3rd. International Conference on September 2017
Information Technology, Computer and Telecommunications
سومین کنفرانس بین المللی شهریورماه ۱۳۹۶
فناوری اطلاعات، مهندسی کامپیوتر و مخابرات

افزایش میزان صحت خلاصه‌سازی ویدیویی با روش ترکیبی صوتی تصویری

[۱] بهاره اسپهبدی، [۲] عبدالکریم الهی

[۱] دانشجوی دانشگاه آزاد اسلامی، واحد بهشهر، گروه کامپیوتر، مازندران، ایران، Bahareh.espahbodi@gmail.com
[۲] عضو هیئت علمی دانشگاه آزاد اسلامی، واحد بهشهر، گروه کامپیوتر، مازندران، ایران،Ak_elahi@yahoo.com

چکیده

در دنیایی زندگی می‌کنیم که وجود دوربین‌های خانگی و قدرت رسانه باعث شده تا با حجم خیره کننده‌ای از داده‌های ویدیویی سر و کار داشته باشیم. این امر بر اهمیت نحوه‌ی دسترسی و پردازش این نوع داده‌ها می‌افزاید. با خلاصه‌سازی ویدیویی، فیلم به یکسری فریم یا کلیپ کوتاه ولی با معنی خلاصه می‌گردد که اهداف فرد را نیز برآورده می‌سازد. در این پژوهش سعی شده در ابتدا داده با کمک الگوریتم(K-Medoids) خوشه‌بندی شود سپس با تشخیص قسمت‌های صوتی مهم در ویدیو و جدا کردن آن از نویزها و بخش‌های کم اهمیت، اطلاعات انطباقی میان صوت و تصویر در رویدادهای مهم به دست آید و میزان تکرار آن‌ها نیز مشخص گردد. این راهکار را بر روی ۵۰ ویدیو از پایگاه داده Open CV تکرار شده است و بطور میانگین ۸۸٪ نرخ صحت در خلاصه‌سازی و ۳۱٪ میزان خطا به دست آمده که به نسبت سایر روش‌ها جزء بالاترین نرخ صحت می‌باشد.

واژه‌های کلیدی : کاوش ویدیویی، خلاصه‌سازی ویدیویی، خوشه‌بندی، کا-میدویدز، انطباق صوتی تصویری

۱- مقدمه

امروزه وجود دوربین‌های فراوان که در دسترس همگان قرار دارد و در کنار آن قدرتمند شدن سیستم‌های ویدیویی، منجر به این شده تا افراد بتوانند در موقعیت‌های مختلف و در لحظه، فیلمبرداری کنند. به دلیل رشد ویدئوهای دیجیتالی در دسترس با سرعت نمایی، کاربران برای دسترسی به ویدئوهای دیجیتال نیاز به کمک دارند[۱]. پژوهش درباره‌ی خلاصه ویدیویی کمک می‌کند تا به این نیازها پاسخ داده شود.

مدت کمی است که به اهمیت کاوش داده‌های ویدیویی[۱] پی برده‌اند و همه روزه جنبه‌ی جدیدی از مزایای آن مشخص می‌گردد یا آنکه روشی نو برای استفاده از آن ارائه می‌شود[۲]. می‌توان با یک کاوش خوب به اطلاعات بالایی رسید و از آنها در ساده‌ترین امور تا مهم‌ترین و پیچیده‌ترین اتفاقات بهره برد. در این بین خلاصه‌سازی ویدیو[۲] مدتی است که مورد توجه قرار گرفته است. در بسیاری از زمینه‌های کاری این‌که بتوان داده‌ی ویدیویی طولانی مدت را به چندین عکس یا یک فیلم کوتاه(به طوری‌که مفهوم و موضوع اصلی فیلم حذف نشود)، خلاصه کرد اهمیت بسیار دارد و به این دلیل که ویدیو در دنیای ما همه‌گیر شده است نیاز به این سیستم بیش از پیش حس می‌گردد.

