مجله علمی پژوهشی «علوم و فناوری‌های پدافند نوین»

سال هفتم، شماره ۴، زمستان ۱۳۹۵؛ ص ۳۲۶-۳۱۵

خلاصه‌سازی مقیاس‌پذیر ویدئویی با استفاده از انتخاب واژه‌نامه تنک

پوریا اعتضادی‌فر[۱]، حسن فرسی[۲*]

۱- دانشجوی دکترا، ۲- استاد، دانشگاه بیرجند

(دریافت: ۹۴/۱۲/۰۸، پذیرش: ۹۵/۰۶/۲۸)

چکیده

یکی از حوزه‌های مهم در پدافند غیرعامل، شناسایی تهدیدات و اعلام هشدار است. یکی از روش‌های پرکاربرد در حوزه شناسایی بررسی داده-های ویدئویی به منظور شناسایی اهداف ناشناس و اعلام هشدار است. به منظور بررسی سریع، با دقت بالا روش‌های خلاصه‌سازی ویدیو ارائه شده است. همچنین در طول سال‌های گذشته، ایجاد ویدیو دیجیتالی منجر به رشد نمایی محتوای ویدئویی شده است. به منظور افزایش قابلیت استفاده از این حجم بالای ویدیو، تحقیقات بسیاری به انجام رسیده و خلاصه‌سازی ویدیو به جهت مرور سریع این مجموعه ویدئویی بزرگ و برای کمک به فهم سریع محتوای داده‌های ویدئویی پیشنهاد شده است. در خلاصه‌سازی ویدیو، تصاویری به عنوان نماینده از هر صحنه انتخاب می‌شود تا مروری تصویری از تمام فیلم به دست آید. اخیراً روش‌هایی با استفاده از فرمول‌بندی تنک برای خلاصه‌سازی ویدیو، داده‌های ویدئویی را به میزان زیادی نسبت به دیگر روش‌ها خلاصه نموده‌اند. در این مقاله به خلاصه‌سازی ویدیو، به عنوان یک مسئله انتخاب واژه‌نامه تنک پرداخته می‌شود. بدین منظور، با استفاده از روشی جدید بر پایه کدینگ تنک، می‌توان به میزان زیادی خلاصه‌سازی داده‌های ویدئویی را نسبت به دیگر روش‌های خلاصه‌سازی ویدیو که با روش تنک و یا روش‌های دیگر پیشنهادشده‌اند بهبود بخشید. اساس این روش بر پایه حل معادله بهینه‌سازی با استفاده از آستانه‌گذاری نرم است که پیچیدگی کمتری نسبت به روش‌های پیشنهادشده اخیر است. این امر را می‌توان با بررسی میزان پیچیدگی روش پیشنهادی با روش‌های متداول اخیر متوجه شد. در انتها نتایج آزمایش برای مجموعه داده‌های معیاری زمین حقیقت و با روش‌های State of the art، ادعای ما در بهبود میزان خلاصه‌سازی روش پیشنهادی را نشان می‌دهد.

کلیدواژه‌ها: خلاصه‌سازی ویدیو، انتخاب واژه‌نامه، فریم کلیدی، آنالیز ویدیو، کدینگ تنک

Scalable Videos Summarization Using Sparse Dictionary Selection

P. Etezadifar, H. Farsi*

University of Birjand

(Received: 27/02/2016; Accepted: 18/09/2016)

Abstract

One of the important topics of passive defense is threats detection and waning alarm. One of the most widely used methods in detection field is video data investigation in order to identify unknown targets and warning alarm. In order to evaluate a fast and high-precision technique, video summarization is presented. Also, during the past years, creation of digital videos has caused exponential growth of video content. To increase the high volume of video usability, a lot of researches have been done and video summarization has been proposed to quick view of large video collection and quick understanding of the content of video data. In the video summarization, pictures are selected as a representative of each scene to obtain a visual overview of whole video. Recently, new methods using sparse formulation are suggested for video summarization being more effective in video data summarization than other methods. In this paper, video summarization is presented as a sparse dictionary selection problem. For this purpose, using a new method based on sparse coding, have been able to improve video data summarization compared to other video summarization methods based on sparse or other coding. Finally, the results for the ground truth data collection and State of the art methods, shows improvement our claim in the video summary on proposed method.

Key words: Video Summarization, Dictionary Selection, Key Frame, Video Analysis, Sparse Coding.

*Corresponding Author E-mail: hfarsi@birjand.ac.ir

۱. مقدمه

همان طور که می‌دانید، یکی از حوزه‌های مهم در پدافند غیرعامل، شناسایی تهدیدات و اعلام خبر است. به منظور پیشگیری از حملات دشمنان می‌بایست هر لحظه اقدامات احتمالی مورد بررسی و جستجو قرار گیرد. یکی از این حوزه‌ها رصد هوایی با استفاده از سامانه‌های هوشمند پردازشی است. با توجه به احتمال تجاوز در هر لحظه از زمان به وسیله هواپیماهای رادار گریز و موشک‌های کروز که اصولاً به دلیل ارتفاع پایین پروازی قادر به ردیابی توسط سامانه‌های راداری نمی‌باشند؛ می‌بایست از روش‌های تصویربرداری برخط (ویدیو) استفاده کرد. به دلیل حجم بالای داده‌های به دست آمده توسط دوربین‌های ویدئویی نمی‌توان اطلاعات آن‌ها را به صورت برخط پردازش کرده و سامانه قادر به اطلاع‌رسانی در زمان خبر نیست. یکی از روش‌هایی که اخیراً مطرح شده است، خلاصه‌سازی ویدیو است که با استفاده از آن بتوان نسبت داده به اطلاعات را افزایش داده و خروجی به دست آمده را بتوان به عنوان داده‌ای با میزان اطلاعات بالا به سامانه تصمیم‌گیر وارد کرده و با استفاده از آن بتوان به وجود و یا عدم وجود اشیای متخاصم هوایی پی برد. همچنین از روش خلاصه‌سازی ویدیو در حملات آفندی به منظور عدم آشکارسازی به وسیله سنسورهای راداری استفاده می‌شود [۳-۱]. به عنوان مثال در موشک‌های کروز و هواپیماهای بدون سرنشین به منظور مسیریابی از سامانه‌های ویدئویی استفاده می‌شود که اساس کار آن‌ها با استفاده از روش‌های خلاصه‌سازی ویدیو است. بنابراین، یکی از روش‌ها به منظور مقابله با یک سامانه آفندی بررسی روش‌های کاری در آن سامانه آفندی است. یکی از روش‌های مورد استفاده به منظور تعیین هدف در فاز نهایی موشک کروز استفاده از ویژگی‌های بصری است. بدین صورت که در فاز انتهایی، موشک کروز شروع به جست‌وجو در فریم‌های ویدئویی اخذشده از سطح زمین می‌کند و پس از تطبیق منطقه مورد نظر با تصاویر موجود در حافظه‌اش، به منطقه هدف اصابت می‌کند [۶-۴]. به‌دلیل نرخ بالای تصاویر به‌دست آمده از دوربین تعبیه شد، در موشک کروز، نمی‌توان تمام فریم‌ها را مورد بررسی قرار داد. یکی از راه‌های کاهش نرخ فریم‌های جست‌وجو استفاده از روش‌های خلاصه‌سازی ویدیو است. در این مقاله به بهبود روش خلاصه‌سازی پرداخته شده است. همچنین، می‌توان در زمینه‌های صنعتی و تجاری نیز استفاده کرد. باید به این نکته اشاره نمود که خلاصه‌سازی ویدیو به یک ابزار کلیدی، برای مرور و دسترسی کارآمد و نیز برای دست‌کاری مجموعه بزرگ ویدئویی، تبدیل شده و توجه بیشمار محققین را به خود جذب کرده است. بیشتر فعالیت‌هایی که در حوزه خلاصه‌سازی سیگنال ویدئویی متمرکز شده‌اند، بر پایه ویژگی‌های گوناگونی مثل حرکت[1] [۷]، شنیداری[2] [۸] و یا روش‌های ترکیبی[3] [۹] است. مقاله مروری در مورد سامانه اتیک از نمونه عملیات نیز ارئه شده اند [۱۰]. در مدیریت مجموعه ویدئویی، ویدیوهای مشابه و یا مرتبط، معمولاً مطابق با محتوای آن‌ها به فصل‌هایی گروه‌بندی می‌شوند. بنابراین، کاربران این ویدیوها، به دیگر روابط بین گروه‌هایی از این فایل‌های ویدئویی علاقه‌مند می‌شوند. برای مثال، در آرشیوهای اخبار ویدئویی، تمام کلیپ‌هایی که در مورد یک موضوع از کانال‌های مختلف و جلسات خبری پخش می‌شود، هر دو با هم تکامل موضوع را نشان می‌دهند. از طرفی با پیشرفت فناوری در زمینه تصویربرداری و افزایش اطلاعات ویدئویی ساخته شده، افزایش تقاضا مردم برای ذخیره‌سازی، به اشتراک گذاشتن اطلاعات ویدئویی در اینترنت و شبکه‌های اجتماعی، بازیابی ویدیو[4] [۱۱]، مراقبت داخلی[5] [۱۲]، آشکارسازی ناهنجاری[6] [۱۳] و ... بیش از پیش، مورد توجه قرار گرفته است. در این مورد خلاصه‌سازی ویدیو با حذف داده‌های حشو، با استفاده از فشرده‌سازی و چگال کردن آن‌ها، کمک بسیاری به افزایش سرعت برنامه‌ها، کاهش فضای ذخیره‌سازی و سادگی ارسال از طریق اینترنت می‌شود. روش‌های خلاصه‌سازی ویدئویی را می‌توان به دو صورت کلی خلاصه‌سازی استاتیکی[7] و دینامیکی[8] مجزا نمود [۱۴]. خلاصه‌سازی استاتیکی ویدیو بر اساس پیدا کردن فریم‌های کلیدی کار می‌کند. منظور از فریم‌های کلیدی، تصاویری ثابت از ویدیو هستند که مهم‌ترین محتوای آن را شامل می‌شوند [۱۵]. ولی خلاصه دینامیکی ویدیو، شامل دنباله‌ای از شات‌های کوچک است که پشت سر هم قرار گرفته‌اند. با این وجود، استخراج فریم کلیدی، انعطاف بیشتری برای نشان دادن محتوای ویدیو دارد. به طور معمول، استخراج فریم کلیدی به دو صورت مبتنی بر متن و محتوا انجام می‌گیرد [۱۶]. در روش مبتنی بر متن، کلمه و یا جمله‌ای دلخواه انتخاب می‌شود و فریمی کلیدی متناسب با آن متن استخراج می‌شود. این روش زمان‌بر و پرهزینه است. روش مبتنی بر محتوا، بر اساس ویژگی‌های استخراج شده از ویدیو صورت می‌پذیرد. این روش نسبت به روش مبتنی بر متن ارزان‌تر و سریع‌تر است ولی به دلیل تفاوت ادراک انسان و ماشین، زیاد قابل‌اطمینان نیست.

