Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: www.jscit.nit.ac.ir
Volume 7, Number 1, Spring & Summer 2018, pp. 13-28
Received: 2018/01/27, Revised: 2018/06/03, Accepted: 2018/10/29

Human Action Recognition Using FREAK-HOG and CSVM

Nacer Farajzadeh[1*], Mahdi Hashemzadeh[2]
1*- Faculty of IT and Computer Engineering Department, Azarbaijan Shahid Madani University, Tabriz, Iran.
2- Faculty of IT and Computer Engineering Department, Azarbaijan Shahid Madani University, Tabriz, Iran.
[1*]n.farajzadeh@azaruniv.ac.ir, [2]hashemzadeh@azaruniv.ac.ir

Corresponding author address: Nacer Farajzadeh, Faculty of IT & Computer Engineering, Azarbijan Shahid Madani University, Tabriz, Iran, Post Code : 5375171379.

Abstract- Recently, human action recognition in videos has become an interesting area of research due to its variety of important applications such as intelligent security supervisions, smart environments, education, health-care monitoring systems, data mining, etc. There are, however, number of challenges that makes the development of these systems a bit harder than the common machine vision systems, both in accuracy and efficiency: changes in illumination, moving background, cluttered backgrounds, camera motions, complexity of the actions, to name a few. One of the commonly used methods for automatic human action recognition is to, firstly, extract some feature points within the video frames, then describe those points locally, and finally, code (cluster) them to feed a learning algorithm to build an action recognition model. In this paper, we aim to increase the accuracy of these methods by introducing the use of texture information extracted using a human retina-inspired algorithm (FREAK) together with the appearance-based information of the moving objects. In order to increase the efficacy and reduces the overhead of furthered texture information in the model building phase and, of course, in hope of increasing the accuracy as well, we propose to use a cascade approach to build the desired model. Experiments on two large datasets namely UCF101 and HMDB51, confirm that the proposed method achieves a very comparable results with the state-of-the-art methods.

Keywords- Human Action recognition, Texture features, FREAK, HOG, Cascade model, Support Vector Machine.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: www.jscit.nit.ac.ir
جلد ۷، شماره ۱، بهار و تابستان ۱۳۹۷ ، صفحه ۱۳-۲۸
دریافت: ۹۶/۱۱/۰۷، بازنگری: ۹۷/۰۳/۱۳، پذیرش: ۹۷/۰۸/۰۷

بازشناسایی فعالیت‌های انسان در ویدیو با استفاده از ویژگی‌های FREAK-HOG و ماشین بردار پشتیبان آبشاری

ناصر فرج زاده*[۱]، مهدی هاشم‌زاده[۲]
*۱- دانشکده فناوری اطلاعات و مهندسی کامپیوتر ، دانشگاه شهید مدنی آذربایجان، ایران.
۲- دانشکده فناوری اطلاعات و مهندسی کامپیوتر ، دانشگاه شهید مدنی آذربایجان، ایران.
[1*]n.farajzadeh@azaruniv.ac.ir, [2]hashemzadeh@azaruniv.ac.ir

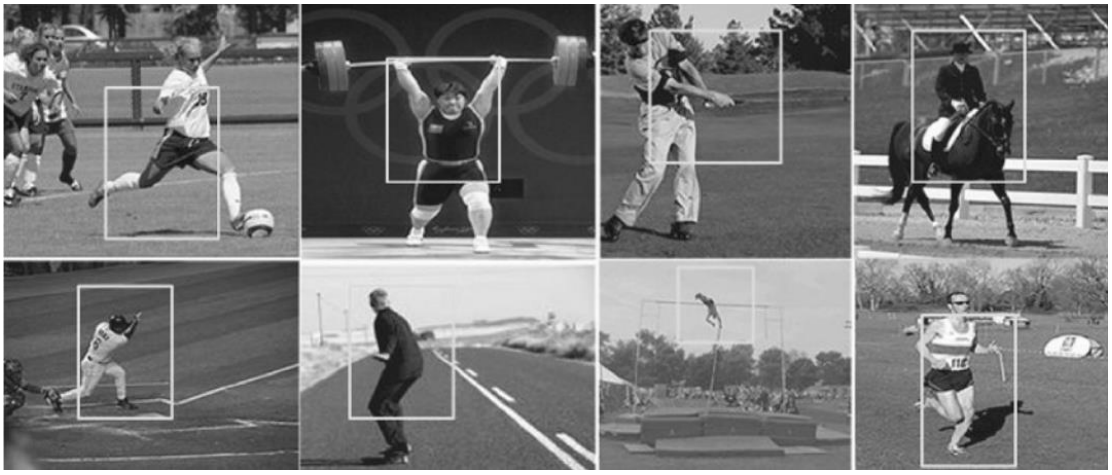
* نشانی نویسنده مسئول: ناصر فرجزاده، تبریز، ۳۵ کیلومتری جاده تبریز مراغه، دانشگاه شهید مدنی آذربایجان، دانشکده فناوری اطلاعات و مهندسی کامپیوتر، کد پستی: ۵۳۷۵۱۷۱۳۷۹

چکیده- در سال‌های اخیر، بازشناسایی خودکار فعالیت‌های انسان در ویدیو تبدیل به یکی از حوزه‌های مهم تحقیقاتی شده است. دامنه کاربرد این تحقیقات گسترده بوده و در سامانه‌هایی نظیر سامانه‌های نظارتی و امنیتی، رابط‌های کاربریِ واکنش‌گرا، استخراج اطلاعات حرکتی-رفتاری، آموزش و مراقبت‌های بهداشتی مورد استفاده و بهره‌برداری قرار گرفته است. اما چالش‌هایی نظیر تغییرات شدت رو شنایی تصاویر، متحرک بودن پس‌زمینه و دوربین، شلوغی و ازدحام، پیچیدگی و تنوع فعالیت انجام شونده باعث شده‌اند تو سعه سامانه‌هایی که از نظر دقت بازشناسایی مورد اطمینان بوده و در عین حال سرعت عمل قابل قبولی داشته باشند، با مشکل مواجه شوند. یکی از روش‌های مر سوم باز شنا سایی فعالیت‌های انسان به این صورت است که ابتدا برخی ویژگی‌های تصویر به همراه تو صیف آن ویژگی‌ها از فریم‌های ویدیویی به‌صورت محلی استخراج می‌شود. سپس، این ویژگی‌ها برای استفاده یک الگوریتم یادگیری جهت ساخت مدل بازشناسایی کننده فعالیت، کدگذاری می‌شوند. در این مقاله با هدف افزایش دقت بازشناسایی فعالیت‌ها، بهره‌گیری از یک تو صیف‌گر بافتیِ الهام گرفته شده از شبکیه چ شم انسان و ترکیب آن با یک تو صیف‌گر ظاهری-حرکتی برای تو صیف نقاطِ ویژگی استخراج شده از توالی فریم‌ها، پیشنهاد می شوند. همچنین برای افزایش سرعت ساخت مدل و کاهش هزینه‌های پرداز شی نا شی از ترکیب ویژگی‌های پیشنهادی، یک رویکرد آبشاری برای ساخت مدل طبقه‌بندی کننده ارائه می‌شود. نتایج آزمایش‌های انجام گرفته بر روی دو پایگاه‌داده‌ی بزرگ UCF101 و HMDB51 نشان می‌دهد که روش پیشنهادی سرعت عمل‌کرد بسیار خوبی دارد و دقت بازشناسایی آن قابل مقایسه با آخرین دستاوردها در این حوزه است.

واژه‌های کلیدی: بازشناسایی فعالیت، ویژگی‌های بافتی، FREAK، HOG، مدل آبشاری، ماشین بردار پشتیبان.

۱- مقدمه

بازشناسایی فعالیت‌های انسان با استفاده از بینایی ماشین، زمینه تحقیقاتی بسیار جذابی است که در چند سال اخیر اهمیت قابل ملاحظه‌ای در کاربردهایی نظیر خانه‌های هوشمند، سامانه‌های نظارتی و امنیتی، رابط‌های کاربری واکنش‌گرا، آموزش و مراقبت‌های بهداشتی و استخراج اطلاعات از ویدیو، پیدا کرده است [۱]. همچنین، آنالیز رفتارهای حرکتی انسان از دیر باز مورد علاقه سایر علوم از قبیل روانشناسی و بیولوژی بوده است. به‌طور کلی فعالیت‌های انسانی را می‌توان نتیجه یکی از سه گروه حرکتی زیر



شکل ۱: نمونه‌هایی از بازشناسایی خودکار فعالیت‌های ورزشی [۷].

دانست: ۱) حرکت‌های ساده[1] نظیر حرکت سر و ابرو، ۲) حرکت‌های ترکیبی[2] نظیر راه رفتن و پریدن، و ۳) حرکت‌های تعاملی[3] نظیر دست دادن و روبوسی کردن [۲].

هدف یک سامانه‌ی خودکار بازشناسایی فعالیت‌های انسان عبارت است از: ۱) مشاهده و یا زیر نظر گرفتن نامحسوس حرکات انجام شده توسط یک فرد، ۲) بازشناسایی فعالیت انجام گرفته و ۳) اتخاذ تصمیم مناسب در پاسخ به فعالیت انجام شده [۳]. به‌عنوان مثال، آگاهی از این موضوع که کاربر یک سامانه در حال انجام چه فعالیتی است، این امکان را به بوجود می‌آورد تا پاسخ مناسبی مانند صدور پیغام‌های مفید و راه‌گشا به کاربر به‌طور هوشمند داده شود؛ مثلا از طریق شناسایی فعالیت‌های خاص ورزشی، می‌توان محیط هوشمندی را توسعه داد که در آن با فراهم آوردن اطلاعات مهم حرکتی، و همراهی با ورزشکار به‌طور پیوسته، حرکات نادرست وی را گوشزد و فرآیند یادگیری را سرعت بخشید. شکل ۱ نمونه‌هایی از فعالیت‌های ورزشی و بازشناسایی برخی حرکات را نشان می‌دهد.

انسان‌ها به سادگی قادر به درک و بازشناسایی فعالیت‌های مختلف هستند. علی‌رغم این‌که عمل بازشناسایی فعالیت‌ها برای یک انسان بسیار آسان و طبیعی به نظر می‌رسد، دارای پردازش‌های پیچیده شناختی بوده [۴] و پیاده‌سازی چنین قابلیتی در ماشین نیازمند عملیاتی نظیر درک محیط، یادگیری از مشاهدات پیشین و ایجاد مدلی دقیق برای تعیین نوع فعالیت است.

یکی از راهکارهای متداول در این حوزه، استفاده از روش‌های یادگیری ماشین برای ساخت مدلی جهت بازشناسایی رفتارها در تصاویر ویدیویی است. به این ترتیب که پس از انجام برخی پیش پردازش‌ها روی تصاویر ویدیویی، ویژگی‌های خاصی از تصویر و یا دنباله‌ای از تصاویر (فریم‌های ویدیویی) استخراج شده و با استفاده از یک الگوریتم یادگیری، طبقه‌بندی کننده‌ای (مدلی) بر اساس این ویژگی‌ها ساخته می‌شود. این طبقه‌بند نمونه‌های مشابه رفتارهای آموزش دیده را در تصاویر ویدیویی که در آینده دریافت می‌کند، بازشناسایی و طبقه‌بندی می‌کند.