در زمینه‌ی کاوش ویدیویی دهه‌ای است که کارهای متفاوتی انجام شده، این علم جنبه‌های گوناگونی دارد؛ یکی از مهم‌ترین آنها که در این مقاله نیز مورد توجه قرار گرفته است کاوش و تشخیص رویداد است. محققین بسیاری در این زمینه کارکرده‌اند. شناسایی رویداد در ویدیوها، فرآیندی سخت و پیچیده است؛ در بیشتر اوقات نرم‌افزارهایی که حرکت را شناسایی می‌کنند در تشخیص درست یک رویداد دچار اشتباه می‌شوند. از کاوش رویداد می‌توان در زمینه‌های مختلفی استفاده کرد، از مساله‌ای بحرانی همچون تشخیص بمب‌گذاری و تشخیص عملیات تروریستی تا ساده‌ترین مسائل همچون خلاصه‌سازی یک فیلم سینمایی به صحنه‌های مهم آن. راهکارهای ارائه شده همگی سعی کردند تا صحت را تا میزان مطلوبی بالا برند اما این راهکارها نتوانستند همزمان هم به ویژگی‌های صوتی و ویژگی‌های تصویری توجه کنند و سازگاری میان این دو را در نظر نگرفتند. در حالی‌که با کمک گرفتن از این ایده می‌توان هم بار محاسباتی را کاهش داد و هم به نتایج مطلوب‌تر و دقیق‌تری رسید.

۲- خلاصه‌سازی ویدیویی

خلاصه‌سازی ویدیویی یکی از مهم‌ترین عناوین موجود در بحث کاوش ویدیویی است که جستجوی داده‌های عظیم را سریع‌تر نموده و همچنین فهرست‌بندی محتوایی و دستیابی به داده را سریع‌تر و البته راحت‌تر می‌سازد[۳]. خلاصه‌سازی می‌تواند از مجموع فریم‌های کلیدی و یا شات‌ها تشکیل شود که این بسته به شرایط متغیر است

خلاصه‌سازی ویدیو به ایجاد یک خلاصه از ویدیوی دیجیتالی اشاره دارد که باید سه اصل را پوشش دهد:

---

[1] Video Mining

[2] video Summarization

- شامل موجودیت‌ها[3] و رخدادهای[4] با اولویت بالا از ویدئو باشد.
- خلاصه، خود باید یک درجه منطقی[5] از پیوستگی را نمایش دهد.
- خلاصه، باید عاری از تکرار باشد.

خلاصه‌سازی کاربردهای فراوان دارد همچون:

- خلاصه‌سازی اخبار[۳]
- خلاصه‌سازی فیلم‌های سینمایی[۴-۶]
- تشخیص فعالیت در ویدیو[۷]
- تشخیص رویداد(مانند بمب‌گذاری، گل در فوتبال، رویداد معمول و غیرمعمول و .... در ویدیو[۸-۱۰]
- و غیره

۲ نوع مختلف از خلاصه ویدیویی وجود دارد:

- خلاصه‌سازی ویدیویی: خلاصه‌های تصاویر ساکن، نمایش تصاویر برجسته یا فریم‌های کلیدی [۱۱] .
- مختصر نمودن فیلم[6]: خلاصه‌های تصاویر متحرک، یک دنباله از سکانس‌ها(تکه‌های ویدئو) است[۱۲].

۲-۱- تعاریف اولیه

برای داده‌های ویدیویی و استفاده از آنها واژگان فنی خاصی وجود دارد؛ زیو زایانگ[7] و همکاران در کتاب خود این واژه‌ها را معرفی نمودند که به ترتیب زیر است[۱۳]:

- شات[8] ویدیویی: سکانس متوالی از فریم‌ها که با یک تک دوربین فیلم‌برداری شده‌اند. در واقع بلوک سازنده‌ی جریان‌های ویدیویی است.
- فریم کلیدی[9]: فریمی است که بیانگر قسمت برجسته‌ی ویدیویی از یک شات است.
- صحنه‌ی ویدیویی[10]: گروهی از شات‌ها که از لحاظ معنایی مرتبط می‌باشند و در واقع شات‌های کنار هم که یک داستان یا مفهوم سطح بالا را بیان می‌کنند.
- گروه ویدیویی: یک موجودیت میانی بین شات‌های فیزیکی و صحنه‌های معنایی بوده و همچون پل رابطی بین این دو عمل می‌کند.
- نشانه‌گذار صوتی: سکانس‌های پیوسته از فریم‌های ویدیویی که بیانگر یک کلاس کلیدی صوتی و بیانگر یک رویداد مورد علاقه در ویدیو است.
- نشانه‌گذار ویدیویی: سکانس‌های متوالی از فریم ویدیویی است که نشان‌دهنده‌ی یک رویداد مورد علاقه در ویدیو است.
- کاندیدهای برجسته[11]: قطعه‌ی ویدیویی است که اهمیت دارد و می‌توان با نشانه‌گذار ویدیویی یا صوتی تعیین گردد.

