٫۰۸	[۳۲] Vemulapalli ۲۰۱۴
RBF Kernel SVM	۹۷٫۵۹	[۵۸] Vemulapalli ۲۰۱۶
Grassmannian manifold	۹۵٫۲۵	[۲۹] Slama ۲۰۱۴
LSTM	۹۷٫۰۰	[۲۲] Liu ۲۰۱۶
Kernel Linearization	۹۸٫۲۰	[۳۶] Koniusz ۲۰۱۶
TriViews probabilistic fusion approach	۹۸٫۰۰	[۲۷] Chen ۲۰۱۵
Action-Snippets and Activated Simplice	۹۶٫۴۸	[۱۹] Wang ۲۰۱۶
Discriminative Activated Simplice	۹۷٫۹۹	[۵۹] Luo ۲۰۱۷
GCA-LSTM network	۹۹٫۰۰	[۲۶] Liu ۲۰۱۷
SSP	۱۰۰	روش پیشنهادی

جدول ۶: نتایج روش پیشنهادی بر روی مجموعه داده Florence ۳D و مقایسه آن با سایر کارها.

خلاصه روش	صحت	مقاله
Graph kernels	۹۱٫۶۳	[۴۰] Wang ۲۰۱۶
RBF Kernel SVM	۹۳٫۰۶	[۵۸] Vemulapalli ۲۰۱۶
Riemannian Trajectories	۸۹٫۶۷	[۲۰] Anirudh ۲۰۱۷
dictionary of soft-quantization	۹۲٫۲۵	[۵۷] Wang ۲۰۱۶
Kernel Linearization	۹۵٫۲۳	[۳۶] Koniusz ۲۰۱۶
Manifold	۸۷٫۰۴	[۲۱] Devanne ۲۰۱۵
Activated Simplices	۹۴٫۲۵	[۱۹] Wang ۲۰۱۶
Discriminative Activated Simplices	۹۵٫۳۰	[۵۹] Luo ۲۰۱۷
SSP	۹۶٫۱۱	روش پیشنهادی

جدول ۷: مقایسه روش‌های مختلف خوشه‌بندی به عنوان مقداردهی اولیه متغیرهای پنهان در صحت بازشناسی برای مجموعه داده‌های CAD −۶۰، UT-Kinect و Florence ۳D.

	CAD −۶۰	Florence ۳D	UT-Kinect
K-means	۹۱٫۶۷	۹۳٫۳۴	۹۹٫۰۰
K-medoids	۹۷٫۶	۹۶٫۱۱	۱۰۰
Guassian Mixture	۹۳٫۳۳	۹۰٫۹۰	۹۱٫۰۰

مختلف گزارش شده است. در هر ردیف، صحت، دقت و بازخوانی را برای روش‌ها بر اساس نحوه گزارش آنها در مقالات مربوط مقایسه کرده‌ایم.

به طور مشابه، در جداول ۵ و ۶ ما صحت تشخیص (Acc) را با روش‌های دیگر به ترتیب در مجموعه داده‌های UT-Kinect و Florence ۳D مقایسه می‌کنیم. در جدول ۵ همان طور که دیده می‌شود، [۴۹] و [۵۰] دقت ۱۰۰٪ را به دست آورده‌اند اما از محاسبات بسیار پیچیده‌ای روی داده‌ها استفاده کرده و داده‌های ورودی را به فضاهای دیگر منتقل می‌کنند، در حالی که ما از ژست ساده و ویژگی‌های اسکلت استفاده کردیم. این نشان می‌دهد که مدل ما بسیار قدرتمند است و نیازی به عملیات پیچیده بر روی داده‌ها ندارد. به علاوه، لیو و سایرین [۲۲] از مدل حافظه بلند کوتاه‌مدت (LSTM)[1] استفاده می‌کند که برترین روش در بسیاری از مسایل است، اما دقت کمتری نسبت به روش ما در این مجموعه داده به دست می‌آورد. علت آن این است که LSTM اغلب به مقدار زیادی از داده نیاز دارد تا به درستی آموزش داده شود در حالی که مدل ما به طور مؤثر حتی با مقدار کم از داده به خوبی آموزش می‌بینند.

