مجله ماشین‌بینایی و پردازش‌تصویر
سال یکم، شماره دو، زمستان۱۳۹۲

مقاله مروری

مروری بر سیستم‌های تحلیل محتوایی و معنایی ویدیو از دیدگاه ساختار سلسله مراتب معنایی در تولید فیلم

محمدحسین سیگاری، حمید سلطانیان‌زاده، حمیدرضا پوررضا

چکیده

سیستم‌های تحلیل معنایی ویدیو به دسته‌ای از سیستم‌ها اطلاق می‌شود که نوعی ابزار هوشمند جهت بررسی و تحلیل محتوایی و معنایی ویدیو می‌باشند. بررسی و دسته‌بندی این سیستم‌ها می‌تواند از دیدگاه‌های مختلف انجام شود. در این مقاله سعی شده بر اساس یک ساختار سلسله مراتبی که از نحوه تولید فیلم توسط یک فیلم‌ساز برگرفته شده است، سیستم‌های تحلیل معنایی ویدیو از نظر شکاف معنایی میان ویژگی‌های سطح پایین و مفاهیم سطح بالا مورد بررسی قرار گیرند. به این ترتیب، پس از معرفی سیستم‌های تحلیل معنایی ویدیو، دو چالش اصلی در این سیستم‌ها مطرح می‌شود: شکاف حسگری و شکاف معنایی. پس از آن با ارائه یک ساختار سلسله مراتبی مبتنی برنحوه ساخت فیلم،چگونگی کاهش شکاف معنایی در سیستم‌های تحلیل ویدیو مورد بررسی قرار گرفته و تحقیقات انجام شده در این زمینه مرور می‌شود. بر اساس این ساختار، سیستم‌های تحلیل ویدیو در سه سطح دسته‌بندی و مرور می‌شوند: پردازش فریم‌ها، استخراج محتوا و استخراج معنا. در نهایت مشکلات و مسائل باز در سیستم‌های کنونی تحلیل معنایی ویدیو بازگو می‌گردد. مهمترین مشکلات در این زمینه عبارتند از: تنوع رویدادها و مفاهیم در یک ویدیو، امکان وجود معانی و مفاهیم متعدد برای یک رویداد معین، پردازش‌های بلندمدت برای استخراج معانی و مفاهیم و استفاده و ترکیب داده‌های چندنوعی. به این ترتیب، با مروری بر مقالات، مسیرهای تحقیقاتی فعلی و مشکلات پیشرو در زمینه سیستم‌های تحلیل معنایی ویدیو معرفی خواهد شد.

کلید واژه ها

آشکارسازی رویداد؛ استخراج مفهوم؛ تحلیل معنایی ویدیو؛ ساختار معنایی سلسله مراتبی؛ شکاف معنایی.

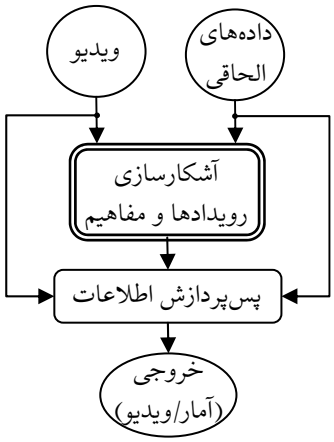
۱ مقدمه

امروزه رسانه‌های دیداری-شنیداری مهمترین رسانه‌های تاثیرگذار بر جوامع بشری می‌باشند و حجم بسیار زیادی از اطلاعات دیجیتال را شامل می‌شوند. با ورود سیستم‌های دیجیتال برای تولید، ضبط و پخش اطلاعات چندرسانه‌ای و همچنین فراهم‌آوری زیرساخت مخابراتی برای انتقال اطلاعات حجیم، سرعت رشد این رسانه‌ها در حال افزایش است. تا چندی پیش که حجم اطلاعات ویدیویی بسیار محدود بود، انجام چنین پردازش‌هایی به صورت دستی مشکلات چندانی را به همراه نداشت. اما هم‌اکنون با افزایش سریع حجم اطلاعات ویدیویی، نیاز به طراحی و تولید سیستم‌های خودکار برای تحلیل و پردازش داده‌های ویدیویی به منظور سهولت در روند نمایه‌سازی، جستجو و بررسی محتوای درونی آنها بسیار لازم به نظر می‌رسد.

این مقاله در مردادماه ۱۳۹۱ دریافت و در خردادماه ۱۳۹۲ بازنگری و پذیرفته شد.

محمد حسین سیگاری، دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر.hoseyn_sigari@ieee.org
حمید سلطانیان زاده، دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر؛آزمایشگاه تحلیل تصویر، بخش رادیولوژی، بیمارستان هنری فورد، دیترویت، میشیگان، ایالات متحده آمریکا.hszadeh@ut.ac.ir
حمید رضا پوررضا، دانشگاه فردوسی مشهد، دانشکده مهندسی، گروه مهندسی کامپیوتر.hpourreza@um.ac.ir

سیستم تحلیل معنایی ویدیو[1]، یک سیستم خودکار یا نیمه‌خودکار پردازش داده‌های ویدیویی جهت استخراج و دسته‌بندی اطلاعات معنایی ویدیو است که می‌تواند کاربردهای مختلفی نظیر جستجو و بازیابی ویدیو[2][۱]، خلاصه‌سازی[3][۲] و نظارت[4][۳]داشته باشد.در یک نگاه ساده، شمای کلی سیستم تحلیل معنایی ویدیو مطابق شکل۱ می‌باشد. ورودی سیستم شامل ویدیو و داده‌های الحاقی می‌باشد. داده‌های الحاقی دربردارنده هرنوع اطلاعات افزونه‌ای است که علاوه بر ویدیو به سیستم در تحلیل معنایی ویدیو کمک می‌کند. مثلا برای ویدیو سینمایی وجود متن زیرنویس می‌تواند به عنوان داده‌های الحاقی تلقی شود.همچنین داده‌های صوتی همراه ویدیو می‌تواند به عنوان داده الحاقی تلقی گردد. توجه به این نکته ضروری است که لزوما داده‌های الحاقی برای تمام سیستم‌های تحلیل معنایی در دسترس نیست، بنابراین فقط در صورت وجود این داده‌ها، از آنها برای افزایش کارایی سیستم استفاده می‌گردد. در برخی مقالات از داده‌های الحاقی به عنوان داده‌های خارجی نام برده می‌شود[۲].



شکل ۱ شمای کلی سیستم تحلیل معنایی ویدیو

ورودی‌های سیستم در بخشی به نام «آشکارسازی رویدادها و مفاهیم» مورد پردازش قرار می‌گیرند. در واقع می‌توان این بخش را مهمترین بخش سیستم تحلیل معنایی ویدیو نامید، چرا که این بخش باید بتواند رویدادها و مفاهیمی که در ویدیو رخ می‌دهد را آشکار و شناسایی نماید. مثلا در ویدیو فوتبال، آشکارسازی رویدادهایی همچون گل، پنالتی، و کرنر در این بخش انجام می‌شود. پس از آشکارسازی رویدادها و مفاهیم درون ویدیو، زمان وقوع و سایر اطلاعات لازم از رویدادها و مفاهیم به بخش دیگری به نام «پس‌پردازش اطلاعات» منتقل می‌گردد. در این بخش رویدادها و مفاهیم ویدیو که در مرحله قبل استخراج شده است، مورد پردازش قرار گرفته تا خروجی نهایی سیستم آماده شود.

[1] Semantic Video Analysis
[2] Video Retrieval
[3] Summarization
[4] Surveillance

معمولا پردازش‌های این مرحله چندان پیچیده نبوده و جزو پردازش‌های اصلی سیستم تحلیل معنایی محسوب نمی‌گردد. در سیستم تحلیل معنایی ویدیو، دو نوع خروجی کلی وجود دارد: (۱) اطلاعات آماری و (۲) ویدیو.نوع اول خروجی، شامل اطلاعات آماری از رویدادهای رخداده در ویدیو است. به عنوان مثال، در ویدیو فوتبال این خروجی شامل آماری همچون تعداد گل‌های زده شده، تعداد کارت‌های قرمز و زرد، و تعداد کرنرها می‌باشد. نوع دوم خروجی، یک ویدیو است که معمولا شامل بخش‌هایی از ویدیو ورودی می‌باشد. این نوع خروجی معمولا به منظور خلاصه‌سازی ویدیو یا پالایش محتوای ویدیو ارائه می‌گردد.

در این مقاله، با مروری بر تحقیقات اخیر در زمینه تحلیل محتوایی و معنایی ویدیو، این تحقیقات مورد بررسی قرار می‌گیرد. این مقاله در پنج بخش سازمان‌دهی شده است که در بخش حاضر، مقدمه‌ای بر موضوع ارائه شد. در بخش دوم چالش‌های اساسی در این زمینه مطرح می‌گردد. در بخش سوم، پس از معرفی یک ساختار سلسله مراتبی محتوایی-معنایی، سیستم‌های تحلیل محتوایی و معنایی ویدیو مورد بررسی و طبقه‌بندی قرار می‌گیرند. بحث و نتیجه‌گیری در بخش چهارم ارائه شده و در نهایت خلاصه‌ای از موضوعات مطرح شده در این مقاله در بخش پنجم ارائه می‌گردد.

۲ چالش‌های اساسی در تحلیل معنایی ویدیو

مشکلات بسیار زیادی در تولید سیستم‌های کارامد برای تحلیل معنایی ویدیو وجود دارد، اما این مشکلات را می‌توان در قالب دو چالش اساسی تقسیم‌بندی نمود: (۱) شکاف حسگری[5] و (۲) شکاف معنایی[6][۴]. در این بخش به بررسی اجمالی این چالش‌ها خواهیم پرداخت.

۲-۱ شکاف حسگری

به شکاف میان موجودیت شی در عالم واقعیت و توصیف آن در تصویر، شکاف حسگری می‌گویند. به عبارت دیگر، شکاف حسگری به مشکلات ابزارهای تصویربرداری و نقش آنها در تحلیل تصویر می‌پردازد. از این رو، برای رفع شکاف حسگری در سیستم‌های تحلیل معنایی ویدیو، روش‌ها و ابزارهای مختلف تصویربرداری استفاده شده است. ویژگی‌های مهم در تصویربرداری که ممکن است عدم توجه به آنها باعث ایجاد شکاف حسگری گردد عبارتند از: (۱) قابلیت کنترل دوربین، (۲) مکان و زاویه دوربین، (۳) تعداد دوربین‌ها، (۴) سرعت تصویربرداری، (۵) کیفیت وضوح تصویر و (۶) طیف تصویربرداری.

[5] Sensory Gap
[6] Semantic Gap

از جمله تحقیقات مهم که تاکید آنها بر کاهش شکاف حسگری است می‌توان به [۵, ۶] اشاره نمود. در این تحقیقات، سیستم‌هایی برای آشکارسازی گل وآفساید در مسابقات فوتبال ارائه شده است که عمده تاکید آنها بر رفع شکاف حسگری می‌باشد. به علت اینکه در این سیستم‌ها دقت بسیار اهمیت دارد، از چندین دوربین با نرخ تصویربرداری بالا در مکان‌های مشخص و با زاویه معلوم استفاده شده است. نرخ تصویربرداری بالا در این سیستم‌ها برای جلوگیری از تارشدگی توپ در هنگام شوت کردن است. همچنین تعداد، مکان و زاویه دوربین‌ها طوری تعیین شده که همپوشانی اشیا به حداقل رسیده و عملکرد سیستم دقیق باشد.

از جمله تحقیقات دیگری که شکاف حسگری در آنها اهمیت ویژه‌ای دارد، تحلیل ویدیوهای نظارتی است. در ویدیوهای نظارتی، علاوه بر این که تعداد دوربین‌ها و مکان و زاویه آنها اهمیت دارد، طیف تصویربرداری نیز بسیار مهم است. بسیاری از دوربین‌های نظارتی کنونی، قابلیت تصویربرداری در طیف مادون قرمز نزدیک[1] را دارند[۳]. به این ترتیب، تحلیل ویدیوهای تصویربرداری شده از فضاهای تاریک نیز امکان‌پذیر خواهد شد. عدم توانایی دوربین برای تصویربرداری در طیف مادون قرمز، به نوعی شکاف حسگری در سیستم ایجاد می‌نماید.

۲-۲ شکاف معنایی

شکاف معنایی، به مسئله عدم تطابق میان مفاهیم استخراج شده توسط یک سیستم تحلیل معنایی ویدیو و مفاهیم قابل درک توسط انسان می‌پردازد. به عبارت دیگر، تفسیر معانی از اطلاعات تصویری در سیستم کامپیوتری و انسان متفاوت است، که این موضوع طراحی سیستم‌های تحلیل معنایی ویدیو را مشکل می‌کند. به عنوان یک مثال ساده، انسان به راحتی می‌تواند وقوع گل در فوتبال را تشخیص دهد، اما آشکارسازی گل توسط کامپیوتر نیازمند پردازش‌های زیادی است. به طور خلاصه این پردازش‌ها عبارتند از آشکارسازی توپ، آشکارسازی دروازه، ردیابی توپ و تعیین این که آیا توپ از خط دروازه عبور کرده یا خیر. بررسی محتوای ویدیو و استخراج اطلاعات معنایی از یک‌سری تصاویر دیجیتال، مشکلات و مسائلی را به دنبال دارد که برای حل آنها نیازمند ارائه الگوریتم‌های کارامد هستیم. این مشکلات و مسائل با عنوان شکاف معنایی مطرح می‌شود.

