**دانشکده مهندسی برق و رباتیک**

**گروه مهندسی الکترونیک**

# تشخیص فعالیت های انسان مبتنی بر پردازش رشته های ویدئویی

**دانشجو : امیر محمودی**

**استاد یا اساتید راهنما :**

**دکتر علیرضا احمدی‌فرد**

**پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد**

**شهریور ۱۳۹۶**

## دانشکده مهندسی برق و رباتیک

## گروه مهندسی الکترونیک

**پایان نامه کارشناسی ارشد آقای امیر محمودی** به شماره دانشجویی: ۹۳۱۵۶۴۴

**تحت عنوان:**

**تشخیص فعالیت‌های انسان مبتنی بر پردازش رشته‌های ویدئویی**

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

| اساتید راهنما | امضاء | اساتید مشاور | امضاء |
|---|---|---|---|
| نام و نام خانوادگی : | | نام و نام خانوادگی : | |
| نام و نام خانوادگی : | | نام و نام خانوادگی : | |

| اساتید داور | امضاء | نماینده تحصیلات تکمیلی | امضاء |
|---|---|---|---|
| نام و نام خانوادگی : | | نام و نام خانوادگی : | |
| نام و نام خانوادگی : | | | |
| نام و نام خانوادگی : | | | |
| نام و نام خانوادگی : | | | |

تقدیم به:

پدر و مادر مهربان و برادرانم.

# سپاس گزاری

سپاس آن عدمی راکه هست ما بربود / ز عشق آن عدم آمد جهان جان به وجود

به هر کجا عدم آید وجود کم گردد / زهی عدم که چو آمد از او وجود فزود

«مولوی»

بر خود لازم می دانم از استاد گرامی ام آقای دکتر علیرضا

احمدی فرد که هدایت این پایان نامه را بر عهده داشته اند

و همواره مشوق اینجانب بوده اند سپاس گزاری نمایم.

# تعهد نامه

اینجانب **امیر محمودی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی الکترونیک** دانشکده **مهندسی برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه **تشخیص فعالیت‌های انسان مبتنی بر پردازش رشته‌های ویدئویی** تحت راهنمائی **دکتر علیرضا احمدی‌فرد** متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده ( یا بافتهای آنها ) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

**تاریخ**

**امضای دانشجو**

## مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است ) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

*** متن این صفحه نیز باید در ابتدای نسخه های تکثیر شده پایان نامه وجود داشته باشد .**

## چکیده

در سال‌های اخیر با افزایش قابل توجه تعداد ویدئوهایی که در اینترنت بارگذاری می‌شوند و هم‌چنین افزایش جنایت و فعالیت‌های تروریستی، نیاز به سامانه‌ی خودکاری که بتواند محتوای فعالیت‌هایی که در ویدئو رخ می‌دهد را شناسایی کند، بیش‌تر از گذشته احساس می‌شود. اولا تارنماهای به اشتراک‌گذاری ویدئو به دنبال راهکاری هستند که بتوانند محتواهای غیرقانونی یا خشونت‌آمیز را شناسایی کرده و حذف کنند. دوما نیروی انسانی به دلیل مشکلاتی مانند خستگی یا بی‌توجهی و هم‌چنین هزینه‌ی بالای آن برای نظارت بر دوربین‌های تصویری مناسب نیست. به دلایلی که گفته شد، زمینه‌ای تحقیقاتی برای شناسایی خودکار فعالیت در رشته‌های ویدئویی ایجاد شده است.

به دلیل تنوع پس‌زمینه و ظاهر افراد، برای توصیف فعالیت به ویژگی‌های توصیف‌کننده‌ی حرکت نیاز داریم. از طرفی این ویژگی‌ها باید نسبت به نویز و تغییرات نامنظم ناگهانی ناشی از لرزش دوربین مقاوم باشند. یکی دیگر از چالش‌هایی که در این زمینه وجود دارد تنوع شیوه‌ی رخداد یک فعالیت است. بنابراین لازم است درکی معنایی از رخداد موجود در ویدئو بدست آید.

در این تحقیق روشی سلسه‌مراتبی برای شناسایی فعالیت پیشنهاد می‌شود که ترتیب زمانی اطلاعات موجود در ویدئو را در نظر می‌گیرد و می‌تواند ویدئوهای با طول نابرابر را با یکدیگر مقایسه کند. در روش پیشنهادی از روش مسیر متراکم بهبودیافته برای استخراج ویژگی استفاده می‌شود و این ویژگی‌ها با استفاده از بردار فیشر بهبودیافته کد می‌شوند. ساختار ترتیبی ویدئو طی دو مرحله با روش‌های ادغام رتبه و تابع هسته‌ی انعطاف‌پذیر با زمان استخراج می‌شود. و سپس با استفاده از ماشین بردار پشتیبان فعالیت‌ها از یکدیگر تفکیک می‌گردند. با اجرای این روش بر روی ترکیب شماره‌یک پایگاه داده‌ی HMDB51 صحت ۵۷,۹۷درصد و بر روی پایگاه داده‌ی Olympic Sports صحت ۸۵,۰۷درصد بدست آمده است.



**کلمات کلیدی**: شناسایی فعالیت، ادغام رتبه، بردار فیشر بهبودیافته، مسیر متراکم بهبودیافته، پایگاه داده HMDB51، پایگاه داده Olympic Sports، تشخیص فعالیت، پردازش ویدئو

# فهرست مطالب

# فهرست اشکال

# فهرست جداول

# فصل اول : مقدمه

## ۱-۱ مقدمه

این تحقیق به موضوع شناسایی فعالیت‌های انسان در ویدئوهای مربوط به عالم واقع می‌پردازد. منظور از شناسایی فعالیت‌های انسان این است که یک سیستم رایانه ای بتواند به صورت خودکار رخدادهای یک ویدیوی نامعلوم را تحلیل کند و دریابد که چه فعالیتی در آن انجام می‌شود[۱]. این ویدئوها می‌تواند شامل فیلم‌ها، ویدئوهایی که دراینترنت بارگذاری می‌شوند، تصاویر دوربین‌های نظارتی و .. باشد. با وجود تحقیقات گسترده در این زمینه در سال‌های اخیر، هنوز موانع بسیاری بر سر پیاده‌سازی آن برای کاربرد عملی وجود دارد. تنوع پس‌زمینه و تفاوت ظاهر افراد انجام‌دهنده‌ی فعالیت موجب می‌شود به دنبال استخراج ویژگی‌هایی باشیم که نسبت به این موارد مقاوم باشند. با توجه به این‌که «حرکت» مهم‌ترین عنصر در شناسایی فعالیت است، مواردی نظیر لرزش دوربین، وجود نویز فراوان، جهت تصویربرداری، تغییرات روشنایی و حرکت دوربین به همراه صحنه در تصاویر واقعی از جمله چالش‌های شناسایی فعالیت هستند که موجب تفاوت زیاد بین ویدئوهای یک کلاس می‌شود[۲]. علاوه بر این، یک فعالیت خود می‌تواند به اشکال گوناگونی انجام شود. به عنوان مثال شکل (۱-۱) برخی از حالت‌های مختلف رخداد فعالیت نوشیدن را نشان می‌دهد.

بنابراین لازم است به‌گونه‌ای درکی معنایی از رخداد موجود در ویدئو بدست آید و شناسایی فعالیت بر مبنای این درک ثانویه انجام شود.

تشخیص فعالیت ویدئویی در حوزه‌های گوناگونی دارای کاربردهای متنوع است. تشخیص هویت افراد از روی ویژگی‌های فیزیکی یا رفتاری آن‌ها، تشخیص محتوای ویدیوهایی که در اینترنت بارگذاری می‌شوند، شناسایی فعالیت‌های مشکوک و پیشگیری از اقدامات تروریستی به وسیله‌ی دوربین‌های نظارتی هوشمند، امکان تعامل کاربر با رایانه از طریق فرامین حرکتی و استفاده از این فناوری در صنعت پویانمایی برای مدل کردن حرکات و رفتارهای واقعی انسان‌ها از جمله‌ی این کاربردها هستند[۳].

شکل (۱ - ۱) برخی از انواع مختلف فعالیت نوشیدن که از پایگاه داده‌ی HMDB51 [۴] استخراج شده است.

## ۱-۱-۱ نظارت تصویری

نظارت تصویری عموما به واسطه‌ی دوربین‌های مداربسته[۱] یا دوربین‌های تحت شبکه[۲] امکان‌پذیر می‌شود. این دوربین‌ها در مکان‌های عمومی مختلفی نظیر زیرساخت‌های حمل‌ونقل، خیابان‌های پرجمعیت، سالن‌های ورزشی، موزه‌ها، فروشگاه‌ها و مراکز خرید نصب می‌شوند تا به نگرانی‌های امنیتی در این اماکن پاسخ دهند.

با مدرن‌تر شدن جامعه و افزایش جمعیت و بالا رفتن سطح جرم و تهدیدات تروریستی، مانند گذشته نمی‌توان با استقرار نیروهای امنیتی در برخی مکان‌های کوچک به نیاز امنیتی جامعه پاسخ داد. دوربین‌های نظارتی می‌توانند نیاز به حضور این نیروها در همه‌جا را برطرف کنند. با این وجود،

[۱] Closed-Circuit Television (CCTV)
[۲] IP Cameras

نظارت بر تصاویر دریافتی این دوربین‌ها به وسیله‌ی نیروی انسانی پرهزینه است. سیستم بینایی انسان قادر است حجم وسیعی از داده‌های بصری سطح پایین را دریافت کرده و اطلاعات منتخب را به مغز ارسال کند تا با پردازش آن‌ها یک درک معنایی سطح بالا بدست آورده و به فرد سطح آگاهی قابل قبولی از موقعیت هر لحظه را ارائه کند. اما نظارت مداوم بر تعداد زیادی دوربین برای انسان ملال‌آور است. از طرفی خستگی ناشی از نظارت طولانی‌مدت می‌تواند باعث کاهش دقت و عدم توجه به برخی رویدادها شود.

برای پردازش داده‌های خام این دوربین‌ها در سیستم‌های سنتی از نیروی انسانی استفاده می‌شود. تحقیقات زیادی در مورد ضعف انسان در چنین نظارتی انجام شده است. به عنوان مثال در تحقیق انجام شده در [۵]، نشان داده شد که دقت نیروی انسانی، پس از ۲۰ دقیقه مشاهده و ارزیابی مداوم تصاویر به پایین‌تر از سطح قابل قبول کاهش پیدا می‌کند. علاوه بر آن در [۶] مطرح گردید که احتمال اشتباه سیستم بینایی انسان در یافتن اهداف بصری نادر مانند غربالگری سرطان در تصاویر پزشکی یا یافتن جسم خطرناک در پویش چمدان مسافرین بسیار زیاد است.

هزینه‌ی بالای نیروی انسانی و ضعف‌های دیگر آن منجر به ایجاد زمینه‌ای پژوهشی برای گذار به سیستم‌های نظارتی خودکار گردیده است.

## ۱-۱-۲ ویدئوهای اینترنتی

در گذشته‌ی نه چندان دور تولید محتوای ویدئویی منحصر و محدود به تعداد اندکی از رسانه‌های شناخته شده بود. اما با گسترش اینترنت و تارنماهای به اشتراک‌گذاری محتوای تولید شده توسط کاربر[۳]، همزمان با رفع این انحصار، امکان کنترل محبوبیت محتواها نیز از اختیار آن‌ها خارج شده است. امروزه صدها میلیون کاربر در حال به اشتراک‌گذاری محتوای تصویری‌شان در اینترنت هستند

[۳] User Generated Content (UGC)

و به همان نسبتی که طول ویدئوها کاهش یافته است، سرعت انتشار محتوا نیر افزایش یافته است[۷].

بارگذاری ویدئوهای خشونت‌آمیز و غیرقانونی، یکی از مشکلات مهم این تارنماهاست که با وجود حجم زیاد محتوا، کنترل آن دشوار است. در حال حاضر این تارنماها از طریق نظارت مداوم بر محتواهایی که توسط کاربران به عنوان محتوای نامناسب گزارش می‌شود به این مسئله پرداخته‌اند. اما در صورت امکان‌پذیر شدن تشخیص خودکار محتوا، این مشکل تا حد زیادی تسهیل خواهد شد. یکی دیگر از کاربردهای تشخیص خودکار محتوا، در موتورهای جست‌وجو است. در صورت کاربردی شدن این روش می‌توان به جای فهرست کردن[۴] ویدئوها بر اساس محتوای متنی همراه آن، آن‌ها را بر اساس محتوای تصویری‌شان فهرست کرده و مورد جست‌وجو قرار داد.

## ۱-۲ معرفی پایگاه‌های داده

این تحقیق بر روی دو پایگاه داده‌ی HMDB51 [۳] و Olympic Sports [۸] اجرا می‌شود و نتایج آن با تحقیقات پیشین مقایسه می‌گردد.

### ۱-۲-۱ پایگاه داده‌ی HMDB51

این پایگاه داده که در سال ۲۰۱۱ در دانشگاه براون[۵] در ایالات متحده‌ی آمریکا ایجاد شده است، شامل ۶۸۴۹ کلیپ است که به ۵۱ فعالیت تقسیم شده‌اند. این فعالیت‌ها در شکل (۱-۲) قابل مشاهده هستند. بیشتر این ویدئوها از فیلم‌های سینمایی یا وبسایت‌های UGC جمع‌آوری شده‌اند. این ۵۱ فعالیت را می‌توان در پنج‌دسته‌ی کلی جای داد[۹]:

الف) حرکت‌های صورت (لبخند، خنده، جویدن، صحبت).

ب) تعامل با اشیا (سیگار کشیدن، خوردن، نوشیدن).

---

[۴] index

[۵] Brown University

پ) حرکات عمومی بدن (دست زدن، بالا رفتن از پله، شنا کردن و ..).

ت) حرکت بدن ضمن تعامل با اشیا (شانه‌کردن مو، شمشیربازی، اسب‌سواری، شلیک با اسلحه و ...).

ث) تعامل بین انسان‌ها (در آغوش گرفتن، بوسیدن، مشت زدن به دیگران، دست دادن و ...).

ارائه‌کننده‌ی دیتابیس سه ترکیب متفاوت از انتخاب کلیپ‌ها برای آموزش و آزمایش را پیشنهاد کرده است. هر ترکیب شامل ۵۱۰۰ ویدئو است که ۷۰ درصد آن برای آموزش و ۳۰ درصد برای آزمایش درنظر گرفته شده است. یعنی به ازای هر یک از ۵۱ فعالیت ۷۰ ویدئو برای آموزش و ۳۰ ویدئو برای آزمایش درنظر گرفته شده است.


شکل (۱ - ۲) فعالیت‌های پایگاه داده‌ی HMDB51 [۴].

### ۱-۲-۲ پایگاه داده‌ی Olympic Sports

این پایگاه داده در سال ۲۰۱۰ ایجاد شده است و شامل ۱۶ فعالیت ورزشی است که در شکل (۱-۳) نمایش داده شده است.


شکل (۱ - ۳) فعالیت‌های پایگاه داده‌ی Olypmic Sports [۸].

فایل‌های این پایگاه داده دارای پسوند seq هستند که می‌توان از جعبه‌ابزار Piotr [۱۰] در محیط متلب[۶] برای خواندن یا نوشتن آن‌ها استفاده کرد. ارائه‌دهنده‌ی پایگاه داده یک ترکیب از انتخاب داده‌ها برای آموزش و آزمایش را به همراه آن پیشنهاد کرده است.

[۶] Matlab

## ۱-۳ هدف و ساختار پایان‌نامه

فصل‌های آتی این پایان‌نامه به این ترتیب سازمان‌دهی شده است: در فصل دوم روش‌های گذشته را مرور و آن‌ها را با یکدیگر مقایسه می‌کنیم. سپس در فصل سوم مبانی نظری را شرح می‌دهیم. در فصل چهارم روش پیشنهادی این تحقیق معرفی خواهد شد. فصلهای پنجم و ششم به نتایج شبیه‌سازی و تحلیل آن‌ها می‌پردازند.

# فصل دوم : مرور روش‌های پیشین

## ۲-۱ مقدمه

پیش از گسترش روش‌های شناسایی فعالیت بر اساس محتوای رشته‌های ویدئویی، شناسایی محتوای ویدئوها تنها از طریق متون همراه ویدئو یا حاشیه‌نویسی[۷] برای آن امکان‌پذیر بود. طی سال‌های اخیر تحقیقات زیادی در زمینه‌ی روش‌های شناسایی فعالیت در ویدئو انجام شده است. عمده‌ی این تحقیقات بر روی فعالیت‌های کوتاه‌مدت و به طول چند ثانیه انجام شده است و تحقیقات کمتری به شناسایی فعالیت‌های در مقیاس دقیقه، ساعت و روز پرداخته‌اند. روش‌های شناسایی فعالیت‌های کوتاه‌مدت هنوز ضعیف هستند و به شناسایی فعالیت‌های بلندمدت نیز کمتر پرداخته شده است[۱۱].

روش‌های شناسایی فعالیت انسان به دو دسته‌ی کلی تک‌لایه[۸] و سلسله‌مراتبی[۹] تقسیم می‌شوند[۱][۱۱][۱۲].

## ۲-۲ روش‌های تک‌لایه در شناسایی فعالیت

در این روش شناسایی فعالیت مستقیما از روی داده‌های ویدئویی انجام می‌شود و توالی مشخصی از فریم‌های تصاویر بیان‌گر یک فعالیت است. این روش‌ها زمانی موثر هستند که بتوان از داده‌های آموزشی الگوی ترتیبی مشخصی استخراج نمود که فعالیت مورد نظر را توصیف کند[۱].

روش‌های تک‌لایه برای تشخیص فعالیت‌های کوتاه‌مدت مانند راه رفتن، پریدن، دست تکان دادن و ... و هم‌چنین فعالیت‌های بلندمدت محدود (مثلا: خودرویی که به سمت چپ گردش می‌کند) مناسب هستند[۱۱].

دو رویکرد در روش‌های تک‌لایه وجود دارد: رویکرد فضا-زمانی[۱۰] و رویکرد ترتیبی[۱۱] [۱][۱۱]. در رویکرد فضا-زمانی، فعالیت به صورت حجمی که از تجمیع فریم‌ها در طول بعد زمان تشکیل

---

[۷] Annotation
[۸] Single-Layered
[۹] Hierarchical

می‌شود، یا ویژگی استخراج شده از این حجم مدل می‌شود. اما در رویکرد ترتیبی، ویژگی‌ها از هر فریم یا تعداد معدودی از فریم‌ها استخراج می‌شود. سپس با کنار هم قرار دادن ویژگی‌های بدست آمده از همه‌ی فریم‌ها، برداری از ویژگی‌ها ساخته می‌شود و این بردار که دربردارنده‌ی ترتیب وقایع ویدئو است، فعالیت را توصیف می‌کند.

