
بسم الله الرحمن الرحیم

بسمه‌تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

پایان‌نامه دکتری

گرایش مخابرات سیستم

عنوان:

تحلیل و مدل‌سازی پارامتریک سیگنال ویدئو

نگارنده:

مونا امیدیگانه

استاد راهنما:

آقای دکتر شاهرخ قائم‌مقامی

استاد مشاور:

آقای دکتر شروین شیرمحمدی

دی‌ماه ۱۳۹۰

این رساله با عنوان

تحلیل و مدل‌سازی پارامتریک سیگنال ویدئو

به عنوان بخشی از شرایط احراز درجه دکترای تخصصی (Ph.D.) با حضور اساتید ذیل در تاریخ ۱۳۹۰/۱۰/۲۸ توسط خانم مونا امیدیگانه ارائه و مورد تأیید قرار گرفت.

۱- جناب آقای دکتر شاهرخ قائم‌مقامی (استاد راهنما)

۲- جناب آقای دکتر فرخ مروستی

۳- جناب آقای دکتر مسعود بابایی‌زاده

٤- جناب آقای دکتر عمادالدین فاطمی‌زاده

۵- سرکار خانم دکتر شهره کسایی (استاد مدعو خارج دانشکده)

٦- جناب آقای دکتر محمد قنبری (استاد مدعو خارج دانشگاه)

۷- جناب آقای دکتر حسن قاسمیان (استاد مدعو خارج دانشگاه)

۸- جناب آقای دکتر شهرام معین (استاد مدعو خارج دانشگاه)

سپاسگزاری

پیش از همه شاکر خدا هستم که قسمتی از زیبایی‌های بیکران علم را به من نشان داد و امیدوارم این فرصت را داشته باشم تا در این راه بی‌پایان گام بردارم.

از استادان محترم و معزز گروه مخابرات سیستم دانشگاه شریف که طی دوران تحصیلم در محضر ایشان تلمذ کردم و از خرمن دانش، راهنمایی و تشویق‌های سودمند آن‌ها بهره بسیار بردم، سپاسگزارم.

از استاد معظم جناب آقای دکتر قائم‌مقامی که با راهنمایی‌های ارزشمند خود بنده را در تدوین و نگارش رساله یاری رساندند و همواره با تواضع و سعه‌صدر اطلاعات با ارزشی در اختیارم گذاشتند، از صمیم قلب تشکر و قدردانی می‌نمایم.

از مساعدت و راهنمایی‌های استاد بزرگوار جناب آقای دکتر شیرمحمدی، استاد مشاور رساله نیز سپاسگزارم.

مراتب سپاس خود را به محضر اساتید گرامی داور جناب آقای دکتر مروستی، جناب آقای دکتر قاسمیان، جناب آقای دکتر قنبری، سرکار خانم دکتر کسایی، جناب آقای دکتر بابایی‌زاده، جناب آقای دکتر معین و جناب آقای دکتر فاطمی‌زاده که قبول زحمت نمودند و داوری این کار را بر عهده گرفتند، اعلام می‌دارم.

در انتهای سخن بر خود واجب می‌دانم از تمامی استادان، دوستان و همکارانی که با راهنمایی‌ها، کمک‌های فکری و سایر مساعدت‌ها مرا در نوشتن این رساله یاری رساندند، صمیمانه تشکر کنم.

مدیون و سپاسگزار پدر، مادر و همسر گرامی‌ام نیز هستم که همواره شرایط و امکانات تحصیل، تحقیق و پژوهش را برایم فراهم آوردند.

در پایان، این رساله را تقدیم می‌کنم به:

تمام انسان‌های آزاده، که در طول تاریخ همواره از راحت و آسایش خود در راه خدا و برای ائتلای انسانیت گذشتند.

چکیده

سیگنال ویدئو در بسیاری از سیستم‌های مخابراتی، پزشکی، آموزشی، تجاری و تفریحی کاربرد دارد. رشد روزافزون استفاده از این سیگنال، نیاز به مطالعه و بررسی دقیق و نظری آن را افزایش می‌دهد. درنتیجه تحلیل ویدئو یکی از مباحث مهم و اساسی در دنیای تحقیقاتی امروزه به شمار رفته، نقش مهمی در بهبود کارایی سیستم‌های نامبرده ایفا می‌کند. سیگنال ویدئوی دیجیتال – شامل نمونه‌های دیجیتال در بعد زمان و مکان – دارای افزونگی زیادی در هر دو بعد مذکور است. با مدل‌سازی و استخراج پارامترهای ویدئو، سیگنال تنکی[1] به دست می‌آید که حاوی اطلاعات مهمی از سیگنال اصلی است و می‌تواند برای بهبود عملکرد سیستم‌های پردازش ویدئو استفاده شود. از کاربردهای مدل‌سازی آماری ویدئو می‌توان فشرده‌سازی ویدئو[2]، تشخیص رفتار انسان[3]، بازیابی ویدئو[4]، چکیده‌سازی ویدئو[5]، انتقال ویدئو، شاخص‌گذاری ویدئو[6] و خوشه‌بندی ویدئو[7] را نام برد.

در این رساله به بررسی مساله تحلیل و مدل‌سازی پارامتریک سیگنال ویدئو پرداخته شده است. برای این منظور سیگنال ویدئو از دو دیدگاه مورد توجه قرار می‌گیرد. در دیدگاه اول، سیر تحول زمانی فریم‌های ویدئویی که حاوی مشخصات مکانی[8] سیگنال هستند در نظر گرفته می‌شوند. تحول زمانی پارامترهای مکانی بر اساس خصوصیات آماری تبدیل دوبعدی موجک[9] فریم‌ها که مطابق با خصوصیات سیستم بینایی انسان است، مورد بررسی قرار می‌گیرد. برای آنالیز روابط زمانی و تحلیل روند تغییرات این پارامترها از سه روش اندازه‌گیری فاصله بین فریم‌های متوالی، تجزیه زمانی و مدل-سازی AR استفاده شده است. در روش اول با کمک معیار فاصله KL[10] بین پارامترهای مکانی، تحولات زمانی فریم‌ها بررسی می‌شود و نتایج این تحلیل‌ها برای انتخاب مرز شات‌های آنی و تدریجی، خوشه‌بندی ویدئو و چکیده‌سازی ویدئو استفاده می‌شوند. روش دوم، به تجزیه زمانی پارامترهای مکانی اختصاص دارد و بدون نیاز به انتخاب مرز شات‌ها و خوشه‌بندی ویدئو به چکیده‌سازی ویدئو بر اساس رخدادهای بصری دنباله ویدئو می‌پردازد. در روش سوم، با کمک مدل AR، روابط زمانی بین پارامترهای مکانی متناظر فریم‌های ویدئو مطالعه می‌شوند و انتخاب مرز شات‌ها، خوشه‌بندی ویدئو و تعیین فریم‌های کلیدی بر اساس خطای تخمین مدل انجام می‌گیرد. برای ارزیابی این روش‌ها از آزمون‌های مختلف ادراکی[11] و تحلیلی[12] بر روی دادگان‌های TRECVID و Hollywood2 استفاده می‌شود. نتایج آزمون‌ها کارایی و قدرت بالای این روش‌ها را تایید می‌نماید.

در دیدگاه دوم، پارامترهای مکانی-زمانی[13] سیگنال ویدئو یکجا از نتایج تحلیل خواص آماری تبدیل سه‌بعدی موجک ویدئو استخراج می‌شوند. پارامترهای آماری حاشیه‌ای[14] و توأم[15] مستخرج از ویدئو برای تشخیص فعالیت انسان و تعیین سطح فعالیت در ویدئو استفاده می‌شوند. نتایج آزمون-های ادراکی و تحلیلی بر روی دادگان‌های Hollywood2 و KTH بیان‌گر توانایی بالای تحلیل ارائه‌شده است. معرفی و استفاده از پارامترهای مکانی بر اساس خواص آماری حاشیه‌ای تبدیل موجک، استفاده از فاصله KL برای بررسی روند تغییرات زمانی ویدئو، معرفی روش تجزیه زمانی برای بیان روابط پیچیده بین رخدادهای دیداری ویدئو و تحلیل و بررسی خواص آماری حاشیه‌ای و توأم تبدیل سه بعدی موجک از جمله نوآوری‌های این رساله هستند که نقش اساسی را در بهبود عملکرد سیستم‌های مورد آزمون داشته‌اند.

کلمات کلیدی: تحلیل و مدل‌سازی پارامتریک ویدئو[16]، تجزیه زمانی[17]، معیار فاصله[18]، مدل AR[19]، مدل آماری تبدیل سه‌بعدی موجک[20]، تعیین فریم-های کلیدی[21] ویدئو، تشخیص مرز شات‌ها[22]، تشخیص رفتار انسان.

---

[1] Sparse
[2] Video compression
[3] Human action recognition
[4] Video retrieval
[5] Abstraction
[6] Video indexing
[7] Video clustering
[8] Spatial
[9] Wavelet
[10] Kullback-Leibler distance
[11] Subjective
[12] Objective
[13] Spatio-temporal
[14] Marginal
[15] Joint
[16] Parametric video analysis and modeling
[17] Temporal decomposition
[18] Distance measure
[19] Autoregressive modeling
[20] 3D wavelet statistical modeling
[21] Keyframe selection
[22] Shot boundary detection

فهرست مطالب

فهرست جداول

فهرست شکل‌ها

فهرست اختصارات

APE	Auto-Regressive Prediction Errors
APE	Auto-Regressive Prediction Errors ratio
AR	Auto-Regressive
BoF	Block of Frames
EM	Expectation Maximization
DC	Direct Current (the transformed coefficient corresponding to zero frequency)
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
GGD	Generalized Gaussian Density
GMM	Gaussian Mixture Model
GoP	Group of Pictures
HMM	Hidden Markov Model
HMT	Hidden Markov Tree
HSV	Hue, Saturation, Value (coordinates of a color space)
HVS	Human Visual System
i.i.d.	Independent and Identically Distributed
ITU-R	International Telecommunication Union- Radio communication Sector
ITU-T	International Telecommunication Union-Telecommunication Sector
JPEG	Joint Photographic Experts Group
KLD	Kullback-Leibler Distance
Lab	CIE specification that attempts to make the luminance scale more perceptually uniform.
LMS	Least Mean Squares algorithm
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
PDF	Probability Density Function
PSNR	Peak Signal to Noise Ratio
RLS	Recursive Least Squares algorithm
SPIHT	Set Partitioning in Hierarchical Trees
TD	Temporal Decomposition
WD-HMM	Wavelet Domain Hidden Markov Model

فهرست واژه‌ها

Bin	نقطه
Dependent	وابسته، نامستقل
Classifier	طبقه‌بند
Confusion Matrix	ماتریس اغتشاش
Correlated	همبسته
Cousin	عموزاده
Cut	برش
Database	دادگان
Dissolve	حل‌شدگی
Fade in/out	محوشدگی
Generalized Gaussian Density	مدل تعمیم‌یافته گوسی
Human Action Recognition	تشخیص رفتار انسان
Joint Statistics	خواص آماری توأم
Lossless	بدون اتلاف
Lossy	با اتلاف
Marginal Statistics	خواص آماری حاشیه‌ای
Mutual Information	اطلاعات متقابل
Objective	تحلیلی
Orthonormal	عمود یکه
Parent	والد
Precision	دقت
Recall	بازخوانی
Sparse	تنک
Spatio-Temporal Parameters	پارامترهای مکانی-زمانی
Subjective	ادراکی
Support Vector	بردار پشتیبان
Video Abstraction	چکیده‌سازی ویدئو
Video Clustering	خوشه‌بندی ویدئو
Video Compression	فشرده‌سازی ویدئو
Video indexing	شاخص‌گذاری ویدئو
Video Retrieval	بازیابی ویدئو
Visual Event	رخداد بصری
Wavelet Transform	تبدیل موجک

فصل اول

مقدمه

انگیزه‌های تحقیق

چالش‌ها

نوآوری‌ها

ساختار رساله

مقالات مستخرج از رساله

با افزایش قدرت پردازش سیستم‌ها و پهنای باند شبکه‌های ارتباطی، چه در کامپیوترهای رومیزی و چه در سیستم‌های قابل حمل و همراه، کاربردهای ویدئو و سرویس‌های ارائه‌شده، مانند آنچه YouTube، Google Video و سرویس‌های دیگر ارائه می‌دهند همه‌روزه در حال پیشرفت هستند و به جزئی از زندگی روزانه مردم تبدیل شده‌اند. همزمان، کاربران توقعات بیشتری از این سرویس‌ها دارند و انتظار دارند که سیستم‌های هوشمندتر و تعامل پذیرتر را در اختیار داشته باشند.

تحلیل محتوای ویدئو[1] یکی از زمینه‌های فعال تحقیق در سال‌های اخیر به شمار می‌رود. سیگنال ویدئو که شامل دنباله‌ای از تصاویر است، دارای ویژگی‌های مکانی و زمانی است. در برخی مدل‌ها روش‌های ترکیبی ویژگی‌های زمانی- مکانی[2]، برای تحلیل محتوای ویدئو به کار رفته است [1, 2, 3]. دیدگاه زمانی- مکانی روشی قدرتمند در بسیاری از کاربردهای تحلیل محتوای ویدئو است. در [4] یک نمایش یک‌بعدی از فریم‌های ویدئو با کمک تبدیل Mojette ارائه شده است. از این روش در کاربردهایی نظیر تخمین حرکت، تشخیص تغییر صحنه و استخراج نواحی دلخواه استفاده شده است. در [5] شات‌ها با کمک موزاییک‌های یک‌بعدی بر اساس تصویر[3] اشعه X بر روی فریم‌های ویدئو – که نمایانگر مجموع مقادیر در جهات عمودی و افقی هستند- تقسیم شده‌اند. مشکل اساسی این روش‌ها انتخاب توصیف‌گرهای زمانی و مکانی مناسب با توجه به کاربرد روش است. راه دیگر مدل‌سازی و توصیف ارتباطات زمانی سیگنال ویدئو، استفاده از ویژگی‌های زمانی مستقیم نظیر جریان نوری[4] [2] و بردارهای حرکت است [6, 7, 8]. در [2] از یک توصیف‌گر محلی برای نمایش نواحی دلخواه و اطلاعات زمانی آن با جریان نور استفاده شده است. در [6] میدان‌های حرکت به صورت یک سیگنال

[1] Video Content Analysis
[2] Spatio-Temporal Features
[3] Projection
[4] Optical Flow

متمایز مانند سری‌های زمانی در نظر گرفته می‌شود و سازوکاری برای فیلترکردن میدان‌های حرکت به کار می‌رود. در [3] برش‌های زمانی- مکانی برای ارائه الگوهای حرکت به کار می‌روند و فریم کلیدی با بردارهای حرکت استخراج می‌شود. بسیاری از این روش‌ها توانایی درنظرگرفتن ارتباطات زمانی طولانی را ندارند و تعیین کمی این ارتباطات بسیار مشکل است. برای کاربردهایی نظیر کلاس‌بندی شات‌ها، بازیابی ویدئو و شاخص‌گذاری ویدئو، اطلاعات زمانی مهم‌تر هستند. برای مدل‌کردن روابط زمانی دنباله ویژگی مکانی، الگوریتم‌های مدل‌کردن دنباله‌های زمانی می‌توانند به کار روند. یکی از روش‌های معمول برای مدل‌کردن سری‌های زمانی، استفاده از مدل‌های آماری است. در [9] یک مدل مخفی مارکف چندسطحی معرفی شده است. در [10] شات‌ها با زنجیره مارکف مدل شده‌اند. در [11, 12] یک مدل درختی مارکف برای مدل کردن ارتباطات تبدیل دوبعدی موجک[1] به کار رفته است که نتایج بسیار خوبی در مدل‌سازی تصویر داشته است. در [13] از میدان تصادفی مارکف برای مدل‌کردن صحنه استفاده شده است. مدل مخفی مارکف به خوبی سیستم‌ها و ارتباطات پیچیده را مدل می‌نماید [14, 15, 16] ولی احتیاج به فرضیاتی درباره نوع توزیع‌ها دارد و برای به دست‌آوردن مدل مناسب نیاز به محاسبات دقیق است. در [17,18] از مدل AR برای مدل‌سازی ویدئو بر اساس ویژگی‌های مکانی هیستوگرام رنگ که ویژگی‌های مناسبی نیستند، استفاده شده است. این روش به خوبی ارتباطات زمانی را مدل کرده است و برای تشخیص فریم کلیدی و لبه شات‌ها به کار رفته است. در برخی کارها نیز به سیگنال ویدئو، به عنوان سیگنال سه‌بعدی نگریسته شده و مدل‌های آماری خوبی برای آن‌ها استخراج شده است. مدل GMM سیگنال سه‌بعدی ویدئو را در حوزه پیکسل در نظر گرفته و اشیاء را در ویدئو جدا کرده است [19-28]، با این کار امکان جداسازی شات، استخراج فریم‌های کلیدی، ویرایش ویدئو و تشخیص حرکت و اتفاق در ویدئو ایجاد شده است. در این کار برای جلوگیری از تاخیر زیاد و کاهش پیچیدگی از مدل دنباله‌ای GMM نیز استفاده شده است. مدل‌های آماری دیگری نیز برای تحلیل ویدئو در حوزه زمان تعریف و استخراج شده‌اند [29].

۱-۱- انگیزه‌های تحقیق

ویدئو واسطی است که امکان انتقال منابع غنی و پرظرفیت اطلاعات را دارد. این واسط در انتقال اطلاعات به مخاطبان مختلف و در بسیاری زمینه‌ها – مخابرات، پزشکی، تفریحی و تحصیلی- نقش مهمی دارد. با افزایش استفاده از سیگنال ویدئو، نیاز به وجود روش‌های کارا، هوشمند و راحت برای بهره‌برداری بهینه از این سیگنال با درنظرگرفتن خصوصیات آن [30-36] بیشتر می‌شود.

[1] Wavelet

یک سیستم پردازش ویدئو به طور کلی به سه قسمت اصلی تقسیم می‌گردد. ابتدا سیگنال ویدئو تحلیل می‌شود، نتایج این قسمت به فرمت از پیش تعیین شده‌ای که مناسب پردازش است، در می‌آیند و در نهایت این اطلاعات بسته به کاربرد، مورد استفاده قرار می‌گیرند. تحلیل و مدل‌سازی ویدئو بخش اول هر سیستم پردازش ویدئو است که در این بخش، با توجه به وجود افزونگی[1] زیاد در بعد زمان و مکان ویدئوی دیجیتال، تلاش می‌شود یکسری پارامتر از سیگنال ویدئو استخراج شود تا سیگنال تنکی به دست آید. در نظر گرفتن خصوصیات سیستم بینایی انسان ابزار مفیدی در این راه خواهد بود؛ چرا که در نهایت خروجی هر پردازش ویدئویی باید مطلوب انسان باشد. لذا تحلیل محتوای ویدئو در این پروژه بر مبنای روش‌های آماری[2] است و ارزیابی‌ها بر اساس معیارهای سیستم بینایی انسان[3] خواهند بود.

در سطح بالاتر، مدل‌سازی و تحلیل ویدئو می‌تواند باعث پیشرفت و بهبود روش‌های پردازش ویدئو شود. تحلیل ویدئو شامل استخراج پارامترهایی از سیگنال ویدئو است که خواص کلیدی این سیگنال را در بر دارند. این خواص کلیدی می‌توانند برای مثال، یک تغییر شات، یک فریم کلیدی، یک شیء در حال حرکت یا یک اتفاق خاص باشند. مدل‌سازی و تحلیل ویدئو ابزار مناسب برای پردازش سیگنال و استخراج اطلاعات لازم جهت رسیدن به خروجی مناسب را فراهم می‌آورد. مثلاً، در فشرده‌سازی ویدئو، مدل‌سازی ویدئو به تشخیص قسمت‌های مهم سیگنال مانند پیش‌زمینه و اشیاء متحرک کمک می‌کند تا بیت‌های بیشتر و در نتیجه کیفیت بهتری به این اجزاء تصویر اختصاص یابد. به این ترتیب تحلیل ویدئو می‌تواند باعث بهبود کارایی و کیفیت کدینگ ویدئو گشته، سیگنال فشرده بهتری را ایجاد کند. همچنین، بازیابی ویدئو هم می‌تواند از مدل‌سازی بهره گیرد. برای مثال سرعت بازیابی با استخراج پارامترهای مناسب از سیگنال و تشکیل بردار ویژگی مناسب در کنار معیار فاصله مقتضی افزایش می‌یابد. نهایتاً، می‌توان از کاربرد جداسازی منابع برای مدل‌سازی ویدئو نام برد به طوری که ویدئوی اصلی دارای مشخصات خاصی است و مجموعه مشخصات معلوم می‌توانند برای جداسازی این سیگنال از یک سیگنال ترکیبی مانند مجموع سیگنال اصلی و نویز استفاده شوند.

با توجه به مسائل بیان شده در این قسمت، در این رساله به مدل‌سازی و آنالیز سیگنال ویدئو بر مبنای روش‌های پارامتریک آماری[4]، با در نظر گرفتن معیار های سیستم بینایی انسان[5] می‌پردازیم، و از نتایج این تحلیل‌ها برای بهبود عملکرد سیستم‌های پردازش ویدئو بهره می‌گیریم.

[1] Redundancy
[2] Statistical Modleing
[3] Human Visual System
[4] Statistical Modelling
[5] Human Visual System

۱-۲- چالش‌ها

سیگنال ویدئوی دیجیتال که شامل نمونه‌های دیجیتال در بعد زمان و مکان است، دارای افزونگی زیادی است. با مدل‌سازی و استخراج پارامترهای ویدئو، سیگنال تنکی[1] به دست می‌آید، که در زمینه-های بسیاری قابل استفاده است. هدف اصلی این رساله استخراج مجموعه‌ای از پارامترها از سیگنال ویدئو است که می‌تواند در کاربردهای مختلف پردازش ویدئو استفاده شود. از کاربردهای پایه استفاده از پارامترها می‌توان جداسازی شات، انتخاب فریم‌های کلیدی شات‌ها، تحلیل نوع فعالیت در ویدئو، تحلیل اتفاق، تشخیص حرکات محلی و کلی و جداسازی اشیاء در ویدئو را نام برد. از طرف دیگر، موارد ذکرشده، در بهبود عملکرد سیستم‌های مختلف پردازش ویدئو نظیر فشرده‌سازی ویدئو، ویرایش ویدئو، انتقال ویدئو، بازیابی و شاخص‌گذاری ویدئو و خلاصه‌سازی ویدئو قابل استفاده هستند.

یکی از مهم‌ترین چالش‌ها در ارائه مدلی مناسب برای سیگنال ویدئو، ارائه مدلی است که به خصوصیات مختلف ابعاد زمانی و مکانی این سیگنال توجه داشته باشد. سیگنال ویدئو در واقع مجموعه‌ای از رخدادهای دیداری پیاپی است که می‌توانند جداگانه، همزمان و یا با مقداری همپوشانی زمانی اتفاق بیافتند. این ساختار همپوشان رخدادهای زمانی، تحلیل سیگنال ویدئو را پیچیده می‌کند، چون مرز مشخصی بین فعالیت‌های بصری جداگانه وجود ندارد. بنابراین ساختارهای پیشنهادی باید علاوه بر درک تحولات زمانی آنی در ویدئو، قدرت تحلیل تغییرات تدریجی ویدئو را هم داشته باشند. درک سرعت زمانی تغییرات محتوای ویدئو نیز از مسائل مهم در این زمینه به حساب می‌آید. علاوه بر این، پیچیدگی محاسباتی تحلیل‌های انجام‌شده بر روی سیگنال ویدئو نیز باید مورد توجه قرار گیرد. این امر در سیستم‌های زمان واقعی بیشتر اهمیت دارد.

از طرف دیگر سیستم بینایی انسان در تمامی زمینه‌های پردازش ویدئو و تصویر مورد توجه است. نکات مهم و اولویت‌دار از نظر چشم انسان، برای سیستم‌های پردازش نیز باید با اهمیت باشد. ارزیابی کیفی نتایج به دو روش آزمون ادراکی[2] و آزمون تحلیلی[3] تقسیم می‌شوند. در آزمون ادراکی از انسان برای بررسی نتایج استفاده می‌شود. از آنجا که چشم انسان پارامترهای بسیاری را در نظر می‌گیرد، این آزمون دقیق و در عین حال پیچیده، گران و وقت‌گیر است. گرچه این معیار نتایج معتبرتری را ارائه می-کند، ولی اجرای آن نیازمند در نظر گرفتن فاکتورهای بسیاری است که می‌تواند آنها را تکرارناپذیر نماید. در مقابل، آزمون تحلیلی تکرارپذیر است و اغلب به سادگی می‌تواند برای بهینه‌سازی در حلقه‌های فیدبک ارزیابی در تحلیل سیگنال مورد استفاده قرار گیرد.

[1] Sparse
[2] Subjective
[3] Objective

با توجه به موارد ذکرشده، در این رساله، به ارائه روش‌هایی برای تحلیل و مدل‌سازی پارامتریک ساختار زمانی- مکانی سیگنال ویدئو پرداخته شده است و در عین حال میزان تاخیر و پیچیدگی محاسباتی عملیات و سازگاری آنها با خصوصیات سیستم بینایی انسان مورد ملاحظه خاص قرارگرفته‌اند.

۱-۳- نوآوری‌ها

ما در این رساله به ارائه مدل‌های ترکیبی و روش‌های نوینی در زمینه تحلیل محتوای ویدئو می‌پردازیم. روش‌های ارائه‌شده را می‌توان به طور کلی به دو گروه تقسیم کرد:

- روش‌های تحلیل تحولات زمانی پارامترهای مکانی

در دسته اول، با سه ایده تجزیه زمانی، استفاده از یک معیار فاصله مناسب و مدل AR به بررسی تغییرات زمانی محتویات مکانی ویدئو پرداخته‌ایم که برای اولین بار مطرح می‌شوند. برای این تحلیل‌ها از پارامترهای مکانی مستخرج از خواص آماری حاشیه‌ای زیرباندهای تبدیل موجک برای نخستین بار استفاده شده است. به این ترتیب که ابتدا تبدیل موجک به فریم‌های ویدئویی اعمال می‌گردد. سپس هیستوگرام‌های حاشیه‌ای زیرباندهای این تبدیل با مدل تعمیم‌یافته گوسی تقریب زده می‌شوند و پارامترهای این مدل به عنوان مشخصات مکانی مستخرج از حوزه موجک در نظر گرفته می‌شوند. در ادامه این رساله از این مشخصات به عنوان پارامترهای مکانی نام برده می‌شود و سیر تحول زمانی این پارامترهای مکانی با کمک یکی از سه روش ارائه شده، بررسی می‌گردد. دلیل انتخاب این پارامترها این است که تبدیل دوبعدی موجک دارای ساختاری مشابه با توزیع سیستم بینایی انسان است و پارامترهای مستخرج از زیرباندهای این تبدیل، ما را به نتایج مطلوبی می‌رساند که آزمون‌های گسترده انجام گرفته، موید این موضوع هستند. این روش‌ها به خوبی سیر تحولات زمانی آنی و تدریجی محتوای مکانی فریم‌ها را بیان می‌کنند. در زیر به توضیح مختصری از هر کدام از این روش‌های تحلیل تحولات زمانی می‌پردازیم:

- ✓ **معیار فاصله مناسب:** روشی برای تعقیب تغییرات زمانی پارامترهای مکانی است. در این روش روند تغییرات بین فریمی با کمک معیار فاصله مناسب KL[1] بین پارامترهای مکانی حاشیه‌ای زیرباندهای تبدیل دوبعدی موجک مورد بررسی قرار می‌گیرد و تغییرات به خوبی شناسایی می‌شوند. مرزهای شات‌ها و خوشه‌های شات‌ها بر اساس ضابطه شباهت و تفاوت انتخاب می‌گردند. نتیجه ارزیابی‌های ادراکی و تحلیلی بیان‌گر دقت بالای این روش در مقایسه با روش‌های متداول است.

[1] Kullback-Leibler Distance

- ✓ **تجزیه زمانی:** روشی برای تحلیل زمانی-مکانی رخدادهای بصری سیگنال ویدئو است. این روش سیگنال ویدئو را به عنوان مجموعه‌ای از مولفه‌های بصری مستقل همپوشان در نظر می-گیرد. رخدادها همان توابع فشرده معمول دارای همپوشانی هستند که سیر تحول زمانی مجموعه‌ای از پارامترهای مکانی سیگنال ویدئو را توصیف می‌کنند. ما از تجزیه زمانی برای حل ساختار همپوشان رخدادها، که از مهمترین مسائل موجود در تحلیل ویدئو است، استفاده می-کنیم. در این روش، مجموعه ای از پارامترهای مکانی، از سیگنال ویدئو استخراج می‌شوند و به صورت ترکیب خطی از مجموعه‌ای از توابع فشرده همپوشان زمانی به نام رخداد، طی یک مرحله بهینه‌سازی، بیان می‌گردند. این روش سریع و دقیق برای تحلیل و مدل‌سازی سیگنال ویدئو استفاده شده است و قابلیت و کارایی بالای این رویکرد در کاربرد چکیده‌سازی[1] ویدئو ارائه شده است.
- ✓ **مدل AR:** مدل پارامتریک AR که دارای قابلیت بیان سیستم‌های زمانی خطی پیچیده و نویزی است، برای بیان ساختار زمانی پارامترهای آماری حاشیه‌ای تبدیل موجک دوبعدی فریم‌های ویدئو استفاده شده است. از پارامترهای تخمین مدل مربوطه به خوبی می‌توان در تشخیص رخداد در ویدئو و تعیین مرز شات‌ها و فریم‌های کلیدی ویدئو بهره جست.

- تحلیل پارامترهای مکانی-زمانی ویدئو

روش آخر ارائه‌شده، استفاده از پارامترهای آماری حاشیه‌ای و توأم تبدیل سه‌بعدی موجک است. این اولین باری است که خواص آماری حاشیه‌ای و توأم موجک سه‌بعدی مورد بررسی قرار می‌گیرد و نتایج این تحلیل به عنوان پارامترهای مکانی-زمانی ویدئو استفاده می‌شوند. در این قسمت، مهمترین نوآوری ما ارائه روش نوینی برای مدل‌سازی و تحلیل ویدئوهای طبیعی بر اساس خواص آماری تبدیل سه‌بعدی موجک است. ضمن مقایسه این روش با سایر روش‌های موجود، نشان داده می‌شود که پارامترهای استخراج‌شده از این خصوصیات آماری، نمایندگان مناسبی برای تفسیر محتوای ویدئو بر اساس درک انسانی هستند. برای این کار روش خود را در دو کاربرد مورد ارزیابی قرار داده‌ایم: گروه‌بندی میزان فعالیت ویدئو و تشخیص رفتار انسان.

۱-٤- ساختار کلی رساله

ساختار کلی این رساله به ۷ فصل تقسیم شده است. در فصل دوم، مروری بر تحقیقات مرتبط در مدل‌سازی و تحلیل پارامتریک ویدئو خواهیم داشت. پس از آن در فصول متوالی راهکارهای مختلف پیشنهادی ارائه گردیده‌اند. در فصل سوم سیر تحولات زمانی پارامترهای مکانی ویدئو با کمک معیار

[1] Video abstraction

فاصله مناسب مورد ارزیابی قرار می‌گیرد. فصل چهارم به ارائه روش تجزیه زمانی پارامترهای مکانی می-پردازد و در فصل پنجم تغییرات زمانی پارامترهای مکانی با کمک مدل‌سازی AR مدل می‌شوند. در فصل ششم تحلیل ویدئو را برمبنای پارامترهای مکانی-زمانی مستخرج از تبدیل سه‌بعدی موجک شرح می‌دهیم و در نهایت آخرین فصل به نتیجه‌گیری و مروری بر کارهای آینده خواهد پرداخت.

۱-۵- مقالات مستخرج از پایان‌نامه

Journal Papers:

1. **M. Omidyeganeh**, S. Ghaemmaghami, and S. Shirmohammadi, "Video Keyframe Extraction based on Marginal Statistics of 2D-Wavelet Transform," IEEE Transaction on Image processing, vol. 20, issue 10, pp. 2730:2737, 2011.
2. **M. Omidyeganeh**, S. Ghaemmaghami, and S. Shirmohammadi, "Group based Visually Sensitive Spatio-Temporal Video Analysis and Abstraction," Signal, Image and Video Processing, Springer, Accepted, to appear, 2012.
3. **M. Omidyeganeh**, S. Ghaemmaghami, and S. Shirmohammadi, "3D-Wavelet Statistics: A New Approach for Video Content Analysis," Multimedia Tools and Applications, Springer, Accepted, to appear, 2012.
4. **M. Omidyeganeh**, S. Ghaemmaghami, and S. Shirmohammadi, "Event based Spatio-Temporal Video Analysis for Keyframe Selection and Video Retrieval," IET Image Processing, Under revision.
5. **M. Omidyeganeh**, S. Ghaemmaghami, A. Javadtalab, and S. Shirmohammadi "A Robust Wavelet based Approach to Fingerprint Identification," Submitted to IET Electronics Letters.

Conference Papers:

1. **M. Omidyeganeh**, A. Javadtalab, S. Ghaemmaghami, S. Shrmohammadi, "Contourlet Based Generalized Gaussian Modeling for Intelligent Fingerprint Identification," Submitted to IEEE I2MTC 2012.
2. **M. Omidyeganeh**, A. Javadtalab, S. Ghaemmaghami, S. Shirmohammadi, "Statistical Modeling of Error Resilient JPEG2000 Decoding", Proc. IEEE Region 10 Annual Conference (TENCON), Bali, Indonesia, November 21-24, 2011.
3. **M. Omidyeganeh**, A. Javadtalab, S. Shirmohammadi, "Intelligent Driver Drowsiness Detection through Fusion of Yawning and Eye Closure", Proc. IEEE Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems, Ottawa, ON, Canada, September 19-21, 2011.
4. **M. Omidyeganeh**, H. Khalilian, S. Ghaemmaghami, S. Shirmohammadi, "Robust Digital Video Watermarking in an Orthogonal Parametric Space", Proc. IEEE Region 10 Annual Conference (TENCON), Fukuoka, Japan, November 21-24, 2010.
5. **M. Omidyeganeh**, S. Ghaemmaghami, S. Shirmohammadi, "Autoregressive Video Modeling through 2D Wavelet Statistics", Proc. Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing, Darmstadt, Germany, October 15-17, 2010.
6. **M. Omidyeganeh**, S. Ghaemmaghami, S. Shirmohammadi, "An Event Based Approach to Video Analysis and Keyframe Selection", Proc. IEEE Workshop on Signal Processing Systems, San Francisco, CA, USA, October 6-8, 2010.
7. **M. Omidyeganeh**, S. Ghaemmaghami, and H. Khalilian, Video activity Analysis Based on 3D Wavelet Statistical Properties, IEEE International Conference on Advanced Communications Technology, ICACT'09, Korea, 2009.
8. H. Khalilian, S. Ghaemmaghami, and **M. Omidyeganeh**, Digital Video Watermarking in 3-D Ridgelet Domain, IEEE International Conference on Advanced Communications Technology, ICACT'09, Korea, Feb. 2009.

فصل دوم

مروری بر تحقیقات مرتبط در زمینه تحلیل و مدل‌سازی پارامتریک ویدئو

مقدمه

مدل ترکیبی گوسی

مدل پارامتریک AR برای بررسی محتوای سیگنال

مدل مخفی مارکف

مروری بر برخی مدل‌های دیگر

کاربردهای تحلیل و مدل‌سازی ویدئو

دادگان‌های ویدئو

جمع‌بندی

۲-۱- مقدمه

در این بخش به برخی روش‌های تحلیل محتوای سیگنال ویدئو اشاره می‌کنیم. هدف کلی این پروژه مدل‌سازی پارامتریک محتوای سیگنال ویدئو است که در آن با استخراج پارامترهایی بر اساس خواص سیستم بینایی انسان و خواص آماری سیگنال ویدئو، سیگنال تنکی به دست آوریم و در کاربردهای مختلف پردازش سیگنال ویدئو از این پارامترهای استخراج‌شده، استفاده نماییم. عموماً فرآیندهای طبیعی[1] موجود، خروجی‌های قابل مشاهده‌ای را تولید می‌کنند که به عنوان سیگنال شناخته می‌شوند. سیگنال‌ها به چند دسته بزرگ تقسیم می‌شوند. دسته اول سیگنال‌هایی هستند که ذاتاً گسسته‌اند و دسته دیگر سیگنال‌ها، طبیعتاً پیوسته‌اند. منابع سیگنال می‌توانند ایستان (مشخصه آماری سیگنال با زمان تغییر نکند) یا غیر ایستان باشند. سیگنال می‌تواند خالص[2] باشد یا توسط سیگنال دیگری مانند نویز تخریب شده و یا در اثر اعوجاج یا تداخل تغییر کرده باشد. سیگنال ویدئوی مورد بحث در این پروژه سیگنال گسسته غیرایستان خالص است.

فواید مدل‌سازی منابع و سیگنال‌ها را می توان به دو دسته مهم تقسیم نمود:

- اولین فایده مدل‌سازی، فراهم آوردن روش مناسب جهت تجزیه و تحلیل سیگنال و پردازش آن به منظور رسیدن به سیگنال خروجی معین است. مثلاً اگر به دنبال بهبود کیفیت یک سیگنال هستیم، ناگزیریم که سیگنال اصلی را از میان سیگنال موجود بازشناسیم که لازمه این کار، داشتن مدلی از

[1] Natural Processes
[2] Pure Signal

سیگنال و پارامترهای آن است. یا اگر هدف فشرده‌سازی دیجیتال سیگنال باشد، فرآیند اختصاص بیت باید با توجه به میزان اهمیت اجزاء و مولفه‌های سیگنال از نظر سیستم بینایی انجام گیرد.

- دومین فایده مدل‌سازی سیگنال، شناسایی منبع تولید سیگنال است؛ بدون این‌که به منبع واقعی دسترسی مستقیم داشته باشیم. این کار زمانی که دسترسی مستقیم به منبع تولید فرآیند تصادفی امکان-پذیر نبوده یا هزینه‌بر است، مقرون به صرفه است. در این حالت با داشتن یک مدل مناسب از منبع می‌توانیم سیگنال منبع را تولید و شبیه‌سازی کنیم.

گاهی منظور از مدل‌سازی، مدل‌سازی سیگنال‌های قطعی[3] است که هدف آن تشخیص پارامترهای مجهول و در عین حال قطعی نظیر فرکانس، دامنه و فاز است؛ که بحث ما شامل این مدل‌سازی نمی‌شود. دو کلاس مهم برای مدل‌سازی سیگنال‌ها و فرآیندهای تصادفی مدل‌سازی پارامتریک و غیرپارامتریک هستند، که هدف از آنها تشخیص پارامترهای آماری سیگنال، تبدیل‌یافته سیگنال و یا ویژگی‌های استخراج شده از سیگنال است:

- در مدل‌سازی‌های غیرپارامتریک نمونه‌های مورد نظر به تنهایی برای نمایش توزیع به کار می-روند و در واقع پایه اصلی مدل‌سازی نقاط به صورت خام هستند. این مدل‌ها می‌توانند با در نظر گرفتن مجموع اطلاعات زمانی و مکانی مقاوم‌تر شوند. مدل‌های هیستوگرام[4]، مدل‌های کرنل و مدل‌های رگرسیون از این دسته هستند. گرچه مدل‌های غیرپارامتریک در حالات دینامیکی شدید محیط خوب عمل می‌کنند ولی پیچیدگی‌های محاسباتی زیاد آنها مانعی برای پیاده‌سازی، خصوصاً در استفاده‌های زمان–واقعی می‌شود. از طرفی مدل‌های غیرپارامتریک نمی‌توانند برای چندین مقیاس زمانی تعمیم داده شوند گرچه این مدل‌ها دقت بیشتری نسبت به روش‌های پارامتریک دارند.
- در مدل‌سازی‌های پارامتریک شدت یا رنگ پیکسل یا مولفه‌های تبدیل‌یافته سیگنال ویدئو با کمک یک توزیع احتمال با پارامتر مشخصی مدل می‌شود. ترکیب چند توزیع گوسی به همراه پارامترهای میانگین و واریانس یک مثال از این مدل‌سازی است. از جمله مدل‌های آماری موجود، فرآیندهای گوسی، مدل تعمیم‌یافته گوسی، مارکف و مدل مخفی مارکف و مدل‌های AR هستند. در این روش‌ها به فرآیند مذکور، یکی از مدل‌ها نسبت داده می‌شود و سپس توسط روش‌های تحلیلی و آماری پارامترهای مدل استخراج می‌شوند. این روش‌ها با تعداد کمی پارامتر، سیگنال و توزیع آن را توصیف می‌کنند و نسبت به مدل‌های غیرپارامتریک تعداد کمی پارامتر مجهول دارند، که باید شناسایی شوند. در نتیجه تخمین‌های به دست آمده کاربردی‌تر خواهند بود و برای کاربردهایی نظیر فشرده‌سازی

[3] Deterministic Signals
[4] Histogram Models

بیشتر به کار می‌روند. روش‌های پارامتریک عموماً مطمئن‌تر و مقاوم‌تر در برابر نویز هستند [37] . عیب این روش‌ها این است که نسبت به روش‌های غیرپارامتریک دارای دقت پایین‌تری هستند.
در ادامه این بخش به معرفی چند مدل پارامتریک برای سیگنال ویدئو می‌پردازیم.

۲-۲- مدل ترکیبی گوسی[5]

مدل ترکیبی گوسی ابتدا برای مدل‌کردن تصویر استفاده شده است [24,25,38]. این مدل به همراه فاصله KL به عنوان معیاری برای مشابهت دو تصویر به کار برده شده است [26,27]. کاربردهای دیگری نظیر بازیابی تصاویر [24] و خوشه‌بندی تصاویر [28] نیز برای این مدل به کار رفته است. در [21,39] یک روش آماری برای بازنمایی و مدل‌کردن ویدئو ارائه شده است. در روش‌های بازنمایی ویدئو، ویدئو به اشیاء بامعنی تقسیم می‌شود و برای بازیابی و شاخص‌دهی از این نتایج استفاده می‌گردد. در این روش، خوشه‌بندی بدون نظارت[6] با کمک مدل GMM انجام شده و نواحی همدوس زمانی-مکانی در فضای ویژگی‌ها و قطعات همدوس متناظر در ویدئو از هم جدا می‌شوند. نکته مهم این کار، تحلیل ویدئوی ورودی به شکل یک هویت مستقل - به جای در نظر گرفتن آن به شکل مجموعه‌ای از فریم‌ها- است. در این روش، زمان و مکان به شکل یک واحد دیده می‌شوند و الگوریتمی نیز به نام GMM تکه‌ای مطرح شده است، که در آن به جای استخراج یک GMM از کل ویدئو، دنباله‌ای از GMMها از دنباله ویدئویی استخراج می‌شود، که این کار باعث می‌گردد، امکان تعریف الگوهای غیر-محدب و غیرخطی هم وجود داشته باشد. نواحی زمانی-مکانی استخراج‌شده از این روش امکان تشخیص و تعریف "رخداد"[7] را هم پدید می‌آورند. روش استفاده‌شده در واقع، یک روش پارامتریک است و درآن هیچ محدودیتی برای مدل‌کردن هندسی یا سختی اشیاء وجود ندارد. فضای ویژگی یک فضای ٦ بعدی درنظرگرفته می‌شود. که ابعاد آن سه بعد فضای رنگ L,a,b ، دوبعد مکانی x,y و یک بعد زمانی t هستند. فرض می‌شود که توزیع شش بعدی اطلاعات، ترکیبی از احتمالات گوسی است. به این ترتیب که هر ناحیه همگن در صفحه تصویر، با یک توزیع گوسی بیان می‌شود و تعدادی ناحیه همگن هم با ترکیبی از این توزیع‌ها قابل بیانند. بنابراین از الگوریتم EM[8] برای تعیین پارامترهای مدل استفاده شده است. در نتیجه توزیع یک متغیر تصادفی $X \in R^d$ ترکیبی از k توزیع گوسی با توزیع زیر است:

$$f(x|\theta) = \sum_{j=1}^{k} \alpha_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\} \qquad (۱-۲)$$

[5] Gaussian Mixture Model
[6] Unsupervised Clustering
[7] Event
[8] Expectation-Maximization

به طوری که مجموعه پارامترهای $\theta=\{\alpha_j,\mu_j,\Sigma_j\}_{j=1}^{k}$ دارای خصوصیات زیر هستند:

$$\alpha_j>0,\sum_{j=1}^{k}\alpha_j=1 \qquad (۲-۲)$$

$\mu_j\in R^d$ و Σ_j یک ماتریس dxd معین مثبت است.

k=3 k=5

شکل ۲-۱ شکل بالا: تصویر اولیه، تصاویر ردیف پایین: اشیاء استخراج شده با مدل GMM. استفاده از مقادیر بزرگتر برای k باعث استخراج اشیاء بیشتری از تصویر و افزایش دقت می‌شود [65].