با توجه به آنچه گفته شد، برای استخراج رویدادها و مفاهیم ویدیو، لازم است پردازش اطلاعات در سطوح معنایی مختلف انجام گیرد. ماهیت پردازش‌ها در هر سطح متفاوت بوده و در نتیجه اطلاعات استخراج شده از آنها در سطح معنایی متفاوتی خواهد بود. بر اساس یک دسته‌بندی کلی، پردازش‌های انجام شده برای استخراج اطلاعات معنایی از ویدیو را می‌توان در سه سطح طبقه‌بندی کرد: (۱) پردازش‌های سطح پایین، (۲) پردازش‌های سطح میانی و (۳) پردازش‌های سطح بالا. برای حذف شکاف معنایی میان ویژگی‌های سطح پایین و مفاهیم سطح بالا باید سعی نمود استخراج ویژگی‌ها بر اساس سطوح معرفی شده انجام شود. به عبارت دیگر، مفاهیم سطح بالا باید طی چند مرحله، به تدریج از ویژگی‌های سطح پایین استخراج شوند. در غیر این صورت، استخراج مفاهیم سطح بالا به تغییرات ویژگی‌های سطح پایین بسیار حساس خواهد بود و در نتیجه سیستم ناکارامد می‌شود.

در پردازش‌های سطح پایین، تصاویر ویدیو خام توسط الگوریتم‌هایی مورد پردازش قرار گرفته و ویژگی‌های سطح پایین تولید می‌شوند که معمولا این ویژگی‌ها، اطلاعات معنایی سطح بالایی با خود به همراه ندارند. پردازش‌های سطح پایین می‌تواند شامل آشکارسازی مرز بین شات‌ها [۷, ۸, ۹, ۱۰]، استخراج فریم کلیدی [۱۱, ۱۲, ۱۳] و تشخیص نوع نما [۱۴, ۱۵]باشد.

پس از استخراج ویژگی‌های سطح پایین، پردازش‌های سطح میانی بر روی آنها انجام شده تا اطلاعات معنایی در سطح بالاتری از ویدیو استخراج گردد. ویژگی‌های سطح میانی معمولا محتوای ویدیو را توصیف می‌کنند. برخی از مهمترین پردازش‌های سطح میانی عبارتند از: آشکارسازی و تشخیص اشیا [۱۶, ۱۷]، آشکارسازی و تشخیص افراد [۱۵, ۱۸, ۱۹, ۲۰]، تحلیل رفتار حرکتی اشیا [۱۹, ۲۱, ۲۲]، تشخیص لوگو [۲۳, ۲۴]، تشخیص نوع صحنه [۱۷, ۱۸]، تشخیص حرکت آهسته و پخش تکراری صحنه‌های مهم [۱۵, ۲۳, ۲۵] و تشخیص نوع حرکت دوربین [۲۶].

آخرین سطح از پردازش‌های انجام شده برای تحلیل معنایی ویدیو، پردازش‌های سطح بالا می‌باشد. هدف نهایی از پردازش‌های سطح بالا استخراج معنا از ویدیو بر اساس ویژگی‌های سطح میانی است. از جمله الگوریتم‌های ارائه شده در این سطح از پردازش‌ها می‌توان به آشکارسازی و تشخیص رویداد [۲۷]، جستجو و بازیابی معنایی ویدیو [۱] و خلاصه‌سازی [۲] اشاره نمود. به جرات می‌توان گفت که آشکارسازی و تشخیص رویداد، اصلی‌ترین و پایه‌ای‌ترین پردازش در میان ویژگی‌های سطح بالا است. چرا که پس از آشکارسازی و تشخیص رویدادها در ویدیو، تقریبا می‌توان هر نوع پردازش سطح بالای ویدیو از جمله جستجو و بازیابی، خلاصه‌سازی، نظارت و ... را انجام داد. بنابراین می‌توان گفت آشکارسازی رویداد اولین مرحله از پردازش‌های سطح بالا است.

برای درک بهتر چگونگی پل زدن میان ویژگی‌ها سطح پایین و مفاهیم سطح بالا، یک مثال در مورد خلاصه‌سازی ویدیو مسابقات فوتبال ارائه می‌شود [۱۵]. در این مقاله، در پردازش‌های سطح پایین، ابتدا رنگ زمین چمن تعیین شده و مرز بین شات‌ها مشخص می‌شود. همچنین نوع نما (نمای دور، نمای متوسط، نمای بسته و نمای خارج از زمین) برای هر شات تعیین می‌گردد. سپس بر اساس ویژگی‌های سطح پایین، ویژگی‌های سطح میانی استخراج می‌شود. این ویژگی‌ها عبارتند از تشخیص حرکت آهسته،

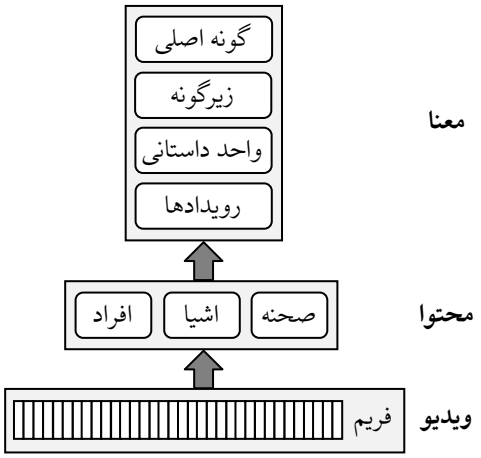
[1] Near Infrared

آشکارسازی داور و آشکارسازی محوطه جریمه. ویژگی‌های سطح میانی بر اساس خروجی پردازش‌های سطح پایین استخراج می‌شود. تشخیص حرکت آهسته وابسته به آشکارسازی مرز بین شات‌ها است. همچنین آشکارسازی داور باید در شات‌هایی با نمای بسته و آشکارسازی محوطه جریمه باید در شات‌هایی با نمای دور انجام شود. پس از استخراج ویژگی‌های سطح میانی، ویژگی‌های سطح بالا استخراج می‌شوند. در این سیستم، ابتدا رویداد گل به عنوان ویژگی سطح بالا استخراج می‌گردد. آشکارسازی رویداد گل بر اساس تشخیص نمای بسته بازیکن موثر در وقوع گل، تشخیص نمای خارج از زمین مربوط به هیجان تماشاچیان و تشخیص پخش حرکت آهسته انجام می‌شود. سپس خلاصه‌سازی ویدیو به عنوان بالاترین رده معنایی در این سیستم، در سه شکل قابل ارائه است. شکل اول خلاصه‌سازی فقط بر اساس بخش‌های حرکت آهسته انجام می‌شود. شکل دوم فقط شامل گل‌های بازی و شکل سوم مبتنی بر شی است. در شکل سوم خلاصه‌سازی، شات‌هایی که شامل داور و محوطه جریمه باشند، مهم تلقی می‌شوند. به این ترتیب، با طی کردن سه سطح پردازش، خلاصه‌سازی ویدیو انجام می‌شود.

همان طور که در مثال بالا ملاحظه می‌گردد، عملکرد هر سطح از پردازش‌ها در تحلیل معنایی ویدیو وابسته به ویژگی‌های استخراج شده از سطح قبل است. مثلا در چنین سیستمی نمی‌توان صرفا بر اساس هیستوگرام رنگ یا لبه‌های تصویر خلاصه‌سازی انجام داد، بلکه لازم است طی انجام پردازش‌های سلسله مراتبی، به تدریج محتوا و معنای ویدیو تحلیل شود.

۳ سیستم‌های تحلیلمحتوایی و معنایی ویدیو

یک ویدیو از دیدگاه تولید فیلم در سه سطح قابل توصیف است: (۱) معنا[1]، (۲) محتوا[2] و (۳) ویدیو. این دیدگاه که در شکل۲مشاهده می‌گردد،یک ساختار سلسله مراتبی می‌باشد که بر اساس نحوه تولید ویدیو توسط یک فیلم‌ساز ایجاد شده است.فیلم‌ساز برای ساخت یک فیلم،ابتدا معنای کلی را که درون ذهن خود دنبال می‌کند در نظر گرفته، سپس با نوشتن فیلمنامه، آن معنای کلی را با جزییات معنایی کوچک‌تر شرح می‌دهد. در هنگام تصویربرداری، انتخاب صحنه،چینش اشیا درون آن و نحوه رفتار افراد (بازیگران) باید طوری باشد که معنای خاصی که در بخش‌های مختلف فیلم‌نامه دنبال می‌گردد، به خوبی به بیننده القا شود. به این ترتیب تصویربرداری انجام می‌گیرد تا ویدیو اولیه بدست آید. پس از آن، تدوین‌گر بر اساس اصول هنری، تغییراتی از قبیل جابجایی بخش‌های از ویدیو و در صورت نیاز افزودن جلوه‌های ویژه را به ویدیو اولیه اعمال می‌کند تا ویدیو نهایی بدست آید.



شکل ۲ ساختار سلسله مراتبی محتوایی-معنایی برای تحلیل ویدیو (برگرفته از دیدگاه نحوه تولید فیلم توسط یک فیلم‌ساز)

بر این اساس، فریم‌ها در سطح پایین، محتوا در سطح میانی و معنا در سطح بالا قرار دارد. محتوا شامل صحنه، افراد و اشیا درون صحنه است که معمولا به تنهایی معنای مشخصی ندارند، بلکه با چینش مناسب و ایجاد ارتباط بین آنها می‌توان معنای خاصی را به بیننده القا کرد. همچنین سطوح معنایی به چهار سطح قابل تقسیم است:

- گونه[3] اصلی: تعیین گونه اصلی همان دسته‌بندی ویدیوها بر اساس شباهت‌های معنایی آنها در دید کلی است. گونه اصلی می‌تواند شامل ویدیو سینمایی، مستند، ورزشی و ... باشد.
- زیرگونه: تعیین زیرگونه به منظور دسته‌بندی دقیق‌تر ویدیوهای یک گونه است. انواع زیرگونه، بر اساس گونه اصلی متغیر است. مثلا برای ویدیو سینمایی، زیرگونه شامل کمدی، درام، تراژدی و ... و برای ویدیو ورزشی، زیرگونه همان نوع ورزش (فوتبال، والیبال و ...) است.
- واحد داستانی: واحد داستانی شامل یک بخش پیوسته از ویدیو می‌باشد که شامل چندین رویداد یا واحد داستانی کوچک‌تر است. واحد معنایی، روایتی از یک معنا یا مفهوم را توصیف می‌کند.
- رویداد: رویداد کوچکترین واحد معنایی ویدیو است که یک اتفاق یا واقعه خاصی را شرح می‌دهد. مدت زمان وقوع یک رویداد معمولا بسیار کوتاه‌تر از یک واحد داستانی است.

هرچند ساختار معنایی برخی از ویدیوها مانند ویدیو ورزشی و خبری، به ظاهر از روند معرفی شده در شکل۲ پیروی نمی‌کند و بر اساس فیلم‌نامه یا داستان معینی تدوین نمی‌شود، اما می‌توان از نظر معنایی آن ویدیو را در قالب شکل۲ متصور شد. مثلا برای یک ویدیو خبری، گونه آن اخبار و زیرگونه آن ممکن است سیاسی، اجتماعی، ورزشی و ... باشد. هر واحد داستانی در ویدیو خبر، یک خبر مستقل است. مثلا خبر مربوط به پیشرفت دانشمندان

[1] Semantic
[2] Content
[3] Genre

کشورمان در تولید داروهای جدید و خبر موفقیت تیم رباتیک یک دانشگاه در مسابقات بین‌المللی، اگر چه هر دو از زیرگونه اخبار علمی هستند، اما دو واحد داستانی مستقل را تشکیل می‌دهند. همچنین برای ویدیو ورزشی، فیلم‌نامه از پیش تعیین شده وجود ندارد، اما می‌توان چنین ساختاری معنایی را برای ویدیوی ورزشی نیز متصور شد.

مثلا برای ویدیو مسابقات فوتبال، گونه ویدیو ورزشی و زیرگونه ویدیو همان نوع مسابقات ورزشی، یعنی فوتبال است. واحد داستانی در این ویدیو را می‌توان بخش‌های بازی-توقف[1] در نظر گرفت که در هر واحد داستانی ممکن است یک یا چند رویداد (مانند گل، خطا، کرنر و ...) به وقوع بپیوندد. بنابراین ملاحظه می‌شود که با داشتن یک نگاه جامع، می‌توان ساختار معنایی معرفی شده در شکل۲ را برای ویدیو اخبار نیز تعریف کرد.