[1] Motion
[2] Audio
[3] Multi-Modality
[4] Video Retrieval
[5] Indoor Surveillance
[6] Anomaly Detection
[7] Static
[8] Dynamic

بنابراین، محققین برای کاهش این تفاوت، از توصیف‌گرهای مختلفی استفاده می‌کنند. در ادامه ساختار کلی مقاله به صورت خلاصه مورد بررسی قرار می‌گیرد. در اولین گام به فرموله کردن مسئله خلاصه‌سازی ویدیو با استفاده از انتخاب واژه‌نامه تنک پرداخته می‌شود. در مرحله دوم به حل مسئله محاسبه‌شده برای خلاصه‌سازی ویدیو، در مرحله سوم به چگونگی پیاده‌سازی حل مسئله و الگوریتم آن پرداخته می‌شود. در مرحله بعد، به بررسی ویژگی‌هایی که برای آموزش واژه‌نامه[1] استفاده می‌شود پرداخته شده و در انتها نتیجه روش پیشنهادی با دیگر روش‌های پیشنهادشده در خلاصه‌سازی ویدیو مقایسه می‌شود. در ادامه به بررسی کارهای صورت گرفته در این زمینه توسط محققین پرداخته می‌شود.

در زمینه خلاصه‌سازی ویدیو، کارهای زیادی انجام شده است[۱۰و ۱۷ و ۱۸] که مجال بررسی همه آن‌ها نیست. اساس روش‌های پیشنهادشده بدین صورت است که ویژگی خاصی را برای مجموعه فریم‌های ویدئویی در نظر گرفته و فریم‌هایی را میزان تغییرات این ویژگی نسبت به فریم قبل خود بیشتر بوده به عنوان فریم کلیدی استخراج می‌شوند. در ادامه به بررسی چند مقاله معروف در این زمینه پرداخته می‌شود. کلبان [۱۹] روشی برای مقابله با مشکلات خلاصه‌سازی ویدیو معرفی کرد که این روش بر اساس به‌کارگیری ویژگی‌های تلفیقی سطح بالا بود. در روش پیشنهادی کلبان، در ابتدا نمونه‌برداری به منظور حذف داده‌های اضافی، انجام می‌شود و تنها زیرمجموعه‌ای از فریم‌های ویدئویی باقی می‌ماند. سپس پنج ویژگی سطح بالا از فریم‌های باقی‌مانده استخراج می‌شود. در مرحله بعد، وزن‌های بهینه‌ای برای ترکیب این ویژگی‌ها، توسط روش گرادیان نزولی استنتاج می‌شوند. در نهایت، از الگوریتم وزنی k-means برای تشخیص مهم‌ترین قسمت‌های یک خلاصه‌سازی نهایی، استفاده شده است. در این مقاله، از روش دسته‌بندی ضعیفی به منظور انتخاب فریم کلیدی استفاده شده است که ضعف این روش محسوب می‌گردد روش پیشنهادی توسط پن [۲۰] روشی برای خلاصه‌سازی ویدیو بر اساس الگوریتم آشکارسازی مرز شات است. در ابتدا، مرز شات توسط مقایسه هیستوگرام رنگ فریم‌های متوالی، به دست می‌آید. پس از آن، با استفاده از گروه‌بندی شات‌های مشابه و انتخاب یک شات به عنوان نماینده از هر گروه و حذف شات‌های تکراری عمل می‌کند. در انتها، قسمت‌های انتخاب‌شده بر اساس اهمیت آن‌ها، خلاصه ویدیو را می‌سازند. در این روش به دلیل استفاده از مرز شات نمی‌توان به طور کلی داده ویدئویی را مورد بررسی قرار دهد. بنابراین یک صحنه نسبتاً مشابه که در زمان‌های متفاوت نمایش داده می‌شود به عنوان صحنه‌ای جدید در نظر گرفته‌شده و فریم‌های کلیدی تکراری تولید می‌گردد. در روشی دیگر که برای خلاصه‌سازی محتوایی ویدیو استریو پیشنهادشده در ابتدا ویدیو را به شات‌هایی تقسیم‌بندی می‌کند. سپس فریم‌هایی که ویژگی بصری شبیه به یک دیگر دارند با یک دیگر جمع می‌شوند. سپس ناهمخوانی زمینه[2]، مناطق مسدود[3] و نقشه عمق[4] تخمین زده می‌شود. در گام بعد، بردار ویژگی با استفاده از کلاس‌بندی فازی ویژگی‌ها ساخته شده و انتخاب شات با استفاده از کلاس بندی شات‌های مشابه صورت می‌گیرد. در انتها فریم‌های کلیدی با استفاده از روش بهینه‌سازی برای مکان فریم‌ها با پیدا کردن کمترین شباهت بین فریمی استخراج می‌شوند [۲۱]. این روش نیز مشکلاتی نزدیک به آنچه برای روش پن گفته شد را دارا است. ساتو و لی [۲۲] الگوریتم دسته‌بندی برای گروه‌بندی فریم‌ها با رنگ‌های مشابه ارائه دادند. در این روش، دنباله ویدئویی به قطعه‌هایی تقسیم می‌شود که این کار با گروه‌بندی فریم‌های مشابه بر اساس ممان‌های رنگ انجام می‌شود. سپس، به دسته‌بندی این قطعه‌ها می‌پردازد. پس از آن، قطعات پشت سرهم که متعلق به یک دسته‌اند با هم ترکیب می‌شوند و قطعات مجاور به منظور کاهش افزونگی، ادغام می‌شوند. در نهایت، زیرمجموعه‌ای از فریم‌ها از طولانی‌ترین قسمت‌ها برای خلاصه ویدئویی سخت، انتخاب می‌شوند. در روشی دیگر، با استفاده از انتخاب واژه‌نامه تنک به استخراج فریم‌های کلیدی پرداخته است، البته این روش به جای استفاده از نرم ۱، نرم ۱و۲ را برای شرط تنکی خود استفاده کرده و با استفاده از روش نسترو[5] [۲۳] به انتخاب واژه‌نامه تنک برای مسئله محدب ولی غیر مسطح[6] پرداخته است [۲۴]. این روش دارای پیچیدگی نسبتاً بالایی است که از معایب آن محسوب می‌شود. روشی دیگر، با استفاده از محاسبه میزان تفاوت بین هیستوگرام رنگ دو فریم متوالی فریم کلیدی استخراج شده است [۲۵]. این روش بسیار نسبت به تغییرات شدت روشنایی حساس بوده و صحنه‌های شبیه به یک دیگر که دارای تغییرات شدت روشنایی هستند را به عنوان صحنه‌های جدید انتخاب کرده و فریم‌ها کلیدی شبیه به یک دیگر انتخاب می‌کند. فو و همکارانش [۲۶]، روشی برای خلاصه‌سازی ویدیوهای چند نمایشی[7] پیشنهاد کرده‌اند. در این