۳- مرور ادبیاتی

در زمینه کاوش ویدیویی و به خصوص خلاصه‌سازی ویدیویی کارهای تحقیقاتی زیادی انجام گرفته است. هوآنگ و همکاران[12] در سال ۲۰۱۲ سعی کردند تا با شناسایی اشیا و موجودات متحرک در ویدیو رویداد را تشخیص داده و نه تنها خلاصه‌سازی ویدیو را انجام دهند بلکه کارایی آن را تا میزان مطلوبی بالا برند[۱۴]. سنگدو[13] و همکاران در سال ۲۰۱۴ نیز روی شناسایی رویداد در ویدیو کار کرده‌اند و تفاوت روش آن‌ها با سایر روش-

---

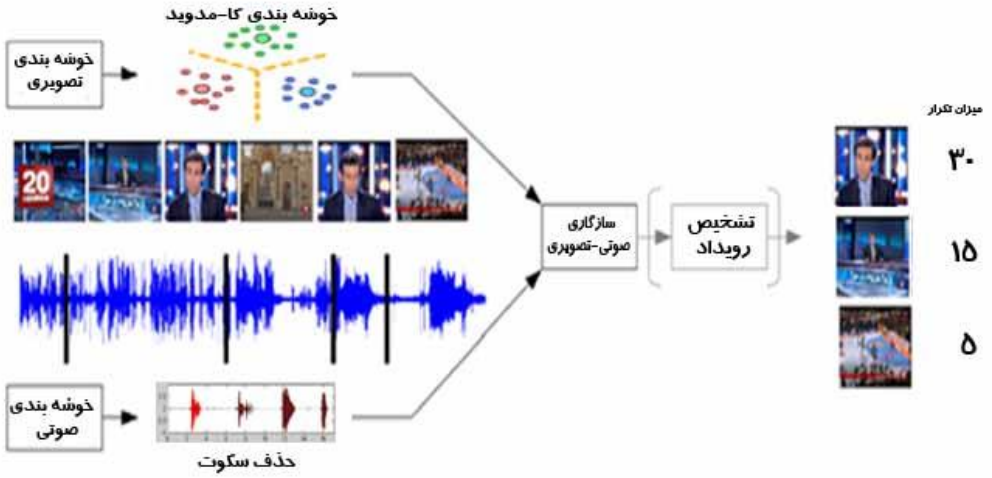
[3] Entities
[4] Events
[5] Logical
[6] Video skimming
[7] Ziyou Xiong
[8] Shot
[9] Key Frame
[10] Video scene
[11] Highlight candidate
[12] Hoang Trinh
[13] Sangdoo Yun

ها در این است که هم به ویژگی‌های ایستای موجودیت‌ها در ویدیو توجه کردند و هم به ویژگی‌های حرکتی آن و بر این اساس خلاصه‌سازی ویدیویی انجام گرفت[۱۵].

راجات[14] و همکاران در سال ۲۰۱۶ به خلاصه‌سازی رویدادها در ویدیو پرداختند، در روش آن‌ها ویدیو به تمام اجزای موجود در محتویات آن تقسیم شد، شامل متن، صدا، صورت‌ها و غیره و سپس بر این اساس خلاصه‌سازی ویدیو انجام گرفت[۱۶]. دیمو[15] و همکاران در سال ۲۰۱۵ با کمک روش کاربر محور سعی در خلاصه‌سازی ویدیو داشتند. در این روش از اطلاعات سطح بالای فریم‌های حذف شده و اطلاعات سطح پایین ویدیو استفاده می‌شود تا خلاصه‌سازی با معنای بهتر ارائه شود[۱۷]. توسلی‌پور و همکاران در سال ۲۰۱۳ با کمک شبکه بیزین، ویدیوهای ورزشی را خلاصه‌سازی نمودند. روش آن‌ها براساس شات و کلاسه‌بندی شات‌ها[16] بود. در سه گام و با کمک مدل مخفی مارکوفی خلاصه‌سازی ویدیو انجام گرفت[۱۸].