در عمل و در سامانه‌های کاربردی، چالش‌های متعددی در هریک از مراحل مختلف پیاده‌سازی وجود دارد که بازشناسایی خودکار فعالیت‌های انسانی را دشوار می‌کنند. برای مثال، مقاومت در برابر خطاهای حاصل از تغییرات ناگهانی حرکت‌ها و رفتارها در مراحل اولیه پردازش، نحوه بازنمایی و ارائه‌ی فعالیت در مراحل میانی پردازش و نحوه ارائه‌ی معنایی آن در سطح بالای پردازش، از جمله‌ی این دشواری‌ها هستند. همچنین موارد دیگری از قبیل وجود پس-زمینه‌های متفاوت، حجم بالای داده‌ها، زوایای مختلف دید و شرایط گوناگون محیط تصویربرداری در تصاویر ویدیویی نیز جزء چالش‌های تاثیرگذار این سامانه‌ها بشمار می‌آیند [۵].

هدف اصلی که در بیشتر سامانه‌های بازشناسایی مبتنی بر بینایی ماشین مورد توجه قرار می‌گیرد، تلاش برای افزایش کارایی سامانه از نظر دقت بازشناسایی و قدرت تعمیم آن در شرایط متفاوت ناشی از تغییرات روشنایی، متحرک بودن پس‌زمینه و دوربین، شلوغی و ازدحام در تصاویر، پیچیدگی و تنوع فعالیت‌های انجام شونده است. راهکاری که اغلب برای افزایش دقت مورد توجه محققین قرار می‌گیرد، استفاده از ویژگی‌های مقاوم نسبت به تغییرات ذکر شده است. از میان راهکارهای موجود، روش شناسایی و ردیابی نقاط ویژگی[4] (که مقاوم در برابر تغییرات و نوسانات تصویر هستند) در توالی فریم‌ها و توصیف رفتار این نقاط، بیشتر از سایر راهکارها مورد

توجه قرار گرفته است [۶].

در کنار روش‌های مبتنی بر نقاط ویژگی، یکی از رویکردهای مرسوم برای دست‌یابی به دقت بازشناسایی مطلوب، استفاده از داده‌های آموزشی متنوع برای ساخت مدلی قوی و کارآمد است. البته مدت زمان طولانی مورد نیاز برای پردازش حجم بسیار زیادی از داده‌ها که در فریم‌های متوالی ویدیویی وجود دارند، می‌تواند یک محدودیت اساسی در این کاربرد به‌حساب بیاید.

راهکاری که اغلب توسط محققین برای کاهش مدت زمان پردازش داده و ساخت مدل مورد استفاده قرار می‌گیرد، استفاده از روش کاهش ابعاد ویژگی‌های استخراج شده است [۴]. اما لازم به ذکر است که به‌کارگیری روش‌های کاهش بُعد که در اکثر موارد صرفاً به‌صورت مهندسی شده و بدون توجه به ماهیت مسئله انجام می‌گیرد، معمولا باعث از دست رفتن اطلاعات ارزشمندی می‌شود که ممکن است در فرآیند بازشناسایی بسیار مفید باشند. از این‌رو، بدیهی است که راهکار کاهش ابعاد ویژگی برای افزایش سرعت پردازش، روش مناسبی برای به‌کارگیری در این سامانه نیست و ممکن است باعث کاهش چشم‌گیر دقت سامانه شود.

در این مقاله، برای افزایش دقت عمل بازشناسایی خودکار فعالیت انسانی، یک توصیف‌گر بافتی الهام گرفته شده از پردازش‌های سطح پایین شبکیه چشم انسان، جهت استخراج اطلاعات بافتی نقاط ویژگی موثر در بازشناسایی حرکات مختلف انسان، معرفی می‌شود. این نوع از توصیف‌گرها به دلیل ماهیت دودویی‌شان، هزینه پردازشی بسیار ناچیزی در فاز بازشناسایی دارند و از این جهت نیز بسیار مناسب سامانه مورد نظر هستند. همچنین، به منظور افزایش سرعت ساخت مدل طبقه‌بندی کننده توام با هدف غلبه بر چالش استفاده از داده‌های آموزشی زیاد، یک الگوریتم آموزشی آبشاری برای ساخت مدل ارائه می‌شود. استفاده از این روش برای ساخت مدل نه تنها زمان بالاسری ایجاد شده در به‌کارگیری ویژگی‌های بافتی را به حداقل می‌رساند، بلکه به میزان قابل توجهی زمان ساخت مدل را کاهش داده و در اغلب موارد باعث افزایش دقت طبقه‌بندی نیز می‌گردد.

نتایج آزمایش‌های انجام گرفته بر روی پایگاه‌داده‌های ویدیویی بزرگی که شامل حرکات متنوع از فعالیت‌های مختلف انسانی هستند، نشان می‌دهند که روش پیشنهادی به دلیل بهره‌گیری از ویژگی‌های بافتی غنی و عدم استفاده از فرآیند کاهش ابعاد، از دقت بازشناسایی بسیار مناسب و قابل مقایسه با روش‌های مرز دانش برخوردار است. همچنین نتایج به‌دست آمده نشان از افزایش چشم‌گیر سرعت ساخت مدل به دلیل استفاده از مدل آبشاری پیشنهادی دارد.

ادامه مقاله به شرح ذیل بخش‌بندی شده است: در بخش ۲ کارهای پیشین انجام شده در زمینه بازشناسایی فعالیت‌های انسانی مرور می‌شود. در بخش ۳ روش پیشنهادی معرفی می‌شود و در بخش ۴ نتایج آزمایش‌های ارزیابی کارایی آن ارایه می‌گردد. در بخش ۵ نتیجه‌گیری و محورهای توسعه و مطالعه بیشتر ارائه می‌شوند.

۲- کارهای پیشین

در پژوهش‌های اولیه برای خودکار کردن بازشناسایی فعالیت‌های انسانی، به‌کارگیری شکل و طرح‌واره بدن انسان مورد توجه قرار گرفته است [۸-۱۰]. در [۸]، یاموتو و همکارانش سعی کردند در هر تصویر، الگویی[۵] شبیه به اندام انسان را استخراج کرده و با استفاده از مدل مارکوف مخفی[۶] عمل بازشناسایی فعالیت را انجام دهند. در تحقیقی دیگر، بابیک و همکارش نیز روشی برای تشخیص الگوهای شبیه به اندام انسان در هر فریم، معرفی کردند [۹]. سپس، با تجمیع اختلاف الگوها در فریم‌های متوالی، ویژگی‌هایی با عناوین MEI[۷] و MHI[۸] استخراج کردند و برای بازشناسایی رفتارها از تطبیق این ویژگی‌ها بهره بردند. بلنک و همکارانش [۱۰] روشی برای ایجاد ساختار سه‌بُعدی از اختلاف الگوها در طول زمان معرفی کردند و با استفاده از معادله پواسن، برخی ویژگی‌های مکانی-زمانی را از اختلاف الگوها استخراج کرده و از روش نزدیکترین همسایه برای عمل بازشناسایی بهره بردند.

گروه دیگری از روش‌ها، نقاط مفصلی، خط سیر و تغییرات آن‌ها را در توالی فریم‌ها، مورد توجه قرار داده‌اند. به‌عنوان مثال، ییلماز و شاه با بررسی مسیر حرکتِ[۹] نقاط مفصلی و مقایسه آن‌ها با یکدیگر، روشی برای بازشناسایی فعالیت‌های انسان ارایه کردند [۱۱]. همچنین علی و همکارانش نیز از مسیر حرکت نقاط مفصلی بهره برده و مجموعه‌ای از ویژگی‌های مقاوم به تغییرات را برای بازشناسایی حرکات مختلف معرفی کردند [۱۲]. با این‌که روش‌های مبتنی بر مسیر حرکت نقاط مفصلی از کارایی بسیار خوبی برخوردار هستند، اما به دلیل عدم شناسایی دقیق نقاط مفصلی در شرایط محیط خارج از آزمایشگاه، استفاده از این روش‌ها با محدودیت‌هایی مواجه است.

در تحقیقات اخیر، روش‌های مبتنی بر استفاده از ویژگی‌های محلی از آن جهت که نیازی به اطلاعات خاصی از قبیل شکل و یا محل قرارگیری شخص در تصویر را ندارند، برای بازشناسایی خودکار فعالیت‌ها بسیار مورد توجه قرار گرفته‌اند. این روش‌ها با استفاده از تشخیص نقاط ویژگی و سپس توصیف آن‌ها به‌صورت مکانی-

زمانی[10] عمل می‌کنند. یکی از اولین کارهایی که مبتنی بر این روش بود، توسط لاپتو در سال ۲۰۰۵ معرفی گردید [۱۳]. آن‌ها روشی بنام هریس[11] سه‌بعدی را که توسعه داده شده روش گوشه‌یابی هریس دوبعدی [۱۴] بود، جهت تشخیص نقاط ویژگی مورد نیاز برای توصیف حرکات، معرفی کردند. در تحقیقی دیگر، استفاده از ویژگی‌های محلی گابور در بُعد زمان-مکان، توسط دلار و همکارانش معرفی شد [۱۵]. برای کدگذاری ویژگی‌های استخراج شده، آن‌ها از روش کیسه واژه‌ها[12] استفاده کردند. اویکونوموپولوس و همکارانش [۱۶] با استفاده از مفهوم آنتروپی و روش استخراج ویژگی SRD[13] الگوریتم جدیدی برای استخراج ویژگی‌های مکانی-زمانی حرکات پیشنهاد کردند [۱۷]. ویلمز [۱۸] روش گوشه‌یابی هسین[14] در بعد زمان-مکان را معرفی کرد که بعداً به‌صورت هسین سه‌بعدی برای بازشناسایی فعالیت‌های انسان به‌کار گرفته شد [۱۹]. در تحقیقی دیگر، یک روش نمونه‌برداری تراکمی از روی مجموعه‌ای از نقاط خاصِ تعیین شده در توالی فریم‌ها، توسط وانگ و همکارانش پیشنهاد شد [۲۰]. نتایج مقایسه این روش با هریس سه‌بعدی، گابور و هسین سه‌بعدی، نشان داده است که روش‌های نمونه‌برداری تراکمی و هریس سه‌بعدی کارایی بهتری نسبت به دو روش دیگر دارند.

استفاده از مسیر[15] حرکت نقاط ویژگی در توالی فریم‌ها نیز جز روش‌های محبوب در بازشناسایی فعالیت بشمار می‌رود. یکی از روش‌های شناخته شده ردیابی نقاط ویژگی در توالی فریم‌ها، الگوریتم KLT[16] [۲۱] است. در [۲۲]، مسینگ و همکارانش استفاده از هریس سه‌بعدی و الگوریتم KLT را برای بازشناسایی فعالیت پیشنهاد دادند. در پژوهشی دیگر در [۲۳]، کانیچه و برموند از روش گوشه‌یابی شی-توماسی[17] [۲۴] برای شناسایی نقاط ویژگی استفاده کردند و برای ردیابی آن نقاط از KLT بهره بردند. همچنین آن‌ها کارایی روش گوشه‌یابی FAST[18] [۲۵] در شناسایی نقاط ویژگی و ردیابی آن نقاط براساس ویژگی‌های HOG[19] [۲۶] را مورد بررسی قرار دادند و به این نتیجه رسیدند که این روش حساسیت کمتری نسبت به روش KLT در برابر نویز دارد. در تحقیقی دیگر، استفاده از ترکیب KLT و SIFT[20] [۲۷] به عنوان روش درون‌یابی برای ایجاد مسیر حرکت نقاط ویژگی پیشنهاد شد. محققین این کار با نمونه‌برداری تراکمی از مسیرهای ایجاد شده و توصیف آن‌ها به روش HOG، HOF[21] [۲۸] و MBH[22] [۲۹] اقدام به بازشناسایی فعالیت‌ها کردند. با این‌که این روش تعداد قابل توجهی مسیر حرکت ایجاد می‌کند، اما با توجه به نتایج به‌دست آمده، کارایی این روش بهتر از روش‌های KLT و SIFT گزارش شده است.