با داشتن مجموعه بردارهای ویژگی $x_1,...,x_n$ پارامترها به صورت ML[۹] به شکل زیر تخمین زده می‌شوند:

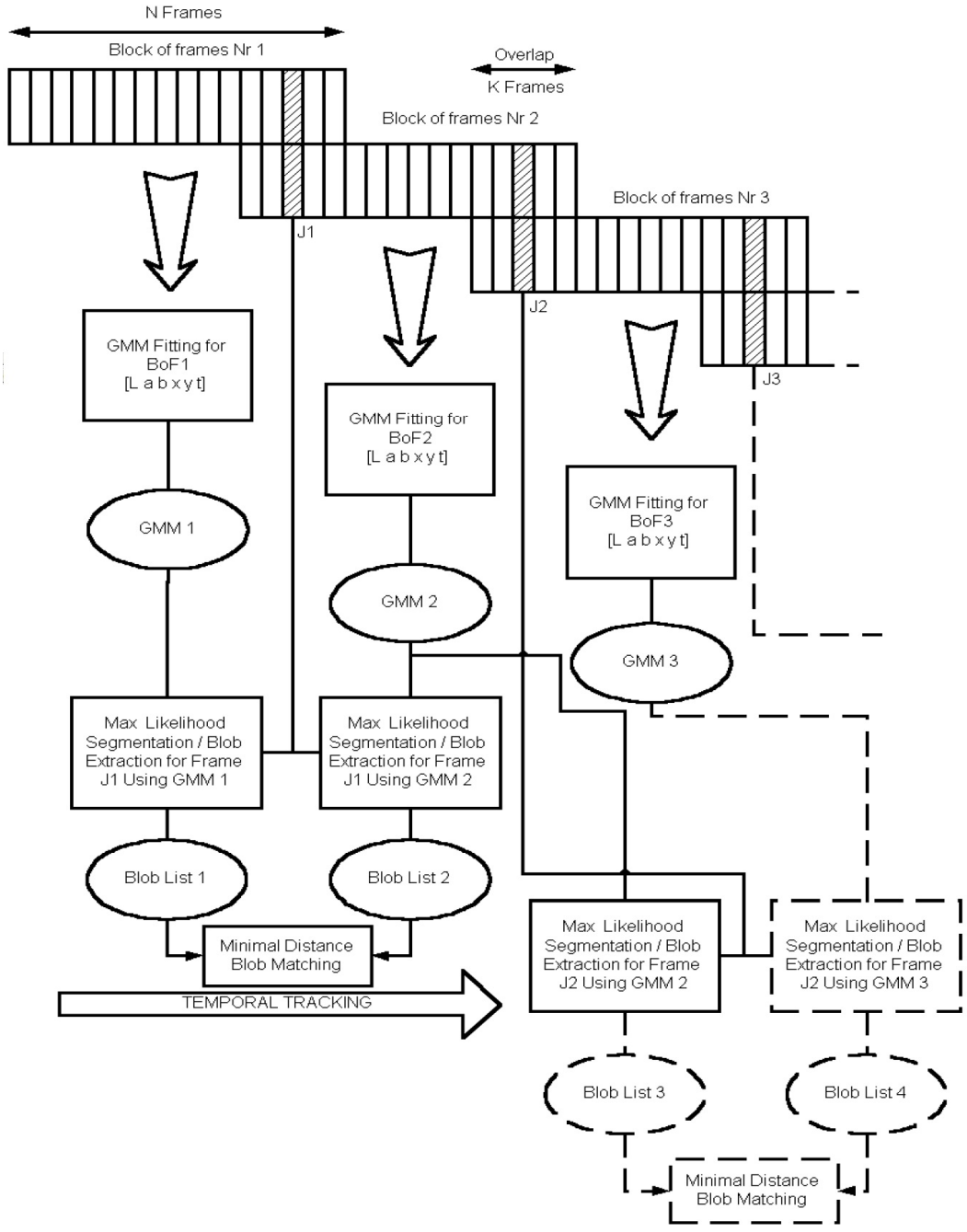
$$\theta_{ML}=\arg\max_{\theta}L(\theta|x_1,...,x_n)=\arg\max_{\theta}\sum_{i=1}^{n}\log f(x_i|\theta) \qquad (۳-۲)$$

که با روش EM این محاسبات انجام می‌شود [23] و با الگوریتم k-mean الگوریتم مقدار اولیه می‌گیرد [40]. تعیین تعداد توزیع‌های گوسی، k، نیز مساله مهمی است. همان‌طور که در شکل ۲-۱ دیده می‌شود، هرچه k بیشتر باشد، اجزای بیشتری از تصویر جدا می‌شوند. ولی نباید مقدار آن هم از حدی بیشتر شود. در [19, 20, 22] از روش MDL[۱۰] برای تخمین k در تصاویر ساکن استفاده شده است. در [21] از MDL برای ویدئو هم استفاده شده است. به این ترتیب که k طوری انتخاب می‌شود که رابطه زیر ماکزیمم شود:

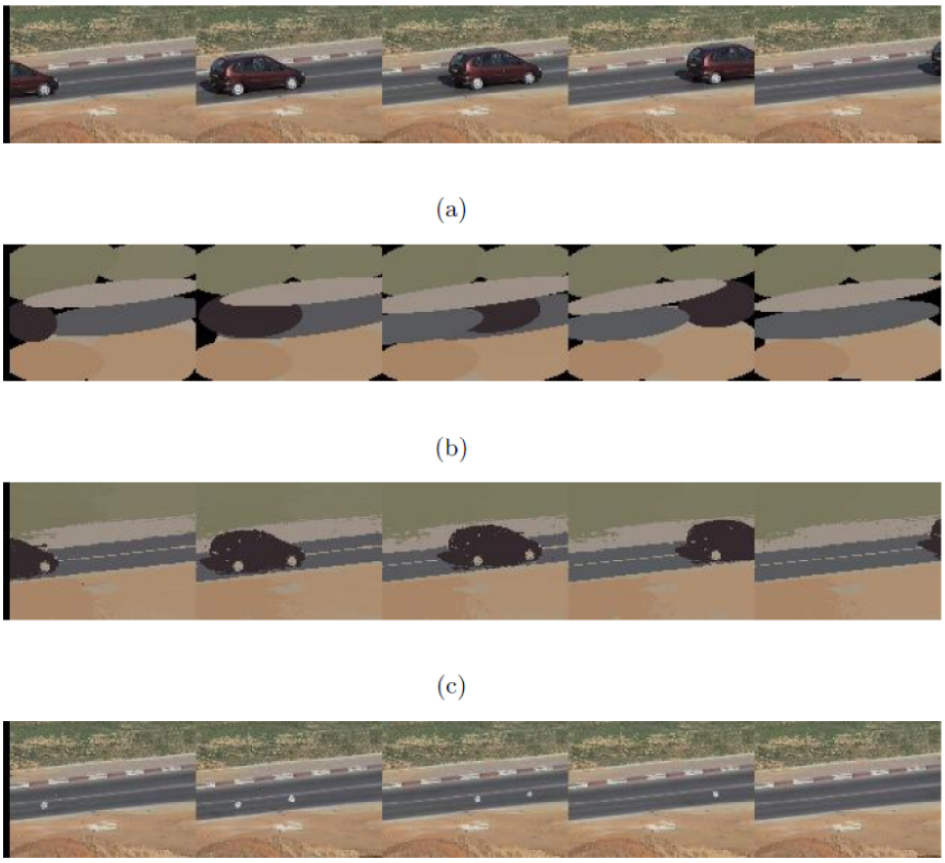
[9] Maximum Likelihood
[10] Minimum Description Length

$$\log L(\theta_{ML}|x_1,...,x_n) - \frac{l_k}{2}\log n \qquad (٤-٢)$$

که $l_k = (k-1) + kd + k(\frac{d(d+1)}{2})$ است.



شکل ۲-۲ بلوک دیاگرام GMM تکه‌ای [21].



شکل۲-۳ a) چند فریم از یک دنباله ویدئویی b) نتایج جداسازی با GMM تکه‌ای c) جداسازی شیء متحرک d) حذف شیء متحرک [21].

از آنجا که برای استفاده از GMM کلی نیاز به داشتن همه ویدئو داریم و این مدل نمی‌تواند اشیاء غیرخطی و غیرمحدب را مدل کند، روش GMM تکه‌ای پیشنهاد شده است. که نمودار آن در شکل ۲-۲ آمده است. در این مدل توجه شده است که در هر بلوک از فریم‌ها همپوشانی بین فریم‌ها در آغاز و پایان گروه رعایت شود تا تسلسل کار حفظ گردد. از کاربردهای این روش، تشخیص حرکت را می‌توان نام برد. به این شکل که از ماتریس کواریانس مدل‌ها، برای استخراج واقعه استفاده کرد. درایه‌های $C_{t,x}, C_{t,y}, C_{t,t}$ در اینجا به ترتیب از راست به چپ نمایانگر پراکندگی حباب در حوزه زمان، وابستگی بین موقعیت افقی و زمان و وابستگی بین موقعیت عمودی و زمان هستند. مقادیر کم این متغیرها نشان-دهنده استاتیک بودن آنهاست. با نرمالیزه‌کردن این مقادیر به صورت زیر برای تشخیص حرکت اشیاء از آنها استفاده می‌شود.

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}, \qquad -1 \le R_{ij} \le 1 \qquad (۵-۲)$$

شکل ۲-۳ نمونه‌ای از این نتایج را نشان می‌دهد. همانطور که در این شکل دیده می‌شود، چرخ‌های اتومبیل جدا نشده‌اند که با بالارفتن دقت این مشکل نیز حل خواهد شد. در [41] هم از این روش برای جداسازی صحنه استفاده شده است. در این‌جا برای کم‌کردن افزونگی و پیچیدگی از فریم‌های برجسته در مدل GMM استفاده می‌شود و نتایج بسیار خوبی نیز به‌دست آمده است.

۲-۲-۱- بحث در مورد روش GMM

در مدل GMM تکه‌ای علاوه بر حفظ ارتباطات زمانی (فریم‌های مشترک بین BOFها)، مشکل تاخیر حل می‌شود. گرچه پیچیدگی سیستم هم تا حدی کاهش می‌یابد ولی پیچیدگی سیستم هنوز خیلی بالا است. نکته مثبت دیگر روش تکه‌ای قابلیت تعریف حرکات غیرمحدب و غیرخطی است. پس از استخراج پارامترهای مدل، می‌توان از آن برای تشخیص نوع حرکت و جهت آن، توصیف واقعه، تغییر شات، استخراج پس‌زمینه و ویرایش ویدئو استفاده کرد. امکان تشخیص حرکات محلی و سراسری از روی پارامترهای استخراجی نیز مزیت این روش است.

معضل اصلی این روش این است که به خصوصیات زمانی و مکانی با یک دید نگریسته شده است و با هر دو بعد مکانی و بعد زمانی رفتار یکسانی داشته است. در صورتی که ابعاد زمانی و مکانی دارای ویژگی و ساختار متفاوتی هستند. مشکل بعدی، پیچیدگی و مقدار محاسبات بسیار بالای این سیستم است. این مدل در فضای ٦ بعدی کار می‌کند و دو مساله بهینه‌سازی را در هر BOF بررسی و حل می‌نماید.

۲-۳- مدل پارامتریک AR برای بررسی محتوای سیگنال ویدئو

در [17, 18] یک مدل پارامتریک برای تحلیل محتوای سیگنال ویدئو ارائه شده است. در این تحقیق، از مدل‌سازی AR[11] برای مدل‌کردن دنباله ویژگی مکانی فریم در طول زمان و تحلیل آن در فضای پارامتریک AR، استفاده شده است. کاربردهای این مدل، شامل تشخیص مرز شات‌ها در دنباله ویدئویی، استخراج فریم‌های کلیدی و ویژگی‌های ترکیبی زمانی- مکانی برای دسته‌بندی شات‌ها است. نتایج آزمون‌ها کارایی بهتر این روش را در مقابل روش سنتی هیستوگرام رنگ نشان می‌دهد. تحلیل محتوای ویدئو، زمینه تحقیق فعالی در سال‌های اخیر به حساب می‌آید. قدم اولیه برای تحلیل محتوای ویدئو تقسیم دنباله ویدئویی به قسمت‌هایی به نام "شات" است، که شامل دنباله‌ای از فریم‌هاست که با یک دوربین بدون وقفه گرفته شده‌اند. تحقیقات زیادی برای تشخیص مرز شات‌ها در سال‌های گذشته انجام شده است. بسیاری از ویژگی‌هایی که در این روش‌ها استفاده می‌شود، توانایی اندازه‌گیری ارتباط زمانی در بازه طولانی را ندارند. نویز سفید و حرکات رایج اشیاء در دیتای ویدئو معمولاً باعث خطا در

[11] Auto-Regressive

تشخیص می‌شوند. در این کار از تخمین پارامترهای سیستم AR به صورت همزمان استفاده می‌شود و مرز شات جایی است که ساختار سیستم تغییر می‌کند؛ این کار با تمرکز بر روی خطای تخمین AR (APE[۱۲]) در روش بازگشتی تخمین پارامترها، انجام می‌شود. پس از تقسیم ویدئو به شات‌ها، فریم‌های اصلی شات به منظور استفاده برای کاربردهایی نظیر بازیابی ویدئو و تشخیص محتوای شات، استخراج می‌شوند. فریم کلیدی می‌تواند با تحلیل ویژگی‌های موجود در شات نظیر حرکت [42]، یا روش‌های خوشه‌بندی [43]، انجام شود. این روش‌ها بر اساس شباهت دو فریم مجاور کار می‌کنند، که با کمک روش پارامتریک فریمی انتخاب می‌شود، که بهتر از بقیه همسایگانش (قبل و بعد) بتواند هم در حوزه ویژگی‌های مکانی و هم در جهت زمان، همسایگانش را توصیف کند.

فضای رنگ HSV با دقت ٤:٤:٨ در نظر گرفته می‌شود، که ۱۲۸ رنگ کوانتیزه‌شده را در پی دارد. تعداد پیکسل‌های دارای هر رنگ، مقدار آن در هیستوگرام هستند. اگر هیستوگرام M مولفه داشته باشد، ویژگی‌های هیستوگرام برای فریم n ام به صورت زیر در نظر گرفته می‌شوند:

$$\{H_n(1),H_n(2),...,H_n(M)\}$$

به این صورت مطابق فرمول مدل AR برای مولفه i ام هیستوگرام به صورت زیر خواهد بود:

$$H_{l+p}(i)=\sum_{j=1}^{p}a_j H_{l+p-j}(i)+\eta(i) \qquad (٦-٢)$$

که $\eta(i)$ نویز جمع‌شونده است. در پروسه آموزش مدل، خطاهای تخمین (APE) الگوریتم RLS، تخمین خوبی برای اندازه‌گیری میزان تطابق اطلاعات حاضر با مدل است و می‌تواند تغییرات ویدئو را تشخیص دهد. مقدار APE فریم n می‌تواند به صورت زیر بیان شود:

$$APE(n)=\sum_{i=1}^{M}|e_i(n)|w_i \qquad (٧-٢)$$

که $e_i(n)$ خطای تخمین AR مولفه i هیستوگرام است و در الگوریتم محاسبه شده است؛ و w_i وزن مولفه iام است. وزن‌ها باید به گونه‌ای نرمالیزه شوند که $\sum_{i=1}^{M}w_i=1$. w_i معمولا برابر اندازه مولفه کنونی در هیستوگرام نرمالیزه‌شده در نظر گرفته می‌شود.

۲-۳-۱- کاربردهای مدل AR درتحلیل ویدئو

تعیین مرز شات ویدئو

برای تشخیص مرز شات، APE به عنوان ملاک سنجش در نظر گرفته می‌شود. APE می‌تواند خطاهای جمع‌شده را محاسبه نماید. وقتی این مقدار زیاد می‌شود، بدین معناست که پارامترهای مدل

[12] AR Prediction Error

نمی‌توانند فریم حاضر را به خوبی مدل نمایند و یک تغییر شات رخ داده است. نحوه تغییر شات را می‌توان به دو دسته آنی و تدریجی تقسیم کرد. در حالت تغییر آنی، تغییر در یک فریم اتفاق می‌افتد و در حالت تغییر تدریجی، در دنباله‌ای از فریم‌ها رخ می‌دهد.

تشخیص مرز شات در حالت تغییر آنی

برای تشخیص مرزهای احتمالی شات، یک سطح آستانه باید در نظر گرفته شود:

$$T_b = km \qquad (۸-۲)$$

که m میانگین APE در روی پنجره زمانی یک‌بعدی به طول p و k بر اساس نتایج تجربی بین ۲ تا ۳ قرار می‌گیرد.

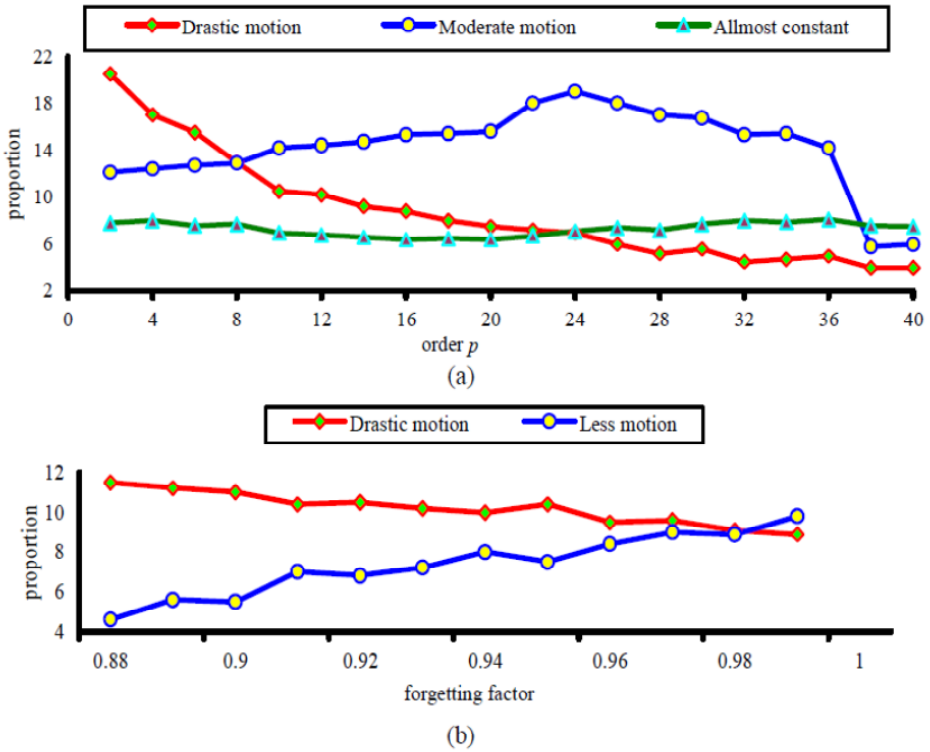
تشخیص مرز شات در حالت تغییر تدریجی

در این حالت دنباله تدریجاً تغییر می‌کند و از روش مقایسه دوقلو[۱۳] به همراه APE برای تشخیص آن استفاده می‌شود.

انتخاب مرتبه مدل AR و فاکتور فراموشی

نتایج آزمون‌ها نشان می‌دهند، که انتخاب مرتبه مدل بین ٤ تا ٢٤ و فاکتور فراموشی بین ۰,۹۲ تا ۰,۹۸ برای اغلب ویدئوهای تحت آزمون جواب بهتری می‌دهد. مرتبه مدل تعیین می‌کند که دنباله فریم‌ها تا چه حد به هم مرتبط هستند. در فیلم‌های دارای حرکات سریع مدل در مرتبه‌های بالاتر بسیار بد عمل می‌کند؛ چون فریم کنونی با فریم‌هایی‌که فاصله زمانی زیادی با آن دارند، دیگر همبستگی زیادی ندارد. شکل ۲-a٤ نسبت APEr[۱٤] که برابر نسبت APE در لبه شات به میانگین APE در طول شات است، را نشان می‌دهد. هنگامی‌که حرکت کلی فریم و حرکت شیء هر دو شدید هستند، فقط فریم‌های خیلی نزدیک می‌توانند فریم حاضر را مدل کنند و APEr با زیاد شدن p افزایش می‌یابد. در حالت تقریباً ثابت، سرعت بازگشت بسیار سریع‌تر است، به طوری‌که APEr با تغییر p، تغییر چندانی نمی‌کند. در شکل ۲-b٤ APEr بر حسب فاکتور فراموشی نشان داده شده است. تاثیر اطلاعات تاریخی با کاهش فاکتور فراموشی کم می‌شود. هنگامی‌که محتوای شات تغییر زیادی می‌کند، APEr با کم شدن λ به تدریج افزایش می‌یابد.

[13] Twin-Comparison Method
[14] APE Ratio



شکل ۲-٤ نمودار نسبت برای تشخیص تغییر شات با تغییر (a) مرتبه و (b) فاکتور فراموشی[18].

استخراج فریم کلیدی

با کمک مدل پارامتریک معرفی‌شده علاوه بر تشخیص لبه شات، می‌توان فریم‌های کلیدی را در دو حوزه زمان و مکان استخراج کرد. در دنباله‌های ویدئویی، محتوای برخی فریم‌ها در فریم‌های بعدی و قبلی ظاهر می‌شوند و فریم کلیدی با تخمین دوطرفه بهتر پیدا می‌شود. پس تخمین‌های رو به جلو[۱۵] و رو به عقب[۱٦] برای تخمین پارامترهای AR به کار می‌روند. در این حالت استخراج فریم کلیدی برابر یافتن فریمی با کمترین APE در شات است، به این معنا که فریم‌های قبلی و بعدی به خوبی فریم کنونی را مدل می‌کنند. به بیان دیگر این فریم نمایش مناسبی از تمام دنباله فریم‌هاست.

یک مثال از استخراج فریم کلیدی در شکل ۲-۵ آورده شده است. در خوشه ۱ بازیکن A در حال گرفتن توپ است، در حالی‌که در خوشه ۲ بازیکن B وارد می‌شود و باعث همپوشانی می‌گردد. در خوشه ۳ هم دوربین روی بازیکن A زوم می‌کند تا از سرویس او فیلم بگیرد. هر خوشه با فریم کلیدی آن نمایش داده می‌شود و فریم‌های کلیدی هم نماینده‌های خوبی برای بیان خوشه‌ها هستند.

[15] Forward prediction
[16] Backward prediction

شکل ۲-۵ فریم‌های کلیدی استخراج‌شده از یک شات [18].

نتایج آزمون‌ها

جدول۲-۱ تعداد فریم‌ها، شات‌ها، برش‌ها، محوشدگی‌ها و حل‌شدگی‌ها در ویدئوهای مورد آزمایش [18].

ویدئو	فریم	شات	برش	محوشدگی	حل‌شدگی
V1	۱۰۰۰۰	۱۸۸	۱۸۸	۰	۰
V2	۷۰۰۰	۱۳۰	۱۳۰	۰	۰
V3	۱۰۵۰۰	٥٤	۲۷	۱٥	۱۲
V4	۸۰۰۰	۱۹۹	۱۹۱	۲	٦
V5	۷٥۰۰	٤۸	٤۲	۰	٦

نتایج اعمال روش‌های ارائه‌شده بر روی ویدئوهای متنوعی بررسی شده‌اند. مشخصات این ویدئوها در جدول ۲-۱ نشان داده شده‌است.

در بررسی‌ها، M1 روش ارائه‌شده بر پایه APE است. M2 روش ارائه‌شده بر اساس تئوری اطلاعات [44] است و M3 روش استفاده از سطح آستانه وفقی برای مرزهای ناگهانی شات و روش مقایسه دوقلو برای تغییر تدریجی مرز شات با کمک شباهت هیستوگرام درون فریمی [45] است. دو معیار برای اندازه‌گیری موفقیت روش‌ها استفاده شده است: بازخوانی R و دقت P که هر دو برای بررسی کارایی روش‌های تشخیص مرز شات استفاده می‌شوند [46] .

$$recall = \frac{\#of\ hits}{\#of\ hits + \#of\ misses}$$
$$precision = \frac{\#of\ hits}{\#of\ hits + \#of\ false\ alarms} \qquad (۹-۲)$$

که بازخوانی[17] بیانگر نسبت تعداد انتخاب‌های درست به کل مرز شات‌ها و دقت[18] برابر نسبت تعداد انتخاب‌های درست به کل انتخاب‌های الگوریتم است. جدول ۲-۲ نتایج آزمون را نمایش می‌دهد. روش

[17] Recall
[18] Precision

پیشنهادی از هیستوگرام رنگ در حوزه مکانی استفاده می‌نماید که به حرکت دوربین و شیء و چرخش و جابجایی حساس نیست. در عین حال الگوریتم RLS نیز در برابر نویز سفید مقاوم است و به وقفه در حرکت حساس نیست. روش پارامتریک ارائه‌شده می‌تواند حرکت را از برخی تغییرات شات با کمک مدل‌سازی طولانی دنباله، تشخیص دهد؛ به همین خاطر در V1 و V2 که حرکات دوربین و شیء شدید است، مدل ارائه‌شده بهتر عمل می کند.

روش ۲ در حالات حل‌شدگی[19] نتایج خیلی بدی دارد، چون مقادیر پیکسل‌ها در این حالت در شرایط وابسته به تئوری اطلاعات صدق نمی‌کنند.

جدول۲-۲ نتایج آزمون تشخیص مرز شات [18].

ویدئو	روش	برش		محو شدگی		حل‌شدگی	
		P(%)	R(%)	P(%)	R(%)	P(%)	R(%)
V1	M1	۹٤,٥	۹۰,۲	N	N	N	N
	M2	۸٥,۸	۷۰,۲	N	N	N	N
	M3	۸۹,۸	۸٦,٤	N	N	N	N
V2	M1	۹٦,٦	۸٥,۰	N	N	N	N
	M2	۸۲,۳	٦۷,۷	N	N	N	N
	M3	۸٦,۹	۸۰,۸	N	N	N	N
V3	M1	۹٦,۲	۹۲,٦	۱۰۰	۹۳,۳	۱۰۰	۸۳
	M2	۱۰۰	۸۸,۹	۹۳,۳	۹۳,۳	٦۷	۳۳
	M3	۹۲,۳	۸۸,۹	۹۳,۳	۸٦	۸۹	٦۷
V4	M1	۹٤,٤	۹٦,۳	۱۰۰	۱۰۰	۸۳	۸۳
	M2	۹٤,۸	۹٥,۸	۱۰۰	۱۰۰	٦۷	۳۳
	M3	۹٥,۲	۹٤,۲	۱۰۰	۱۰۰	٦۷	٦۷
V5	M1	۹۷,٦	۹۷,٦	N	N	۱۰۰	۱۰۰
	M2	۹٥,۳	۹۷,٦	N	N	۱۰۰	٥۰
	M3	۹۳,۳	۱۰۰	N	N	۱۰۰	۱۰۰

۲-۳-۲- بحث در مورد روش AR

روش AR مدل بسیار مناسبی برای مدل‌کردن سلسله زمانی فریم‌ها است و خاصیت زمانی ویدئو به خوبی در این مدل استفاده می‌شود. در موضوع مطرح‌شده از نتایج مدل پارامتریک AR در کاربردهای اندکی استفاده شده است، در صورتی‌که به نظر می‌رسد بتوان کاربردهای دیگری هم برای آن با توجه به ویژگی‌هایش یافت. یکی از معایب این روش، بردار ویژگی به کار رفته در این زمینه است. بردار هیستوگرام رنگ گرچه بسیار ساده است ولی می‌توان با استخراج بردارهای مناسب‌تری نظیر پارامترهای

[19] Dissolve

تبدیل دو بعدی موجک و یا بردار ویژگی پارامترهای مدل مارکف درختی تبدیل‌یافته هر فریم نتایج و کاربردهای بهتری برای این روش توصیف نمود.

۲-٤- مدل مخفی مارکف

فرض کنید سیستمی با N حالت وجود داشته باشد، که در هر لحظه سیستم در یکی از حالات به سر می‌برد. در این صورت می‌توان زنجیره مارکف شکل ۲-٦ را برای انتقال حالت در سیستم فوق تصور کرد [14, 15, 16]. بنابراین پس از گذشت هر گام زمانی، سیستم به یکی از حالات $S_1,S_2,...,S_N$ می‌رود. این انتقال بر اساس احتمال انتقال از حالتی به حالت دیگر بیان می‌شود، که با ضرایب a_{ij} مشخص می‌گردد. فرض کنید حالت سیستم در لحظه t، q_t باشد، در این صورت برای تعیین حالت آینده سیستم در حالت کلی، تمامی اطلاعات گذشته سیستم مورد نیاز است. برای یک حالت خاص که به آن اصطلاحاً زنجیره مارکف یا مدل مارکف مرتبه اول گفته می‌شود، این وابستگی فقط به اطلاعات لحظه قبل معطوف است و حالت بعدی سیستم فقط از روی حالت فعلی تعیین می‌شود:

$$P(q_t = S_j | q_{t-1} = S_i, ..., q_1 = S_k) = P(q_t = S_j | q_{t-1} = S_i) \quad (۱۰-۲)$$

باز هم شرایط را از این محدودتر می‌کنیم و فقط به فرآیندهایی که در مدل فوق صدق می‌کنند و مستقل از زمان (ایستان) هستند می‌پردازیم. به این ترتیب داریم:

$$P(q_t = S_j | q_{t-1} = S_i) = P(q_2 = S_j | q_1 = S_i) \quad (۱۱-۲)$$

در این صورت می‌توان ضرایب a_{ij} را به صورت زیر مستقل از زمان تعریف کرد. یعنی احتمال انتقال از حالت S_i به حالت S_j در زمان دلخواه برابر است با:

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad (۱۲-۲)$$

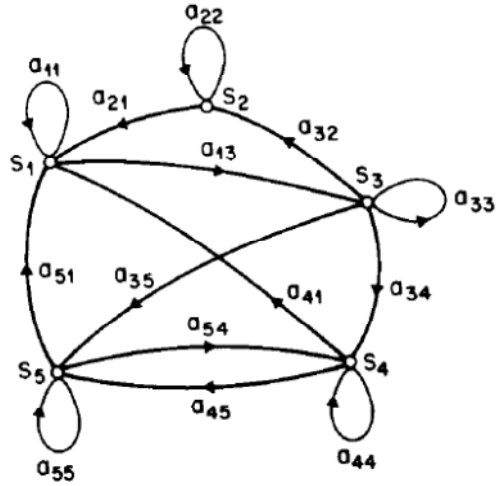
ضرایب فوق از جنس احتمال هستند و در شرایط زیر صدق می کنند:

$$a_{ij} \geq 0, \ \forall i, j$$
$$\sum_{j=1}^{N} a_{ij} = 1, \ \forall i \quad (۱۳-۲)$$

مدل فوق را می‌توان مدل مارکف مشاهده‌پذیر نامید. چون خروجی آن، حالات آن هستند که قابل مشاهده‌اند. بردار شرایط اولیه برای مدل مارکف مرتبه اول به صورت زیر تعریف می‌شود:

$$\pi = \{\pi_i\}, \ \pi_i = P(q_1 = S_i) \quad (۱٤-۲)$$

مدل مارکف پایا اگر فقط دارای یک بردار ایستان باشد، دارای حالت دائمی و ایستان است و هر مدل مارکف مرتبه اول ایستان، ارگادیک است. در مدل مارکف آشکار، هر حالت متناظر با یک پدیده فیزیکی قابل مشاهده است. اما این مدل بسیار محدود است و در خیلی از مسائل کارایی لازم را ندارد. پس لازم است که مدل را کمی توسعه داده و مشاهده در هر زمان، یک تابع احتمال از حالت قبلی سیستم باشد.



شکل ۲-٦ زنجیره مارکف با ۵ حالت [15].

در این حالت فرآیند مارکف را یک فرآیند تصادفی دوگانه نیز می‌نامند. چون هم حضور در حالات مخفی آن تابع احتمال و هم مشاهده آن یک فرآیند تصادفی است. در حقیقت به کمک مشاهده می‌توان حالت فعلی یا قبلی را پیش‌بینی کرد. یک مدل مخفی مارکف از ۵ جزء زیر تشکیل شده است:

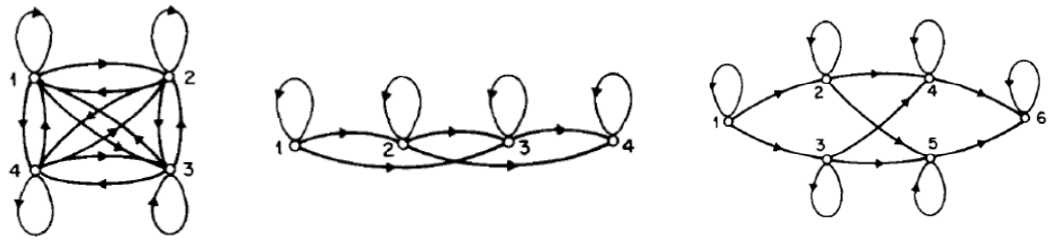
- N: تعداد حالات مخفی در مدل است.
- M: تعداد مشاهدات مستقل در هر حالت یا به عبارت دیگر سایز الفبای مستقل مشاهدات گسسته می‌باشد.
- A: ماتریس احتمال انتقال که با فرمول (۲-۸) تعریف می شود.
- ماتریس توزیع احتمال مشاهده نمونه در هر حالت را با نماد $B = \{b_j(k)\} = \{b_{jk}\}_{N\times M}$ مشخص می‌شود. هر درایه این ماتریس نشانه احتمال مشاهده سمبل k در حالت j است.

$$b_j(k) = P(o_{t+1} = V_k \,|\, q_t = S_j), \quad 1 \le j \le N, \quad 1 \le k \le M, \quad 1 \le t \le T \qquad (۱۵-۲)$$

بردار توزیع احتمال حالت اولیه سیستم: $\pi = \{\pi_i\},\ \pi_i = P(q_1 = S_i), \quad 1 \le i \le N$

با داشتن پارامترهای فوق یا به اختصار $\lambda = (A,B,\pi)$، مدل مارکف به طور کامل شناخته می‌شود و رشته مشاهدات $O = \{O_1, O_2, ..., O_T\}$ به روش زیر قابل تولید است:

- انتخاب حالت اولیه سیستم با توجه به بردار توزیع احتمال اولیه π.
- قرار دادن t=1.
- $O_t = V_k$ با توجه به توزیع احتمال مشاهده ماتریس B.
- انتقال به حالت j با توجه به ماتریس A.
- افزایش t=t+1 تا t=T و بازگشت به مرحله ۳.



(a) (b) (c)

شکل۲-۷ مثال‌هایی از سه نوع HMM، (a) مدل ارگادیک با ٤ حالت، (b) مدل چپ به راست[20] با ٤ حالت و (c) مدل چپ به راست با مسیرهای موازی[21] با ٦ حالت [15].

۲-٤-۱- مدل مخفی مارکف در حوزه تبدیلات

در [44] روش جدیدی برای پردازش سیگنال بر اساس مدل مخفی مارکف در حوزه موجک یک-بعدی ارائه شد و WD-HMM[22] نام گرفت. این روش به خوبی خاصیت آماری غیرگوسی سیگنال‌های طبیعی و همچنین رابطه بین‌سطحی ضرایب تبدیل موجک را مدل می‌کند. در این مدل برای هر ضریب موجک W_i، یک حالت مخفی گسسته S_i با تابع احتمال $P(S_i = m) = p_i^m, m = 1,...,M$ نسبت داده می‌شود. W_i مشروط به $S_i = m$ دارای توزیع گوسی با میانگین $\mu_{i,m}$ و واریانس $\sigma^2_{i,m}$ است. از آنجا که ضرایب موجک ناشی از کانولوشن فیلترهایی با میانگین صفر (فیلترهای بالاگذر) هستند، می‌توان میانگین آنها را صفر فرض کرد. علاوه بر این برای کاهش پارامترهای مدل، فرض می‌شود که تمامی ضرایب هر زیرباند دارای خواص آماری یکسانی باشند. برای هر زیرباند M=2 در نظر گرفته می‌شود و توزیع کناره‌ای به صورت ترکیب گوسی زیر برای زیرباند j ام نوشته می شود:

$$\begin{aligned} & f_j(w) = p_j^1 g(w;\sigma_{j,1}) + p_j^2 g(w;\sigma_{j,2}) \\ & p_j^2 + p_j^1 = 1 \\ & g(w;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-w}{2\sigma^2}) \end{aligned} \qquad (۱٦-۲)$$

در این مدل p_j^1 و p_j^2 به ترتیب احتمال اینکه ضریب W در زیر باند j ام دارای مقدار بزرگ یا کوچکی است، می‌باشد. مقادیر کوچک نتیجه تابع گوسی با واریانس کم و مقادیر بزرگ مربوط به تابع گوسی با واریانس بزرگ هستند. رابطه بین‌سطحی میان ضرایب فرزند موجک و والد[23] آنها وجود دارد. درتبدیل موجک دو بعدی هر ضریب در سطح پایین‌تر، ٤ فرزند در سطحی بالاتر دارد؛ که این ارتباط در شکل ۲-۸ نشان داده شده است. ماتریس انتقال حالت برای رابطه والد به فرزند در حالات مخفی مارکف وجود دارد که به شکل زیر تعریف می‌شود:

[20] Left-right Model
[21] Parallel Path Left-right Model
[22] Wavelet Domain Hidden Markov Model
[23] Parent

$$A_j = \begin{bmatrix} p_j^{1\to1} & p_j^{1\to2} \\ p_j^{2\to1} & p_j^{2\to2} \end{bmatrix}, \qquad j = 2,3,...,J. \tag{۱۷-۲}$$

که $p_j^{m\to m'}$ احتمال این است، که فرزند در لایه j ام در حالت m' باشد، در حالی‌که والدش در حالت m ام بوده است. مشخص است که مجموعه درایه‌های هر سطر ماتریس حالت برابر یک است. $\rho(i)$ را والد نقطه i در درخت ضرایب موجک در نظر بگیرید، داریم:

$$P(S_i = m) = \sum_{m'} P(S_{\rho(i)} = m')P(S_i = m \mid S_{\rho(i)} = m') \tag{۱۸-۲}$$

برای یک WD-HMM مقید[۲۴] که برای همه ضرایب هر زیرباند یک توزیع در نظر گرفته می‌شود، داریم:

$$p_j^m = \sum_{m'} p_{j-1}^{m'} p_j^{m'\to m}, \qquad j = 2,3,...,J. \tag{۱۹-۲}$$

اگر $p_j = [p_j^1 \; p_j^2]$ تعریف شود، رابطه بالا برابر $p_j = p_{j-1}A_j$ می‌شود در نتیجه:

$$p_j = p_1 A_2 A_3 ... A_j, \qquad \forall j = 2,...,J. \tag{۲۰-۲}$$

در نتیجه مدل مخفی مارکف برای ضرایب موجک که به درخت مخفی مارکف نیز مشهور است، به صورت زیر تعریف می‌شود:

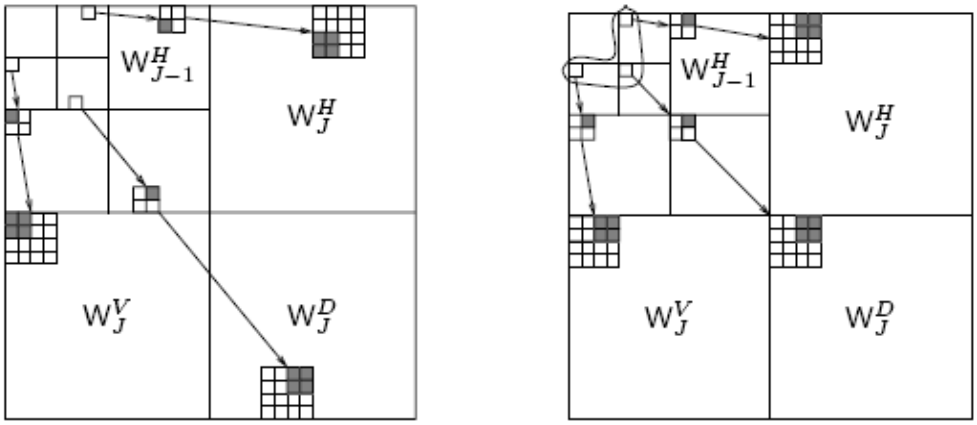
$$\Theta = \{p_1, A_2, ..., A_J; \sigma_{j,1}, \sigma_{j,2} (j = 1,...,J)\} \tag{۲۱-۲}$$

که J تعداد سطوح تبدیل است. این مدل به خوبی خواص آماری حاشیه‌ای[۲۵] و نیز رابطه والد-فرزندی ضرایب را مدل می‌کند. در [12] الگوریتمی کارا بر اساس EM[۲۶] برای یافتن پارامترهای مدل براساس ضرایب آموزش ارائه شده است. برای تصاویر دوبعدی، باید سه درخت تعریف و آموزش داده شوند؛ چون در هر لایه، سه زیرباند اصلی خواهیم داشت [11]. به این ترتیب رابطه والد–فرزندی مد نظر گرفته می‌شود ولی رابطه بین زیرباندها[۲۷] نادیده گرفته می‌شود. به این روش WD-HMM اسکالر (شکل ۲-۸a) می‌گویند و برای هر تصویر، سه دسته درخت تعریف و پارامترهای آنها استخراج می‌گردند. نتایج نشان می‌دهند که مدل مخفی مارکف به خوبی روابط بین سطحی و توزیع‌های حاشیه‌ای را مدل می‌کند.

در [11] برای در نظر گرفتن رابطه بین زیرباندی در تصاویر، از مدل WD-HMM برداری (۲-۸b) استفاده شده است. به طوری که ضرایب سه جهت در هر سطح را یک بردار فرض کرده و روابط به صورت برداری محاسبه می‌شوند . ضرایب موجک در جهت d (d=1,2,3) به ترتیب مربوط به جهات افقی، عمودی و قطری)، لایه j و مکان k را با نماد $w_{j,k}^d$ در نظر گرفته، بردار زیر تشکیل می‌شود:

[24] Tied WD-HMM
[25] Marginal Statistics
[26] Expectation Maximization
[27] Cousins

$$w_{j,k} = [w^1_{j,k}, w^2_{j,k}, w^3_{j,k}]^T \qquad (۲۲-۲)$$



(a) Scalar WD-HMM. (b) Vector WD-HMM.

شکل۲-۸ ساختارهای درختی WD-HMM. در حالت اسکالر سه درخت مدل درختی اسکالر تعریف می‌شود و در حالت برداری یک مدل برداری وجود دارد [47].

و تابع توزیع مخلوط گوسی نیز به صورت مقابل تعریف می‌شود:

$$\begin{aligned} f_j(w) &= p^1_j g(w; C_{j,1}) + p^2_j g(w; C_{j,2}) \\ p^2_j + p^1_j &= 1 \\ g(w; C) &= \frac{1}{\sqrt{(2\pi)^n |\det(C)|}} \exp(-w^T C^{-1} w) \end{aligned} \qquad (۲۳-۲)$$

n=3 و C ماتریس کواریانس است. پارامترهای مدل نیز به صورت زیر در نظر گرفته می‌شوند:

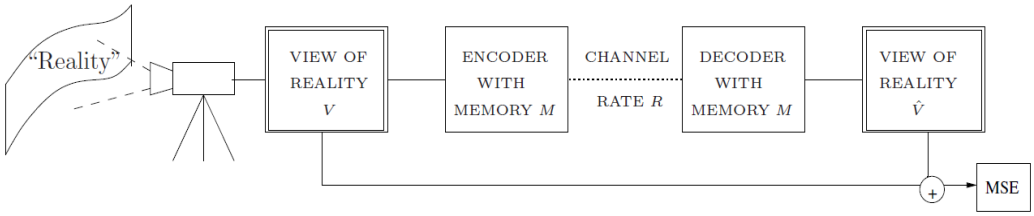
$$\Theta = \{p_1, A_2, ..., A_J; C_{j,1}, C_{j,2} (j = 1, ..., J)\} \qquad (۲٤-۲)$$

از نظر ادراکی نیز روابط درست هستند، زیرا اگر لبه‌ای در تصویر وجود داشته باشد، انتظار داریم ضرایب موجک در تمامی جهات مقدار بزرگی داشته باشند و برعکس در نواحی یکنواخت تصویر، انتظار داریم، ضرایب موجک در سه جهت مقدار کوچکی داشته باشند. رابطه بین جهات در ضرایب غیرقطری ماتریس‌های کواریانس قرار دارد. در این کار از پارامترهای به دست آمده برای بازیابی تصاویر استفاده شده است. در [48] از مدل مخفی مارکف اسکالر برای مدل کردن ضرایب تبدیل کانتورلت نیز استفاده شده است و نتایج آن برای بازیابی تصاویر به کار رفته است.

۲-٤-۲- میدان مارکف برای مدل‌سازی ویدئو

مدلی در [13] بر اساس فرآیندهای تصادفی ارائه گردیده و نرخ اطلاعات این مدل محاسبه شده است. دو منبع اطلاعات درنظر گرفته شده‌اند: حرکت دوربین و اطلاعات دیداری صحنه. منابع اطلاعات با هم ترکیب می‌شوند و یک بردار را تشکیل می‌دهند. در این مقاله نرخ اطلاعات برای حالت‌های بدون

اتلاف[۲۸] و با اتلاف[۲۹] استخراج شده‌اند. صحنه در دو حالت ثابت در زمان و متغیر با زمان مدنظر قرار گرفته است. مساله به صورت شکل ۲-۹ تعریف می‌شود.



شکل ۲-۹ مساله مدنظر برای مدل‌کردن تصادفی ویدئو[13].

V که در صحنه واقعی است، توسط کدکننده با طول حافظه M کد می‌شود و نرخ بیت متوسط R بیت را تولید می‌نماید. در طرف مقابل رشته بیت توسط دکدر با طول حافظه M دکد می‌شود و ویدئوی $\hat{V}$ را ایجاد می‌کند که از نظر معیار MSE شبیه ویدئوی ورودی است. در ابتدا در مدل ساده‌شده فرض می‌شود که صحنه ثابت است و دوربین در جهت افقی با متغیر تصادفی یک‌بعدی برنولی random walk حرکت می‌نماید. random walk فرآیند $W=(W_t : t \in \mathrm{Z}^+)$ است، که $\Pr\{W_0=0\}=1$ می‌باشد و برای $t \geq 1$، $W_t=\sum_{i=1}^{t} N_i$ است. $N_i \in \{-1,1\}$ها i.i.d با احتمال $\Pr\{N_1=1\}=pw$ درنظر گرفته می‌شوند و بدون از دست‌دادن کلیت مساله $pw \leq .5$ فرض می‌شود. در برابر دوربین دیواری است که با یک بردار یک‌بعدی مدل می‌شود (شکل ۲-۱۰) که $X=(X_n : n \in \mathrm{Z})$ هر درایه این بردار متغیرهای تصادفی i.i.d در نظر گرفته می‌شوند که از W هم مستقل هستند. X دارای توزیع px بر روی الفبای Xاست. در هر لحظه t برداری ($V=(V_t : t \in Z^+)$) به طول L>1 از دیوار توسط دوربین گرفته می‌شود. به این ترتیب هر تصویر با تصویر قبلی و بعدی‌اش فقط در یک درایه متفاوت است. در این مدل زوم، گردش دوربین و تغییر زاویه در نظر گرفته نشده است. اگر ویدئوی ورودی ما $V=(V_0,V_1,\ldots)$ باشد، خروجی $\hat{V}=(\hat{V}_0,\hat{V}_1,\ldots)$ است که دکدر آن را بر اساس مقداری تاخیر ایجاد می‌کند. کدر با حافظه M، برای کدکردن V_t می‌تواند از تصاویر $V_{t-1},\ldots,V_{t-M}$ استفاده کند. به این ترتیب کدینگ می‌تواند با اتلاف یا بدون اتلاف درنظر گرفته شود.