در جدول ۷ مقایسه عملکرد روش‌های خوشه‌بندی به منظور مقداردهی اولیه متغیرهای پنهان نشان داده شده است. همان طور که مشخص است، روش خوشه‌بندی تأثیر زیادی در نتایج دارد. روش K-medoids از سایرین نتیجه بهتری را به همراه داشته است زیرا این روش نسبت به نویز مقاومت بیشتری دارد. همچنین روش Guassian Mixture نسبت به K-means در دیتاست CAD −۶۰ نتیجه بهتری را به همراه داشته است. علت این امر این است که روش Guassian Mixture قدرت بیشتری در خوشه‌بندی دارد با این حال در دیتاست‌های Florence ۳D و

1. Long Short-Term Memory

UT-Kinect روش K-means دقت بالاتری را نسبت به روش Guassian Mixture داشته است، زیرا اسکلت‌های مورد استفاده در این دیتاست‌ها بسیار نویزی ضبط شده‌اند و روش Guassian Mixture در برابر نویز نسبت به روش K-means حساس‌تر است.

برای مجموعه داده Florence ۳D، ۳۳ بهترین تعداد حالات مخفی است. برای مجموعه داده CAD- ۶۰، بهترین تعداد حالات پنهان ۲۴ برای ماژول ۱ و ۱۷ برای ماژول ۳ است. برای مجموعه داده UT-Kinect، روش پیشنهادی ما دقت ۱۰۰٪ را هنگامی که تعداد حالات مخفی ۵ است به دست آورد. این نکته مهم است که در اینجا ما از متغیرهای پنهان برای خوشه‌بندی اسکلت‌ها استفاده می‌کنیم، به این صورت که اسکلت‌ها با ژست‌های مشابه در یک گروه قرار خواهند گرفت. در نتیجه، زمانی که تعداد زیادی از ژست‌های مختلف در اسکلت درون مجموعه داده وجود دارد، تعداد متغیرهای پنهان (یعنی تعداد خوشه‌ها) افزایش می‌یابد. این امر به طور مستقیم به یادگیری دسته‌های مجموعه داده‌های فعالیت پیچیده (مانند CAD- ۶۰) کمک می‌کند. با این حال به منظور بازشناسی فعالیت‌های مجموعه داده‌های Florence ۳D نیاز به تعداد زیادی از حالات پنهان دارد، در حالی که فعالیت‌های موجود در این مجموعه داده پیچیده نیستند. دلیل آن این است که اجراکننده‌های این مجموعه داده فعالیت‌ها را به شیوه‌های مختلف و غیر یکسان انجام داده‌اند. علاوه بر این به نظر می‌رسد این مجموعه داده نویزی‌تر از بقیه است به همین دلیل این اسکلت‌ها به راحتی با یکدیگر اشتباه می‌شوند. این واقعیت‌ها نیز در نتایج قابل مشاهده هستند، چون عملکرد دو مجموعه داده UT-Kinect و Florence ۳D کاملاً متفاوت هستند.

همان طور که فعالیت‌های مورد استفاده در CAD- ۶۰ تنها وابسته به دست هستند، ما تنها از مفاصل دست برای بازشناسی فعالیت‌ها استفاده کردیم. الگوریتم‌های ما با استفاده از تکنیک‌های محاسباتی پردازش موازی [۳] اجرا می‌شوند. در طول مرحله آزمایش، ما می‌توانیم بیش از ۳۰۰ فریم بر ثانیه (fps) را در یک پردازشگر Core i۷ ۳۶۳۲QM، با ۵ کلاس فعالیت که برای رباتیک بسیار مناسب است، پردازش کنیم. ما می‌توانیم سرعت پردازش را با استفاده از گره‌های پردازشی بیشتر افزایش دهیم. الگوریتم ما مقیاس‌پذیر است و در نتیجه می‌تواند به سادگی بر روی پلتفرم‌های محاسبات ابری اجرا شود که مورد توجه بسیاری برای کاربردهای روباتیک است.

۵- نتیجه‌گیری

در این مقاله، ما مسئله بازشناسی فعالیت انسانی را در یک چارچوب برچسب‌گذاری توالی فرموله کردیم و ساختارهای مدل گرافی احتمالی جدید را برای بازشناسی فعالیت‌های انسانی ساده و پیچیده (شامل بازه کوتاه و طولانی) از داده‌های اسکلت پیشنهاد کردیم. ما یک مدل را پیشنهاد دادیم و آن را به عنوان یک مدل پیش‌بینی ساختاریافته با وضعیت‌های پنهان مدل‌سازی کرده و پارامترهای مدل را با استفاده از چارچوب پیش‌بینی ساختاریافته توزیع‌شده یاد گرفتیم. همچنین روش تعویض- رده را به منظور سرعت‌بخشیدن به روند آموزش و افزایش دقت پیشنهاد دادیم. بر اساس آزمایش‌ها روی سه مجموعه داده با ویژگی‌های مختلف (شامل دو نوع فعالیت‌های ساده و پیچیده)، ما نتایج بسیار خوبی را به دست آوردیم. همچنین تأثیر روش‌های خوشه‌بندی بر مقداردهی اولیه متغیرهای پنهان مورد بررسی قرار گرفت.

