نوع مقاله: پژوهشی

بازشناسی رفتار حرکتی بین دو فرد با استفاده از روش طبقه‌بندی ترتیبی دومرحله‌ای

سامان نیکزاد[۱]، افشین ابراهیمی[۲]

۱- گروه مخابرات - دانشکده مهندسی برق - دانشگاه صنعتی سهند - تبریز - ایران – s_nikzad@sut.ac.ir
۲- گروه مخابرات - دانشکده مهندسی برق - دانشگاه صنعتی سهند - تبریز - ایران – aebrahimi@sut.ac.ir

چکیده: بازشناسی رفتار حرکتی انسان یکی از مهم‌ترین کاربردهای پردازش تصویر است که جذابیت بسیار زیادی در حوزه‌ی بینایی ماشین پیداکرده است. به دلیل پیچیدگی در تجزیه‌وتحلیل تصویر و ویدئو، روش‌های پیشنهادی به‌روز کماکان از توسعه‌ی روش‌های بدون خطا و کاملاً تعمیم‌یافته فاصله دارند. این مقاله یک روش جدید مبتنی بر حالت‌های کلیدی سایه‌ی قاب به‌منظور تشخیص تعامل دو فرد ارائه می‌دهد. در روش پیشنهادی، از یک توصیفگر شکل مبتنی بر IMTF استفاده شده است که به دلیل توانایی این روش در جمع‌آوری داده‌ها از کلیت تصویر، توصیف تمام عیاری از کل شکل ارائه می‌دهد. این ویژگی برای استفاده در برنامه‌های برخط یا برون‌خط مانند تجزیه‌وتحلیل محتوای ویدئویی و نظارت ویدئویی، کارایی بسیار بالایی دارد. ایده‌ی اصلی بر این اساس است که یک طبقه‌بندی دومرحله‌ای بر پایه‌ی طبقه‌بند الگوکاوی (الگوبرداری) ترتیبی انجام گیرد؛ بنابراین، شکل نمای پیش‌زمینه افراد استخراج شده و با استفاده از ویژگی مثلث چند مقیاسی ثابت، هر قاب با یک واژه‌نامه‌ی تصویری از پیش تعریف‌شده از حالت‌های کلیدی مقایسه می‌شود. پس از این مرحله، هر قاب به‌عنوان یکی از طبقه‌های موجود برچسب‌گذاری می‌شود. خروجی این مرحله، یک توالی از برچسب‌ها است که به‌عنوان ورودی یک طبقه‌بند الگوی ترتیبی استفاده خواهد شد. نتایج به‌دست‌آمده (۹۳/۹٪ برای پایگاه داده‌ی SBU، ۹۲/۴٪ برای پایگاه داده‌ی UT، ۸۶/۶٪ برای پایگاه داده‌ی BIT و ۸۶/۹٪ برای پایگاه داده‌ی NTU)، بیان‌گر دقت و کارآمدی روش پیشنهادی بوده و راه‌حلی جذاب برای مسئله طبقه‌بندی توالی ارائه می‌دهد.

واژه‌های کلیدی: بازشناسی تعامل، سایه‌ی پیش‌زمینه، حالت کلیدی، طبقه‌بند الگوی ترتیبی، الگوکاوی

Recognition of two-person interaction using two-stage sequential pattern classification method

Saman Nikzad[1], Afshin Ebrahimi[2]

1- Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran, Email: s_nikzad@sut.ac.ir
2- Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran, Email: aebrahimi@sut.ac.ir

Abstract: Human action recognition is one of the most important applications of image processing which has gained great attraction in computer vision. Because of the complexity in image and video analysis, state-of-the-art proposed algorithms are still far from developing error-free and fully generalized systems. This work presents a new method based on key-poses of frame silhouettes for the task of two-person interaction recognition. In the proposed algorithm, an IMTF-based shape descriptor which gives a perfect description of the shape due to its ability to collect data from the whole shape. The task is useful for offline or online applications such as video content analysis and video surveillance. The main idea is to perform a two-step classification based on a sequential pattern mining classifier. So, we extract the shape of the foreground silhouette of the persons and describe it using the invariant multiscale triangle feature to compare each frame with a pre-defined dictionary of key-poses. Then, each frame is labeled as one of the existing classes. The output of this step is a sequence of labels which is the input of a sequential pattern classifier. The obtained results (93.9% for SBU database, 92.4% for UT database, 86.6% for BIT dataset and 86.9% for NTU RGB+D dataset) indicate the accuracy and efficiency of the proposed method and provide an attractive solution to the problem of sequence classification.

Keywords: Interaction recognition, foreground silhouette, key-pose, sequential pattern classifier, pattern mining.

تاریخ ارسال مقاله: ۱۴۰۱/۰۶/۲۰
تاریخ اصلاح مقاله: ۱۴۰۲/۰۷/۰۱
تاریخ پذیرش مقاله: ۱۴۰۲/۰۷/۰۱
نام نویسنده مسئول: افشین ابراهیمی
نشانی نویسنده مسئول: گروه مخابرات - دانشکده مهندسی برق - دانشگاه صنعتی سهند - تبریز - ایران.

۱- مقدمه

برای بازشناسی تعامل[1] بین دو فرد در محیط واقعی، کاربردهایی در حوزه‌های مختلف ازجمله تجزیه‌وتحلیل ورزشی [۱، ۲]، خودکارسازی [۳]، خودکارسازی ماشین [۴]، تشخیص خشونت [۵]، تعامل انسان-ماشین [۶] و ... تعریف می‌شوند. تفاوت‌های درون طبقه‌ای و شباهت‌های بین طبقه‌ای، دو عامل اصلی پیچیده‌شدن بازشناسی تعامل هستند. به‌عنوان‌مثال، شباهت بین عمل «آهسته دویدن» و «سریع دویدن»، نمونه‌ای معمولی از شباهت بین طبقه‌ای میان دو طبقه‌ی بسیار مهم در دنیای واقعی است. از سویی دیگر، تنوع درون طبقه‌ای به این معنی است که انجام یک عمل توسط دو فرد مختلف، به روش‌های متفاوتی صورت می‌پذیرد. این مشکل، زمانی بیشتر قابل‌درک می‌شود که بخواهیم رفتارهایی را بشناسیم که متشکل از عمل‌های متفاوتی است که توسط بازیگران متفاوت انجام می‌شود؛ زیرا اکثر ویدیوهای دنیای واقعی، در مکان‌های شلوغ و به‌عنوان ویدیوهای نظارتی ضبط می‌شوند. برای پرداختن به مشکل تفاوت‌های درون طبقه‌ای، سامانه باید بر روی یک مجموعه داده با چندین فرد عملگر آموزش ببیند. پایگاه‌های داده‌ی رایج که به‌طورمعمول از دو فرد عملگر استفاده می کنند ممکن است ابزار مناسبی برای رسیدن به یک راه‌حل جامع نباشند. علاوه بر این، بیشتر پایگاه‌های داده، ویدیوها را از نمای جلو ثبت می‌کنند؛ بنابراین هر دو دست بدون هیچ‌گونه همپوشانی دیده می شوند و این امر تشخیص عمل انجام‌شده را آسان می‌کند. این درحالی‌است‌که در دنباله‌های بازشناسی تعامل، فقط نیمی از بدن فرد عملگر قابل‌رؤیت بوده و طرف دیگر به دلیل زاویه دید، پنهان می‌شود.

هدف اصلی این مطالعه، ارائه‌ی یک طبقه‌بند دومرحله‌ای مبتنی بر الگوکاوی ترتیبی و استخراج اطلاعات مهم به‌منظور عمق بخشیدن و درک اهمیت استفاده از طبقه‌بندهای پیچیده بر اساس طبقه‌بندهای ساده‌تر است. در این مقاله، با الهام از مدل سایه‌ی دوطرفه[2] [۷]، یک توصیفگر ویژگی مثلث چند مقیاسی نامتغیر IMTF[3] [۸] پیشنهاد می شود تا ثمربخش‌ترین ویژگی‌ها به‌منظور استفاده در طبقه‌بند دومرحله‌ای استخراج شوند. نمونه‌ای از یک سایه‌ی پیش‌زمینه در شکل ۱ نشان داده شده است که به تعامل مشت‌زنی بین دو فرد اشاره دارد. پس از استخراج سایه‌ی پیش‌زمینه توسط فرآیند استخراج ویژگی، شکل به‌دست‌آمده با استفاده از توصیفگر IMTF، توصیف می‌شود. ویژگی‌های استخراج‌شده، در یک طبقه‌بند دومرحله‌ای که متشکل از یک طبقه‌بند قاب و یک طبقه‌بند توالی است، پردازش می‌شوند. هر قاب با الگوهای از پیش تعریف‌شده از واژه‌نامه‌ی تصویری، مقایسه شده و سپس به‌عنوان یکی از دسته‌های از پیش تعریف‌شده، برچسب‌گذاری می‌شود. این روش، بر روی تمام قاب‌های متوالی اعمال می‌شود و درنهایت، یک توالی از برچسب‌ها تولیدشده و نتیجه به یک طبقه‌بند توالی اعمال می‌شود که از روش الگوکاوی ترتیبی برای بازتولید برچسب‌های طبقه به‌عنوان خروجی طبقه‌بند نهایی استفاده می‌کند. طبقه‌بند مرحله دوم به روش کمک می‌کند تا مشکلات معمول مانند خطای استخراج سایه‌ی پیش‌زمینه و خطاهای طبقه‌بندی قاب را رفع کند. برای استخراج ویژگی‌ها از سایه‌ی پیش‌زمینه، یک شکل دودویی دوبعدی به جای اشکال سه‌بعدی یا رنگی تحلیل می‌شود؛ بنابراین، بدیهی است که پیچیدگی محاسباتی کم، یکی از مزایای کلیدی استفاده از سایه‌ی پیش‌زمینه است. این در حالی است که اکثر مطالعات سطح بالا در بازشناسی تعامل، بر ویژگی‌های پیچیده‌ای مانند HoF[4]، HoG[5] و STIP[6] ها تکیه دارند که معمولاً روش را با پیچیدگی محاسباتی بالایی مواجه می‌کنند. به‌منظور ایجاد توازن بین دقت و سرعت عملکرد، تعیین دقیق برخی مؤلفه‌ها مانند اندازه واژه‌نامه‌ی تصویری[7]، و ضوح تصویر و حداقل مقدار مؤلفه *minsup* بسیار حیاتی خواهد بود؛ بنابراین، اگر در حالت برون‌خط، به دقت خوبی نیاز داشته باشیم، می‌توانیم تعداد الگوهای واژه‌نامه‌ی تصویری را افزایش دهیم تا بهترین پیش‌بینی از برچسب‌های طبقه را بدون هیچ‌گونه نگرانی در مورد زمان صرف‌شده ارائه دهیم.

۲- مطالعات انجام شده

فعالیت‌های انسانی را می‌توان از نظر مفهومی و بر اساس پیچیدگی، به چهار دسته مختلف طبقه‌بندی کرد: حالت‌ها[8]، کنش‌ها، تعاملات (اندرکنش‌ها) و فعالیت‌های گروهی [۹]. همچنین می‌توان یک حالت را به عنوان حرکتی از یک بخش خاص بدن مانند بالا بردن دست تعریف نمود. یک عمل یا کنش می‌تواند به‌عنوان ترکیبی از یک یا چند حالت توصیف شود. تعامل یا اندرکنش شامل دو فرد است که یک عمل را با یکدیگر انجام می‌دهند و درنهایت، فعالیت‌های گروهی، فعالیت‌هایی هستند که توسط گروهی انجام می‌شوند که از چندین فرد تشکیل شده‌اند. علاوه‌براین، از نقطه‌نظر نوع داده‌ها، می‌توان پایگاه‌های داده را به دو نوع واقعی و مصنوعی دسته‌بندی کرد. پایگاه‌های داده‌ی مصنوعی برای اهداف خاصی آماده می‌شوند، درحالی‌که مجموعه داده‌های واقعی از طریق اینترنت یا فیلم‌ها جمع‌آوری می‌شوند. مشکل اساسی پایگاه‌های داده‌ی مصنوعی این است که نمی‌توان آن‌ها را برای شرایط مختلف آزمایش تعمیم داد. بااین‌حال، پیاده‌سازی روش روی پایگاه داده واقعی، تنگناهایی همچون وجود حرکت دوربین و انسداد تصویر را

[1] Interaction

[2] Bilateral silhouette

[1] Invariant multiscale triangle feature

[4] Histogram of optical flow

[5] Histogram of oriented gradients

[6] Spatio-temporal interest points

[7] Visual vocabulary

[8] Poses

خواهد داشت. بر این اساس، کار با مجموعه داده‌های واقعی می‌تواند پیچیده‌تر و چالش‌برانگیزتر از مجموعه داده‌های مصنوعی باشد.