## ۲-۲-۱ رویکرد فضا-زمانی

در این رویکرد ویدئو یا بخشی از آن که بیان‌گر یک فعالیت است، به صورت حجمی در فضا-زمان در نظر گرفته می‌شود و مدلی برای فعالیت ایجاد می‌شود. شناسایی فعالیت جدید، از طریق محاسبه‌ی میزان شباهت این حجم یا ویژگی استخراج شده از آن حجم، بین ویدئوی آزمایشی و مدل صورت می‌پذیرد. در بیشتر موارد، با استفاده از یک پنجره‌ی لغزان[۱۲] میزان شباهت با مدل آموزش دیده شده برای تمام نقاط ویدئوی آزمایشی محاسبه می‌گردد و معمولا این پنجره‌ها دارای اشتراک هستند[۱۱].

این رویکرد را بر اساس نوع ویژگی استفاده شده می‌توان به سه دسته تقسیم کرد که در ادامه معرفی می‌گردد.

### ۲-۲-۱-۱ حجم فضا-زمانی

در این روش‌ها با فرض در اختیار داشتن دو حجم فضا-زمانی از ویدئوهای آموزش و آزمایش، یک معیار شباهت[۱۳] میان آن‌ها محاسبه می‌شود. و بر اساس این معیار شباهت، فعالیت شناسایی می‌گردد[۱۱]. به عنوان مثال [۱۳] با محاسبه‌ی اختلاف روشنایی بین پیکسل‌های متناظر فریم‌های متوالی ویدئو و تجمیع مقادیر بدست آمده در طول زمان، یک ویژگی دوبعدی بدست آورد که از آن به عنوان معیار شباهت استفاده کرده است. شکل (۲-۱) نشان می‌دهد که آن‌ها از هر ویدئو، تصویر

---

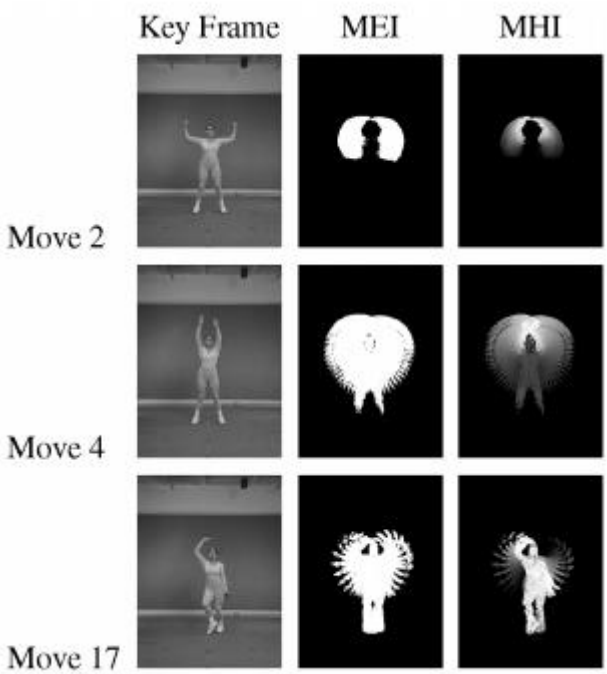
[۱۰] Space-Time
[۱۱] Sequential
[۱۲] Sliding Window
[۱۳] Similarity Measure

انرژی حرکت[۱۴] و تصویر تاریخچه‌ی حرکت[۱۵] را به عنوان معیار شباهت استخراج کرده‌اند. همان‌طور که مشاهده می‌شود ویدئوی وسطی و پایینی انرژی حرکت نسبتا مشابهی دارند و ویدئوی وسطی و بالایی تاریخچه حرکت مشابهی دارند. بنابراین استفاده از هر دو معیار برای شناسایی بهتر فعالیت ضروری است. این روش به دلیل حجم پایین پردازش به صورت زمان‌واقعی[۱۶] اجرا می‌شود.

روش‌های سراسری[۱۷] مانند آن‌چه در [۱۳] اعمال شد، به نویز و تغییرات پس‌زمینه حساس هستند و برای ویدئوهای نویزی یا ویدئوهایی که در آن‌ها دوربین حرکت می‌کند مناسب نیستند. در مقابل، روش‌های محلی[۱۸] به این موارد مقاوم هستند[۱۴]. مقاله‌ی [۱۵] از روش محلی برای شناسایی فعالیت استفاده می‌کند. شکل (۲-۲) روش استفاده شده در این مقاله را نشان می‌دهد. این مقاله، ویژگی جریان نوری[۱۹] را در بخش‌هایی[۲۰] به ابعاد ۳۰*۳۰*۳۰ از ویدئو (۳۰ فریم و در هر فریم پنجره های ۳۰*۳۰) بدست می‌آورد و بر این اساس الگوی T را از روی داده‌های آموزشی ایجاد می‌کند. آن‌گاه از هر نقطه در حجم ویدئوی V تکه‌ی S را استخراج می‌کند و همبستگی بین S و T را محاسبه می‌کند و از آن به عنوان معیار شباهت استفاده می‌کند. علاوه بر این، همه‌ی همبستگی‌های بدست آمده بین T و Sها را با یکدیگر جمع می‌کند تا یک معیار سراسری نیز بدست آورد.

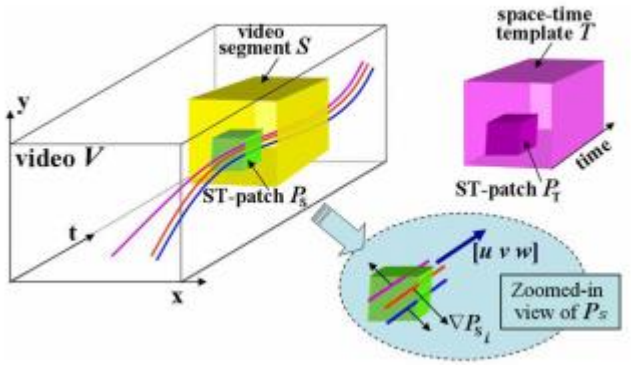
---

۱۴ Motion Energy Image (MEI)
۱۵ Motion History Image (MHI)
۱۶ Real Time
۱۷ Global
۱۸ Local
۱۹ Optical Flow
۲۰ Patches




شکل (۲ - ۱) تصاویر انرژی حرکت و تاریخچه‌ی حرکت [۱۳].

عیب روش [۱۵] نسبت به [۱۳] در حجم و زمان پردازش است؛ اما نسبت به تغییرات مقاوم‌تر است. دلیل اصلی بالا بودن حجم پردازش [۱۵]، تعداد حالات انتخاب بخش‌هایی است که با اشتراک از هر ویدئو استخراج می‌شود.




شکل (۲ - ۲) نمایش روش استفاده شده توسط [۱۵] . در این روش با استخراج جریان نوری و استفاده از یک معیار شباهت، فعالیت‌های مشابه شناسایی می‌گردند[۱۶].

### ۲-۲-۱-۲ مسیر حرکت در فضا-زمان

در این روش‌ها، مسیر حرکت[۲۱] اجسام یا اعضای بدن انسان در فضا-زمان محاسبه می‌شود و شناسایی فعالیت بر اساس این مسیر انجام می‌شود. این مسیرها ممکن است در فضای دوبعدی یا سه‌بعدی اندازه‌گیری شود. به عنوان مثال [۱۷] با تشخیص مسیر حرکت دست انسان در تصویر دوبعدی، فعالیت را شناسایی می‌کند. مقاله‌ی [۱۸] مسیر حرکت مفاصل انسان را در فضای سه‌بعدی بدست می‌آورد و از آن برای شناسایی فعالیت استفاده می‌کند.

مزیت استفاده از مسیر حرکت آن است، که با کاهش حجم ویژگی‌ها و اجتناب از ویژگی‌های کم‌ارزش، حجم پردازش کاهش می‌یابد؛ اما تشخیص مسیر حرکت نیازمند یافتن اعضای بدن و سپس نقاط مفاصل آن در هر فریم است که در تصاویری که تنها با یک دوربین ثبت شده‌اند، پیچیده و مشکل است. اما در صورتی که تصاویر ثبت شده دارای مولفه‌ی عمق باشند، این مشکل تسهیل خواهد شد. به عنوان مثال [۱۹] از تصاویر سه‌بعدی ثبت شده توسط دوربین کینکت[۲۲] برای شناسایی فعالیت استفاده می‌کند.

### ۲-۲-۱-۳ ویژگی‌های محلی در فضا-زمان

در این روش، ویژگی‌ها در پنجره‌های کوچک دوبعدی یا سه‌بعدی محلی استخراج می‌شود. این پنجره‌ها ممکن است به صورت متراکم[۲۳] نمونه‌گیری شوند، یا اینکه صرفا در نقاط مهم فضا-زمان[۲۴] که با الگوریتم‌هایی مانند روش هریس[۲۵] شناسایی می‌شود، انتخاب شوند. ویژگی‌های محلی نسبت به نویز، اغتشاشات پس‌زمینه، حرکت دوربین و انسداد[۲۶] مقاوم هستند و برخی از آن‌ها به چرخش

---

[۲۱] Trajectory
[۲۲] Kinect
[۲۳] Dense
[۲۴] Space-Time Interest Point (STIP)
[۲۵] Harris
[۲۶] Occlusion

یا تغییر مقیاس هم مقاوم هستند. در فصل سوم برخی از ویژگی‌های محلی پرکاربرد در شناسایی فعالیت را شرح خواهیم داد.

معمولا لازم است ویژگی‌های بدست آمده به‌گونه‌ای کد شوند. برخی روش‌های مورد استفاده در پردازش زبان طبیعی[۲۷]، مانند کیف کلمات[۲۸] یا بردار فیشر[۲۹] برای این منظور مورد استفاده قرار می‌گیرد. این روش‌ها در فصل سوم معرفی خواهند شد.

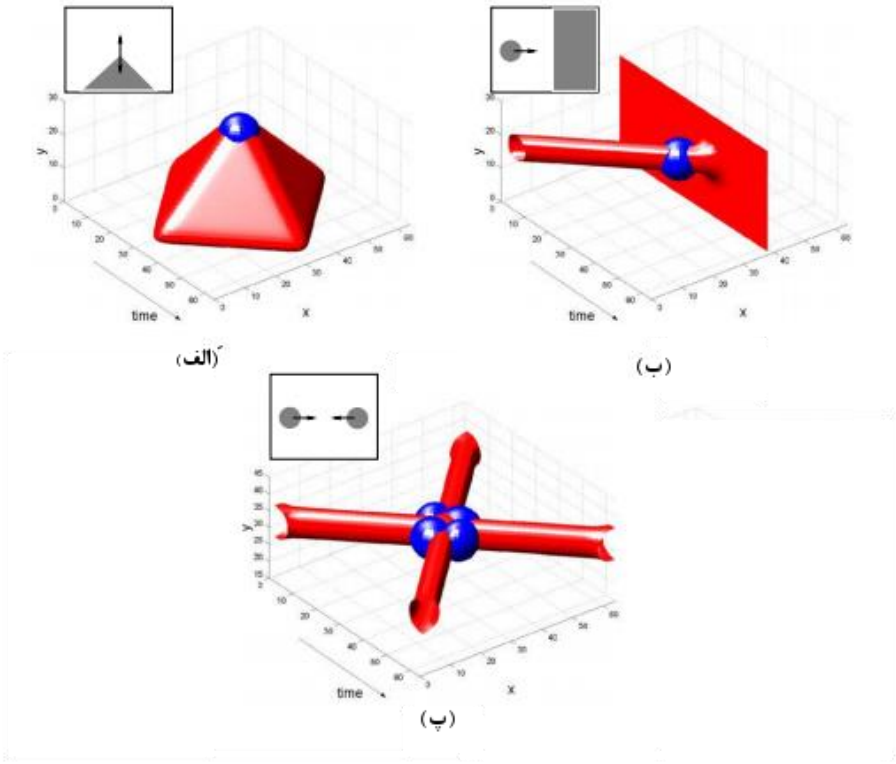
به عنوان مثال [۲۰] در تکه‌های محلی در مدل فضا-زمانی از ویدئو، گرادیان را در جهت‌های مختلف محاسبه می‌کند. این مقاله مقیاس‌های زمانی[۳۰] متفاوتی را در نظر می‌گیرد و بردارهای گرادیان[۳۱] بدست آمده را به هم الحاق می‌کند تا یک بردار 3L بعدی ایجاد کند که L تعداد مقیاس‌های زمانی است. با توجه به اینکه در شناسایی حرکت، تنها جهت[۳۲] گرادیان اهمیت دارد، بردارهای گرادیان نرمالیزه می‌شوند. سپس با بهره‌گیری از روش کیف کلمات، هیستوگرام این بردارها بدست آمده و از آن برای شناسایی فعالیت استفاده می‌شود. شناسایی فعالیت بر اساس معیار فاصله بین هیستوگرام‌ها انجام می‌شود. برای جلوگیری از تشخیص نقاط بی‌حرکت پس‌زمینه، این مقاله نقاطی که میزان تغییرات آن‌ها از یک آستانه کمتر باشد را نادیده می‌گیرد.

مقاله‌ی [۲۱] روشی برای محاسبه‌ی نقاط مهم فضا-زمانی مستقل از مقیاس ارائه می‌کند. شکل (۲-۳) نتیجه‌ی شناسایی این نقاط را در چند ویدئوی ساختگی نمایش می‌دهد. مقاله‌ی [۲۲] با استفاده از روش [۲۱] نقاط مهم فضا-زمانی را شناسایی کرده و در آن‌ها ویژگی‌های محلی استخراج کرده است. سپس با استفاده از ماشین بردار پشتیبان[۳۳]، فعالیت‌ها را شناسایی کرده است. مقاله‌ی [۲۳] نیز از روش [۲۱] برای شناسایی STIP استفاده می‌کند و در آن نقاط ویژگی محلی استخراج

---

۲۷ Natural Language Processing
۲۸ Bag of Words (BoW)
۲۹ Fisher Vector
۳۰ Temporal Scale
۳۱ Gradient
۳۲ Orientation
۳۳ Support Vector Machine (SVM)

می‌کند. سپس با استفاده از تحلیل مولفه‌های اصلی[۳۴] ابعاد ویژگی‌ها را کاهش می‌دهد و با استفاده از روش BoW هیستوگرام ویژگی‌ها را تشکیل می‌دهد. در نهایت با استفاده از SVM فعالیت را شناسایی می‌کند.

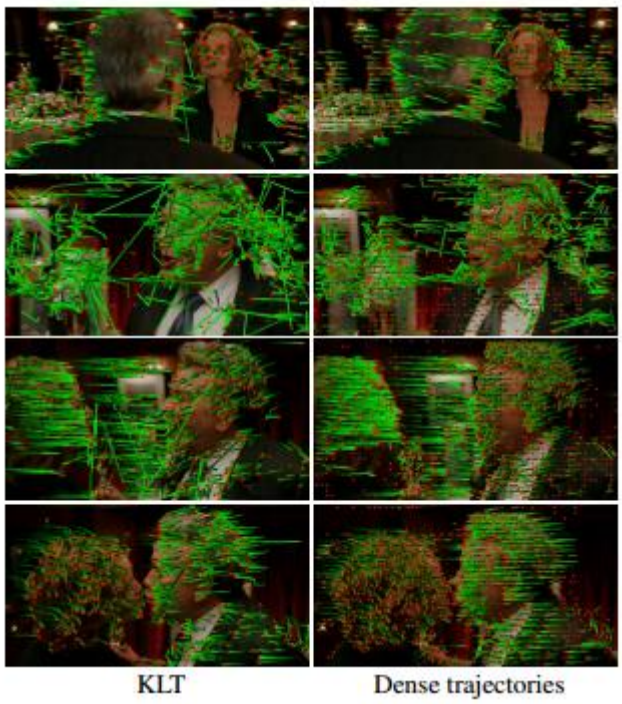



شکل (۲ - ۳) شناسایی نقاط مهم فضا-زمانی با روش [۲۰] در چند ویدئوی ساختگی. الف) حرکت گوشه‌ی یک جسم. ب) برخورد توپ به دیوار. پ) برخورد دو توپ

یکی از روش‌های پرطرفدار استخراج ویژگی محلی برای شناسایی فعالیت که نخستین بار در [۲۴] معرفی شد، مسیر متراکم[۳۵] نام دارد. این روش در نقاطی که به صورت متراکم انتخاب می‌شوند، ویژگی‌های هیستوگرام گرادیان‌های جهت‌دار[۳۶]، هیستوگرام جریان نوری[۳۷] و هیستوگرام مرز حرکت[۳۸] و هم‌چنین مسیر حرکت تا ۱۵ فریم بعدی را بدست می‌آورد. سپس این ویژگی‌ها را به هم

[۳۴] Principle Component Analysis (PCA)
[۳۵] Dense Trajectory
[۳۶] Histogram of Oriented Gradients (HOG)
[۳۷] Histogram of Optical Flow (HOF)
[۳۸] Motion Boundary Histogram (MBH)

الحاق می‌کند تا یک بردار ویژگی به طول ۴۲۶ بدست آورد. این روش در مقایسه با روش‌های تبدیل ویژگی مقیاس‌نابسته[۳۹] [۲۵] و ردیاب کانادی-لوکاس-توماسی[۴۰] [۲۶] نسبت به حرکت‌های نامنظم ناگهانی[۴۱] مقاوم‌تر است. به عنوان مثال شکل (۲-۴) مقایسه‌ی این روش با روش KLT را نمایش می‌دهد.




شکل (۲ - ۴) مقایسه‌ی روش مسیر متراکم با روش KLT . مشاهده می‌گردد که روش مسیر متراکم نسبت به حرکت‌های نامنظم ناگهانی مقاوم‌تر است [۲۴].

با وجود مقاوم بودن روش [۲۴] نسبت به حرکت‌های نامنظم ناگهانی و هم‌چنین استفاده از ویژگی MBH که به حرکت مقاوم است، باز هم در دنیای واقعی حرکت دوربین می‌تواند موجب استخراج ویژگی‌های نامربوط گردد. مقاله‌ی [۲۷] با تخمین حرکت دوربین، تمام ویژگی‌های ناشی از آن را حذف می‌کند. این مقاله نام روش خود را مسیر متراکم بهبودیافته[۴۲] نهاده است.در این پایان‌نامه ما

[۳۹] Scale-Invariant Feature Transform (SIFT)
[۴۰] Kanade–Lucas–Tomasi (KLT) Tracker
[۴۱] Irregular Abrupt Motions
[۴۲] Improved Dense Trajectory (IDT)

از روش [۲۷] برای استخراج ویژگی استفاده خواهیم نمود. روش‌های مسیر متراکم و مسیر متراکم بهبود یافته در فصل سوم شرح داده می‌شود.

## ۲-۲-۲ رویکرد ترتیبی

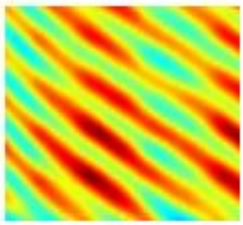
در روش‌های ترتیبی، ویژگی‌ها به گونه‌ای استخراج می‌شود که اطلاعات زمانی ویدئو حفظ شود. در این روش‌ها از هر فریم یا تعداد معدودی از فریم‌ها ویژگی استخراج می‌شود. سپس این بردارهای ویژگی به ترتیب زمانی در کنار هم قرار می‌گیرند. مثلا اگر هر بردار ویژگی D بعد داشته باشد و تعداد بردارهای ویژگی N باشد، ویژگی نهایی یک ماتریس D*N است که اطلاعات زمانی ویدئو را دربردارد. برخی از روش‌ها، شناسایی فعالیت را مستقیما از روی خود ویژگی‌ها و با استفاده از یک معیار شباهت انجام می‌دهند. اما برخی دیگر از مدل‌های حالت[۴۳] (مانند مدل مخفی مارکوف[۴۴] [۲۸] یا شبکه بیزین پویا[۴۵] [۲۹]) برای تخمین احتمال ایجاد شدن یک توالی از ویژگی‌های ورودی توسط فعالیت مورد نظر استفاده می‌کنند.