در ساختار شکل۲، نحوه تولید فیلم توسط فیلم‌ساز از بالا به پایین است، در حالی که برای تحلیل ویدیو باید از پایین به بالا حرکت نمود. یعنی ابتدا فریم‌های ویدیو مورد پردازش قرار گرفته تا محتوای آن تعیین گردد. سپس با تحلیل محتوای ویدیو، معانی نهفته در آن تشخیص داده می‌شود.از طرفی، بخش‌های مختلف این ساختار از لحاظ معنایی، متناظر با لایه‌های مختلف پردازش‌های معنایی است که در بخش دوم (شکاف معنایی) معرفی شد. به عبارت دیگر،معمولا پردازش‌های سطح پایین مربوط به فریم‌ها، پردازش‌های سطح میانی مربوط به محتوا و پردازهای سطح بالا مربوط به معنا است.

در این بخش، با توجه به دیدگاه مطرح شده در شکل۲، سیستم‌های تحلیل ویدیو برای استخراج محتوا و معنای ویدیو مرور خواهد شد. بدین منظور ابتدا پردازش‌های سطح پایین بر روی فریم‌ها برای آشکارسازی مرز بین شات‌ها و استخراج فریم کلیدی مورد مطالعه قرار می‌گیرد. سپس پردازش‌های میانی برای استخراج محتوا از ویدیو بررسی خواهد شد. در نهایت، پردازش‌های سطح بالا برای استخراج معنا از ویدیو معرفی و دسته‌بندی می‌گردد.

۳-۱ پردازش فریم‌ها

پردازش‌های سطح پایین با پردازش فریم‌ها آغاز می‌شود. پردازش‌هایی مانند آشکارسازی مرز بین شات‌ها و استخراج فریم‌های کلیدی از مهمترین پردازش‌های سطح پایین است. در این بخش با مروری اجمالی بر روش‌های آشکارسازی مرز بین شات‌ها و استخراج فریم‌های کلیدی، این روش‌ها را مورد بررسی قرار می‌دهیم.

۳-۱-۱ آشکارسازی مرز بین شات‌ها

«آشکارسازی مرز بین شات‌ها[2]» با عناوین دیگری مانند «آشکارسازی گذار بین شات» نیز در مقالات معرفی شده است. این عملیات به منظور جداسازی شات‌های متوالی یک ویدیو از یکدیگر صورت می‌گیرد و معمولا اولین مرحله از مراحل تحلیل معنایی ویدیو است.اگر گذار بین دو شات به گونه‌ای باشد که یک شات در فریم k یکباره پایان یابد و شات بعدی به یکباره از فریم $k+1$ آغاز شود، این گذار را گذار ناگهانی یا کات می‌گویند. اما در نوع دیگر گذار که به آن گذار تدریجی می‌گویند، یک شات به تدریج و در طول چند فریم پایان یافته و در عین حال، شات بعدی به تدریج آغاز می‌شود. این نوع گذار، به عنوان یکی از ابزارهای هنری در ویرایش و تدوین ویدیو استفاده شده و خود می‌تواند حاوی اطلاعات معنایی باشد. گذار تدریجی می‌تواند در انواع مختلف صورت گیرد که مهمترین آنها عبارتند از: محو و ظهور تدریجی[3]، گذار انحلالی[4] و گذار سایشی[5]. نحوه ظهور مهمترین انواع گذار بین شات‌ها در شکل۳ نشان داده شده است.

یکی از مهمترین و پرکاربردترین ویژگی‌ها در آشکارسازی مرز میان شات‌ها، اندازه‌گیری تغییرات هیستوگرام در فریم‌های متوالی است. دلیل اهمیت و کاربرد فراوان این ویژگی، حجم محاسبات کم و در عین حال دقت نسبتا خوب در آشکارسازی مرز بین شات‌ها می‌باشد. هیستوگرام معمولا در فضای رنگی HSI محاسبه می‌گردد [۲۸, ۲۹, ۳۰]، اما فضای رنگی RGB نیز ممکن است مورد استفاده قرار گیرد [۲۹].

ویژگی‌های مبتنی بر حرکت نیز یکی دیگر از ویژگی‌های مورد استفاده برای آشکارسازی گذار بین شات‌ها می‌باشد. در این روش‌ها، اطلاعات حرکتی بر اساس تغییرات مقادیر پیکسل‌ها یا بلوک‌های تصویر در فریم‌های متوالی [۳۱] و اندازه حرکت بلوک‌های تصویر در فریم‌های متوالی [۱۰] استخراج می‌گردد. این روش‌ها اغلب برای ویدیوهای فشرده استفاده می‌شود، چرا که بخش عمده‌ای از اطلاعات ویدیو فشرده مربوط به اطلاعات حرکتی می‌باشد. به این ترتیب بدون این که ویدیو از حالت فشرده خارج شود، بر اساس اطلاعات حرکتی کد شده در ویدیو، می‌توان آشکارسازی مرز بین شات‌ها را انجام داد که این امر باعث کاهش چشمگیر حجم محاسبات می‌گردد. مهمترین عیب روش‌های مبتنی بر حرکت، ناکارامدی آنها در آشکارسازی دقیق گذار تدریجی میان شات‌ها است.

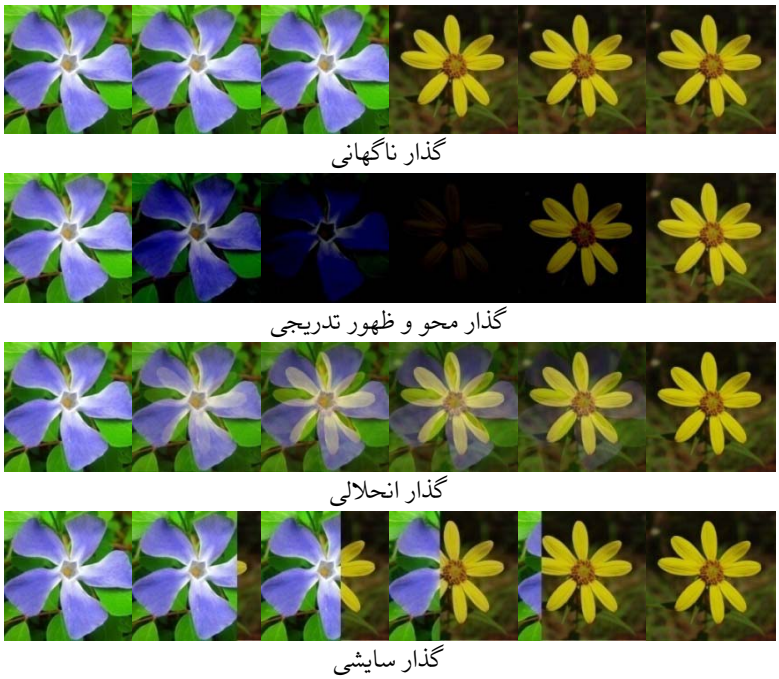
[1] Play-Break

[2] Shot Boundary Detection

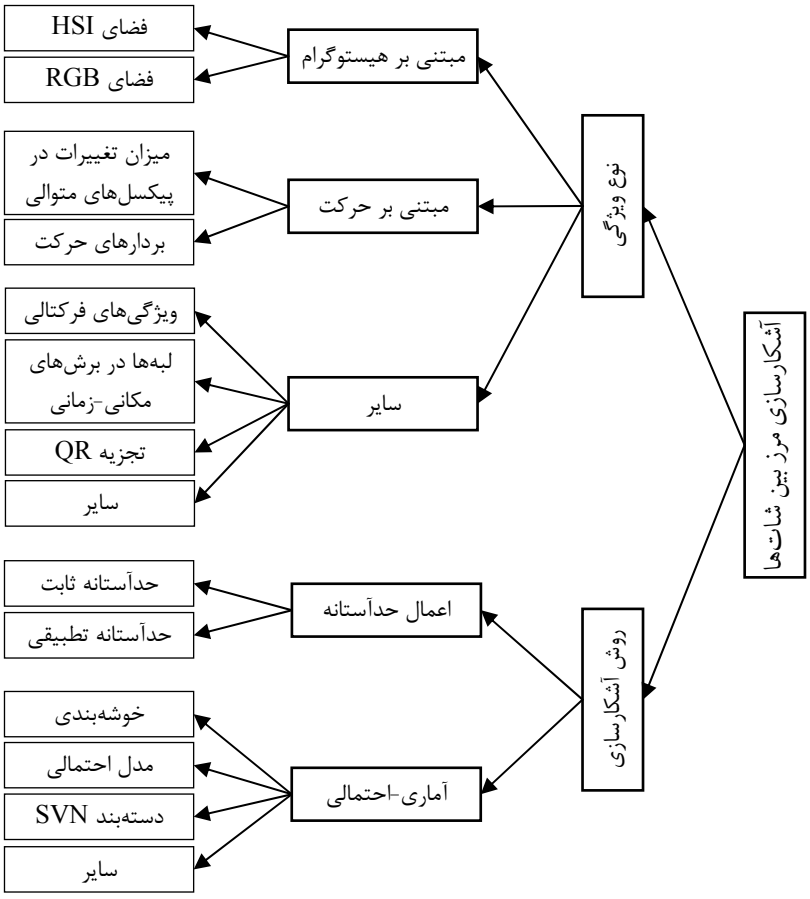
[3] Fade In/Fade Out

[4] Dissolve Transition

[5] Wipe Transition



شکل ۳ نمایش مهمترین انواع گذار بین شات‌ها



شکل ۴ دسته‌بندی روش‌های آشکارسازی مرز میان شات‌ها

بسیاری از روش‌هایی که تاکنون برای آشکارسازی مرز بین شات‌ها معرفی شده‌اند، از قوانین اکتشافی که معمولا بر اساس تجربه فردی بدست می‌آیند، استفاده می‌کنند. از جمله این روش‌ها می‌توان به اعمال حدآستانه به صورت ثابت یا تطبیقی [۲۸] و قوانین فازی [۲۹] اشاره کرد. اما در پژوهش‌های اخیر، بیشتر از روش‌های آماری نظیر خوشه‌بندی [۹]، مدل‌سازی احتمالی با توابع گوسی [۷]، شبکه عصبی [۱۰] و طبقه‌بندی‌کننده SVM[۳۱] استفاده گردیده است. در اغلب این پژوهش‌ها، علاوه بر آشکارسازی مرز بین شات‌ها، نوع گذار نیز تعیین شده است. در شکل۴، روش‌های مختلف آشکارسازی مرز بین شات‌ها از دیدگاه

نوع ویژگی استخراج شده و روش آشکارسازی، دسته‌بندی شده است.

آنچه در آشکارسازی مرز بین شات‌ها بسیار مهم است، آشکارسازی گذار تدریجی بین شات‌ها است. تقریبا تمام روش‌های آشکارسازی مرز بین شات‌ها با دقت بسیار بالایی قادرند تا گذار ناگهانی را آشکارسازی کنند، اما آشکارسازی گذار تدریجی مهمترین چالش در این حوزه از تحقیقات است. روش‌های نوین مانند [۷, ۹] برای آشکارسازی گذارهای تدریجی دقت خوبی دارند، اما معمولا حجم محاسباتی آنها زیاد است. به همین دلیل هنوز در بسیاری از سیستم‌های تحلیل ویدیو، از روش‌های ساده مانند روش‌های مبتنی بر هیستوگرام استفاده می‌شود.

۳-۱-۲ استخراج فریم کلیدی

فریم کلیدی[1] برای یک شات تعریف شده و فریمی است که نسبت به سایر فریم‌های دیگر، اطلاعات بیشتری در مورد معنا و محتوای آن شات در بر دارد [۳۲]. فریم کلیدی می‌تواند به عنوان نماینده‌ای از سایر فریم‌های شات مورد پردازش قرار گرفته و اطلاعات معنایی آن استخراج شود. دو مشکل اصلی در انتخاب فریم کلیدی مطرح است:

۱. آیا انتخاب یک فریم کلیدی برای هر نوع شات کافیست؟ یا این که لازم است بر اساس نوع شات، تعداد فریم‌های کلیدی انتخاب شده تغییر یابد؟
۲. از میان فریم‌های مختلف یک شات، کدام فریم اطلاعات بیشتری را در بر دارد؟

برای پاسخ به سوال اول، ساده‌ترین پاسخ این است که همواره برای هر نوع شات با هر مدت زمان، فقط یک فریم کلیدی انتخاب شود. اما روش‌های دیگری نیز ارائه شده که بر اساس میزان تغییرات فریم‌های یک شات، تعداد فریم‌های کلیدی انتخاب شده تغییر می‌کند. معیار تغییرات فریم‌ها می‌تواند بر اساس تغییرات رنگ [۳۳, ۳۴]، تغییرات بافت تصویر [۳۳, ۳۴] و تغییرات آنتروپی فریم‌های متوالی [۱۱] باشد. یکی دیگر از روش‌های انتخاب فریم کلیدی، خوشه‌بندی فریم‌های یک شات است. این روش‌ها بر خلاف روش‌های قبلی، هیچ توجهی به ترتیب فریم‌ها ندارد، بلکه فقط محتوای هر فریم را مورد توجه قرار می‌دهد. در این روش‌ها معمولا پس از خوشه‌بندی فریم‌ها، خوشه‌های بزرگ‌تر انتخاب شده و فریمی که به مرکز خوشه‌های بزرگ نزدیک‌تر است، به عنوان فریم کلیدی انتخاب می‌شود [۱۲, ۳۳, ۳۵].