مدل‌های گرافی احتمالی (مانند روش ما) ابزاری قدرتمند در بیان روابط بین متغیرها می‌باشند، با این حال این روش‌ها به خصوص زمانی که تعداد متغیرها زیاد است، نیازمند محاسبات سنگینی می‌باشند. به علاوه مشکل دیگر این روش‌ها نیاز به استفاده از ویژگی‌های مهندسی‌شده است، مانند ویژگی اسکلت که در کار ما استفاده شد. اسکلت توصیف سطح بالا از بدن انسان می‌باشد و در مسایل بازشناسی فعالیت انسانی مانند کار ما مفید است ولی قابل استفاده در بسیاری از مسایل دیگر نیست پس به نوعی روش ما وابسته به تعریف دستی ویژگی‌ها است. یادگیری عمیق می‌تواند در این دست مسایل به عنوان راه حلی برای استخراج خودکار ویژگی‌ها مورد استفاده قرار گیرد.

مراجع

[1] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, *Efficient Structured Prediction with Latent Variables for General Graphical Models*, arXiv preprint arXiv:1206.6436, 2012.

[2] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Distributed structured prediction for big data," in *Proc. NIPS Workshop on Big Learning*, 5 pp., 2012.

[3] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Distributed message passing for large scale graphical models," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'11,* pp. 1833-1840, Providence, RI, USA, 20-25 Jun. 2011.

[4] J. Piger, "Econometrics: models of regime changes," *Complex Systems in Finance and Econometrics*, pp. 190-202, Jul. 2009.

[5] H. Tong, *Threshold Models in Non-Linear Time Series Analysis*, Lecture Notes in Statistics, vol. 21, Springer Science & Business Media, 2012.

[6] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica: J. of the Econometric Society*, vol. 57, no. 2, pp. 357-384, Mar. 1989.

[7] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: a review," *Computer Vision and Image Understanding*, vol. 158, pp. 85-105, May 2017.

[8] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys,* vol. 43, no. 3, Article No. 16, 43 pp., Apr. 2011.

[9] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. on Robotics and Automation, ICRA'12,* pp. 842-849, Saint Paul, MN, USA, 14-18 May. 2012.

[10] N. Hu, G. Englebienne, Z. Lou, and B. Krose, "Learning to recognize human activities using soft labels," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 1973-1984, Oct. 2016.

[11] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *The International J. of Robotics Research*, vol. 32, no. 8, pp. 951-970, Jul. 2013.

[12] M. M. Arzani, *et al*., "Structured prediction with short/long-range dependencies for human activity recognition from depth skeleton data," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'17*, pp. 560-567, Vancouver, BC, Canada, 24-28 Sept. 2017.

[13] C. N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *Proc. of the ACM 26th Annual Int. Conf. on Machine Learning,* pp. 1169-1176, Montreal,Canada, Jun. 2009.

[14] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori, "Similarity constrained latent support vector machine: an application to weakly supervised action classification," *Computer Vision-ECCV*, vol. 7578, pp. 55-68, Oct. 2012.

[15] M. Khodabandeh, *et al*., "Discovering human interactions in videos with limited data labeling," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 9-18, Boston, MA, USA, 7- 12 Jun. 2015.

[16] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Trans. on Cybernetics*, vol. 43, no. 5, pp. 1383-1394, Aug. 2013.

[17] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549-1562, Dec. 2012.

[18] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a Riemannian manifold," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4498-4507, Las Vegas, NV, USA, 27- 30 Jun. 2016.