شکل ۱: نمونه‌ای از استخراج سایه‌ی پیش‌زمینه مربوط به تعامل مشت زدن

بر پایه‌ی [۱۰]، می‌توانیم عمل بازشناسی تعامل را بر اساس بازنمایی تصویر به دو دسته‌ی اصلی طبقه‌بندی کنیم: بازنمایی کلی (سراسری) و بازنمایی محلی. بازنمایی‌های کلی و محلی از دیدگاه استخراج ویژگی، تفاوت‌های چشمگیری دارند. بازنمایی کلی، کار شناسایی عمل را با مکان‌یابی افراد حاضر در صحنه آغاز می‌کند. این درحالی‌است که بازنمایی محلی، از سطح پیکسل آغاز می‌شود. روش پیشنهادی این تحقیق یک روش بالا به پایین در دسته‌ی بازنمایی کلی است که از پردازش افراد عملگر به جای پردازش در سطح پیکسل شروع می‌شود. به‌عنوان یک نمونه از روش‌های با بازنمایی کلی، در [۱۱]، نویسندگان یک شبکه کاملاً متصل عمیق به نام R-NKTM، به منظور تعریف یک توصیفگر پیشنهاد می‌کنند. در [۱۲]، نویسندگان مدل اصلی LSTM[1] را توسعه داده و یک شبکه جدید بر پایه‌ی (GCA-LSTM) را پیشنهاد می‌کنند که شامل یک سلول حافظه محتوای سراسری و دولایه LSTM است. این شبکه دارای قابلیت بازشناسی قوی برای تشخیص تعامل مبتنی بر اسکلت است. در [۱۳]، نویسندگان از مفاصل انسان به‌عنوان نقاط کلیدی معنادار استفاده می‌کنند و این بازنمایی حالت حرکت را PoTion می‌نامند. پس از مرحله تخمین حالت، نقشه‌ی حرارتی (heat map) برای مفاصل انسان در هر قاب استخراج می‌شود. بازنمایی مذکور با جمع‌بندی زمانی این نقشه‌های احتمالی و رنگ‌آمیزی هر یک از آن‌ها بسته به زمان نسبی قاب‌ها در کلیپ ویدئویی و جمع‌بندی نهایی آن‌ها به دست می‌آید. این مطالعه را می‌توان در دسته بازنمایی‌های کلی طبقه بندی کرد، چراکه با استفاده از مرحله تخمین‌گر حالت، پردازش قاب را از افراد حاضر درصحنه آغاز می‌کند. در [۱۴]، یک مدل‌سازی مکانی-زمانی از ویدئوهای تعامل انسان-شیء برای تشخیص برخط و برون‌خط پیشنهاد شده است. برای طبقه‌بندی برخط، فواصل بین مفاصل و شیء به‌عنوان ویژگی‌های سطح پایین در نظر گرفته می‌شوند که این امر می‌تواند ما را به طبقه‌بندی این تحقیق به‌عنوان یک روش بر پایه‌ی بازنمایی کلی سوق دهد. از سوی دیگر، به‌عنوان یک نمونه از روش‌های با بازنمایی محلی، رحمانی و همکاران [۱۵] روشی را پیشنهاد کردند که در آن، یک توصیفگر را در یک نقطه با رمزگذاری بافت‌نگار[2] اجزای اصلی جهت‌دار (HOPC[3]) و در یک حجم مکانی-زمانی انطباقی حول آن نقطه استخراج می‌کنند. به‌این‌ترتیب، روش جدید می‌تواند نقاط کلیدی مکانی-زمانی (STKPs[4]) را در توالی‌های ابری نقطه‌ای سه‌بعدی شناسایی نماید. به‌عنوان مثال دیگری از روش‌های با بازنمایی محلی، در [۱۶]، بازنمایی جدیدی از ویدئوها پیشنهاد شده است که داده‌ها را به شیوه‌ای کلی و بدون توجه به محتوا کدگذاری می‌کند که منجر به بازشناسی تعامل به صورت بسیار دقیق می شود. این بازنمایی جدید، اطلاعات حرکتی موجود در تمام قاب‌های ویدئویی را در یک تصویر واحد خلاصه می‌کند که به آن تصویر پویا می گویند.

با توجه به مطالعات اشاره شده در این بخش، می‌توان نتیجه گرفت تمامی تحقیقات بر پایه‌ی دو دسته بندی بازنمایی کلی و محلی و با استفاده از انواع ویژگی‌های مختلف همچون، ویژگی‌های صوتی، بصری و متنی انجام شده‌اند که هرکدام از روش‌های پیشنهادی معمولاً مختص کاربردی خاص است.

۳- روش پیشنهادی

هدف این بخش ارائه‌ی روش پیشنهادی و معرفی آن در سه مرحله اصلی است. روند کار کلی نیز در شکل ۴ نمایش داده شده است. در اولین مرحله، یک واژه‌نامه‌ی تصویری از سایه‌ی پیش‌زمینه ساخته می‌شود (شکل ۲ و شکل ۳ را ببینید). به‌منظور مقایسه هر قاب با واژه‌نامه‌ی تصویری و انجام طبقه‌بندی قاب، از روش کیف ویژگی‌های تصویری که متشکل از حالت‌های کلیدی[5] است استفاده می‌کنیم. با استفاده از توصیفگر IMTF، مقایسه‌ای جهت ارزیابی همبستگی بین هر قاب و حالت‌های کلیدی انجام می‌شود. هر دنباله‌ی ویدئویی به یک توالی از برچسب‌ها تبدیل می‌شود که آماده تجزیه‌وتحلیل با یک طبقه‌بند الگوی ترتیبی هستند؛ بنابراین هدف ما یافتن جواب دو سؤال اصلی خواهد بود: آنچه در قاب انجام می‌شود و آنچه در کل توالی انجام می شود. طبقه‌بند ترتیبی، افزونه‌ای جدید است که به‌طور ایده‌آل برای طبقه‌بندی ویدئویی با طول متغیر مناسب است.

۳-۱-واژه‌نامه‌ی تصویری سایه‌های پیش‌زمینه

در سال‌های اخیر، کاربرد خلاصه‌سازی ویدئو پژوهشگران فراوانی را مجذوب خود نموده است. به طور معمول، هدف از خلاصه‌سازی یک ویدئو، صرفه‌جویی در زمان و حافظه‌ی فرآیند بازیابی داده‌های آن است. این فرآیند خلاصه سازی ممکن است به تصاویر ثابت یا متحرک منجر شود [۱۷]. در واقع قاب‌های کلیدی، زیرمجموعه‌ای از همین تصاویر

[1] Long short-term memosy

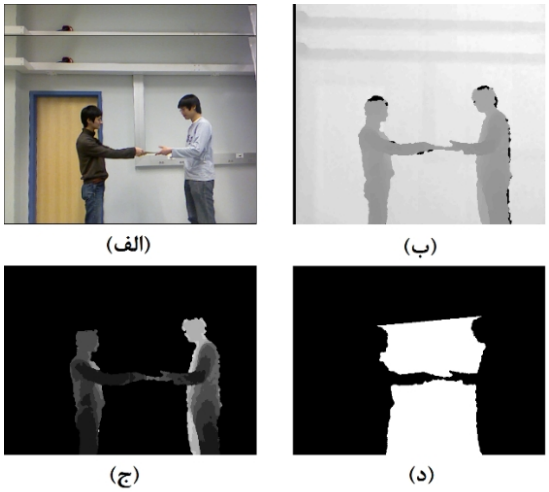
[2] Histogram

[3] Histogram of oriented principal components

[4] Spatio-temporal key-points

[5] Key-poses

ثابت هستند و به تصاویر استخراج‌شده از توالی‌های ویدئویی اشاره دارند که می‌توانند محتوای بصری اصلی دنباله را منعکس کنند [۱۸]. یک روش قابل‌قبول استخراج قاب کلیدی باید دو ویژگی اصلی داشته باشد: جذابیت و تنوع [۱۹].

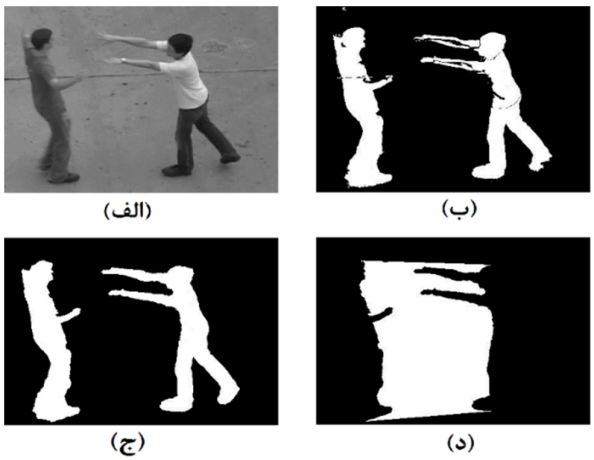


شکل ۲: فرآیند ایجاد سایه‌ی پیش‌زمینه در پایگاه داده SBU (الف) قاب اصلی (ب) تصویر عمق (ج) استخراج پس‌زمینه از تصویر (د) استخراج تصویر سایه پیش‌زمینه

جذابیت را می‌توان دارا بودن جالب‌ترین و بااهمیت‌ترین محتوای ویدیو در قاب‌های کلیدی، تعریف کرد. همچنین منظور از تنوع، نبود قاب‌های اضافی یا تکراری در مجموعه‌ی قاب‌های کلیدی است. در حال حاضر، روش‌های استخراج قاب کلیدی متنوعی به‌منظور بازشناسی تعامل پیشنهاد شده‌اند. روش انتخابی این پژوهش، روش استخراج قاب کلیدی پیشنهاد شده در [۲۰] بوده که بر اساس مطالعه مذکور، می‌توان با استفاده از خوشه‌بندی k-medoids با فاصله اقلیدسی، قاب‌های کلیدی دارای جذابیت و تنوع بالا به‌منظور بازشناسی تعامل استخراج نمود. این فرآیند برای هر طبقه تکرار می‌شود تا مجموعه‌ای از k قاب کلیدی برای هر دسته به دست آید. پس از تولید واژه‌نامه‌ی تصویری، هر دنباله به مجموعه‌ای از قاب‌های کلیدی برای انجام طبقه‌بندی مرحله بعدی تبدیل می‌شود.

۲-۳-طبقه‌بند قاب

به‌منظور استخراج اطلاعات داخلی و خارجی (مرزی) سایه‌ی دوطرفه بین افراد حاضر در صحنه، از توصیفگر مثلث چند مقیاسی نامتغیر (IMTF) پیشنهاد شده در [۲۱]، استفاده شده است. ویژگی IMTF بر روی شکل O و بر پایه‌ی N نقطه مرزی تعریف می‌شود $O=\{p_i,\ \ i\in[1,N]\}$، که در آن p_i، نقطه نمونه iام یک شکل است. نقطه p_i به‌صورت $p_i=(x_i,y_i)$ نشان داده می شود، که در آن x_i به‌عنوان مختصات محور طولی و y_i به‌عنوان مختصات محور عرضی نقطه p_i تعریف می‌شود.