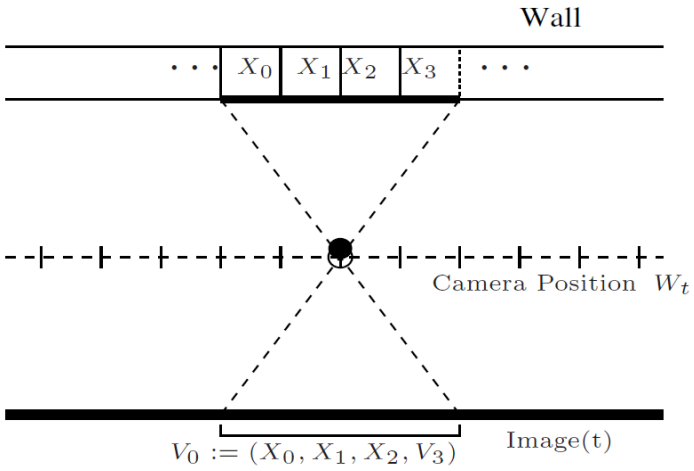
برای حالت بدون اتلاف نرخ اطلاعات به شکل زیر محاسبه می شود:

$$H(V)=\lim_{t\to\infty}\frac{1}{t}H(V^t)=\lim_{t\to\infty}H(V_t|V^{t-1}) \qquad (۲۵-۲)$$

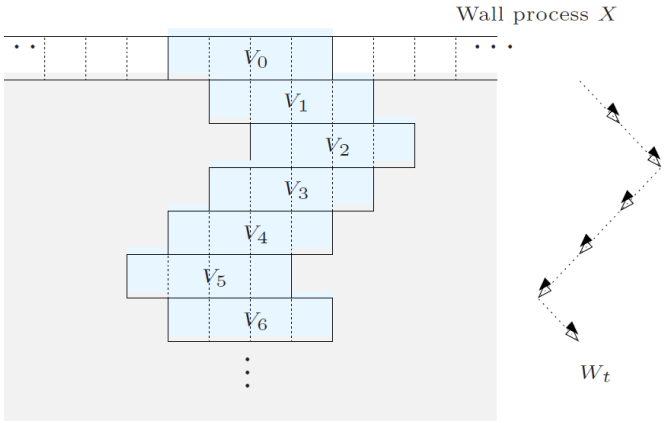
که $V^t=(V_1,\ldots,V_t)$ و فرض می‌شود که V_0 برای دکدر مشخص است و باند بالایی و پایینی برای نرخ اطلاعات برابر است با:

[28] Lossless
[29] Lossy

$$(1-2pw)H(X)+H(pw)\Pr\{\overline{A}_L\}\le H(V)\le(1-2pw)H(X)+H(pw)$$
$$A_L=\{(X_0,...,X_L)=(x_0,x_1,x_0,x_1,...),x_0,x_1\in X\} \qquad (۲۶-۲)$$



(الف)



(ب)

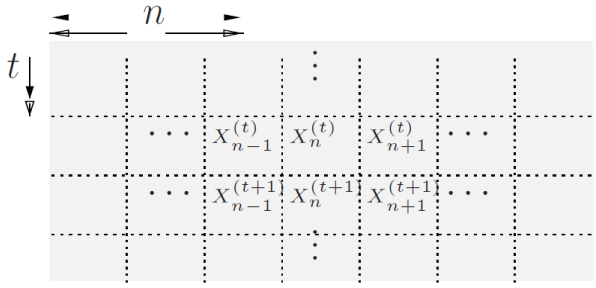
شکل ۲-۱۰ (الف) مدل ساده شده (ب) بردار متغیر تصادفی V [13].

و برای حالت باتلف معیار زیر تعریف می‌شود:

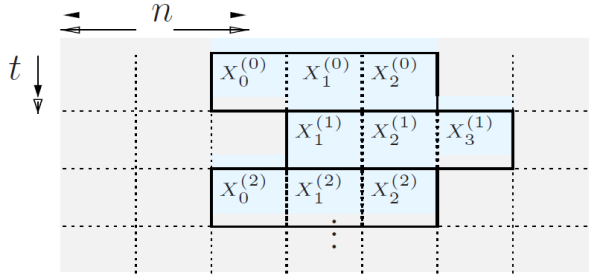
$$R_{V^t}=\inf\{t^{-1}I(V^t;\hat{V}^t):p(\hat{V}^t|V^t)\ such\ that\ d(V^t,\hat{V}^t)\le D\} \qquad (۲۷-۲)$$

و باندهای بالایی و پایینی به شکل زیر محاسبه می‌شوند:

$$(1-2pw)R_X(D) \le R_V(D) \le (1-2pw)R_X(D) + H(pw)$$
$$A_L = \{(X_0,...,X_L) = (x_0,x_1,x_0,x_1,...), x_0,x_1 \in X\} \quad (۲۸-۲)$$



(الف)



(ب)

شکل ۲-۱۱ (الف) مدل دینامیک که در بعد زمان مارکف است و در بعد مکان متغیر تصادفی i.i.d است (ب) حرکت در یک میدان تصادفی رخ می‌دهد[13] .

حال در نظر می‌گیریم که صحنه نیز با زمان تغییر می‌نماید. برای مدل‌کردن این حالت از میدان تصادفی استفاده می‌شود (شکل ۲-۱۱).

$$(1-2pw)R_X(D) \le R_V(D) \le (1-2pw)R_X(D) + H(pw)$$
$$A_L = \{(X_0,...,X_L) = (x_0,x_1,x_0,x_1,...), x_0,x_1 \in X\}$$
$$RF = \{X_{W_t}^{(t)} : (n,t) \in Z \times Z^+\} \quad (۲۹-۲)$$
$$V_t = (X_{W_t}^{(t)}, X_{W_{t+1}}^{(t)}, ..., X_{W_{t+L-1}}^{(t)})$$

و نرخ اطلاعات در حالت بدون تلف برابر است با:

$$H(pw) - H(P_e) + H(V|W) \le H(V) \le H(pw) + H(V|W) \quad (۳۰-۲)$$
$$P_e = \Pr\{\hat{W}_t(Y) \ne W_t\}$$

و برای حالت با اتلاف و میدان تصادفی گوسی AR نیز نرخ اطلاعات محاسبه شده است. نتایج به دست آمده نشان می‌دهند که کدینگ DPCM کدینگ غیربهینه‌ای است.

۲-٤-۳- مروری بر سایر مدل‌های مبتنی بر مدل مارکف

در [49,50] از مدل HMM برای جداسازی و شناسایی احساسات از روی دنباله ویدئویی زمان واقعی که از صورت افراد مختلف گرفته می‌شود، استفاده می‌گردد. در این کار برای هر حالت یک HMM تعریف و آموزش داده می‌شود، سپس دنباله ویدئویی از هر HMM می‌گذرد و دکد می‌گردد. در نهایت HMMای که احتمال شرطی بیشتری داشته باشد، به عنوان خروجی در نظر گرفته می‌شود. انتخاب ساختار HMM نیز مهم است. در [50] برای هر احساس – مانند خشم، شادی و ...– یک HMM دارای سه حالت و چپ به راست تعریف شده است. تعداد پارامترهای این مدل نسبت به مدل [49] کم‌تر و آموزش آن راحت‌تر است ولی آزادی عمل کم‌تری دارد. در [49] از مدل ارگادیک استفاده شده است و ادعا شده است که اگر تعداد داده‌های آموزش زیاد باشند، در صورتی که مدل واقعی چپ به راست باشد، در مرحله آموزش ماتریس انتقال مدل ارگادیک خودبه‌خود به مدل چپ به راست متمایل خواهد شد. مدل ارگادیک دارای پارامترهای بیشتر و آزادی عمل بیشتری است ولی پیچیدگی‌های بیشتری نیز دارد. در این کار، شناسایی حالت در سطح پایین و با ردگیری صورت و استخراج ویژگی-هایی در هر فریم، صورت می‌گیرد و جداسازی بین حالات نیز با یک زنجیره مارکف در سطح بالاتر انجام می‌پذیرد. از مدل‌های مخفی مارکف در سیستم‌های ترکیبی صوت و ویدئو[۳۰] نیز استفاده می-شود [51]. مثلاً در کاربردهایی مانند شناسایی اشارات دست با کمک تحلیل صحبت از مدل HMM استفاده می‌شود [52]. در [53] از مدل HMM برای بهبود کیفیت صحبت از روی مدل‌سازی حرکات لب در تلفن‌های تصویری استفاده شده است. مدل‌سازی حرکات لب و لب‌خوانی با مدل HMM در [54] انجام شده است. کاربرد دیگر این مدل در ردگیری هدف (صورت) در کاربردهای زمان واقعی است [55]. در [56] از مدل مخفی مارکف برای دسته‌بندی ویدئوهای اخبار استفاده شده است. HMMها بر اساس ویژگی‌های استخراج‌شده از متن نوشته‌شده در اخبار و ردگیری صورت خبرنگار آموزش می‌بینند و سپس کلاس‌بندی انجام می‌شود.

۲-٤-٤- بحث در مورد روش HMM

مدل مارکف امکان مدل‌کردن روابط پیچیده را فراهم می‌نماید. این مدل‌ها همانند مدل‌های AR روابط زمانی را به خوبی مدل می‌کنند. می‌توان از مدل مارکف لایه‌ای برای مدل‌سازی ویدئو استفاده نمود که لایه بالا برای مدل‌کردن تسلسل زمانی و مدل پایین برای مدل‌کردن فریم‌ها یا تبدیل‌یافته آن‌ها استفاده شوند. مدل مارکف درختی را می‌توان برای تبدیلات سه‌بعدی تعریف و استفاده کرد. در این حالت می-توان مدل تکه‌ای تعریف نمود و باز هم مشکلی که باقی می‌ماند، همان برخورد یکسان با اطلاعات زمانی

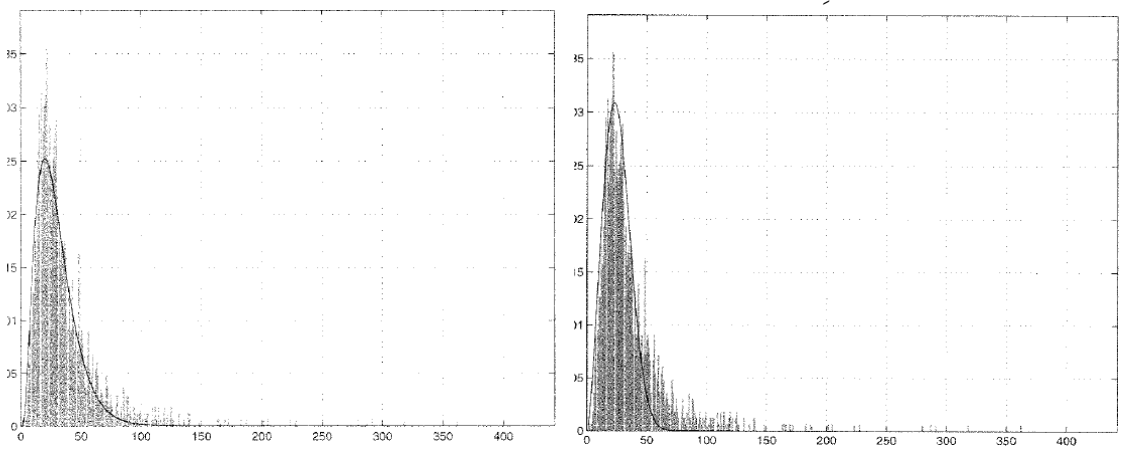
[30] Speech-to-Video Synthesizer

و مکانی است. می‌توان این مدل را با مدل AR ترکیب نمود و از پارامترهای استخراج‌شده از مدل مخفی مارکف از هر فریم، به عنوان بردار ویژگی در مدل زمانی AR استفاده کرد.زنجیره مارکف مطرح‌شده درباره تصویر در مورد ویدئو هم قابلیت استفاده را دارد ولی همانند آن‌چه در مورد GMM مطرح شد، با این کار علاوه بر مشکل تاخیر، مساله ارتباطات زمانی هم مطرح می شود و در ضمن این مدل دارای پیچیدگی‌های بسیار زیادی است. مساله افزونگی نیز در اینجا حل نشده است، چون زنجیره مارکف برای هر پیکسل فقط همبستگی با یک همسایه را در نظر می‌گیرد که در سیگنال سه‌بعدی هر پیکسل ۲۶ همسایه دارد.

۲-۵- مروری بر برخی مدل‌های دیگر

۲-۵-۱- مدل‌های آماری برای استخراج محتوا

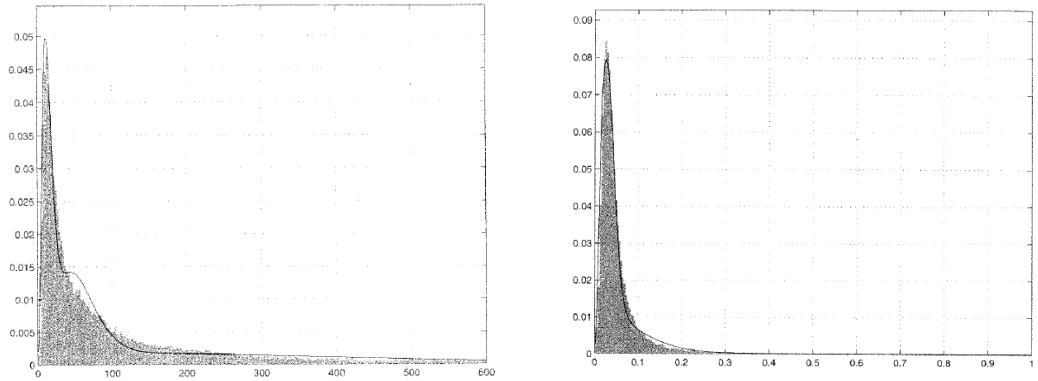
در [29] روشی برای استخراج اطلاعات از دنباله ویدئو، درباره مفهوم آن ارائه شده است. در این روش از دو عنصر طول شات و فعالیت صحنه استفاده می‌شود و برای هرکدام از این دو، مدل آماری ارائه می‌گردد. در نهایت از این دو عنصر استخراج‌شده در فرمول بیزی[31] برای حل مشکلات جداسازی شات استفاده می‌شود و نتایج حاصل با نتایج استفاده از سطوح آستانه ثابت مقایسه می‌شوند؛ که بهبود قابل توجهی در نتایج دیده می‌شود. کاربرد دیگری که برای این روش مطرح می‌شود، استفاده از این دو عنصر به عنوان ویژگی برای بازیابی ویدئو است. نشان داده می‌شود که مدل بیزی به دست آمده در این مورد نیز بسیار کارا است و به خوبی ویژگی‌های محتوا را استخراج می‌کند. دوره شات توسط مدل Erlang و مدل Weibul مدل می‌شود؛ که تقریب خوبی با توجه به نتایج به دست آمده ارائه می‌دهد. ایراد مدل Erlang در فرضِ مربوط به ثابت‌بودن نرخ دریافت است.



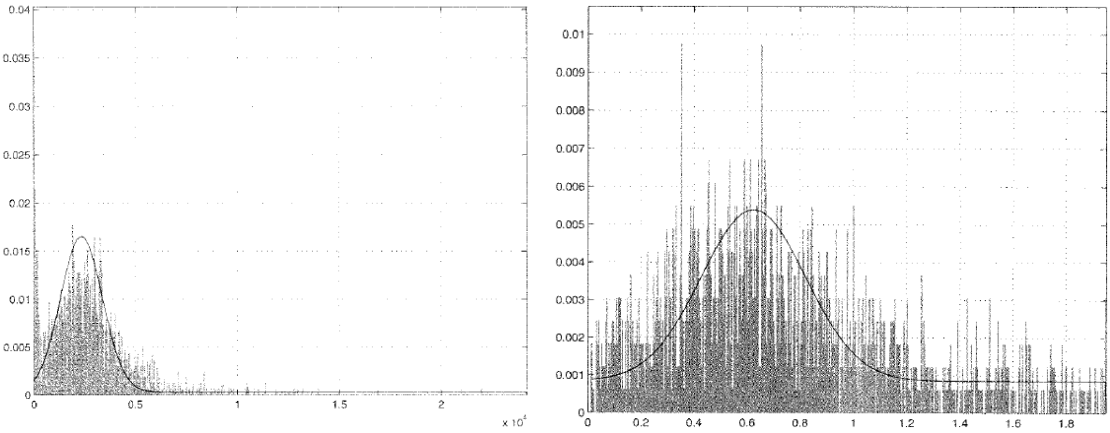
شکل ۲-۱۲ هیستوگرام دوره زمانی و مدل منطبق‌شده به آن، چپ مدل Erlang و راست مدل Weibul [29].

[31] Bayesian

فعالیت در شات نیز با ترکیبی از چند توزیع در نظرگرفته می‌شود. برای فریم‌های عادی، مدلی مرکب از سه توزیع Erlang و یک توزیع یکنواخت منظور می‌گردد و برای فریم‌های حالت گذرا بین دو شات نیز ترکیبی از یک توزیع گوسی و یک توزیع یکنواخت فرض می‌شود. علاوه بر این، دو معیار، فاصله هیستوگرام رنگ و فاصله تانژانتی هم در این مدل استفاده می‌شوند. شکل‌های ۲-۱۳ و ۲-۱٤ نتایج این تقریب‌ها را نشان می‌دهند. این روش در جداسازی شات ویدئو به کار گرفته شده است.



شکل ۲-۱۳ تقریب هیستوگرام فعالیت شرطی برای فریم‌های عادی. چپ فاصله هیستوگرام و راست فاصله تانژانت[29] .



شکل ۲-۱٤ تقریب هیستوگرام فعالیت شرطی برای فریم‌های گذرا. چپ فاصله هیستوگرام و راست فاصله تانژانت[29].

۲-۵-۲- مدل‌سازی حرکات صورت و بدن انسان

یک سیستم کدینگ ویدئو بر مبنای مدل، از مدل‌هایی برای توصیف اشیای سه‌بعدی در ویدئو استفاده می‌کند. مثلاً از مدل‌هایی برای توصیف صورت انسان، مقادیر پارامترهای مش‌های سه‌بعدی و بردارهای حرکات مدل‌ها با تحلیل ویدئو استخراج می شوند. این پارامترها و تخمین‌های حرکت ارسال و در طرف گیرنده سنتز و بازسازی می شوند [57]. مدل‌سازی صورت و بدن و حرکات آن و انیمیشن یک موضوع رایج با سابقه بیش از ۲۵ سال در تحقیقات گرافیک کامپیوتری است [59 ,58]. در کتاب Park و Water مرور کلی بر این مسائل ارائه شده است [60]. در گروهی از کارها تکنیک‌هایی برای تصفیه و

ثبت اطلاعات گرفته‌شده از اسکنرهای لیزری مطرح گردیده‌اند. مدل به دست آمده سپس با کمک روشی فیزیکی به صورت انیمیشن در می‌آید. در روش دیگری، تولید مدل‌های صورت بر اساس پارامترها با اندازه‌گیری‌های تصادفی صورت، بر طبق شاخص‌های آماری بدن‌سنجی انجام می‌گیرد. با کمک این اندازه‌گیری‌ها به عنوان قیود، محدوده گسترده‌ای از ویژگی‌های هندسی صورت تحت پوشش قرار می‌گیرد. دسته دیگری از مدل‌سازی‌های ویدئو، از دو تصویر از زوایای مختلف استفاده می‌کنند. زاویه دو دوربین استفاده‌شده طوری است، که تصاویر آنها بر هم عمود باشند. سیستم دیگری از کانتورهای تصاویر برای تولید مدل‌های هندسی اشیاء استفاده می‌کند [60]. در روش دیگری، کاربر تناظر بین چندین تصویر را تعیین می‌کند و از روش‌های بینایی برای بازسازی تصاویر سه‌بعدی استفاده می‌کند. سپس یک مدل مش سه‌بعدی به این نقاط سه‌بعدی بازسازی‌شده تطبیق داده می‌شود. این روش به خوبی مدل‌های طبیعی صورت را ایجاد می‌کند؛ ولی نیاز به کاربر دارد. در [61] نیز مدلی پارامتریک برای حرکات بدن انسان ارائه شده است.

۲-٦- کاربردهای تحلیل و مدل‌سازی ویدئو

در این قسمت به بیان مختصری درباره کارهای مرتبط انجام شده در دو کاربرد پردازش ویدئو که در این رساله برای ارزیابی روش‌های تحلیل و مدل‌سازی ویدئو به کار گرفته شده‌اند، می‌پردازیم. این کاربردها شامل چکیده‌سازی ویدئو[۳۲] و تشخیص رفتار انسان[۳۳] هستند:

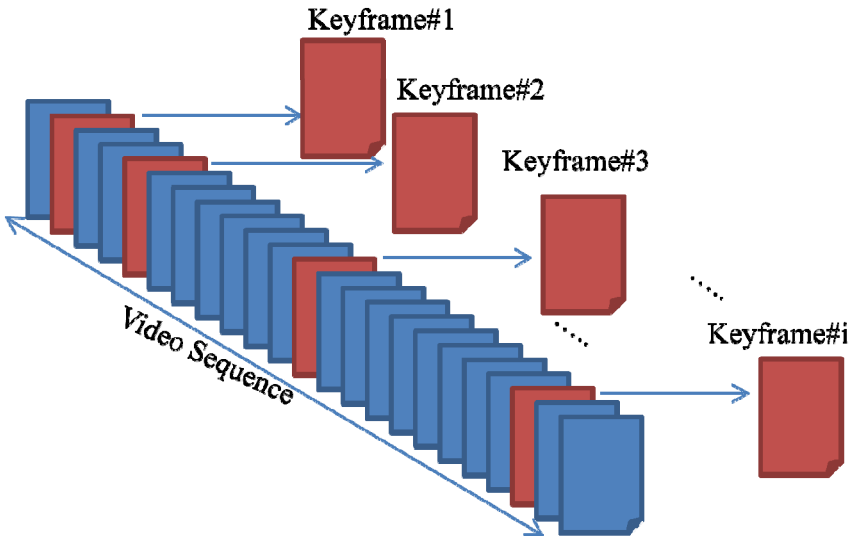
- نتایج چکیده‌سازی ویدئو که شامل انتخاب فریم‌های کلیدی می‌شود، می‌تواند در بسیاری از کاربردهای پردازش ویدئو شامل بازیابی ویدئو، ویرایش ویدئو و فشرده‌سازی ویدئو مورد استفاده قرار گیرد. انتخاب پارامترهای مکانی مناسب و روش تحلیل تحولات زمانی درخور، دو فاکتور مهم در تعیین دقت و کارایی این مساله است.
- تشخیص رفتار انسان یک زمینه تحقیق مهم در مدل‌سازی و تحلیل ویدئو را به خود اختصاص داده است [62-79] و می‌تواند در کاربردهای تحلیل فعالیت انسان، تشخیص ژست، بیومتریک، بازیابی و شاخص‌گذاری ویدئو و سیستم‌های نظارت به کار رود. در تشخیص رفتار انسان، ویژگی‌های سیگنال از دنباله ویدئویی استخراج‌شده برای تعیین رفتار استفاده می‌شوند. بنابراین، ویژگی‌های استخراج‌شده و روش کلاس‌بندی تاثیر به سزایی در کارایی و دقت این سیستم‌ها دارند.

[32] Video abstraction
[33] Human action recognition

۲-۶-۱- چکیده‌سازی ویدئو

چکیده‌سازی ویدئو به فرآیند تولید خلاصه‌ای از محتوای دنباله ویدئویی اطلاق می‌شود [80, 81]. این خلاصه می‌تواند اطلاعات کلیدی درباره سیگنال ویدئو به کاربر دهد که طول زمانی بسیار کوتاه‌تری از سیگنال اولیه دارد. دو نوع طبقه‌بندی اصلی برای روش‌های چکیده‌سازی ویدئو وجود دارد: چکیده-سازی بر پایه تصاویر ساکن (فریم‌های کلیدی) و چکیده‌سازی بر پایه تصاویر متحرک (ویدئوهای مختصرشده[۳۴]). چکیده‌سازی بر پایه تصاویر ساکن بسیار سریع‌تر از چکیده‌سازی بر پایه ویدئوهای مختصر است، زیرا در پروسه انتخاب آنها نیازی به در نظر گرفتن اطلاعات صوتی و متنی یا توجه به مسائل زمان‌بندی و همزمان‌سازی ویدئو وجود ندارد و تنها محتوای بصری سیگنال اهمیت دارد. همچنین، با نمایش فریم‌های کلیدی به ترتیب زمانی استخراج‌شان، طی الگوریتمی سریع و آسان، کاربران دید خوبی درباره محتوای بصری ویدئو به دست می‌آورند. از طرف دیگر، ویدئوهای اختصاریافته، می-توانند حاوی اطلاعات بیشتری بوده و برای کاربران جذاب‌تر باشند [81].

فریم‌های کلیدی حاوی اطلاعات مهمی درباره محتوای ویدئو بوده، سیر محتوایی سیگنال ویدئو را با کمک چند فریم انتخابی نمایش می‌دهند (شکل ۲-۱۵). فریم‌های کلیدی می‌توانند در بسیاری از کاربردهای پردازش ویدئو نظیر شناسایی صحنه[۳۵]، خلاصه‌سازی ویدئو[۳۶]، بازیابی ویدئو، ویرایش ویدئو و فشرده‌سازی به کار گرفته شوند.



شکل ۲-۱۵ دنباله ویدئویی و فریم‌های کلیدی انتخاب‌شده

[34] Video skims
[35] Scene detection
[36] Summarization

در روش‌های سنتی استخراج فریم‌های کلیدی، برخی ویژگی‌ها از هر فریم استخراج می‌شوند و سپس مرزهای شات‌ها و خوشه‌ها تعیین می‌گردند. در ادامه، فریم‌های کلیدی بر اساس یک معیار فاصله در فضای ویژگی انتخاب می‌شوند.

یکی از ساده‌ترین روش‌ها برای انتخاب فریم کلیدی، انتخاب اولین فریم یا فریم وسط از هر شات است [82,83] که لزوماً نتیجه خوبی نمی‌دهد. زیرا معمولاً برای توصیف یک شات بیش از یک فریم کلیدی مورد نیاز است. برخی روش‌های سریع دیگر اولین فریم شات را به عنوان فریم کلیدی انتخاب می‌کنند و فاصله بین آخرین فریم کلیدی منتخب را با فریم کنونی محاسبه می‌نمایند و اگر این فاصله بیش از یک سطح آستانه مشخص باشد، آن فریم به عنوان فریم کلیدی جدید انتخاب می‌شود و روند ادامه می‌یابد، گرچه فریم کلیدی منتخب با این روش ممکن است انتخاب خوبی نباشد و محتوای بصری ویدئو را به خوبی نشان ندهد [82]. روش‌های دیگر انتخاب فریم کلیدی براساس می‌نیمم‌های محلی در منحنی‌های فعالیت یا حرکت نیز روش‌هایی سیستماتیک هستند که معمولاً بار محاسباتی بالایی دارند، ضمن این که سکون معیار دقیقی برای انتخاب فریم کلیدی محسوب نمی‌شود [17,84].

در [17] از روش مدل‌سازی AR[۳۷] برای مدل‌کردن ویدئو بر اساس ویژگی‌های رنگ استفاده شده است. اما ویژگی‌های رنگ، پارامترهای مناسبی برای بیان محتوای بصری و بافتی سیگنال به حساب نمی‌آیند.

در [83] روشی بر اساس تئوری اطلاعات به کار رفته است که برای تقسیم‌کردن دنباله ویدئویی به بلوک‌های کوچک‌تر و استخراج فریم‌های ویدئویی استفاده شده است و با بازخوانی بالا و دقت پایین مرز شات‌ها را تعیین می‌کند. روشی مبنی بر خوشه‌بندی[۳۸] با پیچیدگی‌های محاسباتی زیاد در [82] آمده است که نزدیک‌ترین فریم به مرکز خوشه را به عنوان فریم کلیدی انتخاب می‌نماید. برای بررسی محتوای بصری دنباله ویدئویی جهت تعیین فریم‌های کلیدی، باید ویژگی‌های مکانی مناسبی استخراج شوند. این ویژگی‌ها می‌توانند اطلاعات حرکت [85,86]، اطلاعات هیستوگرام‌های رنگ [82,17] یا ویژگی‌های استخراج‌شده از اعمال یک تبدیل دوبعدی به فریم‌ها [87] باشند که در روش‌های انتخاب فریم کلیدی موجود به کار رفته‌اند.

برای ارزیابی نتایج استخراج فریم‌های کلیدی از آزمون‌های ادراکی استفاده می‌شود. زیرا روش استانداردی برای ارزیابی تحلیلی نتایج استخراج فریم‌های کلیدی نیست [80]. برای این منظور از

[37] Auto-regressive
[38] Clustering-based

دادگان‌های رایج در این زمینه استفاده می‌گردد که در ادامه این فصل به توضیحاتی درباره آن‌ها خواهیم پرداخت.

۲-۶-۲- تشخیص رفتار انسان

تشخیص رفتار انسان یک زمینه فعال تحقیق در تحلیل و مدل‌سازی ویدئو به حساب می‌آید و از نتایج آن می‌توان در کاربردهایی مانند تحلیل فعالیت، تشخیص ژست[۳۹]، بیومتریک‌ها[۴۰] [63]، شاخص-گذاری[۴۱] و بازیابی ویدئو و سیستم‌های نظارتی[۴۲] استفاده کرد. در [65] یک دیدگاه سلسله‌مراتبی برای گروه‌بندی اعمال انسان تعریف شده است:

- حرکت اولیه انسان[۴۳]: هر حرکت کوچک عضو بدن.
- رفتار انسان[۴۴]: مجموعه‌ای از حرکات اولیه که باعث حرکت کل بدن می‌شود.
- فعالیت انسان[۴۵]: مجموعه‌ای از رفتارهای انسان.

در اینجا منظور از رفتار، مانند آنچه در [63,65,66] آمده است، مجموعه‌ای از الگوهای ساده حرکتی است که توسط یک فرد در دنباله ویدئویی انجام می‌شود، مانند راه رفتن و مشت زدن مساله مهم در سیستم‌های تشخیص رفتار انسان، استخراج ویژگی‌های مناسب از دنباله ویدئویی است که یک مساله کلاس‌بندی محسوب می‌شود.

دو دید اصلی برای انتخاب ویژگی‌های رفتاری وجود دارد [65]: نمایش تصویر محلی[۴۶] و نمایش تصویر کلی[۴۷].

- دیدگاه کلی، کل بدن انسان را به عنوان ناحیه مورد نظر[۴۸] محسوب می‌کند و ویژگی‌ها را از این ناحیه استخراج می‌نماید. این روش اطلاعات غنی را برای رسیدن به نتایج عالی به کار می‌گیرد و روند استخراج ویژگی آن پیچیده نیست. این روش معمولاً نیاز به استخراج پس‌زمینه یا

[39] Gesture
[40] Biometrics
[41] Indexing
[42] Surveillance
[43] Primitive action
[44] Human action
[45] Human activity
[46] Local image representation
[47] Global image representation
[48] Region of Interest

تکنیک‌های دنبال‌کردن بدن دارد و به نویز و همپوشانی[49] حساس است [67,68]. برای حل این مشکلات، روشی بر مبنای شبکه توری[50] که نسخه دیگری از این دسته کلی است، استفاده می‌شود [69]. در این روش منطقه مورد نظر به سلول‌های مکانی کوچک‌تری تقسیم و هر سلول به صورت محلی کد می‌شود. یک گروه از فریم‌ها نیز می‌توانند با هم برای تشکیل سلول‌های سه‌بعدی و تولید توصیف‌گرهای مکانی–زمانی استفاده شوند [70]. به هر حال در این روش‌ها، نیاز به داشتن اطلاعات کلی درباره بدن وجود دارد.

- در دیدگاه محلی، توصیف‌گرهای محلی به عنوان مجموعه‌ای از تکه[51]های مستقل حول نقاط مورد نظر[52] محاسبه می‌شوند [71-73]. نقاط مورد نظر، نقاطی هستند که تغییرات ناگهانی در ابعاد زمانی یا مکانی در آنها اتفاق می‌افتد زیرا این نقاط حاوی اطلاعات بیشتری از دیگر نقاط در دنباله ویدئویی هستند. تکه‌ها سپس تشکیل مجموعه ویژگی‌ها[53] را می‌دهند. این دیدگاه در برابر نویز، همپوشانی جزئی و تغییرات پس‌زمینه تا حدی مقاوم‌تر است و به یک مرحله پیش‌پردازش قوی نیاز دارد. همچنین، تعداد متفاوت و معمولاً زیادی از توصیف‌گرها در این روش به کار می‌روند. در نتیجه مقایسه دو دنباله ویدئویی ساده نیست و تکه‌ها معمولاً دسته‌بندی می‌شوند و کتاب‌های کد برای بیان مجموعه‌های ویژگی‌های دنباله ویدئویی به کار می‌روند. علاوه بر این در اطلاعات استخراجی معمولاً افزونگی بالایی هست. برخی از روش‌ها همبستگی بین ابعاد زمانی و مکانی تکه‌ها را برای کاهش افزونگی استفاده می‌کنند [72, 75, 76].

۲-۷- دادگان‌های ویدئو

در این قسمت، به اختصار به معرفی دادگان‌های ویدئو مورد استفاده در این رساله می پردازیم.

۲-۷-۱- دادگان TRECVID[54]

هدف اصلی در TRECVID ایجاد امکان پیشرفت در روش‌های بازیابی محتوای ویدئوی دیجیتال با فراهم نمودن محیط و معیارهای در دسترس برای ارزیابی روش‌هاست [88]. این مجموعه در صدد ایجاد مدل شرایط واقعی برای ارزیابی است. به طور خاص TRECVID2006 شامل ویدئوهای اخبار عربی، انگلیسی و چینی مربوط به دوره زمانی ۲ سال است که برای ارزیابی روش‌های سه هدف مجزا

[49] Occlusion
[50] Grid-based
[51] Patch
[52] Interest points
[53] Bag-of-features
[54] TREC Video Retrieval Evaluation

ایجاد شده است. ما در این رساله از مجموعه TRECVID 2006 مربوط به تعیین انواع مرز شات و انتخاب فریم کلیدی استفاده می‌نماییم.

این مجموعه شامل ۱۳ ویدئوی طولانی با ۵۹۷۰۴۳ فریم ویدئو است که ۳۷۸۵ تغییر شات – که ٤٨,٧٪ آنها به صورت برش[55] و مابقی ۵۱,۳٪ تغییر تدریجی شات[56] هستند- و نرخ نمونه‌برداری زمانی ویدئوها ۲۹ فریم در ثانیه است. جزئیات بیشتری در این زمینه در جدول ۲-۳ آمده است. ویدئوهای این مجموعه دارای انواع ویژگی‌ها و حالات هستند و فعالیت‌های مختلفی را در بردارند. ویدئوهای بایگانی-شده سیاه و سفید که مربوط به سال‌های اولیه استفاده از ویدئو بوده است نیز در این دادگان وجود دارد.

جدول ۲-۳- جزئیات دادگان تعیین مرز شات TRECVID 2006.

تعداد دنباله‌های ویدئویی		۱۳
حجم ویدئوها		٤,٢٤ GB
تعداد فریم‌ها		۵۹۷۰٤۳
تعداد تغییرات[57] شات‌ها		۳۷۸۵
نوع تغییرات	برش[58]	۱۸٤٤(٪٤۸,۷)
	حل‌شدگی[59]	۱۵۰۹(٪۳۹,۹)
	محوشدگی[60]	۵۱(٪۱,۳)
	سایر	۳۸۱(٪۱۰,۱)

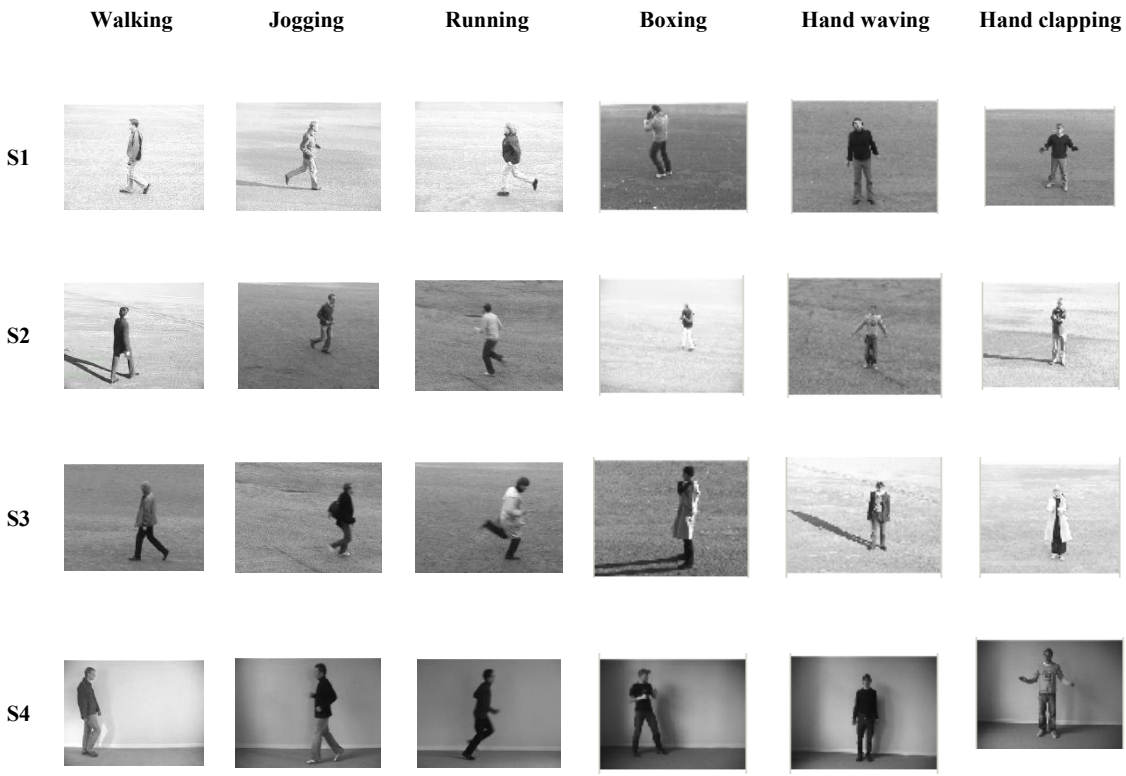
۲-۷-۲- دادگان Hollywood

این دادگان شامل ۱۲ کلاس از رفتار انسان و ۱۰ کلاس متفاوت شات است که دارای بیش از ۳٦٦۹ دنباله ویدئویی است و در مجموع حدود ۲۰ ساعت دوره زمانی این مجموعه است [89]. هدف از ایجاد این مجموعه ارائه بستری برای آزمون و پیشبرد روش‌های شناسایی فعالیت انسان در محیط‌های متفاوت است. محتویات این دادگان از بیش از ٦۹ فیلم ویدئویی انتخاب شده‌اند.

قسمت مورد استفاده از این مجموعه در این رساله مربوط به شات‌های مختلف آن است که شامل ۱۱۵۲ دنباله ویدئویی است. این مجموعه ۱۰۲۵۲۷۸ فریم ویدئویی و ۸۱۹۹ تغییر شات (آنی و تدریجی) می‌دارد. این مجموعه شامل فضاهای متنوعی مانند بیرون خانه، جاده، اتاق، هتل، آشپزخانه، اداره و مرکز خرید است.

[55] Cut transition
[56] Gradual transition
[57] Transition
[58] Cuts
[59] Dissolves
[60] Fade in/out

۲-۷-۳- دادگان KTH



شکل ۲-۱۵- نمونه فریم‌هایی از دیتابیس KTH [79].

این دادگان دارای ۲۳۹۱ دنباله ویدئویی است که توسط ۲۵ نفر در ٤ سناریوی مختلف در حال انجام شش رفتار شامل مشت زدن، کف زدن، تکان‌دادن دست، راه رفتن، دویدن آرام و دویدن سریع فیلمبرداری شده است [79]. سایز هر فریم ویدئو ۱٦۰ در ۱۲۰ پیکسل است و دوره زمانی هر دنباله ویدئویی متفاوت و سرعت نمونه‌برداری زمانی برابر ۲۵ فریم بر ثانیه است. ویدئوها در چهار سناریوی: محیط بیرونی، محیط بیرونی با تغییرات مقیاس، محیط بیرونی با لباس‌های مختلف و محیط داخلی گرفته شده‌اند. نقطه دید دوربین در ویدئوهای مختلف متفاوت است ولی دوربین تقریباً استاتیک می‌باشد. این دادگان به سه قسمت آموزش (۸ نفر) ، تایید (۸نفر) و تست (۹ نفر) تقسیم شده است. شکل ۲-۱٦ برخی از نمونه فریم‌های ٦ رفتار مختلف (ستون‌ها) در ٤ سناریو (سطرها) را نشان می‌دهد. دادگان KTH توسط بسیاری از الگوریتم‌های تشخیص رفتار انسان برای ارزیابی دقت استفاده شده است [71-73, 77-79, 90].

۲-۷-٤- سایر دادگان‌های مورد استفاده

علاوه بر دادگان‌های مطرح شده، نتایج تحلیل‌ها بر روی تعداد زیادی دنباله ویدئویی با طول زمانی بیش از ٤۰۰۰۰ فریم و نرخ نمونه برداری ۱۵، ۲٤ و ۲۹ فریم در ثانیه آزمون شده‌اند. اغلب ویدئوها از دادگان 'Simon Fraser University Video Library and Tools' [91] ، و دادگان 'Open Video Project' [92] انتخاب شده‌اند. ابعاد اغلب فریم‌های ویدئو QCIF (176x144) یا CIF (352x288) است. این ویدئوها در مکان‌های مختلف فیلمبرداری شده‌اند (محوطه داخل و خارج ساختمان) و دارای مشخصات متنوعی هستند. ویدئوها شامل مسابقات ماشین، پخش اخبار، دویدن حیوانات، پرواز هواپیما، شکستن لیوان و ... هستند. همچنین این مجموعه دارای حالات مختلف زوم دوربین[٦۱]، لایه گذاری[٦۲]، حرکتهای انتقالی دوربین[٦۳] و محوشدگی شات‌ها[٦٤] است. اغلب ویدئوهای آزمون بسیار پویا[٦۵] در هر دو بعد مکان و زمان می‌باشند در حالی که نمونه‌های ایستا[٦٦] نیز وجود دارند.

۲-۸- جمع‌بندی

در این فصل تفاوت‌ها و مزایای روش‌های تحلیل و مدل‌سازی پارامتریک و غیرپارامتریک سیگنال بیان شد و دلایل انتخاب روش تحلیلی پارامتریک ارائه گشت. سپس به بررسی و توضیح روش‌های مطرح در تحلیل پارامتریک سیگنال ویدئو پرداختیم و معایب و مزایای هر روش را شرح دادیم. روش‌های ترکیبی گوسی که سیگنال را سه‌بعدی در نظر گرفته و به ترتیب در حوزه پیکسل و تبدیل موجک به مدل‌سازی سیگنال می‌پردازند، معرفی شدند و مورد بحث قرار گرفتند. مدل AR و HMM نیز به عنوان مدل‌هایی که در صدد بیان ارتباطات زمانی فریم‌های ویدئویی هستند، معرفی شدند. و سپس مروری کلی بر برخی مدل‌های آماری دیگر که برای مدل‌سازی حرکات صورت و بدن انسان به کار می‌روند، صورت گرفت.
در ادامه این فصل مختصری درباره کارهای مرتبط با کاربردهای چکیده‌سازی ویدئو و تشخیص رفتار انسان توضیح داده شد. این دو کاربرد برای ارزیابی روش‌های تحلیل و مدل‌سازی ارائه شده در این رساله به کار رفته‌اند و آزمون‌های ادراکی و تحلیلی انجام شده کارایی بالای روش‌های تحلیلی ارائه شده را تایید می‌کند. همچنین، دادگان‌های ویدئویی که برای ارزیابی‌ها استفاده می‌شوند، معرفی شدند.

[61] Zooming
[62] Padding
[63] Translation
[64] Scene dissolves
[65] Dynamic
[66] Static

فصل سوم

تحلیل تحول زمانی پارامترهای مکانی با کمک معیار فاصله KL

مقدمه

انتخاب پارامترهای مکانی

انتخاب معیار فاصله

الگوریتم ارائه‌شده

نتایج شبیه‌سازی و بحث

جمع‌بندی

۳-۱- مقدمه

در این بخش از رساله به ارائه روشی جدید برای تحلیل و مدل‌سازی ویدئو با کمک پارامترهای توزیع تعمیم‌یافته گوسی[1](GGD) زیرباندهای تبدیل دو بعدی موجک با استفاده از معیار فاصله KL[2] می‌پردازیم. این دیدگاه پارامتریک بر اساس خواص آماری حاشیه‌ای زیرباندهای تبدیل موجک که مطابق خواص سیستم بینایی انسان است، به استخراج پارامترهای مکانی از فریم‌ها می‌پردازد. علاوه بر این معیار فاصله KL به خوبی بیان‌گر تفاوت پارامترهای مذکور است. مدل‌سازی تصویر با تبدیل دو بعدی موجک ابتدا در [93] ارائه شد و نتایج شبیه‌سازی‌ها نشان داد که تابع GGD می‌تواند تقریب مناسبی برای توزیع حاشیه‌ای زیرباندهای این تبدیل برای فیلترهای مختلف موجک و سطوح تبدیل متفاوت باشد [93-95,11,147]. در [11] از پارامترهای GGD در کنار فاصله KL برای بازیابی بافت در تصاویر ساکن استفاده شد، که به نتایج خوبی دست یافتند.

ما با کمک پارامترهای مکانی مستخرج از خواص آماری حاشیه‌ای زیرباندهای تبدیل موجک فریم‌ها و فاصله KL بین این پارامترها و میزان شباهت بین فریم‌ها در یک خوشه ویدئو و تفاوت بین فریم‌های خوشه های مختلف، به تحلیل سیر تحولات زمانی ویدئو می‌پردازیم. از نتایج این تحلیل برای تعیین مرز بین شات-های آنی و تدریجی ویدئو استفاده می‌شود و مرز شات‌ها و خوشه‌های شات‌ها بر اساس ضابطه شباهت و تفاوت انتخاب می‌گردند. ارزیابی‌های ادراکی و تحلیلی موید دقت بالای این روش در مقایسه با روش‌های متداول است.

[1] Generalized Gaussian Density
[2] Kullback-Leibler Distance

۳-۲- انتخاب پارامترهای مکانی

در تحلیل و مدل‌سازی سیگنال ویدئو، انتخاب پارامترهای مکانی یکی از مسائل مهم محسوب می‌شود. برای استخراج پارامترهای مناسب از فریم‌های ویدئو، از خواص آماری حاشیه ای تبدیل موجک فریم‌ها استفاده می‌نماییم. به این ترتیب که ابتدا تبدیل موجک به فریم‌های ویدئویی اعمال می گردد. سپس هیستوگرام‌های حاشیه‌ای زیر باندهای این تبدیل با مدل تعمیم‌یافته گوسی تقریب زده می‌شوند و پارامترهای این مدل به عنوان مشخصات مکانی مستخرج از حوزه موجک در نظر گرفته می‌شوند.