[19] C. Wang, J. Flynn, Y. Wang, and A. L. Yuille, "Recognizing actions in 3D using action-snippets and activated simplices," in *Proc. 31st. AAAI Conf. on Artificial Intelligence*, *AAAI'16,* pp. 3604-3610, Phoenix, USA, 12-17 Feb. 2016.
[20] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of riemannian trajectories," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 922-936, May 2017.
[21] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE Trans. on Cybernetics*, vol. 45, no. 7, pp. 1340-1352, Sept. 2015.
[22] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Proc. European Conf. on Computer Vision*, *ECCV'16*, pp. 816-833, Amsterdam, The Netherlands, 8-16 Oct. 2016.
[23] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 1, pp. 1799-1807, Dec. 2014.
[24] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: deep learning on spatio-temporal graphs," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5308-5317, Las Vegas , NV, USA, 27-30 Jun. 2016.
[25] Z. Shi and T. K. Kim, Learning and Refining of Privileged Information-Based RNNs for Action Recognition from Depth Sequences, arXiv preprint arXiv:1703.09625, 2017.
[26] J. Liu, G. Wang, L. Y. Duan, P. Hu, and A. C. Kot, *Skeleton Based Human Action Recognition with Global Context-Aware Attention LSTM Networks*, arXiv preprint arXiv:1707.05740, 2017.
[27] W. Chen and G. Guo, "Triviews: a general framework to use 3d depth data effectively for action recognition," *J. of Visual Communication and Image Representation*, vol. 26, pp. 182-191, Jan. 2015.
[28] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Proc. Asian Conf. on Computer Vision*, pp. 428-443, Singapore, Singapore, 1-5 Nov. 2014.
[29] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3d human gesture and action recognition," in *Proc. 22nd IEEE Int. Conf. on Pattern Recognition, ICPR'14,* pp. 3499-3504, Stockholm, Sweden, 24-28 Aug. 2014.
[30] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image and Vision Computing*, vol. 32, no. 8, pp. 453-464, Aug. 2014.
[31] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for RGB-D action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1054-1062, Boston, MA, USA, 8-10 Jun. 2015.
[32] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 588-595, Columbus, Ohio, USA, 24-27 Jun. 2014.
[33] G. Zhu, L. Zhang, P. Shen, and J. Song, "Human action recognition using multi-layer codebooks of key poses and atomic motions," *Signal Processing: Image Communication*, vol. 42, pp. 19-30, Mar. 2016.
[34] B. Ni, P. Moulin, and S. Yan, "Order-preserving sparse coding for sequence classification," in *Proc. European Conf. on Computer Vision, ECCV'12*, pp. 173-187, Firenze, Italy , 7-13 Oct. 2012.
[35] G. I. Parisi, C. Weber, and S. Wermter, "Self-organizing neural integration of pose-motion features for human action recognition," *Frontiers in Neurorobotics*, vol. 9, 3 pp., 2015.
[36] P. Koniusz, A. Cherian, and F. Porikli, Tensor Representations via Kernel Linearization for Action Recognition from 3D Skeletons (Extended Version), arXiv preprint arXiv:1604.00239, 2016.
[37] D. R. Faria, C. Premebida, and U. Nunes, "A probabilistic approach for human everyday activities recognition using body motion from rgb-d images," *in Proc. 23rd IEEE Int. Symp. on Robot and Human Interactive Communication, RO-MAN'14*, pp. 732-737, Edinburgh, UK, 25-29 Aug. 2014.
[38] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116-124, Jun. 2013.
[39] A. Manzi, P. Dario, and F. Cavallo, "A human activity recognition system based on dynamic clustering of skeleton data," *Sensors*, vol. 17, no. 5, p. 1100, May 2017.
[40] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, "Graph based skeleton motion representation and similarity measurement for action recognition," in *Proc. European Conf. on Computer Vision*, *ECCV'16*, pp. 370-385, Amsterdam, The Netherlands, 8-16 Oct. 2016.
[41] J. Shan and S. Akella, "3D human action segmentation and recognition using pose kinetic energy," in *Proc. IEEE Workshop on Advanced Robotics and Its Social Impacts, ARSO'14,* pp. 69-75, Evanston, IL, USA, 11-13 Sept. 2014.
[42] M. I. Jordan and Y. Weiss, *Probabilistic Inference in Graphical Models*, Handbook of Neural Networks and Brain Theory, 2002.
[43] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848-1852, Oct. 2007.
[44] S. Nowozin, C. H. Lampert, *et al.*, "Structured learning and prediction in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 6, no. 3-4, pp. 185-365, May 2011.
[45] T. Hazan and R. Urtasun, "A primal-dual message-passing algorithm for approximated large scale structured prediction," in *Proc. of the 23rd In. Conf. on Neural Information Processing Systems*, *NIPS'10,* vol. 1, pp. 838-846, Dec. 2010.
[46] H. Tong, Non-Linear Time Series: A Dynamical System Approach, Oxford University Pres, 1990.
[47] L. Xia, C. C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops, CVPRW'12,* pp. 20-27, Rhode Island, USA, 18-20 Jun. 2012.
[48] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, *CVPRW'13,* pp. 479-485, Portland, ON, USA, 23-24 Jun. 2013.
[49] J. Ye, K. Li, G. J. Qi, and K. A. Hua, "Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams," in *Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval*, pp. 99-106, Shanghai, China, 23-26 Jun. 2015.
[50] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a Riemannian manifold," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4498-4507, Las Vegas, NV, USA, 27-30 Jun 2016.
[51] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," *J. of Computer Vision and Image Processing*, vol. 2, no. 4, p. 12, Dec. 2012.
[52] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *J. of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2-11, Jan. 2014.
[53] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features," in *Proc. 8th IEEE Conf. on Industrial Electronics and Applications, ICIEA'13,* , pp. 567-572, Melbourne, Australia, 19-21 Jun. 2013.
[54] R. Gupta, A. Y. S. Chia, and D. Rajan, "Human activities recognition using depth images," in *Proc. of the 21st ACM Int. Conf. on Multimedia*, pp. 283-292, Barcelona, Spain, 21-23 Oct. 2013.
[55] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914 - 927, May 2014.
[56] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-d posture data," *IEEE Trans. on Human-Machine Systems*, vol. 45, no. 5, pp. 586-597, Dec. 2015.
[57] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3d key-pose-motifs for action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2639-2647, Las Vegas, NV, USA, 27-30 Jun. 2016.
[58] R. Vemulapalli, F. Arrate, and R. Chellappa, "R3dg features: relative 3d geometry-based skeletal representations for human action recognition," *Computer Vision and Image Understanding*, vol. 152, pp. 155-166, Nov. 2016.
[59] C. Luo, C. Ma, C. Y. Wang, and Y. Wang, "Learning discriminative activated simplices for action recognition," in *Proc. 32st. AAAI Conf. on Artificial Intelligence*, *AAAI'17,* pp. 4211-4217, Feb. 2017.