شکل ۳: فرآیند ایجاد سایه‌ی پیش‌زمینه در پایگاه داده UT (الف) قاب اصلی (ب) استخراج پس‌زمینه از تصویر (ج) پس‌پردازش تصویر پس‌زمینه (د) استخراج تصویر سایه پیش‌زمینه

بنابراین، می‌توانیم IMTF را به‌صورت زیر تعریف کنیم:

$$IMTF=\{Area_k(i),cDist_k(i),k\in[1,T],i\in[1,N]\} \quad (۱)$$

که در آن، $Area_k$ مساحت نرمال شده و $cDist_k$ فاصله مرکزی $cDist$ در مقیاس k است. N به‌عنوان تعداد کل نقاط نمونه‌برداری و T به‌عنوان تعداد مقیاس‌ها تعریف می‌شوند که به‌طور تجربی روی $T=\log_2(N/2)$ تنظیم شده است. مساحت علامت‌دار مثلث نیز طبق رابطه‌ی زیر تعریف می‌شود:

$$Area_k(i)=\begin{bmatrix} x_i-h(k) & y_i-h(k) & 1\\ x_i & y_i & 1\\ x_i+h(k) & y_i+h(k) & 1\end{bmatrix} \quad (۲)$$

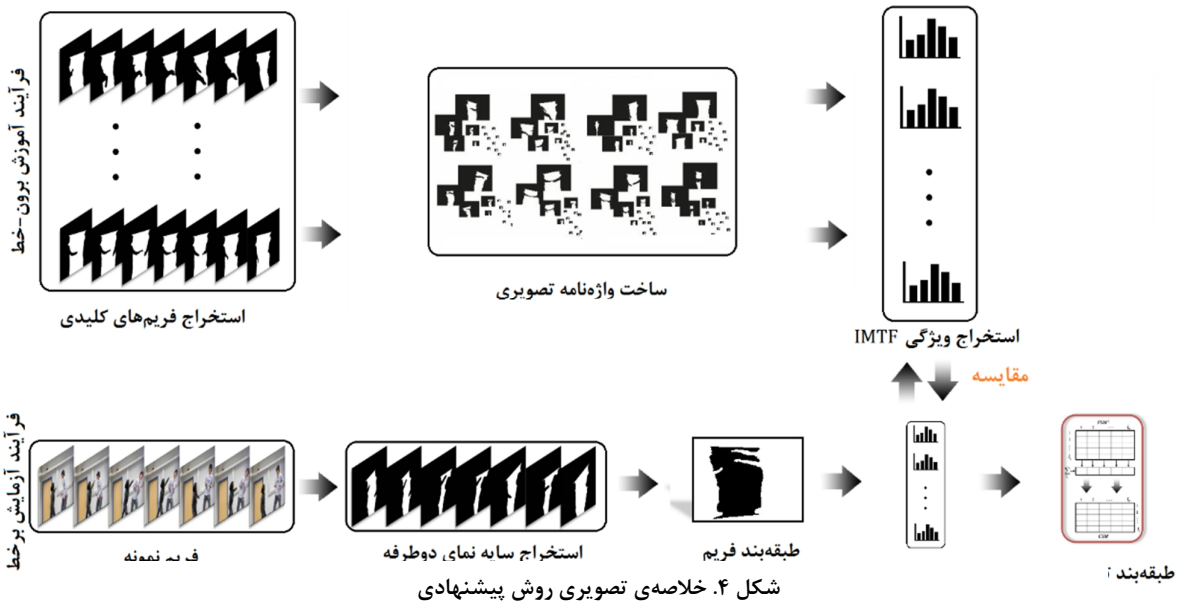
که در آن، دو نقطه مجاور $p_i-h(k)=(x_i-h(k),y_i-h(k))$ و $p_i+h(k)=(x_i+h(k),y_i+h(k))$ برای هر نقطه نمونه p_i از شکل O تعریف می‌شود. مؤلفه‌ی $h(k)$ به‌عنوان فاصله بین نقاط تعریف شده و به‌صورت 2^{k-1} تنظیم می‌شود که مقداری کوچک‌تر یا برابر $N/2$ است.

فاصله مرکزی $cDist_k(i)$ نیز به‌صورت زیر محاسبه می‌شود:

$$cDist_k(i)=\sqrt{(x_i-x_{gik})^2+(y_i-y_{gik})^2} \quad (۳)$$

که در آن (x_{gik},y_{gik}) به‌عنوان مختصات مرکز مثلث $\Delta P_{i-h(k)}P_iP_{i+h(k)}$ $(k\in[1,T],i\in[1,N])$ تعریف شده که به‌صورت زیر تعیین می‌شود:

$$\begin{cases} x_{gik}=(x_{i-h(k)}+x_i+x_{i+h(k)})\\ y_{gik}=(y_{i-h(k)}+y_i+y_{i+h(k)})\end{cases} \quad (۴)$$



شکل ۴. خلاصه‌ی تصویری روش پیشنهادی

نامتغیرهای ذکر شـــده در بالا، می‌توانند تغییرناپذیری انتقال و چرخش شـکل را مدیریت کنند. نرمال‌ســازی $Area_k(i)$ و $cDist_k(i)$ نیز طبق روابط زیر انجام می‌شود:

$$Area_k(i)=\frac{Area_k(i)}{\max_{i=1}^{N}|Area_k(i)|}$$
$$cDist_k(i)=\frac{cDist_k(i)}{\max_{i=1}^{N}|cDist_k(i)|} \quad (۵)$$

پس از انجام نرمال‌ســازی محلی، مقادیر IMTF در محدوده $[-1,1]$ قرار خواهند گرفت.

۳-۳-تطبیق شکل

در مرحله اول، از تشـــخیص لبه به‌منظور نمایش هر دو ســـایه‌ی پیش‌زمینه به‌صورت مجموعه‌ای از نقاط استفاده شده است (شکل ۵ را ببینید):

$$P=\{p_1,p_2,...p_n\} \quad p_i\in\mathbb{R} \quad (۶)$$

برای این کار، پس از استخراج پیش‌زمینه، به یک روش آشکارساز لبه مانند روش canny نیاز داریم. در فرآیند تطبیق محیط شـــکل‌ها، با توجه به دو سایه‌ی پیش‌زمینه A و B با نقاط محیط به ترتیب $\{p_1, p_2, \ldots, p_n\}$ و $\{q_1, q_2, \ldots, q_n\}$ نمونه‌برداری شده و عدم تجانس دو محیط p_i و q_j با استفاده از فاصله L_1 توصیفگرهای IMTF محاسبه می‌شود:

$$d(p_i,q_i)=\sum_{k=1}^{T}(|Area_k^p(i)-Area_k^q(j)|+ |cDist_k^p(i)-cDist_k^q(j)|) \quad (۷)$$

همچنین ماتریس عدم تجانس بین تمام نقاط نمونه نیز به‌صـــورت زیر تعریف می‌شود:

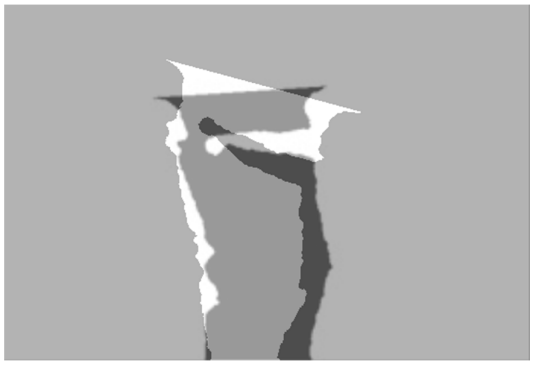
$$Dist(X,Y)\begin{bmatrix} d(p_1,q_1) & d(p_1,q_2) & \cdots & d(p_1,q_N) \\ d(p_2,q_1) & d(p_2,q_1) & \cdots & \cdots \\ \cdots & \cdots & d(p_i,q_j) & \cdots \\ d(p_N,q_1) & d(p_N,q_2) & \cdots & d(p_N,q_N) \end{bmatrix} \quad (۸)$$

با اســـتفاده از برنامه‌نویســـی پویا، نگاشـــت بهینه H بین دو توالی نقطه‌ای به‌گونه‌ای یافت می‌شود که $\sum_{i=1}^{N}d(p_i,q_{H(j)})$ حداقل باشد.

همچنین می‌توانیم اندازه‌گیری عدم تجانس برای شکل‌های X و Y را به‌صورت زیر تعریف کنیم:

$$Dist_{\min}(X,Y)=\sum_{i=1}^{N}d(p_i,q_{H(i)}) \quad (۹)$$

تمام این روش‌های عددی از [۱۹] اقتباس شده‌اند.



شکل ۵: شباهت‌سنجی دو عمل مشابه مشت زدن توسط ویژگی مثلث چند مقیاسی نامتغیر IMTF [۲۱]

۴-۳-طبقه بند توالی

هر تعامل بین دو فرد از عمل‌های متفاوتی تشکیل شده است که توسط آن‌ها انجام می‌شود. به‌عنوان مثال، دست دادن از سه عمل مختلف تشکیل شده است: نزدیک شدن، دست دادن، و عقب رفتن؛ بنابراین، در روش پیشنهادی، یک طبقه‌بند ثانویه به منظور دریافت این اطلاعات و انجام طبقه‌بندی برای کل توالی ویدیو پیشنهاد شده است. به‌این‌ترتیب، تصمیم‌گیری در مورد ماهیت عمل انجام‌شده در ویدیو صرفاً توسط یک یا چند قاب انتخابی انجام نمی‌گیرد. علاوه بر این، با استفاده از طبقه‌بند مرحله دوم، می‌توان خطاهای احتمالی مرحله طبقه‌بندی قاب را نیز حذف کرد. برای انجام این کار، از روش الگو کاوی ترتیبی SPM[1] استفاده شده است که می‌توان آن را به‌عنوان یک ابزار طبقه‌بندی قدرتمند و دقیق برای اصلاح طبقه‌بندی‌های اشتباه قاب، معرفی کرد. الگو کاوی ترتیبی یک حالت خاص از داده کاوی ساختار یافته است که فرآیند استخراج اطلاعات ارزشمند از مجموعه داده‌های نیمه ساختار یافته را انجام می‌دهد. در یک پایگاه داده که حاوی مجموعه‌ای از توالی‌ها است، روش SPM زیرتوالی‌های با بسامد تکرار بالا را به عنوان الگوهایی که حد آستانه‌ی مؤلفه‌ی *minsup* را ارضا می‌کنند، خواهد یافت. حداقل آستانه پشتیبانی یا *minsup* به معنی حداقل آستانه‌ی تعداد دفعات مشاهده شدن توالی است [۲۲].