به عنوان مثال [۳۰] ویژگی جریان نوری را در هر فریم از ویدئوهای آزمایش و آموزش محاسبه می‌کند. سپس معیار شباهت را بین هر فریم از ویدئوی آموزش با فریم متناظر در ویدئوی آزمایش محاسبه می‌کند. در نهایت این معیارهای شباهت را با یکدیگر جمع می‌کند. به گفته‌ی مقاله اگر الگوهای مورب مشابه شکل (۲-۵) در تصویر حاصل وجود داشته باشد، فعالیت‌های دو ویدئو مشابه هستند. این روش حتی در زمانی که دوربین حرکت می‌کند نیز نتیجه‌ی خوبی داشته است.

اما مقاله‌ی [۳۱] پس از استخراج یک توالی از ویژگی‌ها، با استفاده از مدل مارکوف با طول متغیر[۴۶] [۳۲] فعالیت‌ها را شناسایی کرده است.

---

[۴۳] State Models
[۴۴] Hidden Markov Model (HMM)
[۴۵] Dynamic Bayesian Network (DBN)
[۴۶] Variable-Length Markov Models (VLMM)



شکل (۲ - ۵) الگوهای مورب بیان‌گر شباهت دو فعالیت [۲۸].

مقاله‌ی [۳۳] نشان می‌دهد هرچه طول زمانی پنجره‌ها بیشتر شود، مدل میدان تصادفی شرطی[۴۷] برای شناسایی فعالیت از HMM بهتر عمل خواهد کرد.

مقاله‌ی [۳۴] یک تابع هسته[۴۸] برای SVM معرفی می‌کند که در آن توالی زمانی ویژگی‌ها در نظر گرفته شده است. با استفاده از این تابع هسته، رشته‌های ورودی می‌توانند طول‌های نابرابر داشته باشند. شکل (۲-۶) دو فعالیت سوارشدن در خودرو و پیاده‌شدن از خودرو را نشان می‌دهد که این روش توانسته است به خوبی آن‌ها را از یکدیگر تشخیص دهد. روش این مقاله در فصل سوم تشریح خواهد شد.

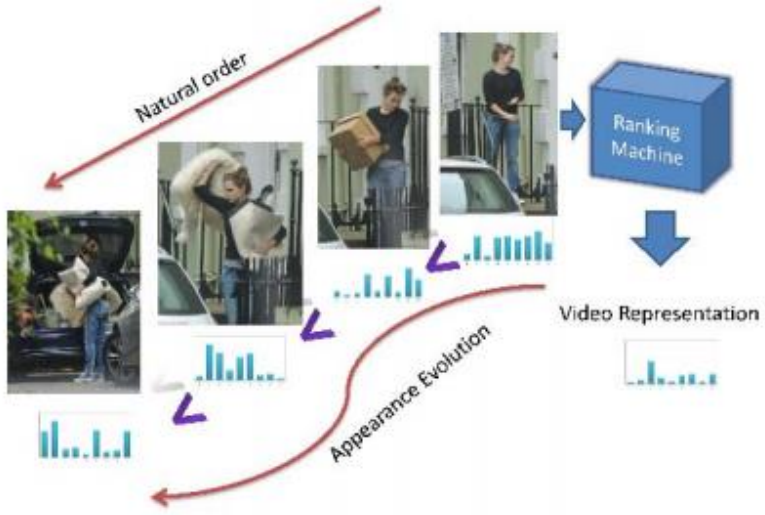

شکل (۲ - ۶) دو فعالیت سوارشدن در خودرو و پیاده‌شدن از آن [۳۴].

مقاله‌ی [۳۵] از هر فریم ویدئو یک بردار ویژگی استخراج می‌کند. سپس بر اساس این ویژگی‌ها و ترتیب آن‌ها، با بکارگیری ماشین رتبه‌بندی[۴۹]، یک تابع رتبه‌بندی[۵۰] بهینه را بدست می‌آورد. و

---

[۴۷] Conditional Random Field (CRF)
[۴۸] Kernel Function
[۴۹] Ranking Machine
[۵۰] Ranking Function

پارامترهای آن تابع به عنوان بردار ویژگی کل ویدئو درنظر گرفته می‌شود. بنابراین هر ویدئو به یک بردار ویژگی تبدیل می‌شود که دربردارنده‌ی اطلاعات زمانی آن ویدئو است. این بردارها به ورودی SVM وارد شده و فعالیت شناسایی می‌شود. مقاله نام روش خود را ویدئوداروین[۵۱] یا ادغام رتبه[۵۲] نهاده است. شکل (۲-۷) نمای کلی این روش را نشان می‌دهد. این روش نیز در فصل سوم شرح داده خواهد شد.




شکل (۲ - ۷) نمایش روش ویدئوداروین. ماشین رتبه‌بندی ترتیب وقایع در ویدئو را یاد می‌گیرد و در نهایت یک بردار ویژگی که دربردارنده‌ی این ترتیب است ارائه می‌کند[۳۵][۳۶].

## ۲-۳ روش‌های سلسله‌مراتبی

در روش‌های سلسله‌مراتبی فعالیت به تعدادی زیرفعالیت تقسیم می‌شود که به آن‌ها عمل‌های بنیادی[۵۳] گفته می‌شود. به عنوان مثال فعالیت «غذا پختن» از زیرفعالیت‌های «ورود به آشپزخانه»، «استفاده از ظروف» و «کار با اجاق گاز» تشکیل می‌شود [۱۱].

مقاله‌ی [۳۷] پس از شناسایی عمل‌های بنیادی، سعی می‌کند با استفاده از ساختار نحوی، با آن‌ها یک رشته جمله بسازد که این جمله بیان‌گر فعالیت مورد نظر است.

---

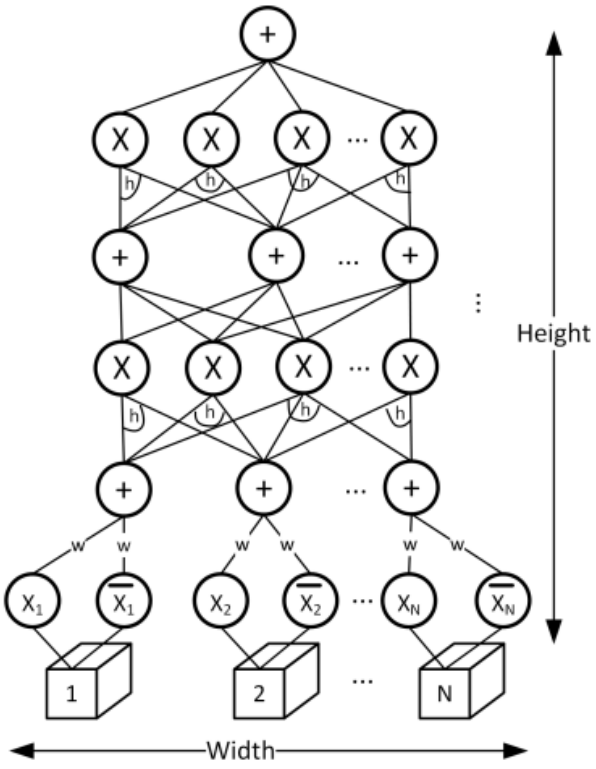
۵۱ VideoDarwin
۵۲ Rank Pooling
۵۳ Atomic Actions

مقاله‌ی [۳۸] با یافتن ارتباط فضا-زمانی عمل‌های بنیادی، فعالیت اصلی را شناسایی می‌کند. روش‌های دیگر، از مدل‌های آماری استفاده می‌کنند. مقاله‌ی [۳۹] از HMM سلسله‌مراتبی برای شناسایی فعالیت‌های روزمره استفاده می‌کند. در لایه‌ی اول عمل‌های بنیادی شناسایی می‌شود. سپس در لایه‌ی بعدی فعالیت اصلی کشف می‌گردد.

مقاله‌ی [۴۰] روش BoW را با یک شبکه‌ی چند لایه‌ی جمع-ضرب[۵۴] [۴۱] ترکیب می‌کند. این مقاله مدل فضا-زمانی ویدئو را به بخش‌های کوچکتر تقسیم می‌کند و در این بخش‌ها که در مقیاس‌های مختلف و به صورت متراکم انتخاب می‌شوند، ویژگی استخراج می‌کند. سپس با استفاده از روش BoW توزیع کلمات موجود در هر تکه را بدست می‌آورد. سپس با استفاده از یک تابع ریاضی از هر یک از این توزیع‌ها یک عدد بدست آورده و آن را به شبکه‌ی SPN می‌دهد.

شکل (۲-۸) ساختار روش این مقاله را نشان می‌دهد. برای هریک از فعالیت‌هایی که قرار است شناسایی شود، باید مدلی مانند شکل (۲-۸) ساخته شود. در پایین‌ترین لایه اعداد بدست‌آمده از توزیع کلمات هریک از N تکه وارد شبکه می‌شود. در لایه‌ی دوم برای هر تکه یکی از گره‌ها صفر و دیگری یک است. این لایه تعیین می‌کند که این تکه به پس‌زمینه تعلق دارد یا به سوژه‌ای که فعالیت را انجام می‌دهد. بالاترین لایه یک امتیاز بدست می‌دهد که این امتیاز برای مدلی که فعالیت به آن تعلق دارد بیشینه می‌شود. در ابتدا تمام گره‌های لایه دوم یک فرض شده و ضمن حرکت از پایین به بالا فعالیت شناسایی می‌شود. سپس با حرکت از بالا به پایین، مقادیر لایه‌ی دوم و در نتیجه تکه‌هایی که فعالیت در آن رخ داده است شناسایی می‌گردد.

[۵۴] Hierarchical Sum-Product Network (SPN)




شکل (۲ - ۸) ساختار روش مقاله‌ی [۴۰].

شکل (۲-۹) تکه‌هایی که فعالیت در آن‌جا انجام می‌شود را نشان می‌دهد که از روش گفته شده شناسایی شده‌اند.

شکل (۲ - ۹) شناسایی تکه‌هایی که فعالیت در آن‌ها رخ داده است[۴۰][۴۲].

## ۲-۴ جمع‌بندی

روش‌های معرفی شده را می‌توان در جدول (۲-۱) خلاصه کرد.

جدول (۲ - ۱) خلاصه‌ی روش‌های معرفی شده.

| روش تک‌لایه | | روش سلسله‌مراتبی |
|---|---|---|
| **رویکرد فضا-زمانی** | **رویکرد ترتیبی** | ساختار نحوی<br>توصیف در فضا-زمان<br>روش آماری |
| **حجم فضا-زمانی**<br>**مسیر حرکت**<br>**ویژگی محلی** | **معیار شباهت**<br>**مدل حالت** | |

در میان رویکردهای فضا-زمانی، استفاده از مسیر حرکت از لحاظ حجم پردازش از روش‌های دیگر به‌صرفه‌تر است؛ اما بدون در اختیار داشتن عمق تصویر مشکل است. ویژگی‌های محلی پرکاربردترین روش در این میان است که اساس خیلی از روش‌های ترتیبی و روش‌های سلسله‌مراتبی نیز قرار می‌گیرد.

اگرچه روش‌های فضا-زمانی هم بعد زمان را در نظر می‌گیرند، اما واقعیت آن است که بعد زمان با بعد مکان تفاوت دارد و لازم است به‌گونه‌ای دیگر مورد بررسی قرار گیرد. بر این اساس روش‌های ترتیبی در فعالیت‌هایی که ترتیب زمانی در آن‌ها اهمیت دارد بهتر عمل می‌کنند. روش‌های سلسله‌مراتبی برای فعالیت‌های طولانی و چندمرحله‌ای نتیجه‌ی بهتری دارند؛ اما موجب افزایش حجم پردازش و زمان آن می‌گردند.

# فصل سوم : مبانی نظری

## ۳-۱ مقدمه

در این فصل ابتدا برخی از روش‌های معمول استخراج ویژگی برای کاربرد شناسایی فعالیت معرفی می‌شود و برخی از روش‌های پردازش زبان طبیعی برای کد کردن ویژگی‌ها شرح داده می‌شود. سپس دو روش اخیر به نام‌های تابع هسته‌ی انعطاف‌پذیر با زمان[۵۵] و ادغام رتبه (یا ویدئوداروین) برای شناسایی فعالیت معرفی و تشریح می‌گردد.

## ۳-۲ ویژگی‌های محلی

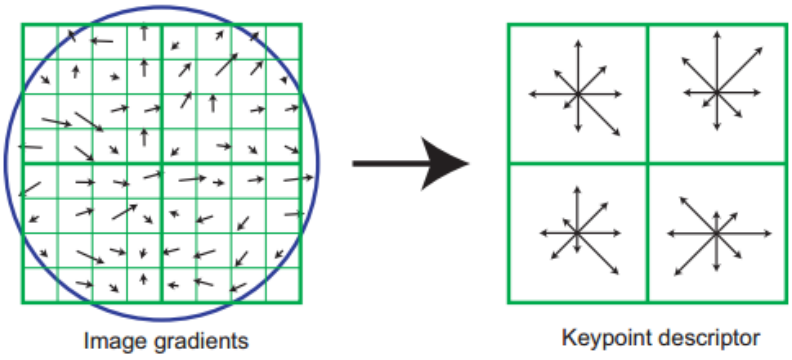
### ۳-۲-۱ هیستوگرام گرادیان‌های جهت‌دار

گردایان یکی از مفاهیم اصلی پردازش تصویر است.گرادیان تصویر در یک جهت، بیان‌گر تغییرات شدت روشنایی یا رنگ تصویر در آن جهت است. اندازه و جهت گرادیان به ترتیب از رابطه‌های (۳-۱) و (۳-۲) بدست می‌آید که در آن I(x,y) شدت روشنایی در پیکسلی به مختصات (x,y) است.

$$\boldsymbol{M(x,y) = \sqrt{[\,I(x+1,y) - I(x-1,y)]^2 + [\,I(x,y+1) - I(x,y-1)]^2}} \qquad (۱-۳)$$

$$\boldsymbol{\theta(x,y) = tan^{-1}\frac{I(x,y+1)-I(x,y-1)}{I(x+1,y)-I(x-1,y)}} \qquad (۲-۳)$$

[۵۵] Time Flexible Kernel (TFK)




شکل (۳ - ۱) کوانتیزه کردن جهت‌های گرادیان [۴۳].

هیستوگرام گرادیان که نخستین بار بدون خواندن آن به این نام توسط [۴۴] معرفی شد، از طریق شمردن تعداد تکرار هریک از جهت‌های گرادیان در یک تکه محلی از تصویر بدست می‌آید. برای این منظور، معمولا مانند شکل (۳-۱) جهت‌ها به ۸ جهت کوانتیزه می‌شوند.

## ۳-۲-۲ هیستوگرام جریان نوری

جریان نوری در تصویر، برداری دوبعدی است که بیانگر جهت حرکت ظاهری اجزای سازنده‌ی تصویر (مانند اجسام، لبه‌ها، اشخاص و ..) بین دو فریم متوالی است. این بردار با این فرض محاسبه می‌شود که شدت روشنایی اجسام در فریم‌های متوالی تغییر نمی‌کند و پیکسل‌های همسایه جهت حرکت یکسانی دارند[۴۵]. فرض کنیم پیکسلی به مختصات (x,y,t) به اندازه‌ی (dx,dy,dt) حرکت کرده است. در این صورت اگر I شدت روشنایی این پیکسل باشد داریم:

$$\boldsymbol{I(x,y,t) = I(x+dx, y+dy, t+dt)} \qquad (۳-۳)$$

از طرفی با استفاده از بسط تیلور می‌توان عبارت سمت راست رابطه‌ی (۳-۳) را به صورت رابطه‌ی (۳-۴) نوشت.

$$\boldsymbol{I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt} \qquad (۳-۴)$$

از مقایسه‌ی روابط (۳-۳) و (۳-۴) داریم:

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt = 0 \qquad (۵-۳)$$

با تقسیم طرفین رابطه‌ی (۳-۵) به dt داریم:

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0 \qquad (۶-۳)$$

که $V_x$ و $V_y$ مولفه‌های سرعت یا همان جریان نوری در جهت‌های x وy هستند. در معادله‌ی (۳-۶) دو مجهول وجود دارد و به تنهایی قابل حل نیست. برای حل آن می‌توان از روش لوکاس-کانادی[۵۶] [۲۶][۴۶] استفاده نمود. در این روش فرض می‌کنیم در یک تکه‌ی ۳*۳ اطراف آن نقطه از تصویر، جریان نوری با آن نقطه مشابه است. بنابراین ۹ نقطه داریم که همگی $V_x$ و $V_y$ یکسانی دارند. لازم است دستگاه معادلات به فرم (۳-۷) را حل نمود.

$$S\begin{bmatrix} v_x \\ v_t \end{bmatrix} + B = 0 \qquad (۷-۳)$$

که در آن:

$$S = \begin{bmatrix} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} \end{bmatrix}, B = \begin{bmatrix} \frac{\partial I}{\partial t} \end{bmatrix} \qquad (۸-۳)$$

پاسخ معادله‌ی (۳-۷) به صورت (۳-۹) خواهد بود.

$$\begin{bmatrix} v_x \\ v_t \end{bmatrix} = (S^T S)^{-1} S^T B \qquad (۹-۳)$$

پس از بدست آوردن بردار جریان نوری، جهت این بردارها به ۹ جهت کوانتیزه می‌شود و هیستوگرام آن‌ها برای هر تکه از تصویر محاسبه می‌شود. جهت نهم زاویه‌ی صفر درجه است که بردارهایی که بزرگی‌شان از آستانه‌ای کمتر باشد به آن کوانتیزه می‌شوند.

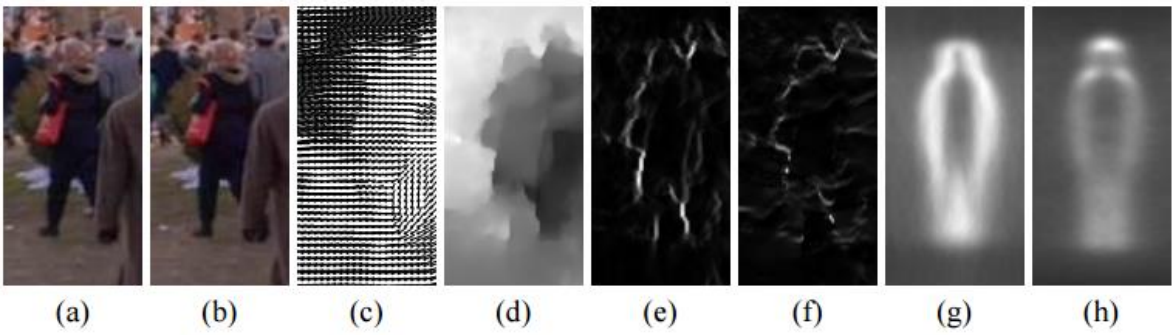
[۵۶] Lucas-Kanade

## ۳-۲-۳ هیستوگرام مرز حرکت

جریان نوری نسبت به حرکت‌های ناشی از لرزش یا حرکت دوربین یا تغییرات نامربوط دیگر مقاوم نیست. اما روش MBH می‌تواند تغییرات ناشی از حرکت خفیف دوربین را حذف کند. این روش با مشتق گرفتن از جریان نوری این تغییرات را حذف می‌کند.