محمد مهدی ارزانی در سال ۱۳۸۵ در رشته مهندسی کامپیوتر گرایش سخت افزار کارشناسی خود را از دانشگاه شاهد تهران دریافت کرد. سپس وی در سال ۱۳۸۷ مدرک کارشناسی ارشد در گرایش هوش مصنوعی خود را از دانشگاه صنعتی شریف اخذ کرد. در حال حاضر وی دانشجوی دکتری دانشگاه علم و صنعت در رشته هوش مصنوعی

می‌باشد. زمینه‌های تحقیقاتی مورد علاقه وی عبارت‌اند از: پردازش تصویر و ویدیو، مدل‌های احتمالی، یادگیری عمیق و پردازش تصاویر RGBD.

محمود فتحی در ۱۳۶۳ کارشناسی الکترونیک خود را از دانشگاه علم و صنعت ایران دریافت نمود. وی کارشناسی ارشد معماری کامپیوتر خود را از دانشگاه برادفورد انگلستان در ۱۳۶۵ دریافت کرد و در ۱۳۷۰ به درجهٔ دکتری معماری کامپیوتر در پردازش تصویر از دانشگاه UMIST انگلستان رسید. ایشان از سال ۱۳۷۰ عضو هیأت علمی دانشکدهٔ کامپیوتر دانشگاه علم و صنعت ایران است. از زمینه‌های مورد علاقهٔ او پردازش برخط تصاویر و یادگیری عمیق و کاربرد آن در مهندسی ترافیک است. همچنین در زمینهٔ QoS در شبکه‌های کامپیوتری در انتقال ویدیو و سیگنال روی خطوط اینترنت در حال تحقیق است.

احمد اکبری ازیرانی کارشناسی و کارشناسی ارشد خود را در رشته الکترونیک به ترتیب در سال‌های ۱۳۶۵ و ۱۳۶۸ از دانشگاه صنعتی اصفهان دریافت کرد. وی در سال‌های ۱۳۷۱ و ۱۳۷۴ در رشته پردازش سیگنال و ارتباطات از دانشگاه رنس فرانسه مدرک کارشناسی ارشد و دکتری خود را اخذ کرد. وی از سال ۱۳۷۵ عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران است. موضوعات تحقیقاتی مورد علاقه وی شبکه‌های کامپیوتری، امنیت شبکه، پردازش صوت، بازشناسی گفتار، به‌سازی گفتار، پیاده‌سازی الگوریتم‌های پردازش سیگنال می‌باشند.