روش SPM به‌طور گسترده در زمینه‌های مختلفی مانند بیوتکنولوژی [۲۳]، پیش‌بینی حساسیت به بیماری [۲۴]، بازاریابی [۲۵]، شبکه‌اینترنت [۲۶]، پایش سلامت موتور [۲۷] و غیره استفاده شده است. بااین‌حال، استفاده از آن در زمینه بینایی کامپیوتر هنوز به‌طور عمیق مورد بررسی قرار نگرفته است. در این پژوهش، به‌عنوان مرحله طبقه‌بند توالی، از یک روش طبقه‌بندی توالی جدید استفاده می‌کنیم که مبتنی بر مدل موفقیت‌آمیز الگو کاوی ترتیبی است که توسط Exarchos و همکارانش معرفی شده است [۲۸]. این مدل از دو مرحله اصلی تشکیل شده است: ایجاد مدل SPM و بهینه‌سازی. درواقع، Exarchos و همکاران از روش داده‌کاوی ترکیبی پیروی کرده‌اند که این روش، از الگو کاوی ترتیبی برای طبقه‌بندی توالی استفاده می کند. مسئله الگو کاوی ترتیبی بدین‌صورت تعریف می‌شود: فرض کنید $I=\{i_1,i_2,...,i_o\}$ مجموعه‌ای از اجزاء باشد. مجموعه اجزای $I_x=\{i_1,i_2,...,i_m\} \subseteq I$ مجموعه‌ای نامنظم از اجزای متمایز با اندازه $|I_x|$ است. توالی $s=(s_1,s_2,...,s_p)$ لیست مرتب‌شده‌ای از مجموعه اجزاء است که در آن $s_i \subseteq I, i\in\{1,...,p\}$ است. *sup* یک توالی، بخشی از یک توالی داده است که حاوی یک توالی فرعی است. الگو کاوی ترتیبی، توالی‌های پرتکرار را استخراج می‌کند که در آن کلمه پرتکرار[2] به‌عنوان داشتن مقداری بالاتر از حد آستانه تعیین‌شده توسط کاربر تعریف می‌شود. یک توالی با طول l به‌صورت یک l-sequence تعریف می‌شود. ورودی روش پیشنهادی مجموعه‌ای از توالی‌های آموزشی برچسب‌گذاری شده است که از یک طبقه‌بند قاب استخراج می‌شوند و خروجی تابعی است که یک توالی آزمایشی را به یکی از هشت طبقه نگاشت می‌کند. مرحله اول مدل SPM را بر اساس الگوهای متوالی استخراج شده ایجاد می‌کند. به‌طور خلاصه، در مرحله‌ی اول، مجموعه داده‌های توالی‌های آموزشی به زیرمجموعه‌های l_c جدا می شوند، جایی که هر زیرمجموعه شامل دنباله‌های متعلق به همان طبقه است $(S_j, j=1,...,l_c)$. سپس، روش SPM اعمال می شود و مجموعه‌های l_c از الگوهای متوالی P^j استخراج می‌شوند که طبقه‌های l_c را مشخص می‌کنند. در مرحله دوم، یک ماتریس امتیاز الگو برای هر طبقه $(PSM^j, j=1,...,l_c)$ تعریف می‌شود که در آن هر ماتریس PSM^j دارای امتیازی است که مفهوم الگوی متعلق به طبقه j را با تمام توالی‌های S_i توصیف می‌کند. این مفهوم بر اساس یک تابع امتیازدهی تعریف شده است. در بخش تابع امتیازدهی، اگر m امین الگوی طبقه j ام P_m^j در توالی S_i موجود باشد، آنگاه عنصر $PSM^j(m,i)$ برابر با طول P_m^j منهای ۱ است. در غیر این صورت اگر در توالی S_i وجود نداشته باشد، $PSM^j(m,i)$ برابر با صفر است. در مرحله سوم، ماتریس‌های امتیاز الگو با ضرب هر ردیف PSM^j در مؤلفه‌ای که وزن یک الگوی خاص را نشان می‌دهد، به‌روز می‌شود. مرحله چهارم از PSM^j به‌روز شده استفاده می‌کند و ماتریس امتیاز طبقه در این مرحله محاسبه می‌شود که در آن هر (j,i) عنصر CSM، امتیاز توالی iام برای طبقه jام است. با استفاده از یک مؤلفه‌ی وزن به‌عنوان وزن هر طبقه (w_c)، ماتریس امتیاز طبقه در مرحله پنجم به‌روز می‌شود. در مرحله آخر، یک ماتریس درهم‌ریختگی با استفاده از ماتریس امتیاز طبقه به‌روز شده محاسبه می شود و برچسب طبقه pc_i برای هر توالی S_i پیش‌بینی می‌شود. طبقه‌ای که بالاترین امتیاز را برای S_i کسب می‌کند به‌عنوان طبقه پیش‌بینی شده (محاسبه‌شده) تعیین می‌شود. مرحله دوم مدل، بهینه‌سازی مدل را انجام می‌دهد که به ما کمک می‌کند دو مجموعه وزن بهینه را پیدا کنیم. مجموعه اول برای طبقه ها w_c و مجموعه دوم برای الگو های متوالی w_p است. درواقع یک تابع هدف با استفاده از ماتریس درهم‌ریختگی طبقه‌بندی تعریف شده و فرآیند بهینه‌سازی روی تابع هدف انجام می‌شود تا آن را به حداقل برساند. اگر $g(CM)$ را به‌عنوان تابعی بر اساس ماتریس سردرگمی CM تعریف کنیم، می‌توانیم معادله ۴ را به‌عنوان یک مسئله بهینه‌سازی مطرح کنیم. پس از فرآیند بهینه‌سازی، وزن‌های بهینه شده را به‌صورت wp^* , wc^* به‌دست خواهیم آورد. معادله ۵ تابع هدف مورد استفاده برای فرآیند بهینه‌سازی را نشان می‌دهد که بین صفر و یک متغیر است. این کمینه سازی تابع هدف، دقت طبقه‌بندی توالی را افزایش می‌دهد. با استفاده از مؤلفه‌های اولیه با مقادیر $(wp=1, wc=1)$، فرآیند بهینه‌سازی با استفاده از یک روش بهینه‌سازی محلی به نام روش بهینه‌سازی Roll آغاز می‌شود [۲۹، ۳۰].

[1] Sequential pattern mining

[2] Frequent

$$f(D,wc,wp)=g(CM) \quad (۱۰)$$

$$g(CM)=l_s-trace(CM) \quad (۱۱)$$

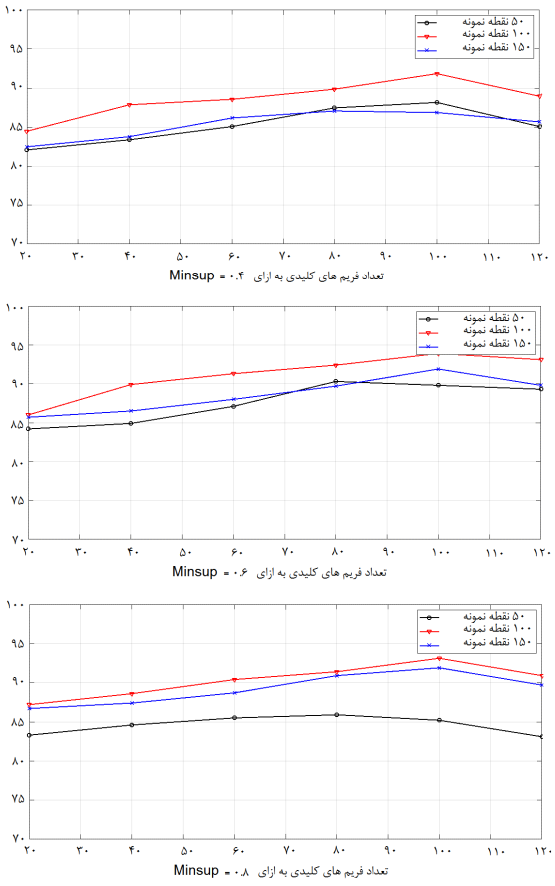
همچنین در روش پیشنهادی، از روش CM-SPADE معرفی شده توسط Fournier-Viger و همکاران [۳۱] که ساختار جدیدی بر اساس نسخه قدیمی SPADE (کشف الگوی متوالی با استفاده از طبقه‌های هم ارزی) [۳۲] است استفاده شده است. نتایج روش در مقایسه با روش‌های پیشرفته مانند SPADE، PrefixSpan، GSP و SPAM) و الگوهای متوالی بسته (ClaSP و CloSpan) بهبود چشمگیری نشان می‌دهد.

۴- پایگاه‌های داده

به‌منظور توسعه یک روش شناخت تعامل کاملاً جامع، روش پیشنهادی این مقاله بر روی چهار پایگاه داده رایج که تو سط مطالعات روز مورد استفاده قرار می‌گیرند، آزمایش شده است: مجموعه داده‌ی تعاملی SBU kinect [۳۳]، مجموعه داده‌ی تعاملی UT [۳۴]، پایگاه داده‌ی BIT [۳۵] و مجمو عه‌ی داده‌ی رف تار حرکتی NTU RGBD [۳۶]. هشت تعامل مختلف در مجموعه داده‌ی SBU وجود دارد که شامل طبقه‌های در آغوش گرفتن، لگد زدن، هل دادن، تبادل، دست دادن، نزدیک شدن، ترک کردن و مشت زدن است که توسط هفت شرکت‌کننده مختلف انجام می‌شود. این پایگاه داده‌ی چالش‌برانگیز، سه ویژگی اصلی دارد: حرکات غیر تکراری، شباهت حرکات بدن و تنوع اجراکنندگان. حرکات غیر تکراری، شناخت را در مقایسه با حرکات تکراری پیچیده می‌کند. در این پایگاه داده، داده‌های عمق با ا ستفاده از حس‌گر کینکت[1] مایکروسافت با وضوح ۶۴۰×۴۸۰ پیکسل به دست می‌آیند. علاوه‌براین، روش پیشنهادی بر روی پایگاه داده‌ی UT نیز ارزیابی شده است که این پایگاه داده، از شش طبقه مختلف شامل ا شاره کردن، هل دادن، لگد زدن، م شت زدن، در آغوش گرفتن و تکان دادن دست تشکیل شده است. این پایگاه داده، حاوی ویدئوهایی از اجرای رفتارهای حرکتی برای شش طبقه معرفی شده (با وضوح ۴۸۰×۷۲۰ پیکسل، ۳۰fps) است که افراد مختلف را در طول یک تعامل به ت صویر می‌ک شد. پایگاه داده‌ی تعاملی BIT-Interaction [۳۵] نیز به منظور بررسی روش پیشنهادی مطالعه شده است که این پایگاه داده شامل ۸ طبقه تعاملات ان سانی از طبقه های تعظیم کردن، م شت زدن، دست دادن، کف زدن، بغل کردن، ضربه زدن، نوازش کردن و هل دادن با ۵۰ ویدیو در هر طبقه ا ست. بخش‌هایی از بدن در ت صویر قابل م شاهده بوده و تغییراتی جزئی در شرایط نوری و مقیاس ت صویر وجود دارد. علاوه بر این، روش پیشنهادی بر روی پایگاه داده‌ی NTU RGB+D [۳۶] که در ا صل یک پایگاه جامع باز شنا سی رفتار حرکتی است، پیاده سازی شده است. در این پایگاه داده، تمام نمونه ها از ۱۰۶ فرد مختلف توسط حس‌گرهای کینکت جمع آوری شد. ویدیوهای RGB، توالی نقشه عمق، داده های اسکلت سه بعدی و فیلم های مادون قرمز (IR) برای هر نمونه ارائه شده است. تنوع شرایط محیطی در مقایسه با مجموعه داده های قبلی، از جمله ۹۶ پس زمینه مختلف به همراه تغییرات روشنایی، بیشتر است. پایگاه‌های داده‌ی مذکور در دسته‌ی پایگاه‌های داده‌ی مقید[2] قرار می‌گیرند که از لحاظ ساختار و زاویه‌ی دوربین و رفتارها و همچنین افراد حاضر در صحنه، با پایگاه‌های داده‌ی نامقید[3] که به طور معمول از و یدیو های فیلم ها یا محتوای اینترنت جمع‌آوری می‌شوند، متفاوت هستند. روش پیشنهادی بازشناسی نیز به منظور بازشناسی رفتار حرکتی در این دسته از پایگاه‌های داده معرفی شده است.

۵- نتایج

هدف این تحقیق، اثبات این موضوع است که شکل سایه‌ی پیش‌زمینه با استفاده از توصیفگر IMTF می‌تواند هر قاب تعامل را در یک شکل دودویی که حاوی توصیفی آموزنده از قاب است، خلاصه کند.