در این روش ابتدا جریان نوری تصویر محاسبه می‌شود. سپس هرکدام از $V_x$ و $V_y$ به عنوان دو تصویر جداگانه درنظر گرفته شده و از هرکدام گرادیان گرفته می‌شود. سپس جهت‌های گرادیان کوانتیزه می‌شود و مطابق آن‌چه در بخش ۳-۲-۱ شرح داده شد، هیستوگرام گرادیان برای هر یک از این دو تصویر محاسبه می‌شود. این دو هیستوگرام که به ترتیب $MBH_x$ و $MBH_y$ نامیده می‌شوند، به عنوان ویژگی نهایی مورد استفاده قرار می‌گیرند.

شکل (۳-۲) این ویژگی‌ها را نمایش می‌دهد.



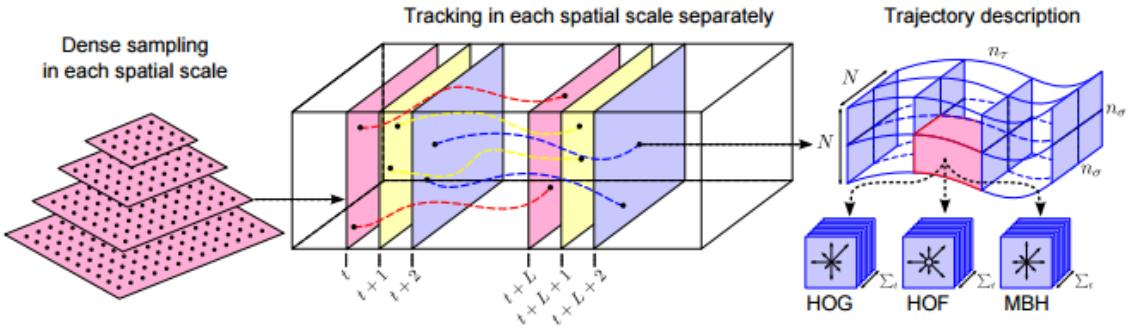

شکل (۳ - ۲) استخراج ویژگی MBH [۴۷]. دو تصویر a و b، دو فریم متوالی از ویدئو هستند. دو تصویر c و d به ترتیب جهت و اندازه‌ی جریان نوری هستند. تصویرهای e و f به ترتیب گرادیان جریان نوری در جهت x و y هستند. دو تصویر g و h هم به ترتیب میانگین گرادیان در جهت x در تمام فریم‌ها و میانگین گرادیان در جهت y در تمام فریم‌ها هستند.

## ۴-۲-۳ مسیر متراکم

در این روش[۲۴]، مسیر حرکت به همراه سه ویژگی معرفی شده در بخش‌های ۳-۲-۱ و ۳-۲-۲ و ۳-۲-۳ به گونه‌ای با یکدیگر ترکیب می‌شوند. این روش ساختار ترتیبی فریم‌ها در ویدئو را به خوبی

در نظر می‌گیرد و نسبت به حرکت‌های شدید ناگهانی نیز مقاوم است. در ادامه جزییات این روش شرح داده می‌شود.

به طور کلی محاسبه‌ی ویژگی‌ها طی سه مرحله انجام می‌پذیرد. شکل (۳-۳) ساختار این روش را نمایش می‌دهد. در وهله‌ی اول از هر فریم ویدئو در مقیاس‌های مکانی متفاوت به صورت متراکم نمونه‌گیری می‌شود. نمونه‌گیری در شبکه‌ای[۵۷] که نقاط آن به اندازه‌ی W از هم فاصله دارند انجام می‌شود. مقدار W به صورت تجربی بدست می‌آید. در همین مرحله، نقاط همگن که احتمالا متعلق به پس‌زمینه هستند حذف می‌شوند تا ضمن کاهش حجم محاسبات، توجه بر نقاطی که مربوط به فعالیت هستند متمرکز گردد. برای این منظور، ماتریس خودهمبستگی[۵۸] در همه‌ی نقاط محاسبه شده و مقدار ویژه و بردار ویژه‌ی آن بدست می‌آید. نقاطی که مقدار ویژه برای آن‌ها از آستانه‌ای کمتر شود در ادامه‌ی کار نادیده گرفته می‌شوند.




شکل (۳ - ۳) ویژگی مسیر متراکم طی سه مرحله استخراج می‌شود [۲۴].

در مرحله‌ی دوم به ازای هر کدام از مقیاس‌های مکانی، مسیر حرکت هر پیکسل تا L فریم بعد از آن کشف می‌شود. برای این‌کار از جریان نوری متراکم استفاده می‌شود. یعنی جریان نوری در هر پیکسل محاسبه می‌شود. با فرض در اختیار داشتن جریان نوری در پیکسلی به مختصات (x,y) از

---

۵۷ Grid
۵۸ Autocorrelation

فریم t به صورت رابطه‌ی (۳-۱۰)، مکان جدید این پیکسل در فریم بعدی را می‌توان با استفاده از رابطه‌ی (۳-۱۱) محاسبه نمود که در آن M هسته‌ی فیلتر میانه[۵۹] است.

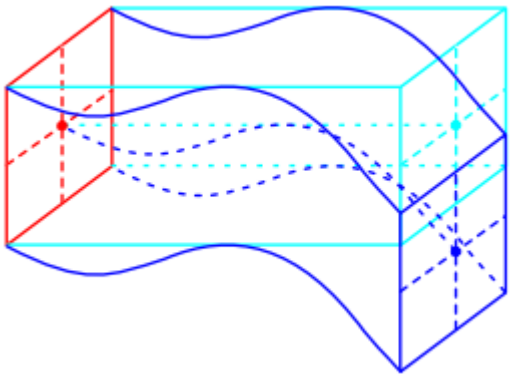
$$\boldsymbol{w(x_t, y_t) = (u_t, v_t)} \qquad (۳-۱۰)$$

$$\boldsymbol{P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{(x_t, y_t)}} \qquad (۳-۱۱)$$

پس از محاسبه‌ی L نقطه‌ی بعدی این نقاط مطابق (۳-۱۲) کنار هم قرار داده می‌شود تا بیان‌گر مسیر حرکت آن نقطه باشد. مقاله‌ی [۲۴] مقدار L=15 را پیشنهاد کرده است.

$$\boldsymbol{(P_t, P_{t+1}, P_{t+2}, \dots)} \qquad (۳-۱۲)$$

در مرحله‌ی سوم در اطراف هر یک از این L نقطه، یک پنجره به ابعاد N*N در نظر گرفته می‌شود تا حجمی خمیده مانند شکل (۳-۴) بدست آید (مقاله [۲۴] مقدار N=32 را پیشنهاد کرده است).



شکل (۳ - ۴) حجم خمیده اطراف مسیر حرکت [۴۸].

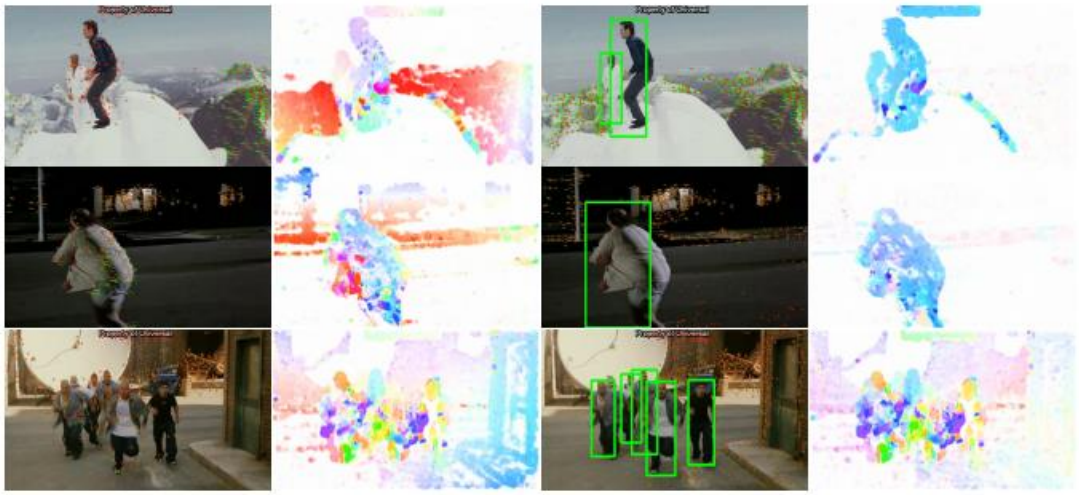
این حجم به پنجره‌هایی کوچک‌تری تقسیم شده و در هریک از این پنجره‌ها ویژگی‌های HOG و HOF و MBH محاسبه می‌شود (پیشنهاد [۲۴]، انتخاب پنجره‌های کوچکی به ابعاد ۲*۲ و به طول ۳ فریم بوده است). در نهایت در هریک از این پنجره‌های کوچک یک بردار ۴۲۶بعدی استخراج می‌شود که ۳۰ مولفه‌ی آن مربوط به مسیر حرکت، ۹۶ مولفه مربوط به HOG، ۱۰۸ مولفه مربوط به HOF، ۹۶ مولفه مربوط به $MBH_x$ و ۹۶ مولفه‌ی دیگر مربوط به $MBH_y$ است.

[۵۹] Median Filtering Kernel

### ۳-۲-۵ مسیر متراکم بهبودیافته

همان‌طور که در بخش ۳-۲-۳ گفته شد، ویژگی MBH نسبت به حرکت خفیف دوربین مقاوم است و استفاده از این ویژگی در روش مسیر متراکم موجب حذف این حرکت‌ها می‌شود. اما هنوز هم ممکن است حرکت دوربین در ویدئوهای واقعی بر ویژگی‌های استخراج شده تاثیر بگذارد. در این روش با فرض ناچیز بودن تغییرات بین دو فریم متوالی، از فریم‌های متوالی برای محاسبه‌ی هموگرافی[۶۰] و در نتیجه حذف حرکت دوربین استفاده می‌شود. محاسبه‌ی هموگرافی با استفاده از روش اجماع تصادفی نمونه[۶۱] [۴۹] انجام می‌شود.

در مواقعی که عمده‌ی تصویر موجود در فریم، تصویر انسانی که در حال انجام فعالیت است باشد، ممکن است حرکت‌های انسان با حرکت دوربین اشتباه گرفته شده و حذف شود. برای جلوگیری از این مشکل، این روش از یک ردیاب انسان[۶۲] در فریم‌ها استفاده می‌کند. شکل (۳-۵) تفاوت استفاده یا عدم استفاده از ردیاب انسان را نشان می‌دهد.



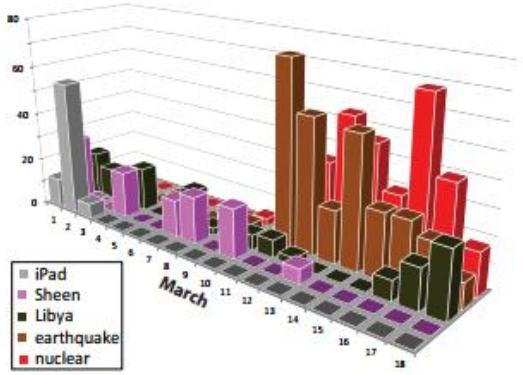
شکل (۳ - ۵) نمایش مزیت استفاده از ردیاب انسان [۲۷]. سمت چپ بدون ردیاب و سمت راست با ردیاب هموگرافی را محاسبه کرده است.

---

۶۰ Homography
۶۱ Random Sample Consensus (RANSAC)
۶۲ Human Tracker

## ۳-۳ کد کردن ویژگی‌ها

یکی از روش‌های پرکاربرد در پردازش زبان طبیعی، تصمیم‌گیری بر اساس دفعات تکرار کلمات مختلف در متن است.




شکل (۳ - ۶) - دفعات تکرار پنج کلمه مختلف در خبرگزاری سی‌ان‌ان در ۱۸ روز ابتدای ماه مارس۲۰۱۱ [۵۰].

معمولا یک متن به صورت کیف کلمات توصیف می‌شود و بر این اساس محتوای متن شناسایی می‌شود. به عنوان مثال شکل (۳-۶) دفعات تکرار پنج کلمه را در تاریخ‌های مختلف در وبلاگ خبرگزاری سی‌ان‌ان نشان می‌دهد.

همین روش‌ها را می‌توان برای کد کردن ویژگی‌های استخراج شده از ویدئو به منظور شناسایی فعالیت به‌کار برد. به طور کلی در این روش‌ها تعداد دفعات تکرار یک ویژگی در تصویر شمرده می‌شود. برای این‌کار لازم است ابتدا یک لغت‌نامه یا کتاب کد[۶۳] ایجاد شود. سپس هر ویژگی به یکی از این لغات منتسب شود. این انتساب می‌تواند به صورت سخت‌گیرانه[۶۴] یا نرم[۶۵] انجام شود. در ادامه سه روش به نام‌های کیف کلمات، کیف کلمات با انتساب نرم و بردار فیشر معرفی می‌شود.

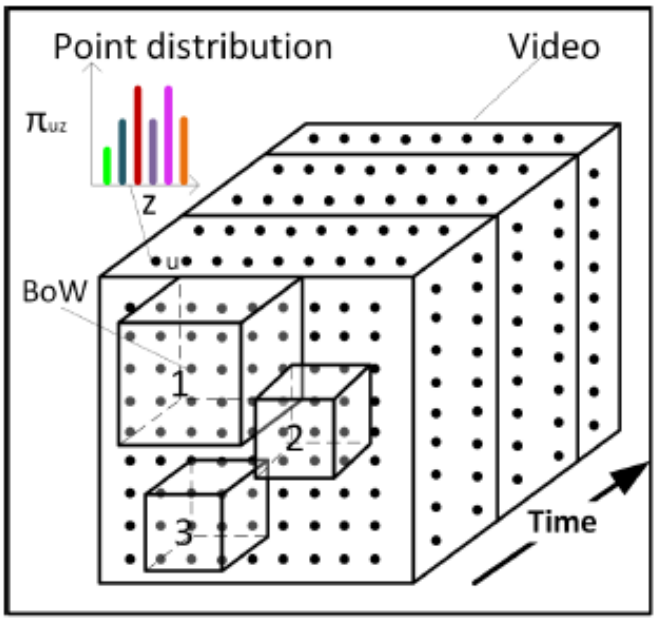
---

۶۳ Codebook
۶۴ Hard Assignment
۶۵ Soft Assignment

## ۳-۳-۱ کیف کلمات

در این روش مجموعه‌ای از بردارهای ویژگی محلی استخراج شده از هر تکه در مدل فضا-زمانی ویدئو در یک بردار ویژگی خلاصه می‌شوند، که این بردار بیان‌گر چگونگی توزیع ویژگی‌ها در آن تکه است. برای منظور لازم است ابتدا از تمام نقاط شبکه در تعداد زیادی از ویدئوهای آموزش ویژگی محلی استخراج نمود. و سپس با یک روش خوشه‌بندی (مانند خوشه‌بندی کی-میانگین[۶۶]) این بردارهای ویژگی محلی را به چند دسته خوشه‌بندی کرد. مراکز این خوشه‌ها کلمات لغت‌نامه هستند. پس از آن هر ویژگی محلی جدیدی با استفاده از معیار فاصله[۶۷] به یکی از این لغات منتسب می‌شود.

فرض کنیم می‌خواهیم ویژگی کیف کلمات را در تکه‌ای به ابعاد $N*N*N$ از مدل فضا-زمانی یک ویدئو بیابیم. شکل (۳-۷) این فرایند را نشان می‌دهد.




شکل (۳ - ۷) - روش کیف کلمات برای کد کردن ویدئو [۴۰].

[۶۶] K-means Clustering
[۶۷] Distance Measure

برای این منظور در هر یک از نقاط $u$ از شبکه که در این تکه قرار دارد پنجره‌هایی به مرکز آن نقطه و به ابعاد $\frac{N}{2^m} * \frac{N}{2^m}$ درنظر گرفته می‌شود؛ که:

$$\boldsymbol{m = 2, 3, \dots, log_2 \frac{N}{4}} \qquad (۱۳-۳)$$

و از هر یک از این پنجره‌ها یک بردار ویژگی محلی بدست می‌آید که به یکی از کلمات دیکشنری منتسب می‌شود. با در نظر گرفتن همه‌ی این پنجره‌ها یک توزیع از کلمات دیکشنری برای نقطه‌ی $u$ بدست می‌آید که به آن $\pi_{u,z}$ می‌گوییم (zها کلمات دیکشنری هستند). در نهایت با جمع کردن $\pi_{u,z}$ها در تکه‌ی موردنظر، توزیع کلمات در آن تکه بدست می‌آید که این توزیع همان ویژگی نهایی در آن تکه است.

## ۳-۳-۲ کیف کلمات با انتساب نرم

ویژگی‌ای که از روش کیف کلمات بدست می‌آید دارای مقادیر پراکنده[۶۸] است و احتمالا خیلی از مولفه‌های آن صفر می‌شود. زیرا در انتساب سخت‌گیرانه هر بردار ویژگی محلی به یک کلمه منتسب می‌شود و همیشه تعداد زیادی از کلمات در هر تکه وجود ندارند. اما روش انتساب نرم[۵۱]، هر بردار ویژگی محلی را به صورت نرم به تمام لغات دیکشنری منتسب می‌کند. در روش انتساب سخت هر بردار ویژگی محلی با مقدار ۱ به نزدیک‌ترین کلمه و با مقدار صفر به کلمات دیگر منتسب می‌شود؛ اما در روش انتساب نرم، هر بردار ویژگی محلی با مقادیر مختلفی که بین صفر و یک قرار دارند به تمام کلمات منتسب می‌شود. این مقدار برای بردار ویژگی محلی S و کلمه‌ی $C_i$ از رابطه‌ی (۳-۱۴) بدست می‌آید که در آن منظور از $d(S,C_i)$ فاصله‌ی بین S و $C_i$ است. مطابق با این رابطه، به نزدیک‌ترین کلمه مقدار یک و به بقیه‌ی کلمات مقداری بین صفر و یک منتسب خواهد شد.

$$\boldsymbol{U(S, C_i) = [\frac{min_j\left(d(S,C_j)\right)}{d(S,C_i)}]^{\beta}} \qquad (۱۴-۳)$$

[۶۸] Sparse

واضح است که اگر β به بی‌نهایت میل کند یا خیلی بزرگ شود، روش انتساب نرم به انتساب سخت‌گیرانه تبدیل می‌شود.