۴- راهکار پیشنهادی

در این راهکار با کمک الگوریتم کا-میدویدز[17] در ابتدا فریم‌های ویدیو به فریم‌های کلیدی که نشان‌دهنده‌ی فعالیت خاصی هستند، تقسیم می‌شوند. جهت بهبود روش کا-میدویدز تلاش شده است تا نحوه انتخاب k با دقت بالاتری انجام گیرد به همین جهت در انتخاب k سعی شده تا با جستجوی سریع در میان فریم‌های موجود، تعداد فریم‌هایی که از لحاظ ویژگی تصویری از یک میزان آستانه، بیشتر تفاوت دارند، شمرده شود؛ این میزان برابر با تعداد K می‌باشد. استفاده از این خوشه‌بندی قبل از به دست آوردن سازگاری باعث بالا رفتن کارایی، حذف نویز و بالا بردن دقت می‌شود. در ادامه برای این‌که حالت مشابهی درباره صوت داشته باشیم، نویز در صوت حذف شده و با کمک خوشه‌بندی صوتی تنها چانک‌های کلیدی باقی می‌مانند. سپس با بهره‌گیری از الگوریتم درختی دو دندوگرام صوتی و ویدویی از روی خوشه‌های موجود حاصل می‌گردد. بعد از آن، با کمک مقیاسی به نام اطلاعات انطباقی[18] سازگاری میان این دو سنجیده می‌شود. موارد با میزان اطلاعات انطباقی بالاتر سازگارتر بوده و بیانگر یک رویداد می‌باشند. با کمک یک شمارشگر تعداد رویدادهای مشابه و تکرار هر مورد مشخص می‌گردد. در آخر فریم‌های کلیدی که نماینده رویدادها می‌باشند، ارائه می‌گردند. نتیجه فریم‌های خلاصه شده با دقت بالا هستند.



شکل ۱: قالب کاری روش پیشنهادی

در این راهکار برای خوشه‌بندی صوتی از انرژی سیگنال و مرکز طیفی استفاده شده است که به ترتیب در فرمول‌های (۱) و (۲) آورده شده‌اند.

14 Rajat Aggarwal
15 Anastasios Dimou
16 Shot
17 K-Medoid
18 Mutual Information

- ***انرژی سیگنال***: از این ویژگی برای تشخیص سکوت در صوت استفاده می‌شود. علاوه بر آن بین کلاس‌های صوتی نیز تفاوت قائل می‌شود. برای هر فریم انرژی سیگنال از فرمول (۱) قابل محاسبه است.

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|x_i(n)|^2 \qquad (۱)$$

x_i که $(i=1,\ldots,N)$ نشانگر نمونه‌ی صوتی از iامین فریم با طول N است.

- ***مرکز طیفی***[۱۹]: مرکز طیفی iامین فریم که با C_i نشان داده می‌شود همان مرکز ثقل یک طیف است.

بنا بر فرمول ۲ $X_{i(k)}$ که $i=1,..,N$ می‌باشد، برابر است با ضریب تبدیل گسسته فوریه[۲۰] از iامین کوچکترین فریم که N برابر با طول فریم است. این میزان برای صداهای مختلف و موقعیت‌های گوناگون به شدت متغیر است.

$$C_i = \frac{\sum_{k=1}^{N}(k+1)x_{t(k)}}{\sum_{k=1}^{N}x_{t(k)}} \qquad (۲)$$

قطعه‌بندی و خوشه‌بندی به طور مستقل روی کیفیت‌های صوتی بصری انجام می‌شوند و در نتیجه دو دندوگرام ایجاد می‌گردد که هر گره در آن برابر با دسته‌ای از قطعه‌ها در کیفیت مشابه است. پیدا کردن یک رویداد تکراری با شباهت بالا هم در محتویات تصویری و هم صوتی به میزان همبستگی صوتی تصویری بین یک خوشه از قطعه‌های صوتی و خوشه‌ای از شات‌های ویدیویی بستگی دارد. همبستگی صوتی بصری با استفاده از "اطلاعات متقابل[۲۱]" که از این به بعد با نام MI قید می‌شود به صورت تصادفی به دست می‌آید. به طور واضح مشخص است که دو خوشه همبسته حداقل در اطلاعات مشترک هستند. C_i^x نشانگر iامین خوشه برای کیفیت $x \in \{A, V\}$ است، که با iامین گره از دندوگرام مشخص می‌گردند. MI بین C_i^A و C_j^V با فرمول (۳) مشخص می‌گردد.