۳-۲-۱- تبدیل موجک در پردازش تصویر

تبدیل موجک به دلیل موفقیت در پیاده‌سازی برای کاربردهای مختلف پردازش سیگنال و خواص تئوری-اش، در سال‌های اخیر مورد توجه زیادی قرار گرفته است[11]. سیگنال x با کمک ترکیب خطی توابع پایه موجک به شکل زیر قابل بسط است:

$$x = \sum_{n=1}^{\infty} C_n \varphi_n, \quad C_n = < x, \varphi_n > \qquad (۱-۳)$$

که $\{\varphi_n\}_{n\epsilon N}$ توابع پایه موجک هستند.

یکی از مزایای بسیار مهم موجک در کاربردهای پردازش تصویر و نیز کاربردهای دیگر پردازش سیگنال، قابلیت بالای آن در نمایش نقاط ناپیوستگی است. موجک‌های با محدوده متناهی این قابلیت را دارند که در نقاط ناپیوستگی سیگنال تمرکز کنند[11]. یعنی حول نقاط ناپیوستگی، ضرایب غیرصفر، و در سایر نقاط، ضرایب صفر دارند. همانطور که گفتیم موجک ناپیوستگی نقطه‌ای را با دقت نمایش می‌دهد. با ضرب پایه‌ها این خصلت به سیگنال‌های دوبعدی هم تعمیم پیدا می‌کند. یعنی موجک دو بعدی در یک تصویر نقاط ناپیوستگی را به خوبی نمایش می‌دهد. قابلیت مهم موجک انطباق نسبی آن با سیستم ادراکی انسان در هر دو حس بینائی و شنوائی است. این ویژگی اولین علت ابداع این تبدیل چندسطحی برای شبیه‌سازی حساسیت فرکانسی بینائی و شنوائی انسان بوده است [96,97].

۳-۲-۲- خواص آماری تبدیل دوبعدی موجک

یک روش مدل‌سازی تصویر بر اساس تبدیل دو بعدی موجک در [11] بر اساس مدل‌سازی آماری ارائه شده است. سیستم بینایی انسان[3] به لبه‌ها و بافت بسیار حساس است. تبدیل موجک به صورت ساختاری با این خصوصیت سیستم بینایی انسان هماهنگ بوده، به خوبی این تغییرات را مد نظر می‌گیرد[96, 97]. به این ترتیب ما پارامترهای استخراج‌شده از این تبدیل را برای ساختن بردارهای ویژگی مکانی ویدئو استفاده

[3] Human visual system (HVS)

کرده‌ایم. در تحقیقات موجود، نشان داده شده است که تابع توزیع تعمیم یافته گوسی به خوبی توزیع ضرایب تبدیل موجک در هر زیرباند را برای انواع فیلترها و تعداد مراتب تبدیل موجک، تخمین می‌زند [96, 93, 11]. تابع GGD به صورت زیر تعریف می‌شود:

$$\mathrm{p}(\mathrm{x};\mu,\alpha,\beta)=\frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})}\mathrm{e}^{-\left(\frac{|\mathrm{x}-\mu|}{\alpha}\right)^{\beta}} \qquad (۲-۳)$$

که $\Gamma(.)$ تابع گاما و α پارامتر مقیاس[4] است که پهنای تابع را تعیین می‌کند و β پارامتر شکل‌دهی[5] است و به صورت معکوس با سرعت نزول تابع توزیع احتمال[6] رابطه دارد و μ میانگین توزیع است که صفر فرض می‌شود. تابع توزیع لاپلاسین است، اگر $\beta=1$ و گوسین است اگر $\beta=2$. به این ترتیب با استخراج این دو پارامتر (α و β) اطلاعات کافی برای تعیین هیستوگرام حاشیه‌ای[7] زیرباند خواهیم داشت.

۳-۳- انتخاب معیار فاصله

مساله مهم دیگر پس از انتخاب پارامترهای مکانی، انتخاب معیار فاصله مناسب برای بیان میزان شباهت یا تفاوت بین پارامترهای مکانی استخراج شده است.

KLD[8] که به عنوان شاخص تباین اطلاعات[9] هم شناخته می‌شود، تفاوت بین دو توزیع احتمال را نشان می‌دهد. این مقیاس اندازه‌گیری نامتقارن و غیرمنفی، بین دو توزیع P و Q به صورت زیر بیان می‌شود:

$$D(P\|\mathrm{Q})=\int_{\mathrm{x}}\mathrm{p}(\mathrm{x})\log\left(\frac{\mathrm{p}(\mathrm{x})}{\mathrm{q}(\mathrm{x})}\right) \qquad (۳-۳)$$

به این ترتیب با کمی محاسبه، فاصله KL بین پارامترهای GGD دو زیر باند معادل تبدیل موجک به صورت زیر در می‌آید [93]:

$$\mathrm{D}(\mathrm{p}(.;\alpha_{\mathrm{i}},\beta_{\mathrm{i}})\,\|\mathrm{p}(.;\alpha_{\mathrm{j}},\beta_{\mathrm{j}}))=\log\left(\frac{\beta_{\mathrm{i}}\alpha_{\mathrm{j}}\Gamma\left(\frac{1}{\beta_{\mathrm{j}}}\right)}{\beta_{\mathrm{j}}\alpha_{\mathrm{i}}\Gamma\left(\frac{1}{\beta_{\mathrm{i}}}\right)}\right)+\left(\frac{\alpha_{\mathrm{i}}}{\alpha_{\mathrm{j}}}\right)^{\beta_{\mathrm{i}}}\frac{\Gamma\left(\frac{\beta_{\mathrm{j}}+1}{\beta_{\mathrm{i}}}\right)}{\Gamma\left(\frac{1}{\beta_{\mathrm{i}}}\right)}-\frac{1}{\beta_{\mathrm{i}}} \qquad (٤-۳)$$

در نتیجه، با فرض واقع‌گرایانه[10] استقلال زیرباندهای متفاوت همسطح تبدیل، فاصله بین دو بردار ویژگی GGD از دو فریم ویدئو برابر با مجموع فاصله‌های KL بین دوبه‌دوی زیرباندهای معادل خواهد بود:

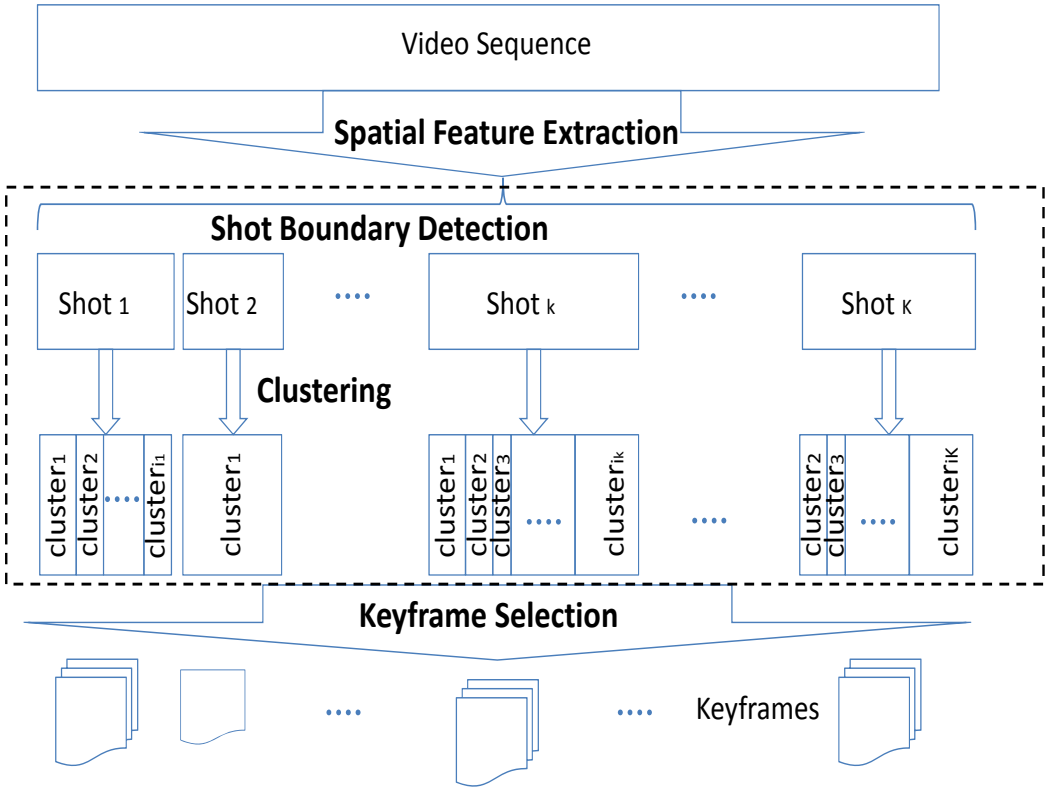
$$D\left(Im_i,Im_j\right)=\sum_{l=1}^{3L}\mathrm{D}(\mathrm{p}(.;\alpha_{\mathrm{i}}^{(l)},\beta_{\mathrm{i}}^{(l)})\,\|\mathrm{p}(.;\alpha_{\mathrm{j}}^{(l)},\beta_{\mathrm{j}}^{(l)})) \qquad (۵-۳)$$

[4] Scale parameter
[5] Shape parameter
[6] Probability distribution function (PDF)
[7] Marginal
[8] Kullback Leibler Distance
[9] Information divergence
[10] Realistic

که L بیان‌گر تعداد سطوح تبدیل موجک است و در هر سطح تبدیل سه زیر باند جهتی افقی، عمودی و قطری مدنظر هستند.

۳-٤- الگوریتم ارائه‌شده

یکی از مهم‌ترین معایب دیدگاه‌های موجود در تعیین فریم‌های کلیدی این است که این روش‌ها فقط از فاصله بین فریم‌های مجاور ویدئو استفاده می‌کنند و به طور سیستماتیک از تفاوت بین فریم‌های کل خوشه ویدئو چشم‌پوشی می‌نمایند. این در حالی است که تحلیل بر مبنای خوشه می‌تواند به شباهت/تفاوت های بین فریم‌های ویدئو، در مقایسه با روش‌های بر مبنای معیارهای اندازه‌گیری بین فریم‌ها – که در برابر تغییرات تدریجی در فریم‌های متوالی بدون تغییر قابل ملاحظه‌ای می‌مانند- حساس‌تر است. این اولین انگیزه مهم برای ارائه روشی برای تعیین فریم‌های کلیدی با تحلیل بر مبنای خوشه است.



شکل ۳-۱. شماتیک الگوریتم ارائه‌شده جهت انتخاب فریم‌های کلیدی.

برای درک دیدگاه تحلیل مبتنی بر خوشه، نیاز به بررسی محتوای بصری دنباله ویدئویی است. برای این کار انتخاب ویژگی‌های مناسب از هر فریم ویدئو یک فاکتور مناسب در روند تحلیل به شمار می‌رود. این ویژگی‌ها می‌توانند اطلاعات حرکتی [85,86]، هیستوگرام رنگ[82,17,18] یا ویژگی‌های استخراج‌شده از تبدیل دوبعدی فریم‌های ویدئوئی [87] باشند، که می‌توانند در روش‌های تعیین فریم‌های کلیدی استفاده شوند.

یکی از منابع اصلی ایجاد خطا در روش‌های موجود استخراج ویژگی، عدم تطابق ویژگی‌ها با ساختار سیستم بینایی انسان است. این دومین مزیت روش ارائه‌شده است که از پارامترهای مستخرج از تبدیل دوبعدی موجک هر فریم، که تطابق خوبی با سیستم HVS دارد، استفاده می‌کند. سومین مزیت روش ارائه-شده، استفاده از معیار فاصله متناسب با ویژگی‌های استخراجی است.

همانطور که قبلاً نیز بیان شد، فریم کلیدی با فریم‌های موجود در خوشه خود مشابه و از دیگر فریم‌های دنباله ویدئویی متفاوت است. به این ترتیب ما هر دو شرط شباهت و تفاوت را مد نظر قرار داده‌ایم و از معیار فاصله KL برای این منظور استفاده کرده‌ایم. الگوریتم ارائه‌شده برای انتخاب فریم کلیدی دارای سه مرحله است که در شکل ۳-۱ نمایش داده شده‌اند. در ابتدا پارامترهای مکانی از هر فریم استخراج می‌شوند و ماتریس ویژگی F تشکیل می‌گردد. سپس مرزهای شات‌ها و خوشه‌ها تعیین می‌گردند. در پایان، از هر خوشه یک فریم بر اساس معیار شباهت و تفاوت اعمال‌شده انتخاب می‌شود. در زیربخش‌های بعدی، هر مرحله با جزئیات توضیح داده می‌شود.

۳-٤-۱- مقدمات

فرض کنید دنباله ویدئویی ، V ، شامل K شات است و شات k ام دارای i_k خوشه است، بنابراین روابط زیر را خواهیم داشت:

$$V = \bigcup_{k=1}^{K} S_k,$$
$$S_k = \bigcup_{j=1}^{i_k} C_{k,j},$$
$$C_{k,j} = \bigcup_{n=1+b_{k,j}}^{b_{k,j}+n_{k,j}} f_n,$$
$$b_{k,j} = \sum_{k_1=1}^{k-1} \sum_{j_1=1}^{i_{k_1}} n_{k_1,j_1} + \sum_{j_1=1}^{j-1} n_{k,j_1} \qquad (٦-٣)$$

که S_k و f_n به ترتیب نشان‌گر k امین شات و n امین فریم هستند. $C_{k,j}$ نشان‌دهنده j امین خوشه از k امین شات است و $n_{k,j}$ فریم دارد. $b_{k,j}$ نیز تعداد فریم‌های قبل از شروع خوشه جاری است. هدف استخراج یک فریم کلیدی از هر خوشه است.

۳-٤-۲- استخراج ویژگی

فرض کنید که دنباله ویدئویی دارای N فریم است و $2P$ پارامتر GGD از زیرباندهای تبدیل موجک هر فریم طبق معادله (۳-۷) استخراج شده است، که $P = 3L$ و L تعداد سطوح تبدیل موجک است. به این ترتیب ماتریس ویژگی F تشکیل می‌شود، که ستون‌هایش بردارهای ویژگی استخراج‌شده از فریم‌ها و ابعادش برابر $2P \times N$ است.

$$F = [fv_1\ fv_2\ \dots fv_n\ fv_N],\quad 1 \le n \le N$$
$$fv_n=[\alpha_{n,1}\ \beta_{n,1}\ \alpha_{n,2}\ \beta_{n,2}\ \dots\ \alpha_{n,p}\ \beta_{n,p}\ \dots\ \alpha_{n,P}\ \beta_{n,P}]^T,\ 1 \le p \le P \qquad (۳-۷)$$

۳-٤-۳- تعیین مرز شات و خوشه‌بندی

در این مرحله تمام دنباله ویدئویی به K شات غیرهمپوشان تقسیم می‌شود. فاصله KL بین دو فریم ویدئوئی همجوار طبق فرمول (۳-٤) محاسبه و یک بردار KL تشکیل می‌شود. منحنی KL کشیده شده، مرزهای تدریجی[11] و آنی[12] بر اساس روش سطح آستانه به صورت تجربی تعیین می‌شوند. سطح آستانه برای مرز شات برابر $T_s = \rho m_w$ انتخاب می‌گردد که m_w میانگین محلی فاصله KL روی پنجره‌ای به طول w فریم است که بین ٤ تا ۱۰ فریم در نظر گرفته می‌شود. و ρ به صورت تجربی، عددی بین ۲ تا ۳ انتخاب می‌گردد.

در مرحله بعد، هر شات به یک یا چند خوشه تقسیم می‌شود. منحنی KL هر شات هموار[13] می‌گردد و ماکزیمم‌های محلی بردار KLD هموار شده با مقادیری بالاتر از سطح آستانه تعیین و به عنوان مرز خوشه‌ها انتخاب می‌گردند. در اینجا سطح آستانه برابر $T_c = \rho m_{w+4}$ در نظر گرفته شده است.

۳-٤-٤- انتخاب فریم کلیدی

در این مرحله میانگین فاصله KL بین فریم n ام از هر خوشه و فریم‌های دیگر خوشه با استفاده از فرمول (۳-٤) محاسبه می‌شود:

[11] Gradual boundary
[12] Abrupt boundary
[13] Smooth

$$m_{dist}(n, C_{k,j}) = \frac{1}{n_{k,j}-1}\sum_{\substack{i\neq n \\ i\subset C_{k,j}}} D(fv_n, fv_i)$$
$$= \frac{1}{n_{k,j}-1}\sum_{\substack{i=1+b_{k,j} \\ i\neq n}}^{b_{k,j}+n_{k,j}} D(fv_n, fv_i) \qquad (۸-۳)$$

این مقدار نشان‌دهنده میانگین فاصله بین فریم n و فریم‌های دیگر خوشه است. از معادله (۳-۷) مشخص است که $m_{dist}(n, C_{k,j})$ با معیار شباهت رابطه عکس دارد و با افزایش این فاصله شباهت فریم با فریم‌های دیگر کاهش می‌یابد.

می‌توان میانگین فاصله فریم n در خوشه $C_{k,j}$ و فریم‌های دیگر از خوشه‌های دیگر را با کمک معادله (۳-٤) به شکل زیر نوشت:

$$m_{dist}(n, \bar{C}_{k,j}) = \frac{1}{N-n_{k,j}}\sum_{i\not\subset C_{k,j}} D(fv_n, fv_i) =$$
$$\frac{1}{N-n_{k,j}}\left(\sum_{i=1}^{b_{k,j}} D(fv_n, fv_i) + \sum_{i=1+b_{k,j}+n_{k,j}}^{N} D(fv_n, fv_i)\right) \qquad (۹-۳)$$

که $\bar{C}_{k,j}$ قسمت مکمل $C_{k,j}$ از دنباله ویدئو است. $m_{dist}(n, \bar{C}_{k,j})$ میانگین فاصله بین هر فریم از خوشه با فریم‌های بیرون خوشه را نشان می‌دهد.

فریم کلیدی باید از فریم‌های بیرون خوشه تا حد ممکن متفاوت باشد، در نتیجه $m_{dist}(n, \bar{C}_{k,j})$ رابطه مستقیم با معیار تفاوت دارد و افزایش $m_{dist}(n, \bar{C}_{k,j})$ باعث افزایش فاکتور تفاوت می‌شود. معیار نهایی با تقسیم یا تفریق معادلات (۳-۷) و (۳-۸) به دست می‌آید.

$$F_{k,j}(n) = \frac{m_{dist}(n, \bar{C}_{k,j})^x}{m_{dist}(n, C_{k,j})^y},$$

$$F_{k,j}(n) = x.m_{dist}(n, \bar{C}_{k,j}) - y.m_{dist}(n, C_{k,j}) \qquad (۱۰-۳)$$

x و y برای تطبیق معیارهای تشابه[14] و تفاوت[15] در الگوریتم انتخاب فریم کلیدی هستند که در اینجا برابر یک در نظر گرفته شده‌اند. به این ترتیب فریم کلیدی هر خوشه با معیار زیر انتخاب می‌شود:

[14] Similarity
[15] Dissimilarity

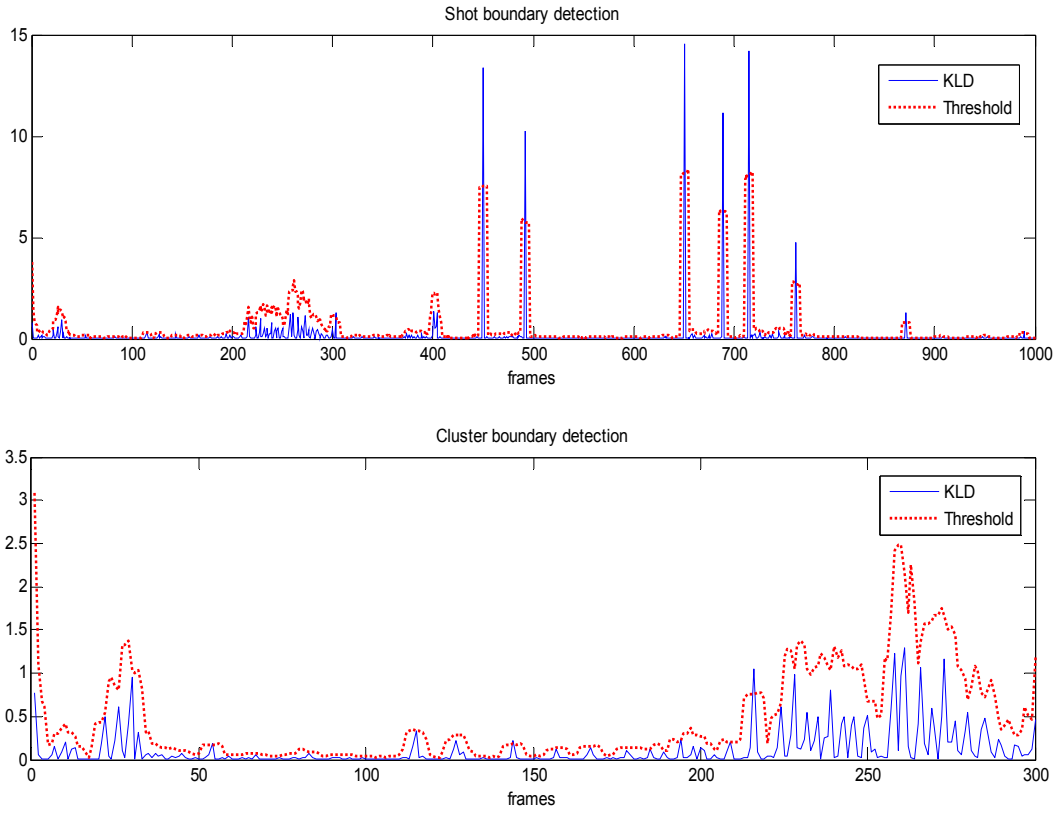
$$keyf(k,j) = argmax_{n \subset C_{k,j}} F_{k,j}(n) \qquad (۱۱-۳)$$

۳-۵- نتایج شبیه‌سازی و بحث

تعداد زیادی دنباله ویدئویی شامل بیش از ۲۰ ساعت ویدئو با خصوصیات متفاوت برای ارزیابی الگوریتم ارائه‌شده به کار رفته است. ویدئوهای آزمون ما از میان دادگان صحنه‌های Hollywood2 [89] ، دادگان تشخیص مرز شات TRECVID [88]، کتابخانه ویدئویی 'Simon Fraser University' [91]، دادگان 'The Open Video Project' [92] و مجموعه‌های دیگر انتخاب شده است. ابعاد مکانی فریم‌های ویدئویی متفاوت بوده و شامل ابعاد QCIF، CIF و 16 CIF هستند. مجموعه‌های ویدئویی بالا در بسیاری از تحقیقات استفاده شده‌اند و دنباله‌های ویدئویی از محیط‌های مختلف –داخلی و خارجی- با خاصیت‌های مختلف فیلمبرداری شده‌اند.

ما تبدیل دوبعدی موجک را بر تمام فریم‌های ویدئو اعمال نموده، پارامترهای GGD را از زیرباندهای تبدیل استخراج می‌نماییم و ماتریس ویژگی F را تشکیل می‌دهیم. دنباله ویدئویی با کمک منحنی فاصله KL بین فریم‌های مجاور و بکارگیری روش سطح آستانه بر اساس میانگین فاصله‌های KL بین فریم‌های همسایه، به شات‌ها تقسیم می‌شوند. سپس، با کمک روش سطح آستانه دیگر، هر شات با توجه به ماکزیمم-های محلی منحنی فاصله KL هموار شده، به یک یا چند خوشه تقسیم می‌گردد. در شکل ۳-۲ مثالی برای این روند نمایش داده شده است. نمودار بالایی روش تعیین مرز شات را نشان می‌دهد که در آن منحنی ممتد منحنی فاصله KL و منحنی منقطع نشان‌دهنده مقدار سطح آستانه در هر فریم است. نمودار پایینی نیز مرحله خوشه‌بندی را نشان می‌دهد که منحنی ممتد منحنی فاصله KL و منحنی منقطع سطح آستانه را نمایش می-دهد.

برای ارزیابی الگوریتم تعیین مرز شات ارائه‌شده، ما نتایج الگوریتم را بر روی دو دادگان ویدئویی مشهور TRECVID 2006 و Hollywood2 ارزیابی کردیم. TRECVID 2006 دارای ۱۳ دنباله ویدئویی طولانی اخبار با ۵۹۷۰۴۳ فریم می‌باشد. اینجا هدف تعیین مرز شات‌ها و نوع آنها (تدریجی یا آنی) است. در ابتدا مرز شات‌ها بر اساس روش سطح آستانه ارائه‌شده تعیین می‌شود و سپس شات‌های نوع برش از میان مرز شات‌ها انتخاب می‌شوند.

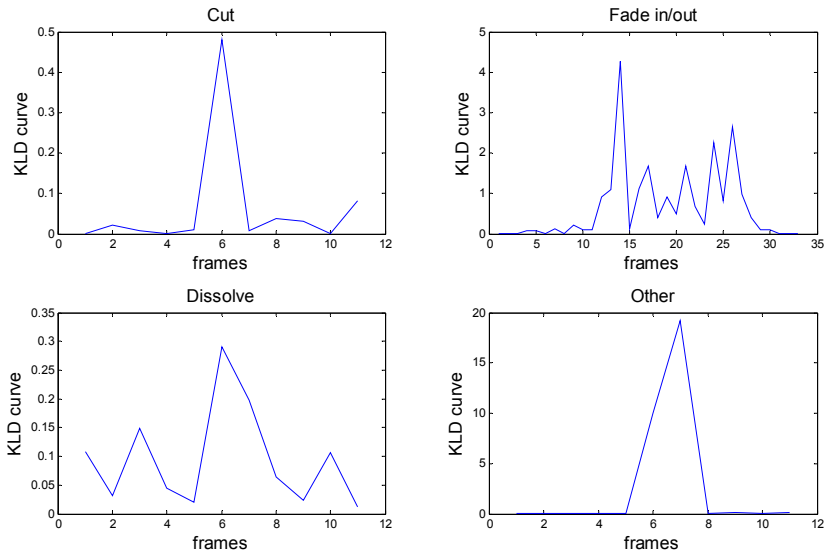


شکل ۳-۲. انتخاب مرز شات و خوشه‌ها برای نمونه ویدئوی 'sceneclipautoautotrain00060.avi' از مجموعه Hollywood2. مرزهای شات انتخابی فریم‌های: 989 871 762 715 689 651 492 450 303 (شکل بالا) و مرزهای خوشه‌ها فریم‌های: 209 194 144 115 54 (شکل پایین).

انتظار ما داشتن یک بیشینه در منحنی فاصله KL در مرز شات‌های برش است به طوری که منحنی سریعاً پس از فریم مرز شات به سمت صفر کاهش یابد. فرض کنید که فریم i به عنوان مرز شات انتخاب شود. نوع تغییر شات "آنی" است، اگر دو شرط زیر همزمان برقرار باشند:

$$|D(fv_{i+2}, fv_{i+1}) - D(fv_{i+1}, fv_i)| < t_p$$
$$|D(fv_{i-1}, fv_{i-2}) - D(fv_{i-3}, fv_{i-2})| < t_p \qquad (۱۲-۳)$$

که سطح آستانه t_p بین ۰,۱ و ۰,۳ به صورت تجربی انتخاب می‌شود. شکل ۳-۳ چند نمونه منحنی فاصله KL را در مرز شات‌ها نمایش می‌دهد. برای یافتن فریم‌های شروع و پایان تغییرات شات تدریجی، نقاط آغاز و پایان صعود و نزول منحنی فاصله KL حول نقطه تغییر شات تعیین می‌شوند.



شکل ۳-۳. منحنی فاصله KL حول فریم‌های تغییر شات برای انواع تغییر شات.

کارایی الگوریتم با محاسبه معیارهای دقت[16] و بازخوانی[17] ارزیابی می‌شود. آزمون‌های تعیین مرز شات و نوع آنها با الگوریتم بیان شده با اعمال فیلترهای موجک متفاوت –Haar، Daubechies و Symlets – و تعداد سطوح تبدیل مختلف –۳، ٤ و ٥ سطح برای فیلتر Daubechies– تغییر پارامترهای ترشلد ρ– و w در انتخاب مرز شات $T_s = \rho m_w$ و t_p در تعیین نوع تغییر شات انجام شده‌اند و فاکتورهای بازخوانی و دقت برای تمامی آزمون‌ها محاسبه می‌شوند و منحنی‌های دقت نسبت به بازخوانی رسم می‌گردند. به این ترتیب روش ارائه‌شده با روش‌های مختلف انتخاب مرز شات در دادگان TRECVID2006 در شکل ۳-٤ مقایسه می‌شوند. همانطور که در منحنی‌ها دیده می‌شود، روش ارائه‌شده به خوبی مرز شات‌های تدریجی و آنی را شناسایی می‌کند.

کارایی بالای روش ارائه‌شده در تعیین مرز شات‌های تدریجی قابل توجه است. به این دلیل که اغلب الگوریتم‌های دیگر همواره حساس به دقت هستند، زیرا در دنباله‌های ویدئویی طبیعی تعداد زیادی رخداد بصری وجود دارد (مانند حرکت اشیاء و دوربین) که می‌توانند با تغییر تدریجی شات‌ها اشتباه گرفته شوند [98-100]. ولی روش پیشنهادی در برابر تغییرات پارامترهای سیستم مقاوم است (شکل ۳-٤.ج) طوری که با تغییر نرخ بازخوانی از ۰,۳ تا ۱، دقت سیستم همواره در حدود ۰,۹ باقی می‌ماند. این مساله نشان‌دهنده

[16] Precision
[17] Recall

کارایی بالای مدل پیشنهادی برای جلوگیری از انتخاب نابجا[18]ی این نوع تغییرات شات است.

اشکال در تعیین نوع تغییر شات در شات‌های با تغییرات تدریجی با طول کمتر از ٤ فریم است که به اشتباه تغییر آنی تشخیص داده می‌شوند. در حالات دیگر، الگوریتم پیشنهادی مرز شات‌ها را با دقت بالا تشخیص می‌دهد. با افزایش تعداد سطوح تبدیل موجک، جزئیات بیشتری از فریم‌ها استخراج می‌شود و نتایج بهبود می‌یابد.

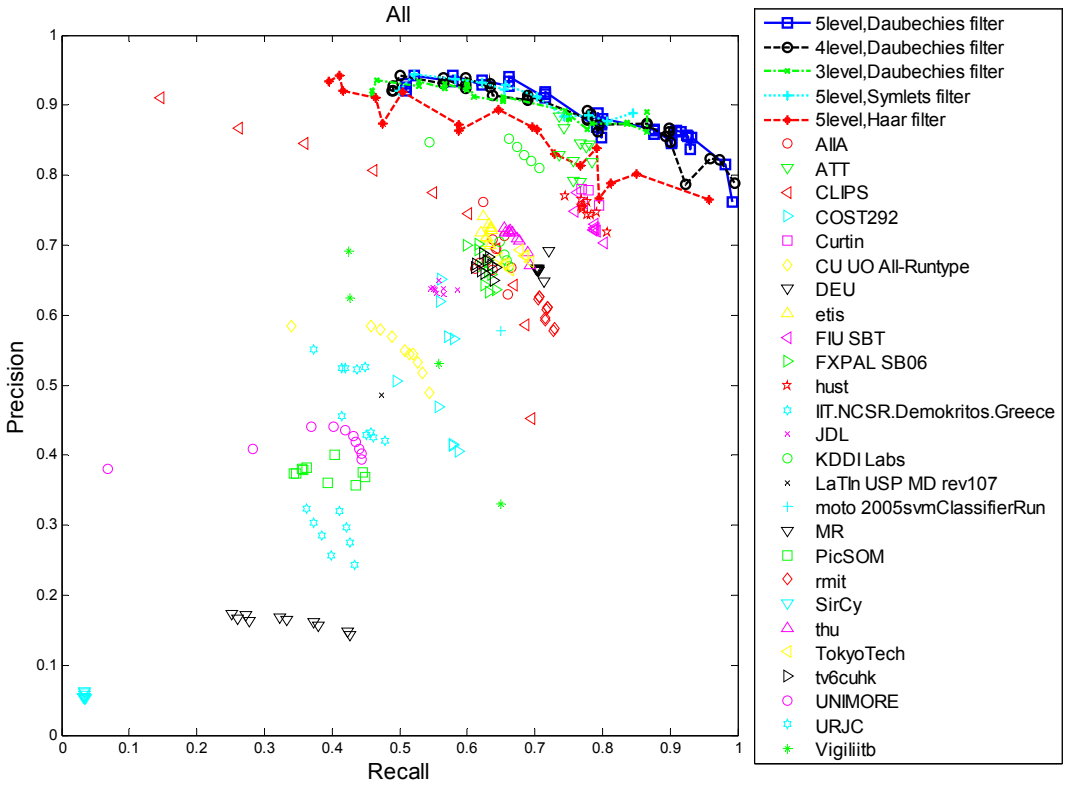
دادگان دیگری که برای ارزیابی نتایج استفاده شده است، دادگان Hollywood2 است که شامل ۱۱۵۲ دنباله ویدئویی است که از بیش از ٦۹ فیلم ویدئویی انتخاب شده‌اند. این مجموعه شامل ۱۰۲۵۲۷۸ فریم ویدئویی و ۸۱۹۹ تغییر شات (آنی و تدریجی) است. شکل ۳-۵ مثالی از منحنی فاصله KL بین فریم‌های مشابه را برای روش ما (بالا) و منحنی خطای تخمین AR برای روش [17] (پایین) نمایش می‌دهد. ویدئوی آزمون ویدئوی 'sceneclipautoautotrain00077.avi' است که از مجموعه Hollywood2 انتخاب شده است. نمودارها نشان می‌دهند که دقت روش پیشنهادی بسیار بالاست، در حالی که مرز شات‌ها در فریم‌های ۳۵۸، ۳۹۰ و ٤۲٤ در روش [17] تشخیص داده نشده‌اند؛ زیرا این روش فقط از ویژگی‌های رنگ به عنوان ویژگی‌های مکانی استفاده می‌کند. در صورتی که این ویژگی‌ها وقتی شات‌ها دارای بافت رنگی یکسان هستند، مرز شات را تشخیص نمی‌دهند. دو مثال از تغییر فریم‌ها در مرز شات در شکل ۳-٦ آمده است که روش [17] تغییر شات را در مثال دوم تشخیص نداده است ولی روش ارائه شده، به خوبی این تغییرات را تعیین می‌کند.

منحنی تغییرات دقت نسبت به بازخوانی برای تعیین مرز شات با تغییر پارامترهای سیستم، ρ و w، و برای فیلترهای مختلف و سطوح متفاوت تبدیل موجک محاسبه و نمایش داده شده است. شکل ۳-۷ این منحنی‌ها را نمایش می‌دهد. همچنین این منحنی‌ها با منحنی تغییر دقت بر حسب بازخوانی با روش [61]، با تغییر مرتبه مدل AR، در این شکل مقایسه شده است. نتایج نشان‌دهنده بهبودی است که با کمک روش پیشنهادی به دست می‌آید.

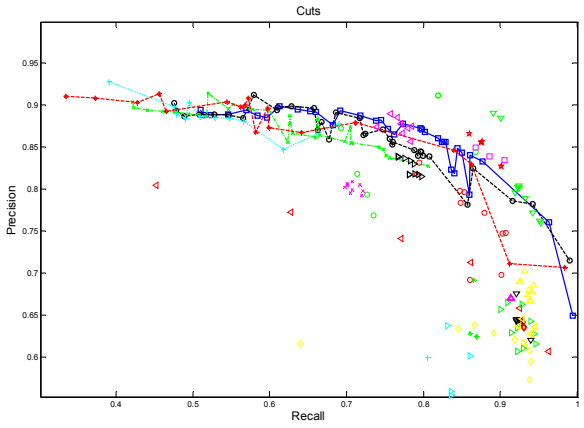
ایراد روش ارائه‌شده در هنگام اعمال آن به ویدئوهای با نرخ فشرده‌سازی بالا ظاهر می‌شود که در آنها اثر بلوکه‌شدن[19] مشهود است. دلیل این مساله نیز، اطلاعات نادرستی است که از پارامترهای تبدیل موجک در مورد لبه‌های تصویر به دست می‌آید. برای مثال ویدئوی 'sceneclipautoautotrain00060.avi' از دادگان Hollywood2 را می‌توان نام برد. در این مواقع استفاده از ویژگی‌های رنگ در کنار ویژگی‌های

[18] False detection
[19] Blocking effect

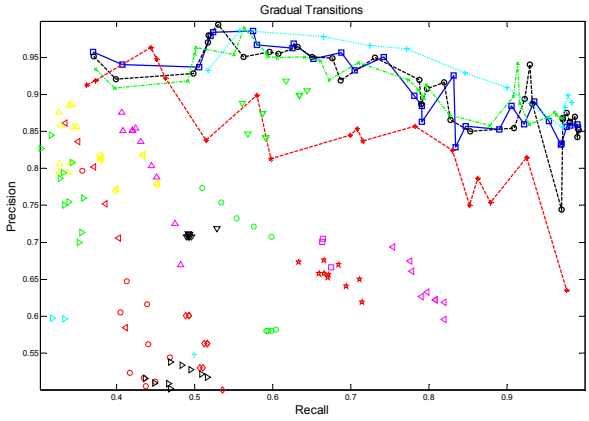
تبدیل موجک می‌تواند مفید باشد.



الف) هر دو نوع تغییر شات

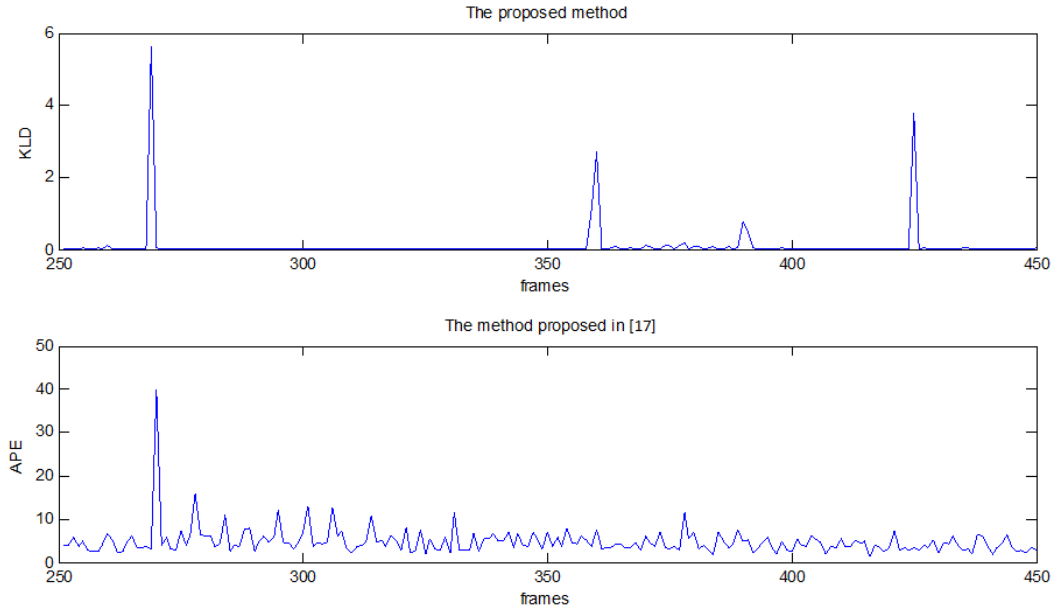


ب) مرز شات‌های آنی

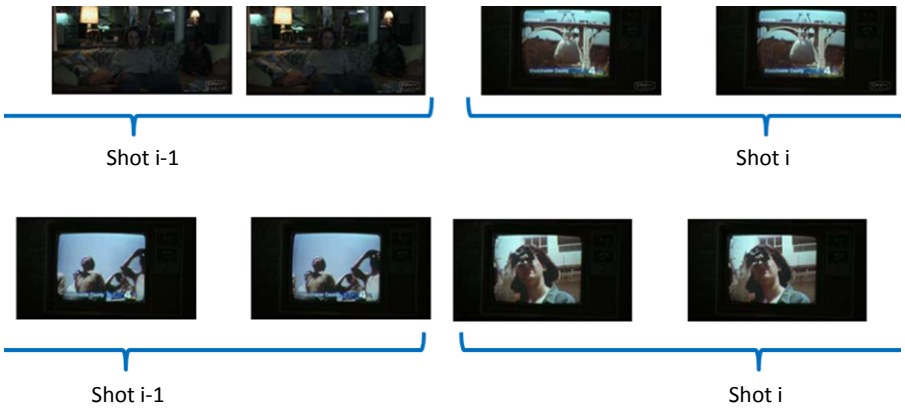


ج) مرز شات‌های تدریجی

شکل ۳-٤. منحنی‌های دقت برحسب بازخوانی برای الگوریتم‌های تعیین مرز شات دادگان TRECVID 2006.

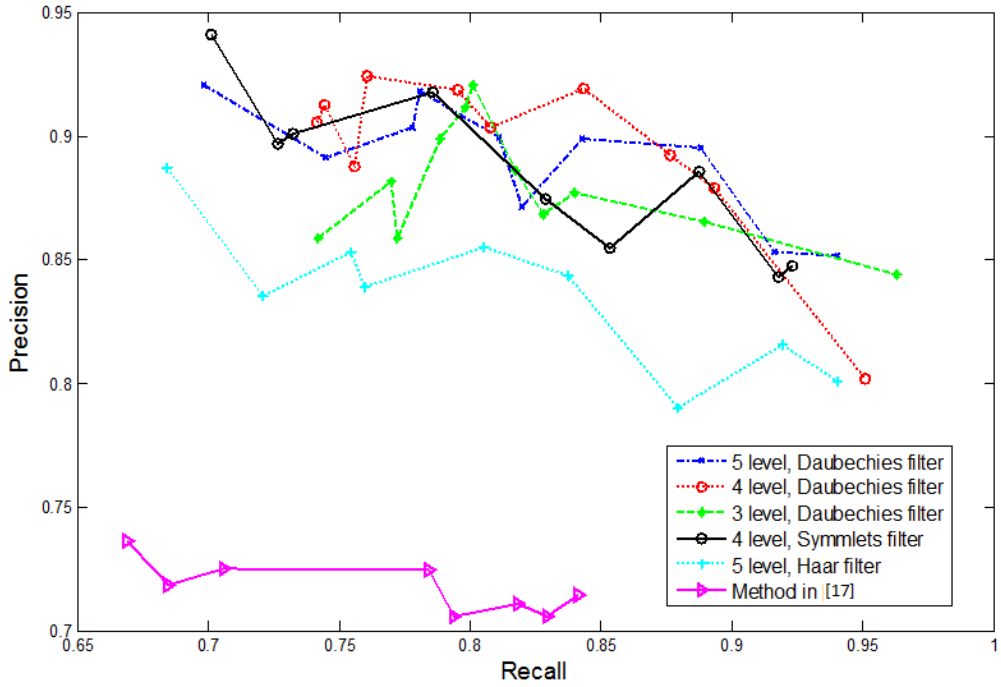


شکل ۳-۵. مرحله پیش‌پردازش. مرزشات‌ها در فریم‌های ۲٦۸، ۳۵۸، ۳۹۰ و ٤۲٤ هستند، منحنی تعیین مرز شات با روش ارائه‌شده (نمودار بالا) و روش در [17] (منحنی پایین).

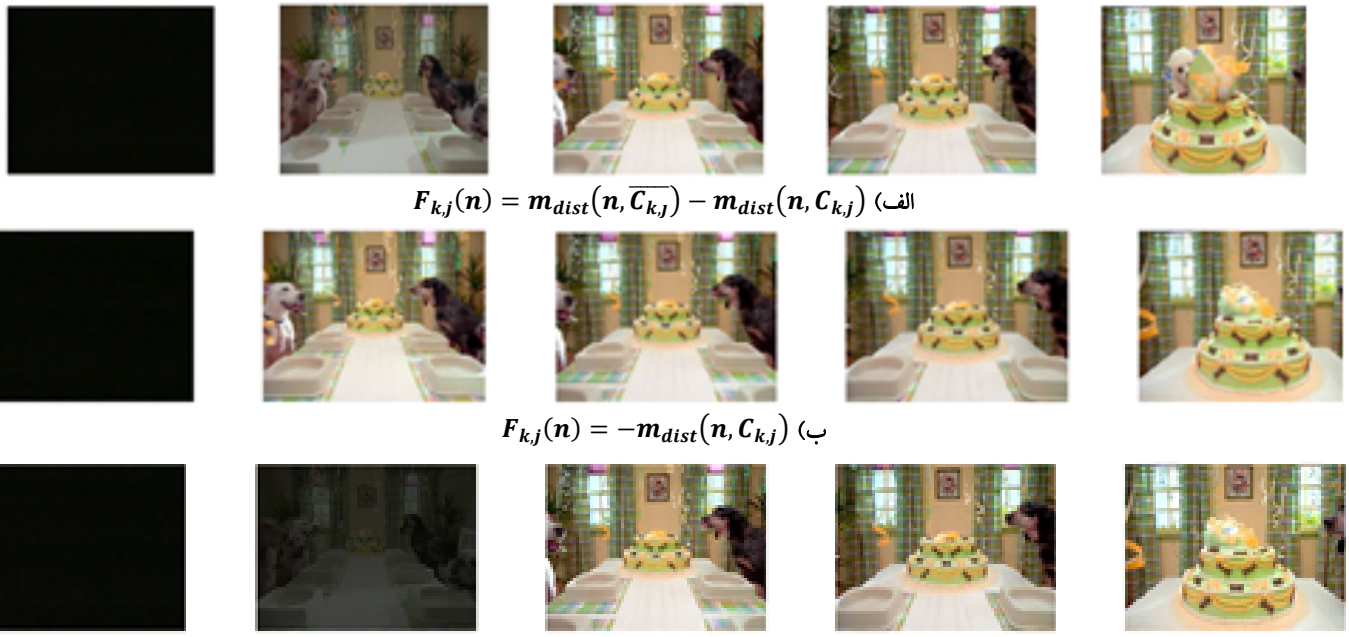


شکل ۳-٦ دو نمونه از مرز شات.