## ۳-۳-۳ بردار فیشر

در این روش[۵۲]، هر یک از کلمات لغت‌نامه یک مدل مخلوط گوسی[۶۹] است و ویژگی‌های محلی با شیوه‌ای احتمالاتی به این مدل‌ها منتسب می‌شوند. بنابراین لازم است در مرحله‌ی اول K مدل مخلوط گوسی از روی نمونه‌های آموزشی، آموزش داده شود. با در اختیار داشتن این مدل‌های مخلوط گوسی می‌توان برای هر مجموعه‌ای از بردارهای ویژگی محلی، یک بردار فیشر بدست آورد که دربردارنده‌ی اطلاعات همه‌ی آن مجموعه ویژگی‌های محلی است و می‌تواند به عنوان ویژگی نهایی جایگزین آن ویژگی‌های محلی مورد استفاده قرار گیرد.

فرض کنید I مجموعه‌ای از بردارهای ویژگی محلی باشد که هر کدامشان دارای $D$ بعد هستند.

$$\boldsymbol{I} = (\boldsymbol{X_1}, \dots, \boldsymbol{X_N}) \qquad (۱۵-۳)$$

و $K$ مدل گوسی در اختیار داریم.

$$\boldsymbol{\theta} = (\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}, \boldsymbol{\pi_k} : \boldsymbol{k} = \boldsymbol{1}, \dots, \boldsymbol{K}) \qquad (۱۶-۳)$$

با در اختیار داشتن میانگین و انحراف معیار مدل‌ها، روابط (۳-۱۷) و (۳-۱۸) را محاسبه می‌کنیم.

$$\boldsymbol{u_{j,k}} = \frac{\boldsymbol{1}}{\boldsymbol{N\sqrt{2\pi k}}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{q_{i,k}} \frac{\boldsymbol{x_{j,i} - \mu_{j,k}}}{\boldsymbol{\sigma_{j,k}}} \qquad (۱۷-۳)$$

$$\boldsymbol{v_{j,k}} = \frac{\boldsymbol{1}}{\boldsymbol{N\sqrt{2\pi k}}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{q_{i,k}} \left[\left(\frac{\boldsymbol{x_{j,i} - \mu_{j,k}}}{\boldsymbol{\sigma_{j,k}}}\right)^{\boldsymbol{2}} - \boldsymbol{1}\right] \qquad (۱۸-۳)$$

که در آن

$$\boldsymbol{q_{i,k}} = \frac{\boldsymbol{exp}[-\frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{X_i - \mu_k})^{\boldsymbol{T}} \boldsymbol{\Sigma_k^{-1}} (\boldsymbol{X_i - \mu_k})]}{\sum_{\boldsymbol{t=1}}^{\boldsymbol{K}} \boldsymbol{exp}[-\frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{X_i - \mu_t})^{\boldsymbol{T}} \boldsymbol{\Sigma_k^{-1}} (\boldsymbol{X_i - \mu_t})]} \qquad (۱۹-۳)$$

[۶۹] Gaussian Mixture Models

و j=1,2,3,…D بیان‌گر بُعد مورد نظر در بردار ویژگی است. در نهایت با تجمیع مقادیر بدست آمده از روابط (۳-۱۷) و (۳-۱۸) یک بردار به طول 2DK بدست می‌آید که همان ویژگی نهایی است:

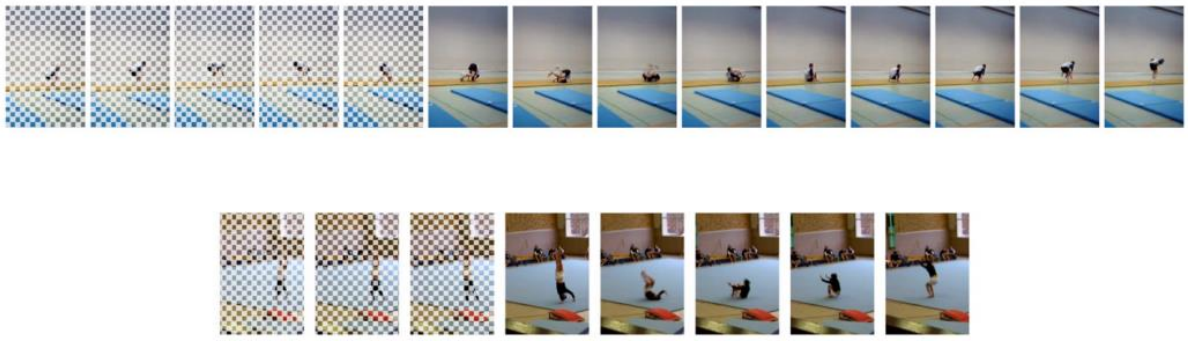
$$\boldsymbol{fisher(I) = [\cdots u_k \ldots v_k \cdots]} \qquad (۳-۲۰)$$

مقاله‌ی [۵۳] با ایجاد دو تغییر در روش بردار فیشر، بردار فیشر بهبودیافته را معرفی می‌کند که کارایی به مراتب بالاتری نسبت به [۵۲] دارد. در روش بهبود یافته: اولا یک تابع غیرخطی مانند (۳-۲۱) به تمام مولفه‌های بردار فیشر اعمال می‌شود. ثانیا، این بردار با استفاده از نُرم دو[۷۰]، نرمالیزه می‌گردد.

$$\boldsymbol{F(z) = |z|sgn(z)} \qquad (۳-۲۱)$$

## ۳-۴ روش تابع هسته‌ی انعطاف‌پذیر با زمان برای شناسایی فعالیت

ویدئوهای ثبت شده در دنیای واقعی، از نظر طول زمانی با هم برابر نیستند و فعالیت مورد نظر نیز همیشه در یک نقطه‌ی مشخص از زمان شروع نشده و به یک نقطه‌ی مشخص ختم نمی‌شود. در واقع دو ویدئو که حاوی فعالیت مشابهی هستند ممکن است تنها در قسمتی از فریم‌هایشان با هم مشابه باشند.



شکل (۳ - ۸) دو ویدئو از پایگاه داده‌ی HMDB51 که فعالیت مشابهی در آن‌ها انجام می‌شود. تنها فریم‌های پایانی این ویدئوها با هم مطابقت دارند[۳۴].

[۷۰] L-2 Norm

به عنوان مثال شکل (۳-۸) دو ویدئو که فعالیت یکسانی در آن‌ها رخ می‌دهد از پایگاه داده‌ی HMDB51 را نشان می‌دهد که فقط فریم‌های پایانی‌شان با هم مطابقت دارد.

در این روش یک تابع هسته‌ی انعطاف‌پذیر با زمان برای SVM معرفی می‌شود که می‌تواند دو دنباله با طول‌های نابرابر را با هم مقایسه کند. این تابع هسته به‌گونه‌ای است که با استفاده از آن، مقایسه‌ی بین دو دنباله با زمان انعطاف‌پذیر می‌گردد و امکان شناسایی فعالیت‌هایی که در زمان فشرده یا کشیده شده‌اند فراهم می‌شود. در ادامه به معرفی این روش خواهیم پرداخت.

## ۳-۴-۱ تابع هسته انعطاف‌پذیر با زمان

فرض کنید X دنباله‌ای شامل N بردار ویژگی باشد که به ترتیب کنار هم قرار گرفته‌اند و هر کدام دارای D بُعد هستند.

$$\boldsymbol{X} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_N}\} \qquad (۲۲-۳)$$

و Y دنباله‌ای شامل M بردار ویژگی متوالی D بُعدی باشد.

$$\boldsymbol{Y} = \{\boldsymbol{y_1}, \dots, \boldsymbol{y_M}\} \qquad (۲۳-۳)$$

و می‌خواهیم دو دنباله‌ی غیرهم‌طول X و Y را با هم مقایسه کنیم.

با توجه به‌اینکه معمولا مرکز و هسته‌ی فعالیت در وسط ویدئو قرار دارد، برای مقایسه‌ی این دو دنباله ابتدا مراکز آن‌ها را مطابق شکل (۳-۹) بر هم منطبق می‌کنیم. سپس دو تابع F(t) و G(t) را به صورت رابطه‌های (۳-۲۴) و (۳-۲۵) تعریف می‌کنیم.

$$\boldsymbol{F(t)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{f_i(t)x_i} \qquad (۲۴-۳)$$

$$\boldsymbol{G(t)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} \boldsymbol{g_j(t)y_j} \qquad (۲۵-۳)$$

که در آن‌ها f و g دو تابع گوسی مانند رابطه‌های (۳-۲۶) و (۳-۲۷) هستند.

$$f_i(t) = \frac{1}{N*\sigma_x\sqrt{2\pi}} e^{-\frac{(t-\mu_i)^2}{2\sigma_x^2}} \quad (۲۶-۳)$$

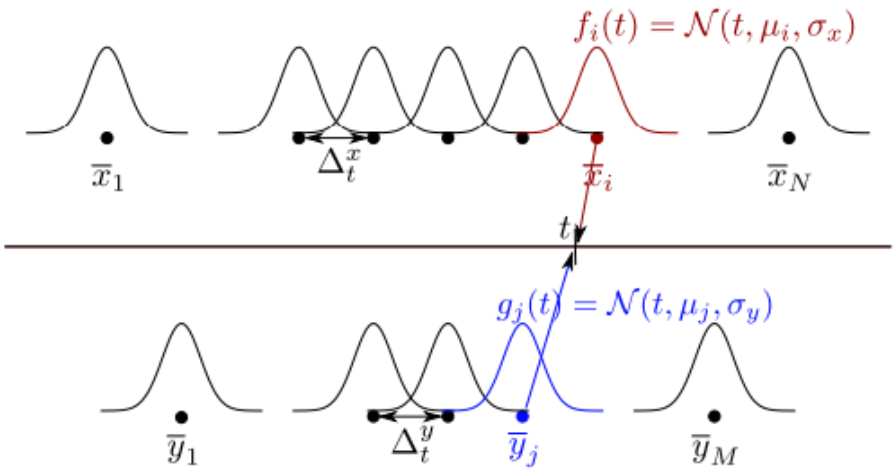
$$g_j(t) = \frac{1}{M*\sigma_y\sqrt{2\pi}} e^{-\frac{(t-\mu_j)^2}{2\sigma_y^2}} \quad (۲۷-۳)$$

که پارامتر میانگین این توابع گوسی از روابط (۳-۲۸) و (۳-۲۹) بدست می‌آید. پارامتر σ که به صورت تجربی انتخاب می‌شود، میزان انعطاف‌پذیری به زمان را تعیین می‌کند که در ادامه شرح داده خواهد شد.

$$\mu_i = (i - \frac{N+1}{2})\Delta_t^x \quad (۲۸-۳)$$

$$\mu_j = (j - \frac{M+1}{2})\Delta_t^y \quad (۲۹-۳)$$

چنان‌که شکل (۳-۹) نشان می‌دهد، با استفاده از توابع F و G در اطراف هر یک بردارهای $x_i$ و $y_i$ یک تابع گوسی که انعطاف پذیری در رخداد بردار ویژگی را مدل می کند درنظر گرفته می‌شود. این توابع گوسی با یکدیگر اشتراک دارند و واضح است که هرچه σ را بزرگ‌تر انتخاب کنیم، این توابع پهن‌تر شده و درنتیجه F و G نسبت به زمان انعطاف‌پذیرتر می‌شوند. از طرفی با کاهش σ توابع گوسی باریک‌تر شده و انعطاف‌پذیری به زمان کاهش می‌یابد. ولی از طرفی قطعی در انطباق دو دنباله را کاهش می دهد. با فرض فاصله‌ی زمانی یکسان بردارها در هر دو دنباله، می‌توان پارامتر Δt را برای هر دو رشته یکسان و برابر با یک درنظر گرفت.




شکل (۳ - ۹) توابع گوسی اطراف هر نمونه‌ی رشته‌های هم‌مرکز شده [۳۴].

تابع هسته‌ی انعطاف‌پذیر با زمان را به صورت رابطه‌ی (۳-۳۰) تعریف می‌کنیم.

$$\boldsymbol{TFK(F,G) = \int F(t)^T G(t) dt} \qquad (۳۰-۳)$$

بنابراین:

$$\boldsymbol{TFK(F,G) = \sum_{i=1}^{N} \sum_{j=1}^{M} K_{GAUSS_p}\left(f_i(t), g_j(t)\right) K_{LIN}(x_i, y_j)} \qquad (۳۱-۳)$$

که در آن

$$\boldsymbol{K_{GAUSS_p}\left(f_i(t), g_j(t)\right) = \frac{(2\pi\sigma_x\sigma_y)^{(1-2\rho)/2}}{NM\sqrt{2\rho}} e^{\frac{-\|\mu_i - \mu_j\|^2}{4\sigma_x\sigma_y/\rho}}} \qquad (۳۲-۳)$$

و

$$\boldsymbol{K_{LIN}(x_i, y_j) = dot(x_i, y_j)} \qquad (۳۳-۳)$$

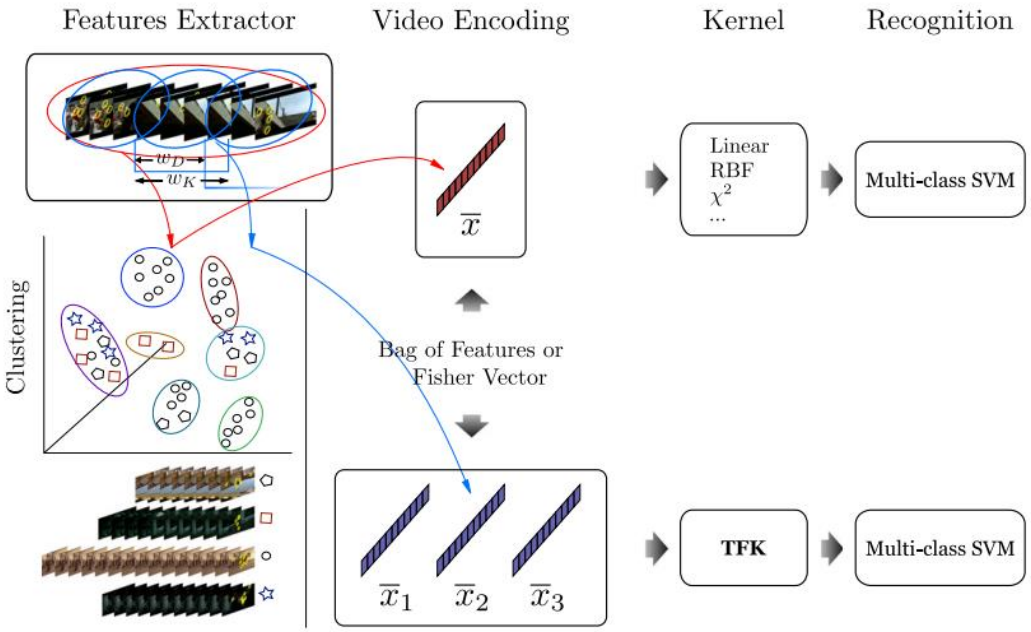
در رابطه‌ی (۳-۳۲)، ρ پارامتر تابع هسته‌ی $K_{GAUSS_\rho}$ می‌باشد. با انتخاب $\rho = 1/2$ تابع هسته‌ی بهاتاچاریا[۷۱] حاصل می گردد[۳۴]. در رابطه (۳-۳۱) $K_{LIN}$ یک تابع هسته‌ی خطی است.

[۷۱] Bhattacharyya

با استفاده تابع هسته (۳-۳۱)، دو دنباله به فضای جدیدی می‌روند که در آن فضا با یکدیگر هم‌طول هستند و ضمن این تبدیل انعطاف‌پذیری نسبت به زمان نیز درنظر گرفته شده است.

## ۳-۴-۲ شناسایی فعالیت

شکل (۳-۱۰) ساختار کلی روش شناسایی فعالیت با استفاده از تابع هسته‌ی انعطاف‌پذیر با زمان را نشان می‌دهد. در اولین مرحله ویدئو به بخش‌هایی تقسیم می‌شود که هرتکه شامل $W_K$ فریم باشد. این بخش‌ها می‌توانند دارای اشتراک باشند یا اینکه بدون اشتراک انتخاب شوند. برای این منظور با استفاده از یک پنجره‌ی لغزان بین فریم‌ها با گام $W_D$ حرکت کرده و هربار تا $W_K$ فریم بعدی را جدا می‌کنیم. اگر گام حرکت با طول بخش‌ها برابر باشد، بخش‌ها اشتراک نخواهند داشت، اما اگر گام حرکت کوچک‌تر از طول بخش‌ها انتخاب شود، بخش‌ها دارای اشتراک خواهند بود.




شکل (۳ - ۱۰) – شناسایی فعالیت با ترکیب تابع هسته‌ی انعطاف‌پذیر با زمان و تابع هسته‌ی خطی[۳۴].

در مرحله‌ی دوم در هریک از این بخش‌ها ویژگی استخراج کرده و با یکی از روش‌های کد کردن ویژگی (مانند کیف کلمات یا بردار فیشر)، این ویژگی‌ها در یک بردار ویژگی جایگزین کد می‌شوند. در نتیجه هر بخش، تنها با یک بردار ویژگی توصیف می‌گردد. با کنار هم قرار دادن این بردارها به

ترتیب زمانی بخش‌ها در ویدئو، دنباله‌ای از بردارها ایجاد می‌شود که آن ویدئو را توصیف می‌کند. به عنوان مثال اگر ابعاد بردار ویژگی بدست‌آمده در هر بخش D باشد و تعداد بخش‌ها N باشد، این دنباله را می‌توان در یک ماتریس D*N ذخیره نمود. این‌کار را برای تمام ویدئوها انجام می‌دهیم. با توجه به تفاوت طول ویدئوها این دنباله‌ها برای ویدئوهای مختلف با هم برابر نخواهند بود.

در مرحله‌ی سوم با در اختیار داشتن داده‌های آموزشی و آزمایشی و استفاده از تابع هسته‌ی معرفی شده در رابطه‌ی (۳-۳۱)، دو تابع هسته برای آموزش و آزمایش محاسبه می‌شود و سپس با استفاده از SVM فرایند آموزش و سپس آزمایش به منظور شناسایی فعالیت اجرا می‌گردد.