$$\mathrm{MI}(C_i^A, C_j^V) = \sum_{(a,v)\in\{(0,0),(1,1)\}} p(a,v)\,ln\left(\frac{p(a,v)}{p(a)p(v)}\right) \qquad (۳)$$

a و v، متغیرهای تصادفی باینری هستند که به ترتیب نشانگر عضویت در C_i^A و C_j^V هستند. در عمل احتمالات $p(a,v)$ ، $p(a)$ و $p(v)$ از قطعه‌بندی موقتی تخمین زده می‌شوند. برای مثال، احتمال $p(a=1, v=1)$ برابر است با مجموع مقدار زمان‌هایی که هر قطعه از C_i^A با یک قطعه C_j^V همزمان اتفاق می‌افتد و با کل زمان ویدیو نرمالیزه می‌شود که با زمان کلی از ویدیو، نرمالیزه می شود. به زبان دیگر معادله‌ی(۳) نشان می‌دهد که تا چه میزانی قطعه‌ها در خوشه صوتی C_i^A همزمان با شات‌های خوشه C_j^V رخ می‌دهند؛ بعلاوه از آنجا که خوشه‌ها از دندوگرام‌ها انتخاب می‌شوند مسلماً تا حدودی همگن هستند. دو خوشه با MI یکسان، نشان می‌دهند که قطعه‌ها همگرایی سمعی بصری دارند. استفاده از مقیاس اطلاعات کیفیتی ناخواسته، چندین مزیت فوق‌العاده نسبت به روش‌های خوشه‌بندی معمولی دارد. اولین مورد، شناسایی بهترین خوشه است. در نظریه‌های معمولی خوشه‌بندی با کمک چند ضابطه اخباری که در مورد نحوه خوشه‌بندی تصمیم می‌گیرند، انجام می‌شود. بعلاوه در میان نتایج خوشه‌ها، هر فرد باید آنهایی که بسته به نیاز بهترین جواب را می‌دهد، پیدا کند. در راهکار پیشنهادی آنپهونگ تا و همکاران با اندازه‌گیری MI بین تمام جفت خوشه‌های صوتی بصری، از این دو مشکل جلوگیری می‌کند[۱۹]. مزیت دیگر نظریه کیفیتی ناخواسته این است که تمام جفت خوشه‌های ممکن می‌توانند از سازگاری بالاتر به سازگاری پایین‌تر رتبه‌بندی شوند، بنابراین چندین رویداد کاندید فراهم می‌آید. اینکه چندین جواب دارند آنها را قادر می‌سازد تا چندین رویداد خصیصه‌بندی شده را کشف کنند؛ به طور مثال هر مهمان در یک برنامه‌ی زنده را تشخیص داده و افراد نامربوط را با استفاده از قوانین اکتشافی شناسایی کند.

۵- ارزیابی راهکار پیشنهادی

بر خلاف روش‌های مشابه کاوش و اغلب روش‌های تحقیق که معیاری برای ارزیابی وجود دارد تا بتوان با این ارزیابی در مورد درجه خوبی یک روش و کیفیت آن تصمیم گرفت، در امر خلاصه‌سازی ویدیویی هنوز استانداردی برای سنجش و ارزیابی نتیجه، وجود ندارد. به دلایل مختلفی چون سلیقه افراد، تخصصی بودن خلاصه در رشته‌های متفاوت و غیره نمی توان نظر درستی در مورد خوبی یا بدی یک خلاصه داد و در نتیجه نمی‌توان به راحتی عمل ارزیابی را انجام رساند.

در اینجا سعی شده تا یکی از راهکارهای تقریبا جدید ارزیابی خلاصه‌سازی به نام کاس استفاده شود که در واقع تغییر در روش f-measure و تبدیل آن به روشی کاربر محور می‌باشد[۱۹]. این روش نتیجه خلاصه‌سازی خودکار را با نظر کاربران می‌سنجد و ارزیابی مناسبی با نظر مستقیم

[19] Spectral Centroid

[20] Discrete Fourier Transform (DFT)

[21] Mutual Information

کاربران انجام می‌دهد. در ابتدا از چندین کاربر خواهش می‌شود تا فیلم را ببینند و بعد به صورت دستی بر طبق میل خود خلاصه‌ای ایستا از آن فیلم تهیه کنند. برای راحتی کار فریم‌های نمونه در اختیارشان قرار می‌گیرد تا از میان آن‌ها خلاصه‌ها را پیدا کنند. کاربر در انتخاب فریم‌ها و تعداد آن‌ها کاملا مختار و آزاد است. در گام دوم این فریم‌های انتخاب شده با فریم‌های خلاصه‌سازی خودکار مورد مقایسه قرار می‌گیرند. در گام سوم کیفیت این فریم‌ها با کمک دو مقیاس CUS_A و CUS_E سنجیده می‌شود.