در مرحله بعدی، با استفاده از یک سطح آستانه، هر شات به یک یا چند خوشه تقسیم می‌شود. سپس فریم‌های کلیدی از دنباله ویدئویی با کمک فرمول (۳-۱۰) انتخاب می‌شوند، طوری که از هر خوشه یک فریم کلیدی داشته باشیم. ما الگوریتم تعیین و انتخاب فریم‌های کلیدی را برای تمام ویدئوهای دادگان اعمال و فریم‌های کلیدی را استخراج نموده‌ایم. شکل ۳-۸ مثالی از نتایج الگوریتم را نمایش می‌دهد. سه معیار انتخاب فریم‌های کلیدی به قطعات ویدئویی اعمال شده‌اند. همانطور که در شکل دیده می‌شود، برخی از فریم‌های کلیدی استخراج‌شده در حالتی که معیار تفاوت برای تعیین فریم کلیدی در نظر گرفته نشده است، شبیه یکدیگر هستند (مثلاً، فریم‌های دوم و سوم در شکل ۳-۸.ب). در حالی‌که دو معیار انتخاب دیگر فریم-های کلیدی متمایزتری انتخاب کرده‌اند. همچنین، مقایسه نتایج معیار تفریقی (شکل ۳-۸.الف) با معیار کسری (شکل ۳-۸.ج) نشان می‌دهد که معیار تفریقی نتایج بهتری دارد.



شکل ۳-۷ منحنی‌های دقت بر حسب بازخوانی برای تعیین مرز شات در دادگان Hollywod2.



الف) $F_{k,j}(n) = m_{dist}(n, \overline{C_{k,j}}) - m_{dist}(n, C_{k,j})$

ب) $F_{k,j}(n) = -m_{dist}(n, C_{k,j})$

ج) $F_{k,j}(n) = \frac{m_{dist}(n,\overline{C_{k,j}})}{m_{dist}(n,C_{k,j})}$

شکل ۳-۸ نتایج استخراج فریم کلیدی برای ویدئویی با ۵ شات و ۹ خوشه. ویژگی‌های GGD برای تبدیل موجک با ٤ سطح و معیار فاصله KL.

در مرحله انتخاب فریم کلیدی، الگوریتم ارائه‌شده، از روش انتخاب برحسب هیستوگرام رنگ بهتر عمل می‌کند. زیرا پارامترهای مکانی منتخب با این روش توانایی تشخیص جزئیات بیشتری را در مقایسه با

پارامترهای رنگ دارند. در شکل ۳-۹ یک مثال آمده است که درآن روش هیستوگرام رنگ، فریم‌های ناصحیح را استخراج کرده است، در حالی که روش ما فریم‌های مناسب‌تری را به عنوان فریم‌های کلیدی انتخاب می‌کند چون توانایی مکان‌یابی لبه‌ها و بافت‌ها را در تصویر دارد و جزئیاتی را که برای درک توسط سیستم بینایی انسان مهم‌تر از رنگ است به کار می‌برد. در شکل ۳-۹ج، که نتایج استخراج فریم کلیدی روش [17] آمده است، برخی از فریم‌های دنباله ویدئویی نماینده‌ای در دسته فریم‌های کلیدی ندارند، مثلاً فریم‌های ردیف ٤ از شکل ۳-۹الف ولی در حالت بکارگیری ویژگی‌های موجک، انتخاب دقیق‌تری برای فریم‌های کلیدی انجام گرفته است، زیرا فضای موجک تقریب دقیق‌تری از HVS ارائه می‌دهد.

الف) دنباله ویدئویی نمونه‌برداری شده زمانی.

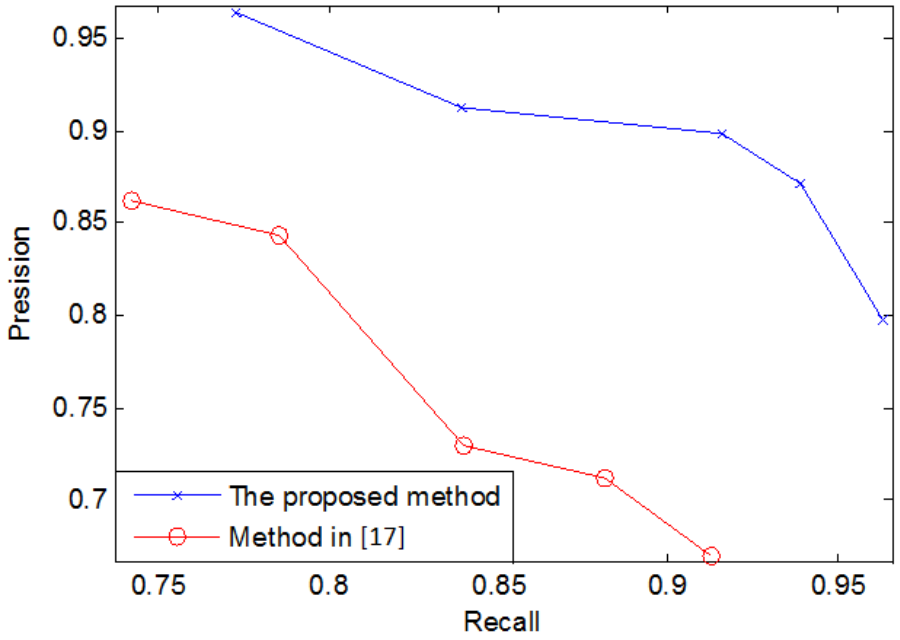
ب) روش پیشنهادی.

ج) روش ارائه‌شده در [17].

ج) روش ارائه‌شده در [87].

شکل ۳-۹. مقایسه روش‌های تعیین فریم کلیدی، ویژگی‌های GGD استخراج‌شده از زیرباندهای تبدیل دوبعدی موجک ٤ سطحی با فیلتر 'Daubechies' و معیار KL.

علاوه بر این، ما کارایی پارامترهای موجک را با پارامترهای DCT بکار رفته در [87] مقایسه کردیم و در شکل ۳-۹د مثالی از نتایج آن آمده است. سیستم بینایی انسان به لبه‌ها و بافت تصویر حساس است که این اطلاعات نمی‌توانند با ضرایب فرکانس پایین تبدیل DCT استنتاج گردند، در حالی که تبدیل موجک این تغییرات را بخوبی دنبال می‌کند. همانطور که در شکل دیده می‌شود، با روش [87] هیچ فریم کلیدی از میان فریم‌های ردیف چهارم شکل ۳-۹الف انتخاب نشده است.



شکل ۳-۱۰ ارزیابی ادراکی روش پیشنهادی. تعداد فریم‌ها : ۶۷۲۰۰، تعداد فریم‌های کلیدی استخراج شده: ۵۸۹.

از آنجا که هیچ روش ارزیابی استانداردی بدون حضور انسان[20] برای نتایج استخراج فریم کلیدی وجود ندارد، علاوه بر مقایسات بالا، یک آزمون ادراکی برای ارزیابی روش ارائه‌شده استفاده کرده‌ایم [101]. از ۱۱ فرد که اطلاعاتی درباره روش‌های پردازش ویدئو داشتند، برای امتیازدهی به نتایج الگوریتم استخراج فریم-های کلیدی با روش ارائه‌شده و روش [17] استفاده شده است. این افراد فریم‌های کلیدی انتخاب‌شده را به دو رده صحیح و ناصحیح امتیازدهی می‌کنند. نمودار دقت برحسب بازخوانی در شکل ۳-۱۰ برای نتایج آزمون‌های متفاوت اعمال‌شده با پارامترهای سیستم متفاوت آورده شده‌اند. نتایج آزمون‌های ادراکی نشان-دهنده کارایی روش ارائه‌شده است.

[20] Human-in-the-loop

۳-٦- جمع‌بندی

در این فصل از رساله به ارائه روشی برای مدل‌سازی آماری ویدئو پرداخته‌ایم و از نتایج این مدل‌سازی برای تشخیص مرز شات‌های آنی و تدریجی و استخراج فریم‌های کلیدی استفاده کرده‌ایم. با استفاده از پارامترهای آماری حاشیه‌ای زیرباندهای تبدیل موجک، پارامترهای مکانی ویدئو از فریم‌های متوالی آن انتخاب می‌شوند و روابط بین این پارامترهای مکانی با کمک معیار فاصله مناسب KL استنتاج می‌گردد. پارامترهای آماری مستخرج از فریم‌ها، برای تشکیل بردارهای ویژگی استفاده می‌شوند و KLD بیان‌گر فاصله بین این بردارهای ویژگی است. در مرحله اول فاصله‌های KL بین فریم‌های مجاور برای تعیین مرز شات‌ها و خوشه‌بندی آنها استفاده می‌گردند. سپس فریم‌های کلیدی بر اساس معیارهای تشابه و تفاوت تعریف‌شده بین فریم‌های داخل و خارج خوشه مربوطه، انتخاب می‌شوند. نتایج آزمون‌ها موید پیشرفت‌های به دست آمده با روش ارائه‌شده، در مقایسه با روش‌های موجود می‌باشد. این پیشرفت به دلیل انطباق ساختار تبدیل دوبعدی موجک با خصوصیات سیستم بینایی انسان است. علاوه بر این، حساسیت بالای معیار فاصله KL به فضای ویژگی انتخاب شده است. این روش تحلیل تحولات زمانی پارامترها با کمک معیار فاصله مناسب می‌تواند در کاربردهای دیگر پردازش ویدئو نظیر بازیابی ویدئو، خلاصه‌سازی ویدئو بر اساس محتوا و ویرایش ویدئو استفاده شود.

فصل چهارم

تجزیه زمانی پارامترهای مکانی

مقدمه
تقریب سیر تحول زمانی سیگنال ویدئو
کاربرد تجزیه زمانی در چکیده‌سازی ویدئو
الگوریتم تعیین فریم کلیدی
نتایج آزمون‌ها
جمع‌بندی

در این فصل روشی برای تحلیل سیگنال ویدئو ارائه می‌شود که بر پایه رخدادهای زمانی- مکانی[1] کار می‌کند. سیگنال ویدئو می‌تواند به عنوان مجموعه‌ای از مولفه‌های بصری مستقل همپوشان، رخداد، در نظر گرفته شود. رخدادها همان توابع فشرده معمول دارای همپوشانی هستند که سیر تحول زمانی مجموعه‌ای از پارامترهای مکانی سیگنال ویدئو را توصیف می‌کنند. ما از تجزیه زمانی برای حل ساختار همپوشان رخدادها، که از مهمترین مسائل موجود در تحلیل ویدئو است، استفاده می‌کنیم. در این روش، مجموعه‌ای از پارامترهای مکانی، از سیگنال ویدئو استخراج می‌شوند و به صورت ترکیب خطی از مجموعه‌ای از تابع‌های فشرده همپوشان زمانی، طی یک مرحله بهینه‌سازی، بیان می‌گردند.

برای کاربرد چکیده‌سازی از ویژگی‌های مکانی تبدیل موجک فریم‌ها استفاده می‌کنیم. ابتدا برای پایین نگه‌داشتن پیچیدگی محاسباتی، سیگنال ویدئو به گروه‌های همپوشان تقسیم می‌گردد و پارامترهای مدل آماری حاشیه‌ای از زیر باندهای تبدیل موجک دوبعدی فریم‌ها استخراج می‌شود. تجزیه زمانی به پارامترهای مکانی که به صورت ماتریسی که ستون‌هایش بردارهای ویژگی استخراج‌شده فریم‌ها هستند، برای محاسبه توابع رخدادها اعمال می‌شود. فریم‌های قرارگرفته در مراکز ثقل رخدادها، که تعدادشان به مراتب کمتر از تعداد کل فریم‌هاست، به عنوان فریم‌های کاندید انتخاب می‌شوند. فریم‌های کلیدی طی یک مرحله بر اساس فاصله بین این کاندیداها در فضای ویژگی انتخاب می‌شوند. یکی از نوآوری‌های این روش عدم لزوم

[1] Spatio-temporal event-based approach

تعیین مرز شات و خوشه‌بندی آنها پیش از انتخاب فریم کلیدی است که مرحله‌ای لازم برای روش‌های متداول می‌باشد. نتایج آزمون‌ها تایید کننده کارایی و دقت بالای روش پیشنهادی است.

٤-١- مقدمه

سیگنال ویدئو می‌تواند به صورت مجموعه‌ای از *رخدادهای دیداری*[٢] مستقل در نظر گرفته شود که این رخدادها به صورت پیاپی اتفاق می‌افتند و رخدادهای همسایه می‌توانند با هم همپوشانی زمانی داشته باشند. منظور از *رخداد دیداری* هر گونه فعالیت بصری مشخص است که در طول زمان گسترش یافته باشد. برای مثال حرکت خودرو در امتداد جاده، پریدن گربه روی میز و واردشدن فردی به اتاق هر یک رخدادی دیداری هستند. بنابراین سیگنال ویدئو شامل ترکیبی از رخدادهای دیداری پیاپی است که می‌توانند جداگانه، همزمان و یا با مقداری همپوشانی زمانی اتفاق بیافتند. رخدادهای ویدئو شامل موارد زیر می‌شود:

١) انواع مختلف تغییر شات (برش[٣] و انواع محوشدگی[٤])، زوم دوربین[٥]، لایه‌گذاری[٦] و پدیدارشدن شئ[٧] – که می‌تواند باعث خوشه‌بندی شات[٨] شود.

٢) اشیاء زمانی که در قطعات زمانی از سیگنال امتداد یافته‌اند و در سه دسته : فعالیت‌ها[٩]، حرکات[١٠] و بافت‌های زمانی[١١] جای می‌گیرند [84,102].

هر رخداد یک فعالیت بصری زمانی است که در بعد زمان گسترش می‌یابد و دوره زمانی و شکل آن برای رخدادهای مختلف متفاوت است. رخدادهای همسایه می‌توانند دارای همپوشانی زمانی باشند. به بیان دیگر رخدادهای ویدئو با چند پارامتر توصیف می‌شوند:

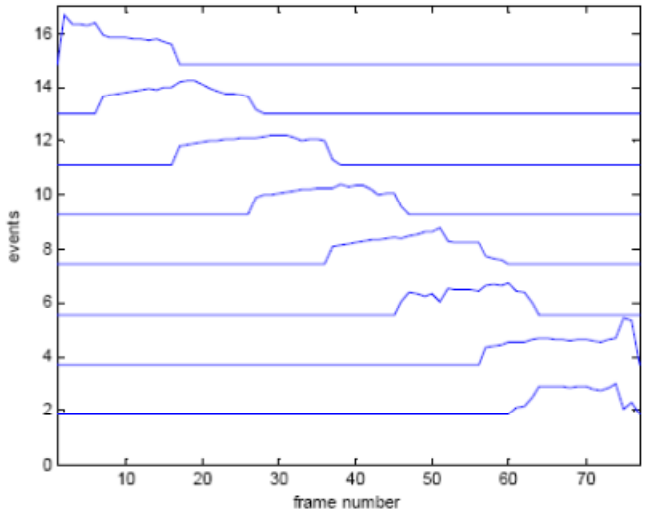
١) دوره زمانی، که می‌تواند از کوتاه – پدیدارشدن شئ- تا طولانی – گوینده خبر که در حال خواندن اخبار در فریم‌های متوالی طولانی است- تغییر کند.

٢) میزان همپوشانی با رخدادهای مجاور، به طور مثال زوم‌کردن دوربین و پدیدارشدن یک شئ به طور همزمان و یا همپوشانی زمانی بین دو شات ویدئو در حین تغییر شات با روش محوشدگی.

---

[2] Visual event
[3] Cut
[4] Fade in/out
[5] Camera zooming
[6] Padding
[7] Object emergence
[8] Shot clustering
[9] Activity
[10] Motion
[11] Temporal textures

این در حالی است که، با فرض معقول، کارگردان درباره هر رویداد منحصراً فکر و تصمیم‌گیری کرده - است و هر رخداد ویدئو می‌تواند از رخدادهای دیگر ویدئو مستقل درنظر گرفته شود. این ساختار همپوشان رخدادهای زمانی، تحلیل سیگنال ویدئو را پیچیده می‌کند، چون مرز مشخصی بین فعالیت‌های بصری جداگانه وجود ندارد. در شکل ٤-١ مثالی از توابع رخداد بصری و همپوشانی زمانی بین آنها نمایش داده شده است. در این مثال هشت رخداد بصری شناسایی شده‌اند که دارای هپوشانی زمانی می باشند. هر ردیف این نمودار، یک رخداد بصری را نمایش می‌دهد که در یک بازه زمانی در دنباله ویدئویی اتفاق افتاده است.



شکل ٤-١- نمونه‌ای از همپوشانی زمانی بین رخدادهای بصری.

در این رساله نشان داده شده است که تجزیه زمانی[١٢] به خوبی قادر به توصیف این ارتباطات پیچیده زمانی با کمک مجموعه‌ای از پارامترهای مکانی می‌باشد.

تجزیه زمانی در ابتدا در [103] برای استخراج رخدادهای همپوشان صحبت با کمک مجموعه‌ای مناسب از ویژگی‌های طیفی به کار گرفته شد. این روش تجزیه بعدها به عنوان رویکردی موثر در پردازش صوت با تعیین رخدادهای صوتی شناخته شد [104]. قابلیت تشخیص همپوشانی زمانی بین رخدادهای مجاور می-تواند در آنالیز ویدئو نیز به کار رود. ما در این رساله از این روش استفاده نموده‌ایم و قابلیت بالای آن را در چکیده‌سازی نشان داده‌ایم.

٤-٢- تقریب سیر تحول زمانی سیگنال ویدئو

TD بردارهای ویژگی استخراج‌شده از فریم‌های یک بلوک ویدئو را بر حسب جمع وزن‌دار تابع‌های رخداد – که هر یک متناظر با یک رخداد بصری هستند- بیان می‌کند. تجزیه زمانی در واقع به دنبال مدل-

[12] Temporal Decomposition (TD)

کردن سیر تحول زمانی بردارهای ویژگی مکانی سیگنال ویدئو است و این کار را بر اساس تحلیل جداگانه نقش توابع پایه متوالی در مشخصات آماری پارامترهای مکانی انجام می‌دهد. نتیجه رویکرد تحلیلی یک ماتریس متعامد از بردارهای ویژگی، *بردارهای هدف*[۱۳]، و یک ماتریس تنک از برادارهای پایه، *رخدادها* یا *اهداف*، است. این روش می‌تواند به صورت زیر بیان شود:

$$Y = A\Phi \qquad (۱-٤)$$

که Y ماتریس پارامترهای مکانی با ابعاد $q \times N$ است و N تعداد فریم‌ها و q تعداد پارامترهای مکانی استخراج شده از هر فریم هستند. Φ ماتریس حاوی توابع رخداد با ابعاد $m \times N$ و A ماتریس بردارهای اهداف با ابعاد $q \times m$ است و m تعداد توابع رخداد در فاصله فریم‌های n=1 تا n=N است.

معادله (٤-١) می‌تواند به صورت عددی[۱٤] به شکل مجموعه‌ای از معادلات خطی، که هر کدام سیر تحول زمانی پارامتر i ام ستون‌های ماتریس Y را بیان می‌کنند، نوشته شود:

$$\hat{y}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \qquad 1 \le n \le N, 1 \le i \le q \qquad (۲-٤)$$

که $\hat{y}_i(n)$ پارامتر i ام از فریم n است که با کمک این مدل تخمین زده می‌شود. $\phi_k(n)$ تابع رخداد k ام در فریم n و a_{ik} فاکتور وزن می‌باشند. در این معادلات، فقط ماتریس Y معلوم است و این ماتریس باید از طریق متعامدسازی[۱۵] تجزیه شود تا مقادیر ماتریس‌های A و Φ محاسبه شوند [104].

روند تجزیه زمانی در دو مرحله مهم انجام می‌شود. در مرحله نخست، ماتریس Y به زیربلوک‌های همپوشان، با طول l_{sb}، تقسیم می‌شود. سپس محل توابع رخداد هر زیر بلوک با تجزیه[۱٦]SVD ماتریس پارامترهای مکانی زیربلوک تعیین می‌شود. این تقسیم ماتریس Y به زیر بلوک‌ها باعث ساده‌تر شدن و افزایش سرعت الگوریتم در مقایسه با اعمال تجزیه SVD به کل ماتریس می‌شود. برای تکمیل این مرحله، توابع رخداد در یک الگوریتم تکرارشونده در جهت کمینه‌کردن اختلاف بین پارامترهای تخمین زده شده و پارامترهای اصلی اصلاح می‌شوند؛ و الگوریتم بهترین توابع رخداد را برای توصیف سیر تحول زمانی بردارهای ویژگی مکانی فریم‌های متوالی انتخاب می‌نماید.

طول هر زیربلوک l_{sb} در واقع نشان‌دهنده دقت زمانی روش TD می‌باشد. هر چه این مقدار کوچکتر باشد، احتمال از دست‌دادن رخدادهای بصری کوتاه کمتر شده، احتمال قطعه‌قطعه‌شدن یک رخداد بصری طولانی بیشتر می‌گردد. ما مقدار این متغیر را به صورت تجربی در حدود سرعت نمونه‌برداری زمانی[۱۷]

[13] Target vectors
[14] Scalar
[15] Orthogonalization
[16] Singular Value Decomposition
[17] Temporal sampling rate

سیگنال ویدئو درنظر گرفته‌ایم تا تمامی رخدادهای دیداری که در حدود یک ثانیه به درازا می‌کشند را پوشش دهیم و در عین حال پیچیدگی محاسباتی الگوریتم را هم در حد پایین نگه‌داشته باشیم.

پس از انتخاب رخدادهای هر زیربلوک، رخدادهای بهینه و محل آنها در دنباله ویدئویی اصلی بر اساس مکان‌های گذر از صفر منفی تابع زمانی $v_k(n_{c_k})$ تعیین می‌شود [103]:

$$v_k(n_{c_k}) = \frac{\sum_n (n-n_{c_k})\emptyset_k^2(n)}{\sum_n \emptyset_k^2(n)} \qquad 1 \le k \le m \qquad (۳-٤)$$

که n_{c_k} شاخص مرکز رخداد k ام مربوط به تابع رخداد k است و حاصل جمع روی کل بلوک ویدئو انجام می‌گیرد و m تعداد رخدادهای استخراج‌شده می‌باشد.

فاز اول طبق روش توضیح داده شده به هر زیر بلوک از بلوک ویدئو اعمال می‌گردد و مراکز رخدادها تعیین می‌شود. سپس طی یک مرحله اصلاح[18] خطای تجزیه کاهش می‌یابد که این مرحله دوم تاثیری در مکان مراکز رخدادها نخواهد داشت.

٤-۳- کاربرد روش تجزیه زمانی در چکیده‌سازی ویدئو

در حالی که روش‌های مطرح‌شده دارای موفقیت نسبی در انتخاب فریم‌های کلیدی هستند، این روش‌ها از کمبودهای مهمی رنج می‌برند که در روش پیشنهادی به حل آن‌ها پرداخته‌ایم:

۱) اغلب روش‌های موجود به مرحله تعیین مرز شات/خوشه قبل از مرحله انتخاب فریم کلیدی احتیاج دارند [87, 82-85, 105, 17]. این مرحله، مرحله‌ای زمان‌بر است و باعث افزایش بار محاسباتی الگوریتم چکیده‌سازی می‌شود.

۲) اغلب روش‌های متداول بر اساس فاصله بین فریم‌های مجاور به انتخاب فریم کلیدی می‌پردازند که این معیار نمی‌تواند به خوبی بیان‌گر میزان متمایز بودن فریم از فریم‌های دیگر باشد [87, 82-85, 105, 17]. این اتفاق اغلب هنگامی که محتوای بصری ویدئو به آرامی در طی فریم‌های متوالی تغییر می‌کند، می‌افتد.

۳) برخی از روش‌های موجود ویژگی‌های مناسبی برای پارامترهای مکانی انتخاب نمی‌کنند که باعث انتخاب اشتباه یا نادیده گرفته‌شدن فریم‌های کلیدی می‌شود [82, 17, 105].

این سه مورد در روش پیشنهادی حل شده است. روش ارائه‌شده نیازی به انتخاب مرز شات یا خوشه‌بندی شات‌ها ندارد. این روش برخلاف روش‌های متداول اختلاف بین تمام فریم‌های کاندید با فریم موجود را برای انتخاب بهتر فریم کلیدی در شرایط تغییر زمانی سریع و کند محتوای بصری فریم‌ها، به کار می‌برد. در

[18] Refinement

نهایت ویژگی‌های مکانی مناسب براساس خصوصیات سیستم بینایی انسان برای بهبود نتایج استفاده می‌شود. این پیشرفت‌ها باعث بهبود نتایج انتخاب فریم‌های کلیدی در مقایسه با روش‌های متداول می‌گردد.

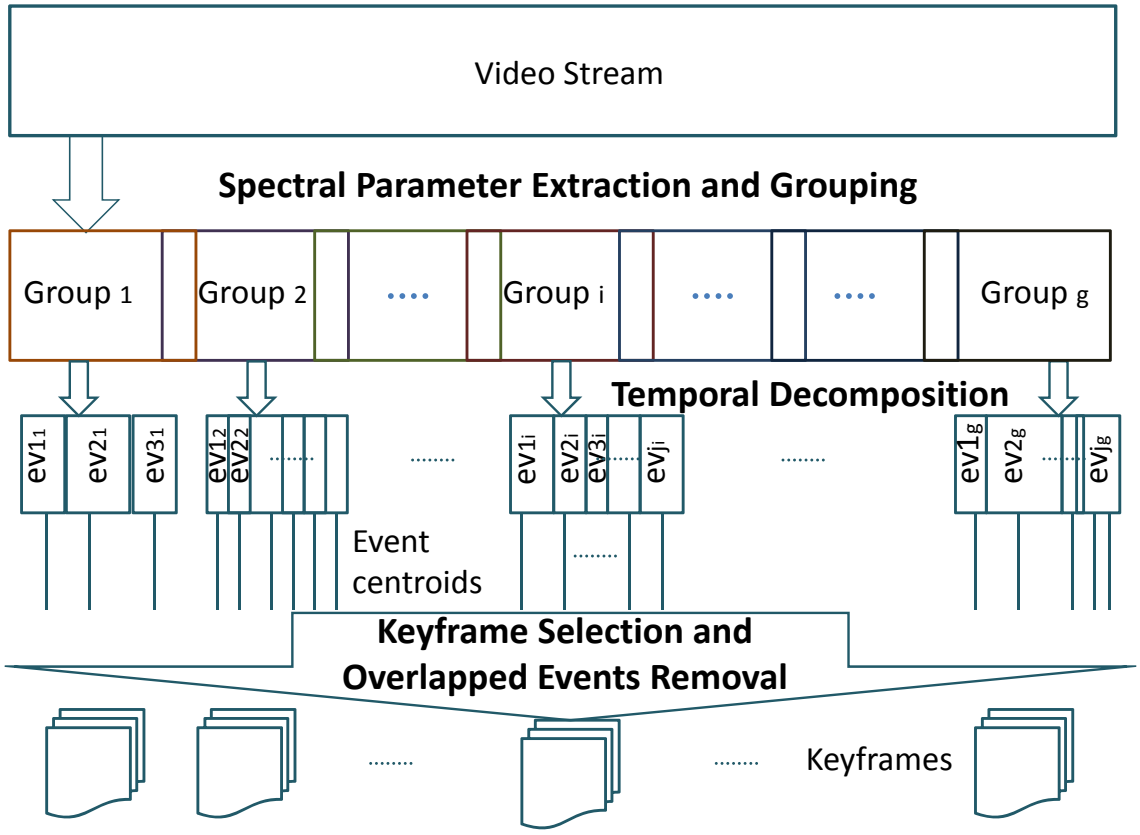
٤-٤- الگوریتم تعیین فریم کلیدی

در این رساله ما روشی نوین برای چکیده‌سازی ویدئو ارائه نموده‌ایم که بر پایه آنالیز همبستگی زمانی بین فریم‌های متوالی ویدئو، رخدادهای بصری را در دنباله ویدئویی می‌یابد. این عمل با روش تجزیه زمانی انجام می‌گیرد که در آن وابستگی‌های محلی در مجموعه‌ای از ویژگی‌های مکانی که از فریم‌های ویدئویی استخراج شده‌اند طی فرآیند تعامدسازی در فضای ویژگی‌ها تعیین می‌گردند.

در این روش ابتدا دنباله ویدئویی به گروه‌های همپوشانی از فریم‌ها –هر گروه دارای ۲۰۰-۳۰۰ فریم- تقسیم می‌شود. ایده گروه‌بندی امکان استفاده از چکیده‌سازی برای ویدئوهای طولانی در حین سریع و ساده بودن الگوریتم را فراهم می‌آورد. در مرحله بعدی TD به گروه‌ها اعمال می‌شود و فریم‌های کاندید برای مرحله انتخاب فریم‌های کلیدی تعیین می‌گردند. این مرحله تعیین فریم‌های کاندید جانشین مرحله انتخاب مرز شات و خوشه‌بندی شات‌ها در روش‌های سنتی است که قادر به حل و فصل طبیعت همپوشان رخدادهای بصری نیستند. روش ارائه‌شده پارامترهای مکانی مناسب را بر پایه خواص آماری حاشیه‌ای زیرباندهای تبدیل دوبعدی موجک استخراج می‌نماید و یک فضای پارامتریک بر پایه خواص سیستم بینایی انسان[19] تولید می‌نماید. برای استخراج ویژگی‌ها از هر فریم، تبدیل موجک انتخاب شده است زیرا که نمایشی تنک از سیگنال تصویر فراهم می‌نماید و با تابع توزیع فرکانسی سیستم بینایی انسان منطبق است [96, 97, 93, 11]. در مرحله بعدی فریم‌های کلیدی از بین مراکز رخدادهای بصری هر گروه براساس فاصله KL[20] بین پارامترهای متناظر مکانی GGD انتخاب می‌شوند. این پارامترهای مکانی در کنار روش-های گروه‌بندی و تجزیه زمانی و همچنین معیار فاصله مناسب KL فاکتورهای اصلی در رسیدن به کارایی بالای روش پیشنهادی هستند.

همچنین، ایده تحلیل زمانی، باعث انتقال سیگنال با افزونگی بالا به مجموعه‌ای متراکم از فریم‌های کلیدی می‌گردد که هیچ اطلاعات بصری درباره سیگنال از دست نمی‌رود. در نتیجه هر تحلیل برپایه محتوای بصری سیگنال می‌تواند بر روی نتیجه به دست آمده از این روند اعمال گردد.

[19] Human visual system (HVS)
[20] Kull-back Leibler Distance



شکل٤-٢. الگوریتم ارائه‌شده برای استخراج فریم‌های کلیدی.

روش ارائه‌شده در این رساله با روش‌های متداول استخراج فریم‌های کلیدی متفاوت است و طی یک پروسه شبه‌بهینه[٢١] که به وابستگی زمانی[٢٢] بین فریم‌های ویدئو بسیار حساس است، به دنبال فریم‌های کلیدی می‌گردد. مزیت دیگر این روش استفاده از تجزیه زمانی بر پایه گروه‌بندی[٢٣] ویژگی‌های مکانی سیگنال ویدئو است که به خوبی ساختار همپوشان تحولات بصری سیگنال‌های ویدئو طولانی را تحلیل می‌نماید و امکان استفاده از روش تجزیه زمانی را در کاربردهای زمان-واقعی[٢٤]، طی یک مرحله ساده تقسیم‌بندی دنباله ویدئویی به گروه‌های همپوشان، فراهم می‌آورد.

سیگنال ویدئو در ابتدا به گروه‌های همپوشان تقسیم می‌شود، سپس هر گروه فریم‌های ویدئویی با تعیین همبستگی‌های مکانی در فضای ویژگی که نمایانگر محتوای فریم‌هاست، به رخدادهای همپوشان بصری

[21] Near optimum
[22] Temporal correlation
[23] Group based TD
[24] Real time

تجزیه می‌شوند. با این روش آنالیز، فقط تعداد کمی از فریم‌های ویدئو، فریم‌هایی که در مرکز رخدادهای بصری قرار دارند، برای پروسه استخراج فریم کلیدی استفاده می‌شوند.

به این ترتیب الگوریتم استخراج فریم‌های کلیدی به چهار مرحله تقسیم می‌شود که در شکل ٤-٢ دیاگرام آن آمده است. ابتدا ویژگی‌های مکانی از فریم‌های ویدئو استخراج می‌شوند و ماتریس پارامترها تشکیل می‌شود. در مرحله دوم، سیگنال ویدئو به گروه‌های ویدئویی همپوشان تقسیم می‌شود که هر گروه شامل مجموعه‌ای از رخدادهای بصری است. لازم به ذکر است که مراحل اول و دوم می‌توانند جابجا شوند و تاثیری در نتایج یکدیگر ندارند. سپس، با کمک تجزیه زمانی، مراکز رخدادهای هر گروه از فریم‌ها انتخاب می‌شوند و در آخر، رخدادهای همپوشان حذف شده، فریم‌های کلیدی از بین مراکز رخدادها براساس ضوابط خاصی انتخاب می‌شوند. هر کدام از مراحل فوق با جزئیات در ادامه آورده شده‌اند و سپس پیچیدگی محاسباتی این رویکرد بررسی شده است.

٤-١-٤- استخراج ویژگی‌های مکانی

ما از پارامترهای تبدیل دوبعدی موجک به عنوان پارامترهای مکانی که مشخصات بصری فریم‌ها را بیان می‌کنند، استفاده نموده‌ایم. سیستم بینایی انسان[25] به لبه‌ها و بافت بسیار حساس می‌باشد. تبدیل موجک به صورت ساختاری با این خصوصیت سیستم بینایی انسان هماهنگ بوده، به خوبی این تغییرات را در نظر می‌گیرد[96,97] . به این ترتیب ما پارامترهای استخراج‌شده از این تبدیل برای ساختن بردارهای ویژگی مکانی ویدئو استفاده کرده‌ایم. بنابراین با کمک پارامترهای استخراج‌شده از سیگنال ویدئو ماتریس پارامتر Y به صورت زیر تعریف می‌شود:

$$\mathrm{Y} = [\mathrm{fv}_1 \ \mathrm{fv}_2 \ \dots \ \mathrm{fv}_\mathrm{i} \ \ \mathrm{fv}_\mathrm{N}],\ \mathrm{fv}_\mathrm{i} = \left(\mathrm{fv}_{\mathrm{GGD}_\mathrm{i}}\right)^\mathrm{T} \qquad (٤\text{-}٤)$$

و فاصله بین دو فریم i و j برابر خواهد بود با:

$$\mathrm{dist}(\mathrm{i},\mathrm{j}) = \mathrm{dist}_{\mathrm{KLD}}\left(\mathrm{fv}_\mathrm{i}, \mathrm{fv}_\mathrm{j}\right) =$$
$$D\left(\mathrm{fv}_\mathrm{i}, \mathrm{fv}_\mathrm{j}\right) = \sum_{l=1}^{3L} \mathrm{D}(\mathrm{p}(.;\alpha_\mathrm{i}^{(l)}, \beta_\mathrm{i}^{(l)}) \left\| \mathrm{p}(.;\alpha_\mathrm{j}^{(l)}, \beta_\mathrm{j}^{(l)})\right) \qquad (٥\text{-}٤)$$

که L تعداد سطوح تبدیل، fv بردار ویژگی استخراج‌شده و i و j شاخص فریم‌ها هستند.

[25] Human visual system (HVS)

٤-٢-٤- گروه‌بندی

جهت تعمیم و مناسب‌سازی ایده تجزیه زمانی به عنوان یک روش تحلیل زمانی-مکانی ویدئو، برای استفاده در کاربردهای متنوع پردازش ویدئو، روش تجزیه زمانی برپایه گروه‌بندی[٢٦] ارائه شده است. دنباله ویدئویی به گروه‌های ویدئویی همپوشان با طول فریم برابر، تقسیم می‌شود و تجزیه زمانی بر هر گروه از فریم‌ها جداگانه اعمال می‌گردد. همپوشانی بین فریم‌های مجاور مانع از دست‌دادن رخدادهای بصری در مرز بین فریم‌ها می‌شود. طول هر گروه، N_g ، در حدود ٢٠٠ تا ٣٠٠ فریم در نظر گرفته می‌شود تا پیچیدگی محاسباتی روش TD نیز در حد پایین نگهداشته شود. همچنین میزان همپوشانی بین گروه‌های مجاور، N_{ov}، بین ١٥ تا ٢٥ فریم براساس نرخ نمونه‌برداری زمانی ویدئو، در نظر گرفته می‌شود تا هیچ رخداد بصری که در حدود یک ثانیه طول می‌کشد، نادیده گرفته نشود.

به این ترتیب دنباله ویدئویی V، به g گروه همپوشان تقسیم می‌شود تا پروسه چکیده‌سازی ویدئو سریع‌تر و ساده‌تر گردد. سیگنال ویدئو می‌تواند به صورت اجتماع g گروه ویدئو توصیف شود:

$$\mathrm{V} = \mathrm{U}_{\mathrm{i}=1}^{\mathrm{g}}\,\mathrm{G}_{\mathrm{i}} \qquad (٦\text{-}٤)$$

که $\mathrm{G_i}$ نمایانگر گروه i ام از دنباله ویدئویی است. اکنون با درنظر گرفتن ماتریس Y به عنوان ماتریس حاوی بردارهای ویژگی کل دنباله ویدئویی از فریم ١ تا N، خواهیم داشت:

$$\mathrm{Y} = [\mathrm{fv}_1 \;\; \mathrm{fv}_2 \;\; \dots \;\; \mathrm{fv}_{\mathrm{j}} \dots \;\; \mathrm{fv}_{\mathrm{N}}] \qquad (٧\text{-}٤)$$

با اعمال گروه‌بندی به دنباله ویدئویی، ماتریس ویژگی هر گروه به صورت زیر در می‌آید:

$$\mathrm{Y_i} = [\mathrm{fv}_{\mathrm{i}\times\mathrm{N_g}-\mathrm{i}\times\mathrm{N_{ov}}+1} \;\; \mathrm{fv}_{\mathrm{i}\times\mathrm{N_g}-\mathrm{i}\times\mathrm{N_{ov}}+2} \dots \mathrm{fv}_{(\mathrm{i}+1)\times\mathrm{N_g}-\mathrm{i}\times\mathrm{N_{ov}}}] \qquad (٨\text{-}٤)$$

که $\mathrm{Y_i}$ ماتریس پارامترهای مکانی گروه i ام می‌باشد. با درنظرگرفتن این که طول هر گروه برابر N_g فریم است و تعداد فریم‌های همپوشان بین گروه‌های مجاور برابر N_{ov} فریم می‌باشد، گروه i ام حاوی فریم‌های شماره $\mathrm{i}\times\mathrm{N_g}-\mathrm{i}\times\mathrm{N_{ov}}+1$ تا $(\mathrm{i}+1)\times\mathrm{N_g}-\mathrm{i}\times\mathrm{N_{ov}}$ خواهد بود.

٤-٣-٤- تجزیه زمانی ویدئو

در این مرحله روش TD به هر گروه $\mathrm{G_i}$ اعمال می‌شود تا مراکز رخدادهای ماتریس پارامترهای $\mathrm{Y_i}$ گروه i ام با کمک معادله (٤-١) شناسایی گردد. فرض کنید $\mathrm{n_i}$ نشان‌دهنده تعداد رخدادها در گروه i، و $\mathrm{ind}_{\mathrm{i},1}, \mathrm{ind}_{\mathrm{i},2}, \dots, \mathrm{ind}_{\mathrm{i},\mathrm{n_i}}$ شاخص مراکز رخدادهای گروه i ام که با روش TD انتخاب شده‌اند، و $\mathrm{fv}_{\mathrm{ind}_{\mathrm{i},1}}, \mathrm{fv}_{\mathrm{ind}_{\mathrm{i},2}}, \dots, \mathrm{fv}_{\mathrm{ind}_{\mathrm{i},\mathrm{n_i}}}$ نمایانگر بردارهای ویژگی مربوط به این مراکز رخداد هستند. این مراکز

[26] Group based temporal decomposition

رخدادها به عنوان فریم‌های کاندید برای فریم‌های کلیدی استفاده می‌شوند. از آنجا که تعداد رخدادهای استخراج‌شده از تجزیه زمانی به مراتب کمتر از تعداد فریم‌های گروه‌ها می‌باشد، معمولاً کمتر از ۱۵٪ کل فریم‌ها، تعداد فریم‌هایی که در مرحله انتخاب فریم‌های کلیدی دخیل خواهند بود و در نتیجه حجم محاسبات به مراتب کاهش خواهد یافت.

مراکز رخدادها به عنوان فریم‌های کاندید برای انتخاب فریم‌های کلیدی استفاده می‌شوند. ولی این احتمال وجود دارد که این فریم‌ها فریم‌های کلیدی نباشند؛ زیرا:

۱) ممکن است رخداد بصری مشابهی در یک گروه ویدئویی وجود داشته باشد که در فواصل زمانی متفاوت رخ دهند و این رخداد به عنوان رخدادهای مختلف با تجزیه زمانی شناسایی شده باشد.
۲) ممکن است رخدادهای بصری طولانی به چند رخداد تقسیم شوند. برای رفع این مشکل، ما این رخدادها را به عنوان کاندید برای تهیه چکیده ویدئویی در نظر می‌گیریم.

چون برخی رخدادها ممکن است بسیار طولانی باشند و فریم‌های مشابه به عنوان مراکز رخداد انتخاب شوند. ما طول زیربلوک‌ها – دقت زمانی– را در مرحله تجزیه زمانی در حدود سرعت نمونه‌برداری زمانی ویدئو در نظر می‌گیریم تا تمام رخدادهای بصری را که در حدود یک ثانیه ادامه دارند، پوشش دهیم و همچنین میزان پیچیدگی محاسباتی را پایین نگهداریم. در نتیجه امکان وجود افزونگی در رخدادهای استخراج شده وجود دارد و ممکن است تمام مراکز رخدادها از هم مستقل نباشند.

بنابراین برای حفظ سادگی و سرعت الگوریتم، این مراکز رخدادها به عنوان فریم‌های کاندید برای فریم‌های کلیدی در نظر گرفته می‌شوند و افزونگی موجود در آن‌ها در مرحله بعدی حذف می‌شود.

٤-٤-٤- ویدئوی خلاصه‌شده[۲۷]

برای انتخاب فریم‌های کلیدی از بین فریم‌های کاندید گروه i ام، میانگین فاصله بین هر فریم کاندید و دیگر فریم‌های کاندید آن گروه محاسبه می‌شود:

$$\mathrm{D_i(j)} = \frac{1}{\mathrm{n_i}-1}\sum_{\substack{\mathrm{k}=1 \\ \mathrm{k}\neq\mathrm{j}}}^{\mathrm{n_i}} \mathrm{D}\left(\mathrm{fv}_{\mathrm{ind_{i,j}}}, \mathrm{fv}_{\mathrm{ind_{i,k}}}\right) \qquad (٩-٤)$$

که $\mathrm{j} = 1, \dots, \mathrm{n_i}$ و $\mathrm{D}(.)$ فاصله KL بین بردارهای ویژگی $\mathrm{fv}_{\mathrm{ind_{i,j}}}$ و $\mathrm{fv}_{\mathrm{ind_{i,k}}}$ است. به‌این ترتیب تعداد $\frac{n_i}{2} \times (n_i - 1)$ محاسبه $\mathrm{D}(.)$ لازم خواهد بود. با درنظر گرفتن این‌که تعداد مراکز ثقل رخدادهای استخراج‌شده درمرحله تجزیه زمانی حداکثر برابر ۱۵٪ تعداد کل فریم‌هاست، $n_i = 0.15N$ برقرار خواهد بود طوری که N تعداد کل فریم‌ها و n_i تعداد رخدادهای گروه i ام باشند. درنتیجه تعداد محاسبات تقریباً

[27] Vide summary

برابر $\frac{0.15N}{2}(0.15N-1) \cong 0.0225N^2$ می‌شود که در مقایسه با N^2 محاسبه مورد نیاز در هنگام عدم استفاده از تجزیه زمانی، بار محاسباتی در حدود $\frac{1}{0.0225} = 44$ مرتبه کاهش می‌یابد.

فریم‌های کلیدی فریم‌هایی هستند که مقدار فاصله محاسبه‌شده برای آنها طبق فرمول (٤-١٢) بزرگتر از سطح آستانه $T = 2m_w$ باشد، که این سطح آستانه به صورت تجربی با آزمون بر روی ویدئوهای مختلف انتخاب شده است. m_w میانگین محلی $\mathrm{D_i}$ روی پنجره‌ای با اندازه w و مرکزیت فریم مورد محاسبه است که اینجا برابر با سه در نظر گرفته می‌شود. به این ترتیب، فریم‌های کلیدی هر گروه، فریم‌هایی خواهند بود که بیشینه فاصله را با مراکز رخداد دیگر در آن گروه داشته باشند.

نهایتاً، اولین و آخرین فریم‌های کلیدی استخراج‌شده از هر گروه با فریم‌های کلیدی انتخاب‌شده در مرز گروه‌های مجاورشان مقایسه می‌شوند و اگر فریم‌های کلیدی مجاور در گروه‌های متفاوت مشابه باشند، یکی از آنها حذف می‌شود. سطح آستانه برای این کار برابر $T = \frac{m_{g_i}+m_{g_{i+1}}}{2}$ انتخاب می‌شود که m_{g_i} میانگین فاصله بین فریم‌های کلیدی گروه i ام است.

فریم‌های کلیدی استخراج‌شده، نماینده سیگنال ویدئویی خواهند بود و می‌توانند برای بهبود کارایی بسیاری از کاربردهای پردازش ویدئو نظیر تشخیص صحنه[۲۸]، تحلیل، ویرایش و بازیابی ویدئو استفاده شوند. همچنین این فریم‌های کلیدی می‌توانند برای مدیریت دادگان‌های عظیم به کار روند و عملیات جستجو را در این سیستم‌ها سریع‌تر و آسان‌تر نمایند.

٤-٤-٥- بررسی پیچیدگی محاسباتی سیستم

برای تحلیل پیچیدگی محاسباتی روش TD، ابتدا پیچیدگی محاسباتی الگوریتم SVD را بیان می‌نماییم که برای یک ماتریس با ابعاد p در n برابر $O(pn^2)$ می‌باشد [106]. به این ترتیب می‌توانیم یک تابع مثبت غیرنزولی خطی[۲۹] $f(.)$ را در نظر بگیریم به طوری که f از مرتبه O از[۳۰] n^2 باشد. پس پیچیدگی می‌تواند به صورت $f(pn^2)$ نوشته شود و با توجه به خطی بودن تابع، به صورت $p.f(n^2)$ بیان گردد.