در پاسخ به سوال دوم نیز روش‌های متعددی وجود دارد که ساده‌ترین آنها انتخاب فریم میانی شات است. اما این روش به صورت کورکورانه عمل کرده و غالبا عملکرد خوبی ندارد. در روش بهتر، نمودار تغییرات فریم‌های متوالی بر اساس یک معیار مشخص رسم می‌گردد. سپس بر اساس میزان تغییرات فریم‌های متوالی، تعداد فریم‌های کلیدی تعیین می‌گردد. چنین روشی در [۱۳, ۳۴] مورد استفاده قرار گرفته است. در روش‌های مبتنی بر خوشه‌بندی، معمولا یکی از فریم‌های هر خوشه به عنوان فریم‌کلیدی انتخاب شود که به مرکز خوشه نزدیک‌تر باشد [۱۲, ۳۳, ۳۵]. در این حالت، معمولا تعداد فریم‌های انتخاب شده متناسب با تعداد خوشه‌ها خواهد بود، ضمن این که توالی فریم‌ها برای انتخاب فریم کلیدی مورد توجه واقع نمی‌شود.

در میان روش‌های بررسی شده، عملکرد روش‌های مبتنی بر خوشه‌بندی نسبت به سایر روش‌ها بهتر است، اما این روش‌ها حجم محاسباتی زیادی داشته و بدون توجه به ترتیب زمانی فریم‌ها، فریم کلیدی را انتخاب می‌کنند. در عوض روش‌های مبتنی بر بررسی تغییرات فریم‌های متوالی، معمولا حجم محاسبات کمتری داشته و نسبت به روش‌های کورکورانه دقت بهتری دارند. به همین دلیل نسبت به سایر روش‌ها بیشتر مورد توجه قرار گرفته‌اند.

یک نکته مهم در ارزیابی روش‌های استخراج فریم کلیدی این است که اصولا معیاری دقیق برای انتخاب فریم کلیدی از یک شات وجود ندارد. بلکه ارزیابی‌ها بر اساس معیارهای سلیقه‌ای فردی است. بنابراین مقایسه میان روش‌های استخراج فریم کلیدی بسیار دشوار است. برای مطالعه بیشتر در مورد روش‌های استخراج فریم کلیدی به [۳۲] مراجعه شود. در این مقاله روش‌های مختلف استخراج فریم کلیدی با معیارهای ریاضی مورد بررسی و مقایسه قرار گرفته است.

۳-۲ تحلیل محتوای ویدیو

محتوای ویدیو را می‌توان به سه بخش اصلی تقسیم نمود: (۱) صحنه، (۲) اشیا و (۳) افراد. در این بخش به روش‌های آشکارسازی و تشخیص این سه دسته اصلی از محتوای ویدیو خواهیم پرداخت.

۳-۲-۱ تشخیص نوع صحنه

تعیین نوع صحنه در ویدیو می‌تواند نقش به سزایی در تحلیل معنایی آن داشته باشد. در اغلب روش‌های ارائه شده برای تشخیص نوع صحنه، ابتدا فریم(های) کلیدی استخراج شده و بر اساس تحلیل فریم‌های کلیدی نوع صحنه تعیین می‌گردد. با توجه به تعدد انواع صحنه و پیچیدگی مسئله، تحقیقات انجام شده در زمینه تشخیص صحنه در سه دسته کلی قابل تقسیم‌بندی است. در هر دسته از تحقیقات، محدودیت خاصی برای کاهش پیچیدگی لحاظ شده است.

در دسته اول تحقیقات، دقیقا نوع صحنه تشخیص داده نمی‌شود، بلکه میزان شباهت آن با سایر صحنه‌ها مورد ارزیابی قرار می‌گیرد [۳۰, ۳۶]. به عبارت دیگر به‌جای تشخیص و طبقه‌بندی صحنه، خوشه‌بندی آن انجام می‌شود. این دسته از تحقیقات معمولا بر اساس خوشه‌بندی فریم‌های کلیدی انجام شده و در

[1] Keyframe

جستجو و بازیابی ویدیو کاربرد دارند. دسته دوم تحقیقات به طبقه‌بندی صحنه از لحاظ منظره خارجی یا داخلی می‌پردازد [۳۷, ۳۸]. در این تحقیقات طبقه‌بندی انواع صحنه به دو دسته کلی محدود می‌شود. در [۳۷, ۳۸]، پس از استخراج ویژگی‌های سطح پایین از فریم‌ها، به ترتیب با استفاده از درخت تصمیم و شبکه‌های بیز، نوع صحنه طبقه‌بندی می‌گردد. در دسته سوم، تشخیص صحنه برای ویدیوهای خاص انجام می‌گیرد. این دسته از روش‌ها که غالبا برای ویدیو مسابقات ورزشی طراحی می‌شوند، معمولا بر اساس نوع نما و برخی ویژگی‌های تصویری دیگر، نوع صحنه را تشخیص می‌دهند [۱۷, ۱۸, ۳۹]. در اغلب این روش‌ها، نوع صحنه بر اساس تعریف یک سری قوانین اکتشافی بر روی نوع نما و رنگ اشیا و زمین مسابقه تعیین می‌شود.

۳-۲-۲ آشکارسازی و تشخیص اشیا

منظور از آشکارسازی اشیا، بررسی حضور یا عدم حضور یک شی خاص در تصویر است، در حالی که منظور از تشخیص اشیا، طبقه‌بندی آنها به منظور شناسایی نوع آنها است. یکی از بخش‌های مهم در تحلیل محتوای ویدیو، آشکارسازی و تشخیص اشیا است، به طوری که در برخی تحقیقات مانند [۱۷, ۴۰]، تاکید اصلی بر روی همین موضوع می‌باشد. روش‌های مختلفی برای آشکارسازی و تشخیص اشیا وجود دارد که بحث بر روی آنها خارج از موضوع است. برای مطالعه بیشتر در این زمینه به [۴۱] مراجعه شود. در این بخش کاربرد آشکارسازی و تشخیص اشیا در تحلیل معنایی ویدیو مورد بررسی قرار می‌گیرد.

در زمینه آشکارسازی و تشخیص اشیا برای ویدیوهای خاص، ویدیو مسابقات ورزشی بسیار مورد توجه محققین بوده است که از جمله آن می‌توان به آشکارسازی توپ در مسابقات فوتبال [۶, ۱۶, ۴۲, ۴۳]، آشکارسازی دروازه و محوطه جریمه در مسابقات فوتبال [۱۵, ۲۱, ۴۴] و آشکارسازی تابلو امتیازات مسابقات بسکتبال [۱۷] اشاره نمود. در اکثر این تحقیقات آشکارسازی اشیا بر اساس دانش زمینه و تعریف مدل‌های ابتکاری[1] در مورد رنگ و شکل اشیا انجام شده است. هرچند در تحقیقات معدودی مانند [۴۵] از روش‌های یادگیری ماشین برای آشکارسازی اشیا استفاده شده است.

یکی از روش‌های مرسوم در تحلیل معنایی ویدیو، آشکارسازی و تشخیص متن زیرنویس تصویر است. متن زیرنویس تصویر به دو حالت ممکن است مورد استفاده قرار گیرد. در حالت اول، متن زیرنویس به صورت یک فایل همراه با ویدیو ارائه می‌شود که با بررسی اطلاعات آن می‌توان اطلاعات مفیدی را برای تحلیل معنایی ویدیو استخراج نمود [۴۶]. در این حالت، متن زیر نویس به عنوان یک داده الحاقی تلقی می‌شود. اما در حالت دوم، زیرنویس در قالب یک شی در تصویر درج شده و قبل از بررسی محتوای آن باید آشکارسازی و تشخیص داده شود. با توجه به اینکه تشخیص محتوای زیرنویس درج شده در تصویر پیچیده است، معمولا از این روش برای تشخیص متن‌های کوتاه که حاوی اطلاعات زیادی هستند، استفاده می‌شود. از جمله این تحقیقات می‌توان به آشکارسازی رویداد در مسابقات فوتبال بر اساس جدول امتیازات نمایش داده شده در گوشه ویدیو اشاره نمود [۴۴, ۴۷].

از دیگر روش‌های آشکارسازی اشیا در ویدیو، روش‌های مبتنی بر مدل‌سازی پس‌زمینه و آشکارسازی اشیا متحرک در ویدیو است. عموما کاربرد این روش‌ها در سیستم‌های نظارتی است و معمولا از دوربین‌های ثابت برای تصویربرداری استفاده می‌کنند. در چنین روش‌هایی، با فرض ثابت بودن پس‌زمینه یا معلوم بودن مدل حرکتی آن، اشیا متحرک آشکارسازی می‌شود [۳].

آشکارسازی و تشخیص اشیا در ویدیوهای عمومی چندان مورد توجه نیست، چرا که در این ویدیوها تنوع اشیا بسیار زیاد است و آشکارسازی و تشخیص آنها پیچیده می‌باشد. در ویدیوهای عمومی، آشکارسازی و تشخیص اشیا فقط برای انواع معین و محدودی انجام می‌شود. از جمله تحقیقات در زمینه آشکارسازی و تشخیص اشیا در ویدیو می‌توان به آشکارسازی چهره در ویدیو اخبار [۴۸] اشاره کرد. همچنین در [۴۰] آشکارسازی و ردیابی اشیا (با تاکید بر آشکارسازی و ردیابی چهره، متن و خودرو) در ویدیو مورد توجه بوده است. در این تحقیق، ضمن گردآوری یک مجموعه داده استاندارد و ارائه معیارهای مفید برای ارزیابی روش‌ها، چند روش مختلف برای آشکارسازی و ردیابی اشیا نیز مورد مقایسه قرار گرفته است.

۳-۲-۳ آشکارسازی و شناسایی افراد

روش‌های آشکارسازی افراد در ویدیو می‌تواند شبیه به آشکارسازی اشیا باشد. چرا که می‌توان انسان را همانند یک شی مدل کرد و به جستجوی آن در تصویر پرداخت. از جمله کارهای انجام شده در زمینه آشکارسازی افراد، آشکارسازی بازیکنان و داور در مسابقات ورزشی است [۶, ۱۹, ۲۰, ۲۱, ۲۲, ۴۹]. در اغلب این روش‌ها، آشکارسازی بازیکنان بر اساس مدل‌های ابتکاری از قبیل تشخیص رنگ لباس بازیکنان یا داور در زمین مسابقه است. اما در برخی تحقیقات مانند [۱۹] از روش‌های یادگیری ماشین برای آشکارسازی افراد استفاده شده است.

یکی دیگر از روش‌های آشکارسازی افراد در ویدیو، روش‌های مبتنی بر مدل‌سازی پس‌زمینه و آشکارسازی اشیا متحرک است. این روش‌ها در سیستم‌های نظارتی کاربرد دارند و اساس عملکرد آنها آشکارسازی و مدل‌سازی حرکت در ویدیو است [۳]. این روش‌ها برای تشخیص فعالیت‌های انسان[2] نیز قابل استفاده هستند [۵۰].

شناسایی افراد نیز موضوع بسیار مهمی در تحلیل معنایی ویدیو است. شناسایی افراد، نه به معنی شناسایی دقیق هویت آنها،

[1] Heuristic

[2] Human Activity Recognition

بلکه به معنی تعیین نقش آنها در ویدیو می‌باشد. به عنوان مثال می‌توان چهره گوینده خبر در ویدیو اخبار را بر اساس بررسی تعداد تکرار ظهور چهره افراد مختلف در ویدیو خبر تعیین نمود [۴۸]. از دیگر تحقیقات انجام شده در زمینه آشکارسازی و شناسایی افراد می‌توان به تشخیص بازیکنان هر تیم در مسابقات ورزشی [۶, ۱۴, ۱۹, ۵۱, ۵۲]، تشخیص دروازه‌بان در مسابقه فوتبال [۱۴] و تشخیص داور در مسابقات ورزشی [۱۴, ۱۵, ۲۳, ۵۱] بر اساس رنگ لباس آنها اشاره نمود. همان طور که ملاحظه می‌شود، آشکارسازی و شناسایی افراد معمولا در ویدیوهای خاص و بر اساس دانش زمینه (مثلا رنگ لباس آنها) و مدل‌های ابتکاری انجام می‌گیرد. نمونه‌ای از نتایج آشکارسازی بازیکنان در مسابقات فوتبال و شناسایی تیم آنها در شکل(۵) نشان داده شده است [۲۱].