$$CUS_A = \frac{n_{mAS}}{n_{Us}} \quad (۴)$$

$$CUS_E = \frac{n_{\bar{m}AS}}{n_{Us}} \quad (۵)$$

که n_{mAS} برابرست با تعداد فریم‌های کلیدی مشابه بین خلاصه‌ی خودکار(AS) و خلاصه‌ی کاربران که دقیقا برعکس $n_{\bar{m}AS}$ یعنی تعداد فریم-های کلیدی که در این دو با هم برابر نیستند می‌باشد. n_{Us} نیز تعداد فریم‌های موجود در خلاصه کاربر (US) می‌باشد.

۶- ارزیابی روش پیشنهادی

نتایج حاصل از راهکار پیشنهادی توسط روش کاس بر روی ۵۰ ویدیو از سایت Open CV مورد بررسی قرار گرفت.
نرخ خطا و نرخ صحت به دست آمده در ابتدا با ۵ روش دیگر که آن‌ها نیز روی ویدیوهای Open CV اعمال شده‌اند، مقایسه می‌گردد که نتیجه این مقایسه در جدول ۱ مشخص است.

جدول ۱: مقادیر ارزشیابی روش‌های موجود[۱۹]

	OV	DT	STIMO	$VSUMM_1$	$VSUMM_2$	راهکار پیشنهادی
CUS_A	۰,۷۰	۰,۵۳	۰,۷۲	۰,۸۵	۰,۷۰	۰,۸۸
CUS_E	۰,۵۷	۰,۲۹	۰,۵۸	۰,۳۸	۰,۲۷	۰,۳۱

مقایسه‌ی دو به دوی راهکارهای پیشنهادی با هر یک از ۵ روش با بازه اطمینان ۹۸٪ انجام گرفته است. که به نوعی برتری راهکار پیشنهادی را نشان می‌دهد.
مطابق جدول ۱، روش پیشنهادی دارای بالاترین میزان نرخ صحت نتایج است و نرخ خطای نسبتاً پایینی نیز دارد که این بیانگر برتری این روش می‌باشد.
در ادامه جداول ۲ تا ۵، مقایسه‌ی دو به دوی نتایج حاصل از راهکار پیشنهادی که توسط روش کاس به دست آمده است با هر یک از ۵ روش و در بازه اطمینان ۹۸٪ و ۹۵٪ انجام گرفته است و به نوعی برتری راهکار پیشنهادی را نشان می‌دهد.

جدول ۲: تفاوت میان **CUS_A** در سطح اطمینان ۹۸٪ روش‌ها با روش پیشنهادی

تفاوت‌ها	بازه اطمینان(۹۸٪)	
	مینیمم	ماکزیمم
روش پیشنهادی- $VSUMM_1$	0.01	0.03
روش پیشنهادی- $VSUMM_2$	0.16	0.18
روش پیشنهادی-OV	0.12	0.21
روش پیشنهادی-DT	0.3	0.38
روش پیشنهادی-STIMO	0.13	0.18

جدول ۳: تفاوت میان CUS_E در سطح اطمینان ۹۸٪ روش‌ها با روش پیشنهادی

بازه اطمینان(۹۸٪)		تفاوت‌ها
ماکزیمم	مینیمم	
-0.01	-0.15	روش پیشنهادی -$VSUMM_1$
0.10	-0.01	روش پیشنهادی-$VSUMM_2$
-0.1	-0.42	روش پیشنهادی-OV
0.07	-0.03	روش پیشنهادی-DT
-0.18	-0.4	روش پیشنهادی-STIMO

جدول ۴: تفاوت میان CUS_A در سطح اطمینان ۹۵٪ روش‌ها با روش پیشنهادی

بازه اطمینان(۹۵٪)		تفاوت‌ها
ماکزیمم	ماکزیمم	
0.03	0.01	روش پیشنهادی -$VSUMM_1$
0.18	0.16	روش پیشنهادی-$VSUMM_2$
0.20	0.13	روش پیشنهادی-OV
0.38	0.30	روش پیشنهادی-DT
0.18	0.13	روش پیشنهادی-STIMO