[1] Dictionary

[2] Disparity Field
[3] Occluded Areas
[4] Depth Map
[5] Nesterov
[6] Non-Smooth
[7] Multi-View Videos

مقاله در گام اول گراف شات فضا- زمانی[1] ساخته می‌شود و با فرموله کردن عملیات برچسب‌گذاری گراف به خلاصه‌سازی ویدیو می‌پردازد. در روشی دیگر ژانگ و همکارانش [۲۷]، ابتدا کل فضای ویدیو کلاس‌بندی کرده و پس از آن فریم‌هایی که در مرکز کلاس قرار دارند را به عنوان فریم‌های کلیدی انتخاب کرده؛ ولی اگر فریمی در مرکز خوشه قرار نداشته باشد، نزدیک‌ترین فریم به مرکز به عنوان فریم کلیدی استخراج می‌شود. در این روش، کلاس‌بندی استفاده‌شده ضعیف بوده و به خوبی قادر به مجزا نمودن صحنه‌ها از یک دیگر ندارد. در روشی دیگر، با استفاده از محاسبه آنتروپی توأم و اطلاعات مشترک بین فریم‌های متوالی به آشکارسازی محوشدگی و برش شات می‌پردازد و سپس در هر شات فریم‌های کلیدی را با استفاده از ویژگی‌های آنتروپی محاسبه می‌کند [۲۸]. این روش نیز شبیه به روش پن قادر به بررسی کلی فیلم ناست و نمی‌تواند به درستی صحنه‌های مشابه را حذف نماید.

۲. الگوریتم پیشنهادی

در این مقاله، هدف خلاصه‌سازی ویدیو با استفاده از انتخاب واژه‌نامه تنک است. در ادامه به صورت خلاصه فرضیات اولیه استفاده‌شده در این مقاله را توضیح داده خواهد شد و پس از آن به فرموله کردن مسئله پرداخته می‌شود. در این مقاله بردار با حرف کوچک و به صورت $\underline{x}$ و همچنین ماتریس با حرف بزرگ و به صورت X نشان داده می‌شود. همچنین فرض شده است که سایز پایگاه داده ویدئویی n است به این معنا که در این پایگاه داده n فریم ویدئویی وجود دارد ($D \in \Re^{d\times n}$). همچنین، پایگاه داده به صورت $D=\{\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_n\}$ نمایش داده می‌شود. در این پایگاه داده فریم‌های دوبعدی به بردارهای ویژگی یک‌بعدی تبدیل شده‌اند. بنابراین، $\underline{d}_i \in \Re^d$، نشان‌دهنده i امین تصویر از پایگاه داده با ابعاد $d\times 1$ است. برای این کار می‌توان از روش کیف کلمات[2] Gist [۲۹]، وCENTRIST [۳۰ و ۳۱] استفاده کرد و یا اینکه پیکسل‌ها را به صورت برداری زیر هم قرار داد. هدف این مقاله استخراج فریم‌های کلیدی $D'_{key} = \Im_{key}(D) = \{\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_k\}$ به طول k است. منظور از $\Im_{key}(D)$ پروسه استخراج فریم‌های کلیدی و همچنین D'_{key} پایگاه داده فریم‌های کلیدی استخراج‌شده به صورت $D'_k \in \Re^{d\times k}$ است که باید تا حد امکان مقدار $k = n$ باشد. این شرط باعث ارضای تنک بودن فریم‌های کلیدی می‌شود.

در این مسئله، فرض می‌شود مشاهده‌ای از بردار x به صورت زیر در دسترس باشد:

$$\underline{y} = D\underline{x} + \underline{e} \quad \& \begin{cases} \underline{x} \in \Re^n \\ \underline{y} \in \Re^m \\ D \in \Re^{m\times n} \end{cases} \qquad (۱)$$

در معادله (۱) منظور از $\underline{y}$، مشاهدات و منظور از $\underline{e}$، نویز است. همچنین، D به عنوان واژه‌نامه و معلوم بوده و $\underline{x}$ بردار مجهول است. همچنین، توزیع پیشین هر یک از درایه‌های $\underline{x}$ در این مسئله یک توزیع لاپلاس با میانگین صفر است و درایه‌ها مستقل از هم[3] (i.i.d) می‌باشند. دلیل استفاده از توزیع لاپلاس به این لحاظ است که این توزیع حول مقادیر صفر تیز بوده و به جواب مسئله که باید جوابی تنک باشد نزدیک‌تر است تا بقیه توزیع‌ها از قبیل گوسی و ... همچنین نویز دارای توزیع گوسی با میانگین صفر و واریانس $\sigma^2 I$ است. در این مسئله چون توزیع پیشین متغیر مجهول $\underline{x}$ در دسترس است بهترین تخمین MAP است [۳۲]. بنابراین برای تخمین $\underline{x}$ باید معادله (۲) حل گردد.

$$\hat{x}_{map} = arg\max_{\underline{x}} \left\{ Ln\left(P\left(y \mid x\right)\right) + Ln\left(P\left(\underline{x}\right)\right) \right\} \qquad (۲)$$

با توجه به فرضیات مسئله می‌دانید که توزیع $\underline{y} \mid \underline{x}$، همان توزیع نویز با میانگین $D\underline{x}$ است که برابر با $N\left(Dx, \sigma^2 I\right)$ است. همچنین می‌دانید اگر $\underline{e} \sim N\left(\mu, \Sigma\right)$, $\underline{e} \in \Re^m$ آنگاه دارید [۳۴]:

$$P\left(y \mid x\right) = \frac{1}{\left(2\pi\right)^{m/2}\sigma^m} e^{-\frac{\|y-Dx\|_2^2}{2\sigma^2}} \qquad (۳)$$

با استفاده از معادله‌های (۲ و ۳) و با حذف مقادیر ثابت، با توجه به اینکه مقادیر ثابت در بهینه‌سازی تأثیری ندارد، دارید:

$$\hat{\underline{x}}_{map} = arg\min_{\underline{x}} \left\{ \left\| \underline{y} - D\underline{x} \right\|_2^2 + \lambda \|\underline{x}\|_1 \right\} \qquad (۴)$$

می‌توان معادله (۴) را با استفاده از مدل‌سازی واژه‌نامه استفاده‌شده در Group Lasso [۳۳ و ۳۴]، به صورت زیر بازنویسی کرد:

$$\hat{X} = arg\min_{X} \left\{ \|Y - DX\|_F^2 + \lambda \|X\|_1 \right\} \qquad (۵)$$

در معادله (۵)، منظور از X، ماتریس ضرایب تعقیب است، نرم فروبنیوس $\|X\|_F$، به صورت $\|X\|_F \triangleq \left(\sum_{i,j} X_{ij}^2\right)^{1/2}$ تعریف می‌شود.