توضیح این‌که، روش استخراج ویژگی HOF شباهت زیادی به روش HOG دارد با این تفاوت که محاسبه HOG روی شبکه‌ی متراکمی از سلول‌ها در یک تصویر انجام می‌گیرد در حالی‌که محاسبه HOF روی شبکه‌ی متراکمی از شار نوری (جهت گرادیان‌های تصاویر پشت سر هم) انجام می‌شود. در روش MBH، میدان شار نوری به مولفه-های افقی و عمودی آن تفکیک می‌شود. سپس مشتقات مکانی به طور جداگانه برای هریک از مولفه‌ها محاسبه شده و هیستوگرام جهت گرادیان‌ها (شبیه به روش HOG) محاسبه می‌شود.

پنگ و همکارانش ترکیب مناسبی از ویژگی‌های تصویری را معرفی کردند که اطلاعات غنی در مورد انواع حرکات انسان را در بر داشتند [۳۰]. آن‌ها از ترکیب HOF و HOG برای توصیف نقاط ویژگی استفاده کرده و سپس با استفاده از روش VQ[23]، ابعاد ویژگی‌های به‌دست آمده را کاهش دادند. سپس با به‌کارگیری روش کدگذاری فیشر [۳۱]، اقدام به کدگذاری ویژگی‌ها کرده و با استفاده از ماشین بردار پشتیبان، مدل یادگیر را ایجاد کردند.

در [۳۲]، با توجه به این نکته که استفاده از روش‌های استخراج ویژگی مبتنی بر عملگرهای تفاضلی باعث خواهد شد که حرکت‌های ریز و جزیی تشکیل دهنده‌ی یک فعالیت نادیده گرفته شوند، لان و همکارانش روشی بنام MIFS را پیشنهاد داده‌اند که هدف آن تعدیل تاثیرات روش‌های استخراج ویژگی مبتنی بر هرم گاووسی بر روی جزییات تصویر است. آن‌ها ترکیب ویژگی‌های خط سیر، HOG، HOF، MBHx و MBHy را برای توصیف ویژگی‌ها و روش PCA را برای کاهش ابعاد ویژگی‌ها پیشنهاد کردند. همچنین آن‌ها برای کدگذاری و ساخت مدل به ترتیب از روش کدگذاری فیشر و الگوریتم طبقه‌بندی کننده ماشین بردار پشتیبان استفاده کردند.

در [۳۳]، باقری و همکارانش به بررسی این موضوع پرداختند که استفاده از طبقه‌بندهای ساده و ویژگی‌های نه‌چندان پیچیده، ممکن است راه‌گشای چالش‌های موجود در بازشناسایی فعالیت‌های انسانی باشد. از این‌رو آن‌ها روشی را پیشنهاد دادند که در آن مجموعه‌ای از مدل‌های ساده‌ی ساخته شده با استفاده از ویژگی‌های متفاوت، عمل بازشناسایی را انجام می‌دهند. به همین منظور، یک استراتژی ترکیبی بر اساس نظریه دمپستر-شافر[24] پیشنهاد کردند که می‌تواند به طور موثری از مدل‌های پایه‌ای متنوعی که با منابع مختلف آموزشی ساخته شده‌اند، بهره ببرد.

در برخی از پژوهش‌های اخیر [۳۴, ۳۵]، استفاده از روش‌های مبتنی بر یادگیری عمیق در حوزه بازشناسایی فعالیت انسان نیز مورد توجه قرار گرفته است. در اکثر کاربردهای مبتنی بر یادگیری عمیق، فاز استخراج ویژگی‌ها به‌صورت خودکار انجام می‌گیرد و نیازی به مهندسی و یا ارایه روشی مشخص برای شناسایی و استخراج

ویژگی‌های مورد نظر نیست. البته طراحی مدل‌های مبتنی بر یادگیری عمیق، نیاز مبرم به مجموعه وسیعی از داده‌های آموزشی دارد. این مسئله می‌تواند در زمینه تشخیص فعالیت‌های انسان که تنوع زیادی هم دارند، به یک چالش بسیار جدی تبدیل شود. همچنین، برای ساخت چنین مدل‌هایی مدت زمان پردازش بسیار طولانی مورد نیاز است.

همچنان که پیش‌تر نیز اشاره شد، یکی از چالش‌های پیش روی سامانه‌های بازشناسایی فعالیت‌های انسان، تغییرات شدت نور حاصل از تغییرات حرکتی اندام و دوربین در توالی فریم‌ها است. در اثر این تغییرات محیطی، اطلاعات بافت نقاط ویژگی به شدت دست‌خوش تغییرات مخرب می‌گردد. از این‌رو لازم است علاوه بر اطلاعات ظاهری و حرکتی، اطلاعات بافتی این نقاط نیز کدگذاری و توصیف شده و سپس مورد بهره‌برداری قرار بگیرند. بر همین اساس، در این پژوهش افزایش کارایی سامانه بازشناسایی فعالیت انسان با به‌کارگیری یک توصیف‌گر بافت قوی و ترکیب آن با یک توصیف‌گر ظاهری است به‌عنوان هدف اصلی دنبال می‌شود.

۳- روش پیشنهادی

دیاگرام شکل ۲ روند راهکار پیشنهادی برای بازشناسایی فعالیت‌های انسان در ویدیو را نشان می‌دهد. در روش پیشنهادی ابتدا با استفاده از یک روش استخراج ویژگی، ویژگی‌های مکانی-زمانی در ویدیوهای آموزشی استخراج می‌شوند. سپس با اعمال یک روش کدگذاری روی هر یک از این ویژگی‌ها، بردار ویژگی نهایی برای بازنمایی حرکت‌های موجود در ویدیو ایجاد می‌شود. در انتها، بردارهای آموزشی حاصل برای آموزش و ساخت یک مدل طبقه‌بندی کننده استفاده می‌شوند. استخراج این مدل براساس یک رویکرد آبشاری از ماشین‌های بردار پشتیبان انجام می‌شود. مدل به‌دست آمده قادر خواهد بود فعالیت-های انسان را در ویدیوهای آزمایشی بازشناسایی کند. در ادامه، جزئیات مربوط به هر یک از این مراحل در زیربخش‌های جداگانه ارایه می‌شود.

۳-۱- استخراج ویژگی‌های مکانی-زمانی محلی

در رسانه‌ی ویدئو، علاوه بر بُعد مکان، بُعد زمان نیز وجود دارد که در اثر توالی فریم‌ها و ارتباط معنایی در این توالی شکل می‌گیرد. بنابراین، علاوه بر استخراج ویژگی‌های مکانی از هر فریم مستقل، ویژگی‌های زمانی سودمندی نیز در توالی فریم‌ها قابل استخراج است. در روش پیشنهادی، برای استخراج نقاط ویژگی در توالی فریم‌ها از الگوریتم هَریس سه‌بعدی [۱۳] استفاده می‌شود.

روش گوشه‌یابی هریس (هریس دوبُعدی)، نقاطی از تصویر را تشخیص می‌دهد که تغییرات روشنایی در آن نقاط نسبت به همسایگان‌شان قابل توجه است. اما در روش گوشه‌یابی هریسِ سه‌بعدی، فقط آن دسته از نقاطی که در توالی تصاویر خاصیت گوشه‌بودن را دارند، تشخیص داده می‌شوند. این الگوریتم در ابتدا ماتریس مُمان درجه دوم مکانی-زمانی هر فریم را محاسبه کرده و سپس مقادیر ویژه‌ی[۲۵] تاثیرگذار در توالی ماتریس‌های ایجاد شده را به‌دست می‌آورد. در انتها، نقاط نهایی مکانی-زمانی هر فریم بر اساس بیشنه محلی مثبت در بین نقاط موجود انتخاب می‌شود. در شکل ۳، مثالی از اجرای الگوریتم هریس سه‌بعدی بر روی نمونه‌ای از فریم‌های متوالی نشان داده شده است.

در روش پیشنهادی، برای اولین بار ویژگی FREAK[۲۶] [۳۶] و ترکیب آن با HOG به ترتیب برای توصیف اطلاعات بافتی و توصیف اطلاعات ظاهری و حرکتی به منظور توصیف هر یک از نقاط ویژگی استخراج شده با استفاده از هَریس سه‌بعدی به‌کار گرفته می‌شوند. در ادامه، توضیح کوتاهی در مورد توصیف‌گرهای HOG و FREAK ارایه می‌شود. جزییات استفاده و پیاده‌سازی این الگوریتم‌ها در بخش ۲-۴ ارایه شده است.

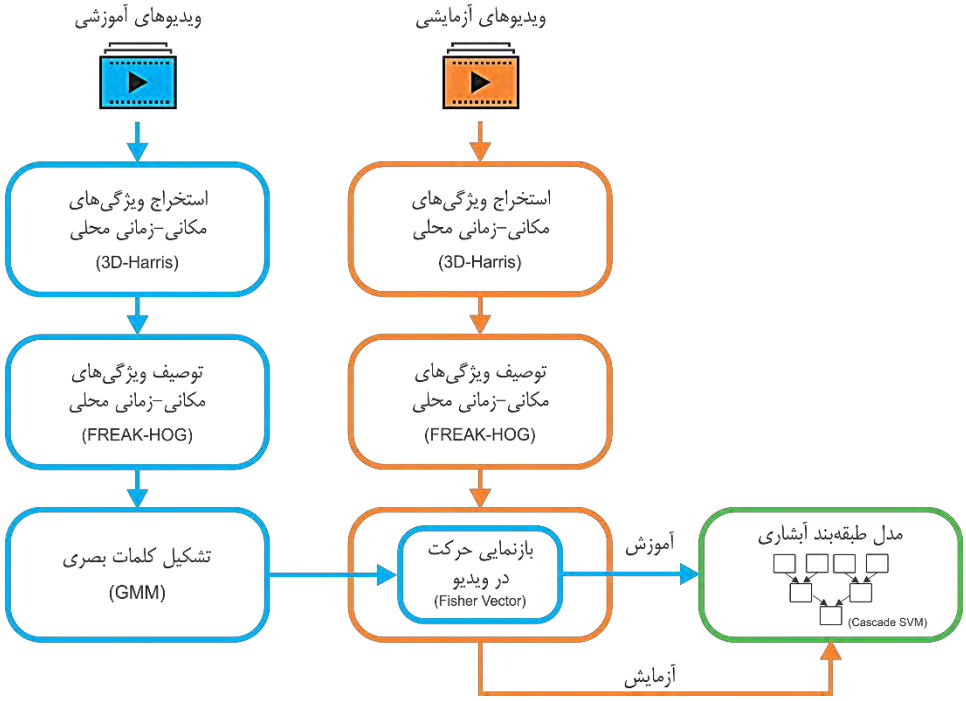
۳-۲- توصیف‌گر HOG

در تحقیقات متعددی نشان داده شده است که توصیف‌گر HOG از کارایی بسیار مطلوبی در تشخیص اشیا برخوردار است. ایده اصلی و در عین حال ساده در توصیف‌گر HOG عبارت است از این که اطلاعات محلی اغلب توسط توزیع شدت گرادیان یا جهت لبه‌های یک شکل حتی بدون اطلاع دقیق از محل حضور لبه‌ها به خوبی قابل توصیف است. برای جداسازی اطلاعات ساختاری یک نقطه، نقاط همسایه آن نقطه به شبکه‌ای از سلول‌های مکانی-زمانی تقسیم می‌شود. سپس برای هر یک از سلول‌ها، هیستوگرام جهت لبه‌ها به‌صورت مجزا محاسبه شده و در نهایت، همه‌ی هیستوگرام‌ها با هم ترکیب و بردار نهایی به‌دست می‌آید. به این ترتیب، HOG اطلاعات نمای ظاهری و جهت لبه‌های موجود در اطراف نقاط ویژگی را می-تواند کدگذاری کند.