با اعمال الگوریتم TD به ماتریس پارامتر ویدئو با ابعاد $p \times n$ ، این ماتریس به زیرماتریس‌هایی با ابعاد $p \times l_{sb}$ تقسیم می‌شود، پس پیچیدگی محاسباتی برای این زیربلوک‌ها برابر $p.f(l_{sb}^2)$ می‌باشد و پیچیدگی محاسباتی کل تحلیل ویدئو با روش تجزیه زمانی بر پایه گروه‌بندی برای تعیین مراکز ثقل رخدادها برای هر گروه می‌تواند به صورت $n_{sub-blocks}.p.f(l_{sb}^2)$ بیان شود که $n_{sub-blocks}$ نمایانگر تعداد

[28] Scene detection
[29] Positive non-decreasing linear function
[30] Big-O of n^2

زیربلوک‌هاست و از آنجا که تعداد زیر بلوک‌ها کمتر از تعداد رخدادها می‌باشد، این پیچیدگی کمتر از $n_{ev}.p.f(l_{sb}^2)$ خواهد بود که n_{ev} تعداد رخدادهای استخراج شده است.

مراکز ثقل رخدادها به عنوان کاندید برای تعیین چکیده ویدئویی نهایی استفاده می‌شوند. پیچیدگی محاسباتی باقیمانده در روش ارائه‌شده مربوط به انتخاب فریم‌های کلیدی از بین n_{ev} فریم خواهد بود و با انتخاب فریم کلیدی از بین n فریم کل سیگنال ویدئو که در روش‌های متداول انتخاب فریم کلیدی استفاده می‌شود، مقایسه خواهد شد.

در اینجا ما پیچیدگی روش خود را با دو روش انتخاب فریم کلیدی مقایسه می‌نماییم:

- در روش اول فاصله بین بردارهای ویژگی دو فریم مجاور برای انتخاب فریم کلیدی استفاده می‌شود. بنابراین مرتبه محاسبات مورد نیاز برای انتخاب فریم‌های کلیدی می‌تواند با یک تابع خطی مثبت غیر نزولی $g(.)$ از مرتبه O g – n – بیان شود. به این ترتیب، پیچیدگی محاسباتی روش سنتی انتخاب فریم کلیدی $g(n)$ خواهد بود و پیچیدگی محاسباتی روش ارائه‌شده $g(n_{ev})$ می‌باشد. پیچیدگی محاسباتی روش ارائه‌شده می‌تواند به صورت $n_{ev}.p.f(l_{sb}^2)+g(n_{ev})$ بیان گردد. نتایج آزمایش‌ها و بررسی‌ها نشان می‌دهند که $n_{ev}<0.15n$، پس با توجه به خطی‌بودن $g(.)$ می‌توان گفت $g(n_{ev})<0.15g(n)$. بخش اول الگوریتم تجزیه زمانی هنوز نیاز به بررسی دارد: با درنظر گرفتن شرط $l_{sb}^2<n$، و خطی بودن $f(n)$ خواهیم داشت $n_{ev}.p.f(l_{sb}^2)<0.15np_{max}f(n)$. علاوه بر این $f(n)$ و $g(n)$ در یک مرتبه هستند و همچنین توابع خطی، مثبت و غیرنزولی هستند. پس می‌توان عدد مثبت a را پیدا کرد، به گونه‌ای که شرط $f(n)\leq ag(n)$ برقرار باشد. به این ترتیب پیچیدگی محاسباتی سیستم ارائه‌شده به صورت زیر بیان می‌گردد:

$$(0.15+a\,(n_{ev})_{max}p_{max})g(n)\approx \left(0.15+a\left(\frac{n}{l_{sb}}\right)_{max}p_{max}\right)g(n) \qquad (۱۰-٤)$$

که هدف ما کوچک‌تر بودن این پیچیدگی از پیچیدگی محاسباتی روش سنتی است که برابر $g(n)$ می‌باشد. این هدف با شرط ساده و قابل دسترس زیر به دست می‌آید:

$$l_{sb}^2<n,\ \frac{n_{max}\cdot p_{max}}{l_{sb_{min}}}<0.85a \qquad (۱۱-٤)$$

- دیدگاه دوم مقایسه تمام فریم‌های موجود در یک بازه معین از سیگنال با یکدیگر، جهت انتخاب فریم با بزرگترین فاصله از فریم‌های دیگر به عنوان فریم کلیدی می‌باشد. برای این کار مرتبه

محاسبات می‌تواند با تابع خطی مثبت غیرنزولی $g(.)$ – که g از مرتبه O از n^2 است– بیان گردد. به این ترتیب پیچیدگی محاسباتی روش سنتی با $g(n^2)$ و روش ارائه‌شده با $g(n_{ev}^2)$ بیان می‌شود. همان طور که قبلاً هم مطرح شد، رابطه $n_{ev} < 0.15n$ برقرار است، پس $g(n_{ev}^2) < 0.0225g(n^2)$ و مجموع محاسبات مدل ارائه‌شده به طور تقریبی به صورت زیر خواهد بود:

$$10^{-2}f(n^2) + 10^{-2}g(n^2) \leq 10^{-2}(1+a)g(n^2) \qquad (۱۲-٤)$$

با فرض این‌که $l_{sb} \approx \frac{n}{n_{sb}}$ و $n_{ev}.p \approx 1$ که n_{sb} تعداد زیربلوک‌ها در گروه می‌باشد و همیشه بیشتر از ۱۰ است. باز با توجه به این حقیقت که $f(n^2)$ و $g(n^2)$ توابع خطی مثبت غیر نزولی و در یک مرتبه هستند، عدد مثبت a را می‌توان یافت طوری که $f(n^2) \leq ag(n^2)$. بنابراین اگر $a < 99$ باشد، پیچیدگی محاسباتی کلی روش ارائه‌شده کمتر از روش سنتی است؛ و در غیر اینصورت با اعمال شرط $l_{sb}^2 < n$ و حذف جمله اول نامعادله، این نامساوی برقرار خواهد بود. گرچه مقدار a در عمل به مراتب کوچک‌تر از این سطح آستانه است زیرا دو تابع از یک مرتبه‌اند. بنابراین پیچیدگی محاسباتی روش ارائه‌شده کمتر از ۱۰٪ پیچیدگی محاسباتی روش سنتی می‌گردد.

این در حالی است که ما محاسبات مربوط به انتخاب مرز شات وخوشه‌بندی شات را برای روش سنتی در نظر نگرفته‌ایم؛ که با درنظر گرفتن آنها بار محاسباتی روش‌های سنتی بالاتر خواهد رفت.

٤-۵- نتایج آزمایش‌ها

روش ارائه‌شده، پیاده‌سازی و نتایج آن در نمونه‌های مختلف انتخاب فریم کلیدی بررسی شده است. در اینجا جزئیات آزمایش‌ها و نتایج مهم بیان می‌شوند.

برای انتخاب مقادیراولیه برای برخی پارامترها نظیر l_{sb}, N_g, N_{ov} و همچنین سطوح آستانه موجود در الگوریتم، دادگان Hollywood2 استفاده شده است. این مجموعه ۱۱۵۲ دنباله ویدئویی شامل ۱۰۲۵۲۷۸ فریم ویدئو و ۸۱۹۹ تغییر شات است، که از بین ۹۶ فیلم سینمایی انتخاب شده است. روش ارائه‌شده بر روی این ویدئوها اعمال و نتایج بررسی و بر اساس نتایج انتخاب فریم‌های کلیدی، مقادیر متغییرها تعیین شده‌اند.

برای ارزیابی الگوریتم چکیده‌سازی ویدئو، به طور خاص از دادگان تعیین مرز شات TRECVID 2006 استفاده کرده‌ایم. این مجموعه شامل ۱۳ ویدئوی طولانی با ۵۹۷۰٤۳ فریم ویدئو است که ۳۷۸۵

تغییر شات – که ٤٨,٧٪ آنها به‌صورت برش[31] و مابقی ٥١,٣٪ تغییر تدریجی شات[32] هستند- و نرخ نمونه-برداری زمانی ویدئوها ٢٩ فریم در ثانیه است. اطلاعات بیشتر درباره این مجموعه در جدول ٤-١ آمده است. در این جدول منظور از *نسبت شات*[33] نسبت تعداد شات‌ها به کل فریم‌های ویدئویی است؛ و *میانگین فریم بر شات*[34] نشان‌دهنده میانگین تعداد فریم‌ها در هر شات است. همان طور که در جدول دیده می‌شود، بیشتر نمونه‌های ویدئویی پویا بوده مقدار *میانگین فریم بر شات* آنها حدود ١٥٨ فریم است؛ که بیان‌گر وجود تعداد شات‌های کوتاه زیادی است. این در حالی است که ویدئوی اول شات‌های طولانی‌تری دارد.

جدول ٤-١. جزئیات دادگان TRECVID 2006.

ردیف	نام ویدئو	تعداد فریم‌ها	تعداد شات‌ها	نسبت شات	میانگین فریم بر شات
١	20051101_142800_LBC_NAHAR_ARB.mpg	١١٢٠٨	٢٤٣	٪٠,٢٢	٤٦١
٢	20051114_091300_NTDTV_FOCUSINT_CHN.mpg	٣١١٣٨	١٩٨	٪٠,٦٤	١٥٧
٣	20051115_192800_NTDTV_ECONFRNT_CHN.mpg	٣١١٦٩	١٦٨	٪٠,٥٤	١٨٦
٤	20051129_102900_HURRA_NEWS_ARB.mpg	٢٢٦٥٩	١٤٨	٪٠,٦٥	١٥٣
٥	20051205_185800_PHOENIX_GOODMORNCN_CHN.mpg	٥٨١٤٢	٣٦٠	٪٠,٦٢	١٦٢
٦	20051208_125800_CNN_LIVEFROM_ENG.mpg	٥٨١٤٢	٣٦٧	٪٠,٦٣	١٥٨
٧	20051208_145800_CCTV_DAILY_CHN.mpg	٥٨١٤٢	٤٤١	٪٠,٧٦	١٣٢
٨	20051208_182800_NBC_NIGHTLYNEWS_ENG.mpg	٥٨١٤٢	٦٣٨	٪١,١	٩١
٩	20051209_125800_CNN_LIVEFROM_ENG.mpg	١٢٨٦٤	٩٤	٪٠,٧٣	١٣٧
١٠	20051213_185800_PHOENIX_GOODMORNCN_CHN.mpg	٥٨١٤٢	٢٧١	٪٠,٤٧	٢١٥
١١	20051227_105800_MSNBC_NEWSLIVE_ENG.mpg	٥٨١٤٢	٥٢٠	٪٠,٨٩	١١٢
١٢	20051231_182800_NBC_NIGHTLYNEWS_ENG.mpg	٢٨٥٠٢	٢٦٦	٠,٩٣	١٠٧
١٣	20051227_125800_CNN_LIVEFROM_ENG.mpg	٩٧٧٢	٧١	٪٠,٧٣	١٣٨
مجموع		**٥٩٧٠٤٣**	**٣٧٨٥**	**٪٠,٦٨**	**١٥٨**

پس از استخراج پارامترهای GGD از زیرباندهای تبدیل موجک دوبعدی فریم‌ها و تشکیل ماتریس ویژگی ویدئو، مرحله گروه‌بندی بر روی ویدئوهای طولانی اعمال می‌گردد. سپس مراکز ثقل رخدادهای بصری هر گروه با روش TD انتخاب می‌گردند. در انتها فریم‌های کلیدی از میان مراکز ثقل رخدادهای هر

[31] Cut transition
[32] Gradual transition
[33] Shot ratio
[34] Mean frame/shot

گروه انتخاب می‌شوند و فریم‌های کلیدی مرزی با فریم‌های کلیدی مرزی گروه مجاورشان مقایسه و فریم-های مشابه حذف می‌شوند.

برای ارزیابی روش ارائه‌شده، از ویژگی‌های هیستوگرام رنگ در مختصات فضای رنگ HSV[35] استفاده نموده‌ایم. ویژگی‌های رنگ HSV در بسیاری از کاربردهای پردازش ویدئو به عنوان ویژگی‌هایی که حاوی محتوای بصری فریم‌ها هستند به کار گرفته شده‌اند [17,101]. در این روش‌ها، هیستوگرام‌های رنگ HSV نرمالیزه برای فریم‌ها محاسبه می‌گردد و به‌ترتیب ۸، ٤ و ٤ ویژگی از مختصات Hue، Saturation و Value انتخاب می‌شوند. برای محاسبه فاصله بین بردارهای ویژگی HSV از فاصله اقلیدسی[36] استفاده می-شود.

الف. دنباله ویدئویی نمونه‌برداری‌شده زمانی.

ب. فریم‌های کلیدی استخراج‌شده بر اساس روش پیشنهادی (پارامترهای مکانی GGD + روش تجزیه زمانی).

د. فریم‌های کلیدی استخراج‌شده بر اساس پارامترهای مکانی رنگ و فاصله اقلیدسی.

ج. فریم‌های کلیدی استخراج‌شده بر اساس پارامترهای مکانی رنگ و روش تجزیه زمانی.

شکل٤-۳. نتایج استخراج فریم کلیدی.

[35] Hue Saturation Value
[36] Euclidean Distance

الف. دنباله ویدئویی نمونه‌برداری‌شده زمانی.

ب. فریم‌های کلیدی استخراج‌شده بر اساس روش پیشنهادی (پارامترهای مکانی GGD + روش تجزیه زمانی).

ج. فریم‌های کلیدی استخراج‌شده بر اساس پارامترهای مکانی رنگ و روش تجزیه زمانی.

د. فریم‌های کلیدی استخراج‌شده بر اساس پارامترهای مکانی رنگ و فاصله اقلیدسی.

شکل٤-٤. نتایج استخراج فریم کلیدی.

مثال‌هایی از نتایج انتخاب فریم کلیدی در شکل‌های ٤-٣ و ٤-٤ آمده است. برای نمایش فریم‌ها در قسمت الف این تصاویر، نمونه‌برداری زمانی انجام داده‌ایم. همان طور که در شکل‌ها دیده می‌شود، روش ارائه‌شده بر اساس تجزیه زمانی (قسمت ب از تصاویر) در مقایسه با روش‌های دیگر، بخوبی دنباله ویدئویی را نمایش می‌دهد و تفاوت‌ها را تشخیص می‌دهد.

برای نشان دادن کارایی روش ارائه‌شده، ویژگی‌های رنگ HSV نیز برای استخراج فریم‌های کلیدی استفاده شده‌اند. نتایج استفاده از این پارامترها به عنوان پارامترهای مکانی و اعمال روش TD برای انتخاب فریم‌های کلیدی در قسمت ج از دو شکل ٤-٣ و ٤-٤ آمده‌اند. همچنین، برای مقایسه روش ارائه‌شده با روش‌های موجود، پارامترهای رنگ HSV با کمک فاصله اقلیدسی برای پیدا کردن مرز شات‌ها و خوشه-بندی آنها و انتخاب فریم کلیدی به صورت سنتی به کار گرفته شده‌اند و نتایج در قسمت د شکل‌های ٤-٣ و

٤-٤ آمده است. همانطور که پیش‌تر هم مطرح شد، ویژگی‌های رنگ در بسیاری کاربردهای ویدئو استفاده شده‌اند [17, 101]. روش ما به مراتب بهتر از این روش‌ها عمل می‌کند، زیرا ویژگی‌های مکانی به کار گرفته‌شده در این روش، ما را قادر به تشخیص جزئیات بافت تصویر می‌نماید که مهم‌تر از رنگ آن است. برای مثال همانطور که در شکل ٤-٤ دیده می‌شود، برخی از فریم‌های دنباله ویدئویی نماینده‌ای بین فریم-های کلیدی در قسمت‌های ج و د شکل ٤-٤ ندارند - فریم‌های موجود در سطر سوم قسمت الف شکل ٤-٤. این در حالی است که روش ما انتخاب‌های بهتری کرده است و این به دلیل استفاده از ویژگی‌های فضای تبدیل موجک است که تشابه بیشتری با سیستم بینایی انسان دارد [96, 97].

علاوه بر نتایج بالا، برای ارزیابی روش ارائه‌شده، از یک تست ادراکی استفاده نموده‌ایم. زیرا روش تجزیه زمانی روشی نوین می‌باشد و هیچ روش مشابهی که برای استخراج فریم‌های کلیدی نیازی به تعیین شات و خوشه‌بندی آنها نداشته باشد، تا کنون گزارش نشده است؛ و از طرفی روش استانداردی برای ارزیابی نتایج استخراج فریم‌های کلیدی نیست [80]. برای این منظور از ۹ شخص مستقل که آگاهی اندکی درباره روش-های پردازش ویدئو داشتند، برای سنجش نتایج روش ارائه‌شده استفاده نمودیم. این افراد نتایج انتخاب فریم-های کلیدی را به سه رده *خوب*[37]، *بد*[38] و *قابل قبول*[39] براساس آزمون ارائه‌شده در [107] امتیازدهی کردند. هر بار یک شات ویدئو انتخاب و در قسمت بالای صفحه نمایش‌گر به شخص تحت آزمایش نمایش داده می‌شود و سپس فریم‌های کلیدی انتخاب‌شده در قسمت پایین صفحه نشان داده شده، از شخص خواسته می‌شود که درباره کیفیت این انتخاب نظر داده، یکی از سه دسته مذکور را برای نتایج انتخاب فریم‌های کلیدی آن شات انتخاب نماید. در ادامه شات‌های بعدی به همین ترتیب به فرد نمایش داده شده، از او درخواست می‌شود که امتیاز هر شات را وارد کند. این کار تا انتهای دنباله ویدئو ادامه می‌یابد.

لازم به ذکر است که این مدل ارزیابی بر اساس نتایج هر شات بر اساس فریم‌های کلیدی آن، نحوه اجرای روش ما را تغییر نمی‌دهد و روش ما همانطور که توضیح داده شد، وابسته به تعیین شات‌ها نیست. بلکه نمایش نتایج به این شکل است. فریم‌های کلیدی در روش ما بر اساس تجزیه زمانی بر پایه گروه‌بندی دنباله ویدئویی است.

جدول ٤-٢ نتایج آزمون‌های ادراکی روی دادگان TRECVID را نمایش می‌دهد. 'Kf%' نشان‌گر درصد تعداد فریم‌های کلیدی به‌کل فریم‌های ویدئو، و 'Kf/shot' میانگین تعداد فریم‌های کلیدی به شات است. نتایج نشان می‌دهند که 'Kf%' برای ویدئوهای حاوی شات‌های کوتاه و در نتیجه پویا بیشتر است

[37] Good
[38] Bad
[39] Acceptable

(ویدئوهای ۷ و ۸). همچنین 'kf/sh' برای ویدئوهای با فعالیت و حرکت کمتر، مقدار کمتری دارد (ویدئوی ۱). این مساله توانایی روش ارائه‌شده در تعقیب سیر تحول زمانی محتوای بصری سیگنال ویدئو را نشان می‌دهد. علاوه بر این، بر اساس نتایج آزمون ادراکی، ۸۵٪ شات‌ها امتیاز خوب گرفته‌اند که این بیان‌گر قدرت بالای مدل مطرح‌شده برای انتخاب و تعیین دقیق مکان فریم‌های کلیدی است.

جدول ٤-٢. نتایج تست ادراکی روی دادگان TRECVID. Gr# تعداد گروه‌ها، Sh# تعداد شات‌ها، Ev# تعداد رخدادهای انتخاب‌شده، Kf# تعداد فریم‌های کلیدی استخراج‌شده. پارامترهای سیستم: فیلتر موجک: 'Daubechies4'، تعداد سطوح تبدیل:٤، طول گروه (N_g): ۲۵۰ ، همپوشانی بین گروه‌های مجاور(N_{ov}): ۲۵ و طول زیربلوک (l_{sb}): ۲۵.

	#Gr	#Sh	#Ev	#Kfs	خوب	قابل‌قبول	بد	Kf%	Kf/shot
۱	٤٩٨	٢٤٣	١٦٥١٨	٣٢٣	٢١٥(٪٨٨,٤٨)	١٥(٪٦,١٧)	١٣(٪٥,٣٥)	٪٠,٢٩	١,٣٢٩
٢	١٣٨	١٩٨	٤٨٧٠	٣٣٢	١٦٢(٪٨١,٨٢)	١٧(٪٨,٥٩)	١٩(٪١١,٧٣)	٪١,٠٧	١,٦٧٧
٣	١٣٨	١٦٨	٤٨٠٨	١٩٣	١٤١(٪٨٣,٩٣)	١٢(٪٧,١٤)	١٥(٪١٠,٦٤)	٪٠,٦٢	١,١٤٩
٤	١٠٠	١٤٨	٣٤٣٦	٢٣١	١٢٣(٪٨٣,١١)	١٧(٪١١,٤٩)	٨(٪٦,٥)	٪١,٠٢	١,٥٦١
٥	٢٥٨	٣٦٠	٨٧٣٠	٤٧٢	٢٨٤(٪٧٨,٨٩)	٤٣(٪١١,٩٤)	٣٣(٪١١,٦٢)	٪٠,٨١	١,٣١١
٦	٢٥٨	٣٦٧	٨٣٦٧	٥٤٥	٣٢٧(٪٨٩,١٠)	٢٥(٪٦,٨١)	١٥(٪٤,٥٩)	٪٠,٩٤	١,٤٨٥
٧	٢٥٨	٤٤١	٨٦٥٥	٧٦٧	٣٦٣(٪٨٢,٣١)	٤٢(٪٩,٥٢)	٣٦(٪٩,٩٢)	٪١,٣٢	١,٧٣٩
٨	٢٥٨	٦٣٨	٨٦١٨	٨٧٨	٥٤٤(٪٨٥,٢٧)	٤٧(٪٧,٣٧)	٤٧(٪٨,٦٤)	٪١,٥١	١,٣٧٦
٩	٥٧	٩٤	١٨١٦	١٩٢	٨٣(٪٨٨,٣٠)	٨(٪٨,٥١)	٣(٪٣,٦١)	٪١,٤٩	٢,٠٤٣
١٠	٢٥٨	٢٧١	٨٧٤٤	٣٩٦	٢١٦(٪٧٩,٧٠)	٣٣(٪١٢,١٨)	٢٢(٪١٠,١٩)	٪٠,٦٨	١,٤٦١
١١	٢٥٨	٥٢٠	٨٨٩٨	٦١٧	٤٦٣(٪٨٩,٠٤)	١٥(٪٢,٨٨)	٤٢(٪٩,٠٧)	٪١,٠٦	١,١٨٧
١٢	١٢٦	٢٦٦	٤٢٨٧	٤٤٧	٢١٩(٪٨٢,٣٣)	٢٦(٪٩,٧٧)	٢١(٪٩,٥٩)	٪١,٥٧	١,٦٨
١٣	٤٣	٧١	١٤٢	١١٥	٦٢(٪٨٧,٣٢)	٤(٪٥,٦٣)	٥(٪٨,٠٦)	٪١,١٨	١,٦٢

همچنین روش ارائه‌شده را با سه روش دیگر در یک آزمون ادراکی مقایسه نموده‌ایم. در این آزمون، علاوه بر ویژگی‌های رنگ HSV، از ویژگی‌های رنگ L*a*b*[40] نیز استفاده نموده‌ایم. فضای رنگ L*a*b* براساس آزمون‌های ادراکی فراوان طراحی شده است و تمامی رنگ‌های مرئی را در بر دارد و ادعا شده است که فاصله اقلیدسی مناسب‌ترین معیار فاصله برای اندازه‌گیری تفاوت ادراکی بین رنگ‌های دو تصویر در این فضای رنگ است [108]. مانند روش بکار گرفته‌شده برای انتخاب ویژگی‌های رنگ HSV، هیستوگرام‌های رنگ نرمالیزه برای هر مختصات برای فریم مورد نظر محاسبه می‌شود و به ترتیب ۸، ٤ و ٤ پارامتر از مختصات L* (روشنایی)، a* (مکان بین رنگ قرمز/سبز)، و b* (مکان بین رنگ آبی/زرد) انتخاب می-

[40] CIELab

شوند. همچنین فاصله اقلیدسی برای اندازه‌گیری فاصله بین بردارهای ویژگی انتخاب می‌گردد. ویدئوی ۹ از دادگان TRECVID برای این آزمون ادراکی انتخاب شده و آزمون مانند قسمت قبل تکرار می‌شود. جدول ٤-۳ نتایج ارزیابی را نمایش می‌دهد که بیان‌گر قدرت بسیار بالای روش ارائه‌شده است. نتایج همچنین تاییدکننده تاثیر بسیار مثبت پارامترهای GGD استخراج‌شده از زیرباندهای تبدیل موجک هستند.

روش‌های به کار رفته در آزمون به شرح زیر هستند:

روش ۱: پارامترهای مکانی GGD + روش تجزیه زمانی با استفاده از فاصله KL.

روش ۲: پارامترهای مکانی HSV + روش تجزیه زمانی با استفاده از فاصله اقلیدسی.

روش ۳: پارامترهای مکانی HSV + روش سنتی با استفاده از فاصله اقلیدسی.

روش ٤: پارامترهای مکانی L*a*b* + روش سنتی با استفاده از فاصله اقلیدسی.

جدول٤-۳. نتایج ارزیابی ادراکی روی ویدئوی ۹. تعداد فریم‌ها:۱۲۸٦٤، تعداد گروه‌ها: ۵۷، تعداد شات‌ها: ۹٤، میانگین تعداد فریم بر شات: ۱۳۲.

	# رخداد	# فریم‌های کلیدی	خوب	قابل‌قبول	بد	Kf%	Kf/sh
روش ۱	۱۸۱٦	۱۹۲	۸۳(٪۸۸,۳)	۸(٪۸,۵۱)	۳(٪۳,٦۱)	٪۱,٤۹	۲,۰٤۳
روش ۲	۱۷۲۰	۱۷٤	۷۵(٪۷۹,۷۹)	۱۱(٪۱۱,۷)	۸(٪۸,۵۱)	٪۱,۳۵	۱,۸۵۱
روش ۳	-	۱۵۳	۷۰(٪۷٤,٤۷)	۱٤(٪۱٤,۸۹)	۱۰(٪۱۰,٦٤)	٪۱,۱۹	۱,٦۲۸
روش ٤	-	۱۵۷	۷۲(٪۷٦,٦)	۱۱(٪۱۱,۷)	۱۱(٪۱۱,۷)	٪۱,۲۲	۱,٦۷۰

همان طور که قبلاً هم بیان شد یکی از مزایای روش ارائه‌شده، عدم نیاز به تعیین مرز شات‌ها و خوشه-بندی آنها برای انتخاب فریم‌های کلیدی است و روش مذکور فریم‌های کلیدی را با انتخاب رخدادهای هر گروه و حذف مراکز رخداد مشابه در گروه و مرز بین گروه‌ها انجام می‌دهد. این ساختار گروه‌بندی قابلیت استفاده از الگوریتم را در کاربردهای زمان واقعی فراهم می‌آورد.

از آنجا که انتظار می‌رود حداقل یک فریم در هر شات از ویدئو انتخاب شده باشد، آزمون‌های بسیاری بر روی دادگان TRECVID با تغییردادن پارامترهای مختلف سیستم – سطوح تجزیه مختلف و فیلترهای موجک متفاوت، طول‌های مختلف گروه‌ها و زیربلوک‌ها و میزان همپوشانی‌های متفاوت بین گروه‌ها- انجام داده‌ایم. تعداد فریم‌ها و تعداد شات‌های دادگان به‌ترتیب ۵۹۷۰٤۳ و ۳۷۸۵ هستند. نتایج این آزمون‌ها در جداول ٤-٤ تا ٤-۸ نمایش داده شده‌اند. ’P‘ در این جداول مقیاسی برای نمایش درصد شات‌هایی است که الگوریتم ما حداقل یک فریم کلیدی برای آنها انتخاب نموده است و به عنوان معیاری برای ارزیابی دقت الگوریتم ارائه می‌شود. ’Event%‘ هم درصد تعداد رخدادها و در نتیجه مراکز رخدادها را نسبت به کل

فریم‌ها نمایش می‌دهد و نشان‌گر میانگین نسبت فریم‌های کلیدی به کل فریم‌ها و معیاری برای نمایش مزایای کاهش پیچیدگی روش ارائه‌شده است.

جدول ٤-٤، تاثیر تغییر طول گروه را در نتایج نمایش می‌دهد. نسبت تعداد فریم‌های کلیدی به تعداد شات‌ها، درصد فریم‌های کلیدی و 'p' با کاهش اندازه گروه افزایش می‌یابد در حالی که تعداد گروه‌ها و در نتیجه بار محاسباتی افزایش می‌یابد.

جدول ٤-٤. نتایج ارزیابی برای طول گروه‌های (N_g) مختلف. فیلتر موجک: 'Daubechies'، تعداد سطوح تبدیل: ٤، همپوشانی بین گروه‌ها (N_{ov}): ٢٥، طول زیربلوک (l_{sb}): ٢٥.

N_g	# گروه	# رخداد	# فریم کلیدی	Kf/sh	Kf%	Event%	P
١٥٠	٤٧٧٣	٩٦٢٨٤	٦٥٤٩	١,٧٣٠٣	٪١,١	٪٠,١٦١٣	٪95.25
٢٠٠	٣٤٠٦	٩١٣٧٨	٥٧٦٥	١,٥٢٣١	٪٠,٩٧	٪٠,١٥٣١	٪94.15
٢٥٠	٢٦٤٨	٨٩١٧٥	٥٥٠٨	١,٤٥٥٢	٪٠,٩٢	٪٠,١٤٩٤	٪93.70
٣٠٠	٢١٦٤	٨٧٠٠٩	٥١١٣	١,٣٥٠٩	٪٠,٨٦	٪٠,١٤٥٧	٪92.56

تاثیر تغییر تعداد فریم‌های همپوشان بین گروه‌ها N_{ov} بر روی نتایج چکیده‌سازی ویدئو در جدول ٤-٥ آمده است. همان طور که از نتایج دیده می‌شود، با افزایش همپوشانی بین گروه‌های مجاور، نسبت فریم‌های کلیدی به شات‌ها، درصد فریم‌های کلیدی و 'p' به آرامی افزایش می‌یابند. به این ترتیب، نتایج با تغییر همپوشانی چندان تغییری نمی‌کنند و عددی حدود ١٥ نتایج قابل قبولی دارد.

جدول ٤-٥. نتایج ارزیابی برای تعداد فریم‌های همپوشان متفاوت (N_{ov}). : فیلتر موجک: 'Daubechies4'، تعداد سطوح تبدیل:٤، طول گروه (N_g): ٢٥٠ و طول زیربلوک (l_{sb}): ٢٥.

N_{ov}	# گروه	# رخداد	# فریم کلیدی	Kf/sh	Kf%	Event%	P
١٥	٢٥٣٤	٨٥٣٦٤	٥١٦٦	١,٣٦٤٩	٠,٨٧	٠,١٤٣٠	٪٩٢,٩١
٢٠	٢٥٨٧	٨٧١٣٤	٥٢٨٨	١,٣٩٧١	٠,٨٩	٠,١٤٥٩	٪٩٣,٢٧
٢٥	٢٦٤٨	٨٩١٧٥	٥٥٠٨	١,٤٥٥٢	٠,٩٢	٠,١٤٩٤	٪٩٣,٧

جدول ٤-٦. نتایج ارزیابی برای اندازه‌های مختلف زیربلوک (l_{sb}). فیلتر موجک: 'Daubechies4'، تعداد سطوح تبدیل:٤، طول گروه (N_g): ٢٥٠ ، همپوشانی بین گروه‌های مجاور(N_{ov}): ١٥ و طول زیربلوک (l_{sb}): ٢٥، تعداد گروه‌ها: ٢٥٣٤.

l_{sb}	# رخداد	# فریم کلیدی	Kf/sh	Kf%	Event%	P
١٥	١٣٤٣٥٥	٥٨١١	١,٥٣٥٣	٠,٩٧	٠,٢٢٥	٪٩٣,٢٣
٢٥	٨٥٣٦٤	٥١٦٦	١,٣٦٤٩	٠,٨٧	٠,١٤٣٠	٪٩٢,٩١
٣٥	٥٨٨١١	٤٥٥٦	١,٢٠٣٧	٠,٧٦	٠,٠٩٨٥	٪٩١,٦٧
٤٥	٤٤١٥٠	٤٠١٥	١,٠٦٨٠	٠,٦٧	٠,٠٧٣٩	٪٩١,٩٤

در جدول ٤-٦ نتایج مربوط به آزمون‌ها برای ابعاد مختلف زیربلوک‌ها ارائه شده است. دقت روش، درصد فریم‌های کلیدی، event% و تعداد فریم‌های کلیدی بر شات به صورت معکوس با طول زیربلوک‌ها مرتبط است. با کاهش دقت زمانی[41]، رخدادهای بصری کوتاه‌تر تشخیص داده می‌شوند و سیستم شات‌های کوتاه را بهتر تشخیص می‌دهد.

جدول ٤-٧. نتایج ارزیابی برای تعداد سطوح تبدیل موجک متفاوت. فیلتر موجک: 'Daubechies4'، طول گروه (N_g): ٢٥٠ ، همپوشانی بین گروه‌های مجاور(N_{ov}): ٢٥ و طول زیربلوک (l_{sb}): ٢٥، تعداد گروه‌ها: ٢٦٤٨.

Levels	# رخداد	# فریم کلیدی	Kf/sh	Kf%	Event%	P
٣	٨٩٨٦٨	٤٧٢٤	١,٢٤٨١	٠,٧٩	٠,١٥٠٥	٪٩١,٤٨
٤	٨٩١٧٥	٥٥٠٨	١,٤٤٩٤	٠,٩٢	٠,١٤٩٤	٪٩٣,٧
٥	٩٠١٦٩	٦٦٩٦	١,٧٦٩١	١,١٢	٠,١٥١٠	٪٩٦,٣٥

جدول ٤-٨. نتایج ارزیابی برای فیلترهای موجک متفاوت. تعداد سطوح موجک:٤ ، طول گروه (N_g): ٢٥٠ ، همپوشانی بین گروه‌های مجاور(N_{ov}): ٢٥ و طول زیر بلوک (l_{sb}): ٢٥، تعداد گروه‌ها: ٢٦٤٨.

Levels	# رخداد	# فریم کلیدی	Kf/sh	Kf%	Event%	P
Haar	٨٩٤٣٦	٤٧٩١	١,٢٦٥٨	٠,٨	٠,١٤٩٨	٪٩٢,٥٢
Daubechies4	٨٩١٧٥	٥٥٠٨	١,٤٤٩٤	٠,٩٢	٠,١٤٥٩	٪٩٣,٧
Symlet4	٩٠٨٦٢	٦٥٤٧	١,٧٢٩٧	١,١	٠,١٥٢٢	٪٩٥,٩٥

همچنین تاثیر ویژگی‌های مکانی بر کارایی روش استخراج فریم‌های کلیدی در جداول ٤-٧ و ٤-٨ بررسی شده است. با افزایش سطوح تبدیل، اطلاعات ظریف‌تر و دقیق‌تری از جزئیات فریم‌ها استخراج می‌گردد و نتایج بهبود می‌یابد. همچنین نتایج به نوع فیلتر موجک وابسته هستند و بهترین نتایج مربوط به اعمال فیلتر 'Symlet' است که فیلتری شبه متقارن است.

آزمایش‌های ارائه‌شده کارایی بالای روش پیشنهادی برای چکیده‌سازی ویدئو را ثابت می‌کند. این توانایی مرهون دو فاکتور بسیار مهم در کنار استفاده از روش تجزیه زمانی و گروه‌بندی سیگنال ویدئو می‌باشد:

١) انتخاب ویژگی‌های مکانی مناسب تبدیل موجک (پارامترهای GGD) که به خوبی با سیستم بینایی انسان تطبیق دارند [96, 97].
٢) استفاده از معیار فاصله مناسب (فاصله KL)

[41] Temporal resolution

با در نظرگرفتن این واقعیت که هیچ مرحله انتخاب و تعیین مرز شات و خوشه‌بندی شاتی در روش پیشنهادی وجود ندارد، این روش می‌تواند به عنوان روشی کارا، دقیق و سریع برای تحلیل صحنه نمایش در بسیاری از کاربردهای پردازش ویدئو به کار رود. نتایج آزمایش‌ها در جدول ٤-٩ خلاصه شده‌اند. در این جدول تاثیر تغییرات هر کدام از متغیرها بر نتایج نشان داده شده و بهترین مقادیر آنها هم بیان شده‌اند.

جدول٤-٩. تاثیر پارامترهای مختلف بر نتایج چکیده‌سازی.

	Kf/sh	Kf%	P	بار محاسباتی	بهترین نتایج
$N_g\uparrow$	↓	↓	↓	↓	۲۰۰-۲۵۰
$N_{ov}\uparrow$	↑	↑	↑	↑	۲۰-۲۵
$l_{sb}\uparrow$	↓	↓	↓	↓	۱۵-۲۰
↑تعداد سطح تبدیل	↑	↑	↑	↑	٤

٤-٦- جمع‌بندی

در این فصل روشی بر مبنای تجزیه زمانی پارامترهای مکانی برای تحلیل سیگنال ویدئو ارائه شد که با کمک آن رخدادهای بصری در دنباله ویدئو تعیین می‌شوند. سیگنال ویدئو به عنوان دنباله‌ای از رخدادهای بصری همپوشان در نظر گرفته می‌شود و روش تجزیه زمانی سیر تحولات زمانی این سیگنال را به صورت تقریب خطی از رخدادهای بصری که به وجود آورنده سیگنال ویدئویی هستند، توصیف می‌کند. پارامترهای مکانی GGD که از هیستوگرام حاشیه‌ای ضرایب زیرباندهای تبدیل دوبعدی موجک استخراج شده‌اند، به عنوان ویژگی‌های مکانی در تحلیل رخدادهای بصری استفاده می‌شوند. به منظور کاهش بار محاسباتی و صرفه‌جویی در زمان، ابتدا دنباله ویدئویی به گروه‌های متوالی دارای همپوشانی تقسیم می‌شود. برخلاف روش‌های متداول، استخراج فریم‌های کلیدی در روش پیشنهادی نیازی به تعیین شات‌ها و خوشه‌بندی آنها ندارد و با یک مرحله گروه‌بندی ساده، دنباله به گروه‌های کوچک‌تر برای پایین نگه‌داشتن بار محاسباتی و امکان استفاده از الگوریتم برای کاربردهای زمان واقعی تقسیم می‌شود. نتایج شبیه‌سازی‌ها و آزمون‌های فراوان و تحلیل ویدئو، توانایی و قابلیت بسیار بالای روش پیشنهادی را در مقایسه با روش‌های متداول نشان می‌دهد.

فصل پنجم

تحلیل تحول زمانی پارامترهای مکانی با کمک مدل AR

مقدمه

ساختار مدل پارامتریک AR

الگوریتم ارائه‌شده

نتایج آزمون‌ها

جمع‌بندی

۵-۱- مقدمه

در این بخش از رساله به مدل‌سازی AR بر اساس خواص آماری تبدیل دوبعدی موجک جهت تحلیل سیگنال ویدئو می‌پردازیم. سیگنال ویدئو مجموعه‌ای از سری‌های زمانی پارامترهای مکانی است که می‌توانند با مدل AR تقریب زده شوند. مدل AR ساختار زمانی خطی فرآیندهای تصادفی را توصیف می‌کند. در اینجا از پارامترهای تعمیم‌یافته گوسی برای تقریب خواص آماری حاشیه‌ای زیرباندهای تبدیل دوبعدی موجک فریم‌ها به عنوان پارامترهای مکانی استفاده می‌شود. تبدیل دوبعدی موجک همانطور که پیش‌تر نیز بیان شد، با خصوصیات سیستم بینایی انسان تطبیق دارد و در مقایسه با ویژگی‌های دیگر مانند رنگ ویژگی-های مناسب‌تری را ارائه می‌دهد. مدل AR هر بردار ویژگی مکانی را به صورت ترکیب خطی بردارهای پیشین با طول زمانی مشخص بیان می‌کند. در این بخش از رساله از کاربرد انتخاب مرز شات و استخراج فریم‌های کلیدی برای نشان دادن توانایی این مدل استفاده نموده‌ایم. با کمک خطای تخمین AR تعیین مرز شات‌ها و استخراج فریم کلیدی انجام گرفته است. نتایج شبیه‌سازی بیان‌گر کارایی بالای این روش است.

۵-۲- ساختار مدل پارامتریک AR

مدل AR مدل ساده‌ای برای سری‌های زمانی است، که به منظور تخمین و مدل‌کردن استفاده می‌شود. یک مدل AR با مرتبه p به صورت زیر بیان می‌گردد:

$$x_n = \sum_{j=1}^{p} a_j x_{n-j} + \eta_n \qquad (۱-۵)$$

که η_n نویز ناهمبسته با واریانس σ است. در حوزه تبدیل z، تابع تبدیل به صورت زیر است:

$$H(z)=\frac{\sigma}{A(z)}=\sigma\frac{1}{\sum_{j=0}^{p}a_j z^{-j}}=\sigma\frac{1}{\prod_{i=1}^{p}(1-\alpha_i/z)} \qquad (۲-۵)$$

که $a_0=1$ در نظر گرفته می‌شود و α_i ها هم قطب‌های سیستم هستند. تغییر محتوای ویدئو تدریجی نیست. در نتیجه روش مناسب برای آموزش مدل AR باید دارای شرایط زیر باشد:

- سرعت همگرایی بازگشتی الگوریتم باید به اندازه کافی زیاد باشد، تا پارامترهای مدل بتوانند ساختار ویدئوی حاضر را ارائه دهند.
- پارامترهای مدل باید به سیگنال جدید حساس باشند.

در این کار از روش مستقیم RLS[1] فیلتر وفقی FIR برای تخمین پارامترها استفاده شده است. عملکرد RLS از الگوریتم می‌نیمم مربعات با فاکتور فراموشی به صورت زیر استفاده می‌کند.

- Initial: $a(0)=0$, $P(0)=\delta^{-1}I$, where δ is a small positive number (usually 0.01), $\boldsymbol{I}$ is an identity matrix.
- Update: for $n = 1, 2, \ldots$

$$e(n)=d(n)-a^T(n-1)u(n)$$

$$k(n)=\frac{P(n-1)u(n)}{\lambda+u^T(n)P(n-1)u(n)}$$

$$P(n)=\frac{1}{\lambda}[P(n-1)-k(n)u^T(n)P(n-1)]$$

$$a(n)=a(n-1)k(n)e(n)$$

λ فاکتور فراموشی، $a(n)$ بردار ضرایب AR، $u(n)$ بردار ورودی، $d(n)$ بردار خروجی، $k(n)$ بهره مرحله بازگشتی، $e(n)$ خطای تخمین و $P(n)$ معکوس ماتریس کواریانس ورودی است. در حالت تحلیل زمانی ویدئو، که به صورت سیستم چند متغیره تغییرپذیر با زمان است، مدل پارامتریک بر روی پارامترهای مکانی اعمال می‌گردد.

مزایای استفاده از مدل پارامتریک AR برای مدل‌سازی ویدئو، به شرح زیر است:

- این مدل توانایی بیان سیستم‌های زمانی خطی پیچیده و نویزی را دارد.
- این مدل برای بیان ساختار زمانی ویدئو مناسب است.
- از پارامترهای سیستم، به خوبی می‌توان در تشخیص رخداد[2] در ویدئو استفاده کرد.

[1] Recursive Least Squares
[2] Event

- فاصله بین پارامترهای بدست آمده برای دنباله‌های ویدئویی بسادگی امکان‌پذیر است.

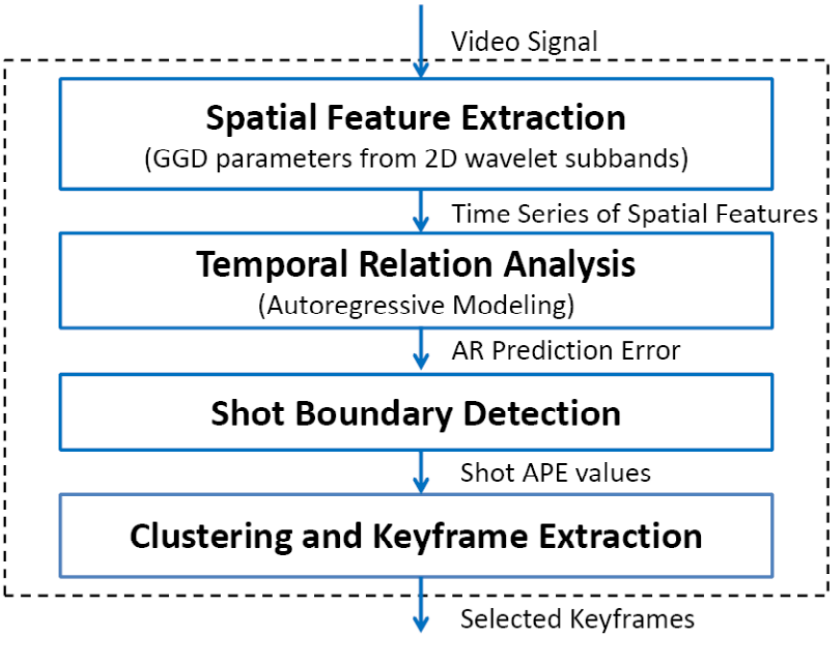
مقدار APE در هر مرحله آموزش به عنوان معیاری برای اندازه‌گیری خطا در نظر گرفته می‌شود و به صورت زیر بیان شود:

$$APE(n) = \sum_{i=1}^{M} |e_i(n)| w_i \qquad (۳-۵)$$

که $e_i(n)$ خطای تخمین AR مولفه i ام و w_i وزن مولفه i ام است. وزن‌ها باید به گونه‌ای نرمالیزه شوند که $\sum_{i=1}^{M} w_i = 1$ برقرار باشد. انتخاب مرتبه مدل AR و الگوریتم مناسب برای آموزش آن مرحله مهمی در بکارگیری این مدل در هر سیستمی به حساب می‌آید.

در این بخش از مدل AR برای بیان روابط زمانی ویژگی‌های مکانی ویدئو استفاده می‌شود. سیگنال ویدئو به عنوان مجموعه‌ای از سری‌های زمانی ویژگی‌های مکانی در نظر گرفته می‌شود و مدل AR برای بیان تحول زمانی آن بکار می‌رود. این مدل تغییرات زمانی دنباله ویدئو را به صورت خطی بیان می‌کند. از خطای تخمین در هر فاصله زمانی برای تعیین مرز شات و خوشه‌بندی آن و انتخاب فریم کلیدی استفاده می‌شود. در ادامه به توضیح الگوریتم بکار گرفته شده می پردازیم.

۵-۳- الگوریتم ارائه‌شده



شکل ۵-۱- سیستم استخراج فریم کلیدی بر اساس مدل AR.