[1] Spatio-Temporal Shot Graph
[2] Bag of Words (BOW)
[3] Independent and Identically Distributed

همچنین، نرم ۱ برای ماتریس به صورت $\|X\|_1 = \sum_{i,j}|X_{ij}|$ محاسبه می‌شود. جواب معادله (۵) در صورت کوچک بودن مقدار λ به صورت حدی به ماتریس همانی I میل می‌کند. این بدان معناست که حل معادله به سمتی پیش می‌رود تا فریم‌های کلیدی برابر با کل فریم‌های ویدیوی ورودی شوند، اما با افزایش مقدار λ، میزان عناصر صفر ماتریس X افزایش پیدا می‌کند و معادله (۵) به سمتی می‌رود که تمامی مقادیر X به سمت صفر میل کنند. این بدان معناست که فریم‌های کلیدی استخراج‌شده را به کمترین تعداد فریم‌ها میل می‌دهد. در این مقاله مقادیر واژه‌نامه D با $D_{key}^{'} \in \Re^{d\times k}$ که واژه‌نامه آموزشی نامیده می‌شود، Y با $D \in \Re^{d\times n}$ که همان واژه‌نامه اولیه ساخته شده با استفاده از ویژگی‌ها و همچنین $X \in \Re^{k\times n}$ ماتریس ضرایب پیگیری[1] می‌باشند. قسمت دوم معادله (۵) که تنک بودن ماتریس X را با استفاده از نرم ۱ محاسبه می‌کند؛ به دنبال آن است تا سطرهای زیادی از ماتریس X را تمام صفر کند و با استفاده از آن سطرهای غیر صفر را به عنوان فریم‌های کلیدی انتخاب کند. بنابراین می‌توان به جای استفاده از نرم ۱ از نرم ۲ و ۱ استفاده کرد [۳۵] که به صورت زیر تعریف می‌شود:

$$\|X\|_{2,1} = \sum_{i=1}^{k}\left|\left(\sum_{j=1}^{n}\underline{x}_{ij}^{\,2}\right)^{1/2}\right| \qquad (۶)$$

با توجه به توضیحات قبلی منظور از $\|\underline{x}_i\|_2$ نرم ۲، سطر iام ماتریس X است. بنابراین با جایگزینی مقدار نرم ۲و۱ با نرم ۱ معادله (۵) به صورت زیر بازنویسی می‌شود:

$$\hat{X} = \underset{X}{arg\,min}\left\{\|Y - DX\|_F^2 + \lambda\|X\|_{2,1}\right\} \qquad (۷)$$

در این مقاله با استفاده از الگوریتم آستانه‌گذاری نرم[2] از خانواده الگوریتم‌های انقباضی[3] به حل این موضوع پرداخته می‌شود. الگوریتم آستانه‌گذاری نرم به حل معادله (۷) می‌پردازد. می‌توان معادله (۷) را به صورت زیر به منظور حذف ضریب D از X نوشت:

$$\tilde{f}(\underline{x}, \underline{x}_0) = f(\underline{x}) + d(\underline{x}, \underline{x}_0)$$
$$if \begin{cases} f(\underline{x}) = \|Y - DX\|_F^2 \\ \qquad +\lambda\|X\|_{2,1} \\ d(\underline{x}, \underline{x}_0) = c/2\|\underline{x} - \underline{x}_0\|_2^2 \\ \qquad + 1/2\|D\underline{x} - D\underline{x}_0\|_2^2 \end{cases} \qquad (۸)$$

بنابراین به جای حل معادله (۷) به حل معادله زیر پرداخته می‌شود:

$$\hat{X} = \underset{X}{arg\,min}\left\{\begin{matrix} \|Y - DX\|_F^2 + \lambda\|X\|_{2,1} \\ +c/2\|\underline{x} - \underline{x}_0\|_2^2 \\ +1/2\|D\underline{x} - D\underline{x}_0\|_2^2 \end{matrix}\right\} \qquad (۹)$$

در ادامه، با محاسبه $\nabla\tilde{f}(\underline{x}, \underline{x}_0) = 0$، مقدار $\underline{x}$ را به دست می‌آید.

$$\begin{aligned} &\nabla\tilde{f}(\underline{x}, \underline{x}_0) = -D^T(Y - DX) + \\ &\lambda\,\mathrm{sgn}(\underline{x}) + C(\underline{x} - \underline{x}_0) - \\ &D^T(D\underline{x} - D\underline{x}_0) = 0 \\ &\Rightarrow -D^TY + D^TDX + \lambda\,\mathrm{sgn}(\underline{x}) + \\ &C(\underline{x} - \underline{x}_0) - D^TD\underline{x} + D^TD\underline{x}_0 = 0 \\ &\Rightarrow \underline{x} = \left(1/C\,D^T(Y - DX) + \underline{x}_0\right) + \\ &1/C\,\mathrm{sgn}(\underline{x}) \end{aligned} \qquad (۱۰)$$

با توجه به [۳۶ و ۳۷] می‌توان مقدار $\underline{x}$ را به صورت زیر به دست آورد:

$$\underline{x} = soft_{1/C}\left[1/C\,D^T(Y - DX) + \underline{x}_0\right] =$$
$$\begin{cases} 1/C\,D^T(Y - DX) + \underline{x}_0 - 1/C & ;1/C < 1/C\,D^T(Y - DX) + \underline{x}_0 \\ 0 & ;1/C \geq \left|1/C\,D^T(Y - DX) + \underline{x}_0\right| \\ 1/C\,D^T(Y - DX) + \underline{x}_0 + 1/C & ;1/C > 1/C\,D^T(Y - DX) + \underline{x}_0 \end{cases} \qquad (۱۱)$$

در معادله (۱۱)، C مقداری ثابت است که در این آزمایش برابر با $\lambda\left(D^TD\right)$ در نظر گرفته شده است. باید به این نکته اشاره کرد که برای همگرا شدن الگوریتم آستانه‌گذاری نرم باید $C \geq \lambda\left(D^TD\right)$ باشد [۳۶]. بنابراین مقدار هر سطر از ماتریس X با استفاده از رابطه (۱۱) محاسبه می‌شود.

در الگوریتم پیشنهادی، شرط همگرایی کوچک بودن مقدار نرم ۲ سطرهای ماتریس X است. پس از اجرای الگوریتم پیشنهادی، نرم ۲ سطرهای ماتریس X محاسبه‌شده و طبقه‌بندی[4] می‌شوند. در مرحله بعد مقادیر بیشتر به عنوان فریم‌های کلیدی ویدیو مورد نظر استخراج می‌گردند. در این مقاله با انجام هم‌زمان آموزش و انتخاب تنک واژه‌نامه به خروجی بهتری نسبت به روش‌های دیگر خلاصه‌سازی ویدیو دست پیدا کرده است. شبه کد الگوریتم پیشنهادی در ادامه نشان داده شده است.