[1] Kinect

[2] Constrained

[3] Unconstrained

شکل ۶. نرخ بازشناسی به ازای مقادیر مختلف قاب‌های کلیدی، سه مقدار مختلف نقاط نمونه‌برداری و سه مقدار *minsup* مختلف در پایگاه داده‌ی SBU

به لطف روش طبقه‌بندی دومرحله‌ای، می‌توان اطمینان حاصل نمود که روش، در برابر اکثر طبقه‌بندی‌های اشتباه انجام شده در مرحله طبقه‌بندی اول، مقاوم است. در آزمایش‌های انجام شده، میانگین دقت طبقه‌بندی به‌عنوان معیار عملکرد تعریف می‌شود:

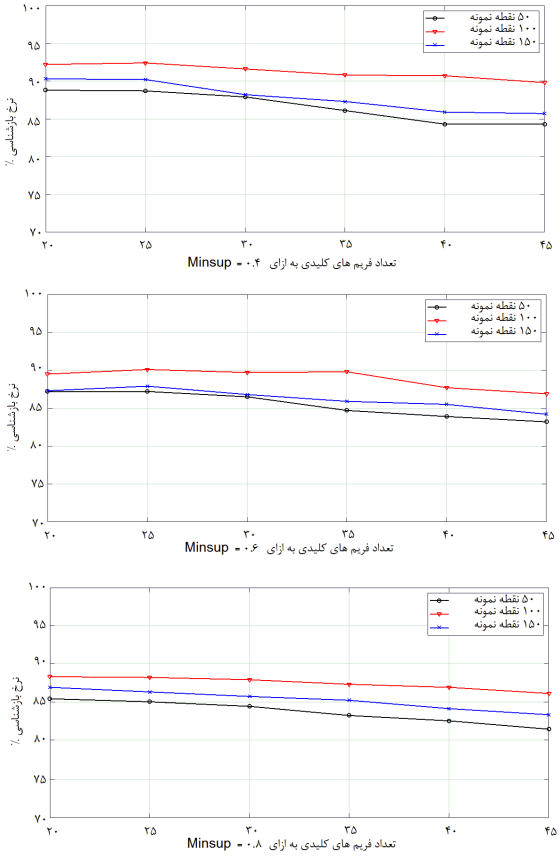
$$Accuracy = \frac{1}{C}\sum_{l=1}^{C}\frac{r_l}{t_l}\times 100\% \qquad (۱۲)$$

که در آن C تعداد دسته‌ها، r_l تعداد کل نمونه‌های طبقه‌بندی صحیح و t_l تعداد کل نمونه‌ها در یک دسته خاص است.

درحالی‌که سرعت و دقت سامانه‌ی طبقه‌بندی دومرحله‌ای پیشنهادی، با توجه به تعداد الگوهای واژه‌نامه‌ی تصویری تغییر می‌کند، تأثیر اندازه این واژه‌نامه نیز در فرآیند آزمایش‌ها مورد بررسی قرار گرفته است. با هدف بررسی چالش‌های ذکر شده در مؤلفه‌های ارزیابی، تأثیر مقادیر مختلف *minsup*، تعداد قاب‌های کلیدی در هر تعامل و سه دسته تعداد نقاط نمونه‌برداری شده از سایه‌ی پیش‌زمینه ۵۰، ۱۰۰ و ۱۵۰ تایی بررسی شده است. به‌منظور یافتن بهترین مقدار برای تعداد حالت‌های کلیدی، دامنه تعریف مؤلفه را به‌گونه‌ای تنظیم می‌کنیم که از ۲۰ حالت کلیدی در هر طبقه شروع شود و به ۱۲۰ حالت کلیدی در هر طبقه برسد.

۱-۵-بررسی تاثیر طبقه‌بند

به منظور مشاهده‌ی تاثیر طبقه‌بند پیشنهادی در مقایسه با طبقه‌بندی‌کننده‌های رایج مورد استفاده برای شناسایی تعامل بین افراد، روش استخراج ویژگی پیشنهادی با طبقه بند SVM تلفیق شده و نتایج این طبقه‌بندی به ازای هر پایگاه داده در بخش مقایسه‌ی نرخ بازشناسی با روش‌های دیگر آورده شده است. برای این منظور، از هسته ی تابع پایه شعاعی[1] استفاده شده و نتایج بدست آمده، نشان‌گر دقت بالاتر طبقه‌بند توالی SPM در مقایسه با این طبقه‌بند رایج است. همچنین به منظور نمایش تاثیر مرحله‌ی دوم طبقه‌بند پیشنهادی، این مرحله حذف شده و نتایج مرحله‌ی طبقه‌بندی قاب در همان مرحله‌ی استخراج ویژگی با استفاده از یک طبقه‌بند K-NN، خوشه‌بندی شده است. نتایج بدست‌آمده نشان می‌دهد که روش پیشنهادی دومرحله‌ای با حاشیه‌ای قابل توجه از هر دو طبقه‌بند رایج، عملکرد بهتری دارد. مهمترین عامل این امر را می‌توان اصلاح خطاهای طبقه‌بند مرحله‌ی اول با استفاده از طبقه‌بند ترتیبی مرحله‌ی دوم دانست. همچنین نتایج بدست‌آمده، نشان‌دهنده‌ی دقت بسیار پایین روش K-NN در تعامل‌هایی همچون دست دادن و تبادل اشیا و عدم تشخیص صحیح این دو طبقه است.



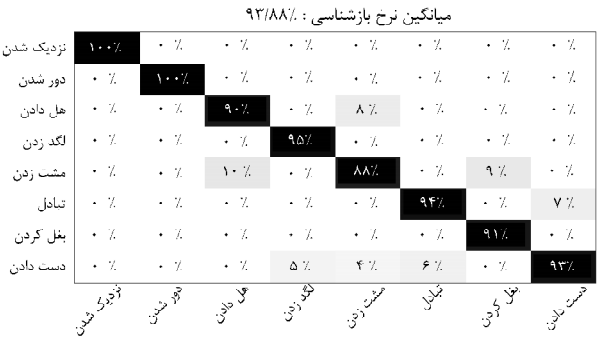
شکل ۷. نرخ بازشناسی به ازای مقادیر مختلف قاب‌های کلیدی، سه مقدار مختلف نقاط نمونه‌برداری و سه مقدار *minsup* مختلف در پایگاه داده‌ی UT

۲-۵-مجموعه داده‌ی تعامل SBU

در این پایگاه داده، به لطف اطلاعات دقیق عمق، نیازی به استخراج پیش‌زمینه و روش‌های تشخیص لبه که به‌طور دقیق در بخش سه مورد بحث قرارگرفته‌اند، وجود نخواهد داشت. درحالی‌که تمام اطلاعات موردنیاز برای استخراج سایه‌ی پیش‌زمینه در این محدوده‌ی عمق قرار دارد، با استفاده از یک فیلتر آستانه‌ای ساده، سایه‌ی پیش‌زمینه به‌طور دقیق استخراج می‌شود.

[1] Radial basis function

مقدار *minsup* یک مؤلفه‌ی کلیدی اســـت که فرآیند یادگیری را در طبقه‌بند SPM کنترل می‌کند؛ بنابراین، مجموعه‌ای از مقادیر از ۰،۴ تا ۰،۸ به‌منظور یافتن یک مقدار *minsup* بهینه بررسی شده است.

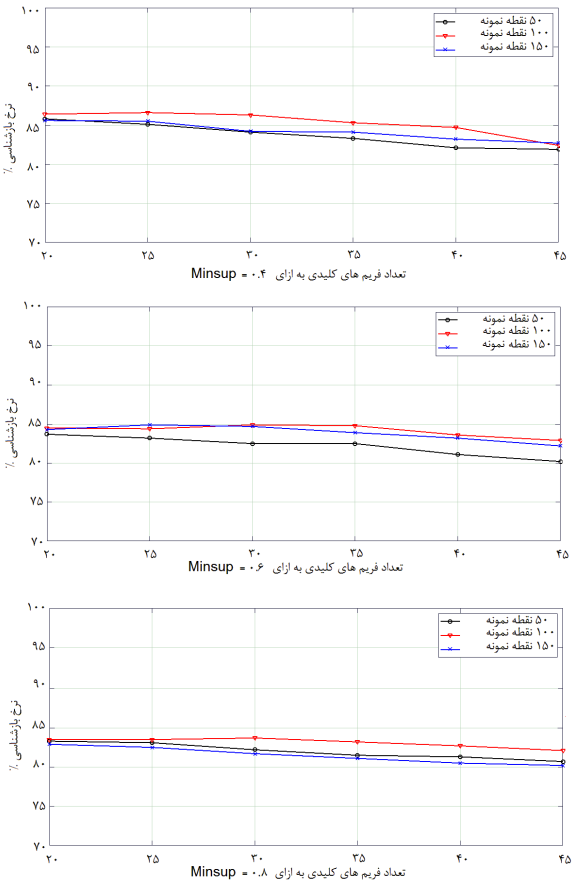


شکل ۸. ماتریس درهم ریختگی در پایگاه داده‌ی SBU

همانطور که در شکل ۶ نیز قابل مشاهده است، در پایگاه داده‌ی SBU، طبقه‌بند SPM با مقدار *minsup* برابر با ۰،۶ و ۱۰۰ قاب کلیدی در هر طبقه و همچنین ۱۰۰ نقطه‌ی نمونه‌برداری شده از سایه‌ی پیش‌زمینه، بهترین نتایج را ارائه می‌دهد. با افزایش تعداد قاب‌های کلیدی از ۲۰ به ۱۰۰، دقت باز شنا سی افزایش مح سو سی یافته و پس‌ازآن کاهش می یابد (یا در برخی موارد ثابت می‌ماند). یکی از دلایل احتمالی، ا ستخراج تعداد زیاد و غیرضـــروری قاب‌های کلیدی با بالا رفتن تعداد حالت‌های کلیدی اســت که باعث ســردرگمی در انتخاب برچســب طبقه مناســب برای قاب تحت آزمایش می‌شود.

از سوی دیگر، با ا ستفاده از تعداد قاب‌های کلیدی کمتر برای هر طبقه، انتظار می‌رود شاهد افزایش تعداد طبقه‌بندی‌های نادر ست باشـــیم، چراکه واژه‌نامه‌ی به‌دســـت‌آمده، جامعیت کافی نخواهد داشت. همچنین نتایج نشان می‌دهد که دقت شناخت با استفاده از نقاط نمونه بیشـــتر در مجموعه مؤلفه‌ی ســـوم کاهش می‌یابد. دلیل این امر، وجود داده‌های نامرتبط در محدوده تو صیفگر ا ست که می‌تواند باعث سردرگمی در روند تطبیق نقاط شود. ا ستفاده از نقاط نمونه کمتر نیز وضوح کمتری را ارائه می‌دهد که می‌تواند جزئیات را در فرآیند حذف کند. نتایج به‌دســـت‌آمده از طبقه‌بند SPM با ا ستفاده از روش اعتبار سنجی ۵-fold بر روی مجموعه داده‌های SBU در جدول ۱ نمایش داده شده است.

این پروتوکل اعتبارســـنجی به‌منظور انجام اعتبار ســـنجی قابل مقایســـه با روش‌های پیشـــرفته انتخاب شـــده اســـت. نتایج به‌د ست‌آمده ن شان می‌دهد که میانگین عملکرد (نرخ موفقیت ٪) با نرخ ۹۱/۹۴٪ برای مقدار *minsup* برابر با ۰،۴، ۹۳/۹٪ برای م قدار *minsup* برابر با ۰،۶ و ۹۳/۱۲٪ برای م قدار *minsup* برابر با ۰،۸ اســـت که در آن تعداد قاب‌های کلیدی به ازای هر طبقه را روی ۱۰۰ و تعداد نقاط نمونه‌برداری شده را بر روی ۱۰۰ قرار می‌دهیم که بهترین مؤلفه‌ها براســـاس نتایج بدســـت آمده هستند.



شکل ۹. نرخ بازشناسی به ازای مقادیر مختلف قاب‌های کلیدی، سه مقدار مختلف نقاط نمونه‌برداری و سه مقدار *minsup* مختلف در پایگاه داده‌ی BIT

از آزمایش‌های انجام‌شـــده، مشـــهود اســـت که نرخ تشـــخیص صـــحیح رفتارهای حرکتی بیش از ۹۰٪ ا ست (به جز یک د سته)، درحالی‌که دو طبقه با دقت ۱۰۰٪ طبقه‌بندی شـــده‌اند. شـــکل ۸ نمایش دهنده‌ی نرخ بازشـــناســـی طبقه‌های مختلف پایگاه داده‌ی SBU در یک ماتریس درهم‌ریختگی اســـت که نشـــان‌دهنده‌ی نقاط ضـــعف و قوت روش پیشـــنهادی در تشـــخیص تعامل‌های مختلف موجود در این پایگاه داده اســـت. علاوه بر این، برخی از خطاها ممکن اســـت زمانی رخ دهند که رفتارهایی نظیر لگد زدن و در آغوش گرفتن با سایر موارد اشتباه گرفته شوند، زیرا زمان اجرای عمل اصلی در مقایسه با اقدامات قبل/بعد بسیار کوتاه است (نزدیک شدن، انجام عمل اصلی و عقب‌نشینی).