۳-۳ سطوح معنایی ویدیو

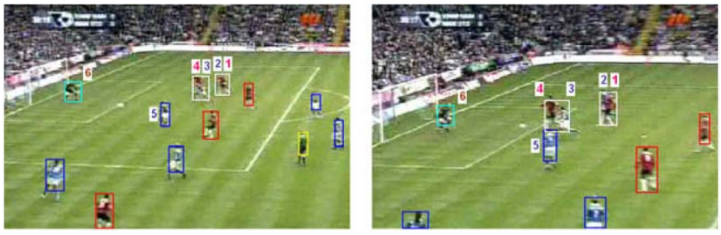
در دیدگاه مطرح شده در شکل(۲)، سطوح معنایی به چهار سطح قابل تقسیم است: گونه اصلی، زیرگونه، واحد داستانی و رویداد. در این بخش به بررسی تحقیقات انجام شده در سطوح معنایی مختلف خواهیم پرداخت.

۳-۳-۱ آشکارسازی رویداد

آشکارسازی رویداد در سطح معنایی بالا قرار می‌گیرد، اما ابتدایی‌ترین پردازش در سطح معنایی می‌باشد. به طور خلاصه، در هر رویداد ۶ پرسش مطرح می‌شود [۲۷]:

- چه کسی (Who): اشاره به فاعل و عامل رویداد
- چه زمانی (When): اشاره به زمان وقوع رویداد
- چه مکانی (Where): اشاره به محل وقوع رویداد
- چه چیزی (What): اشاره به آنچه متاثر از رویداد است
- چرا (Why): اشاره به علت وقوع رویداد
- چگونه (How): اشاره به شرح چگونگی وقوع رویداد

در ادبیات خبرنگاری، به این شش پرسش اساسی برای توصیف یک رویداد، 5W1H می‌گویند [۲۷]. در تحلیل معنایی ویدیو، برای کاهش پیچیدگی الگوریتم‌ها، معمولا به تمام این سوالات پاسخ داده نمی‌شود. در جدول ۱، سیستم‌های آشکارسازی رویداد از نظر توانایی پاسخ به سوالات شش‌گانه مورد مقایسه قرار گرفته‌اند. همان طور که در این جدول مشاهده می‌شود، در میان سوالات مطرح شده برای توصیف یک رویداد، آنچه نسبت به سایر سوالات مهم‌تر به نظر می‌رسد، پاسخ به سوال «چه زمانی» می‌باشد و تقریبا در تمام سیستم‌های آشکارسازی رویداد مورد توجه بوده است. هرچند پاسخ‌گویی به سوالات شش‌گانه برای توصیف یک رویداد وابسته به کاربرد است. مثلا در تحلیل ویدیوهای نظارتی، علاوه بر تعیین زمان وقوع رویداد، تعیین فاعل رویداد و مکان آن نیز اهمیت دارد. همچنین در سیستم‌های تشخیص فعالیت‌های انسان، پاسخ‌گویی به سوال «چگونه» بیشتر مورد تاکید است [۵۰].



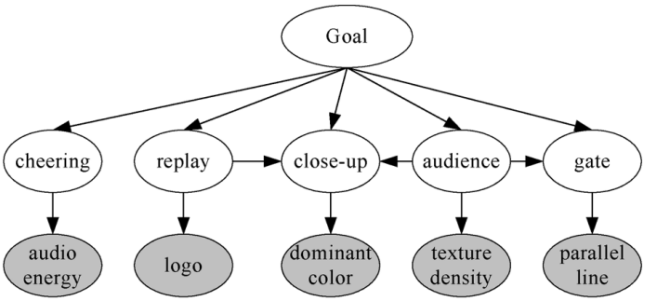
شکل ۵: نمونه‌ای از نتایج آشکارسازی بازیکنان در مسابقات فوتبال و شناسایی تیم آنها [۲۱]

جدول ۱ مقایسه توانایی سیستم‌های آشکارسازی رویداد برای پاسخ‌گویی به سوالات شش‌گانه 5W1H

مرجع	کاربری سیستم	سیستم به چه سوالاتی پاسخ می‌دهد؟					
		چه کسی	چه زمانی	چه مکانی	چه چیزی	چرا	چگونه
[۱۴]، [۲۳]، [۲۴]، [۴۷]،[۵۳]	ویدیو پخش تلویزیونی مسابقات فوتبال		√	√†			
[۵۳]، [۵۴]	ویدیو پخش تلویزیونی مسابقات بسکتبال		√				
[۳۹]	ویدیو پخش تلویزیونی مسابقات بیس‌بال		√				
[۶]	آشکارسازی آفساید در فوتبال با استفاده از دوربین‌های خاص	√	√	√		√	√
[۴۹]	بررسی رفتار حرکتی ورزشکاران				√		√
[۲۲]، [۴۳]	بررسی تاکتیک تیم‌ها در بازی فوتبال		√	√			√
[۵۰]	تشخیص فعالیت‌های انسان						√
[۵۵]، [۵۶]	ویدیوهای عمومی		√				
[۵۷]، [۵۸]	ویدیوهای نظارتی	√	√	√	√		

† در برخی از تحقیقات در زمینه آشکارسازی رویدادهای مهم در مسابقات فوتبال، تیم گل زننده نیز قابل تشخیص است. از این رو می‌توان گفت مکان تقریبی وقوع گل مشخص می‌شود. هرچند مکان وقوع سایر رویدادهای قابل تشخیص نیست.

به دلیل مشکلات موجود در سیستم‌های فعلی برای پوشش شکاف معنایی، آشکارسازی رویداد معمولا برای ویدیوهای خاص انجام می‌شود. یکی از انواع ویدیوهایی که آشکارسازی رویداد در آن بسیار مورد توجه محققان قرار گرفته، ویدیوهای ورزشی است. از جمله این سیستم‌ها می‌توان به سیستم‌های آشکارسازی رویدادهای مهم در بازی فوتبال [۱۴, ۲۳, ۴۲, ۵۳]، بسکتبال [۵۳, ۵۴] و بیس‌بال [۳۹] اشاره کرد. در بخش عمده‌ای از این تحقیقات، آشکارسازی رویداد بر اساس دانش زمینه و قوانین اکتشافی است که توسط طراح سیستم تعریف می‌شود. به عنوان مثال در [۱۴, ۱۵]، آشکارسازی گل در مسابقه فوتبال بر اساس ایجاد یک رابطه منطقی میان چند ویژگی سطح پایین و سطح میانی شامل آشکارسازی دهانه دروازه، تشخیص نوع نما، آشکارسازی دروازه‌بان و داور، تشخیص بخش‌های پخش مجدد و تشخیص هیجان تماشاچیان انجام می‌شود، که این روابط منطقی توسط طراح تعریف شده است. هرچند در تحقیقات اخیر مانند [۲۳, ۲۴, ۴۷]، سعی گردیده رابطه منطقی میان ویژگی‌ها بر اساس روش‌های یادگیری ماشین ایجاد گردد. نمونه‌ای از مدل منطقی ایجاد شده میان ویژگی‌های سطح پایین و سطح میانی با استفاده از شبکه بیز[1] برای آشکارسازی رویداد گل در مسابقات فوتبال در شکل۶ نشان داده شده است [۲۳]. در مدل شبکه بیز، احتمالات شرطی وقوع متغیرهای تصادفی در قالب یک گراف جهت‌دار نمایش داده می‌شود.



شکل ۶: نمونه‌ای از مدل منطقی ایجاد شده میان ویژگی‌های سطح پایین و سطح میانی با استفاده از شبکه بیز برای آشکارسازی رویداد گل در مسابقات فوتبال [۲۳]

سیستم آشکارسازی رویداد می‌تواند به صورت عمومی نیز تعریف شود که از جمله این تحقیقات می‌توان به سیستم‌های معرفی شده در [۵۵, ۵۶] برای آشکارسازی رویدادها و مفاهیم عمومی اشاره نمود. در این سیستم‌ها، ساختار به گونه‌ای طراحی شده تا اطلاعات مربوط به آشکارسازی رویداد، بدون وابستگی سیستم به دانش زمینه، فقط بر اساس نمونه‌های آموزشی استخراج گردد. چنین سیستم‌هایی معمولا آشکارسازی رویداد را بر اساس ویژگی‌های سطح پایین انجام می‌دهند.

نکته قابل توجه در آشکارسازی رویداد این است که معمولا در سیستم‌هایی که به صورت خاص‌منظوره طراحی می‌شوند، می‌توان به طور وسیع از دانش زمینه استفاده نمود. به عنوان مثال، در آشکارسازی رویدادها در بازی‌های ورزشی، تعداد رویدادها کاملا محدود و مشخص است. همچنین صحنه بازی مشخص و تعداد افراد درون صحنه محدود است. بنابراین طراحی سیستم‌های آشکارسازی رویداد برای کاربردهای خاص پیچیدگی کمتری دارد. اما در سیستم‌های عمومی برای آشکارسازی رویدادها و مفاهیم، تعداد رویدادها بسیار زیاد و نحوه رخ دادن آنها بسیار متنوع است. اصولا برای چنین سیستم‌هایی، تعداد رویدادها محدود می‌شود، اما نیاز به داده‌های آموزشی غنی همچنان وجود دارد. چرا که رویدادها می‌توانند در صحنه‌های مختلفی به وقوع بپیوندند و از این لحاظ باید تنوع کافی در داده‌های آموزشی وجود داشته باشد.

برای مطالعه بیشتر در زمینه روش‌های آشکارسازی ویدیو به[۵۹] مراجعه شود. در این مقاله،ضمن دسته‌بندی و ارزیابی روش‌های مختلف آشکارسازی رویداد در کاربردهای مختلف، این روش‌ها از نظر دقت، انعطاف‌پذیری، هزینه و میزان استفاده از دانش زمینه مقایسه شده‌اند.

۳-۳-۲ تشخیص واحد داستانی

تشخیص واحدهای داستانی، بر اساس رابطه معنایی میان رویدادهای آن بخش از ویدیو می‌باشد. با توجه به این که هنوز مشکلات فراوانی در آشکارسازی رویدادها از ویدیو وجود دارد، تشخیص معنای هر واحد داستانی بسیار مشکل است.به همین دلیل، تحقیقات در این زمینه بسیار محدود بوده و در برخی مقالات، تشخیص معنای واحدهای داستانی به صورت ضمنی انجام شده است. به این ترتیب که معنای دقیق واحد داستانی تعیین نشده، بلکه شباهت محتوایی آن با واحدهای داستانی دیگر مورد ارزیابی قرار گرفته است [۳۶, ۶۰]. عمده کاربرد این مقالات در جستجو و بازیابی ویدیو می‌باشد.

واحد داستانی در ویدیو مسابقات ورزشی همان بخش‌های بازی-توقف است. در این زمینه کارهای نسبتا زیادی انجام شده که از جمله آنها می‌توان به [۴۴, ۵۳] اشاره نمود. در اغلب این تحقیقات، بخش‌های بازی-توقف بر اساس نوع نما و حرکت دروبین تشخیص داده می‌شود.

۳-۳-۳ تشخیص گونه و زیرگونه

گونه ویدیو، دسته اصلی ویدیو را در مجموعه‌های ویدیویی تعیین می‌کند. به عنوان مثال، گونه ویدیو می‌تواند در یکی از دسته‌های سینمایی، ورزشی، خبری، آموزشی و مستند قرار گیرد. زیرگونه ویدیو، دسته فرعی ویدیو را در گونه اصلی تعیین می‌نماید. مثلا در ویدیو مسابقات ورزشی، تعیین زیرگونه معادل با تعیین نوع مسابقه ورزشی و در ویدیو اخبار، تعیین زیرگونه به معنی تعیین نوع خبر (سیاسی، اقتصادی، ورزشی، و...) است.

[1] Bayesian Network

گونه ویدیو بالاترین سطح معنایی از دیدگاه تولید فیلم است، بنابراین تعیین گونه ویدیو مستلزم انجام پردازش‌های سطح میانی و سطح بالا برای تحلیل محتوا و استخراج معنا از ویدیو است. به جز تحقیقات محدودی نظیر [۶۱, ۶۲] که در آن سعی شده پردازش‌ها در سطوح میانی و بالا برای تعیین گونه ویدیو انجام شود، سایر روش‌های ارائه شده برای تعیین گونه ویدیو مستقیما بر اساس ویژگی‌های سطح پایین و بدون پردازش‌های سطح میانی عمل می‌کنند [۶۳]. در [۶۲]، برای تعیین گونه ویدیو علاوه بر استفاده از ویژگی‌های تصویری و صوتی سطح پایین، از ویژگی‌هایی مانند کانتور اشیا و ساختار حرکتی ویدیو نیز به عنوان ویژگی‌های سطح میانی استفاده شده است. این در حالیست که در [۶۱]، بعد از انجام یک سری پردازش‌های سطح پایین و سطح میانی، تعیین گونه بر اساس آشکارسازی رویدادها در سطح بالا انجام می‌شود. بنابراین شکاف معنایی در این سیستم کمتر است.

تعیین زیرگونه ویدیو معمولا با فرض مشخص بودن گونه ویدیو انجام می‌شود. به عنوان نمونه می‌توان به تعیین زیرگونه ویدیو اخبار [۶۴]، ویدیو موسیقی [۶۵] و ویدیو مسابقات ورزشی [۶۳, ۶۶] اشاره نمود. سیستم ارائه شده در [۶۱] علاوه بر تعیین گونه ویدیو، زیرگونه ویدیو را نیز تشخیص می‌دهد و از این نظر توانایی کم‌نظیری دارد. علت اصلی این توانایی، وجود سلسله مراتب پردازشی در سطوح معنایی مختلف می‌باشد، به طوری که پس از استخراج ویژگی‌های سطح پایین، رویدادهای مهم آشکارسازی شده و بر اساس آن، زیرگونه و سپس گونه ویدیو مشخص می‌شود.