[1] Pursuit Coefficient Matrix
[2] Soft Thresholding
[3] Shrinkage Algorithms
[4] Sort

الگوریتم پیشنهادی: انتخاب واژه‌نامه تنک

Input: *The Video Set* $D \in \Re^{d\times n}, \lambda, c, \underline{x}_0 = 0$

Initialization: $X = X^{(0)}$

Output: *The Summurized Frames, Which is Used from* $X \in \Re^{k\times n}$

repeat

for $j = 1, 2, \cdots, k$ *do*

$$\underline{x} = soft_{\frac{\lambda}{c}}\left[\frac{1}{c} D^T (Y - DX) + \underline{x}_0\right]$$

if $\left(\frac{\lambda}{c} < \frac{1}{c} D^T (Y - DX) + \underline{x}_0\right)$

$$\underline{x}_l = \frac{1}{c} D^T (Y - DX) + \underline{x}_0 - \frac{\lambda}{c}$$

Else if $\left(\frac{\lambda}{c} > \frac{1}{c} D^T (Y - DX) + \underline{x}_0\right)$

$$\underline{x}_l = \frac{1}{c} D^T (Y - DX) + \underline{x}_0 + \frac{\lambda}{c}$$

Else

$$\underline{x}_l = 0$$

end if

end for

$X \leftarrow X^{(k+1)}$

Until $\|X\|_{2,1} < \delta$

همان طور که در شبه‌کد الگوریتم مورد نظر نشان داده شده است، به ازای مقادیر مختلف δ، فریم‌های کلیدی به دست آمده متفاوت می‌باشند. این نکته به عنوان مقیاس‌پذیری[1] روش پیشنهادی است. هرچه مقدر δ، بزرگ‌تر باشد، تعداد فریم‌های کلیدی خروجی بیشتر و هرچه مقدار δ، کوچک‌تر باشد، تعداد فریم‌های کلیدی کمتر می‌باشند.

۳. معرفی ویژگی مبتنی بر هیستوگرام تبدیل Sensus[2] با ترکیب ویژگی رنگ و ویژگی آماری استفاده‌شده در این مقاله

ویژگی که در این مقاله مورد استفاده قرار گرفته است ترکیبی از ویژگیCENTRIST [۳۸]، رنگ [۳۹] و آماری [۴۲-۴۰] است. ویژگی‌های به دست آمده با استفاده از روش CENTRIST با قسمت‌بندی تصویر به چند بلوک و اعمال تبدیل Sensus روی آن‌ها به دست می‌آید. البته در این روش از فیلترهای مشتق‌گیر نیز برای فیلتر کردن تصویر مورد نظر استفاده می‌شود. اساس این روش برای ساختن قالبی[3] از روی تصاویر بدون در نظر گرفتن فضای رنگ آن‌ها است. این الگوریتم از ساختار هرم مکانی استفاده می‌کند.

[1] Scalable
[2] CENsus Transform hISTogram (CENTRIST)
[3] Template

در این مقاله از دو سطح مکانی استفاده شده است. در یک سطح ۵ تکه[4] از فریم ویدئویی و در سطح دیگر ۱ تکه فریم ویدئویی در نظر گرفته شده است. در این روش به هر تکه از تصاویر ۴۲ ویژگی نسبت داده می‌شود که ۴۰ ویژگی مربوط به مقادیر ویژه هر تکه است و ۲ ویژگی برای محاسبه میانگین و واریانس مقادیر ویژه مورد استفاده قرار می‌گیرد. تا این مرحله برای هر فریم ویدئویی ۲۵۲=۴۲ × (۵+۱) ویژگی استخراج شد. در مرحله بعد با توجه به اینکه روش CENTRIST مستقل از ویژگی‌های رنگ است،

در برخی گزارش‌ها [۳۶] از سه ممان مرکزی[5] برای محاسبه توزیع احتمال رنگ در هر فریم استفاده شده است. این سه ممان برابر با میانگین، واریانس و کجی[6] هستند در این مقاله علاوه بر سه ممان مرکزی که در بالا گفته شد از ممان درجه ۴ نیز استفاده می‌شود. دلیل استفاده از این ممان آن است که سه ممان قبلی قادر به تشخیص میزان کشیدگی[7] در توزیع ویژگی‌های استخراج‌شده ندارند. بنابراین به سه ممان گفته‌شده در [۴۳] ممان درجه ۴ را نیز اضافه می‌شود. می‌توان میزان کشیدگی را با استفاده از معادله (۱۲) محاسبه کرد.

لازم به ذکر است که معادله (۱۲) بر روی فضای رنگ سه‌بعدی HSI[8] محاسبه می‌شود. این سامانه رنگی، اطلاعات رنگ c را از اطلاعات شدت روشنایی آن جدا می‌کند. در این سامانه رنگ، اطلاعات رنگ در قالب پرده رنگ[9] و اشباع[10] ارائه می‌شود درحالی‌که روشنایی[11]، مقدار نور تصویر را بیان می‌کند. این سامانه رنگی توانایی بالایی در ارائه درک رنگ توسط انسان را دارد. زیرا سامانه بینایی انسان به راحتی قادر به تشخیص پرده‌های رنگ متفاوت، بدون توجه به مقدار روشنایی و اشباع است. فضای رنگ HSI با استفاده از روابط زیر از فضای رنگ RGB به دست می‌آید[۴۳].

برای محاسبه ویژگی، فریم ویدئویی را به (۴×۳) قسمت تقسیم می‌شود. بردار ویژگی ساخته شد با استفاده از ویژگی‌های رنگی و آماری برابر با ۱۴۴=۲(۴×۳) می‌شود. بردار ویژگی جدیدی با استفاده از CENTRIST و ویژگی‌هایی رنگی و آماری ساخته می‌شود که این مقدار برابر با ۳۹۶ ویژگی می‌شود. در ادامه به ارزیابی روش پیشنهادی پرداخته می‌شود:

[4] Patch
[5] Central Moments
[6] Skewness
[7] Kurtosis
[8] Hue, Saturation, Intensity (HSI)
[9] Hue
[10] Saturation
[11] Intensity

$$Kurtosis: \ k_i = \left\{\frac{N(N+1)}{(N-1)(N-2)(N-3)}\sum_{j=1}^{N}\left(\frac{p_{ij}-\bar{X}_i}{\sigma_i}\right)^4\right\} - \frac{3(N-1)^2}{(N-2)(N-3)} \quad (۱۲)$$

۴. ارزیابی روش پیشنهادی

در ابتدا به بررسی پایگاه داده‌های استفاده‌شده در این مقاله پرداخته می‌شود.

۴-۱. پایگاه داده

در این مقاله از ۲ پایگاه داده ویدئویی استفاده شده است.

پایگاه داده VSUMM: این پایگاه داده، شامل ۵۰ ویدیو انتخاب‌شده از OVP [۴۴] است و شامل صحنه‌های مختلف ویدئویی که در قسمت قبل توضیح داده شد، است [۴۵]. مدت زمان این کلیپ‌ها بین ۱ تا ۴ دقیقه است. خلاصه زمین حقیقت[1] این پایگاه داده در با استفاده از ۵۰ کاربر به دست آمده است [۴۶]. برای هر ویدیو از ۵ کاربر مختلف کمک گرفته شده است. بنابراین برای هر ویدیو ۵ مجموعه خلاصه وجود دارد. در کل به تعداد ۲۵۰ داده خلاصه‌شده ویدئویی برای ۵۰ کلیپ ویدئویی در دسترس است.

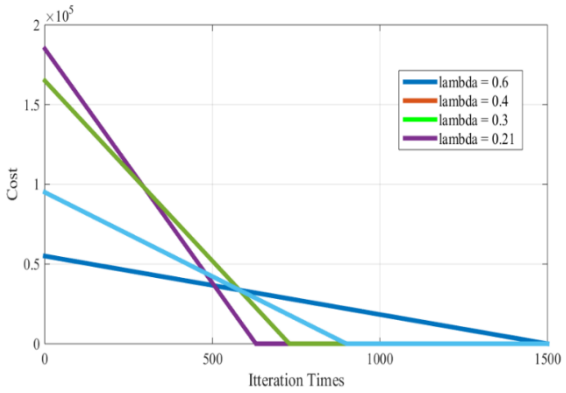
پایگاه داده OVP[2]: این پایگاه داده از مرجع [۴۳] به دست آمده است که شامل ۶ گروه کلیپ ویدئویی است که شامل: مستند[3]، آموزشی[4]، ویدیوهای روزمره[5]، تاریخی[6]، کنفرانس[7] و خدمات عمومی[8] است. برای انجام آزمایش‌ها بر روی این پایگاه داده از ۱۰ ویدیوکلیپ که در جدول (۱) از مرجع [۴۷] معرفی شده‌اند، استفاده شده است. مجموعه فریم‌های کلیدی شبیه به مرجع [۴۸]، به صورت دستی و با استفاده از ۵ دانشجو با سابقه کار در زمینه پردازش ویدیو انتخاب شده است. در ادامه به بررسی مقدار بهینه λ پرداخته می‌شود.