در این قسمت الگوریتم استخراج فریم‌های کلیدی با کمک مدل‌سازی AR بیان می‌شود. این الگوریتم دارای چهار مرحله است، که در شکل ۵-۱ نشان داده شده است. ابتدا پارامترهای مکانی از فریم‌های ویدئو استخراج شده، ماتریس ویژگی‌ها تشکیل می‌شود. سپس با اعمال مدل AR رابطه زمانی پارامترهای مکانی بررسی می‌گردد. در مرحله بعد مرز شات‌ها بر اساس خطای تخمین AR انتخاب می‌شوند و نهایتا یک یا چند فریم کلیدی از هر شات ویدئو انتخاب می‌گردند. در ادامه این بخش هر مرحله با جزئیات بیشتر توضیح داده می‌شوند.

۵-۳-۱- استخراج ویژگی‌های مکانی

دو نوع ویژگی مکانی براساس رنگ (روش‌های موجود) و براساس تبدیل موجک (روش ارائه‌شده) استفاده شده‌اند:

ویژگی‌های هیستوگرام رنگ: برای ارزیابی روش ارائه‌شده، از ویژگی‌های هیستوگرام رنگ استخراج‌شده از فریم‌ها استفاده شده است. این ویژگی‌ها از هیستوگرام رنگ در فضای رنگ HSV[3] که در بسیاری از کاربردهای پردازش ویدئو برای نمایش محتوای بصری فریم‌های ویدئو استفاده شده‌اند، استخراج می‌شوند. در این روش‌ها هیستوگرام‌های نرمالیزه‌شده در هر فضا محاسبه می‌شود و به ترتیب ۸، ٤ و ٤ ویژگی از مختصات Hue، Saturation و Value استخراج می‌گردد.

ویژگی‌های برمبنای تبدیل دوبعدی موجک: همانطور که در بخش‌های دیگر این رساله نیز بیان شده، سیستم بینایی انسان به لبه‌ها و بافت‌ها حساس است و ویژگی‌های رنگ نمی‌توانند این اطلاعات را منتقل نمایند در صورتی که تبدیل موجک دارای ساختاری است که با خصوصیات سیستم بینایی انسان تطابق دارد و تغییرات لبه‌ها را به خوبی درک می‌کند [96,97]. به این ترتیب ما از پارامترهای موجک برای تشکیل بردارهای ویژگی استفاده نموده‌ایم. با کمک تابع تعمیم‌یافته گوسی، توزیع ضرایب زیرباندهای تبدیل موجک دوبعدی فریم‌های ویدئویی را تقریب زده، بردار ویژگی فریم‌ها را تشکیل می‌دهیم.

۵-۳-۲- تحلیل روابط زمانی

روش AR برای توصیف سیر تحول زمانی ویژگی‌های مکانی استفاده می‌شود. روش RLS برای تخمین پارامترهای مدل انتخاب شده است. فرض کنید $FV_i = [fv_{i,1}\ fv_{i,2}\ \dots\ fv_{i,j}\ fv_{i,N}]^T$ بردار ویژگی استخراج‌شده از فریم i ام باشد. هدف مدل‌سازی AR پیدا کردن پارامترهای $a_{j,k}$ است، به گونه‌ای که رابطه زیر برقرار باشد:

[3] Hue Saturation and Value

$$fv_{i,j} = \sum_{k=1}^{p} a_{j,k} fv_{i-k,j} + e_{i,j} \qquad (٤-٥)$$

در این صورت خطای تخمین APE که معیار ما برای شناسایی مرز شات‌ها و انتخاب فریم‌های کلیدی است، برابر است با :

$$APE_i = \sum_{j=1}^{N} |e_{i,j}| w_j \qquad (٥-٥)$$

که w_j وزن هر ویژگی مکانی است و باید طوری تعیین شود که $\sum_{j=1}^{N} w_j = 1$ برقرار باشد.

۵-۳-۳- انتخاب مرز شات و فریم کلیدی

خطای تخمین AR که در قسمت قبلی محاسبه شد، برای انتخاب مرز شات‌ها به کار می‌رود. منحنی AR کشیده می‌شود و مرز شات‌های آنی و تدریجی براساس سطح آستانه‌ای که در [17] نیز استفاده شده است انتخاب می‌شوند. سطح آستانه برابر $T = 3m_p$ انتخاب می‌شود که m_p میانگین مکانی خطای APE بر روی پنجره‌ای به سایز p است و به صورت تجربی برابر با ۳ انتخاب شده است.

ابتدا هر شات بر اساس ماکزیمم‌های محلی منحنی APE آن شات به یک یا چند خوشه تقسیم می‌شود و فریمی که می‌نیممم خطای تخمین را در هر خوشه دارد، به عنوان فریم کلیدی آن خوشه انتخاب می‌گردد. در دنباله‌های ویدئویی، محتوای برخی فریم‌ها در فریم‌های بعدی و قبلی ظاهر می‌شوند و فریم کلیدی با تخمین دوطرفه بهتر پیدا می‌شود. پس تخمین‌های رو به جلو و رو به عقب برای تخمین پارامترهای AR به کار می‌روند. در این حالت استخراج فریم کلیدی برابر یافتن فریمی با کمترین APE در شات است، که به این معناست که فریم‌های قبلی و بعدی به‌خوبی فریم کنونی را مدل می‌کنند. به بیان دیگر این فریم نمایش مناسبی از تمام دنباله فریم‌هاست. پروسه استخراج فریم کلیدی را می‌توان به صورت زیر خلاصه کرد:

۱- APE با یک فیلتر یکنواخت‌کننده، فیلتر می‌شود و ماکزیمم محلی آن برای تقسیم شات به خوشه‌ها استفاده می‌شود.

۲- در یک خوشه بهترین فریم، فریم با می‌نیمم APE است. این فریم به عنوان کاندید در نظر گرفته می‌شود و پارامترهای AR از روی فریم‌های دیگر خوشه برای آن فریم دوباره تخمین‌زده می‌شوند.

۳- فاصله میان تمام فریم‌های کاندید کلیدی محاسبه می‌شود و اگر فاصله فریم‌های کلیدی دو خوشه از سطح آستانه‌ای کوچکتر باشد، آن دو خوشه با هم ترکیب شده، تولید یک خوشه می‌نمایند.

۵-٤- نتایج آزمایشات

روش ارائه‌شده برای استخراج فریم‌های کلیدی از ویدئوهای طبیعی به کار رفته است. در اینجا به ارائه نتایج شبیه‌سازی‌ها می‌پردازیم.

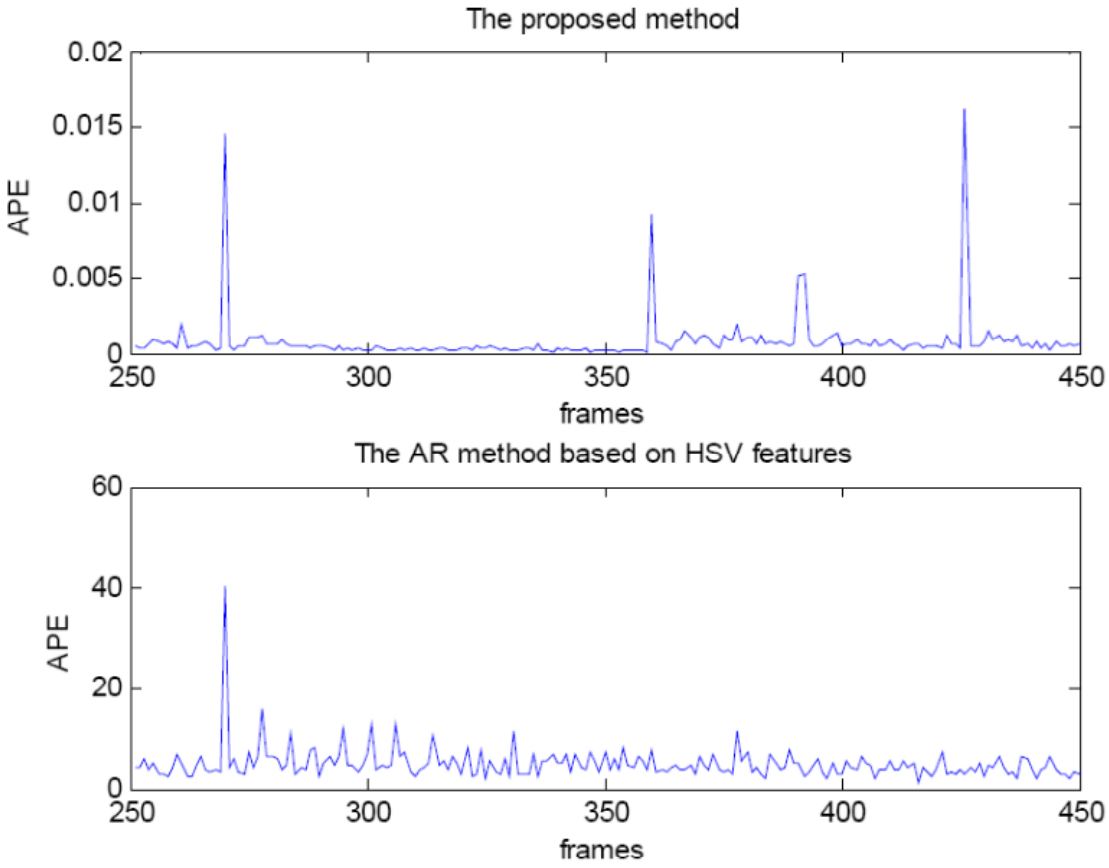
۵-٤-۱- مجموعه آزمون

تعداد زیادی دنباله ویدئویی شامل بیش از ۳۸۰۰۰ فریم ویدئویی با خصوصیات مختلف برای ارزیابی روش به کار گرفته شده است. اغلب این ویدئوها از دادگان Hollywood2 [89] انتخاب شده‌اند. ابعاد فریم‌ها CIF/QCIF است و نرخ نمونه‌برداری زمانی آنها ۱۵ یا ۲٤ فریم در ثانیه می‌باشد. ویدئوهای تست در مکان‌های مختلفی فیلمبرداری شده‌اند و دارای خواص متنوعی هستند. اغلب ویدئوها دینامیک می‌باشند ولی تعدادی ویدئوی استاتیک هم در دادگان وجود دارد.

۵-٤-۲- نتایج شبیه‌سازی

تبدیل دوبعدی موجک به همه فریم‌های ویدئو اعمال می‌شود و دو پارامتر GGD از هر زیرباند استخراج می‌گردند. سپس با کمک الگوریتم RLS مدل AR برای دنباله زمانی هر پارامتر مکانی محاسبه می‌گردد و خطای APE به دست می‌آید. مقدار مناسب برای مرتبه مدل (p) به صورت تجربی بین ۵ تا ۲۵ و مقدار پارامتر فراموشی الگوریتم RLS بیش از ۰,۹ انتخاب شده‌اند. در نهایت مرز شات‌ها و فریم‌های کلیدی تعیین می‌گردند.

شکل ۵-۲ بردار APE را برحسب فریم‌های ویدئو برای روش ارائه‌شده و روش پیشنهادی در [17] را نمایش می‌دهد. ویدئوی آزمون 'sceneclipautoautotrain00077.avi' از دادگان Hollywood2 است. نمودارها نشان‌دهنده دقت روش پیشنهادی هستند. که مرز شات‌ها در فریم‌های ۳۵۸، ۳۹۰ و ٤۲٤ با روش [17] شناسایی نشده‌اند، زیرا این روش بر اساس ویژگی‌های رنگ است و تغییر شات را هنگامی که توزیع رنگ در شات‌های مجاور تقریبا یکسان است، تشخیص نمی‌دهد. دو مثال از تغییر شات‌ها در شکل ۵-۳ نشان داده شده است. که روش [17] تغییر شات را در مثال دوم تشخیص نمی‌دهد در حالی که روش پیشنهادی هر دو مرز شات را به خوبی تعیین می‌کند. تغییر شات‌های مختلفی شامل تغییر آنی و تدریجی در دادگان وجود دارد و روش ارائه‌شده آنها را تشخیص می‌دهد. جدول ۵-۱ شامل ارزیابی نتایج تعیین مرز شات‌ها می‌باشد که برای برخی از ویدئوهای دادگان Hollywood2 محاسبه شده‌اند. نتایج تاییدکننده بهبود چشمگیر کارایی سیستم می‌باشد.

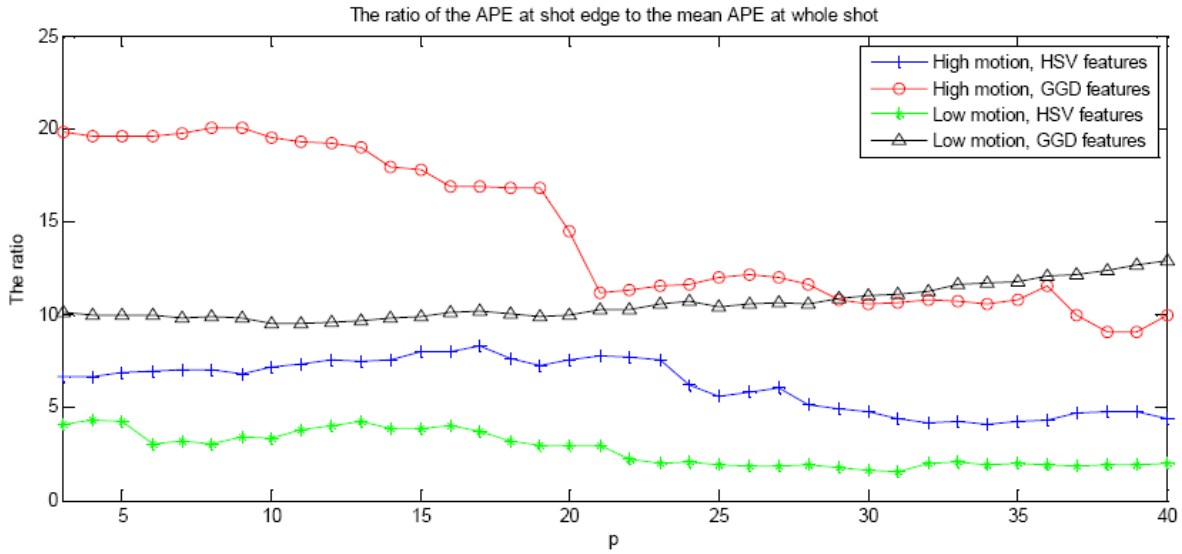


شکل ۵-۲- مرحله پیش‌پردازش. مرز شات‌ها در فریم‌های 268, 358, 390 و 424، با روش پیشنهادی (نمودار بالا) و روش [17] (نمودار پایین).

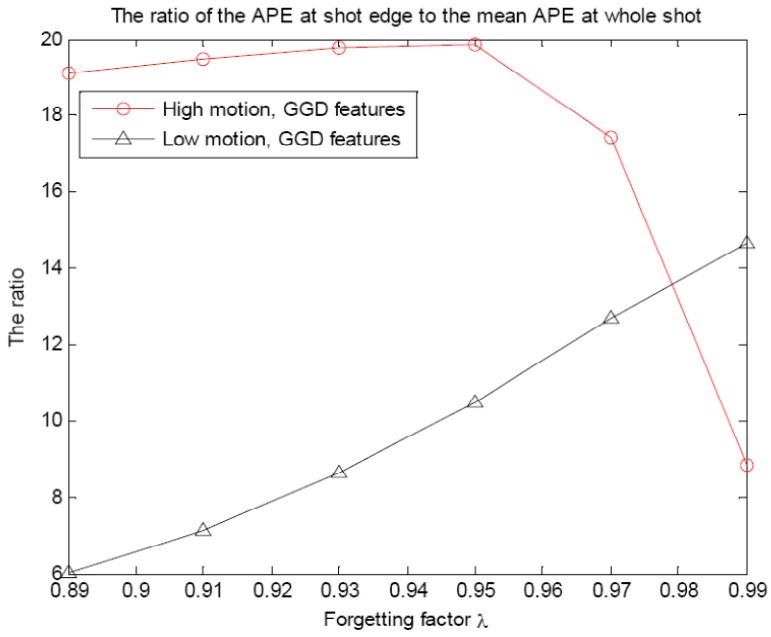
شکل ۵-۳- دو نمونه تغییر شات.

علاوه بر این، نسبت APE در فریم مرز شات به میانگین APE در تمام فریم‌های شات، که با APEr نشان داده می‌شود، محاسبه شده است و در شکل ۵-٤-الف این نسبت را برای مراتب مختلف مدل AR، p، برای دو شات ویدئو با میزان فعالیت متفاوت محاسبه نموده است. ما این نسبت‌ها را برای روش [17] نیز محاسبه نموده‌ایم. همان‌طور که از نمودارها دیده می‌شود، نسبت‌های محاسبه‌شده با روش ارائه‌شده، به

مراتب بالاتر از نسبت‌ها برای روش [17] است. همین طور دیده می‌شود که با افزایش مرتبه مدل، برای ویدئوی با سطح فعالیت بالا، مقدار این نسبت کاهش می‌یابد، زیرا رابطه بین فریم‌های دورتر در این نوع از شات‌ها کاهش می‌یابد.



(الف) نسبت APEr برای مراتب مختلف مدل AR.



(ب) نسبت APEr به ازای مقادیر متفاوت پارامتر فراموشی.

شکل ٥-٤- نسبت مقدار APE در مرز شات به میانگین APE در طول شات.

(الف) دنباله ویدئویی نمونه‌برداری شده زمانی.

(ب) نتایج انتخاب فریم‌های کلیدی بر اساس روش ارائه‌شده.

(ج) استخراج فریم‌های کلیدی بر اساس روش هیستوگرام رنگ.

شکل ۵-۵- مقایسه نتایج انتخاب فریم‌های کلیدی.

همچنین، این نسبت را برای مقادیر مختلف فاکتور فراموشی در الگوریتم RLS در شکل ۵-٤-ب محاسبه نموده‌ایم. این نسبت به طور مستقیم به فاکتور فراموشی در شات‌های با سطح فعالیت پایین وابسته است، زیرا تاریخچه روند تغییر پارامترها به تخمین صحیح پارامتر بعدی در حالات فعالیت کم کمک می‌نماید.

یک مثال از استخراج فریم کلیدی در شکل ۵-۵ آمده است. همانطور که در شکل دیده می‌شود، فریم‌های کلیدی استخراج‌شده با روش ارائه‌شده (شکل ۵-۵-ب) نمایندگان مناسبی از محتوای ویدئو هستند.

جدول ۵-۱- ارزیابی نتایج تعیین مرز شات، ویدئوهای تست از مجموعه [89] انتخاب شده‌اند: تعداد فریم‌های تست: ۱٦٤٥٥، تعداد مرز شات‌ها: ۸۵

روش ارائه‌شده	دقت	**٪۸۹,۵**
	بازخوانی	**٪۹۷,۷**
روش ارائه‌شده در [17]	دقت	٪٦٥,٤
	بازخوانی	٪۹۰,٤

جدول ۵-۲- نتایج تست ادراکی روش ارائه شده.

ویدئو	# فریم‌ها	# فریم‌های کلیدی	# فریم‌های کلیدی ناصحیح	# فریم‌های کلیدی انتخاب‌نشده[4]
1	۱۳۹٦	۵۹	٦	۲
2	۱۷۸۹	۷۳	۷	۵

همانطور که قبلا نیز اشاره شده است، پارامترهای رنگ فضای HSV در بسیاری از کارهای تحقیقاتی استفاده شده‌اند [105,96]. ما روش خود را با این روش مقایسه می‌نماییم. روش ارائه‌شده، از روش HSV بهتر عمل می‌کند زیرا پارامترهای مکانی موجک می‌تواند جزئیات مفیدتر از رنگ برای سیستم بینایی انسان را به‌کار برد. یک مثال در شکل ۵-۵-ج از استخراج فریم کلیدی با روش [17] آمده است که برخی از فریم های ویدئو در شکل ۵-۵-الف (ردیف چهارم) نماینده‌ای در بین فریم‌های کلیدی ندارند. در حالت استفاده از پارامترهای موجک، انتخاب فریم‌های کلیدی با دقت بیشتری انجام می‌گیرد؛ زیرا فضای موجک دارای تقریب بهتری از ساختار بینایی انسان است [96, 97]. از آنجا که روش ارزیابی استانداردی برای ارزیابی نتایج استخراج فریم‌های کلیدی وجود ندارد، علاوه بر مقایسه بالا با ویژگی‌های HSV، از یک آزمون ادراکی نیز استفاده نموده‌ایم. از ۵ فرد که دارای اطلاعاتی از روش‌های پردازش ویدئو هستند، خواسته شده - است، که فریم‌های کلیدی انتخاب شده را به دو رده صحیح و ناصحیح تقسیم نمایند که نتایج در جدول ۵-۲ آمده است.

۵-۵- جمع‌بندی

در این بخش از رساله، مدل AR برای تحلیل تغییرات زمانی ویژگی‌های مکانی موجک دو بعدی ویدئوهای طبیعی، به کار رفته است. نتایج برای کاربرد میانی انتخاب مرز شات و استخراج فریم‌های کلیدی استفاده شده است. مدل AR به خوبی روابط زمانی بین پارامترهای مکانی دنباله سیگنال ویدئو را بیان می-نماید. با استفاده از خطای تخمین AR به عنوان معیاری برای بیان میزان تغییرات بین فریمی، مرز شات‌ها و فریم‌های کلیدی استخراج می‌شوند. این روش که خصوصیات مکانی و زمانی ویدئو را در نظر می‌گیرد، برای تحلیل و مدل‌سازی ویدئو روش مناسبی می‌باشد.

[4] Missing

فصل ششم

تحلیل پارامترهای مکانی- زمانی ویدئو در حوزه تبدیل موجک

مقدمه

تبدیل سه‌بعدی موجک

خواص آماری تبدیل موجک

تخمین اطلاعات متقابل و آنالیز فعالیت در ویدئو

تحلیل فعالیت بر اساس خواص آماری تبدیل سه‌بعدی موجک

جمع‌بندی

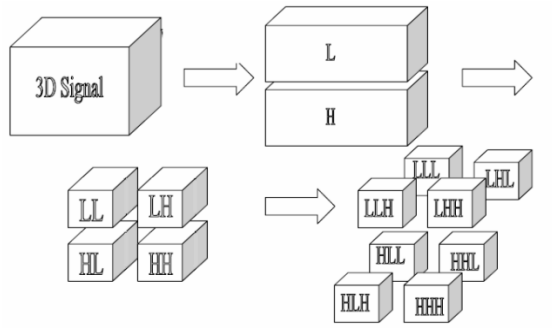
٦-١- مقدمه

در این بخش از رساله، مهم‌ترین نوآوری ما ارائه روش نوینی برای مدل‌سازی و تحلیل ویدئوهای طبیعی به کمک خواص آماری تبدیل سه‌بعدی موجک است. بر اساس بررسی انجام‌شده بر روی مقالات منتشره تا کنون، این اولین باری است که خواص آماری حاشیه‌ای و توأم موجک سه‌بعدی برای تحلیل و مدل‌سازی ویدئو مورد استفاده قرار می‌گیرد. سیگنال ویدئو از حوزه دارای افزونگی بالای مکانی-زمانی با این تبدیل به حوزه تبدیل موجک که افزونگی کمی دارد منتقل می‌شود که در این حوزه سیگنال ویدئو با تعداد ضرایب کمتری قابل بیان است. تبدیل سه‌بعدی موجک برای تشخیص تغییرات شات در دنباله ویدئویی در [109] به کار رفته‌است. ابتدا، تبدیل دوبعدی موجک برای هر فریم محاسبه می‌شود. سپس، تبدیل یک‌بعدی موجک به دنباله زمانی با طول ٨-فریم ضرایب تبدیل اعمال می‌گردد و همبستگی بین ضرایب فریم‌های مجاور و سه ویژگی ساده‌تر محاسبه می‌شوند. این ویژگی‌ها برای تشخیص تغییر شات به کار رفته‌اند.

مقایسه با روش‌های دیگر موجود، نشان می‌دهد که پارامترهای استخراج‌شده از خصوصیات آماری با روش ارائه‌شده، نمایندگان مناسبی برای تفسیر محتوای ویدئو بر اساس درک انسانی هستند. برای این کار روش خود را در دو کاربرد مورد ارزیابی قرار داده‌ایم: گروه‌بندی میزان فعالیت ویدئو و تشخیص رفتار انسان. برای تشخیص رفتار انسان، روش ارائه‌شده با دقت بالای ٩٣,٤٪ در کلاس‌بندی رفتارهای انسانی دادگان KTH [79]، بهتر از روش‌های موجود عمل می‌نماید. همچنین، ما تعریف جدیدی برای سطح فعالیت در ویدئو ارائه می‌کنیم. فعالیت‌ها به دو دسته حرکات سریع و آهسته، بسته به سرعت تغییرات در حوزه زمان و به دو دسته محلی و کلی بسته به درصد درگیری فریم تقسیم می‌شوند. بنابراین، چهار سطح فعالیت خواهیم داشت. سپس پارامترهای زمانی-مکانی مستخرج از توزیع‌های توأم برای به دست آوردن اطلاعات درباره سطح فعالیت در ویدئوها استفاده می‌شوند و با دقت ٨٧,٣٪ کلاس‌بندی می‌گردند.

٦-٢- تبدیل سه‌بعدی موجک

تبدیل فوریه سیگنال را به مولفه‌های فرکانسی‌اش تجزیه می‌نماید، در حالی‌که تبدیل کسینوسی تخمین بهتری از سیگنال با تعداد ضرایب کمتری ارائه می‌دهد. تبدیل فوریه برای سیگنال‌های ایستان مناسب است ولی برای حالات غیرایستان مناسب نیست و اطلاعات کلی درباره سیگنال ارائه می‌دهد که در بسیاری از کاربردهای پردازش سیگنال کافی نیست[110] . تبدیل موجک یک سیگنال را به صورت مجموعه‌ای از بردارهای پایه بیان می‌کند [111]. این تبدیل بر خلاف تبدیل کسینوسی می‌تواند به بلوک‌های سیگنال با ابعاد بزرگ اعمال شود و بنابراین مشکل بلوکی‌شدن را ندارد. این تبدیل نمایش تنکی از سیگنال را خصوصاً برای حالت تک‌بعدی ایجاد می‌نماید. همچنین دارای نسبت فشرده‌سازی بالاتری در قیاس با تبدیل کسینوسی است و از لحاظ ساختاری نشان داده شده است که تطابق خوبی با سیستم ادراکی انسان دارد [96, 97]. تبدیل موجک در بسیاری از کاربردهای پردازش سیگنال مانند پردازش صوت، مهندسی پزشکی، حذف نویز از تصویر، تفسیر تصویر، نهان‌نگاری تصویر/ویدئو و پردازش ویدئو استفاده شده است.



شکل ٦-١. پیاده‌سازی یک سطح تبدیل سه‌بعدی موجک. بازسازی‌شده از [113].

تبدیل موجک با کمک توابع تحلیل عمود یکه[1] که از روی توابع پایه تبدیل موجک ایجاد می‌گردند، یک نمایش چنددقتی از سیگنال ارائه می‌دهد [96, 112]. یکی از مزایای این تبدیل توانایی آن در توصیف‌کردن ارتباط زمانی-مکانی بین اجزای سیگنال در تحلیل ما است که نکته مهمی در این تحلیل به حساب می‌آید. موجک سه‌بعدی طیف ویدئو را به زیرباندهای چندسطحی برای ابعاد مکانی-زمانی و به زیرباندهای جهتی برای بعد مکانی –عمودی، افقی، قطری- تقسیم می‌نماید. این تبدیل جدایی‌پذیر است و تجزیه با عبور از میان یک کانال بانک فیلتر سه‌بعدی انجام می‌گیرد. هر بانک فیلتر سه‌بعدی را می‌توان به صورت ضرب سه بانک فیلتر یک‌بعدی درنظر گرفت.

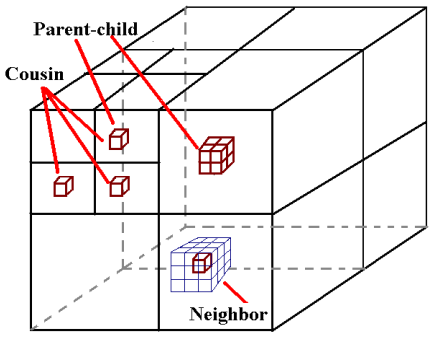
در هر سطح تبدیل، ٨ زیر باند تشکیل می‌شود. تعداد کل ضرایب زیرباندها برابر با ابعاد سیگنال اولیه است و هر بار زیر باند LLL که حاوی تقریبی از سیگنال اصلی است کوچکتر می‌شود. تجزیه مرحله

[1] Orthonormal

بعدی فقط بر روی پایین‌ترین زیرباند انجام می‌شود [114]. پیاده‌سازی تبدیل سه‌بعدی موجک در شکل ٦-١ نشان داده شده است، که L و H به ترتیب بیان‌گر زیرباندهای پایینی و بالایی هستند.

٦-٢-١- روابط ضرایب تبدیل سه‌بعدی موجک

هر زیر باند تبدیل سه‌بعدی موجک نمایان‌گر نسخه نمونه‌برداری شده از فیلترشده سیگنال اصلی است، بنابراین روابطی بین ضرایب زیرباندهای مختلف با قسمت مرتبط از سیگنال اصلی وجود دارد [11, 96, 115]. ضرایب در زیرباندهای سطوح مختلف و جهت یکسان، دارای رابطه والد-فرزندی هستند. ضرایب در یک سطح تبدیل و مکان با جهات مختلف عموزاده[2] هستند، و ضرایب مجاور در یک زیرباند با هم رابطه همسایگی دارند.



شکل ٦-٢. روابط ضرایب تبدیل سه‌بعدی موجک، برداشت از تبدیل دوبعدی [11].

روابط ضرایب تبدیل سه‌بعدی موجک در شکل ٦-٢ نشان داده شده‌اند. هر ضریب X در هر سطح تبدیل دارای ٦ عموزاده است که با CX نمایش داده می‌شود. این ضریب دارای ٢٦ همسایه با نماد NX، در همان زیرباند است و در جهت یکسان و سطح تبدیل بعدی ٨ فرزند دارد. بنابراین هر ضریب X در سطح تبدیل پایین‌تر یک والد PX دارد.

ضریب موجک $w_{l,\boldsymbol{P}}^{o}$ را در جهت o که $o = 1,2,\dots,7$ به ترتیب برای زیرباندهای $LLH, LHL, LHH, HLL, HLH, HHL, HHH$ است و l بیان‌گر سطح تبدیل است و $l = 1,\dots,lev$ و مکان $\boldsymbol{P}$=(x,y,t) است، را در نظر بگیرید. پس عموزاده‌های این ضریب در مکان و سطح تبدیل یکسان و جهات مختلف $\{w_{l,\boldsymbol{P}}^{1}, w_{l,\boldsymbol{P}}^{2}, \dots, w_{l,\boldsymbol{P}}^{7}\}$ هستند. همچنین ضرایب $\cong (i,j,k)$ $\{w_{l,(x+i,y+j,t+k)}^{o}, i = -1,0,+1, j = -1,0,+1, k = -1,0,+1,$ $\boldsymbol{0}\}$ نمایان‌گر همسایه‌های ضریب مذکور در مکان $\boldsymbol{P}$=(x,y,t) و جهت o سطح تبدیل l هستند. نهایتاً ضریب $w_{l,\boldsymbol{P}}^{o}$ والد ضرایب $\{w_{l+1,(2x+i,2y+j,2t+k)}^{o}, i = 0,1, j = 0,1, k = 0,1\}$ در سطح تبدیل دقیق‌تر و جهت یکسان است.

[2] Cousin

٦-٣- خواص آماری تبدیل موجک

٦-٣-١- ویدئوهای مورد استفاده در بررسی خواص آماری

برای مطالعه خواص تبدیل سه‌بعدی موجک و میزان فعالیت در ویدئو، ما از آزمون‌های بسیاری بر روی ویدئوهای متفاوت با میزان فعالیت و بافت متفاوت استفاده نموده‌ایم. ویدئو ها از مجموعه آزمون-های دادگان [89] Hollywood2، دادگان [88] TRECVID، مجموعه SFU Video Library [91]، و دادگان[90] The Open Video Project انتخاب شده‌اند. ما به تصادف ۷۵۰ دنباله ویدئویی از این مجموعه‌ها برای ارزیابی مطالعاتی که بر اساس خواص آماری تبدیل سه‌بعدی ویدئو انجام می-گیرد، انتخاب نمودیم. ویدئوهای انتخاب شده در مجموع دارای بیش از ۱۲۰۰۰۰ فریم با خصوصیات متفاوت هستند. مجموعه ویدئوهای مورد آزمون به صورت گسترده در تحقیقات استفاده می‌شوند. ابعاد هر فریم ویدئو بین QCIF (176x144) و CIF (352x288) می‌باشد و طول زمانی آنها بین ۷۵ تا ۳۰۰ فریم است. نرخ نمونه‌برداری زمانی این ویدئوها ۱۵، ٢٤ و ۲۹ فریم در ثانیه است. هر نمونه ویدئویی دارای یک شات است. این ویدئوها در مکان‌های مختلف فیلمبرداری شده‌اند –محوطه داخل و خارج ساختمان– و دارای مشخصات متنوعی هستند. اغلب ویدئوهای آزمون بسیار پویا[3] در هر دو بعد مکان و زمان می‌باشند در حالی که نمونه‌های ایستا[4] نیز وجود دارند.

ما از فیلترهای معروف موجک –Haar، Daubechies و Symlets– برای تجزیه ویدئو استفاده نمودیم و تعداد سطوح تجزیه ۳، ٤ و ٥ سطح در نظر گرفته شد. برای کلاس‌بندی میزان فعالیت در ویدئو، دو میدان برای فعالیت‌ها در نظر گرفته می‌شود:

- حوزه زمانی: که شامل سرعت تغییرا ت در بعد زمان است و می‌تواند آرام یا سریع باشد.
- حوزه مکانی: که در صد سطح فریم که درگیر تغییرات است را در نظر می‌گیرد و می‌تواند کلی یا محلی باشد.

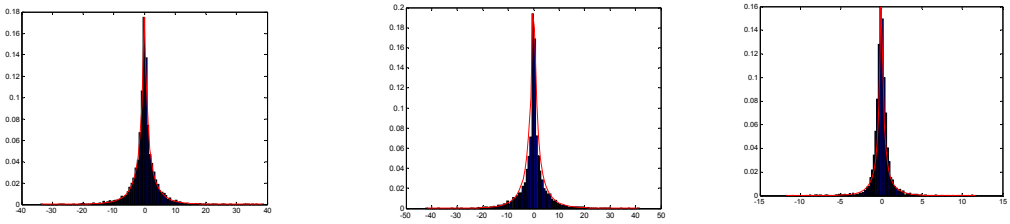
سپس خواص آماری تبدیل سه‌بعدی موجک را در بخش‌های بعدی این فصل مورد تحلیل قرار می-دهیم.

٦-٣-٢- خواص آماری حاشیه‌ای

تبدیل سه‌بعدی موجک با فیلترهای متفاوت و سطوح تبدیل مختلف بر روی ویدئوهای آزمون اعمال شده است و خواص آماری حاشیه‌ای زیرباندهای حاصل مورد مطالعه قرار گرفته‌اند. بطور خاص ۷۵۰ ویدئو که در بخش ٦-٢-١ در باره آنها صحبت شد، درنظر گرفته شده‌اند و پارامترهای GGD تمام زیرباندهای آنها استخراج و منحنی‌های GGD با توزیع‌های ضرایب زیرباندها مقایسه شده‌اند. نتایج

[3] Dynamic
[4] Static

نشان می‌دهد که ۱۰۰٪ منحنی‌ها بطور کامل به توزیع‌ها منطبق شده‌اند و GGD می‌تواند تخمین بسیار مناسبی برای هیستوگرام‌های حاشیه‌ای باشد.



$\alpha = .7588, \beta = .5819,$
$Kurtosis = 15.6221$
$LHH - Subband$

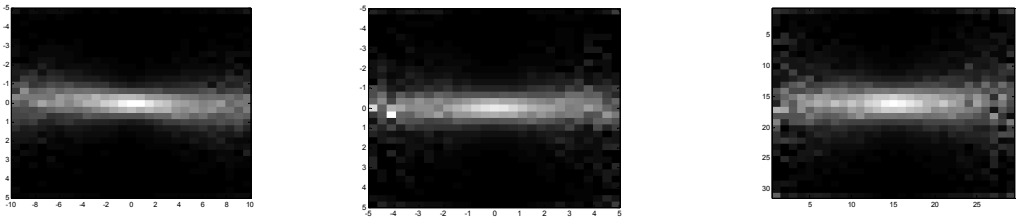
$\alpha = .8591, \beta = .5788,$
$Kurtosis = 11.8763$
$HHH - Subband$

$\alpha = .1203, \beta = .5002,$
$Kurtosis = 25.7525$
$HHL - Subband$

شکل ٦-٣. هیستوگرام‌های حاشیه‌ای زیرباندهای بالاترین سطوح موجک و منحنی‌های منطبق‌شده به آنها.

همچنین مقادیر kurtosis – به عنوان معیاری از تیزی[5] تابع توزیع احتمال متغیرهای تصادفی زمان واقعی- برای این منحنی‌ها از بیش از ۲۱۰۰۰ زیرباند محاسبه شده است و دیده شده است که تمامی این مقادیر بالای ۹ هستند که این بیان‌گر خاصیت غیرگوسی این توزیع‌ها است. سه نمونه از این هیستوگرام-ها و منحنی‌های تخمین شده GGD در شکل ٦-٣ آمده است. همانطور که در شکل دیده می‌شود، هر هیستوگرام دارای یک پیک در صفر است و مقدار آن به سرعت با دور شدن از صفر کاهش می‌یابد. این بدان معنی است که بیشتر ضرایب در زیرباندها صفر یا نزدیک به صفر هستند، پس تبدیل سه‌بعدی موجک بسیار تنک است. مقادیر α و β و kurtosis در این شکل آمده‌اند. مقادیر kurtosis برای این سه هیستوگرام برابر ۱۵,٦۲۲۱، ۱۱,۸۷٦۳ و ۲۵,۷۵۲۵ است که بیان‌گر خاصیت شدید غیرگوسی این توزیع‌هاست (kurtosis برای توزیع‌های گوسی برابر ۳ است). با تخمین دو پارامتر GGD از هر زیرباند، اطلاعات کافی برای بیان توزیع آن زیرباند فراهم می‌شود و توزیع حاشیه‌ای زیرباند مربوطه به دقت بازسازی می‌گردد، در صورتی که برای بیان هیستوگرام زیرباند، نیاز به صدها عدد است.

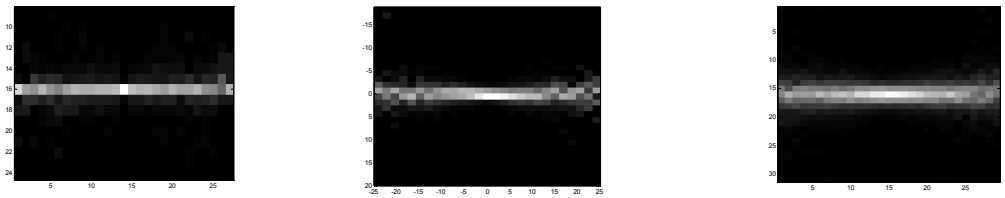
٦-٣-٣- خواص آماری توأم



(ج) (ب) (الف)

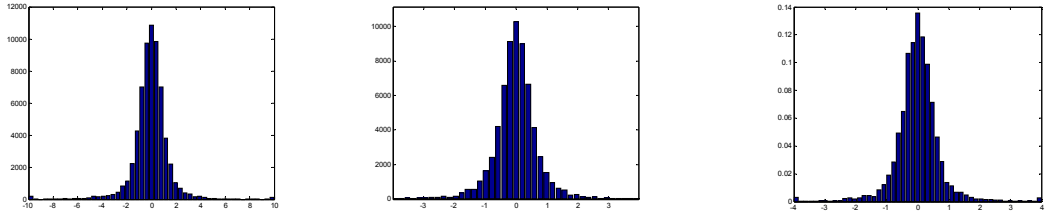
شکل ٦-٤. نمودارهای توزیع شرطی ضرایب مشروط بر (الف) والدین (ب) همسایه‌ها و (ج) عموزاده‌ها. توزیع‌ها فقط برای یکی از همسایه‌ها (همسایه سمت راست در جهت X) و یک عموزاده (ضریب در زیر باند HLH برای ضریب در زیر باند HHH).

[5] Peakness



(ج) (ب) (الف)

شکل ٦-٥. نمودارهای توزیع شرطی ضرایب مشروط بر ضرایب دور (الف) والدین (ب) همسایه‌ها و (ج) عموزاده‌ها. توزیع‌ها فقط برای یکی از همسایه‌ها (همسایه سمت راست در جهت X) و یک عموزاده (ضریب در زیر باند HLH برای ضریب در زیر باند HHH).



kurtosis=7.8338 (ج) **kurtosis= 4.0841 (ب)** **kurtosis=5.8791 (الف)**

شکل ٦-٦. قطع عمودی نمودارهای توزیع توأم (الف) والد، (ب) همسایه و (ج) عموزاده.

خواص آماری توأم تبدیل دوبعدی موجک و کانتورلت بر روی تصاویر دوبعدی در [48, 93] مورد مطالعه قرار گرفته‌اند. در اینجا خواص آماری توأم تبدیل سه‌بعدی بر روی ویدئوهای طبیعی بررسی می‌شود. گرچه تبدیل سه‌بعدی موجک سیگنال ویدئو را به خوبی غیرهمبسته می‌نماید، هنوز وابستگی‌هایی بین ضرایب زیرباندهای مختلف در یک سطح و ضرایب زیرباندهای یکسان در سطوح تبدیل مختلف وجود دارد. الگوریتم‌های پردازش ویدئو می‌توانند بر اساس خواص آماری توأم این تبدیل بهبود یابند. یکی از توزیع‌های آماری توأم ضرایب تبدیل سه‌بعدی موجک، مشروط بر والد، همسایه و عموزاده آنها در شکل ٦-٤ نشان داده شده‌اند.

توزیع‌های شرطی نشان داده شده، دارای شکل پاپیون[6] می‌باشند که واریانس و اندازه ضریب شرطی آنها با هم رابطه دارند، علاوه بر این امیدهای ریاضی شرطی تقریبا صفر هستند. در نتیجه ضرایب تقریباً ناهمبسته‌اند، ولی مستقل نیستند. در شکل ٦-٥ توزیع‌های آماری توأم ضرایب تبدیل سه‌بعدی موجک، مشروط بر والد، همسایه و عموزاده آنها با فاصله دو پیکسل نشان داده شده است. نتایج تایید می‌کنند که این ضرایب از هم مستقل هستند. برای مثال وابستگی بین ضریب و والد، همسایه و عموزاده آن محلی بوده و با افزایش فاصله این وابستگی کاهش می‌یابد. مقادیر kurtosis قطع‌های عمودی نشان‌دهنده خاصیت غیر گوسی توزیع‌های شرطی هستند که در شکل ٦-٦ دیده می‌شوند.

[6] Bow-tie

٦-٤- تخمین اطلاعات متقابل و تحلیل فعالیت در ویدئو

٦-٤-١- پیش‌زمینه

در این بخش، ما از تخمین اطلاعات متقابل[7] (MI) به عنوان یک معیار کمی وابستگی استفاده می‌نماییم. گرچه همبستگی معیار خوبی برای بیان وابستگی در توزیع‌های گوسی است، این مساله برای حالات غیرگوسی صادق نیست [116]، که این شامل مورد بحث ما نیز می‌شود، پس ما از MI استفاده می‌نماییم. محاسبه MI بین دو متغییرتصادفی پیوسته X و Y به شکل زیر است [48, 108] :

$$I(X;Y) = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dydx$$
$$= E_{XY}\{\frac{p(x,y)}{p(x)p(y)}\} = D(p(x,y) \| p(x)p(y)) \qquad (١-٦)$$

که $p(x,y)$ تابع توزیع توأم بین X و Y و $p(x)$ و $p(y)$ توابع توزیع حاشیه‌ای X و Y هستند. $E\{\}$ بیان‌گر امید ریاضی و $D(.)$ نشان‌دهنده فاصله KL هستند. این مقدار نشان‌دهنده مقدار اطلاعات یک متغیر در رابطه با متغیر دیگر است. ما از پایه ٢ برای لگاریتم استفاده می‌نماییم و $I(X;Y)$ برحسب بیت محاسبه می‌شود. مقدار MI میزان اطلاعاتی است که متغیر X درباره متغیر Y انتقال می‌دهد و بالعکس. بنابراین MI متقارن و غیر منفی می‌باشد. علاوه براین MI میزان وابستگی بین دو متغیر را نشان می‌دهد و اگر X و Y مستقل باشند، برابر صفر می‌شود. از طرف دیگر با افزایش وابستگی بین متغییرها مقدار MI نیز افزایش می‌یابد.

برای تخمین MI، هیستوگرام‌های توزیع مطابق فرمول زیر استفاده می‌شوند [48, 117]:

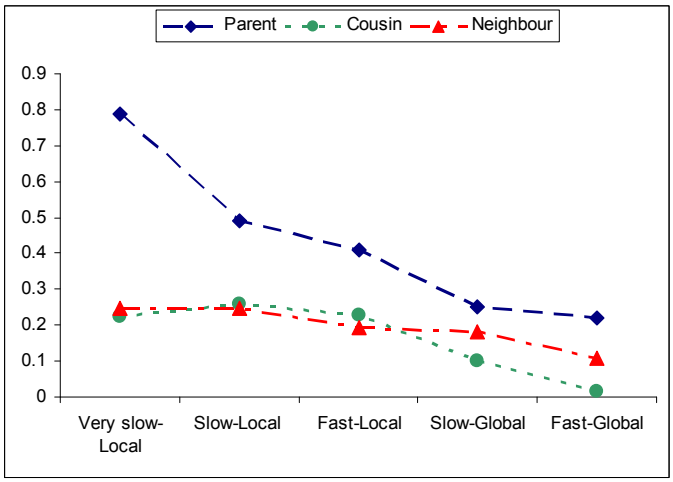
$$\hat{I}(X;Y) = \sum_{i,j} \frac{h_{ij}}{N} \log \frac{h_{ij}N}{h_i h_j} - \frac{(J-1)(K-1)}{2N} \qquad (٢-٦)$$

که h_{ij} مقدار سلول (i,j) از هیستوگرام توأم است و $h_i = \sum_j h_{ij}$ و $h_j = \sum_i h_{ij}$ برابر هیستوگرام‌های حاشیه‌ای هستند، N تعداد همه ضرایب و J و K تعداد بین‌ها در جهات X و Y می‌باشند. جمله دوم رابطه بالا یک بایاس جزئی است و سعی در کاهش خطای تخمین دارد، ولی نمی‌تواند این خطا را بطور کلی کاهش دهد. در نتیجه رابطه بالا در واقع یک کران پایین برای MI است [48, 117]. در [48] مقادیر J و K به صورت تجربی برابر مقادیر پایین انتخاب شده‌اند و خطا در رابطه بالا با افزایش تعداد متغییرها افزایش می‌یابد:

[7] Mutual Information

$$J = K = round(\frac{N}{3000}) + 1 \qquad (۳-٦)$$

٦-٤-٢- نتایج و تحلیل



شکل ٦-٧. تخمین MI بین ضریب X و والد(PX)، همسایه (NX)، عموزاده (CX) (تبدیل موجک سه‌سطحی و فیلتر موجک Daubechies).