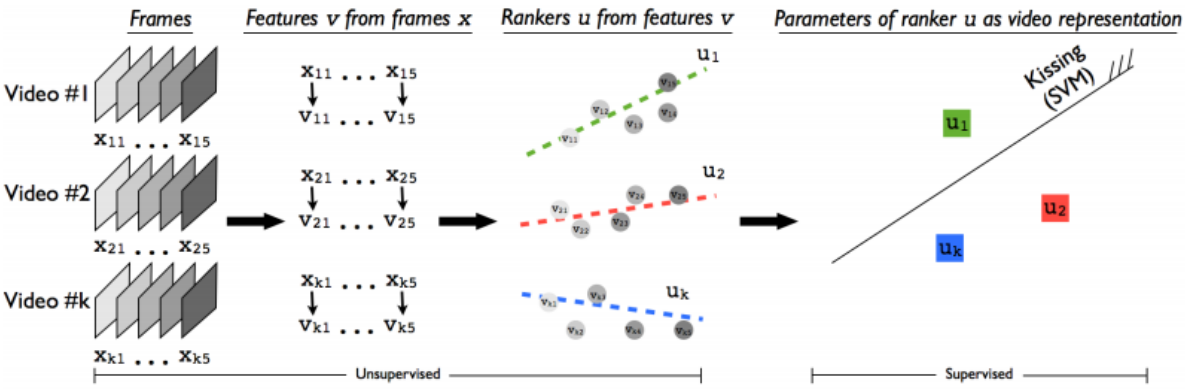
با توجه به اینکه برخی از فعالیت‌ها با درنظر گرفتن انعطاف‌پذیری نسبت به زمان بهتر شناسایی می‌شوند و برخی دیگر با استفاده از تابع هسته‌ی خطی بهتر شناسایی می‌شود، بهتر است یک بار هم بدون تقسیم ویدئو به تکه‌های کوچک، از کل ویدئو ویژگی استخراج کرده و ویژگی‌ها را کد کنیم تا هر ویدئو با یک بردار ویژگی توصیف شود. سپس تابع هسته‌ی خطی آموزش و آزمایش را برای آن‌ها بدست آورده و با تابع هسته‌ی بدست آمده از رابطه‌ی (۳-۳۱) جمع کنیم. سپس از این تابع هسته‌ی حاصل‌جمع برای شناسایی فعالیت به‌وسیله‌ی SVM استفاده کنیم. رابطه‌ی (۳-۳۴) این تابع هسته‌ی حاصل‌جمع را نشان می‌دهد.

$$\boldsymbol{CombK(v1, v2) = TFK(F, G) + K_{LIN}(X, Y)} \qquad (۳۴-۳)$$

## ۳-۵ روش ادغام رتبه برای شناسایی فعالیت

روش ادغام رتبه، روشی برای استخراج ویژگی از ویدئو برای شناسایی فعالیت است. این ویژگی‌ها با حل یک مسئله‌ی بهینه‌سازی بدست می‌آید. این مسئله‌ی بهینه‌سازی بهینه‌ترین ویژگی‌ای که بیان‌گر رتبه یا همان ترتیب فریم‌ها در ویدئو است را بدست می‌دهد.

شکل (۳-۱۱) ساختار این روش را نشان می‌دهد. ابتدا از هر فریم چندین ویژگی محلی استخراج می‌شود. سپس این ویژگی‌ها با استفاده از بردار فیشر کد می‌شوند. در نتیجه هر فریم با یک بردار ویژگی توصیف می‌شود. سپس این ویژگی‌ها در کنار هم قرار داده می‌شود تا دنباله‌ای از ویژگی‌های ویدئو بدست آید و این دنباله با استفاده از روش میانگین متحرک[۷۲]، نرم[۷۳] می‌شود.




شکل (۳ - ۱۱) نمایش ساختار روش ادغام رتبه[۳۵].

سپس با حل یک مسئله‌ی بهینه سازی یک بردار ویژگی نهایی برای کل ویدئو بدست می‌آید که دربردارنده‌ی اطلاعات مربوط به ترتیب زمانی فریم‌ها در ویدئو است.

مسئله‌ی بهینه‌سازی برای رگرسیون[۷۴] به صورت رابطه‌ی (۳-۳۵) را نظر بگیرید. که در آن $x_i$ نمونه ورودی و $y_i$ برچسب آن نمونه است. $w$ پارامتر بهینه‌سازی است.

$$\min_{w} \frac{1}{2} w^T w + C \sum_{i=1}^{l} (max(0, |y_i - w^T x_i| - \varepsilon))^2 \qquad (۳۵-۳)$$

می‌خواهیم این مسئله را با این شرط حل کنیم که: هرگاه بردار $x_j$ که ویژگی کد شده در فریم $j$ است، از بردار $x_i$ متأخرتر بود، $w^T x_j$ از $w^T x_i$ بزرگ‌تر شود. رابطه‌ی (۳-۳۶) این شرط را نشان می‌دهد.

---

[۷۲] Moving Average (MA)
[۷۳] Smooth
[۷۴] Regression

$$s.t \ \ \forall i,j: \ \ x_j \ggg x_i \ \Rightarrow \ w^T x_j > w^T x_i \qquad (۳۶-۳)$$

برای برآورده کردن شرط (۳-۳۶) کافیست در رابطه‌ی (۳-۳۵) به ازای فریم‌های متاخرتر، مقدار $y_i$ را نسبت به فریم‌های قبلی بزرگتر اختیار کنیم. به عنوان مثال می توان آن را برابر با شماره‌ی فریم انتخاب کرد:

$$y_i = i \qquad (۳۷-۳)$$

پس از حل مسئله‌ی بهینه‌سازی فوق، پارامتر آن یعنی w به عنوان ویژگی نهایی ویدئو مورد استفاده قرار می‌گیرد.

بهتر است یک بار دیگر نیز ویژگی‌های کد شده در هر فریم از انتها به ابتدا مرتب شده و بر آن MA اعمال شودو مجددا با حل مسئله‌ی بهینه سازی فوق، w برای حالت معکوس نیز بدست آید. در این صورت ویژگی نهایی از الحاق w بدست آمده در حالت مستقیم، $w_f$، و w بدست آمده از حالت معکوس، $w_r$، حاصل می‌شود.

$$final_{feature} = [w_f, w_r] \qquad (۳۸-۳)$$

پس از محاسبه‌ی بردار ویژگی نهایی برای تمام ویدئوهای آموزش و آزمایش، این بردارها با استفاده از SVM از هم تفکیک شده و فعالیت شناسایی می‌گردد.

## ۳-۶ جمع‌بندی

در این فصل روش‌های معمول استخراج ویژگی و کد کردن ویژگی‌ها معرفی و تشریح شد. سپس دو روش شناسایی فعالیت به نام‌های روش تابع هسته‌ی انعطاف‌پذیر با زمان و روش ادغام رتبه (یا ویدئوداروین) به تفصیل شرح داده شد.

در میان روش‌های استخراج ویژگی، روش مسیر متراکم بهبود یافته نسبت به روش‌های دیگر به نویز و حرکت‌های نامنظم ناگهانی مقاوم‌تر است. در میان روش‌های کد کردن ویژگی، روش کیف کلمات

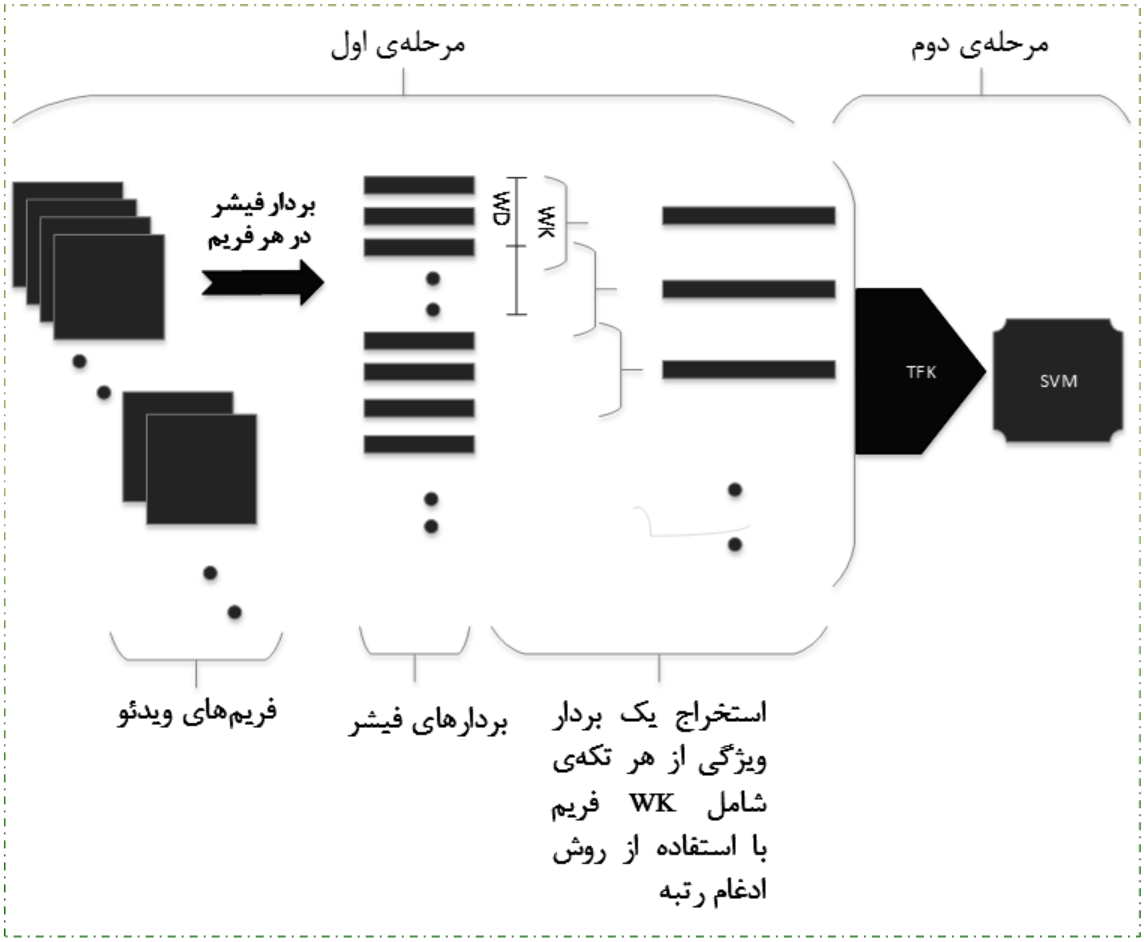
صرفا تعداد دفعات تکرار کلمات در هر تکه را می‌شمارد و در نتیجه اطلاعات بدست آمده به اندازه‌ی کافی غنی نیستند. بنابراین روش‌های کیف کلمات با انتساب نرم و بردار فیشر نسبت به روش کیف کلمات با انتساب سخت کارآمدتر هستند.

در فصل آینده روش پیشنهادی این تحقیق برای شناسایی فعالیت معرفی می‌شود.

# فصل چهارم : روش پیشنهادی

## ۱-۴ مقدمه

هدف از این تحقیق ارائه روشی است که مقایسه ویدئویی فعالیت‌های انجام شده با طول زمانی نابرابر را امکان‌پذیر نماید. لذا ما بر ترتیب زمانی وقایع انجام شده در یک فعالیت تکیه داریم  این فعالیت می تواند برای یک فرد در زمانهای مختلف و یا افراد دیگر با طول زمانی متفاوت انجام شود. این تحقیق روشی سلسه‌مراتبی برای شناسایی فعالیت در رشته‌های ویدئویی پیشنهاد می‌کند. شکل (۴-۱) ساختار کلی روش پیشنهادی را برای یک ویدئو نشان می‌دهد.




شکل (۴ - ۱) ساختار روش پیشنهادی برای یک ویدئو.

در مرحله‌ی اول ابتدا از ویدئوها ویژگی مسیر متراکم بهبودیافته استخراج می‌گردد. سپس به ازای هر فریم، ویژگی‌های مربوط به آن فریم (ویژگی‌هایی که مسیر حرکتشان به آن فریم ختم می‌شود) با روش بردار فیشر بهبود یافته کد می‌شوند. در نتیجه یک بردار ویژگی برای توصیف هر یک از فریم‌ها

بدست می‌آید. سپس با یک پنجره‌ی لغزان با گام $W_D$ در امتداد فریم‌ها حرکت کرده و ویدئو به بخش‌هایی به طول $W_K$ فریم تقسیم می‌شود. مقادیر $W_D$ و $W_K$ به گونه‌ای انتخاب می‌شود که این بخش‌ها ۵۰ درصد با یکدیگر اشتراک داشته باشند. آن‌گاه در هریک از این بخش‌ها با روش ادغام رتبه یک بردار ویژگی بدست می‌آید. بنابراین در پایان مرحله‌ی اول، هر یک از ویدئوها با یک دنباله از بردارها توصیف می‌گردد. در مرحله‌ی دوم با در اختیار داشتن توالی بردارهای توصیف‌کننده‌ی هر ویدئو، ماتریس تابع هسته‌ی انعطاف‌پذیر با زمان برای هر دو مجموعه‌ی ویدئوهای آزمایشی و آموزشی محاسبه می‌شود وسپس با استفاده از SVM فرایندهای آموزش و آزمایش اجرا می‌گردند.

## ۴-۲ شرح مراحل

روش پیشنهادی شامل دو مرحله است. در ادامه به شرح جزییات هریک از این مراحل می‌پردازیم.

### ۴-۲-۱ مرحله اول

در مرحله‌ی اول از همه‌ی ویدئوهای آموزشی و آزمایشی ویژگی مسیر متراکم بهبودیافته استخراج می‌شود. همان‌طور که در بخش‌های ۳-۲-۴ و ۳-۲-۵ شرح داده شد، برای استخراج ویژگی مسیر حرکت پیکسل‌ها در فریم‌های رشته ویدئویی بدست آمده و به ویژگی‌های HOG وHOF و MBH الحاق می‌شود. هر بردار ویژگی، ۴۲۶ مولفه دارد که ۳۰ مولفه‌ی آن مربوط به مسیر حرکت، ۹۶ مولفه مربوط به HOG، ۱۰۸ مولفه مربوط به HOF، ۹۶ مولفه مربوط به $MBH_x$ و ۹۶ مولفه‌ی دیگر مربوط به $MBH_y$ است. پس از استخراج ویژگی، با استفاده از الگوریتم PCA ابعاد بردارهای ویژگی به نصف (یعنی ۲۱۳) کاهش داده می‌شود. آن‌گاه با استفاده از مدل مخلوط گوسی توزیع احتمالاتی ویژگی‌های مربوط به ویدئوهای آموزشی را مدل می‌نماییم، خروجی این الگوریتم پارامتر های مربوط به z تابع گوسی در این مدل می باشد. با در اختیار داشتن پارامتر های مدل مخلوط گوسی و با استفاده از روش فیشر بهبودیافته، ویژگی‌های مربوط به هریک از فریم‌ها ( ویژگی‌هایی که مسیر حرکتشان به آن فریم ختم می‌شود) در هریک از ویدئوهای آموزش و آزمایش به یک بردار

فیشر تبدیل می‌گردند. در نتیجه هر فریم یک ویدئو با یک بردار فیشرتوصیف می گردد. با توجه به مفاهیم بخش ۳-۳-۳، واضح است که طول برداری که برای توصیف هر فریم بدست می‌آید 2*213*z خواهد بود.

سپس با یک پنجره‌ی لغزان با گام $W_D$ در امتداد فریم‌های ویدئو حرکت کرده و ویدئو به بخش‌هایی با طول $W_K$ فریم تقسیم می‌شود. بخش‌ها به گونه‌ای انتخاب می‌شوند که ۵۰ درصد با یکدیگر اشتراک داشته باشند. در هر یک از این بخش‌ها با استفاده از روش ادغام رتبه، یک بردار ویژگی بدست می‌آید که بیان‌گر فریم‌های آن بخش و ترتیب زمانی آن فریم‌ها است. برای این منظور، بردارهای فیشر بدست آمده از فریم‌های هر بخش یک‌بار در جهت مستقیم، یعنی از ابتدا به انتها، و یک‌بار دیگر در جهت معکوس، یعنی از انتها به ابتدا، به شکل دنباله در کنار هم قرار می‌گیرند و در هر بار دنباله‌ی مذکور با استفاده از روش میانگین متحرک نرم می‌شود. سپس برای هر بخش دو رگرسیون بردار پشتیبان[۷۵] مطابق رابطه‌ی (۳-۳۵)، یک بار برای دنباله‌ی نرم شده با ترتیب مستقیم و بار دیگر برای دنباله‌ی نرم شده با ترتیب معکوس آموزش داده شده و پارامتر بدست آمده از این دو بار آموزش به یکدیگر الحاق شده و به عنوان بردار ویژگی در آن بخش در نظر گرفته می‌شود.

بنابراین در پایان مرحله‌ی اول هر یک از بخش‌ها، در یک بردار ویژگی خلاصه می‌شوند و در نتیجه هر ویدئو به صورت دنباله‌ای از بردارها نمایش داده می‌شود.

## ۴-۲-۲ مرحله‌ی دوم

در مرحله‌ی دوم با در اختیار داشتن دنباله‌ی بردارهای نماینده‌ی هر ویدئو و با استفاده از رابطه‌ی (۳-۳۱)، تابع هسته‌ی انعطاف پذیر به زمان یک بار برای ویدئوهای آموزش و بار دیگر برای ویدئوهای آزمایش محاسبه می‌گردد. اگر تعداد کل ویدئوهای آموزش S و تعداد کل ویدئوهای آزمایش T باشد، تابع هسته‌ی آموزش ماتریسی S*S و تابع هسته‌ی آزمایش یک ماتریس T*S

[۷۵] Support Vector Regression

خواهد بود. به عبارتی تابع هسته همه‌ی داده‌ها را به فضای جدیدی می‌برد که در آن فضا، هر داده طولی برابر با S دارد.

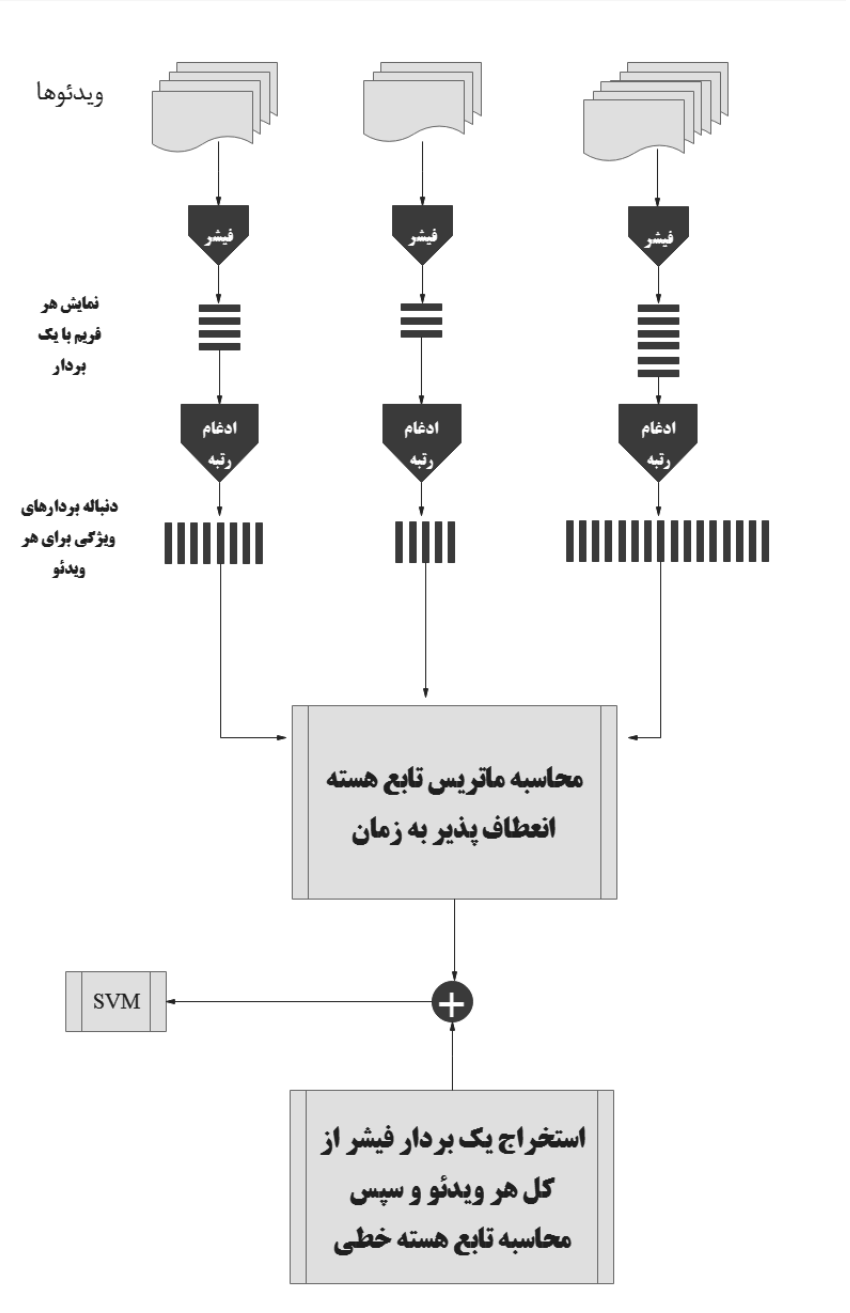
هر سطر از ماتریس تابع هسته‌ی آموزش بیان‌گر یکی از ویدئوهای آموزش در فضای جدید و هر سطر از ماتریس تابع هسته آزمایش بیان‌گر یکی از ویدئوهای آزمایش در فضای جدید است. با استفاده از رابطه‌ی (۳-۳۱) دنباله‌ی بردارهای بیان‌گر هر یک از ویدئوهای آموزش یکی‌یکی با دنباله‌ی بردارهای بیان‌گر ویدئوهای آموزشی دیگر مقایسه می‌شود و هر بار یک عدد بدست می‌آید که این عدد به سطر مربوط به ویدئوی آموزشی اول و ستون مربوط به ویدئوی آموزشی دوم در ماتریس تابع هسته آموزش تعلق دارد.