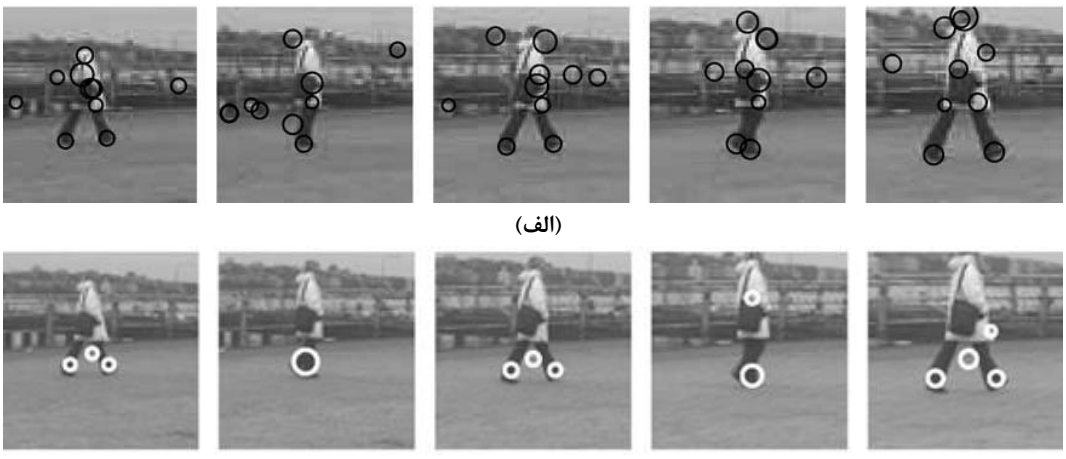
۳-۳- توصیف‌گر FREAK

محققین علوم اعصاب بر این باورند که شبکیه چشم انسان اطلاعات مربوط به جزییات تصویر را با استفاده از اختلاف گاووسی نواحی همسایه محاسبه و کدگذاری می‌نماید [۳۸]. از این‌رو می‌توان گفت که توپولوژی و چیدمان سلول‌های حسگر[۲۷] موجود در شبکیه نقش بسیار مهمی در ایجاد چنین قابلیتی ایفا می‌کنند. سلول‌های شبکیه

به چهار ناحیه تقسیم شده‌اند (شکل ۴): ناحیه‌ی مرکزی[۲۸]، ناحیه حفره‌ای[۲۹]، ناحیه پس-حفره‌ای[۳۰] و ناحیه محاطی[۳۱]. هر ناحیه کارکرد خاص خود را در کشف و بازشناسایی اشیا ایفا می‌نماید؛ تشکیل قسمت‌هایی با وضوح بالا در ناحیه‌ی حفره‌ای اتفاق می‌افتد در حالی‌که بخش‌های با وضوح کمتر در قسمت محاطی شبکیه شکل می‌گیرد.



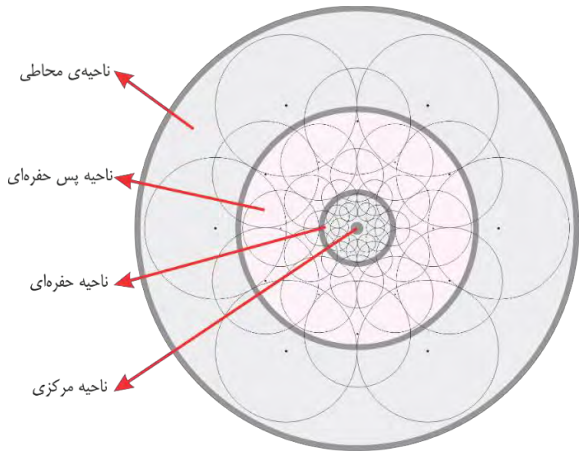
شکل ۲: دیاگرام روش پیشنهادی برای بازشناسایی خودکار فعالیت‌های انسان در ویدیو. مراحل مشخص شده به رنگ آبی و نارنجی به ترتیب فازهای آموزش و آزمایش سامانه را نشان می‌دهند.



(الف)

(ب)

شکل ۳: اجرای الگوریتم تشخیص نقاط ویژگی هریس. الف) نقاط ویژگی تشخیص داده شده در یک فریم توسط هریس دو بعدی، ب) نقاط ویژگی تشخیص داده شده در توالی فریم‌ها توسط هریس سه‌بعدی. همانطور که ملاحظه می‌شود، هریس سه‌بعدی نسبت به نقاط متحرک در توالی تصاویر حساس بوده در حالی‌که هریس دو بعدی در کل صفحه اقدام به شناسایی نقاط ویژگی کرده است [۳۷].



شکل ۴: نحوه‌ی نمونه‌برداری در روش FREAK الهام گرفته شده از شبکیه چشم انسان [۳۶].

برای ساختن یک رشته بیتی (دودویی) که بتواند بافت تشکیل دهنده‌ی اطراف یک نقطه خاص (نقطه ویژگی) را توصیف کند، باید همه‌ی نقاط همسایگی آن نقطه بررسی شوند به‌طوری‌که، برای هر جفت (P_1،P_2)، اگر شدت نور در نقطه P_1 بیشتر از شدت نور در نقطه P_2 باشد، به رشته دودویی عدد یک و در غیر اینصورت عدد صفر افزوده می‌شود. لذا طول رشته (بردار ویژگی) و کارایی آن بستگی به نحوه‌ی انتخاب همسایه‌ها دارد. عملگرهای توصیف بافت دودویی از روش‌های متعددی برای مقایسه شدت نور پیکسل‌ها برای ایجاد بردار ویژگی استفاده می‌کنند. بعنوان مثال روش ORB [۳۹] با استفاده از انتخاب تصادفی جفت پیکسل‌ها این کار را انجام می‌دهد و یا روش DAISY [۴۰] از یک الگوی دایره‌ای شکل استفاده می‌کند.

در روش FREAK استفاده از روش نمونه‌برداری شبیه به آنچه که در شبکیه چشم وجود دارد، پیشنهاد شده است. نحوه‌ی نمونه‌برداری از یک ناحیه در شکل ۴ نشان داده شده است. در این شکل، هر دایره نماینده محدوده‌ی حسی شدت نور، و شعاع آن نشان دهنده‌ی میزان انحراف معیار هسته گاووسی هست که در ناحیه مورد نظر برای به‌دست آوردن میانگین شدت نور اعمال می‌شود. همانطور که ملاحظه می‌شود، توزیع مکانی سلول‌های حسگر متناسب با فاصله‌شان نسبت به مرکز شبکیه به‌صورت نمایی کاهش پیدا می‌کند و تراکم نقاط در اطراف مرکز نسبت به حاشیه‌ی آن زیاد است.

توصیف‌گر F با آستانه‌گیری از تفاضل نواحی حسگرها (دایره‌ها) و هسته‌های گاووسی متناظرشان، یک ناحیه را به شکل یک رشته‌ی دودویی به‌صورت زیر توصیف می‌کند:

$$F = \sum_{0 \le a < N} 2^a T(P_a), \qquad (۱)$$

در رابطه ۱، P_a جفت ناحیه‌های دریافت‌کننده و N بزرگی توصیف‌گر است. T نیز از رابطه‌ی زیر محاسبه می‌شود:

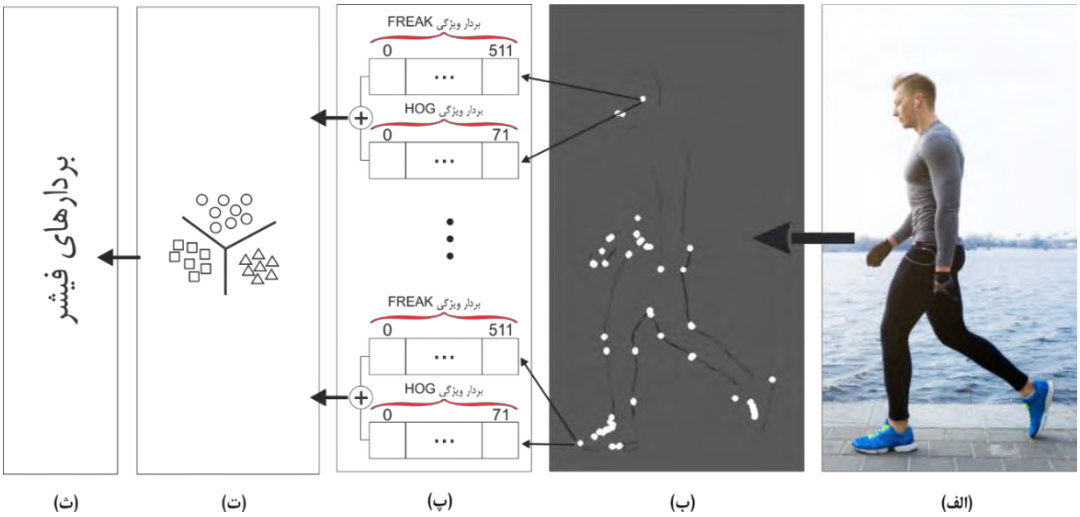
$$T(P_a) = \begin{cases} 1 & \text{if}(I(P_a^{r_1}) - I(P_a^{r_2}) > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (۲)$$

با افزایش تعداد نواحی دریافت کننده، می‌توان به تعداد بیشتری از جفت نواحی مقایسه شونده دست یافت، که این کار در نهایت باعث افزایش طول بردار توصیف کننده خواهد شد. با وجود این، گفته می‌شود که بیشتر نواحی تولید شده برای مقایسه مفید نیستند و از این‌رو باید با استفاده از راهکاری سعی در انتخاب آن دسته از نواحی کرد که تفاضل آن‌ها بتواند اطلاعات مفیدتری را تولید نماید [۳۶]. برای این منظور، در روش پیشنهادی از الگوریتم ارائه شده در [۳۹] استفاده می‌شود. در این الگوریتم، ماتریسی از تفاضل تمامی ترکیبات ممکن از نواحیِ در نظر گرفته شده ساخته می‌شود. سپس میانگین هر ستون برای پیدا کردن بیشترین انحراف معیار محاسبه می‌شود. هر چقدر انحراف معیار ستون‌ها بیشتر باشد، نشان دهنده‌ی موثر بودن نواحی شرکت کننده در تفاضل است. در نهایت، ستون‌ها بر اساس میزان انحراف به‌دست آمده‌شان، به‌صورت نزولی مرتب شده و نواحی مرتبط با هر ستون بعنوان نواحی برگزیده به‌کار برده می‌شوند.

۴-۳- کدگذاری ویژگی‌های محلی

پس از این‌که ویژگی‌های مکانی-زمانی استخراج شدند، لازم است که از آن‌ها برای توصیف حرکت‌های موجود در ویدیو استفاده شود. یکی از روش‌های شناخته شده برای اینکار استفاده از کیسه ویژگی‌ها[۳۲] است. این روش جز روش‌های رایج و کارآمد در زمینه‌هایی نظیر پردازش زبان‌های طبیعی، بازیابی اطلاعات و همچنین بینایی ماشین محسوب می‌شود، که برای اولین بار جهت بازیابی اسناد بر اساس متن با عنوان کیسه واژه‌ها[۳۳] مورد استفاده قرار گرفت [۴۱].