استفاده از سایر انواع داده نظیر داده‌های صوتی و متنی می‌توانند در تعیین گونه و زیرگونه بسیار موثر باشند، به طوری که بیشتر تحقیقات موفق در زمینه تعیین گونه و زیرگونه ویدیو، مبتنی بر استفاده از داده صوتی و متنی در کنار داده‌های تصویری می‌باشد. استفاده از سایر انواع داده، به ویژه داده‌های متنی، می‌تواند تاحدود زیادی شکاف معنایی میان ویژگی‌های سطح پایین و ویژگی‌های سطح بالا را از میان برداشته و باعث افزایش کارایی سیستم شود. در [۶۴, ۶۵, ۶۶] از داده‌های تصویری و صوتی و در [۶۷] از داده‌های تصویری و متنی برای تعیین گونه یا زیرگونه ویدیو استفاده گردیده است. در برخی تحقیقات مانند [۶۸]تعیین گونه فقط بر اساس داده‌های متنی می‌باشد.بررسی تحقیقات نشان می‌دهد که تاکنون استفاده از داده‌های تصویری نتوانسته کارایی خوبی در تعیین گونه و زیرگونه ویدیو داشته باشد.

۴ بحث و نتیجه‌گیری

مطالعه مقالات علمی نشان می‌دهد که تحلیل معنایی ویدیو یک موضوع تحقیقاتی بسیار جدید و جذاب است که در چند سال گذشته ظهور پیدا کرده و در همین مدت توجه بسیاری از محققین را به خود جلب نموده است. بر اساس مطالعات انجام شده، دو چالش اصلی در سیستم‌های تحلیل معنایی ویدیو وجود دارد که عبارتند از: (۱) شکاف حسگری و (۲) شکاف معنایی. در این میان آنچه بیشتر نظر محققین را به خود جلب کرده است، شکاف معنایی میان ویژگی‌های سطح پایین و مفاهیم سطح بالا است. در این بخش ابتدا مسائل و مشکلات باز که در بحث شکاف معنایی در سیستم‌های تحلیل معنایی ویدیو مطرح است، بازگو می‌گردد. سپس تعامل میان سه عامل دقت، سرعت و میزان خودکار بودن سیستم‌های تحلیل معنایی ویدیو مورد بررسی و بحث قرار خواهد گرفت.

۴-۱ مسائل و مشکلات باز در سیستم‌های تحلیل ویدیو

در این بخش با توجه به راهکارها و نتایج تحقیقاتی که مورد مطالعه قرار گرفت، مهمترین مسائل و مشکلات مطرح در زمینه شکاف معنایی معرفی می‌شود. این مشکلات عبارتند از: (۱) تنوع رویدادها و مفاهیم در یک ویدیو، (۲) امکان وجود معانی و مفاهیم متعدد برای یک رویداد معین، (۳) پردازش‌های بلندمدت برای استخراج معانی و مفاهیم و (۴) استفاده و ترکیب داده‌های چندنوعی. هرچند به دلیل جدید بودن زمینه سیستم‌های تحلیل معنایی ویدیو، می‌توان گفت تقریبا بسیاری از مسائل تحقیقاتی مرتبط با این بخش از دانش، به عنوان مسائل باز مطرح هستند، اما برخی از مشکلات فوق نظیر موارد ۱ و ۲ بسیار عمیق بوده و تاکنون روشی برای حل آنها ارائه نشده، به جز این که مسئله بر اساس دانش زمینه و برای ویدیوهای خاص حل شده است. مسئله شماره ۳ یک مسئله مهم در تحلیل معنایی ویدیو است، اما عملا تحقیقی در این زمینه انجام نشده و تقریبا تمام سیستم‌های فعلی، مبتنی بر پردازش‌های کوتاه مدت می‌باشند. در میان مشکلات مطرح شده، تاکنون مسئله شماره ۴ بیشتر مورد توجه محققین بوده و پیشرفت‌های خوبی در این زمینه بدست آمده است، اما هنوز هم این مسئله نیازمند تحقیقات بیشتر است. در جدول ۲ مسائل باز در این زمینه به طور مختصر مطرح شده و راهکارهای فعلی، میزان اهمیت موضوع و میزان پیچیدگی مسئله به صورت کلی ارائه شده است. در ادامه، این مسائل بیشتر مورد بحث قرار خواهد گرفت.

۴-۱-۱ تنوع رویدادها و مفاهیم در یک ویدیو

در حالت کلی، یک ویدیو می‌تواند شامل رویدادها و مفاهیم متعددی باشد که تشخیص تمام آنها کار پیچیده‌ای است. به همین دلیل عمده تحقیقات انجام شده در زمینه تحلیل معنایی ویدیو، بر روی ویدیوهای خاص (معمولا ویدیوهای ورزشی) تمرکز داشته‌اند. چرا که در ویدیو خاص، تعداد رویدادها و مفاهیم محدود می‌باشد. به عنوان مثال در ویدیو مسابقه فوتبال، تعداد رویدادهایی که احتمال وقوع دارند به مواردی همچون گل، پنالتی، کرنر و ... محدود می‌شود که تعداد آنها انگشت‌شمار است. ضمن این که، صحنه، افراد و اشیا در ویدیوهای ورزشی کاملا محدود و مشخص است. اما در یک ویدیو عمومی (مانند ویدیو سینمایی)،

رویدادهایی بسیار زیادی وجود دارد که ممکن است به وقوع بپیوندد. به همین دلیل برای کاهش پیچیدگی سیستم‌های تحلیل معنایی ویدیوهای عمومی، تعداد رویدادهای قابل آشکارسازی محدود می‌شود، با این وجود غالبا این سیستم‌ها دقت مناسبی ندارند. چرا که ممکن است محتوای ویدیو (افراد، اشیا و صحنه) در هنگام وقوع نمونه‌های مختلفی از یک رویداد معین کاملا متفاوت باشد. بنابراین علت گرایش تحقیقات فعلی به تحلیل ویدیوهای خاص (به ویژه ویدیوهای ورزشی) را می‌توان در دو عامل جستجو کرد: (۱) محدودیت تنوع رویدادها و (۲) محدودیت محتوا. این دو عامل باعث می‌شود استفاده از دانش زمینه و تهیه نمونه‌های آموزشی متنوع و کافی برای سیستم‌های تحلیل معنایی ویدیوهای خاص بسیار ساده‌تر از ویدیوهای عمومی باشد.

۴-۱-۲ امکان وجود معانی و مفاهیم متعدد برای یک رویداد معین

یکی دیگر از بزرگترین مشکلاتی که در بحث شکاف معنایی مطرح می‌شود، امکان وجود چند معنی مختلف برای یک رویداد معین است. علاوه بر این، ممکن است یک رویداد معین در شرایط مختلف، معانی مختلفی به همراه داشته باشد. به عنوان یک مثال ساده، خروج توپ از زمین در مسابقه فوتبال توسط یک بازیکن را مورد بحث قرار می‌دهیم. خروج توپ از خطوط طولی زمین فوتبال به عنوان «اوت» شناخته می‌شود و عموما به دلیل عدم قابلیت بازیکن در کنترل توپ رخ می‌دهد. اما ممکن است همین رویداد مفهوم دیگری نیز داشته باشد. مثلا هنگام آسیب‌دیدن یکی از بازیکنان، به منظور انجام بازی جوانمردانه و مداوای فرد آسیب‌دیده، یکی از بازیکنان به طور عمدی توپ را به خارج از زمین بفرستد. بنابراین نمی‌توان برای رویدادهای به ظاهر یکسان، مفاهیم یکسانی در نظر گرفت. هرچند مثال ذکر شده برای یک ویدیو خاص مطرح شد، اما تعدد معانی و مفاهیم برای یک رویداد معین در یک ویدیو عمومی بسیار وسیع‌تر و پیچیده‌تر خواهد بود. با توجه به تحقیقات فعلی، به نظر می‌رسد رفع چنین مشکلاتی در حال حاضر میسر نباشد.

۴-۱-۳ پردازش‌های بلندمدت برای استخراج مفاهیم

عمده روش‌هایی که برای تحلیل معنایی ویدیو ارائه شده، از پردازش‌های کوتاه‌مدت برای استخراج معانی و مفاهیم استفاده نموده‌اند، در حالی که برای استخراج برخی معانی و مفاهیم لازم است ده‌ها دقیقه قبل از وقوع آن را مورد توجه قرار داد [۶۹]. به عبارت دیگر، معمولا برای استخراج معانی و مفاهیم یک یا چند شات متوالی مورد پردازش قرار می‌گیرد که توالی این شات‌ها شاید بیش از یک دقیقه نباشد، در حالی که برای استخراج دقیق معانی و مفاهیم گاه لازم است تمام شات‌های یک ویدیو بررسی شوند.

گاهی ممکن است وقوع دو رویداد متفاوت با فاصله زمانی بسیار زیاد (بیش از چند دقیقه) در یک ویدیو به یکدیگر وابسته باشد و ترکیب اطلاعات آنها منجر به استخراج معانی و مفاهیم جدید شود. توجه به این نکته، به ویژه برای استخراج معانی و مفاهیم سطح بالا، بسیار مهم به نظر می‌رسد، اما تقریبا در هیچ یک از تحقیقات حال حاضر مورد توجه قرار نگرفته است. تنها در برخی تحقیقات مانند [۷۰]، پردازش‌های نچندان بلند مدت (حدود چند دقیقه) برای تحلیل مسابقات ورزشی ارائه شده است.

۴-۱-۴ استفاده و ترکیب داده‌های چندنوعی

در زمینه استخراج اطلاعات چندنوعی (تصویر، صوت، متن و ...) و ترکیب آنها برای تحلیل معنایی ویدیو تحقیقات زیادی انجام شده که نشان از اهمیت این موضوع دارد. به ویژه استفاده از اطلاعات متنی در کنار داده‌های تصویری، می‌تواند در کاهش شکاف معنایی بسیار موثر باشد. اما در اکثر این تحقیقات، ویژگی‌های سطح پایین از داده‌های چند نوعی استخراج و با یکدیگر ترکیب می‌شود. این موضوع در ترکیب اطلاعات تصویری و صوتی بیشتر به چشم می‌خورد. هرچند ترکیب اطلاعات چندنوعی می‌تواند تا حدود زیادی شکاف معنایی را کاهش دهد، اما ترکیب ویژگی‌های چند نوعی سطح پایین نمی‌تواند کارایی چندانی داشته باشد. بنابراین بحث استخراج ویژگی‌های سطح میانی و سطح بالا از داده‌های تصویری و صوتی و ترکیب آنها، همچنان به عنوان یک موضوع باز مطرح است. به نظر می‌رسد استخراج اطلاعات متنی از ویدیو، به دلیل این که معمولا این اطلاعات حاوی اطلاعات معنایی سطح بالا هستند، بیشتر می‌تواند در کاهش شکاف معنایی موثر باشد.

جدول ۲ عناوین اصلی مهمترین مسائل تحقیقاتی باز در زمینه سیستم‌های تحلیل ویدیو و ویژگی‌های آنها

عنوان مسئله	راهکارهای فعلی	میزان اهمیت	میزان پیچیدگی
تنوع رویدادها و مفاهیم در یک ویدیو	- تحلیل ویدیوهای خاص (به ویژه مسابقات ورزشی) - تحلیل ویدیوهای عمومی با اعمال محدودیت زیاد	●●●●●	●●●●
امکان وجود معانی و مفاهیم متعدد برای یک رویداد معین	- راهکاری ارائه نشده است	●●●●	●●●●●
پردازش‌های بلندمدت برای استخراج معانی و مفاهیم	- به جز موارد بسیار خاص، تقریبا راهکاری ارئه نشده است	●●●	●●●●
استفاده و ترکیب داده‌های چندنوعی	- استفاده و ترکیب داده‌های صوتی و متنی - استفاده از بازخورد کاربر	●●●	●●●

یکی دیگر از راه‌های استفاده از داده‌های چند نوعی، استفاده از اطلاعات مبتنی بر کاربر در کنار داده‌های تصویری است.اطلاعات مبتنی بر کاربر به دلیل این که مستقیما از کاربر گرفته می‌شود، اطلاعات معنایی بسیار غنی داشته و بهترین راه‌حل برای کاهش شکاف معنایی است، اما معایبی دارد که استفاده از این روش را بسیار محدود می‌کند. مهمترین این عیوب عبارتند از: (۱) زمان‌بر بودن دریافت اطلاعات و (۲) خسته‌کننده بودن برای کاربر. یکی دیگر از ویژگی‌های روش استخراج اطلاعات معنایی از طریق کاربر، استخراج اطلاعات مطابق با سلیقه و نظر کاربر است. به عبارت دیگر همان طور که قبلا گفته شد، هر انسانی می‌تواند از یک رویداد معین درک متفاوتی داشته باشد. بنابراین ممکن است اطلاعات معنایی متفاوتی توسط دو فرد از یک رویداد معین استنباط گردد. بنابراین دریافت بازخورد از کاربر به نوعی شخصی‌سازی سیستم تحلیل معنایی را به دنبال دارد. این ویژگی با توجه به نوع کاربرد می‌تواند مزیت یا عیب تلقی شود.