دنباله‌ی بردارهای بیان‌گر هریک از ویدئوهای آزمایش نیز با استفاده از رابطه‌ی (۳-۳۱) با هریک از دنباله‌های بیانگر ویدئوهای آموزش مقایسه می‌شود و عدد بدست آمده در سطر مربوط به ویدئوی آزمایشی و ستون مربوط به ویدئوی آموزشی مربوطه از ماتریس تابع هسته‌ی آزمایش قرار می‌گیرد.

در آخر با در اختیار داشتن مقادیر دو ماتریس تابع هسته آموزش و تابع هسته آزمایش، با استفاده از ماشین بردار پشتیبان فرایند آموزش و سپس آزمایش اجرا می‌گردد.

## ۳-۴ شرح روش

شکل (۴-۲) الگوریتم کلی روش پیشنهادی را نشان می‌دهد. چنان‌که در این روندنما مشاهده می‌شود، دراین روش اطلاعات زمانی ویدئوها در دو لایه استخراج می‌گردد. در لایه‌ی اول با استفاده از روش‌های فیشر بهبودیافته و ادغام رتبه ساختار ترتیبی فریم‌ها در هر بخش استخراج می‌شود. سپس در لایه‌ی دوم با محاسبه‌ی تابع هسته انعطاف‌پذیر با زمان، ضمن درنظر گرفتن ترتیب بخش‌ها، امکان مقایسه‌ی ویدئوهایی که در زمان کشیده یا فشرده شده‌اند فراهم می‌گردد.




شکل (۴ - ۲) الگوریتم کلی روش پیشنهادی برای همه ویدئوها.

برخی از فعالیت‌ها با درنظرگرفتن ساختار ترتیبی بهتر شناسایی می‌گردند، اما برخی دیگر بدون درنظرگرفتن توالی‌های زمانی بهتر شناسایی می‌گردند. بنابراین لازم است به گونه‌ای هر دو حالت مذکور در نظر گرفته شود. برای این منظور یک بار نیز کلیه‌ی ویژگی‌های مسیر متراکم بهبودیافته استخراج شده از هریک از ویدئوهای آموزش و آزمایش با روش فیشر بهبودیافته به یک بردار تبدیل می‌شوند. در نتیجه هر ویدئو تنها با یک بردار نمایش داده شود. سپس دو تابع هسته‌ی خطی

آزمایش و آموزش مطابق با رابطه‌ی (۳-۳۳) برای هر یک از دو مجموعه‌ی ویدئوهای آموزش و آزمایش محاسبه می‌گردد. در نهایت تابع هسته‌ی خطی آموزش با تابع هسته‌ی انعطاف‌پذیر با زمان آموزش و تابع هسته‌ی خطی آزمایش با تابع هسته‌ی انعطاف‌پذیر با زمان آزمایش جمع شده، و توابع هسته‌ی حاصل‌جمع به SVM داده می‌شود. به این ترتیب قادر خوهیم بود فعالیت‌هایی که بدون درنظر گرفتن ساختار ترتیبی بهتر شناسایی می‌شوند را نیز به خوبی شناسایی نماییم.

## ۴-۴ اجرای روش

برای استخراج ویژگی IDT از پیاده سازی انجام شده توسط مرجع [۲۷] استفاده گردید. بردار فیشر با استفاده از پیاده‌سازی موجود در کتابخانه VLFEAT در C++ [۵۴] استخراج گردید. برای آموزش رگرسیون بردار پشتیبان از کتابخانه‌ی LibLinear [۵۵] استفاده شد. SVM نیز با استفاده از کتابخانه‌ی LIBSVM [۵۶] پیاده‌سازی گردید.

بیش‌ترین هزینه‌ی پردازش مربوط به قسمت محاسبه‌ی تابع هسته است. با توجه به رابطه‌ی (۳-۳۱) در این قسمت لازم است تعداد زیادی بردار طولانی به صورت دو به دو در هم ضرب نقطه‌ای شوند. طول هریک از این بردارها 2*2*213*z خواهد بود ( یعنی دو برابر طول هر بردار فیشر؛ زیرا همان‌طور که در بخش ۳-۵ گفته شد، در روش ادغام رتبه، پارامتر w در حالت مستقیم و معکوس به یکدیگر الحاق می‌گردد). با توجه به اینکه ضرب نقطه‌ای عملی ترتیبی نیست می‌توان با استفاده از پردازنده‌های موازی یا پردازنده‌ی گرافیکی، سرعت این قسمت را افزایش داد. با استفاده از کارت گرافیک GEFORCE GT 730 با ۲ گیگابایت حافظه‌ی GDDR5، سرعت این قسمت از برنامه نسبت به پردازنده‌ی مرکزی ۲,۷ گیگاهرتز با ۶ گیگابایت حافظه‌ی DDR3، به طور متوسط ۴ برابر افزایش پیدا کرد.

# فصل پنجم : ارزیابی و مقایسه

## ۵-۱ مقدمه

روش پیشنهادی بر روی دو پایگاه داده‌ی HMDB51 و Olympic Sports که در بخش ۱-۲ معرفی شده‌اند اعمال گردید. گام پنجره‌ی لغزان و طول هر بخش مطابق رابطه‌ی (۵-۱) اختیار گردید. در نتیجه هر بخش با بخش بعدی ۵۰ درصد اشتراک خواهد داشت.

$$\boldsymbol{W_D = 15\,, W_K = 30} \qquad (۱-۵)$$

در محاسبه‌ی تابع هسته‌ی انعطاف‌پذیر با زمان، پارامترهای رابطه‌ی (۳-۳۲) را مطابق (۵-۲) انتخاب نمودیم. تعداد مدل‌های گوسی آموزش داده شده برای استخراج بردار فیشر نیز z=256 اختیار شد.

$$\boldsymbol{\rho = 0.5\,, \sigma_x = \sigma_y = 1} \qquad (۲-۵)$$

در این فصل نتایج حاصل از شبیه‌سازی بیان شده و با روش‌های دیگر مقایسه می‌گردد.

## ۵-۲ معیار ارزیابی

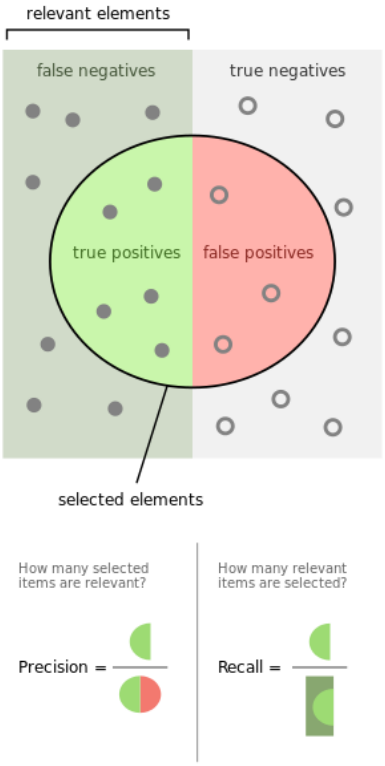
برای ارزیابی نتایج این تحقیق، همانند تحقیق‌های پیشین، از معیارهای صحت[۷۶] و دقت متوسط[۷۷] استفاده می‌شود تا امکان مقایسه‌ی این نتایج با تحقیقات گذشته فراهم گردد. علاوه بر این دقت[۷۸] و بازیافت[۷۹] در هر کلاس نیز گزارش خواهد شد. در ادامه این معیارهای ارزیابی معرفی می‌گردند.

در هر مسئله‌ی دسته‌بندی باینری[۸۰]، برخی از داده‌ها به دسته‌ی مورد نظر تعلق دارند و بقیه‌ی داده‌ها متعلق به آن دسته نیستند. به داده‌هایی که به دسته‌ی مورد نظر تعلق دارند عناصر مربوطه[۸۱] گفته می‌شود. پس از حل مسئله‌ی دسته‌بندی، بخشی از عناصر مربوطه به درستی به دسته‌ی

[۷۶] Accuracy
[۷۷] Average Precision (AP)
[۷۸] Precision
[۷۹] Recall
[۸۰] Binary Classification
[۸۱] Relevant Elements

موردنظر منتسب می‌گردند که به آن‌ها مثبت‌های صحیح[۸۲] گفته می‌شود. به آن بخش از عناصر مربوطه که مسئله‌ی دسته‌بندی آن‌ها را به دسته‌ی موردنظر منتسب نمی‌کند، منفی‌های اشتباهی[۸۳] گفته می‌شود.

از میان داده‌هایی که به دسته موردنظر تعلق ندارند، برخی به اشتباه به دسته‌ی موردنظر منتسب می‌شوند که به آن‌ها مثبت‌های اشتباهی[۸۴] گفته می‌شود. به آن بخش از داده‌های غیرمتعلق به دسته‌ی مورد نظر که به دسته‌ی موردنظر منتسب نمی‌گردند، منفی‌های صحیح[۸۵] گفته می‌شود. شکل (۵-۱) این موارد را نشان می‌دهد.




شکل (۵- ۱) یک مسئله‌ی دسته‌بندی باینری[۵۷].

با توجه به آن‌چه گفته شد، دقت و بازیافت برای هر دسته با روابط (۵-۳) و (۵-۴) تعریف می‌گردند.

---

[۸۲] True Positives (TP)
[۸۳] False Negatives (FN)
[۸۴] False Positives (FP)
[۸۵] True Negatives (TN)

$$\text{دقت} = \frac{\text{تعداد مثبت‌های صحیح}}{\text{تعداد مثبت‌های اشتباهی+تعداد مثبت‌های صحیح}} \quad (۵-۳)$$

$$\text{بازیافت} = \frac{\text{تعداد مثبت‌های صحیح}}{\text{تعداد منفی‌های اشتباهی+تعداد مثبت‌های صحیح}} \quad (۵-۴)$$

دقت متوسط عبارت است از میانگین دقت‌های بدست‌آمده از همه‌ی دسته‌ها. صحت نیز مطابق با رابطه‌ی (۵-۵) تعریف می‌گردد.

$$\text{صحت} = \frac{\text{تعداد داده‌هایی که دسته‌ی آن‌ها به‌درستی شناسایی شد}}{\text{تعداد کل داده‌ها}} \quad (۵-۵)$$

## ۵-۳ ارزیابی

همان‌طور که در بخش ۱-۲-۱ گفته شده، برای پایگاه داده‌ی HMDB51 سه ترکیب متفاوت از داده‌های آموزش و آزمایش وجود دارد. روش پیشنهادی را بر روی ترکیب شماره‌یک این پایگاه داده[۸۶] اجرا کردیم. جدول (۵-۱) دقت و بازیافت بدست آمده برای هر یک از فعالیت‌های این پایگاه داده را نشان می‌دهد.

جدول (۵ - ۱) مقادیر دقت و بازیافت برای هریک از فعالیت‌ها در ترکیب شماره‌یک پایگاه داده‌ی HMDB51.

| نام فعالیت | دقت % | بازیافت % |
|---|---|---|
| **'brush-hair'** | ۸۶,۶۶ | ۸۶,۶۶ |
| **'cartwheel'** | ۵۹,۲۵ | ۵۳,۳۳ |
| **'catch'** | ۹۲,۳۰ | ۸۰ |
| **'chew'** | ۵۴,۷۶ | ۷۶,۶۶ |
| **'clap'** | ۵۲,۵ | ۷۰ |
| **'climb-stairs'** | ۵۴,۵۴ | ۴۰ |
| **'climb'** | ۷۵ | ۹۰ |
| **'dive'** | ۶۵,۳۸ | ۵۶,۶۶ |
| **'draw-sword'** | ۳۶,۶۶ | ۳۶,۶۶ |
| **'dribble'** | ۷۰,۲۷ | ۸۶,۶۶ |

[۸۶] Split 1 of HMDB51

| ۶۶,۶۶ | ۵۸,۸۲ | 'drink' |
|---|---|---|
| ۶۶,۶۶ | ۴۰,۸۱ | 'eat' |
| ۵۰ | ۳۸,۴۶ | 'fall-floor' |
| ۵۰ | ۷۱,۴۲ | 'fencing' |
| ۶۰ | ۷۵ | 'flic-flac' |
| ۹۰ | ۸۴,۳۷ | 'golf' |
| ۶۶,۶۶ | ۵۷,۱۴ | 'handstand' |
| ۳۶,۶۶ | ۶۸,۷۵ | 'hit' |
| ۸۰ | ۶۴,۸۶ | 'hug' |
| ۶۰ | ۶۲,۰۶ | 'jump' |
| ۴۰ | ۶۶,۶۶ | 'kick-ball' |
| ۶۳,۳۳ | ۵۴,۲۸ | 'kick' |
| ۷۰ | ۶۳,۶۳ | 'kiss' |
| ۵۶,۶۶ | ۴۰,۴۷ | 'laugh' |
| ۴۰ | ۲۴ | 'pick' |
| ۷۶,۶۶ | ۵۸,۹۷ | 'pour' |
| ۲۰ | ۱۰۰ | 'pullup' |
| ۲۶,۶۶ | ۳۸,۰۹ | 'punch' |
| ۷۶,۶۶ | ۸۲,۱۴ | 'push' |
| ۷۶,۶۶ | ۹۲ | 'pushup' |
| ۷۰ | ۸۰,۷۶ | 'ride-bike' |
| ۵۳,۳۳ | ۷۶,۱۹ | 'ride-horse' |
| ۶۰ | ۳۶ | 'run' |
| ۵۶,۶۶ | ۵۳,۱۲ | 'shake-hands' |
| ۵۶,۶۶ | ۶۵,۳۸ | 'shoot-ball' |
| ۶۳,۳۳ | ۹۵ | 'shoot-bow' |
| ۳۶,۶۶ | ۶۸,۷۵ | 'shoot-gun' |
| ۸۰ | ۷۲,۷۲ | 'sit' |
| ۷۰ | ۸۷,۵ | 'situp' |
| ۱۶,۶۶ | ۵۰ | 'smile' |
| ۵۰ | ۵۷,۶۹ | 'smoke' |
| ۶۳,۳۳ | ۳۳,۹۲ | 'somersault' |
| ۷۳,۳۳ | ۷۵,۸۶ | 'stand' |
| ۱۶,۶۶ | ۳۱,۲۵ | 'swing-baseball' |
| ۴۳,۳۳ | ۲۷,۶۵ | 'sword-exercise' |
| ۳۰ | ۳۰ | 'sword' |

| | | |
|---|---|---|
| **'talk'** | ۷۰,۳۷ | ۶۳,۳۳ |
| **'throw'** | ۵۰ | ۳۰ |
| **'turn'** | ۶۴,۷ | ۷۳,۳۳ |
| **'walk'** | ۴۷,۳۶ | ۶۰ |
| **'wave'** | ۵۴,۵۴ | ۴۰ |

برخی فعالیت‌ها با درنظر گرفتن ترتیب وقایع بهتر شناسایی می‌شوند و برخی دیگر بدون در نظر گرفتن این ترتیب بهتر شناسایی می‌شوند. بنابراین همانطور که در بخش ۴-۲ گفته شد، یک‌بار تمام ویژگی‌های هر ویدئو به یک بردار فیشر تبدیل می‌شود و دو تابع هسته خطی برای آموزش و آزمایش محاسبه می‌شود و یک‌بار دیگر روش پیشنهادی اعمال می‌گردد و تابع هسته‌ی انعطاف‌پذیر با زمان برای آموزش و آزمایش محاسبه می‌گردد. سپس توابع هسته‌ی بدست آمده از این دو روش با هم جمع می‌شوند. جدول (۵-۲) صحت و دقت متوسط بدست آمده در هریک از این مراحل را نشان می‌دهد. همان‌طور که مشاهده می‌شود روش سلسله‌مراتبی صحت را نسبت به روش استفاده از تابع هسته‌ی خطی افزایش داده است ولی دقت متوسط تغییری نکرده است. اما با جمع کردن دو تابع هسته‌ی بدست آمده از این دو روش، صحت و دقت متوسط هر دو افزایش پیدا کردند.

جدول (۵ - ۲) صحت و دقت متوسط در ترکیب شماره‌یک HMDB51.

| روش/ معیار سنجش | صحت ٪ | دقت متوسط ٪ |
|---|---|---|
| **تابع هسته خطی** | ۵۷,۳۲ | ۵۹,۶۷ |
| **روش سلسه مراتبی** | ۵۷,۸۴ | ۵۹,۶۷ |
| **ترکیب دو تابع هسته** | ۵۷,۹۷ | ۶۱,۱۴ |

در مرحله‌ی دوم روش پیشنهادی به پایگاه داده‌ی Olympic Sports اعمال گردید. جدول (۵-۳) صحت و دقت بدست آمده برای هریک از فعالیت‌های این پایگاه داده را نشان می‌دهد.