روش کیسه واژه‌ها با استفاده از اطلاعات آماری نقاط ویژگی، که به‌صورت محلی استخراج شده‌اند، هیستوگرامی از رخ داد نقاط را در توالی فریم‌های ویدیو ایجاد می‌کند. بدین صورت که در ابتدا با به‌کارگیری یک الگوریتم بدون ناظر نظیر k-Means، خوشه‌هایی (واژه‌هایی) با استفاده از نقاط ویژگی جمع‌آوری شده از ویدیوهای آموزشی ایجاد می‌کند. سپس هر یک از ویژگی‌های محلی به یکی از واژه‌ها نسبت داده می‌شود و به این ترتیب، هیستوگرام رخدادها تشکیل می‌شود. همچنین برای افزایش کارایی، یک عمل



شکل ۵: نحوه استخراج بردارهای فیشر. الف) فریمی از یک ویدیو، ب) استخراج نقاط ویژگی هریس سه‌بعدی، پ) توصیف نقاط ویژگی بوسیله FREAK و HOG، ت) تشکیل کیسه واژه‌ها بوسیله الگوریتم GMM، ث) ایجاد بردارهای فیشر

پس‌پردازشی به منظور نرمال‌سازی اندازه‌ی ویدیوها و نقاط ویژگی محلی استخراج شده، اعمال می‌شود.

از روش‌های رایج برای کدگذاری ویژگی‌ها، کدگذاری فیشر [۳۱] است که سعی در افزایش کارایی کیسه واژه‌ها بوسیله جایگزینی نحوه استفاده از اطلاعات آماری دارد. در این روشِ کدگذاری، بجای استفاده از هیستوگرام رخدادهای نقاط ویژگی محلی، از تفاوت آن ویژگی‌ها و واژه‌ها استفاده می‌شود.

کدگذاری کرنل[۳۴] [۴۲] نیز نوعی روش کدگذاری مبتنی بر کیسه واژه‌هاست، به‌طوری‌که ویژگی‌های محلی براساس یک معیار عضویت به واژه‌ها منتسب می‌شوند. از این‌رو، هر ویژگی به‌جای این‌که فقط عضو یک واژه شود، عضو چندین واژه شده و در نهایت، تبدیل به ترکیبی وزن‌دار از واژه‌ها می‌شود.

یکی دیگر از روش‌های مبتنی بر کیسه واژه‌ها، روش کدگذاری خطیِ محدود شده‌ی محلی[۳۵] [۴۳] است. در این روش، هر ویژگی محلی به دستگاه مختصات محلی خودش نگاشت می‌شود. سپس، مختصات به‌دست آمده به روش گزینش حداکثرها[۳۶] ، باهم ترکیب شده تا بازنمایی نهایی را تولید نمایند. در نهایت، ویژگی‌ها به زیر فضای خطی-محلی افراز شده توسط چندین واژه، نگاشت می‌شوند.

کدگذاری فوقِ بردار[۳۷] [۴۴] نوعی از روش کدگذاری فیشر است. در این روش ابتدا با استفاده از روش k-Means، کیسه واژه‌ها ایجاد می‌شود. سپس، بر اساس تفاضل مرتبه‌ی اول بین ویژگی‌ها و واژگان و همچنین براساس ویژگی‌های قرار گرفته در مرکز ثقل هر واژه، عمل کدگذاری انجام می‌گیرد.

مقایسه روش‌های کدگذاریِ فیشر، کرنل، خطیِ محدود شده‌ی محلی و فوقِ بردار، نشان داده است که روش کدگذاری فیشر از کارایی بهتری از نظر دقت بازشناسایی برخوردار است [۴۵]. بر همین اساس، در روش پیشنهادی از این روش برای بازنمایی نقاط ویژگی محلی در ویدیو استفاده می‌شود و برای ایجاد کیسه واژه‌ها (خوشه‌بندی ویژگی‌های استخراج شده) از مدل ترکیبی گاووسی[۳۸] (GMM) بهره برده می‌شود.

با توجه به این‌که در روش پیشنهادی از دو روش HOG و FREAK برای توصیف نقاط تشخیص داده شده توسط هریس سه‌بعدی استفاده کرده‌ایم، ابتدا بردارهای به‌دست آمده از دو روش HOG و FREAK به هم الحاق شده و بردار ویژگی واحدی برای ایجاد واژه‌ها توسط الگوریتم GMM به‌دست می‌آید و سپس براساس این واژه‌ها، بردار فیشر حاصل می‌شود. شکل ۵ نحوه‌ی ایجاد بردارهای فیشر را نشان می‌دهد.

۵-۳- ساخت مدل طبقه‌بندی کننده

پس از این‌که حرکات ناشی از فعالیت‌های مختلف انسانی استخراج و بازنمایی گردید، با استفاده از یک الگوریتم یادگیری اقدام به ساخت مدلی برای طبقه‌بندی حرکات و بازشناسایی فعالیت می‌شود. ازجمله الگوریتم‌های یادگیری که می‌توان برای این منظور استفاده کرد، عبارتند از الگوریتم نزدیکترین همسایه [۱۰]، شبکه‌های عصبی مصنوعی [۴۶]، ماشین بردار پشتیبان [۱۳] و مدل مارکوف

مخفی [۴۷]. با این‌که هر یک از این الگوریتم‌ها مزایا و معایبی دارند، بر اساس تحقیقی که در [۴۸] انجام شده است، به‌طور کلی روش ماشین بردار پشتیبان در کاربردهایی نظیر سامانه مورد نظر، کارایی بهتری نسبت به سایر روش‌ها از خود نشان داده است.

ایده ماشین بردار پشتیبان در سال ۱۹۹۵ میلادی توسط وپنیک [۴۹] معرفی شد. این ایده مبتنی بر نظریه استفاده از اطلاعات آماری برای ساخت مدل (فرضیه) بوده و اساس کار آن بر حل یک مسئله برنامه‌ریزی درجه دوم استوار است. برنامه‌ریزی درجه دوم نوعی مسئله بهینه‌سازی در ریاضیات است که سعی در بهینه کردن تابع هدفی از نوع درجه دوم با محدودیت‌های خطی را دارد. در طی این فرآیند، نقاطی از داده‌های آموزشی تحت عنوان بردارهای پشتیبان از سایر نقاط موجود شناسایی می‌شوند. با به‌کارگیری این نقاط، فرضیه‌ای به شکل خط، صفحه و یا اَبَر صفحه جهت طبقه‌بندی داده‌ها (شواهد) شکل می‌گیرد. این الگوریتم در کاربردهایی که تعداد داده‌های آموزشی آن کم یا متوسط است، به‌طور گسترده‌ای مورد استفاده قرار گرفته و نتایج بسیار خوبی حاصل شده است. اما زمانیکه حجم داده‌ها در کاربردهایی نظیر بازشناسایی فعالیت انسان خیلی زیاد باشد، استفاده از این الگوریتم هم در زمان ساخت مدل و هم در زمان استفاده از آن برای بازشناسایی، بسیار زمان‌بر می‌شود [۵۰]. برای مقابله با این چالش، در روش پیشنهادی یک ساختار آبشاری از ماشین‌های بردار پشتیبان مورد استفاده قرار می‌گیرد. در ادامه توضیح این روش ارایه می‌شود.

۶-۳- ماشین بردار پشتیبان آبشاری

روش ماشین بردار پشتیبان آبشاری از یک فرآیند مرحله‌ای برخوردار است. در هر مرحله نتایج (بردارهای پشتیبان) ماشین‌های مرحله‌ی قبلی برای انجام عمل طبقه‌بندی به‌کار برده می‌شود. ایده اصلی این روش در کاهش مکرر فضای داده‌های آموزشی تا رسیدن به مرحله‌ای نهایی در ساختار آبشاری است. این کار با شناسایی بردارهای پشتیبان در هر مرحله و دور ریختن مابقی داده‌ها در مرحله‌ی بعدی انجام می‌گیرد. مراحل زیر این فرآیند را توصیف می‌نماید [۵۱]:

۱) تقسیم داده‌های آموزشی به k دسته مستقل و ترجیحاً هم اندازه؛

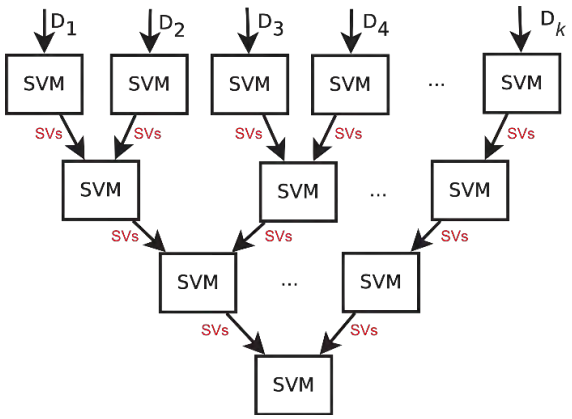
۲) آموزش ماشین‌های بردار پشتیبان مستقل برای هر یک از k زیر مجموعه؛

۳) ادغام بردارهای پشتیبان شناسایی شده توسط ماشین‌های بردار پشتیبان همجوار (برای مثال: دو ماشین بردار پشتیبان همجوار)؛

۴) تکرار مراحل ۲ و ۳ تا رسیدن به آخرین مرحله ما قبل آخر در ساختار آبشاری؛

۵) آموزش ماشین بردار پشتیبان نهایی با استفاده از بردارهای پشتیبان به‌دست آمده از آخرین مرحله.

شکل ۶ شمایی از این روش آبشاری را نشان می‌دهد. جزییات پیاده‌سازی این روش نیز در بخش ۴-۲ ارایه شده است.



شکل ۶: مدل ماشین بردار پشتیبان آبشاری. بردارهای پشتیبان هر مجموعه داده توسط ماشین مربوطه شناسایی و برای آموزش ماشین‌های مرحله بعد استفاده می‌شوند.

۴- نتایج آزمایش‌ها

در این بخش نتایج آزمایش‌های انجام گرفته جهت ارزیابی کارایی روش پیشنهادی ارائه می‌شود. ابتدا پایگاه‌داده‌های مورد استفاده برای انجام آزمایش‌ها، معرفی می‌شوند. سپس نتایج به‌دست آمده از آزمون‌ها بر روی پایگاه‌داده‌ها، ارائه و نهایتا نتایج مقایسه‌های صورت گرفته با سایر روش‌های موجود مورد بحث و بررسی قرار می‌گیرند.

۴-۱- پایگاه داده‌ها

پایگاه داده‌های مورد استفاده جهت ارزیابی راهکار پیشنهادی در این پژوهش، پایگاه‌داده‌های UCF101 [۵۲] و HMDB51 [۵۳] هستند.

UCF101: این پایگاه‌داده یکی از بزرگترین، متنوع‌ترین و چالش برانگیزترین مجموعه‌ی داده‌ی ویدیویی است که در دانشگاه فلوریدای مرکزی جمع‌آوری گردیده و تاکنون الگوریتم‌های بسیاری با استفاده از آن مورد ارزیابی قرار گرفته‌اند [۶]. این پایگاه‌داده در سال ۲۰۱۲ ایجاد شده و شامل ویدیوهایی از ۱۰۱ گروه فعالیت انسانی با تعداد فریم‌های متفاوت است. هر کدام از ویدیوها نشان دهنده‌ی یک عمل خاص از انسان (مانند بازی بیلیارد، حرکات نمایشی روی یخ، دوچرخه سواری، نواختن پیانو، بازی پینگ‌پنگ و

غیره) در محیط‌های مختلف است. همچنین، ویدیوهای این پایگاه‌داده شامل صحنه‌های واقعی است که تنوع زیادی در حرکت دوربین، زاویه دید، مقیاس شی، پس‌زمینه متغیر، ظاهر فرد، ژست فرد، اندازه تصویر فرد، درهم برهمی[۳۹] و شرایط روشنایی مختلف در ویدئو است. از هر گروه فعالیت خاص، چندین ویدیو ثبت شده است. جدول شماره ۱ جزییات تعداد ویدیوهای پایگاه‌داده‌ی UCF101 را در هر گروه نشان میدهد.