۴-۲ تعامل میان دقت، سرعت و خودکار بودن

میزان پیچیدگی و هزینه طراحی سیستم‌های تحلیل معنایی ویدیویی به عوامل زیادی وابسته است، اما از این میان سه عامل به عنوان عوامل اصلی میزان پیچیدگی سیستم معرفی می‌شوند: (۱) دقت، (۲) سرعت پردازش و (۳) میزان خودکار بودن. سه عامل دقت، سرعت و خودکار بودن را می‌توان به ترتیب متضاد با خطا، تاخیر و میزان دخالت نیروی انسانی دانست. در یک سیستم ایده‌آل انتظار داریم میزان خطا، تاخیر و میزان دخالت نیروی انسانی به صفر نزدیک باشد، اما با توجه به فناوری کنونی، دستیابی به چنین سیستمی میسر نیست. این سه عامل به نحو موثری با یکدیگر ارتباط دارند، به طوری که کاهش یکی، باعث افزایش دو عامل دیگر می‌شود. با توجه به این توضیحات، سیستم تحلیل معنایی را می‌توان طوری طراحی نمود که بر اساس کاربرد، یک مصالحه مناسب میان دقت، سرعت و خودکاربودن سیستم برقرار گردد.

۵ جمع‌بندی

در این مقاله، تحقیقات انجام شده در زمینه تحلیل محتوایی و معنایی ویدیو از دیدگاه ساختار سلسله مراتب معنایی در تولید فیلم مورد بررسی قرار گرفت. دیدگاه اصلی حاکم برای دسته‌بندی و بررسی مقالات، مبتنی بر نحوه تولید فیلم توسط فیلم‌ساز است. بر این اساس، انواع پردازش ویدیو در سه سطح دسته‌بندی گردید که عبارتند از: سطح پایین (پردازش فریم)، سطح میانی (استخراج محتوا) و سطح بالا (استخراج معنا).

در سطح پردازش فریم‌ها، مواردی همچون آشکارسازی مرز بین شات‌ها و انتخاب فریم‌های کلیدی مورد بحث قرار گرفت. به نظر می‌رسد تحقیقات انجام شده در این سطح پردازش با موفقیت‌هایی خوبی همراه بوده است. هرچند هنوز مشکلاتی مانند افزایش دقت آشکارسازی گذار تدریجی در پردازش‌های سطح پایین وجود دارد، اما تحقیقات در این زمینه نزدیک به اشباع است.

پردازش‌های سطح میانی تاثیر بسزایی در کاهش شکاف معنایی میان ویژگی‌های سطح پایین و مفاهیم سطح بالا دارد. موضوع اصلی تحقیقات در این سطح بر آشکارسازی، ردیابی و تشخیص اشیا و افراد متمرکز شده است. مهمترین مشکلات در این زمینه، آشکارسازی و ردیابی اشیا و افراد در شرایط مختلف است. روش‌های فعلی در مواجهه با این مشکلات هنوز از دقت و سرعت مناسب برخوردارنیستند. هرچند تحقیقات خوبی در این زمینه‌ها انجام شده، اما در اغلب موارد، الگوریتم‌های ارائه شده وابسته به دانش زمینه یا شرایط خاص هستند و نمی‌توانند به اندازه کافی کارامد باشند. بنابراین هنوز تحقیقات بیشتری در زمینه پردازش‌های سطح میانی با هدف ارائه الگوریتم‌های کارامد لازم است.

پردازش‌های سطح بالا شامل آشکارسازی و تشخیص واحدهای معنایی ویدیو و تشخیص گونه و زیرگونه ویدیو است. با توجه به این که هنوز الگوریتم‌های فعلی برای پردازش‌های سطح میانی چندان کارامد نیست، پردازش سطح بالا و استخراج اطلاعات معنایی از ویدیو با مشکلات زیادی روبرو است. چرا که کارایی الگوریتم‌های سطح بالا وابسته به کارایی الگوریتم‌های سطح میانی در پردازش و استخراج اطلاعات هستند. بنابراین، عمده تحقیقات فعلی در زمینه استخراج اطلاعات سطح بالا، بدون استفاده از پردازش‌های سطح میانی (استخراج محتوا) صورت گرفته است. به عبارت دیگر، معمولا مفاهیم سطح بالا بر اساس ویژگی‌های سطح پایین استخراج می‌شوند که این رویکرد باعث ایجاد شکاف معنایی شده است. بنابراین مهمترین چالش در تحقیقات فعلی، شکاف معنایی میان ویژگی‌های سطح پایین و سطح بالاست.

مسائل و مشکلات باز در زمینه شکاف معنایی به چهار دسته کلی تقسیم گردید: (۱) تنوع رویدادها و مفاهیم در یک ویدیو، (۲) امکان وجود معانی و مفاهیم متعدد برای یک رویداد معین، (۳) پردازش‌های بلندمدت برای استخراج معانی و مفاهیم و (۴) استفاده و ترکیب داده‌های چندنوعی. این مسائل و مشکلات در سطح پردازش میانی و سطح بالا تعریف می‌شوند. به عبارت دیگر، برای حل مسئله شکاف معنایی، ابتدا باید به دنبال ارائه الگوریتم‌های کارامد در سطح میانی بود. سپس بر اساس نوع اطلاعات استخراج شده از پردازش‌های سطح میانی، الگوریتم‌های مناسب در سطح بالا ارائه شود.

قدردانی

نویسندگان مقاله از سرکار خانم محبوبه کهخائی جوان بابت نظرات ایشان در ویرایش ادبی مقاله سپاس‌گزاری می‌نمایند

مراجع

[1] Cees G. M. Snoek, Marcel Worring, "*Multimodal Video Indexing: A Review of the State-of-the-art*", Multimedia Tools and Applications, vol. 25, no. 1, pp. 5-35, 2005.

[2] Arthur G. Money, Harry Agius, "*Video Summarisation: A Conceptual Framework and Survey of the State of the Art*", Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121-143, 2008.

[3] Tomi D. Raty, "*Survey on Contemporary Remote Surveillance Systems for Public Safety*", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 5, pp. 493-515, 2010.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang, "*Image Retrieval: Ideas, Influences, and Trends of the New Age*", ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, April, 2008.

[5] T. D'Orazio, M. Leo, P. Spagnolo, M. Nitti, N. Mosca, "*A Visual System for Real Time Detection of Goal Events During Soccer Matches*", Computer Vision and Image Understanding, vol. 113, no. 5, pp. 622-632, May, 2009.

[6] T. D'Orazio, M. Leo, P. Spagnolo, P. L. Mazzeo, N. Mosca, M. Nitti, A. Distante, "*An Investigation into the Feasibility of Real-Time Soccer Offside Detection From a Multiple Camera System*", IEEE Transactions on Circuits And Systems for Video Technology, vol. 19, no. 12, pp. 1804-1818, December, 2009.

[7] Ali Amiri, Mahmood Fathy, "*Video Shot Boundary Detection Using QR-Decomposition and Gaussian Transition Detection*", EURASIP Journal on Advances in Signal Processing, doi: 10.1155/2009/5094382009.

[8] Aissa Saoudi, Hassane Essafi, "*Spatio-Temporal Video Slice Edges Analysis for Shot Transition Detection and Classification*", World Academy of Science, Engineering and Technology, vol. 28, pp. 45-50, 2007.

[9] زینب زینال‌پور تبریزی، امیرفرید امینیان‌مدرس، محمود فتحی، محمدرضا جاهدمطلق، محسن سریانی، "**خوشه‌بندی فریم‌های ویدیو به کمك ویژگی فرکتالی به منظور تشخیص مرز شات**"، پانزدهمین کنفرانس ملی انجمن کامپیوتر ایران، تهران، ۱۳۸۸.

[10] Partha Pratim Mohanta, Sanjoy Kumar Saha, Bhabatosh Chanda, "*A Model-Based Shot Boundary Detection Technique Using Frame Transition Parameters*", IEEE Transactions on Multimedia, vol. 14, no. 1, pp. 223-233, 2012.

[11] Markos Mentzelopoulos, Alexandra Psarrou, "*Key-Frame Extraction Algorithm using Entropy Difference*", International Multimedia Conference, New York, NY, USA, pp. 39-45, 2004.

[12] Xianglin Zeng, WeimingHu, Wanqing Liy, Xiaoqin Zhang, Bo Xu, "*Key-Frame Extraction using Dominant-Set Clustering*", International Conference on Multimedia and Expo, Hannover, Germany, pp. 1285-1288 2008.

[13] Ali Amiri, Mahmood Fathy, "*Hierarchical Keyframe-based Video Summarization Using QR-Decomposition and Modified k-Means Clustering*", EURASIP Journal on Advances in Signal Processing, doi: 10.1155/2010/8921242010.

[14] Maheshkumar H. Kolekar, Kannappan Palaniappan, Somnath Sengupta, Gunasekaran Seetharaman, "*Semantic Concept Mining Based on Hierarchical Event Detection for Soccer Video Indexing*", Journal of Multimedia, vol. 4, no. 5, pp. 298-312, October, 2009.

[15] Ahmet Ekin, A. Murat Tekalp, Rajiv Mehrotra, "*Automatic Soccer Video Analysis and Summarization*", IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 796-807, July, 2003.

[16] V. Pallavi, Jayanta Mukherjee, Arun K. Majumdar, Shamik Sural, "*Ball Detection From Broadcast Soccer Videos Using Static and Dynamic Features*", Journal of Visual Communication & Image Representation, vol. 19, pp. 426–436, 2008.

[17] Eui-Jin Kim, Gwang-Gook Lee, Cheolkon Jung, Sang-Kyun Kim, Ji-Yeun Kim, Whoi-Yul Kim, "*A Video Summarization Method for Basketball Game*", Pacific Rim Conference on Multimedia, Jeju Island, Korea, pp. 765-775, November, 2005.

[18] Soo-Chang Pei, Fan Chen, "*Semantic Scenes Detection and Classification in Sports Videos*", International Conference on Computer Vision, Graphics and Image Processing, Kinmen, China, pp. 210-217, August, 2003.

[19] Jia Liu, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, Hongqi Wang, "*Automatic Player Detection, Labeling and Tracking in Broadcast Soccer Video*", Pattern Recognition Letters, vol. 30, no. 2, pp. 103-113, 2009.

[20] کاوه کنگرلو، احسان ا... کبیر، "**آشکارسازی بازیکنان در تصاویر فوتبال**"، چهارمین کنفرانس ماشین بینایی و پردازش تصویر ایران، مشهد، بهمن ۱۳۸۵.

[21] H. S. Khatoonabadi, M. Rahmati, "*Automatic Soccer Players Tracking in Goal Scenes by Camera Motion Elimination*", Image and Vision Computing, vol. 27, pp. 469-479, 2009.

[22] Pascual J. Figueroa, Neucimar J. Leite, Ricardo M. L. Barros, "*Tracking Soccer Players Aiming Their Kinematical Motion Analysis*", Computer Vision and Image Understanding, vol. 101, no. 2, pp. 122-135, 2006.

[23] Chung-Lin Huang, Huang-Chia Shih, Chung-Yuan Chao, "*Semantic Analysis of Soccer Video Using Dynamic Bayesian Network*", IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 749-760, August, 2006.

[24] Monireh-Sadat Hosseini, Amir-Masoud Eftekhari Moghadam, "*Fuzzy Rule-based Reasoning Approach for Event Detection and Annotation of Broadcast Soccer*", Applied Soft Computing, vol. 13, no. 2, pp. 846-866, 2013.

[25] Vahid Kiani, Hamid Reza Pourreza, "*An Effective Slow-motion Detection Approach for Compressed Soccer Videos*", ISRN Machine Vision, doi: 10.5402/2012/9595082012.

[26] Vahid Babaee-Kashany, Hamid Reza Pourreza, "*Camera Pan and Tilt Estimation in Soccer Scenes Based on Vanishing Points*", UKSim Fourth European Modelling Symposium on Computer Modelling and Simulation, pp. 152-157, 2010.

[27] Lexing Xie, Hari Sundaram, Murray Campbell, "*Event Mining in Multimedia Streams*", Proceedings of the IEEE, vol. 96, no. 4, pp. 623-647, April, 2008.

[28] Colin O'Toole, Alan Smeaton, Noel Murphy, Sean Marlow, "*Evaluation of Automatic Shot Boundary Detection on a Large Video Test Suite*", UK Conference on Image Retrieval: The Challenge of Image Retrieval, Newcastle, UK, pp., February, 1999.