جدول ۵: تفاوت میان CUS_E در سطح اطمینان ۹۵٪ روش‌ها با روش پیشنهادی

بازه اطمینان(۹۵٪)		تفاوت‌ها
ماکزیمم	ماکزیمم	
-0.01	-0.14	روش پیشنهادی -$VSUMM_1$
0.09	-0.01	روش پیشنهادی-$VSUMM_2$
-0.13	-0.4	روش پیشنهادی-OV
0.07	-0.02	روش پیشنهادی-DT
-0.2	-0.35	روش پیشنهادی-STIMO

بر طبق قاعده، در سطوح اطمینان ۹۵٪ و ۹۸٪ در صورتیکه ارزش مقادیر برابر با صفر باشد نشان‌دهنده‌ی عدم کیفیت می‌باشد. در هیچکدام از مقایسه‌های انجام شده نتیجه صفر حاصل نشده که این نشانگر کیفیت بالای روش پیشنهادی است. از سوی دیگر همانطور که مشخص است نرخ صحت الگوریتم پیشنهادی از تمام روش‌ها بالاتر و بهتر می‌باشد.

در مورد نرخ خطا همانطور که نتایج نیز نشان می‌دهند، نرخ خطای روش پیشنهادی از سه روش $VSUMM_1$، OV و STIMO پایین‌تر بوده و بهتر عمل می‌کند اما نسبت به دو روش دیگر $VSUMM_2$ و DT ضعیف‌تر عمل می‌کند. با این وجود تفاوت میان نرخ خطای روش پیشنهادی و این دو روش کم بوده و نرخ صحت بالاتری نیز نسبت به این دو دارد.

۷- نتیجه‌گیری

در این مقاله در ابتدا با جداسازی صوت و تصویر، خوشه‌بندی جدا در هر دو انجام می‌گیرد سپس با کمک معیار تغییر یافته اطلاعات انطباقی میزان سازگاری میان این دو مشخص شده و بنابراین رویدادها با میزان تکرارهای آن‌ها مشخص می‌شود.

در انتها سعی شده با الگوریتم کاس و تغییر در آن، به ارزیابی درستی از نتیجه خلاصه‌سازی ویدیویی رسید. با این روش می‌توان نتایج خلاصه‌سازی خودکار را به صورت مستقل و جدا، با نظر تک تک کاربران سنجید و در مورد کیفیت ویدیوها و نزدیک بودن آن به ادراک انسانی تصمیم گرفت.

این راهکار بر روی ۵۰ ویدیو تکرار شد، میانگین ۸۸٪ نرخ صحت در خلاصه‌سازی دارد و میزان خطای آن ۳۱٪ می‌باشد که به نسبت سایر روش‌ها جزء بالاترین نرخ صحت می‌باشد و نرخ خطای آن نیز نسبت به اغلب روش‌ها پایین و نسبت به سایر روش‌ها چندان بالا نیست.
نتیجه روش‌ها در بازه اطمینان ۹۵٪ و ۹۸٪ نیز سنجیده شدند که نتیجه خوبی را در بر داشته و نسبت به سایر روش‌ها کیفیت بالاتری را از خود نشان دادند.