HMDB51: این پایگاه داده که در سال ۲۰۱۱ در دانشگاه براون تهیه شده است، شامل ۵۱ فعالیت روزمره انسانی است. برای هر یک از فعالیت‌ها، حداقل ۱۰۱ ویدیو از منابع مختلف از جمله سایت یوتیوب و ویدیوهای گوگل با کیفیت‌های متنوع گردآوری شده است. برای سادگی ارزیابی، گردآوردندگان این پایگاه‌داده فعالیت‌ها را در پنج گروه حرکات ساده صورت (خندیدن، جویدن، حرف‌زدن و ...)، حرکات صورت درگیر با یک شی (سیگار کشیدن، خوردن، آشامیدن و ...)، فعالیت‌های مرتبط با اشیا (شانه‌زدن، بازی گلف، ضربه زدن به توپ و ....)، فعالیت‌های عادی بدون دخالت اشیا (دویدن، دست‌زدن، نشستن و ...)، و فعالیت‌های تعاملی (دست‌دادن، روبوسی کردن، هل دادن و ...) تقسیم بندی کرده‌اند. جدول شماره ۲ جزییات تعداد ویدیوهای هر گروه را نشان می‌دهد.

ویدئوهای موجود در هر دو پایگاه‌داده به دو دسته‌ی ویدئوهای آموزشی و ویدئوهای آزمایشی تقسیم می‌شوند. نحوه این تقسیم‌بندی در جدول ۱ و جدول ۲ نشان داده شده است. ویدئوهای آموزشی جهت ایجاد مدل استفاده شده و ویدئوهای آزمایشی جهت ارزیابی و اطمینان از صحت عمل‌کرد روش پیشنهادی مورد استفاده قرار می‌گیرند. در تحقیقات انجام شده، نسبت این تقسیم‌بندی عموماً بدین صورت است که حدود ۶۰ تا ۸۰ درصد از ویدئوهای پایگاه‌داده برای آموزش و ۲۰ تا ۴۰ درصد آن برای آزمایش و ارزیابی روش‌ها اختصاص می‌یابد. علت انتخاب این نوع تقسیم‌بندی و تعداد بیشتر ویدئوهای آموزشی نسبت به ویدئوهای آزمایشی، اطمینان از مدل‌سازی دقیق توزیع ویژگی‌ها در ویدئوهای آموزشی است. نحوه‌ی انتخاب ویدئوها برای این دو دسته نیز معمولاً به‌صورت تصادفی است تا احتمال هر گونه همبستگی بین داده‌ها و تاثیرگذاری آن‌ها در نتایج نهایی به حداقل برسد.

توضیح این‌که مبنای این تقسیم‌بندی بر اساس پیشنهاد تهیه کنندگان پایگاه‌داده‌های مذکور است و در تمامی ارزیابی‌های انجام گرفته در تحقیقات مختلفی که از این پایگاه‌داده‌ها استفاده کرده‌اند، این تقسیم‌بندی رعایت شده است.

۲-۴- پیاده‌سازی

در این پژوهش، پیاده‌سازی الگوریتم پیشنهادی با زبان برنامه‌نویسی C++ صورت گرفته و بدنه‌ی اصلی راهکار پیشنهادی از جمله پردازش فریم‌ها با استفاده از توابع کتابخانه‌ای OpenCV انجام شده است. در برخی موارد از متدهای پیاده‌سازی شده توسط محققین مقالات مربوطه و یا سایر کتابخانه‌های نوشته شده به زبان C++ جهت جلوگیری از هرگونه خطا در پیاده‌سازی، بهره برده شده است. آزمایش‌ها با استفاده از یک کامپیوتر با CPU دو هسته‌ای ۳/۶ گیگا هرتز و ۴ گیگا بایت RAM انجام گرفته است. هر نقطه ویژگی استخراج شده در حوزه مکانی-زمانی توسط الگوریتم هریس سه‌بعدی، با روش‌های HOG و FREAK به ترتیب با بردارهایی بطول ۷۲ و ۵۱۲ توصیف می‌شود.

استخراج ویژگی‌های HOG با استفاده از توابع کتابخانه‌ی VLFeat[۴۰] انجام می‌شود. همانطور که در بخش ۳-۲ تشریح شد، HOG اطلاعات نمای ظاهری و جهت لبه‌های موجود در اطراف نقاط ویژگی را کدگذاری می‌کند. توابع کتابخانه‌ی VLFeat حاوی توابع کد منبع باز به زبان C است که اغلبِ الگوریتم‌های پردازش تصویر و بینایی ماشین به ویژه استخراج ویژگی‌های استاندارد را شامل می‌شوند؛ پیاده‌سازی بهینه ویژگی‌های HOG نیز جزء توابع این کتابخانه است. این توابع از آدرس http://www.vlfeat.org قابل دریافت است.

استخراج ویژگی‌های FREAK نیز بوسیله توابع پیاده‌سازی شده برای این ویژگی‌ها توسط نویسندگان مقاله مربوطه [۳۶] انجام می‌شود. این توابع از آدرس https://github.com/kikohs/freak قابل دریافت است.

توضیح این‌که، به علت حجم بالای ویدیوهای پایگاه داده‌های مورد استفاده، عمل استخراج ویژگی‌های مورد استفاده بسیار زمان‌بر بوده و حجم بسیار زیادی برای ذخیره‌سازی داده‌های استخراج شده نیاز است. از این‌رو، ابتدا تمامی ویژگی‌های مورد نیاز از ویدئوهای آموزشی این پایگاه‌داده‌ها استخراج شده و در قالب فایل‌هایی با فرمت متنی ذخیره می‌گردند. در مراحل بعدی پردازش‌ها، صرفا داده‌های ذخیره شده در فایل‌های متنی استفاده می‌شوند.

برای تشکیل کیسه واژه‌ها به روش خوشه‌بندی GMM و ایجاد بردارهای فیشر نیز از کتابخانه VLFeat استفاده شده است. برای پیاده‌سازی ماشین بردار پشتیبان آبشاری، از توابع توسعه داده‌شده‌ی نویسندگان مرجع [۵۱] استفاده شده است[۴۱]. همچنین بر اساس آزمایش‌های انجام گرفته، ملاحظه شد که بیشینه دقت برای طبقه‌بند آبشاری روی داده‌های ارزیابی زمانی است که $k = 18$ در نظر گرفته شود. توضیح این‌که هسته استفاده شده در هر یک از

جدول ۱: جزییات داده‌های آموزشی و آزمایشی پایگاه‌داده‌ی UCF101

	ورزشی	موسیقی	انسان-اشیا	ساده	انسان-انسان	کل	مجموع
آموزشی	۴۷۷۱	۱۰۲۷	۱۸۷۲	۱۳۷۰	۴۹۷	۹۵۳۷	۱۳۳۲۰
آزمایشی	۱۹۰۲	۴۰۱	۷۴۷	۵۴۰	۱۹۳	۳۷۸۳	

جدول ۲: جزییات داده‌های آموزشی و آزمایشی پایگاه‌داده‌ی HMDB51

	حرکات صورت-ساده	حرکات صورت-پیچیده	انسان-اشیا	ساده	انسان-انسان	کل	مجموع
آموزشی	۳۰۶	۲۵۴	۱۹۷۵	۱۳۸۸	۵۸۷	۴۵۱۰	۶۷۶۶
آزمایشی	۱۵۳	۱۲۷	۹۸۸	۶۹۴	۲۹۳	۲۲۵۵	

جدول ۳: مقایسه میانگین دقت بازشناسایی، مدت زمان ساخت مدل و میانگین زمان اجرای روش پیشنهادی با و بدون در نظر گرفتن ویژگی‌های FREAK و ماشین بردار پشتیبان آبشاری روی پایگاه‌داده‌ی UCF101

روش	میانگین دقت بازشناسایی (٪)						زمان ایجاد مدل (ساعت)	میانگین زمان اجرا (ثانیه)
	کل	ورزشی	موسیقی	انسان-اشیا	ساده	انسان-انسان		
روش پیشنهادی (HOG-FREAK + CSVM)	۸۸/۱	۹۲/۴	۸۵/۶	۸۲/۸	۸۸/۳	۹۱/۴	۲۳	۱۱/۳
روش پیشنهادی بدون FREAK (HOG + CSVM)	۷۶/۶	۷۶/۸	۷۳/۲	۷۶/۵	۷۴/۴	۸۲/۱	۱۹	۱۰/۵
روش پیشنهادی با MSVM (HOG-FREAK + MSVM)	۸۶/۹	۸۸/۳	۸۴/۲	۸۲/۶	۸۷/۵	۹۰/۳	۳۷	۱۲/۱

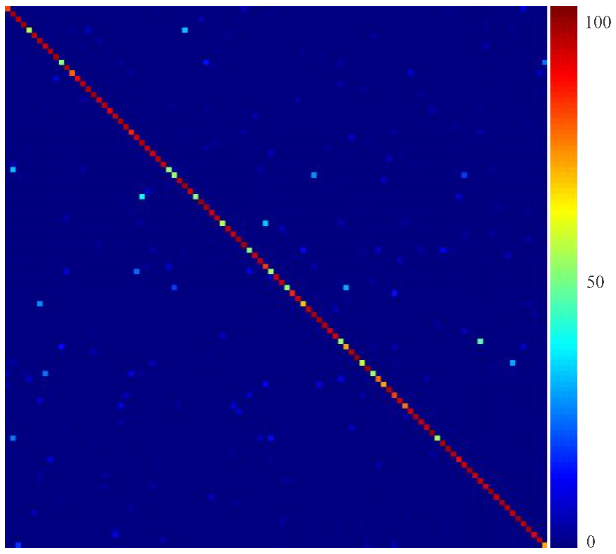
جدول ۴: مقایسه میانگین دقت بازشناسایی، مدت زمان ساخت مدل و میانگین زمان اجرای روش پیشنهادی با و بدون در نظر گرفتن ویژگی‌های FREAK و ماشین بردار پشتیبان آبشاری روی پایگاه‌داده‌ی HMDB51

روش	میانگین دقت بازشناسایی (٪)						زمان ایجاد مدل (ساعت)	میانگین زمان اجرا (ثانیه)
	کل	حرکات صورت-ساده	حرکات صورت-پیچیده	انسان-اشیا	ساده	انسان-انسان		
روش پیشنهادی (HOG-FREAK + CSVM)	۶۴/۷	۶۹/۱	۶۵/۴	۵۷/۲	۵۸/۷	۷۳/۶	۱۰	۷/۴
روش پیشنهادی بدون FREAK (HOG + CSVM)	۵۱/۲	۵۵/۱	۵۲/۳	۴۳/۵	۴۵/۸	۵۹/۵	۹	۶/۱
روش پیشنهادی با MSVM (HOG-FREAK + MSVM)	۶۲/۲	۶۵/۱	۶۴/۴	۵۲/۲	۵۸/۷	۷۰/۶	۱۸	۸/۱

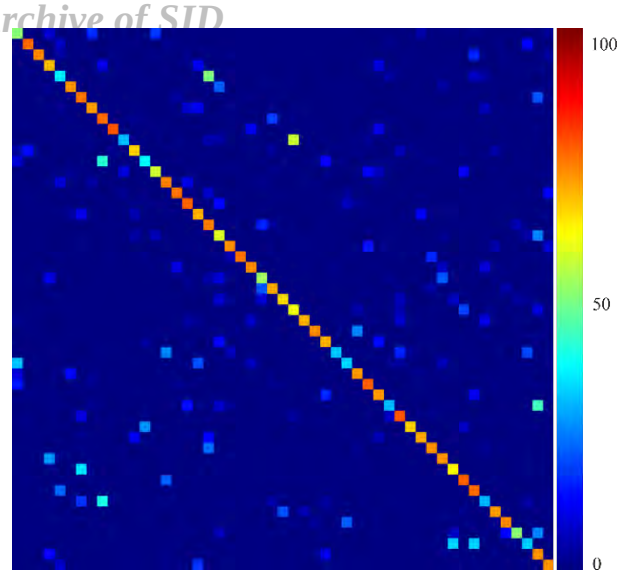
ماشین‌های مستقل در این ساختار، تابع پایه شعاعی[۴۲] است.