۴-۲. مقایسه روش پیشنهادی با روش‌های دیگر

در اولین گام به بررسی مقدار انتخابی برای پارامتر λ پرداخته می‌شود. با بررسی مقادیر مختلف λ بر روی مقادیر مختلف ویدئویی که در شکل (۱) نشان داده شده است به این نتیجه رسید که بهترین مقدار λ برای خلاصه‌سازی ویدیو $\lambda = ۰/۲۱$ در نظر گرفته می‌شود.

در ادامه به مقایسه روش پیشنهادی با روش‌های پیشنهادشده در زمینه خلاصه‌سازی ویدیو پرداخته می‌شود.

در اولین آزمایش، روش پیشنهادی را با روش DSVS[9] [۳۱] مقایسه می‌شود. این مرجع بر روی پایگاه داده VSUMM۰ برای انجام آزمایش‌های خود بهره برده است. علاوه بر روش فوق، روش پیشنهادی با چند روش دیگر مقایسه می‌شود که عبارت‌اند از روش OCFE[10] [۴۸] است که با استفاده کلاس‌بندی برخط با استفاده از ویژگی‌های استفاده‌شده در [۳۱] ویدیو را خلاصه می‌کند. همچنین، روشUCF [۴۹] که این روش در چند بخش به خلاصه‌سازی ویدیو می‌پردازد.



شکل ۱. راندمان الگوریتم پیشنهادی با استفاده از تغییر مقدار λ به صورت تجربی که محور افقی آن مربوط به تعداد تکرار الگوریتم و محور عمودی آن مربوط به تابع ارزش $\lambda\|X\|_{2,1}$ است

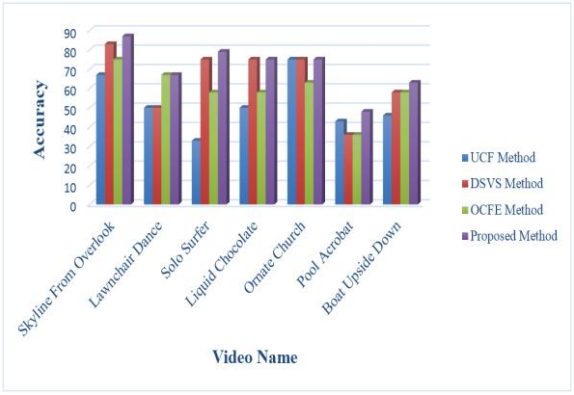
در بخش اول با استفاده از ویژگی محتوایی بر مبنای حرکت شات و شباهت رنگ بین شات‌ها به آشکارسازی شات و پیدا کردن شات‌های مشابه پرداخته است. در مرحله بعد استفاده از ساختن گراف شات‌های مشابه و تشخیص چهره به بازسازی و آشکارسازی صحنه به خلاصه‌سازی ویدیو منجر می‌شود. در این آزمایش به دلیل برابر بودن شرایط آزمایش از ویژگی مبتنی بر هیستوگرام تبدیل Sensus با ترکیب ویژگی رنگ و ویژگی آماری استفاده شده است. در ادامه به نحوه امتیازدهی پرداخته می‌شود. برای امتیاز دادن به فریم‌های خلاصه استخراج‌شده بدین گونه عمل می‌شود که اگر فریم انتخاب‌شده درست باشد به خروجی ۱ واحد و اگر اشتباه باشد صفر واحد اضافه می‌شود. البته برای تطبیق‌های ضعیف نیز مقدار ۰/۵ واحد به خروجی اضافه می‌شود.

[1] Ground Truth
[2] Open Video Project (OVP)
[3] Documentary
[4] Educational
[5] Ephemeral
[6] Historical
[7] Lecture
[8] Public Service

[9] Dictionary Selection Based Video Summarization (DSVS)
[10] Online Clustering Key Frames Extraction (OCFE)

در ادامه خروجی روش پیشنهادی با ۳ روش گفته‌شده در شکل (۲) نشان داده شده است.

به عنوان مثال نتیجه روش پیشنهادی روی مجموعه ویدئویی Liquid Chocolate که در مرجع نیز مورد بررسی قرارگرفته [۳۱] و در شکل (۳) نشان داده شده است.



شکل ۲. مقایسه روش پیشنهادی با روش‌های OCFE، UCF و DSVS برای ۱۸ فایل ویدئویی از پایگاه داده VSUMM۰

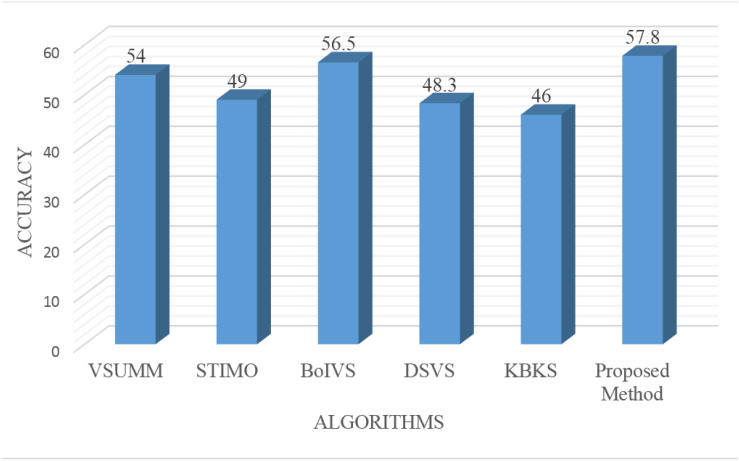
در آزمایش دوم، روش پیشنهادی با استفاده از پایگاه داده OVP و بر روی ۱۰ فایل ویدئویی صورت پذیرفته است که در جدول (۱) از مرجع [۴۹] معرفی شده‌اند. در این آزمایش به مقایسه روش پیشنهادی با روش‌های BoIVS، DT[1] [۵۰]، STIMO[2] [۵۱]، VSUMM [۴۶]، KBKS[3] [۵۲] و DSVS [۳۱] می‌پردازیم. این مقایسه بر روی تمام پایگاه داده VSUUM صورت پذیرفته است و نتیجه آن در شکل (۴) نشان داده شده است.

همان‌طورکه در شکل (۴) نشان داده شده است، روش پیشنهادی به میزان ۱,۳٪ خلاصه‌های به دست آمده از پایگاه داده VSUMM را بهبود داده است. در ادامه یک نمونه از نتایج به دست آمده برای پایگاه داده VSUMM در شکل (۵) نشان داده شده است.