روش پیـشنهادی از لحاظ میانگین نرخ باز شنا سی بسیار نزدیک به روش‌های [۳۷] و [۳۸] عمل نموده اســت. روش پیشـــنهادی، در پایگاه

جدول ۱: مقایسه‌ی نرخ بازشناسی روش پیشنهادی با مطالعات روز بر اساس پایگاه داده‌ی SBU

روش	نزدیک شدن	دور شدن	هل دادن	لگد زدن	مشت زدن	تبادل	بغل کردن	دست دادن	میانگین نرخ بازشناسی
روش Deep Structured [۳۷]	۰/۹۵	۰/۹۳	۰/۹۵	۰/۹۵	۰/۸۹	۰/۹۵	۰/۹۷	۰/۹۲	۹۳/۴
روش GBSWC [۳۸]	۱	۱	۰/۹	۰/۸۶	۰/۸۹	۰/۹۵	۱	۰/۹۱	۹۳/۸۴
روش HGN+KNN [۳۹]	–	–	–	–	–	–	–	–	۹۰/۸
روش RV+HS+GCNs [۴۰]	–	–	–	–	–	–	–	–	۹۴/۵۴
روش MTLN [۴۱]	–	–	–	–	–	–	–	–	۹۵/۴
روش SVM	۰/۸۹	۰/۸۶	۰/۹	۰/۹۲	۰/۸۶	۰/۸۴	۰/۹۱	۰/۸۷	۸۸/۱
روش K-NN	۰/۸۴	۰/۷۳	۰/۸۶	۰/۸۶	۰/۸۱	۰/۷۲	۰/۸۷	۰/۷۴	۸۰/۳
روش پیشنهادی	۱	۱	۰/۹	۰/۹۵	۰/۸۸	۰/۹۴	۰/۹۱	۰/۹۳	۹۳/۹

داده‌ی SBU از اطلاعات عمق فقط به منظور استخراج پیش‌زمینه استفاده نموده است که این عمل صرفا به دلیل افزایش سرعت بوده و این روش بر روی تصاویر RGB نیز قابل پیاده سازی است. این در حالی است که مطالعات یاد شده، اساسا بر پایه‌ی اطلاعات عمق بوده و برای انجام بازشناسی، نیاز به داده‌های عمق دارند. بدیهی است که نصب دوربین عمق در عمل، هزینه‌ی بالاتری نسبت به دوربین‌های RGB داشته و از لحاظ بار محاسباتی نیز دارای هزینه‌ی بیشتری خواهد بود. علاوه بر این، نرخ بازشناسی نسبت به روش [۳۹]، ۳/۱٪ بهبود را نشان می‌دهد. این بهبود نمایان‌گر دقت نسبی بالاتر شکل سایه‌نمای دوطرفه در استخراج اطلاعات رفتار نسبی افراد به هم، در مقایسه با مطالعه‌ی یاد شده، علیرغم استفاده از روش اسکلت دو بعدی و استخراج مفاصل است.

همچنین در جدول ۱ مشاهده می شود که نرخ بازشناسی بدست آمده در روش پیشنهادی در مقایسه با روش‌های اعلام شده در مطالعات [۴۰] و [۴۱]، پایین‌تر است. در هر دوی این مطالعات، از اطلاعات عمق موجود در پایگاه داده استفاده شده است و اطلاعات به یک شبکه‌ی عصبی کانولوشنی اعمال شده است که از لحاظ بار محاسباتی، پیچیده‌تر از روش پیشنهادی است.

۳-۵-مجموعه داده‌ی تعاملی UT

جدول ۲ شامل نتایج بازشناسی برای مجموعه داده‌ی UT با استفاده از طبقه‌بندی SPM و روش اعتبارسنجی 5-fold است که میانگین نرخ بازشناسی ۹۲/۴٪ را نشان می‌دهد.

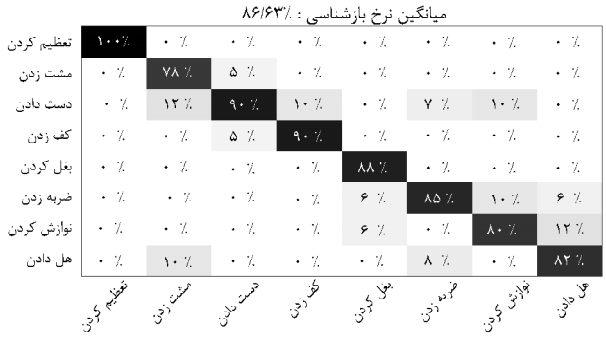
شکل ۱۰. ماتریس درهم ریختگی در پایگاه داده‌ی UT

باید توجه داشته باشیم که این مطالعه با هدف طبقه‌بندی رفتارهای حرکتی دو نفره انجام می‌شود، درحالی‌که یکی از طبقه‌های مجموعه داده‌ی UT مربوط به حرکت اشاره‌کردن است و توسط یک نفر انجام می‌شود که در فرآیند و ارزیابی‌های این تحقیق لحاظ نشده است. جدول ۲ شامل مقایسه دقیقی بین روش پیشنهادی و برخی از روش‌های روز با استفاده از این پایگاه داده است. توضیحات احتمالی برای خطاهای طبقه‌بندی، مشابه دلایلی است که در بخش قبلی توضیح داده شد.

همان‌طور که در شکل ۱۰ نیز نمایش داده شده است، در تعامل‌هایی که حرکت دست یک فرد به سمت فرد دیگر، عامل اصلی رفتار حرکتی است، نرخ بازشناسی کاهش یافته و خطای بازشناسی بین این طبقه‌ها رخ می‌دهد. در مجموعه داده‌ی تعاملی UT نیز بهترین نتایج با استفاده از ۱۰۰ نقطه‌ی نمونه‌برداری شده و حداقل مقدار *minsup* برابر با ۰،۴ و با استفاده از ۲۵ قاب کلیدی در هر طبقه گرفته شده است (شکل ۷). نتایج به‌دست‌آمده توسط روش پیشنهادی، دقت طبقه‌بندی بسیار رقابتی ۹۲،۴٪ را در مجموعه داده‌ی UT نشان می‌دهد که یکی از بالاترین دقت‌های به‌دست‌آمده توسط مطالعات روز در مجموعه داده‌های تعاملی UT است. نزدیک‌ترین نرخ بازشناسی در میان روش‌های مقایسه شده، توسط روش ارائه شده در مطالعه‌ی [۴۲] بدست آمده است که به میزان ۰،۷٪ کمتر بوده و از روش کیف کلمات[1]

[1] Bag of Words

نیز استفاده نموده است. استفاده از کیف کلمات، این مطالعه را از دید کلی به روش پیشنهادی نزدیک می‌سازد. اما در این روش از هیچ‌گونه ویژگی مشترک بین افراد استفاده نشده و ویژگی‌های اصلی استخراج شده در این تحقیق، HoG[1] و HoF[2] از هر فرد و بصورت جداگانه می‌باشد. این ویژگی‌ها، ویژگی‌های شناخته‌شده‌ای بوده و در بسیاری از مطالعات، کارایی خود را در استخراج اطلاعات از ویدیو و به خصوص از ویدیوهای بازشناسی رفتار حرکتی نشان داده‌اند. از مزایای این روش نسبت به روش پیشنهادی نیز می‌توان به تک مرحله‌ای بودن طبقه بند اشاره نمود. در روش‌های [۴۴] و [۴۵] نیز مانند روش پیشنهادی در



شکل ۱۱. ماتریس درهم ریختگی در پایگاه داده‌ی BIT

این مقاله، از طبقه‌بندهای با پیچیدگی کمتر مانند SVM[3] و RaF[4] استفاده شده است که بر خلاف طبقه‌بند معرفی شده در [۴۳]، امکان پیاده‌سازی بر روی سخت افزارهای ساده تر را دارا می‌باشد که این امر، به هرچه نزدیک‌تر نمودن تحقیق به ساخت سخت‌افزار کمک شایانی می‌نماید.

۴-۵-مجموعه داده‌ی تعاملی BIT-Interaction

در جدول ۳ نتایج بازشناسی برای مجموعه داده‌ی BIT-Interction با استفاده از طبقه‌بندی SPM و روش اعتبارسنجی ۱۰-fold نمایش داده شده است که میانگین نرخ باز شناسی ٪۸۶/۶ را نشان می‌دهد. دلیل استفاده از روش ۱۰-fold، استفاده‌ی مطالعات مقایسه شده از این روش ارزیابی است. از آنجا که این پایگاه داده، مشخصات مشابه پایگاه داده‌ی SBU-Interaction را دارد، انتظار نرخ بازشناسی مشابه این پایگاه داده دور از ذهن نخواهد بود. اما به دلیل وجود چند دسته حرکت مشابه هم و با فاصله‌ی بین طبقه‌ای نزدیک، نرخ بازشناسی بدست‌آمده، کمتر از حد انتظار بوده‌است. مقدار و نوع خطاهای بازشناسی نیز در ماتریس درهم‌ریختگی شکل ۱۱ قابل مشاهده است. در این پایگاه داده نیز بیشترین خطای باز شناسی مربوط به رفتارهای حرکتی انجام شده با دست و به دلیل شباهت این رفتار ها بوده و کمترین میزان خطای بازشناسی مربوط به رفتار تعظیم کردن است که از لحاظ فیزیکی تفاوت بسیار زیادی با رفتارهای دیگر دارد. همچنین جدول ۳ شامل مقایسه دقیقی بین روش پیشنهادی و برخی از روش‌های روز با استفاده از این پایگاه داده است. شباهت بیش از حد تعامل‌های نوازش، مشت زدن و هل دادن را می‌توان به عنوان یکی از دلایل وقوع تشخیص نادرست در این پایگاه داده توجیه کرد. در مجموعه داده‌ی تعاملی BIT-Interaction نیز بهترین نتایج با استفاده از ۱۰۰ نقطه‌ی نمونه‌برداری‌شده و حداقل مقدار *minsup* برابر با ۰/۴ و با استفاده از ۲۵ قاب کلیدی در هر طبقه گرفته شده است. به منظور انتخاب بهترین مقادیر این متغیر ها، مجموعه‌ای از اعداد نقاط نمونه‌برداری، تعداد قاب‌های کلیدی و مقادیر *minsup* آزمایش شده‌اند که برخی از آن‌ها در شکل ۹ نمایش داده شده اند. نتایج به‌دست‌آمده توسط روش پیشنهادی، دقت ٪۸۶.۶ را در مجموعه داده‌ی BIT-Interaction نشان می‌دهد که یکی از نتایج قابل قبول در میان نتایج به‌دست‌آمده توسط مطالعات روز در مجموعه داده های تعاملی BIT-Interaction است. مطالعه‌ی [۴۸]، علیرغم استفاده از طبقه‌بند پیچیده‌تر شبکه‌های عصبی کانولوشنی، نرخ بازشناسی پایین‌تری نسبت به روش پیشنهادی و دو مطالعه‌ی مقایسه شده‌ی دیگر ارائه می‌دهد. یک دلیل احتمالی برای این نرخ بازشناسی پایین را می‌توان استفاده از ویژگی‌های ساده تر و انجام تصمیم‌گیری بر روی قاب‌ها دانست.

۵-۵-مجموعه داده‌ی NTU RGB+D

در این پایگاه داده نیز مقدار *minsup* را می‌توان به عنوان یک مؤلفه‌ی کلیدی معرفی نمود که فرآیند یادگیری را در طبقه‌بند SPM کنترل می‌کند؛ بنابراین، همانند پایگاه‌های داده‌ی قبلی، مجموعه‌ای از مقادیر در بازه‌ی ۰،۴ تا ۰،۸ به‌منظور یافتن یک مقدار *minsup* بهینه بررسی شده است. در پایگاه داده‌ی NTU، طبقه‌بند SPM با مقدار *minsup* برابر با ۰،۶ و ۶۰ قاب کلیدی در هر طبقه به ازای ۱۰۰ نقطه‌ی نمونه‌برداری‌شده از سایه‌ی پیش‌زمینه، نتایج بهتری نسبت به سایر مقادیر ارائه می‌دهد. شکل ۱۲ نشان دهنده‌ی نرخ باز شناسی به ازای مقادیر مختلف قاب‌های کلیدی، سه مقدار مختلف نقاط نمونه‌برداری و سه مقدار *minsup* برای این پایگاه داده می‌باشد.