جدول‌های ۳ و ۴ به‌ترتیب نتایج ارزیابی کارایی روش پیشنهادی بر اساس درصد دقت تشخیص نوع فعالیت‌های مختلف در ویدیوهای آزمایشی پایگاه‌داده UCF101 و HMDB51 را نشان می‌دهند. در این جدول‌ها، روش پیشنهادی با و بدون در نظر گرفتن توصیف‌گر FREAK، که به مجموعه ویژگی‌های HOG برای توصیف بافت نقاط ویژگی اضافه شده‌اند، ارایه شده است. همچنین جهت ارزیابی میزان تاثیر استفاده از ماشین بردار پشتیبان آبشاری (CSVM) در سرعت و دقت تشخیص، نتایج روش پیشنهادی بدون در نظر گرفتن آن نیز گزارش شده است. برای ساخت مدل طبقه‌بندی کننده در حالت غیر آبشاری، از الگوریتم ماشین بردار پشتیبان چندکلاسه (MSVM[۴۳]) [۵۲] استفاده شده است.

در نتایج ارایه شده، زمان اجرا با ثبت زمان کل سپری شده (بر حسب ثانیه) برای بازشناسایی فعالیت در تمامی ویدیوهای آزمایشی و تقسیم آن به تعداد ویدیوها به‌دست آمده است. زمان ایجاد مدل نیز برابر با زمانی است که جهت ساخت مدل توسط روش CSVM و MSVM صرف می‌شود. همچنین نتایج به‌دست آمده به تفکیک گروه‌های از پیش تعریف شده فعالیت‌ها (ورزشی، نواختن آلت موسیقی، تعامل انسان با اشیا، فعالیت‌های ساده، تعامل انسان با انسان و ...) جهت مطالعه موردی کارایی روش پیشنهادی ارایه شده



شکل ۷: ماتریس اغتشاش حاصل از اِعمال روش پیشنهادی بر روی پایگاه داده UCF101



شکل ۸: ماتریس اغتشاش حاصل از اِعمال روش پیشنهادی بر روی پایگاه داده HMDB51

جدول ۵: مقایسه کارایی روش پیشنهادی با سایر روش‌ها

مرجع (سال)	روش	میانگین تشخیص (٪) روی UCF101	میانگین تشخیص (٪) روی HMDB51	متوسط تشخیص (٪) روی دو پایگاه‌داده
[۳۰] (۲۰۱۶)	Hybrid-BoW	۸۷/۹	۶۱/۱	۷۴/۵
[۳۲] (۲۰۱۵)	Multi-Skip Feat. Stacking	۸۹/۱	۶۵/۱	۷۷/۱
[۳۳] (۲۰۱۵)	Ensemble Learning	۷۵/۱	-	-
[۳۴] (۲۰۱۵)	LRCN- Weighted Average of RBG + Flow	۸۲/۹	-	-
[۳۵] (۲۰۱۶)	RBG + Opt Flow Networks	۹۲/۴	۶۲	۷۷/۲
روش پیشنهادی	(HOG-BSIF) + (GMM-FV) + CSVM	۸۸/۱	۶۴/۷	۷۶/۴

است. لازم به ذکر است که در تمامی آزمایش‌ها، از روش کدگذاری فیشر به همراه GMM استفاده شده است.

همانطورکه نتایج ارایه شده در جدول‌های ۳ و ۴ نشان می‌دهند، استفاده از ویژگی‌های توصیف‌گر بافت دودویی FREAK به همراه ویژگی‌های HOG به ترتیب باعث افزایش ٪۱۱/۵ و ٪۱۳/۴ کارایی در دقت بازشناسایی در پایگاه‌داده‌های UCF101 و HMDB51 شده است. چنین افزایش‌هایی در دقت، تنها با هزینه بالاسری به ترتیب ۰/۸ و ۱/۳ ثانیه‌ای برای استخراج ویژگی‌های FREAK در زمان اجرا همراه بوده است که در مقایسه با میزان افزایش دقت، به راحتی قابل چشم پوشی است. همچنین، در مقایسه دو روش ساخت مدل CSVM و MSVM، ملاحظه می‌شود استفاده از الگوریتم آبشاری در ساخت مدل به ترتیب حدود ٪۶۳ و ٪۵۵ زمان ساخت مدل را کاهش داده است. مقایسه دقت بازشناسایی این دو روش نیز نشان می‌دهد روش CSVM تقریبا به ترتیب ٪۱/۲ و ٪۲/۵ دقت بازشناسایی را بهبود داده است. این نتایج نشان می‌دهند استفاده از الگوریتم آبشاری در ساخت مدل نه تنها باعث افزایش چشم‌گیر سرعت ساخت آن می‌شود، بلکه باعث افزایش دقت بازشناسایی نیز می‌گردد.

مقایسه نتایج به‌دست آمده در گروه‌های مختلف رفتارها از هر دو پایگاه‌داده، نشان می‌دهد که تشخیص فعالیت‌های انسان-اشیا دشوارتر از بقیه گروه‌ها است. شاید دلیل این امر را در شباهت فعالیت‌ها و همچنین جزییات زیاد انجام شونده در این گروه از فعالیت‌ها دانست. در مقابل، فعالیت‌های ورزشی و یا انسان-انسان، با دقت بیشتری نسبت به سایر فعالیت‌ها بازشناسایی شده‌اند. بدیهی است که این دقت بالا در نتیجه‌ی متمایز بودن حرکات در وزش‌های متفاوت و تعامل بین انسان-انسان است.

شکل‌های ۷ و ۸ به ترتیب ماتریس‌های اغتشاش به‌دست آمده از روش پیشنهادی روی پایگاه‌داده‌های UCF101 و HMDB51 را نشان می‌دهند. در ماتریس‌های مصور شده در این شکل‌ها نیز عمل‌کرد بسیار مناسب روش پیشنهادی برای بازشناسایی رفتارهای بسیار

متنوع و متعدد در این دو پایگاه‌داده، مشهود است.

۳-۴- مقایسه با سایر روش‌ها

در این بخش، عمل‌کرد روش پیشنهادی با سایر روش‌های اخیری که نتایج کارایی‌شان را با پایگاه‌داده‌های UCF101 و HMDB51 سنجیده‌اند، مقایسه می‌شوند. توضیح این‌که، در مقالاتی که نتایج‌شان گزارش شده است، از مجموعه داده آموزشی و آزمایشی یکسانی استفاده شده است. متاسفانه به دلیل عدم ارایه مدت زمان ساخت مدل و همچنین مدت زمان صرف شده برای تشخیص فعالیت در ویدیوهای آزمایشی در مقالات مورد مقایسه و یا عدم همگن بودن معیار سنجش، زمان ساخت و زمان پردازش در نتایج این بخش قابل مقایسه نیست. جدول ۵ نتایج گزارش شده مقالات مرتبط و روش پیشنهادی را نشان می‌دهد.

با مقایسه کارایی روش پیشنهادی از نظر دقت بازشناسایی با روش پیشنهاد شده در [۳۰] که یک روش مبتنی بر ترکیب ویژگی‌ها و کیسه واژه‌ها است، نتیجه می‌شود که این دو روش نتایج تقریبا یکسانی را روی پایگاه‌داده UCF101 به‌دست آورده‌اند. اما به این دلیل که در روش پیشنهادی فاز استخراج ویژگی به مراتب سریع‌تر از این روش است، استفاده از آن در کاربردهای بلادرنگ اولویت بیشتری خواهد داشت. با مقایسه نتایج دو روش روی پایگاه‌داده HMDB51، مشاهده می‌شود روش پیشنهادی دقت بازشناسایی بهتری را به‌دست آورده است.

مقایسه دقت بازشناسایی روش پیشنهادی با روش پیشنهاد شده در [۳۲] نشان می‌دهد که متوسط دقت به‌دست آمده توسط روش پیشنهادی روی هر دو پایگاه‌داده کمتر از ٪۱ از آن روش است. با توجه به این‌که روش [۳۲] یک روش پیچیده و زمانبر هم از نظر ساخت مدل و هم از نظر سرعت بازشناسایی (به دلیل استفاده از ترکیب پنج ویژگی) است، از این‌رو، استخراج ویژگی در این روش زمان پردازشی زیادی را در فاز بازشناسایی به سامانه تحمیل خواهد کرد. در حالی‌که در روش پیشنهادی فقط از ترکیب دو ویژگی استفاده شده است و فاز کاهش ابعاد نیز در آن وجود ندارد. بنابراین با ملاحظه هزینه محاسباتی بسیار کم روش پیشنهادی، این تفاوت اندک در دقت تشخیص قابل چشم‌پوشی است.

مقایسه روش پیشنهادی با [۳۳]، نشان می‌دهد که روش پیشنهادی ٪۱۲/۸ دقت بازشناسایی را افزایش داده است. لازم به ذکر است که ویژگی‌های استفاده شده در [۳۳] شامل HOG، HOF و MBH بوده و روش BoW برای کدگذاری ویژگی‌ها به‌کار برده شده است.

همانطور که جدول ۵ نشان می‌دهد، دقت تشخیص راهکار پیشنهادی روی پایگاه‌داده‌ی UCF101 فاصله نزدیکی با روش ارایه شده در [۳۵] داشته و ٪۵/۲ بهتر از روش ارایه شده در [۳۴] عمل کرده است. روش‌های ارائه شده در [۳۴, ۳۵] روش‌های مبتنی بر یادگیری عمیق هستند. همانطور که پیش‌تر نیز اشاره شد، مرحله استخراج ویژگی در روش‌های مبتنی بر یادگیری عمیق با استفاده از شبکه‌های عصبی چندلایه و بدون مهندسی کردن روش استخراج ویژگی انجام می‌شود. با توجه به ساختار پیچیده‌ی رویکرد یادگیری عمیق، فرآیند ساخت مدلی مبتنی بر آن فوق‌العاده زمان‌بر بوده و همچنین نیاز به فراهم بودن داده‌های آموزشی بسیار زیادی دارد. با این حال، نتایج ارزیابی روش‌ها روی پایگاه‌داده‌ی HMDB51، نشان می‌دهد که روش پیشنهادی روی این پایگاه داده بهتر از روش ارایه شده در [۳۵] عمل کرده و دقت آن‌را تقریبا ٪۳ افزایش داده است.