جدول (۵ - ۳) صحت و دقت در هریک از فعالیت‌های پایگاه داده‌ی Olympic Sports.

| نام فعالیت | دقت % | بازیافت % |
|---|---|---|
| 'basketball-layup' | ۹۰ | ۹۰ |
| 'bowling' | ۸۰ | ۸۸,۸۹ |
| 'clean-and-jerk' | ۸۰ | ۸۰ |
| 'discus-throw' | ۶۸,۷۵ | ۱۰۰ |
| 'diving-platform-10m' | ۱۰۰ | ۱۰۰ |
| diving-springboard-' '3m | ۸۸,۸۹ | ۱۰۰ |
| 'hammer-throw' | ۱۰۰ | ۶۲,۵ |
| 'high-jump' | ۸۴,۶۲ | ۱۰۰ |
| 'javelin-throw' | ۱۰۰ | ۷۵ |
| 'long-jump' | ۶۶,۶۷ | ۱۰۰ |
| 'pole-vault' | ۱۰۰ | ۸۷,۵ |
| 'shot-put' | ۸۸,۸۹ | ۸۰ |
| 'snatch' | ۷۷,۷۸ | ۷۷,۷۸ |
| 'tennis-serve' | ۸۳,۳۳ | ۸۱,۴۳ |
| 'triple-jump' | ۱۰۰ | ۲۵ |
| 'vault' | ۱۰۰ | ۸۰ |

جدول (۵-۴) صحت و دقت متوسط بدست آمده را نشان می‌دهد. همان‌طور که مشاهده می‌شود، در این پایگاه داده روش سلسله‌مراتبی تعداد فعالیت‌های کمتری را نسبت به روش تابع هسته خطی شناسایی کرده است، ولی دقت متوسط آن بیشتر شده است. این‌بار هم با جمع کردن دو تابع هسته صحت و دقت متوسط نسبت به هریک از دو روش دیگر به تنهایی، بیشتر شده است.

جدول (۵ - ۴) صحت و دقت متوسط در پایگاه داده‌ی Olympic Sports.

| روش/ معیار سنجش | صحت ٪ | دقت متوسط ٪ |
|---|---|---|
| تابع هسته خطی | ۸۲,۸۳ | ۸۵,۹۰ |
| روش سلسه مراتبی | ۸۲,۰۸ | ۸۷,۲۵ |
| ترکیب دو تابع هسته | ۸۵,۰۷ | ۸۸,۰۵ |

## ۵-۴ مقایسه

جدول (۵-۵) مقایسه‌ی صحت و دقت متوسط روش پیشنهادی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، روش پیشنهادی نسبت به روش‌های دیگر فعالیت‌های بیشتری را به‌درستی شناسایی کرده است. و روش [۲۷] نسبت به روش‌های دیگر دقت متوسط بالاتری بدست آورده است.

جدول (۵ - ۵) مقایسه‌ی روش پیشنهادی با روش‌های قبلی در پایگاه داده‌ی Olympic Sports.

| روش/ معیار سنجش | صحت ٪ | دقت متوسط ٪ |
|---|---|---|
| **مرجع [۸]** | - | ۷۲,۱ |
| **مرجع [۵۸]** | - | ۷۶,۲ |
| **مرجع [۵۹]** | ۸۲,۷ | - |
| **مسیر متراکم بهبود یافته [۲۷]** | - | **۹۰,۲** |
| **تابع هسته انعطاف‌پذیر با زمان [۳۴]** | ۸۴,۳ | ۸۹,۹ |
| **روش پیشنهادی این تحقیق** | **۸۵,۰۷** | ۸۸,۰۵ |

# فصل ششم : بحث و نتیجه‌گیری

## ۶-۱ بحث و نتیجه‌گیری

در این تحقیق روشی سلسله‌مراتبی برای شناسایی فعالیت در رشته‌های ویدئو معرفی گردید که موجب افزایش صحت شناسایی فعالیت نسبت به روش‌های قبلی گردید (جدول ۵-۵). یکی از مهم‌ترین قسمت‌های یک سیستم شناسایی فعالیت، قسمت استخراج ویژگی از رشته‌های ویدئویی است. این ویژگی‌ها باید به تغییرات ظاهر و رنگ در تصویر و هم‌چنین به حرکت‌های شدید ناگهانی مانند حرکت‌های ناشی از لرزش دوربین مقاوم باشد. بر این اساس استفاده از روش مسیر متراکم بهبودیافته برای استخراج ویژگی، به دلیل دارا بودن خصوصیات ذکر شده، انتخابی مناسب است. استفاده از بردار فیشر بهبودیافته برای کد کردن، موجب فشرده شدن ویژگی‌ها در یک بردار می‌شود؛ اما این روش اطلاعات مربوط به ارتباط زمانی بین ویژگی‌ها را حذف می‌کند. بنابراین در روش پیشنهادی بردار فیشر از روی هر مجموعه از ویژگی‌هایی که از نظر زمانی به یکدیگر برتری ندارند استخراج شده است. هم‌چنین ترکیب دو روش اخیر به نام‌های ادغام رتبه و تابع هسته‌ی انعطاف‌پذیر به‌زمان، موجب بهره‌گیری از مزیت‌های هر دو روش می‌گردد. در روش ادغام رتبه ترتیب فریم‌ها در نظر گرفته می‌شود. روش TFK ضمن ایجاد انعطاف‌پذیری به زمان امکان مقایسه‌ی دو دنباله‌ی غیرهم‌طول را فراهم می‌کند. اگرچه در روش TFK نیز ترتیب تکه‌های ویدئو درنظر گرفته شده است، اما اطلاعات مربوط به ترتیب‌های زمانی درون هر تکه حذف شده است. اما در روش پیشنهادی با ادغام این دو روش، اطلاعات مربوط به ترتیب زمانی در دو مرحله استخراج می‌گردد.

با توجه به اینکه بعضی از فعالیت‌ها با روش ترتیبی بهتر شناسایی شده و برخی دیگر بدون استفاده از این روش بهتر شناسایی می‌شوند، هر دو روش اجرا شده و درپایان توابع هسته‌ی بدست آمده از آن‌ها با یکدیگر جمع می‌شود.

## ۶-۲ کار آینده

روش TFK با وجود کارایی مناسب، از نظر پردازش هزینه‌ی زیادی دارد. با فرض اینکه طول ویدئوها تقریبا با هم برابر باشد، زمان محاسبه‌ی مقادیر تابع هسته آموزش با مجذور تعداد ویدئوهای آموزشی ارتباط دارد. به عنوان مثال محاسبه‌ی تابع هسته‌ی انعطاف‌پذیر با زمان برای داده‌های آموزش و آزمایشی در ترکیب شماره‌یک پایگاه داده‌ی HMDB51 با کارت گرافیک GEFORCE GT730 حدود ۵۰ روز به‌طول انجامید. بنابراین یافتن راهی برای کاهش حجم این پردازش موجب بهبود این روش از نظر زمان اجرا خواهد گردید.

# واژگان و اصطلاحات

| | |
|---|---|
| Closed-Circuit Television (CCTV) | دوربین مدار بسته |
| IP Camera | دوربین تحت شبکه |
| User Generated Content (UGC) | مجتوای تولید شده توسط کاربر |
| Index | فهرست کردن |
| Brown University | دانشگاه براون |
| Matlab | متلب |
| Annotation | حاشیه نویسی |
| Single-Layered | تک‌لایه |
| Hierarchical | سلسله‌مراتبی |
| Space-Time | فضا-زمان |
| Sequential | ترتیبی |
| Sliding Window | پنجره‌ی لغزان |
| Similarity Measure | معیار شباهت |
| Motion Energy Image (MEI) | تصویر انرژی حرکت |
| Motion History Image (MHI) | تصویر تاریخچه حرکت |
| Real Time | زمان‌واقعی |
| Global | سراسری |
| Local | محلی |
| Optical Flow | جریان نوری |
| Patches | بخش‌ها |
| Trajectory | مسیر حرکت |
| Kinect | کینکت |
| Dense | متراکم |
| Space-Time Interest Point (STIP) | نقاط مهم فضا-زمانی |
| Harris | هریس |
| Occlusion | انسداد |
| Natural Language Processing | پردازش زبان طبیعی |
| Bag of Words (BoW) | کیف کلمات |
| Fisher Vector | بردار فیشر |
| Temporal Scale | مقیاس زمانی |
| Gradient | گرادیان |
| Orientation | جهت |
| Support Vector Machine (SVM) | بردار ماشین پستیبان |
| Principle Component Analysis (PCA) | تحلیل مؤلفه‌ی اصلی |
| Dense Trajectory | مسیر متراکم |

| | |
|---|---|
| Histogram of Oriented Gradients (HOG) | هیستوگرام گرادیان‌های جهت‌دار |
| Histogram of Optical Flow (HOF) | هیستوگرام جریان نوری |
| Motion Boundary Histogram (MBH) | هیستوگرام مرز حرکت |
| Scale-Invariant Feature Transform (SIFT) | تبدیل ویژگی مقیاس‌نابسته |
| Kanade–Lucas–Tomasi (KLT) Tracker | ردیاب کانادی-لوکاس-توماسی |
| Irregular Abrupt Motions | حرکت نامنظم ناگهانی |
| Improved Dense Trajectory (IDT) | مسیر متراکم بهبودیافته |
| State Models | مدل‌های حالت |
| Hidden Markov Model (HMM) | مدل مخفی مارکوف |
| Dynamic Bayesian Network (DBN) | شبکه بیزین پویا |
| Variable-Length Markov Models (VLMM) | مدل مارکوف با طول متغیر |
| Conditional Random Field (CRF) | میدان تصادفی شرطی |
| Kernel Function | تابع هسته |
| Ranking Machine | ماشین رتبه‌بندی |
| Ranking Function | تابع رتبه‌بندی |
| VideoDarwin | ویدئوداروین |
| Rank Pooling | ادغام رتبه |
| Atomic Actions | عمل‌های بنیادی |
| Hierarchical Sum-Product Network (SPN) | شبکه چندلایه‌ی جمع-ضرب |
| Time Flexible Kernel (TFK) | تابع هسته انعطاف‌پذیر با زمان |
| Lucas-Kanade | لوکاس-کانادی |
| Grid | شبکه |
| Autocorrelation | خودهمبستگی |
| Median Filtering Kernel | هسته‌ی فیلتر میانه |
| Homography | هموگرافی |
| Random Sample Consensus (RANSAC) | اجماع تصادفی نمونه |
| Human Tracker | ردیاب انسان |
| Codebook | کتاب کد |
| Hard Assignment | انتساب سخت‌گیرانه |
| Soft Assignment | انتساب نرم |
| K-means Clustering | خوشه‌بندی کی-میانگین |
| Distance Measure | معیار فاصله |
| Sparse | پراکنده |
| Gaussian Mixture Models | مدل‌های مخلوط گوسی |
| L-2 Norm | نُرم دو |
| Bhattacharyya | بهاتاچاریا |
| Moving Average (MA) | میانگین متحرک |
| Smooth | نَرم |
| Regression | رگرسیون |

| | |
|---|---|
| Support Vector Regression | **رگرسیون بردار پشتیبان** |
| Accuracy | **صحت** |
| Average Precision (AP) | **دقت متوسط** |
| Precision | **دقت** |
| Recall | **بازیافت** |
| Binary Classification | **دسته‌بندی باینری** |
| Relevant Elements | **عناصر مربوطه** |
| True Positives (TP) | **مثبت‌های صحیح** |
| False Negatives (FN) | **منفی‌های اشتباهی** |
| False Positives (FP) | **مثبت‌های اشتباهی** |
| True Negatives (TN) | **منفی‌های صحیح** |

# مراجع


[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Sureys,* vol. 43, no. 3, pp. 1-43, 2011.

[2] H. V. Chandrashekhar, "HUMAN ACTIVITY REPRESENTATION, ANALYSIS AND RECOGNITION," Master thesis, DEPARTMENT OF ELECTRICAL ENGINEERING, INDIAN INSTITUTE OF TECHNOLOGY, KANPUR, India, 2006.

[3] A. Kumar, "Panic Detection in Human Crowds using Sparse Coding," Master Thesis, Systems Design Engineering Department, University of Waterloo, Ontario, Canada, 2012.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556-2563.

[5] M. W. Green, "The Appropriate and Effective Use of Security Technologies in US Schools. A Guide for Schools and Law Enforcement Agencies," Technical Report NCJ 178265, Sandia National Laboratories, 1999.

[6] J. M. Wolfe, T. S. Horowitz, and N. M. Kenner, "Cognitive psychology Rare items often missed in visual searches," *Nature,* 10.1038/435439a vol. 435, no. 7041, pp. 439-440, 2005.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," presented at the Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, San Diego, California, USA, 2007.

[8] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," in *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 392-405.

[9] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding,* vol. 117, no. 6, pp. 633-659, 2013.

[10] (May 2017). *P. Doll´ar. Piotr's image and video matlab toolbox.* Available: http://vision.ucsd.edu/~pdollar/toolbox/doc/

[11] G. T. Pusiol, "Discovery of human activities in video,"Ph.D dissertation, Computer Science Department , Université Nice Sophia Antipolis,Nice, France, 2012.

[12] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer,* journal article vol. 29, no. 10, pp. 983-1009, October 01 2013.

[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, 2001.

[14] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield, "Combining Local and Global Image Features for Object Class Recognition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 2005, pp. 47-47.

[15] E. Shechtman and M. Irani, "Space-Time Behavior-Based Correlation-OR-How to Tell If Two Underlying Motion Fields Are Similar Without Computing Them?," *IEEE*

*Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 11, pp. 2045-2056, 2007.

[16] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 405-412.

[17] C. Rao and M. Shah, "View-invariance in action recognition," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, 2001, vol. 2, pp. II-316-II-322 .

[18] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *Tenth IEEE International Conference on Computer Vision (ICCV'05)* ,2005, vol. 1, pp. 150-157.

[19] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM,* vol. 56, no. 1, pp. 116-124, 2013.

[20] L. Zelnik-Manor and M. Irani, "Statistical analysis of dynamic actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 1530-1535, 2006.

[21] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, vol 1, pp. 432-439.

[22] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition( ICPR'04), 2004,*. vol. 3, pp. 32-36.

[23] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision,* journal article vol. 79, no. 3, pp. 299-318, September 01 2008.

[24] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Computer Society Conference on Computer Vision and Pattern Recognition,* 2011, pp. 3169-3176.

[25] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150-1157.

[26] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *proc 7th Intl Joint conf on artificial Intelligence(IJCAI)*, 1981, pp 674-679.

[27] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551-3558.

[28] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine,* vol. 3, no. 1, pp. 4-16, 1986.

[29] P. Dagum, A. Galper, and E. Horvitz, "Dynamic network models for forecasting," presented at the Proceedings of the Eighth international conference on Uncertainty in artificial intelligence, Stanford, CA, 1992.

[30] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, vol 2, pp. 726-733.

[31] A. Galata, N. Johnson, and D. Hogg, "Learning Variable-Length Markov Models of Behavior," *Computer Vision and Image Understanding,* vol. 81, no. 3, pp. 398-413, 2001.

[32] I. Guyon and F. Pereira, "Design of a linguistic postprocessor using variable memory length Markov models," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 454-457.

[33] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understanding,* vol. 104, no. 2, pp. 210-220, 2006.

[34] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, "A Time Flexible Kernel framework for video-based activity recognition," *Image and Vision Computing,* vol. 48, pp. 26-36, 2016.

[35] B. Fernando, E. Gavves, J. O. M, A. Ghodrati, and T. Tuytelaars, "Rank Pooling for Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 773-787, 2017.

[36] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378-5387.

[37] M. Brand, "Understanding manipulation in video," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 94-99.

[38] J. F. Allen and G. Ferguson, "Actions and Events in Interval Temporal Logic," *Journal of Logic and Computation,* vol. 4, no. 5, pp. 531-579, 1994.

[39] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 838-845 vol. 1.

[40] M. R. Amer and S. Todorovic, "Sum Product Networks for Activity Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 38, no. 4, pp. 800-813, 2016.

[41] H. Poon and P. Domingos, "Sum-product networks: A new deep architecture," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 689-690.

[42] M. R. Amer, "Hierarchical graphical models for activity recognition in videos," Ph.D Diss, School of Electrical Engineering and Computer Science, Oregon State University,oregan, USA, 2014..

[43] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* journal article vol. 60, no. 2, pp. 91-110, November 01 2004.

[44] R. K. McConnell, "Method of and apparatus for pattern recognition," U.S. Patent No. 4,567,610. 28 Jan. 1986.

[45] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM computing surveys (CSUR),* vol. 27, no. 3, pp. 433-466, 1995.

[46] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation,* vol. 5, no. 1-10, p. 4, 2001.

[47] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," in *Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 428-441.

[48] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision,* vol. 103, no. 1, p. 60, 2013.

[49] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* vol. 24, no. 6, pp. 381-395, 1981.

[50] N. Jojic and A. Perina, "Multidimensional counting grids: Inferring word order from disordered bags of words," *arXiv.org preprint arXiv:1202.3752,* 2012.

[51] J. C. v. Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual Word Ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 7, pp. 1271-1283, 2010.

[52] F. Perronnin and C. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[53] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 143-156.

[54] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," presented at the Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 2010.

[55] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of machine learning research,* vol. 9, no. 8, pp. 1871-1874, 2008.

[56] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST),* vol. 2, no. 3, p. 27, 2011.

[57] Walber (https://commons.wikimedia.org/wiki/File:Precisionrecall.svg), https://creativecommons.org/licenses/by-sa/4.0/legalcode

[58] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos, "Dynamic pooling for complex event recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2728-2735.

[59] A. Gaidon, Z. Harchaoui, and C. Schmid, "Recognizing activities with cluster-trees of tracklets," in *BMVC 2012 - British Machine Vision Conference*, Guildford, United Kingdom, 2012, pp. 30.1-30.13.

## ABSTRACT

With the increasing number of videos being uploaded to the internet and also with the growing rate of crimes and terrorist attacks in recent years, the need for an automatic system aimed at activity recognition is felt more than before. Firstly, Video sharing websites are looking for a way to recognize and remove illegal or violent content automatically. And secondly, human-based video surveillance is not much efficient because of tiredness and distraction as well as its cost. So, a research field for automatic activity recognition was created to address these issues.

On account of variations in video background and people's appearance in the video, we need descriptors that can represent the motion and are also robust to noise and irregular abrupt motions caused by camera motion. Another challenge in this field is to deal with the variations in the way one single activity can be performed. Accordingly we need to obtain a semantic understanding of what is occurring in the video so as to address it.

The present thesis proposes a hierarchical approach for activity recognition, which considers the succession of temporal information in the video and is able to compare videos of different lengths. Improved dense trajectory features are used and then encoded by means of improved fisher vector. Temporal information of the video is extracted in two layers by employing rank pooling and time-flexible kernel methods. Afterwards activities are separated using a support vector machine. The implementation of the proposed method resulted in 57.97% accuracy on split 1 of hmdb51 dataset and 85.07% accuracy on Olympic sports dataset.

**Keywords**: Activity Recognition, Activity Detection, HMDB51 Dataset, Olympic Sports Dataset, Rank Pooling, Time Flexible Kernel, Improved Dense Trajectories, Video Processing

**Shahrood University of Technology**

**Faculty of electrical and robotic engineering**

**MSc Thesis in Electronics**

# Human activity recognition based on video processing

By: Amir Mahmoudi

Supervisor(s):

Dr Alireza Ahmadyfard

September 2017