با توجه به این که این مطالعه با هدف طبقه‌بندی رفتارهای حرکتی دو نفره انجام می‌شود و بسیاری از طبقه‌های رفتار حرکتی موجود در این پایگاه داده (۹۴ طبقه از ۱۲۲ طبقه) مربوط به رفتارهای حرکتی یک نفره است، ممکن است مقایسه‌ی میانگین نرخ باز شناسی نهایی روش پیشنهادی بر روی این پایگاه داده با مطالعات دیگر بر روی این پایگاه (که به طور معمول مطالعات بر روی بازشناسی رفتار حرکتی تک نفره هستند)، مقایسه‌ی دقیقی نباشد. جدول ۴ شامل نتایج بازشناسی رفتار حرکتی بر روی طبقه‌های دو نفره از این پایگاه داده می‌باشد. در

[1] Histogram of Oriented Gradients

[2] Histogram of Optical Flows

[3] Support Vector Machines

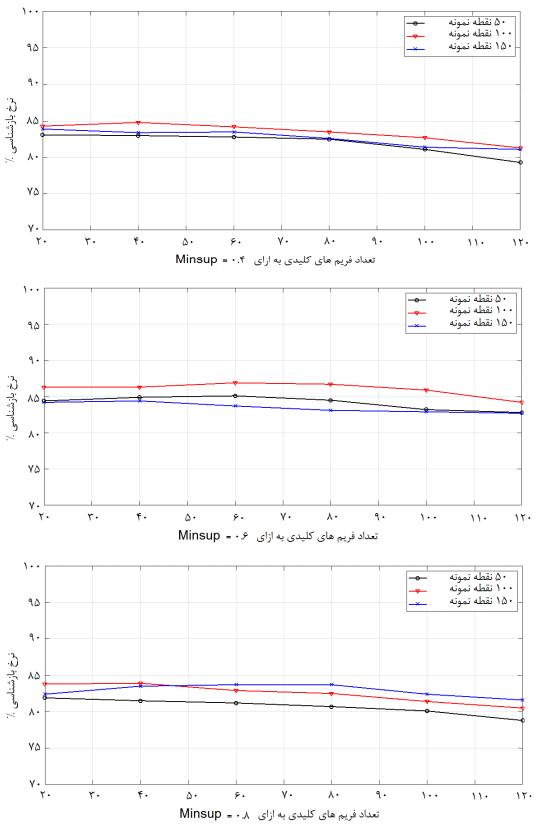
[4] Random Forests

این پایگاه داده، در برخی موارد، شـــباهت بین طبقه‌ای کمتر از دیگر پایگاه‌های داده است.

برای مثال، به دلیل زاویه‌ی پایین‌تر دوربین نســـبت به ســـایر پایگاه‌های داده‌ی معرفی شده، بازشناسی طبقه‌های دست دادن و سیلی زدن، با دقت بالاتری انجام می پذیرد. در نهایت می‌توان گفت روش پیشـنهادی با میانگین نرخ بازشـناسـی ۸۱/۲٪ در حالت بین زاویه‌ای و ۸۶/۹٪ در حالت بین فردی، نرخ قابل قبولی را در مقایســه با روش‌ها و پایگاه‌های داده‌ی دیگر ارائه می‌دهد. توجه به این نکته ضـروری اسـت که در هیچ کدام از مطالعاتی که با روش پیشـنهادی در این پایگاه داده مقایسه شده‌اند، نتایج بازشناسی به تفکیک طبقه‌ها بررسی نشده است.

جدول ۲: مقایسه‌ی نرخ بازشناسی روش پیشنهادی با مطالعات روز بر اساس پایگاه داده‌ی UT

روش	در آغوش گرفتن	لگد زدن	اشاره کردن	مشت زدن	هل دادن	دست دادن	میانگین نرخ بازشناسی
روش Discriminative key component models [۴۲]	–	–	–	–	–	–	۹۱/۷
روش DWT+Harris [۴۳]	۱	۱	۰/۶۵	۰/۹	۱	۱	۹۱/۵
روش Joint+AS [۴۴]	–	–	–	–	–	–	۹۰/۳
روش Motion trajectory occurrences [۴۵]	–	–	–	–	–	–	۶۸/۳
روش LMDI [۴۶]	۰/۷۵	۰/۷۵	۱	۰/۸۷	۰/۸۷	۱	۸۷/۵
روش SVM	۰/۷۸	۰/۸۹	–	۰/۸۱	۰/۸۳	۰/۷۸	۸۱/۸
روش K-NN	۰/۷۳	۰/۸۱	–	۰/۷۶	۰/۷۳	۰/۷۱	۷۴/۸
روش پیشنهادی	۱	۰/۹۵	–	۰/۹	۰/۸۸	۰/۸۹	۹۲/۴



شکل ۱۲. نرخ بازشناسی به ازای مقادیر مختلف قاب‌های کلیدی، سه مقدار مختلف نقاط نمونه‌برداری و سه مقدار *minsup* برای پایگاه داده‌ی NTU RGBD

بنابراین جدول ۴ شامل نرخ نهایی میانگین روش پیشنهادی در طبقه‌های رفتار دونفره با نرخ میانگین اعلام شده توسط مطالعات مذکور در تمامی طبقه‌ها می‌باشد. به منظور ارزیابی روش پیشنهادی، دو نوع ارزیابی بین فردی و بین زاویه‌ای بر پایه‌ی مطالعه‌ی اولیه انجام شده است. در ارزیابی بین فردی، تعامل های انجام شده توسط نیمی از افراد به عنوان داده‌ی آموزشی و تعاملات انجام شده توسط دیگر افراد به عنوان داده‌ی آزمایشی انتخاب شده است.

در ارزیابی زاویه‌ای، تمامی تصاویر ضبط شده توسط دوربین ۱ به عنوان داده‌های آموزشی و داده‌های ضبط شده توسط دوربین‌های شماره‌ی ۲ و ۳ به عنوان داده‌ی آزمایشی در نظر گرفته شده‌اند. بدیهی است که دلیل افت نرخ بازشناسی در ارزیابی بین زاویه‌ای، حساسیت روش پیشنهادی به زاویه‌ی دوربین و شرایط محیط فیلم‌برداری بوده است. حال آن‌که در حالت ارزیابی بین فردی به دلیل عدم حساسیت به وضعیت فیزیکی افراد(همچون قد و لباس)، نتایج به مراتب بهتری بدست آمده است.

۶-نتیجه‌گیری

در این مقاله، شکل سایه‌ی پیش‌زمینه به‌منظور شناخت تعامل بین افراد ارائه شده و دقت بازشناسی روی دو مجموعه داده‌ی رایج ارزیابی شده است. توانایی استخراج اطلاعات حیاتی یک حرکت در حین حذف اطلاعات اضافی، سایه‌ی پیش‌زمینه را به ابزاری قدرتمند برای تجزیه‌وتحلیل رفتارهای حرکتی تبدیل می‌کند. این مسئله به‌عنوان یک فرآیند یادگیری نظارت شده بر اساس طبقه‌بند الگو کاوی ترتیبی (SPM) و توصیف‌کننده شکل IMTF مطرح شده و با پیروی از رویکرد طبقه‌بند الگو کاوی ترتیبی، نرخ بازشناسی ۹۳/۹٪ برای پایگاه داده SBU، ۹۲/۴٪ برای مجموعه داده‌های تعاملی UT، ۸۶/۶٪ برای پایگاه داده‌ی BIT و ۸۶/۹٪ برای مجموعه‌ی داده‌های NTU RGBD به‌دست‌آمده است. مقایسه با مطالعات روز نشان می‌دهد که توصیفگر ویژگی پیشنهادی، مفید بوده و به ما کمک می‌کند تا عملکرد شناخت تعامل را در چندین مجموعه داده‌ی چالش‌برانگیز بهبود بخشیم. بااین‌وجود، روش همچنان دارای محدودیت‌هایی است که ما را به ادامه تحقیق و تلاش برای بهبود عملکرد روش ترغیب می‌کند. از جمله‌ی این محدودیت ها می‌توان به عدم قطعیت طبقه بازشناسی شده در مرحله‌ی پردازش قاب‌ها اشاره نمود.

در واقع بر پایه‌ی دو مرحله‌ای بودن طبقه‌بند، طبقه‌بندی نهایی طبقه یک رفتار حرکتی به بازبینی کلی یک ویدیو به جای تک‌قاب‌ها نیاز دارد. بنابراین طبقه‌بند دومرحله‌ای، علیرغم مزیت افزایش دقت بازشناسی، نیاز به زمان بیشتری برای تحلیل داده‌ها و تعیین طبقه نهایی خواهد داشت. از دیگر محدودیت‌های موجود در روش پیشنهادی می‌توان به وابستگی روش به زاویه‌ی دید دوربین اشاره نمود. البته کاهش نرخ بازشناسی در صورت تغییرات جدی در زاویه‌ی دید در روش پیشنهادی و روش‌های مشابه پیشنهاد شده توسط مقالات دیگر امری اجتناب ناپذیر بوده است، اما تلاش‌ها در این زمینه جهت کاهش حداکثری افت نرخ بازشناسی با تغییر زاویه‌ی دوربین، همچنان یکی از چالش‌های اصلی تحقیقات در این زمینه است.

یکی از نقاط ضعف روش پیشنهادی، نیاز به استخراج پس‌زمینه از تصویر بوده که علاوه بر بار محاسباتی بالاتر نسبت به روش‌هایی که از استخراج پس زمینه بی‌نیاز هستند، امکان وقوع خطا در طی این مرحله از پیش‌پردازش را افزایش می‌دهد. البته در پایگاه های داده‌ای مانند SBU که دارای اطلاعات عمق تصویر هستند، این مشکل به طور کامل حل شده و می‌توان گفت مزایای استفاده از پس‌زمینه بر عدم استفاده از آن غلبه می‌کند.

همچنین استخراج شکل سایه‌نمای دوطرفه که در این روش به عنوان مولفه‌ی اصلی استخراج ویژگی تصویر معرفی شده است، می‌تواند به یکی از چالش‌های اساسی در ثبت نرخ بازشناسی قابل‌قبول معرفی

جدول ۳: مقایسه‌ی نرخ بازشناسی روش پیشنهادی با مطالعات روز بر اساس پایگاه داده‌ی BIT-Interaction

روش	تعظیم کردن	مشت زدن	دست دادن	کف زدن	بغل کردن	ضربه زدن	نوازش کردن	هل دادن	میانگین نرخ بازشناسی
روش STAG [۴۷]	۰/۸۴	۰/۸۶	۰/۸۴	۰/۹۲	۰/۸۲	۰/۹۶	۰/۸۴	۰/۹۲	۸۷/۵
روش Multi-Features [۴۸]	–	–	–	–	–	–	–	–	۸۴/۶۳
روش SAFARRI [۴۹]	–	–	–	–	–	–	–	–	۸۷/۱۲
روش پیشنهادی	۱	۰/۷۸	۰/۹	۰/۹	۰/۸۸	۰/۸۵	۰/۸۰	۰/۸۲	۸۶/۶

شود. به گونه‌ای که اگر بین دو فرد حاضر در صحنه، شیء متحرکی وجود داشته باشد که به عنوان پیش‌زمینه استخراج شود، می‌تواند در تشخیص قاب در مرحله‌ی اول طبقه‌بندی ایجاد اشکال نموده و در کاهش نرخ بازشناسی نهایی تاثیرگذار باشد.