با توجه به ماهیت روش‌های مبتنی بر یادگیری عمیق، ممکن است مدت زمان استخراج ویژگی‌ها، آموزش و ساخت مدلِ یادگیری در آن‌ها روزها و یا هفته‌ها بطول بیانجمد که این امر نیز ممکن است در کاربردهای واقعی و روزمره که معمولا سخت‌افزار و زیرساخت‌های لازم مهیا نیست، باعث عملیاتی نشدن این روش‌ها شود. بعنوان مثال، یکی از کاربردهای بسیار متداول سامانه‌های بازشناسایی حرکات انسان، به‌کارگیری آن‌ها در دستگاه‌های ورزشی هوشمند است که برای تحلیل درستی و نحوه انجام حرکات ورزش‌کاران استفاده می‌شود. در چنین دستگاه‌هایی معمولا به خاطر وجود محدودیت‌های سخت‌افزاری متعددی، لازم است چنین سامانه‌هایی در قالب کیت‌های الکترونیکی بسیار کوچکی پیاده‌سازی و اجرا شوند. تطبیق الگوریتم‌های مبتنی بر یادگیری عمیق در چنین بسترهای سخت‌افزاریِ محدودی تقریبا عملی ناممکن است. این محدودیت‌های سخت‌افزاری در بسیاری از کاربردهای دیگر این سامانه‌ها از قبیل خانه‌های هوشمند، تجهیزات پزشکی رفتار درمانی، ادوات و تجهیزات جنگی کوچک نیز مشهود است. علاوه بر این، در دسترس بودن حجم زیادی از داده‌های آموزشی نیز جزء ملزومات ساخت مدل‌های مبتنی بر یادگیری عمیق به شمار می‌رود و در مواردی که دسترسی به داده‌های آموزشی کافی مقدور نباشد، چنین روش‌هایی از کارایی قابل قبولی برخوردار نخواهند بود.

۵- نتیجه‌گیری و کارهای آینده

در این مقاله، روش جدیدی برای بازشناسایی خودکار فعالیت‌های انسان در ویدیو ارایه گردید. در روش ارایه شده با تاکید بر استفاده از اطلاعات بافتی نقاط ویژگی استخراج شده از تصاویر ویدیویی و ترکیب آن‌ها با یک توصیف کننده‌ی حرکتی، و همچنین به‌کارگیری روش آبشاری در ساخت مدل طبقه‌بندی کننده، سامانه‌ای با کارایی

قابل قبول ارایه گردید. نشان داده شد که استفاده از توصیف‌گر FREAK در مقایسه با سایر توصیف‌گرهای به‌کار گرفته شده در تحقیقات قبلی، هم از نظر دقت و هم از نظر زمان پردازش از کارایی بهتری برخوردار است. نتایج به‌دست آمده از آزمایش‌ها روی پایگاه-داده‌های بزرگ و حاوی رفتارهای متنوع و واقعی از فعالیت‌های انسان‌ها، نشان دادند که استفاده از الگوریتم آبشاری در قالب رویکرد پیشنهادی، می‌تواند علاوه بر افزایش قابل ملاحظه سرعت ساخت مدل، دقت عمل‌کرد قابل مقایسه با روش‌های بسیار پیشرفته را نیز داشته باشد. با مقایسه کارایی روش پیشنهاد شده با سایر تحقیقات بویژه روش‌های مبتنی بر یادگیری عمیق، نشان داده شد که روش پیشنهادی با این‌که بر اساس استخراج مهندسی شده‌ی ویژگی‌ها عمل می‌کند، اما همچنان می‌تواند بعنوان گزینه‌ی مناسبی در مقایسه با روش‌های پر هزینه مبتنی بر یادگیری عمیق باشد و در کاربردهای واقعی مورد استفاده قرار بگیرد.

یافته‌های این تحقیق نویسندگان را بر آن داشته است که روش استخراج ویژگیِ FREAK را در راستای توصیف بافت با در نظر گرفتن توالی فریم‌ها در بُعد زمان (3D-FREAK) توسعه داده و کارایی آن را مورد سنجش قرار دهند. پیش‌بینی می‌شود در صورت توسعه چنین روشی، دقت سامانه در بازشناسایی فعالیت‌ها افزایش یابد.

مراجع

[1] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications", Machine Vision and Applications, Vol. 23, No. 2, pp. 255-281, 2012.

[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review", ACM Computing Surveys (CSUR), Vol. 43, No. 3, p. 16, 2011.

[3] R. Poppe, "A survey on vision-based human action recognition", Image and vision computing, Vol. 28, No. 6, pp. 976-990, 2010.

[4] D. Marr and L. Vaina, "Representation and recognition of the movements of shapes", Proceedings of the Royal Society of London B: Biological Sciences, Vol. 214, No. 1197, pp. 501-524, 1982.

[5] Y. M. Lui and J. R. Beveridge, "Tangent bundle for human action recognition", in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, 2011, pp. 97-102: IEEE.

[6] D. D. Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector", The Visual Computer, Vol. 32, No. 3, pp. 289-306, 2016.

[7] K. Anuradha and N. Sairam, "Spatio-temporal based approaches for human action recognition in static and dynamic background: a survey", Indian Journal of Science and Technology, Vol. 9, No. 5, 2016.

[8] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model", in Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92, 1992 IEEE Computer Society Conference on, 1992, pp. 379-385: IEEE.

[9] 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.

[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes", in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, Vol. 2, pp. 1395-1402: IEEE.

[11] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras", in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, Vol. 1, pp. 150-157: IEEE.

[12] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition", in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8: IEEE.

[13] I. Laptev, "On space-time interest points", International journal of computer vision, Vol. 64, No. 2-3, pp. 107-123, 2005.

[14] C. Harris and M. Stephens, "A combined corner and edge detector", in Alvey vision conference, 1988, Vol. 15, No. 50, p. 10.5244: Manchester, UK.

[15] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features", in Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, 2005, pp. 65-72: IEEE.

[16] T. Kadir and M. Brady, "Scale saliency: A novel approach to salient feature and scale selection", 2003.

[17] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 36, No. 3, pp. 710-719, 2005.

[18] P. R. Beaudet, "Rotationally invariant image operators", in Proc. 4th Int. Joint Conf. Pattern Recog, Tokyo, Japan, 1978, 1978.

[19] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector", Computer Vision–ECCV 2008, pp. 650-663, 2008.

[20] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition", in BMVC 2009-British Machine Vision Conference, 2009, pp. 124.1-124.11: BMVA Press.

[21] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", 1981.

[22] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", in Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 104-111: IEEE.

[23] M. B. Kaaniche and F. Brémond, "Tracking hog descriptors for gesture recognition", in Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on, 2009, pp. 140-145: IEEE.

[24] J. Shi, "Good features to track", in Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on, 1994, pp. 593-600: IEEE.

[25] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection", Computer Vision–ECCV 2006, pp. 430-443, 2006.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, Vol. 1, pp. 886-893: IEEE.

[27] D. G. Lowe, "Object recognition from local scale-invariant features", in Computer vision, 1999. The proceedings of the

seventh IEEE international conference on, 1999, Vol. 2, pp. 1150-1157: Ieee.

[28] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1-8: IEEE.

[29] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance", in European conference on computer vision, 2006, pp. 428-441: Springer.

[30] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice", Computer Vision and Image Understanding, Vol. 150, pp. 109-125, 2016.

[31] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification", Computer Vision–ECCV 2010, pp. 143-156, 2010.

[32] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 204-212.

[33] M. Bagheri et al., "Keep it accurate and diverse: Enhancing action recognition performance by ensemble learning", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 22-29.

[34] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625-2634.

[35] X. Wang, A. Farhadi, and A. Gupta, "Actions~ transformations", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2658-2667.

[36] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint", in Computer vision and pattern recognition (CVPR), 2012 IEEE conference on, 2012, pp. 510-517: Ieee.

[37] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context", in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 2929-2936: IEEE.

[38] G. D. Field et al., "Functional connectivity in the retina at the resolution of photoreceptors", Nature, Vol. 467, No. 7316, pp. 673-677, 2010.

[39] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF", in Computer Vision (ICCV), 2011 IEEE international conference on, 2011, pp. 2564-2571: IEEE.

[40] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo", IEEE transactions on pattern analysis and machine intelligence, Vol. 32, No. 5, pp. 815-830, 2010.

[41] G. Salton, "Automatic information organization and retrieval", 1968.

[42] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases", in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1-8: IEEE.

[43] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification", in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 3360-3367: IEEE.

[44] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors", in European conference on computer vision, 2010, pp. 141-154: Springer.

[45] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods", in BMVC, 2011, Vol. 2, No. 4, p. 8.

[46] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks", IEEE Transactions on Neural Networks and Learning Systems, Vol. 23, No. 3, pp. 412-424, 2012.

[۴۷] ا. فیضی، ع. آقاگل‌زاده، و م. سیدعربی، "شناسایی و دسته‌بندی رفتارها به منظور آشکارسازی رفتارهای غیر معمول با استفاده از مدل مارکوف مخفی"، مجله علمی-پژوهشی رایانش نرم و فناوری اطلاعات، جلد ۵، شماره ۲، تابستان ۱۳۹۵.

[48] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", ed, 2007.

[49] C. Cortes and V. Vapnik, "Support-vector networks", Machine learning, Vol. 20, No. 3, pp. 273-297, 1995.

[50] X. Ke, H. Jin, X. Xie, and J. Cao, "A distributed SVM method based on the iterative MapReduce", in Semantic Computing (ICSC), 2015 IEEE International Conference on, 2015, pp. 116-119: IEEE.

[51] O. Meyer, B. Bischl, and C. Weihs, "Support vector machines on large data sets: Simple parallel approaches", in Data Analysis, Machine Learning and Knowledge Discovery: Springer, 2014, pp. 87-95.

[52] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild", arXiv preprint arXiv:1212.0402, 2012.

[53] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition", in Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 2556-2563: IEEE.

زیرنویس‌ها:

۱ Gestures

۲ Actions

۳ Interactions

۴ Feature points

۵ معادل برای واژه Mask استفاده شده است.

۶ Hidden Markov Models (HMMs)

۷ Motion-Energy Image (MEI)

۸ Motion-History Image (MHI)

۹ Trajectory

۱۰ Spatio-Temporal Interest Point (STIP)

۱۱ Harris

۱۲ Bag-of-Words (BoW)

۱۳ Salient Region Detector (SRD)

۱۴ Hessian

۱۵ Trajectory

۱۶ Kanade-Lucas-Tomasi

۱۷ Shi-Thomasi

۱۸ Features from Accelerated Segment Test (FAST)

[۱۹] Histogram of Oriented Gradient (HOG)
[۲۰] Scale-Invariant Feature Transform (SIFT)
[۲۱] Histogram of Optical Flow (HOF)
[۲۲] Motion Boundary Histogram (MBH)
[۲۳] Vector Quantization
[۲۴] Dempster–Shafer
[۲۵] Eigenvalues
[۲۶] Fast Retina Keypoint
[۲۷] Ganglion
[۲۸] Foveal
[۲۹] Fovea
[۳۰] Parafoveal
[۳۱] Perifoveal
[۳۲] Bag-of-features
[۳۳] Bag-of-words
[۳۴] Kernel coding
[۳۵] Locality-constrained Linear Coding
[۳۶] Max Pooling
[۳۷] Super coding
[۳۸] Gaussian Mixture Model
[۳۹] Clutter
[۴۰] http://www.vlfeat.org
[۴۱] https://github.com/tzulitai/distributed-svm
[۴۲] Radial Basis Function
[۴۳] Multiclass SVM