[29] Bing Han, Xinbo Gao, Hongbing Ji, "*A Shot Boundary Detection Method for News Video Based on Rough-Fuzzy Sets*", International Journal of Information Technology, vol. 11, no. 7, pp. 101-111, 2005.

[30] Zeeshan Rasheed, Mubarak Shah, "*Scene Detection In Hollywood Movies and TV Shows*", Conference on Computer Vision and Pattern Recognition, Wisconsin, USA, pp., June, 2003.

[31] Jianrong Cao, Anni Cai, "*A Robust Shot Transition Detection Method based on Support Vector Machine in Compressed Domain*", Pattern Recognition Letters, vol. 28, no. 12, pp. 1534–1540, 2007.

[32] C. Sujatha, Uma Mudenagudi, "*A Study on Keyframe Extraction Methods for Video Summary*", International Conference on Computational Intelligence and Communication Networks, Gwalior, India, pp. 73-77, 2011.

[33] Evaggelos Spyrou, Giorgos Tolias, Phivos Mylonas, Yannis Avrithis, "*Concept Detection and Keyframe Extraction using a Visual Thesaurus*", Multimedia Tools and Applications, vol. 41, no. 3, pp. 337–373, 2009.

[34] Matko Saric, Hrvoje Dujmic, Domagoj Baricevic, "*Shot Boundary Detection in Soccer Video using Twin-comparison Algorithm and Dominant Color Region*", Journal of Information and Organizational Sciences, vol. 32, no. 1, pp. 67-73, 2008.

[35] Vasileios T. Chasanis, Aristidis C. Likas, Nikolaos P. Galatsanos, "*Scene Detection in Videos Using Shot Clustering and Sequence Alignment*", IEEE Transactions on Multimedia, vol. 11, no. 1, pp. 89-100, January, 2009.

[36] Ja-Hwung Su, Yu-Ting Huang, Hsin-Ho Yeh, Vincent S. Tseng, "*Effective Content-based Video Retrieval using Pattern-Indexing and Matching Techniques*", Expert Systems with Applications, vol. 37, no. 7, pp. 5068-5085, July, 2010.

[37] Milind Ramesh Naphade, Igor V. Kozintsev, Thomas S. Huang, "*A Factor Graph Framework for Semantic Video Indexing*", IEEE Transactions on Circuits And Systems for Video Technology, vol. 12, no. 1, pp. 40-52, January, 2002.

[38] Milind Ramesh Naphade, Thomas S. Huang, "*A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval*", IEEE Transactions on Multimedia, vol. 3, no. 1, pp. 141-151, March, 2001.

[39] Cheng-Chang Lien, Chiu-Lung Chiang, Chang-Hsing Lee, "*Scene-based Event Detection for Baseball Videos*", Journal of Visual Communication and Image Representation, vol. 18, no. 1, pp. 1-14, February, 2007.

[40] V. Vijayakumar, R. Nedunchezhian, "*A Study on Video Data Mining*", International Journal of Multimedia Information Retrieval, vol. 1, no. 3, pp. 153-172, October, 2012.

[41] Marco Alexander Treiber, *An Introduction to Object Recognition: Selected Algorithms for a Wide Variety of Application*, 1st ed.: Springer, 2010.

[42] Jila Hosseinkhani, Hamid Soltanian-Zadeh, Mahmoud Kamarei, Oliver Staadt, "*Ball Detection with the Aim of Corner Event Detection in Soccer Video*", International Symposium on Parallel and Distributed Processing with Applications Workshops Busan, South Korea, pp. 147-152, May, 2011.

[43] Zhenxing Niu, Xinbo Gao, Qi Tian, "*Tactic Analysis based on Real-World Ball Trajectory in Soccer Video*", Pattern Recognition, vol. 45, no. 5, pp. 1937–1947, 2012.

[44] Hossam M. Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien, Tai-hoon Kim, "*Event Detection Based Approach for Soccer Video Summarization Using Machine learning*", International Journal of Multimedia and Ubiquitous Engineering, vol. 7, no. 2, pp. 1-18, 2012.

[45] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, Mubarak Shah, "*High-Level Event Recognition in Unconstrained Videos*", International Journal of Multimedia Information Retrieval, vol. 2, no. 2, pp. 73-101, June, 2013.

[46] Haoran Yi, Deepu Rajan, Liang-Tien Chia, "*Semantic Video Indexing and Summarization Using Subtitles*", Pacific Rim Conference on Multimedia-Advances in Multimedia Information Processing, Tokyo, Japan, pp. 634-641, December, 2004.

[47] Xueming Qian, Huan Wang, Guizhong Liu, Xingsong Hou, "*HMM based Soccer Video Event Detection using Enhanced Mid-level Semantic*", Multimedia Tools and Applications, vol. 60, no. 1, pp. 233-255, 2012.

[48] محمدحسین سیگاری، حمید سلطانیان زاده،حمیدرضا پوررضا، "**تحلیل ویدیو اخبار به منظور آشکارسازی و تشخیص چهره**

گوینده خبر و آشکارسازی مرز بخش‌های خبری"، شانزدهمین کنفرانس ملی انجمن کامپیوتر ایران، تهران، بهمن ۱۳۸۹.

[49] Haojie Li, Jinhui Tang, Si Wu, Yongdong Zhang, Shouxun Lin, "*Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences*", IEEE Transactions on Circuits And Systems for Video Technology, vol. 20, no. 3, pp. 351-364, March, 2010.

[50] J. K. Aggarwal, M. S. Ryoo, "*Human Activity Analysis: A Review*", ACM Computing Surveys, vol. 43, no. 3, pp. 1-47, April, 2011.

[51] Juan Rafael Nunez, Jacques Facon, Alceu de Souza Brito Junior, "*Soccer Video Segmentation: Referee and Player Detection*", International Conference on Systems, Signals and Image Processing, Bratislava, Slovakia, pp. 279-282, June, 2008.

[52] احسان پازوکی، محمد رحمتی، "**ردیابی خودکار بازیکنان در یک مسابقه فوتبال مبتنی بر استخراج مدل رنگی یکپارچه**"، پنجمین کنفرانس ماشین بینایی و پردازش تصویر، تبریز، آبان ۱۳۸۷.

[53] Dian W. Tjondronegoro, Yi-Ping Phoebe Chen, "*Knowledge-Discounted Event Detection in Sports Video*", IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, vol. 40, no. 5, pp. 1009-1024, September, 2010.

[54] Yifan Zhang, Changsheng Xu, YongRui, Jinqiao Wang, Hanqing Lu, "*Semantic Event Extraction from Basketball Games using Multi-Modal Analysis*", International Conference on Multimedia and Expo, Beijing, China, pp. 2190-2193, July, 2007.

[55] Meng Wang, Xian-Sheng Hua, Tao Mei, Richang Hong, Guojun Qi, Yan Song, Li-Rong Dai, "*Semi-supervised Kernel Density Estimation for Video Annotation*", Computer Vision and Image Understanding, vol. 113, no. 3, pp. 384-396, March, 2009.

[56] Mei-Ling Shyu, Zongxing Xie, Min Chen, Shu-Ching Chen, "*Video Semantic Event/Concept Detection Using a Subspace-Based Multimedia Data Mining Framework*", IEEE Transactions on Multimedia, vol. 10, no. 5, pp. 252-259, February, 2008.

[57] Do Joon Jung, Se Hyun Park, Hang Joon Kim, "*Event-Based Surveillance System for Efficient Monitoring*", Pacific Rim Conference on Multimedia, Tokyo, Japan, pp. 641-648, 2004.

[58] Juan C. SanMiguel, Jose M. Martinez, "*A Semantic-based Probabilistic Approach for Real-Time Video Event Recognition*", Computer Vision and Image Understanding, vol. 116, no. 9, pp. 937-952, 2012.

[59] Maryam Koohzadi, Mohammad Reza Keyvanpour, "*An Analytical Framework for Event Mining in Video Data*", Artificial Intelligence Review, doi: 10.1007/s10462-012-9315-52013.

[60] A. Dyana, Sukhendu Das, "*MST-CSS (Multi-Spectro-Temporal Curvature Scale Space), a Novel Spatio-Temporal Representation for Content-Based Video Retrieval*", IEEE Transactions on Circuits And Systems for Video Technology, vol. 20, no. 8, pp. 1080-1094, August, 2010.

[61] Junyong You, Guizhong Liu, Andrew Perkis, "*A Semantic Framework for Video Genre Classification and Event Analysis*", Signal Processing: Image Communication, vol. 25, no. 4, pp. 287-302, April, 2010.

[62] Mostafa Tavassolipour, Mahmood Karimian, Shohreh Kasaei, "*Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula*", IEEE Transactions on Circuits And Systems for Video Technology, 2013.

[63] Mohamad Hoseyn Sigari, Samaneh Abbasi Sureshjani, Hamid Soltanian-Zadeh, "*Sport Video Classification using an Ensemble Classifier*", Iranian Machine Vision and Image Processing, Tehran, Iran, pp. 1-4, November, 2011.

[64] Edda Leopold, Jorg Kindermann, "*Content Classification of Multimedia Documents using Partitions of Low-Level Features*", Journal of Virtual Reality and Broadcasting, vol. 3, no. 6, 2006.

[65] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian, "*Affective Visualization and Retrieval for Music Video*", IEEE Transactions on Multimedia, vol. 12, no. 6, pp. 510-522, October, 2010.

[66] Jinjun Wang, Changsheng Xu, Engsiong Chng, "*Automatic Sports Video Genre Classification using Pseudo-2D-HMM*", International Conference on Pattern Recognition, Hong Kong, pp. 778-781, August, 2006.

[67] Xiaoyu Zhang, Changsheng Xu, Jian Cheng, Hanqing Lu, Songde Ma, "*Effective Annotation and Search for Video Blogs with Integration of Context and Content Analysis*", IEEE Transactions on Multimedia, vol. 11, no. 2, pp. 272-285, February, 2009.

[68] Stanislas Oger, Mickael Rouvier, Georges Linares, "*Transcription-based Video Genre Classification*", International Conference on Acoustics Speech and Signal Processing, Dallas, TX, USA, pp. 5114-5117, 2010.

[69] Nevenka Dimitrova, "*Context and Memory in Multimedia Content Analysis*", IEEE MultiMedia, vol. 11, no. 3, pp. 7-11 2004.

[70] Guangyu Zhu, Changsheng Xu, Qingming Huang, Yong Rui, Shuqiang Jiang, Wen Gao, Hongxun Yao, "*Event Tactic Analysis Based on Broadcast Sports Video*", IEEE Transactions on Multimedia, vol. 11, no. 1, pp. 49-67, January, 2009.

محمدحسین سیگاری مدرک کارشناسی و کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر، به ترتیب در سال ۱۳۸۵و۱۳۸۷ از دانشگاه فردوسی مشهد و دانشگاه علم و صنعت ایران با کسب رتبه اول دریافت نمود. هم اکنون ایشان دانشجوی مقطع دکتری تخصصی در رشته مهندسی کامپیوتر، گرایش هوش مصنوعی در دانشکده مهندسی برق و کامپیوتر دانشگاه تهران می‌باشد. علاقه‌مندی‌های علمی ایشان شامل پردازش تصویر، بینایی ماشین، شناسایی الگو، بیومتریک، سیستم‌های تحلیل معنایی ویدیو و کاربردهای پردازش تصویر در صنعت می‌باشد.

حمید سلطانیان‌زاده مدرک کارشناسی پیوسته خود را در رشته مهندسی برق گرایش الکترونیک در سال ۱۳۶۵ از دانشگاه تهران با کسب رتبه اول دریافت نمود. سپس ایشان مدارک کارشناسی‌ارشد و دکتری خود را در رشته مهندسی برق به ترتیب در گرایش های کنترل و پردازش سیگنال و بیوالکتریک در سال‌های ۱۳۶۹و۱۳۷۱ از دانشگاه میشیگان آمریکا دریافت نمود. هم اکنون ایشان استاد دانشکده مهندسی برق و کامپیوتر دانشگاه تهران و محقق ارشد گروه رادیولوژی بیمارستان هنری فورد آمریکا می‌باشد. علاقه‌مندی‌های علمی ایشان شامل تصویربرداری پزشکی، پردازش تصاویر و سیگنال‌های پزشکی، پردازش تصویر، بینایی ماشین و شناسایی الگو می‌باشد.

حمیدرضا پوررضا مدرک کارشناسی خود را در رشته مهندسی برق گرایش الکترونیک در سال ۱۳۶۸ از دانشگاه فردوسی مشهد کسب کرد. سپس ایشان مدارک کارشناسی‌ارشد و دکتری خود را به ترتیب در رشته مهندسی برق گرایش الکترونیک و مهندسی کامپیوتر گرایش هوش مصنوعی در سال‌های ۱۳۷۲و۱۳۸۲ از دانشگاه امیرکبیر دریافت نمود. هم اکنون ایشان دانشیار گروه مهندسی کامپیوتر دانشکده مهندسی دانشگاه فردوسی مشهد می‌باشد. علاقه‌مندی‌های علمی ایشان پردازش تصویر، بینایی ماشین، شناسایی الگو و سیستم‌های حمل و نقل هوشمند می‌باشد.