۸- مراجع

[1] Ian H.Written.,Eibe Frank.,Mark A.Hall." Data Mining:Practical Machine Learning Tools and Techniques," **Elsevier** ISBN 978-0-12-374856-0, 2011.
[2] Brandon Heung, e. a. "An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping," **Elsevier Geoderma**, 256:62-77, 2016 .
[3] Chengde Zhang, Xiao Wu, Mei-Ling Shyu, Qiang Peng, "Integration of Visual Temporal Information and Textual Distribution Information for News Web Video Event Mining," **IEEE Transactions on Human-Machine Systems**, 46: 124 – 135,2015.
[4] Li-Wei Kang, Chia-Ming Tsai, Chia-Wen Lin. "Learning-based movie summarization via role-community analysis and feature fusion," **IEEE Multimedia Signal Processing (MMSP)**, 17:1-6, 2016.
[5] Hesham Farouk , Kamal A. El Dahshan , Amr Abozeid . "Context-Aware Joint Video Summarization and Streaming (CVSS) Approach Multimedia (ISM), " **IEEE International Symposium**, 18: 597- 602, 2016.
[6] Dongping Zhao, Jiapeng Xiu, Yu Bai, Zhengqiu Yang. "An improved item-based movie recommendation algorithm," **Cloud Computing and Intelligence Systems (CCIS)**, 44: 30-38, 2016.
[7] Chengde Zhang, e. a. Nadaraya. "Near-Duplicate Segments based news web videoevent mining,' **Elsevier Signal Processing**, 120:26-35, 2016 .
[8] Prashant G. Shambharkar, M N Doja. "Automatic classification of movie trailers using data mining techniques: A review," **Computing, Communication & Automation (ICCCA)**, 15: 88-94, 2015.
[9] Chia-Ming Tsai, Li-Wei Kang, Chia-Wen Lin, Weisi Lin. "Scene-Based Movie Summarization Via Role-Community Networks," **IEEE Transactions on Circuits and Systems for Video Technology**, 23:1927- 1940, 2013 .
[10] Vahid Bastani, Lucio Marcenaro, Carlo S. Regazzoni, "Online Nonparametric Bayesian Activity Mining and Analysis From surveillance Video," **IEEE Transactions on Image Processing**, 25: 2089 – 2102, 2016.
[11] P. M. Ashok Kumar, V. Vaidehi, E. Chandralekha. "Video traffic analysis for abnormal event detection using frequent item set mining," **Recent Trends in Information Technology (ICRTIT)**, 25: 551-556, 2013
[12] Adway Mitra, Soma Biswas, Chiranjib Bhattacharyya . "Bayesian Modeling of Temporal Coherence in Videos for Entity Discovery and Summarization," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 3:430 – 443,2017.
[13] Kimber, H. Zhou and D. "Unusual Event Detection via Multicamera Video Mining," **International Conf on Pattern Recognition**, 36:1161–1166,2006.
[14] Hoang Trinh, Jun Li, Sachiko Miyazawa, Juan Moreno, Sharath Pankanti. "Efficient UAV Video Event Summarization," **IEEE 21st International Conference on Pattern Recognition**, 9: 2226- 2229, 2012 .
[15] Sangdoo Yun, Kimin Yun, Soo Wan Kim, Youngjoon Yoo, Jiyeoup Jeong. "Visual Surveillance Briefing System: Event-based Video Retrieval and Summarization," **11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)**, 20: 204-209,2014.
[16] Rajat Aggarwal, Brijesh Singh Butola. "Event Summarization in Videos," **IEEE International Conference on Computing, Communication and Automation**, 16: 1150-156,2016.
[17] Anastasios Dimou, Dimitra Matsiki, Apostolos Axenopoulos, Petros Daras. "A user-centric approach for event-driven summarization of surveillance videos," **IEEE Imaging for Crime Prevention and Detection (ICDP-15),** 15:450-456, 2015.
[18] Mostafa Tavassolipour, Mahmood Karimian, and Shohreh Kasaei. "Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula," **IEEE Transactions on Circuits and Systems for Video Technology**, 24: 291 – 304,2013.
[19] Anh-Phuong Ta, Mathieu Ben., and Guillaume Gravier. "Improving Cluster Selection and Event Modeling in Unsupervised Mining for Automatic AudiovisualVideo Structuring," **Springer-Verlag**,4: 529-540,2012.

3rd. International Conference on September 2017
Information Technology, Computer and Telecommunications
سومین کنفرانس بین المللی شهریورماه ۱۳۹۶
فناوری اطلاعات، مهندسی کامپیوتر و مخابرات

Increasing the precision of Video summarization algorithm with the help of video audio mutual information

Bahareh Spahbodi, Abdolkarim Elahi

Department of comueter, Faculty of Computer, University of Islamic Azad University of Behshahr, E-mail: Bahareh.espahbodi@gmail.com
Department of comueter, Faculty of Computer, University of Islamic Azad University of Behshahr, E-mail: Ak_elahi@yahoo.com

Abstract. Today with the cameras in almost every where we reached to a large amount of video data. This shows the importance of accessing and manipulating these data. With the help of summarization, Video will convert into a group of key frames which is a good representation for the whole video.
This paper attempts to make a video summarization with the help of a K_medoids algorithm which is a clustering algorithm. At the first step the video will be clustered into k groups and with processing the audio and eliminating the silient parts and calculating the mutual information between audio and video the events will be detected. The result was examined and it has reached to %88 of precision which is a great number between other methods and %31 of error rate.

Keywords: Video mining; video Summarization, Clustering, K-Medoids, Video audio mutual information