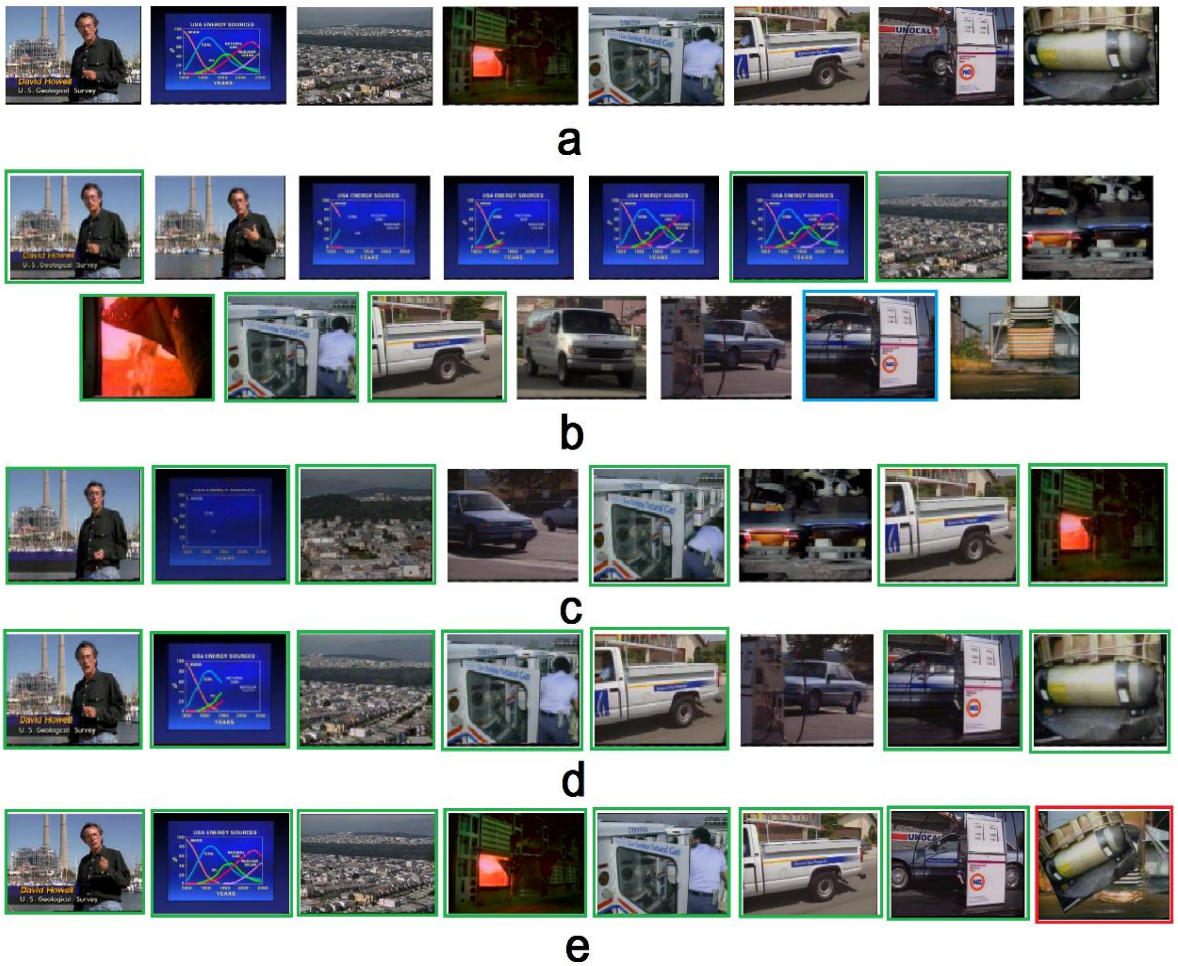
همان طور که در شکل (۵) نشان داده شده است فریم‌های خلاصه‌شده به وسیله روش پیشنهادی به فریم‌های زمین مرجع نزدیک‌تر می‌باشند. روش‌های STIMO و BoIVS نسبتاً دارای خطای بیشتری نسبت به روش VSUMM می‌باشند. ولی در مجموع روش پیشنهادی توانسته است روش‌های گفته شده را بهبود بخشد.

شکل ۳. نمونه‌ای از فریم‌های خلاصه‌شده برای نمونه ویدئویی Liquid Chocolate که سطر اول نمونه زمین مبنا، سطر دوم، سوم، چهارم و پنجم به ترتیب فریم‌های به دست آمده از OCFE، DSVS، UCF و روش پیشنهادی می‌باشند. در سطرهای دوم تا پنجم، فریم‌های استخراج‌شده‌ای که درست می‌باشند با رنگ سبز و آن دسته از فریم‌هایی که نزدیک به زمین مبنا است با رنگ قرمز مشخص شده است.

[1] STIll and MOving video storyboard (STIMO)
[1] Keypoint Based Keyframe Selection (KBKS)



شکل ۴. مقایسه عملکرد روش پیشنهادی با روش‌های VSUMM، STIMO، BoIVS، DSVS و KBKS بر روی تمامی ۵۰ فایل ویدئویی پایگاه VSUMM



شکل ۵. نمونه‌ای از فریم‌های خلاصه‌شده برای ۲ نمونه ویدئویی از پایگاه داده VSUMM که a نمونه زمین مبنا، b، c، d و e به ترتیب فریم‌های به دست آمده از STIMO ، BoIVS، VSUMM و روش پیشنهادی می‌باشند. در b تا e، فریم‌های استخراج‌شده‌ای که درست می‌باشند با رنگ سبز و آن دسته از فریم‌هایی که در نمونه زمین مبنا وجود نداشته ولی از لحاظ محتوایی به عنوان فریم خلاصه‌شده می‌توان در نظر گرفت با رنگ قرمز مشخص شده است.

[1]Delaunay Triangulation (DT)

۵. نتیجه‌گیری

با توجه به اینکه روش‌های مراقبت ویدئویی، جزو روش‌های پر استفاده در حوزه مراقبت و آشکارسازی است. در این مقاله با استفاده از روش‌های جدید در حوزه پردازش ویدیو به بهبود عملکرد این سامانه‌ها پرداخته شد. بنابراین با استفاده از انتخاب واژه‌نامه تنک با الگوریتم انقباضی آستانه نرم می‌توان میزان خلاصه‌سازی ویدیو را بهبود بخشید. بدین منظور ابتدا کدینگ تنک را برای خلاصه‌سازی ویدیو فرموله شد و پس از آن با توجه به آنچه از حل معادلات کلاسیک به دست آمد می‌توان الگوریتمی برای خلاصه‌سازی با استفاده از این روش پیشنهاد داد. برخلاف عملیات ریاضی در حل معادلات تنک؛ پس از به دست آمدن عبارت نهایی، مشاهده می‌شود که دارای پیچیدگی کمی نسبت به روش‌های دیگر بر پایه تنک است. در انتها، روش پیشنهادی به صورت یک الگوریتم نمایش داده شده است که منجر به سادگی پیاده‌سازی آن می‌شود. پس از اجرای الگوریتم بر روی ۲ پایگاه داده OVP، VSUMM و مقایسه فریم‌های کلیدی به دست آمده از روش پیشنهادی و روش‌های STIMO، VSUMM، DSVS، BoIVS و KBKS می توان میزان خلاصه‌سازی را بهبود بخشید که این امر ادعای فرض شده در این مقاله، مبنی بر بهبود روش پیشنهادی نسبت به دیگر روش‌ها را نشان می‌دهد.

۶. مراجع

[1] Weiming, H.; Tieniu, T.; Iang, W.; Steve, M. "A Survey on Visual Surveillance of Object Motion and Behaviors"; IEEE Trans. Syst. Cybernetics 2004, 3, 334-352.

[2] Hong, L.; Ruan, Y.; Wicker, D.; Layne, J. "Energy-Based Video Tracking Using Joint Target Density Processing with an Application to Unmanned Aerial Vehicle Surveillance"; IET Image Processing 2008, 2, 1-12.

[3] Hampapur, A.; Connell, J.; Haas, N.; Merkl, H.; Shu, C. F. "Smart Video Surveillance"; IEEE Signal Magazin 2005, 38-51.

[4] Conte, G.; Dohetry, P. "Vision-based Unmanned Aerial Vehicle Navigation Using Geo-Referenced Information"; J. on Advances in Signal Processing 2009, 10, 1-18.

[5] Carr, J. R.; Sobek, J. S. "Digital Scene Matching Area Correlator (DSMAC)"; Proc. Conf. Image Processing for Missile Guidance, 1980, 36.

[6] Tsai, S. X. "Introduction to the Scene Matching Missile Guidance Technologies"; Guided Missile, 1996.

[7] Liu, T.; Zhang, H. J.; Qi, F. "A Novel Video Key Frame Extraction Algorithm Based on Perceived Motion-Energy Model"; IEEE Trans. on Circuits and Syst. for Video Tech. 2003, 13, 1006-1013.

[8] Taskiran, C. M.; Pizlo, Z.; Amir, A.; Ponceleon, D.; Delp, E. "Automated Video Program Summarization Using Speech Transcripts"; IEEE Trans. on Multimedia 2006, 8, 775-791.

[9] Ma, Y. F.; Lu, L.; Zhang, H. J.; Li, M. "A User Attention Model for Video Summarization"; ACM Multimedia'02 2002, 533-542.