تحقیقات آتی، می‌تواند شامل پیشنهاداتی جهت تبدیل روش فعلی به یک سامانه‌ی سخت‌افزاری کاملا مستقل[1] و بی‌نیاز از رایانه (به طور مثال قابل پیاده‌سازی بر روی سامانه‌های نهفته) باشد که این امر با افزایش سرعت بازشناسی و مطالعه روش‌های جایگزین برای افزایش دقت شناخت تعامل جهت ساده سازی هرچه بیشتر میسر خواهد بود. طراحی و ساخت یک سامانه‌ی مستقل باز شناسی رفتار حرکتی، یک قدم مهم در راستای کاربردی شدن هر چه بیشتر این تحقیق و ساخت تجهیزات پردازش رفتار حرکتی و بازشناسی رفتارهای غیرعادی در محیط‌های شلوغ شهری به عنوان یک سامانه‌ی کمکی-حفاظتی خواهد بود. علاوه‌براین، رفع ضعف‌های اشاره شده در روش پیشنهادی فعلی نیز می‌تواند یکی از موارد مهم جهت تحقیق و بررسی در مطالعات آتی باشد. رفع مشکل کاهش سرعت و دقت ناشی از استخراج پس‌زمینه و جایگزینی آن با یک روش بهینه می‌تواند منجر به دستیابی به نرخ بازشناسی بالاتری شود.

همچنین کاهش وابستگی نرخ بازشناسی به زاویه‌ی دوربین با ارائه‌ی روش‌های مبتنی بر ورودی چنددوربینه و استفاده از ویژگی‌های کمکی جهت بازشناسی دقیق‌تر نیز یکی از مواردی است که اهمیت بالایی در ارائه‌ی یک سامانه‌ی کاربردی دارد.

جدول ۴: مقایسه‌ی نرخ بازشناسی روش پیشنهادی با مطالعات روز بر اساس پایگاه داده‌ی NTU RGB+D

روش	میانگین نرخ بازشناسی CV	میانگین نرخ بازشناسی CS
روش ST-GCN [۵۰]	۸۸/۳	۸۳/۵
روش HAR [۵۱]	۵۷/۷	۶۸/۹
روش Cooperative [۵۲]	۸۹/۰۸	۸۶/۴۲
روش پیشنهادی	۸۱/۲	۸۶/۹

[1] Standalone

تحقیقات آتی می‌تواند علاوه بر رفع موارد اشاره شده، به برخی از مسائل پیشرفته‌تر مانند تشخیص تعامل انسان با پس‌زمینه‌های به‌هم‌ریخته یا دوربین‌های متحرک و همچنین تشخیص تعامل بیش از دو فرد نیز بپردازد.

مراجع

[1] J. Chen, RD. Samuel and P. Poovendran, "LSTM with bio inspired algorithm for action recognition in sports videos", Image and Vision Computing, vol. 112, pp. 104214, 2021.

[2] P.E. Martin, J. Benois-Pineau, R. Péteri and J. Morlier, "Fine grained sport action recognition with twin spatio-temporal convolutional neural networks", Multimedia Tools and Applications, vol. 79, no. 27, pp. 20429-20447, 2020.

[3] V. Voronin, M. Zhdanova, E. Semenishchev, A. Zelenskii,Y. Cen and S. Agaian, "Action recognition for the robotics and manufacturing automation using 3-D binary micro-block difference", The International Journal of Advanced Manufacturing Technology, vol. 117, no. 7, pp. 2319-2330, 2021.

[4] L. Yang, K. Dong, Y. Ding, J. Brighton, Z. Zhan and Y. Zhao, "Recognition of visual-related non-driving activities using a dual-camera monitoring system", Pattern Recognition, vol. 116, pp. 107955, 2021.

[5] Y. Chen, B. Zhang and Y. Liu, "ESTN: Exacter spatiotemporal networks for violent action recognition", In 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), October 2021, Nanjing, China, pp. 44-48.

[6] P. Vogiatzidakis and P. Koutsabasis, "Address and command: two-handed mid-air interactions with multiple home devices", International Journal of Human-Computer Studies, vol. 159, pp. 102755, 2022.

[7] S. Nikzad and H. Ebrahimnezhad, "Two-person interaction recognition from bilateral silhouette of key poses", Journal of Ambient Intelligence and Smart Environments, vol. 9, no. 4, pp. 483-499, 2017.

[8] C. Yang and Q. Yu, "Invariant multiscale triangle feature for shape recognition", Applied Mathematics and Computation, vol. 15, no. 403, pp. 126096, 2021.

[9] JK. Aggarwal and MS. Ryoo, "Human activity analysis: a review", ACM Computing Surveys (CSUR), vol. 43, no. 3, pp.16, 2011.

Computer Physics Communications, vol. 46, no. 3, pp. 401–415, 1987.

[30] D.G. Papageorgiou, I.N. Demetropoulos and I.E. Lagaris, "MERLIN-3.1.1 A new version of the Merlin optimization environment, Computer Physics Communications, vol. 159, pp. 70-71, 2004.

[31] P. Fournier-Viger, A. Gomariz, M. Campos and R. Thomas, "Fast vertical mining of sequential patterns using co-occurrence information", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 2014, Tainan, Taiwan, pp. 40–52.

[32] M.J. Zaki, "Spade: An efficient algorithm for mining frequent sequences", Machine learning, vol. 42, no. 1, pp. 31–60, 2001.

[33] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning", In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, Providence, RI, USA, pp. 28-35.

[34] MS. Ryoo and JK. Aggarwal, "UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)" In IEEE International Conference on Pattern Recognition Workshops, August 2010, Istanbul, Turkey.

[35] Y. Kong, Y. Jia and Y. Fu, "Learning human interaction by interactive phrases", In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, October 7-13, 2012, Italy, pp. 300-313, Springer Berlin Heidelberg.

[36] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.Y. Duan and A.C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding", IEEE Transactions on Pattern Analysis and Machine Intelligence 42(10), 2684–2701, 2020

[37] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo and L. Zhang, "A deep structured model with radius–margin bound for 3D human activity recognition", International Journal of Computer Vision, vol. 118, no. 2, pp. 256-273, 2016.

[38] B. Liu, Z. Ju and H. Liu, "A structured multi-feature representation for recognizing human action and interaction", Neurocomputing, vol. 318, pp. 287-296, 2018.

[39] A. Mottaghi, M. Soryani and H. Seifi, "Action recognition in freestyle wrestling using silhouette-skeleton features", Engineering Science and Technology, an International Journal, vol. 23, no. 4 , pp. 921-930, 2020.

[40] X. Liu,Y. Li, T. Guo and R. Xia, "Relative view based holistic-separate representations for two-person interaction recognition using multiple graph convolutional networks", Journal of Visual Communication and Image Representation, vol. 70, pp. 102833, 2020.

[41] Mehmood, F., Chen, E., Akbar, M.A. and Alsanad, A.A., 2021. Human action recognition of spatiotemporal parameters for skeleton sequences using MTLN feature learning framework. Electronics, 10(21), p.2708.

[42] Y.S. Sefidgar, A. Vahdat, S. Se and G. Mori, "Discriminative key-component models for interaction detection and recognition", Computer Vision and Image Understanding, vol. 135, pp.16-30, 2015.

[43] S. Berlin and M. John, "Particle swarm optimization with deep learning for human action recognition", Multimedia Tools and Applications, vol. 79, pp. 17349-17371, 2020.

[44] Z. Wang, J. Jin, T. Liu, S. Liu, J. Zhang, S. Chen, Z. Zhang and D. Guo, "Understanding human activities in videos: A joint action and interaction learning approach", Neurocomputing, vol. 321, pp. 216-26, 2018.

[45] G. Garzón and F. Martínez, "A fast action recognition strategy based on motion trajectory occurrences", Pattern Recognition and Image Analysis, vol. 29, no. 3, pp. 447-56, 2019.

[46] S. Sahoo and S. Ari, "On an algorithm for human action recognition", Expert Systems with Applications, vol. 115, pp. 524-34, 2019.

[47] M. Mahmood, A. Jalal and K. Kim, "WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors", Multimedia Tools and Applications, vol. 79, pp. 6919-6950, 2020

[48] A. Jalal, M. Mahmood and A.S. Hasan, "Multi-features descriptors for human activity tracking and recognition in Indoor-outdoor environments", In 16th International Bhurban Conference

[10] R. Poppe, "A survey on vision-based human action recognition", Image Vision Computing, vol. 28, no. 6, pp. 976–990, 2010.

[11] H. Rahmani, A. Mian and M. Shah, "Learning a deep model for human action recognition from novel viewpoints", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 667–681, 2018.

[12] J. Liu, G. Wang, L. Duan, K. Abdiyeva and A.C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks", IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1586–1599, 2018.

[13] V. Choutas, P. Weinzaepfel, J. Revaud and C. Schmid, "Potion: pose motion representation for action recognition", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2018, Salt Lake City, UT, USA, pp. 7024–7033.

[14] M. Meng, H. Drira and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition", Image and Vision Computing, vol. 70, pp. 32–45, 2018.

[15] H. Rahmani, A. Mahmood, D.Q. Huynh and A. Mian, "Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition", In European Conference on Computer Vision, September 2014, Zurich, Switzerland, pp. 742–757.

[16] H. Bilen, B. Fernando, E. Gavves and A. Vedaldi, "Action recognition with dynamic image networks", IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pp. 2799–2813, 2018.

[17] B.T. Truong and S.Venkatesh, "Video abstraction: A systematic review and classification", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 3, no. 1, pp. 3-es, 2007.

[18] C. Kim and J.N. Hwang, "Object-based video abstraction for video surveillance systems", IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 12, pp.1128–1138, 2002.

[19] M. Fei, W. Jiang and W. Mao, "Memorable and rich video summarization", Journal of Visual Communication and Image Representation, vol. 42, pp. 207–217, 2017.

[20] S. Baysal, M. C. Kurt and P. Duygulu, "Recognizing human actions using key poses", In 20th International Conference on Pattern Recognition (ICPR), August 2010, Istanbul, Turkey, pp. 1727–1730.

[21] C. Yang and Q. Yu, "Invariant multiscale triangle feature for shape recognition", Applied Mathematics and Computation, vol. 403, pp. 126096, 2021.

[22] R. Agrawal and R. Srikant, "Mining sequential patterns", In Proceedings of the Eleventh International Conference on Data Engineering, March 1995, Taipei, Taiwan, pp. 3-14.

[23] K. Wang, Y. Xu and J. X. Yu. "Scalable sequential pattern mining for biological sequences", In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, November 2004, Washington D.C. USA, pp. 178–187.

[24] I. Batal, H. Valizadegan, G.F. Cooper and M. Hauskrecht, "A pattern mining approach for classifying multivariate temporal data", In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2011 , pp. 358–365.

[25] Y. Chen, M. Kuo, S. Wu and K. Tang, "Discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data", Electronic Commerce Research and Applications, vol. 8, no. 5, pp. 241–251, 2009.

[26] S. Kim, S. Park, J. Won and S. Kim, "Privacy preserving data mining of sequential patterns for network traffic data", Information Sciences, vol. 178, no. 3, pp. 694–713, 2008.

[27] A. Palacios, A. Martínez, L. Sánchez and I. Couso, "Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data", Engineering Applications of Artificial Intelligence, vol. 44, pp. 10–24, 2015.

[28] T.P. Exarchos, M.G. Tsipouras, C. Papaloukas and D. Fotiadis, "A two-stage methodology for sequence classification based on sequential pattern mining and optimization", Data and Knowledge Engineering, vol. 66, no. 3, pp. 467–487, 2008.

[29] G. Evangelakis, J. Rizos, I. Lagaris and I. Demetropoulos, "Merlin-a portable system for multidimensional minimization",

on Applied Sciences and Technology (IBCAST), 2019, pp. 371-376.

[49] C. Chattopadhyay and S. Das, "Supervised framework for automatic recognition and retrieval of interaction: a framework for classification and retrieving videos with similar human interactions", IET Computer Vision, vol. 10(3), pp. 220-227, 2016.

[50] S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition", In *Proceedings of the AAAI conference on artificial intelligence*, April 2018, vol. 32(1).

[51] E. Cippitelli, E. Gambi, S. Spinsante and F. Flórez-Revuelta, " Evaluation of a skeleton-based method for human activity recognition on a large-scale RGB-D dataset", IET.

[52] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu,2018, April. "Cooperative training of deep aggregation networks for RGB-D action recognition", In Proceedings of the AAAI conference on artificial intelligence, April 2018, vol. 32(1).