[10] Truong, B.; Venkatesh, S. "Video Abstraction: A Systematic Review and Classification"; ACM Trans. on Multimedia Computing, Communication and Applications 2007, 3, 1–37.

[11] Hu, W.; Xie, N.; Zeng, X.; Maybank, S. "A Survey on Visual Content-Based Video Indexing and Retrieval"; IEEE Trans. Syst., Man, and Cyber. 2011, 41, 797–819.

[12] Tsai, D. M.; Lai, S. C. "Independent Component Analysis-Based Background Subtraction for Indoor Surveillance"; IEEE Trans. Image Process 2009, 18, 158–167.

[13] Xiang, T.; Gong, S. "Video Behavior Profiling for Anomaly Detection"; IEEE Trans. Pattern Anal. Mach. Intell. 2008, 30, 5, 893–908.

[14] Li, Y.; Zhang, T.; Tretter, D. "An Overview of Video Abstraction Techniques"; Technical Report HPL-2001-191, 2001.

[15] Ciocca, G.; Schettini, R. "Innovative Algorithm for Key Frame Extraction in Video Summarization"; J. Real-Time Image Process 2006, 1, 1, 69–88.

[16] Massimiliano, A. "Extracting and Summarizing Information from Large Data Repositories"; Ph.D. Dissertation, University of Naples Federico II, Italia, 2006.

[17] Money, A. G.; Agius, H. "Video Summarization: A Conceptual Framework and Survey of the State of the Art"; J. Vis. Com. Image Rep. 2008, 19, 121–143.

[18] Li, Y.; Lee, S. H.; Yeh, C. H.; Kuo, C. C. "Techniques for Movie Content Analysis and Skimming: Tutorial and Overview on Video Abstraction Techniques"; IEEE Signal Process. Magazine 2006, 23, 79–89.

[19] Kleban, J.; Sarkar, A.; Moxley, E.; Mangiat, S.; Joshi, S.; Kuo, T.; Manjunath, B. S. "Feature Fusion and Redundancy Pruning for Rush Video Summarization"; Proc. of the ACM Int. Workshop on Video Summarization, 2007, 84–88.

[20] Pan, C. M.; Chuang, Y. Y.; Hsu, W. H. "Fast Rushes Summarization System"; Proc. of the ACM Int. Workshop on Video Summarization, 2007, 74–78.

[21] Doulamis,N. D.; Doulamis, A. D.; Avrithis, Y. S.; Ntalianis, K. S.; Kollias, S. D. "Efficient Summarization of Stereoscopic Video Sequences"; IEEE Trans. Circuits Syst. Video Tech. 2000, 10, 501–517.

[22] Le, D. D.; Satoh, S. "National Institute of Informatics"; Proc. of the ACM Int. Workshop on Video Summarization, 2007, 70–73.

[23] Nesterov, Y. "Gradient Methods for Minimizing Composite Objective Function"; Louvain-la-Neuve, Belgium, CORE, 2007.

[24] Cong, Y.; Yuan, J.; Luo, J. "Towards Scalable Summarization of Consumer Videos *via* Sparse Dictionary Selection"; IEEE Trans. Multimedia 2012, 14, 66–75.

[25] Hanjalic, A.; Langendijk, R. L.; Biemond, J. "A New Key Frame Allocation Method for Representing Stored Video Streams"; 1st Int. Workshop on Image Databases & Multi, Search, Amsterdam, 1996, 67–74.

[26] Fu, Y.; Guo, Y.; Zhu, Y.; Liu, F.; Song, Ch.; Zhou, Zh. "Multi View Video Summarization"; IEEE Trans. Multimedia. 2010, 12, 717–729.

[27] Zhuang, Y.; Rui, Y.; Huang, T.; Mehrotra, S. "Adaptive Key Frame Extraction Using Unsupervised Clustering"; Proc. Int. Conf. Image Process, 1998, 866–870.

[28] Cernekova, Z.; Pitas, I.; Nikou, C. "Information Theory-Based Shot Cut/Fade Detection and Video Summarization"; IEEE Trans. Circuits & Sys for Video Tech. 2006, 16, 82–91.

[29] Oliva, A.; Torralba, A. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope"; Int. J. Comput. Vis. 2001, 42, 145–175.

[30] Wu, J.; Christensen, H.; Rehg, J. "Visual Place Categorization: Problem, Dataset, and Algorithm"; Proc. IROS, 2009.

[31] Wu, J.; Rehg, J. "Centrist: A Visual Descriptor for Scene Categorization"; IEEE Trans. Pattern Anal. Mach. Intell. 2010, 33, 1489–1501.

[32] Gallager, R. G. "Information Theory and Reliable Communication"; John Wiley and Sons, New York, 1968.

[33] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso"; J. Roy. Statist. Soc. Series B (Methodological), 1996, 58, 267–288.

[34] Yuan, M.; Lin, Y. "Model Selection and Estimation in Regression with Grouped Variables"; J. Roy. Statist. Soc. 2006, 68, 49–67.

[35] Cong, Y.; Yuan, J.; Liu, J. "Sparse Reconstruction Cost for Abnormal Event Detection"; Proc. IEEE Conf. Com. Vision & Pattern Recognition (CVPR), 2011, 3449–3456.

[36] Fornasier, M.; Rauhut, H. "Iterative Thresholding Algorithms"; Science Direct, Applied and Computational Harmonic Analysis 2008, 25, 187–208.

[37] Selesnick, I. "A Derivation of the Soft-Thresholding Function"; Ph.D. Thesis, Polytechnic Institute of New York University, 2009.

[38] Wu, J.; Christensen, H.; Rehg, J. "Visual Place Categorization: Problem, Dataset, and Algorithm"; Proc. Intelligent Robots and Systems, 2009, 4763–47760.

[39] Stricker, M.; Orengo, M. "Similarity of Color Images"; Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995, 2420, 381–392.

[40] Rayner, J. C. W.; Best, D. J.; Matthews, K. L. "Interpreting the Skewness Coefficient"; Communications in Statistics Theory and Methods 1995, 24, 593–600.

[41] Groeneveld, R. A.; Meeden, G. "Measuring Skewness and Kurtosis"; J. Roy. Statist. Soc. 1984, 33, 391–399.

[42] Arnold, B. C.; Groeneveld, R. A. "Measuring Skewness with Respect to the Mode"; Am. Statistician J. 1995, 49, 34–38.

[43] Gonzalez, C; Woods, R. E. "Digital Image Processing"; Prentice-Hall, Second Edition, 2002.

[44] "The Open Video Project"; http://www.open-video.org.

[45] "VSUMM Database"; https://sites.google.com/site/vsummsite/results.

[46] Avila, S. E. F.; Lopes, A. P. B.; Daluz, A.; Araújo, A. "Vsumm: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method"; Pattern Rec. Let. 2011, 32, 56–68.

[47] Lu, S.; Wnag, Z.; Mei, T.; Guan, G.; Feng, D. D. " A Bag of Important Model with Locality Constrained Coding Based Feature Learning for Video Summarization"; IEEE Trans, Multimedia 2014, 16, 1497–1509.

[48] Luo, J.; Papin, C.; Costello, K. "Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers"; IEEE Trans. Circuits Syst. Video Tech. 2009, 19, 289–301.

[49] Rasheed, Z.; Shah, M. "Detection and Representation of Scenes in Videos"; IEEE Trans. Multimedia. 2005, 7, 1097–1105.

[50] Mundur, P.; Rao, Y.; Yesha, Y. "Keyframe Based Video Summarization Using Delaunay Clustering"; Int. J. Digit. Libraries 2006, 6, 2, 219–232.

[51] Furini, M.; Geraci, F.; Montangero, M.; Pellegrini, M. "Stimo: Still and Moving Video Storyboard for the Web Scenario"; Multimedia Tools Appl. 2010, 46, 47–69.

[52] Guan, G.; Wang, Z.; Lu, S.; Deng, J. D.; Feng, D. "Keypoint Based Keyframe Selection"; IEEE Trans. Circuits Syst. Video Tech. 2013, 23, 729–734.