ما از تخمین‌های MI برای بررسی وابستگی بین ضریب و والد، همسایه و عموزاده آن در تبدیل سه-بعدی موجک ویدئوهای طبیعی استفاده نمودیم و سپس روابط بین میزان فعالیت در ویدئو و مقدار MI را تفسیر نموده‌ایم. دادگان ویدئوی مورد استفاده در این قسمت، با دادگان مورد استفاده در بخش قبلی یکسان است. برای این کار تخمین‌ها در سه مرحله مورد مطالعه قرار گرفته‌اند. در مرحله اول، تخمین MI بین ضریب در بالاترین سطح تبدیل و والد، همسایه و عموزاده آن محاسبه و برخی از نتایج در جدول ٦-١ و شکل ٦-٧ نمایش داده شده است. در این جدول تخمین‌های MI بین ضریب و 6 عموزاده آن و MI بین ضریب و ٢٦ همسایه آن محاسبه و میانگین گرفته شده‌اند. نتایج می‌توانند منتج به تفاسیر زیر شوند:

جدول ٦-١. تخمین MI بین ضریب X و والد(PX)، همسایه (NX)، عموزاده (CX) (تبدیل موجک سه‌سطحی و فیلتر موجک Daubechies).

	فعالیت بسیار کند-جزئی	فعالیت کند-جزئی	فعالیت سریع-جزئی	فعالیت کند-کلی	فعالیت سریع-کلی
$I(X;PX)$	۰,۷۸٦۳	۰,٤۸۸۷	۰,٤۱	۰,۲۵۱٦	۰,۲۲۱۷
$I(X;CX)$	۰,۲۲۵٤	۰,۲۵۹	۰,۲۲۷۷	۰,۱۰۰۳	۰,۰۱۵۲
$I(X;NX)$	۰,۲۳٦۳	۰,۲٤٦۷	۰,۱۹۲۷	۰,۱۸۲۵	۰,۱۰۹٦

- همان‌طور که از جدول ٦-١ و شکل ٦-٧ دیده می‌شود، مقدار MI زیاد است که این تاییدکننده نتایج به‌دست آمده در بخش ٦-٢-٣ می‌باشد، که نتایج کیفی به دست آمده از هیستوگرام‌های توأم بیان‌گر وجود وابستگی بین ضرایب و والد، همسایه و عموزاده‌شان بود.
- MI بین ضرایب و عموزاده آنها و MI بین ضرایب و همسایه‌شان کمترین مقدار را برای ویدئوهای با سطح فعالیت بالا دارد. این تاییدکننده وجود وابستگی کم بین زیرباندهای مختلف در یک سطح تبدیل است.
- MI بین ضریب و والد آن همواره بیشترین مقدار را دارد. یعنی وابستگی مهم بین والد و ضریب فرزند است و با افزایش سطوح تبدیل، اطلاعات دقیق‌تر و ظریف‌تری از سیگنال استخراج می‌گردد.
- مقدار MI با کاهش فعالیت در ویدئو افزایش می‌یابد. بنابراین وابستگی بین ضریب و والد، همسایه و عموزاده آن با افزایش سطح فعالیت در ویدئو کاهش می‌یابد.
- تغییرات MI بین ضرایب و همسایه آنها براساس تغییرات میزان فعالیت از تغییرات مقدار MI بین ضرایب و والد یا عموزاده آنها کمتر است.

در مرحله دوم، MI برای انواع مختلف تبدیل موجک تخمین زده شده است (جدول ٦-٢). از نتایج چنین استنتاج می‌شود که MI به نوع فیلتر وابسته است. برای مثال با تعویض 'Haar' با 'Daubechies' میزان MI و در نتیجه وابستگی بین ضرایب کاهش می‌یابد.

جدول ٦-٢. تخمین MI بین ضریب X و والد(PX)، همسایه (NX)، عموزاده (CX)، برای فیلترهای موجک مختلف (ویدئو با فعالیت بالا و تبدیل موجک سه‌سطحی).

Symlet	Daubechies	Haar	
٠,١٢٣٣	٠,٢٢١٧	٠,٢٤٥٤	$I(X;PX)$
٠,٠١٦٨	٠,٠١٥٢	٠,٠٢٣٨	$I(X;CX)$
٠,١٦٤٥	٠,١٠٩٦	٠,٢٠٢٣	$I(X;NX)$

در سومین و آخرین مرحله، MI برای تعداد مختلف سطوح تبدیل موجک تخمین زده شده است و نتایج در جدول ٦-٣ آمده‌اند. همانطور که دیده می‌شود، تخمین‌های بین ضریب و همسایه/عموزاده به تعداد سطوح تبدیل وابستگی زیادی ندارد.

جدول ٦-٣. تخمین MI بین ضریب X و والد(PX)، همسایه (NX)، عموزاده (CX)، برای تعداد سطوح تبدیل مختلف (ویدئو با فعالیت بالا و فیلتر موجک Daubechies).

تعداد سطوح موجک	٢	٣	٤
$I(X;PX)$	٠,١٩٦٠	٠,٢٢١٧	٠,٢٧٩٧
$I(X;CX)$	٠,٠١٥٢	٠,٠١٥٢	٠,٠١٥٢
$I(X;NX)$	٠,١٠٩٦	٠,١٠٩٦	٠,١٠٩٦

٦-٥- تحلیل فعالیت بر اساس خواص آماری تبدیل سه‌بعدی موجک

٦-٥-١- خواص آماری توأم و منحنی‌های kurtosis

در این بخش از نمونه‌های ویدئویی که در بخش ٦-٢-١ درباره آنها توضیح داده شد، استفاده شده است و تبدیل سه‌بعدی موجک به تمام دنباله‌های ویدئویی اعمال گشته است. ابعاد زمانی و مکانی به - صورت جداگانه برای تحلیل سطح فعالیت در این ویدئوها بررسی شده‌اند. تغییرات به دو دسته آرام و سریع، بسته به سرعت تغییرات در ویدئو در بعد زمان، و دو دسته محلی و کلی، با توجه به درصد درگیری سطح فریم دسته‌بندی شده‌اند.

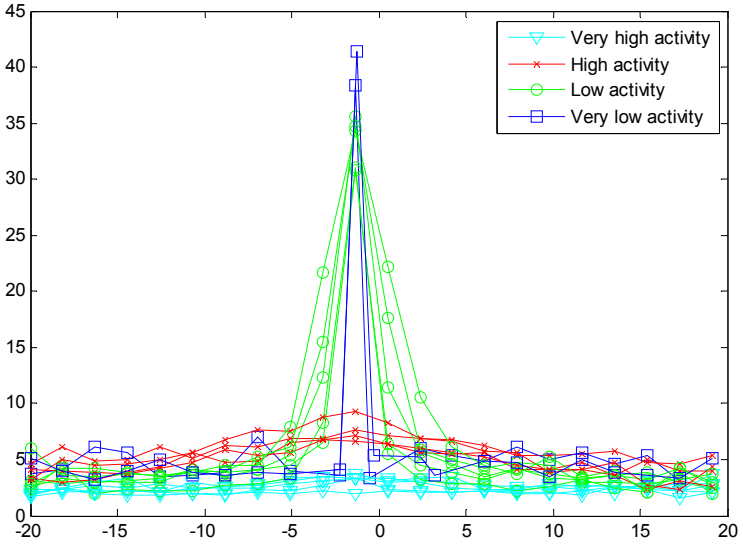
خواص آماری توأم تبدیل سه‌بعدی موجک در بخش ٦-٢-٣ بررسی شده است. توزیع‌های ضرایب مشروط بر والدین آنها برای دقیق‌ترین سطح تبدیل موجک در اینجا برای کلاس‌بندی ویدئو با توجه به سطح فعالیت آنها مورد استفاده قرار گرفته است. برای تشکیل منحنی‌های kurtosis توزیع‌های مشروط ضرایب نسبت به والدین آنها محاسبه شده است و مقادیر kurtosis هیستوگرام‌های قطع عمودی این توزیع‌ها محاسبه شده است. این مقادیر kurtosis تشکیل منحنی kurtosis را می‌دهند. این منحنی‌ها برای ٧ زیرباند بالاترین سطح موجک هر نمونه ویدئویی محاسبه می‌شوند و مقادیر متناظر آنها با هم جمع و در نهایت میانگین‌گیری می‌شوند و در نهایت یک منحنی kurtosis برای هر نمونه ویدئویی خواهیم داشت. منحنی‌های kurtosis هر نمونه ویدئویی استخراج شده است و چند نمونه از آنها در شکل ٦-٨ آورده شده‌اند. همان‌طور که در شکل دیده می‌شود، چهار نوع منحنی kurtosis داریم، که هر یک با یک سطح فعالیت ویدئویی متناظر است. در نتیجه می‌توان ویدئوها را با توجه به این منحنی‌ها به چهار دسته تقسیم نمود:

- گروه اول: ویدئوهای با سطح فعالیت بسیار بالا- ویدئوهایی که شامل ظهور یک شیء، حرکت سریع کلی، تغییرات بسیار سریع یا نویز فراوان هستند. منحنی kurtosis این دسته تقریبا هموار است و مقدار kurtosis همواره زیر ٥ می‌باشد. بنابراین، شکل منحنی تقریبا صاف است و پیک واضحی در صفر ندارد و مقادیر kurtosis با دور شدن از صفر در هر دو جهت با افزایش قدر مطلق مقدار والد کاهش می‌یابد تا به مقدار ٣ برسد.
- گروه دوم: ویدئوهای با سطح فعالیت بالا: ویدئوهایی با تغییرات آرام کلی. منحنی kurtosis در این دسته زیر ١٠ و بالای ٥ در صفر است. یک ماکزیمم در صفر وجود دارد و منحنی با افزایش قدر مطلق والد به آرامی به سمت ٣ پیش می‌رود.
- گروه سوم: ویدئوهای با سطح فعالیت پایین-ویدئوهایی با تغییرات آرام محلی. در اینجا ماکزیمم منحنی در صفر بین ٣٠ تا ٤٠ می‌باشد. و منحنی بعد از گذر از صفر در هر دو جهت به سرعت

کاهش می‌یابد تا به ۳ (گوسی) برسد، بنابراین منحنی در اطراف صفر تند و در باقی حالات تقریبا صاف است.

- گروه چهارم: ویدئوهای با فعالیت بسیار پایین- ویدئوهای با تغییرات محلی و بسیار آرام. شکل منحنی این گروه مانند گروه سه است، ولی این منحنی در صفر بسیار تیزتر است. منحنی یک پیک در صفر برابر حدود ٤٠ دارد و به سرعت به سمت سه کاهش می‌یابد.

منحنی‌های شکل ٦-٨ نشان می‌دهند که با افزایش سطح فعالیت در ویدئو مقادیر kurtosis کاهش می‌یابند و توزیع‌های شرطی ضرایب نسبت به والدشان به توزیع گوسی نزدیک‌تر می‌شوند. همانطور که در بخش ٦-٣ مطرح شد، وابستگی بین ضریب و والد آن با افزایش سطح فعالیت در ویدئو کاهش می‌یابد. بنابراین مقدار kurtosis با کاهش وابستگی و افزایش فعالیت کاهش می‌یابد.



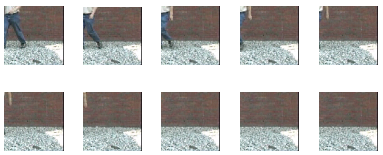
شکل ٦-٨ بردارهای kurtosis چهار کلاس مختلف.

برای دسته‌بندی سطوح فعالیت، توزیع‌های شرطی ضرایب مشروط به والد، همسایه و عموزاده آنها برای ۷ زیر باند دقیق‌ترین سطح تبدیل موجک محاسبه می‌شوند و منحنی‌های kurtosis همان‌طور که پیش از این توضیح داده شد، محاسبه می‌گردند. ما برای یکسان بودن طول بردارهای ویژگی ۹ نقطه[8] برای هر منحنی kurtosis در نظر گرفته‌ایم. به این ترتیب سه منحنی kurtosis با طول ۹ و در کل ۲۷ ویژگی توأم برای هر نمونه ویدئویی خواهیم داشت. برای کلاس‌بندی منحنی‌های kurtosis از طبقه‌بند[9] SVM [118] استفاده شده است که بردار ویژگی توأم را به یکی از چهار کلاس بالا اختصاص دهد. برای این منظور ۵۰۰ دنباله ویدئویی برای آموزش انتخاب و به صورت دستی گروه‌بندی شده‌اند. از

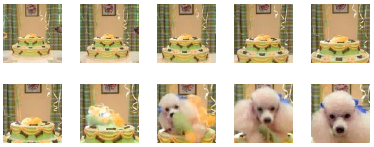
[8] Bin
[9] Classifier

ویدئوهای این مجموعه آزمون بردارهای ویژگی توأم استخراج‌شده و با کمک آنها طبقه‌بند آموزش دیده است. سپس ۵۸ ویدئو برای آزمون انتخاب شده‌اند و با کمک طبقه‌بند، سطح فعالیت آنها تعیین شده است. شکل ٦-۹ چند نمونه از ویدئوها را نشان می‌دهد.

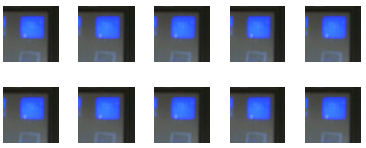
برای ارزیابی الگوریتم گروه‌بندی ارائه‌شده، از آزمون‌های ادراکی استفاده می‌کنیم. ۱۵ فرد غیر آشنا با روش‌های پردازش ویدئو برای گروه‌بندی ۵۸ ویدئوی آزمون به ٤ سطح فعالیت بر اساس توصیه‌نامه ITU-R BT.500-11 استفاده شده‌اند [119]. از این افراد درخواست شده است که ویدئوها را با توجه به میزان تغییرات آنها به چهار دسته تقسیم کنند. آزمون ادراکی بکار رفته یک آزمون تک‌محرک[10] است [120] که در آن پرسشنامه‌ای تشکیل می‌شود و از بیننده درخواست می‌گردد که پس از دیدن هر ویدئو جدول مربوط به آن را تکمیل نماید. افراد تحت آزمایش دارای درک بصری نرمال هستند و اطلاعات و توضیحات لازم قبل از آزمون به آنها داده شده است. ابتدا دنباله ویدئویی به آنها در بالای صفحه نمایش‌گر نشان داده می‌شود و سپس از آنها در خواست می‌گردد که دنباله ویدئویی را با توجه به میزان فعالیت در آن با عدد صحیحی بین ۱ تا ٤ –برای فعالیت‌های زیاد تا کم– نمره دهد. میانگین نتایج این آزمون ادراکی در شکل ٦-۱۰ آمده است.



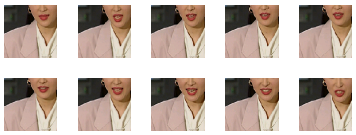
(الف) ویدئو با بافت زیاد –گروه ۲.



(ب) ویدئو با پدیدار شدن شیء– گروه ۱.



(ج) ویدئو با تغییرات بسیار اندک– گروه ٤.

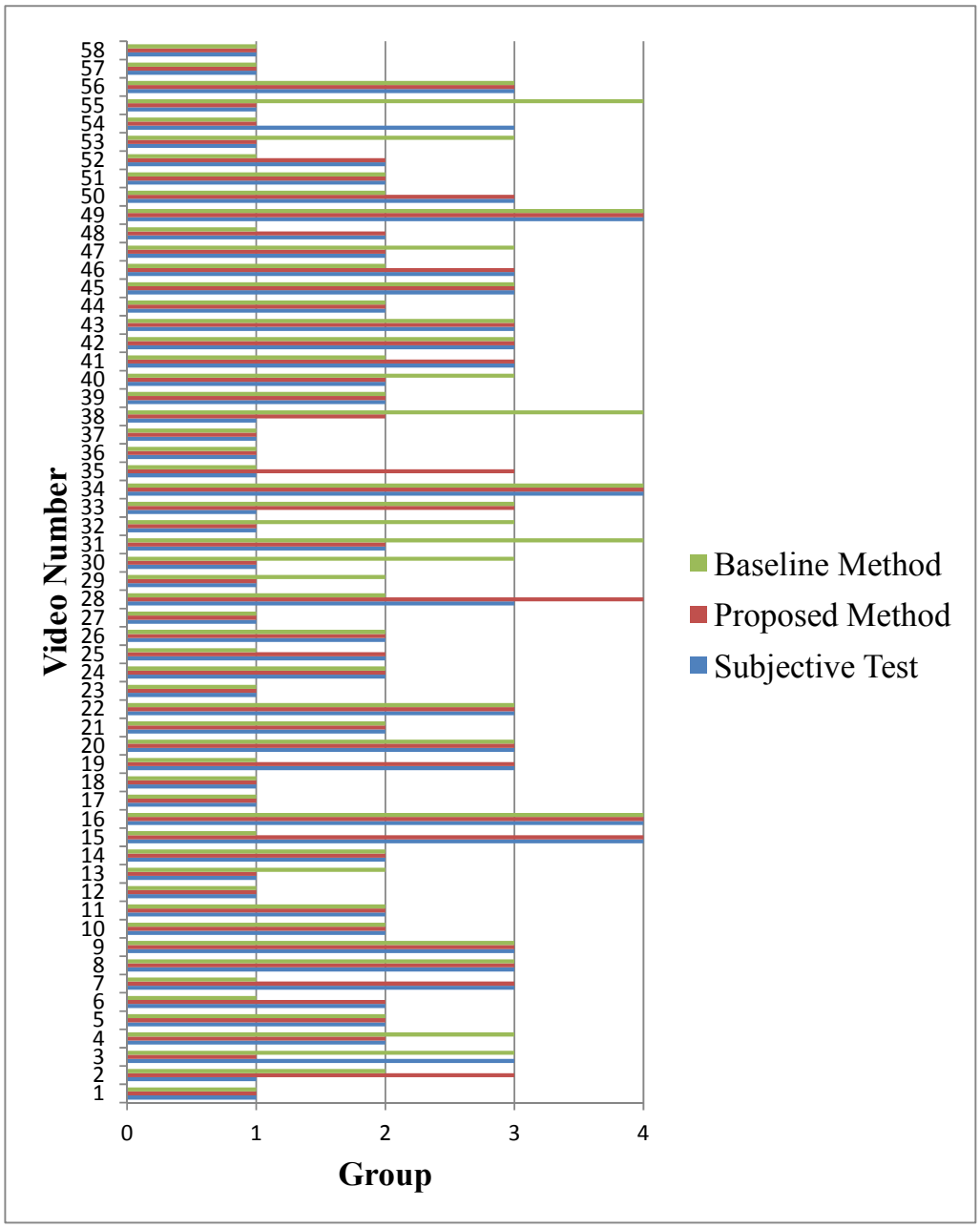


(د) ویدئو با حرکت محلی آهسته– گروه ۳.

شکل ٦-۹. چند نمونه از فریم‌های ویدئویی و گروه‌های تعیین‌شده برای آنها. فریم‌های ویدئو با نرخ ۳ نمونه‌برداری شده‌اند.

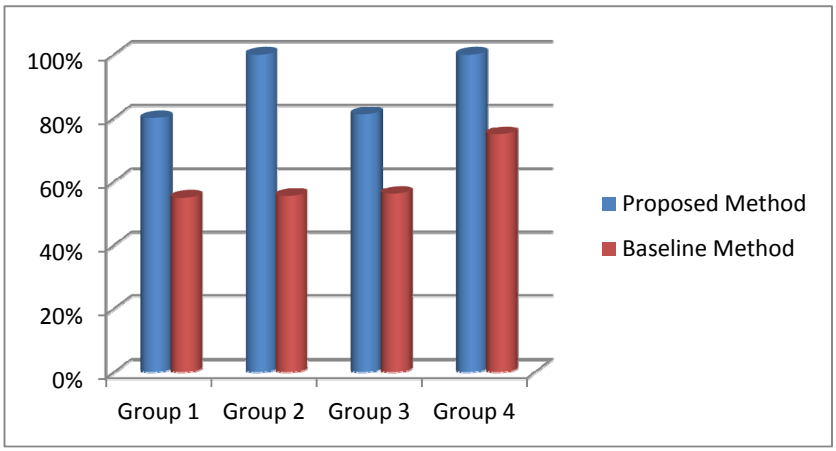
ما همچنین نتایج روش ارائه‌شده را با یک روش پایه مقایسه نموده‌ایم. این روش میانگین و انرژی ضرایب هر زیر باند تبدیل سه‌بعدی موجک را برای تشکیل بردار ویژگی استفاده می‌کند [90]. در [90] از این ویژگی‌ها برای تشخیص رفتار انسان با کمک آزمون ساده خارج‌کردن یک نمونه[11] استفاده شده است. در اینجا باز هم از طبقه‌بند SVM برای آموزش و آزمون این روش مطابق آنچه قبلاً ذکر شد، استفاده می‌شود. نتایج کلاس‌بندی در شکل ٦-۱۰ آمده است.

[10] Stimulus
[11] leave-one-out



شکل ٦-١٠. مقایسه نتایج.

همانطور که در شکل دیده می‌شود، روش ارائه‌شده قابلیت گروه‌بندی دنباله ویدئویی را با دقت ٪۸۷,۳ در مقایسه با کلاس‌بندی ادراکی دارد. همچنین با مقایسه روش ارائه‌شده با روش پایه، این روش به مراتب بهتر عمل می‌نماید. این مساله نشان می‌دهد که پارامترهای آماری توأم حاوی اطلاعات مهم درباره سرعت و مقدار تغییرات در دنباله ویدئویی هستند. دقت گروه‌بندی برای هر گروه در شکل ٦-١١ آمده است.



شکل ٦-١١. دقت گروه‌بندی برای هر کلاس.

٦-٥-٢- تشخیص رفتار انسان با کمک ویژگی‌های آماری حاشیه‌ای و توأم

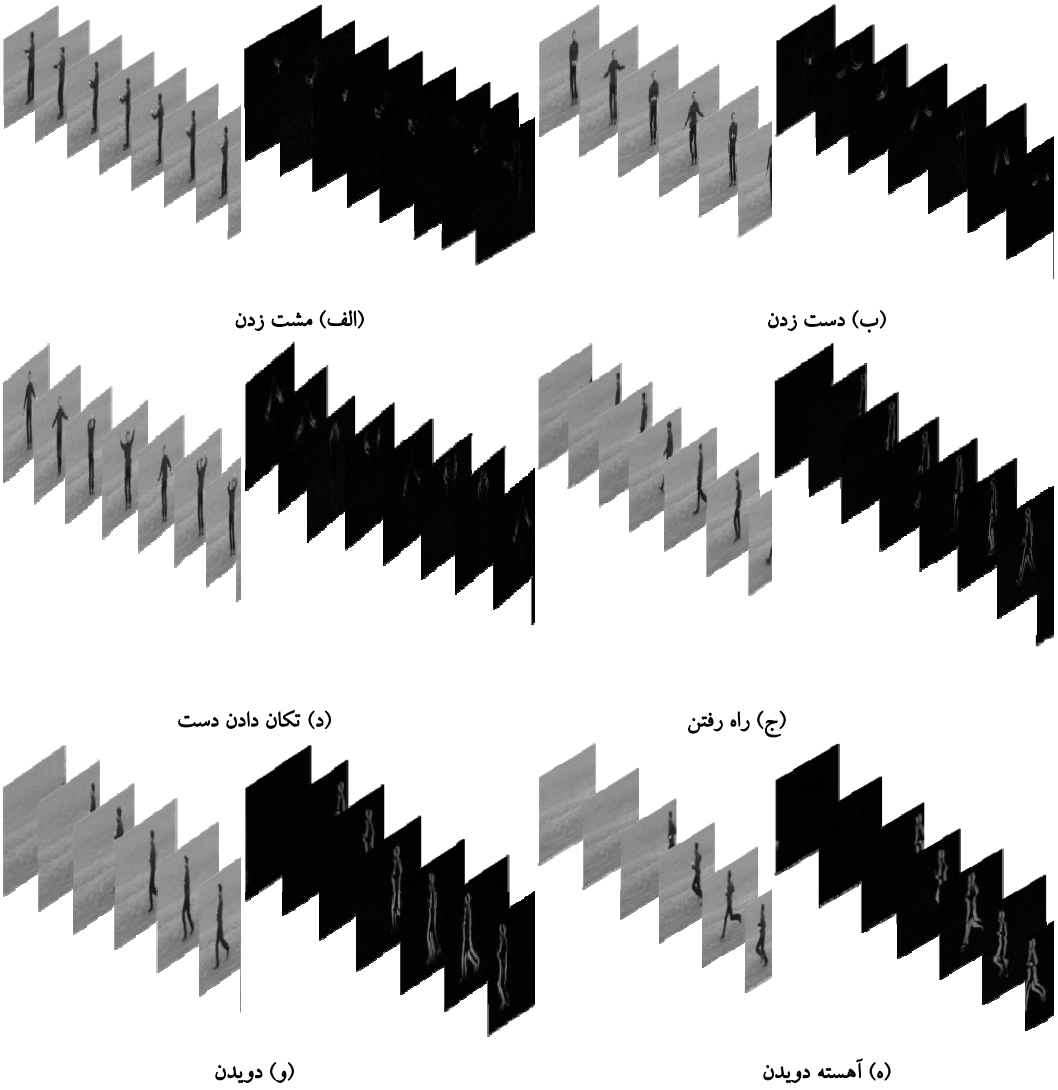
ما از پارامترهای تبدیل سه‌بعدی موجک برای تشخیص رفتار انسان استفاده نموده‌ایم و دادگان رفتار انسان KTH برای ارزیابی روش ارائه‌شده بکار رفته است [79]. این دادگان دارای ٢٣٩١ دنباله ویدئویی است که توسط ٢٥ نفر در ٤ سناریوی مختلف در حال انجام ٦ رفتار شامل مشت زدن، دست زدن، تکان دادن دست، راه رفتن، دویدن آرام و دویدن انجام شده است. اندازه هر فریم ویدئو ١٦٠ در ١٢٠ پیکسل است و دوره زمانی هر دنباله ویدئویی متفاوت است و سرعت نمونه‌برداری زمانی برابر ٢٥ فریم بر ثانیه است. ویدئوها در چهار سناریوی: محیط بیرونی، محیط بیرونی با تغییرات مقیاس، محیط بیرونی با لباس‌های مختلف و محیط داخلی گرفته شده‌اند. نقطه دید دوربین در ویدئوهای مختلف متفاوت است ولی دوربین تقریبا استاتیک است. این دادگان به سه قسمت آموزش (٨ نفر) ، تایید (٨نفر) و آزمون (٩ نفر) تقسیم شده است. دادگان KTH توسط بسیاری از الگوریتم‌های تشخیص رفتار انسان برای ارزیابی دقت استفاده شده است[71-73, 77-79, 90] ، که [71, 77, 90] آزمون‌های ساده‌تری را به‌کار برده‌اند و ما روش خود را با آنها مقایسه نمی‌کنیم.

٦-٥-٢-١- روش ارائه‌شده

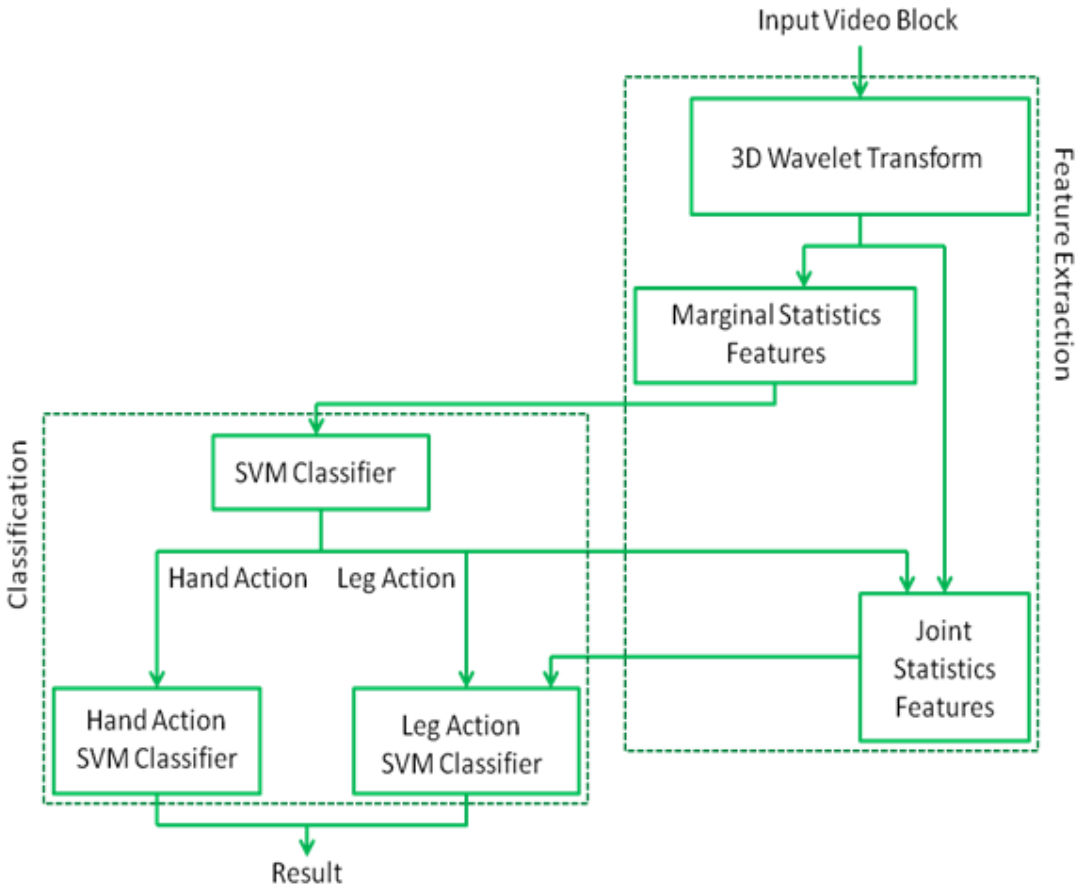
روش ارائه‌شده برای تشخیص رفتار انسان در این رساله، از دسته روش‌های نمایش تصویر کلی[12] است. مفهوم کلی این دیدگاه حذف پس‌زمینه و استخراج ویژگی از بدن انسان[13] است که اطلاعات حرکت و شکل بدن را دارد. پس‌زمینه دادگان KTH غیر استاتیک است و استخراج پس‌زمینه مرحله ساده‌ای نیست. با توجه به خصوصیات تبدیل سه‌بعدی موجک و حساسیت این تبدیل به لبه‌ها و تغییرات آنها در طول زمان، ما از اختلاف فریم‌های مجاور به جای استخراج بدن انسان مانند آنچه در

[12] Global image representation approach
[13] Human silhouette

[66,121] انجام گرفته است، استفاده نموده‌ایم. دنباله ویدئویی حاصل، حاوی اطلاعات مطلوب درباره حرکت است. در مرحله بعدی تبدیل موجک به دنباله ویدئوی حاصل از تفاوت فریم‌های مجاور اعمال می‌شود و پارامترهای حاشیه‌ای و توأم از هر دنباله ویدئویی مانند آنچه در بخش ٦-٢ آورده شد، استخراج شده، بردارهای ویژگی تشکیل می‌گردند. در نهایت، طبقه‌بندهای SVM برای تشخیص رفتار استفاده می‌شوند. ما از کتابخانه LIBSVM [118] برای آموزش و آزمون طبقه‌بند استفاده نمودیم.



شکل ٦-١٢. دنباله ویدئویی نمونه‌برداری‌شده (چپ) و دنباله ویدئویی از هم کم‌شده (راست).



شکل ٦-١٣. الگوریتم ارائه‌شده.

همانطور که در بخش قبلی دیده شد، انتظار می‌رود که ضرایب تبدیل سه‌بعدی موجک حاوی اطلاعات مهم درباره فعالیت در ویدئو باشند. این اطلاعات شامل سرعت و جهت حرکت‌ها است. شکل ٦-١٢ دنباله‌های ویدئویی و دنباله‌های ناشی از اختلاف فریم‌ها را برای هر کلاس ویدئو نمایش می‌دهد. همانطور که در شکل‌ها می‌توان دید نوع لبه‌ها و حرکت‌ها در رفتارهای با دست – مشت زنی، تکان دادن دست و کف زدن- به طور کلی با رفتارهای پا – راه رفتن، دویدن و آهسته دویدن- متفاوت است. برای تشخیص رفتارهای دستی، نوع حرکات مهمترین عوامل برای تشخیص رفتار هستند در صورتی که در تشخیص رفتارهای پا سرعت حرکات مهم‌ترین فاکتور هستند. بنابراین، یک طبقه‌بند سلسله‌مراتبی برای تشخیص رفتار استفاده شده است. ابتدا رفتارها به دو کلاس پا و دست بر اساس ویژگی‌های آماری حاشیه‌ای نمونه‌های ویدئویی با طبقه‌بند SVM تقسیم می‌شوند. در مرحله بعد، دو طبقه‌بند مختلف برای تشخیص رفتارهای دست و پا به طور جداگانه استفاده می‌شود. برای رفتارهای پا ویژگی‌های توأم نیز در کنار ویژگی‌های حاشیه‌ای برای کلاس‌بندی استفاده می‌شوند در حالی که برای رفتارهای دست، فقط ویژگی‌های حاشیه‌ای بکار می‌روند.

الگوریتم ارائه‌شده برای تشخیص رفتار انسان در شکل ٦-١٤ آمده است. ابتدا تبدیل سه‌بعدی موجک به دنباله ویدئوی منهاشده اعمال می‌شود. سپس، پارامترهای آماری حاشیه‌ای – ویژگی‌های GGD- از هر زیرباند ویدئو بر اساس توزیع ضرایب موجک در هر زیرباند تبدیل استخراج می‌شوند. پارامترهای GGD استخراج‌شده، بردار ویژگی حاشیه‌ای را تشکیل می‌دهند. تعداد پارامترهای حاشیه‌ای استخراج-شده برابر $14\times lev$ است زیرا از هر زیرباند دو پارامتر α و β استخراج می‌شود و هر سطح تبدیل موجک ٧ زیر باند دارد و lev تعداد سطوح تبدیل موجک است. در مرحله بعدی این ویژگی‌های حاشیه‌ای ابتدا با طبقه‌بند SVM به حرکات دست یا پا کلاس‌بندی می‌شوند. در صورتی که رفتار دست تشخیص داده شود، این بردار ویژگی‌های حاشیه‌ای برای کلاس‌بندی رفتارهای دستی[14] استفاده می‌شوند.

اگر رفتار پا تشخیص داده شده باشد، پارامترهای آماری توأم از سیگنال ویدئو استخراج می‌شود. به این ترتیب که توزیع ضرایب مشروط به والدین، همسایه‌ها و عموزاده‌ها برای ٧ زیرباند دقیق‌ترین سطح تبدیل موجک محاسبه می‌شوند و مقادیر kurtosis برای قطع‌های عمودی این توزیع‌ها به دست می‌آیند و نهایتاً بردار ویژگی توأم شامل ٢٧ ویژگی تشکیل می‌گردد –همانطور که در بخش ٦-٤-١ توضیح داده شد. همچنین تخمین MI بین ضرایب و والدین برای ٧ زیر باند دقیق‌ترین سطح تبدیل موجک تخمین زده می‌شود و در نهایت بردار ویژگی توأم با طول ٣٤ را تشکیل می‌دهد. نهایتاً این پارامترها در کنار پارامترهای حاشیه‌ای برای کلاس‌بندی رفتارهای پا استفاده می‌شوند.

٦-٥-٢-٢-نتایج تشخیص رفتار انسان

جدول ٦-٤. میانگین نرخ تشخیص رفتار انسان روی دادگان KTH.

الف) تبدیل سه‌بعدی موجک، فیلتر "Symlet" تعداد سطوح تجزیه متفاوت.

تعداد سطوح تجزیه	دقت
٢ سطح	٪٨٢,١
٣ سطح	٪٨٩,٧
٤ سطح	**٪٩٣,٤**
٥ سطح	٪٩١,٨

ب) تبدیل سه‌بعدی موجک با ٤ سطح تجزیه، انواع فیلتر موجک.

فیلتر موجک	دقت
Haar	٪٨٨,٣
Daubechies	٪٩٠,٨
Symlet	**٪٩٣,٤**

جدول٦- ٥. مقایسه روش‌های مختلف تشخیص رفتار انسان بر روی دادگان KTH.

روش	روش [79]	روش [73]	روش [78]	روش [72]	روش ارائه شده
دقت	٪٧١,٧	٪٨٣,٣	٪٨٦,٧	٪٩١,٧	**٪٩٣,٤**

[14] Hand action classification

نرخ‌های تشخیص رفتار برای تعداد مختلف سطوح تبدیل موجک و فیلترهای متفاوت موجک در جداول ٦-٤.الف و ٦-٤.ب آورده شده است. نتایج نشان‌گر این است که بهترین نتایج با فیلتر موجک "Symlet" به دست می‌آید و با افزایش تعداد سطوح تبدیل، نرخ‌های تشخیص افزایش می‌یابد، ولی تبدیل موجک با ٥ سطح اینگونه نیست. دلیل افزایش دقت تشخیص با افزایش سطوح تبدیل این است که با بالارفتن تعداد سطوح، جزئیات لبه و حرکت بیشتری استخراج می‌گردند ولی از آنجا که دادگان KTH دارای پس‌زمینه غیراستاتیک است، با افزایش بیش از حد تعداد سطوح، اطلاعات غیردلخواه از تغییرات و نویز زمینه با اطلاعات لازم و تاثیرگذار مداخله می‌نمایند و باعث کاهش نرخ تشخیص رفتار می‌شوند. همچنین با افزایش سطوح تبدیل تعداد ویژگی‌های مورد استفاده در کلاس‌بندی افزایش می‌یابد و دقت پایین می‌رود. نتایج بهترین نرخ تشخیص روش ارائه‌شده در جدول ٦-٥ با سایر روش‌های موجود مقایسه شده‌اند.

٦-٥-٢-٣-بحث و بررسی

جدول ٦-٦. مقایسه ماتریس‌های اغتشاش روش ارائه‌شده (جداول ٦.الف و ٦.ب) و روش ارائه‌شده در [72] (جدول ٦.ج) برای دادگان KTH

٦-٦.الف ویژگی‌های حاشیه‌ای و توأم

	walk	Jog	Run	Box	Hclp	Hwav
Walk	١٠٠	٠	٠	٠	٠	٠
Jog	٢,١	٩١,٧	٦,٢	٠	٠	٠
Run	٠	١٥,٣	٨٤,٧	٠	٠	٠
Box	٠	٠	٠	٩٩,٣	٠	٠,٧
Hwav	٠	٠	٠	٢,٨	٩٣,٧	٣,٥
Hclp	٠	٠,٧	٠	٨,٣	٠	٩١

٦-٦.ب ویژگی‌های حاشیه‌ای

	walk	Jog	Run	Box	Hclp	Hwav
Walk	١٠٠	٠	٠	٠	٠	٠
Jog	٢,١	٩١	٦,٩	٠	٠	٠
Run	٠	٢٠,١	٧٩,٩	٠	٠	٠
Box	٠	٠	٠	٩٩,٣	٠	٠,٧
Hwav	٠	٠	٠	٢,٨	٩٣,٧	٣,٥
Hclp	٠	٠,٧	٠	٨,٣	٠	٩١

٦-٦.ج روش ارائه‌شده در [72]

	walk	Jog	Run	Box	Hclp	Hwav
Walk	٩٩	١	0	٠	٠	٠
Jog	٤	٨٩	٧	٠	٠	٠
Run	٠	١٩	٨٠	٠	٠	٠
Box	٠	٠	٠	٩٧	٠	٣
Hwav	٠	٠	٠	٠	٩١	٩
Hclp	٠	٠,٧	٠	٥	٠	٩٥

در این بخش، نتایج تشخیص رفتار انسان با جزئیات بیشتری مورد بررسی قرار می‌گیرند و دلایل کارایی الگوریتم ارائه‌شده بیان می‌شود. همچنین پیچیدگی محاسباتی روش ارائه‌شده بحث می‌شود.

ماتریس‌های اغتشاش[15] روش ارائه‌شده (۹۳,٤٪) و روش ارائه‌شده بدون استفاده از پارامترهای آماری توأم در کلاس‌بندی رفتارها (۹۲,۵٪) به ترتیب در جداول ٦-٦.الف و ٦-٦. ب آورده شده‌اند. همچنین ماتریس اغتشاش روش [72] که بهترین نتیجه موجود کلاس‌بندی رفتار انسان روی دادگان KTH است در جدول ٦-٦.ج آورده شده است. با مقایسه این سه ماتریس، معلوم می‌شود که رفتارهای دست و پا در هر سه روش بخوبی از هم تمیز داده می‌شوند و مشکل اصلی در کلاس‌بندی رفتارهای پا ایجاد می‌شود، مخصوصاً بین دویدن و جهیدن[16]، که روش ما در این کلاس‌بندی بخصوص، بهتر عمل کرده است. دلیل این مساله، توانایی تبدیل موجک در تعیین مکان لبه‌ها و حرکات و تغییرات آنها، بافت در ویدئو و جزئیات مهم برای سیستم بینایی انسان است [11,96]. یک فاکتور مهم در تشخیص بین فعالیت‌های پا سرعت تغییرات در دنباله ویدئویی است که به خوبی با ویژگی‌های مکانی-زمانی مستخرج از تبدیل موجک بیان می‌شود. علاوه براین، روش ارائه‌شده، ویژگی‌های کلی را به جای ویژگی‌های محلی به کار می‌برد که باعث ساده‌ترشدن پیاده‌سازی و کلاس‌بندی می‌گردد.

برای بحث درباره پیچیدگی محاسباتی روش ارائه‌شده، ابتدا در نظر بگیرید که پیچیدگی محاسباتی تبدیل گسسته یک‌بعدی موجک برداری با طول N از مرتبه $O(N)$ است [122]. در نتیجه می‌توان تصور کرد که تابع خطی غیر نزولی مثبت $f_1(N)$ می‌تواند وجود داشته باشد که بیان‌گر پیچیدگی تبدیل موجک است. از آنجا که تبدیل سه‌بعدی موجک جداگانه به هر بعد ویدئو اعمال می‌شود، برای بلوک ویدئویی با ابعاد مکانی $X \times Y$ و طول زمانی T فریم، XY، XT و YT تبدیل موجک یک‌بعدی بر روی بردارهایی با طول به ترتیب T، Y و X داریم. بنابراین پیچیدگی محاسباتی تبدیل سه‌بعدی موجک با عبارت $XYf_1(T) + XTf_1(Y) + YTf_1(X)$ قابل بیان است. به بیان دیگر، برای هر سطح از تبدیل موجک تک‌بعدی برداری به طول N تعداد $N \times l$ ضرب لازم است که l برابر طول فیلتر موجک است. پس تعداد کل ضربها برابر است با $N \times l + \frac{N \times l}{2} + \cdots + \frac{N \times l}{2^{lev-1}} \leq N \times l\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{lev-1}} + \cdots + \frac{1}{2^n}\right)$ که طرف دوم نامساوی برابر $2N \times l$ است وقتی $n \to \infty$. پس پیچیدگی محاسباتی کل تبدیل سه‌بعدی موجک برای ماتریسی با سایز XYT دارای باند بالایی برابر $6lev(XYT)$ است که اگر $X = Y$ برابر با $6l(TX^2)$ خواهد شد.

الگوریتم تخمین پارامترهای GGD دارای پیچیدگی محاسباتی از مرتبه $O(N)$ است که N برابر تعداد نمونه‌ها است[11] . به این ترتیب تابع خطی مثبت غیرنزولی $f_2(.)$ را می‌توان درنظر گرفت، به

[15] Confusion
[16] Jogging

طوری که f_2 از مرتبه O از N باشد. تعداد نمونه‌ها در هر زیرباند از اولین سطح تجزیه موجک برابر $\frac{XYT}{8}$ است و برای سطوح دیگر این تعداد تقسیم بر 8^{lev} می‌شود. از آنجا که ۷ زیر باند در هر سطح تجزیه داریم، $7.\frac{XYT}{8}(1+\frac{1}{8}+\cdots+\frac{1}{8^{lev-1}})$ بیان‌گر تعداد نمونه‌هاست که این یک تصاعد حسابی است و برابر خواهد بود با $7.\frac{\frac{XYT}{8}}{1-\frac{1}{8}} < 7.\frac{\frac{XYT}{8}\left(1-\frac{1}{8^{lev}}\right)}{1-\frac{1}{8}}$ که سمت راست نا مساوی برابر با XYT است. پس پیچیدگی محاسباتی این مرحله نیز از مرتبه پیچیدگی محاسباتی مرحله قبل است که این مساله با آنچه در [123] درباره پیچیدگی محاسباتی تخمین GGD ادعا شده است، یکسان است. پیچیدگی محاسباتی تخمین پارامترهای توأم و مقادیر kurtosis آنها نیز برابر $O(XYT)$ است. که باز می‌توان تابع مثبت خطی غیرنزولی $f_3(.)$ از مرتبه O از N را برای بیان این پیچیدگی در نظر گرفت. به این ترتیب پیچیدگی کل مرحله استخراج ویژگی برابر با $f_2(TXY)+f_3(TXY)$ است. باز تابع خطی مثبت غیر نزولی $g_2(.)$ را می‌توان در نظر گرفت که رابطه نامساوی $f_2(TXY)+f_3(TXY)\leq g_2(TX^2)$ هنگامی که $X=Y$ است، برقرار باشد. این پیچیدگی محاسباتی برابر پیچیدگی محاسباتی محاسبه ویژگی‌های l_2-norm از زیرباندهای تبدیل است که در [90]، استفاده-شده، در حالی که نتایج ما بسیار بهتر است. با در نظر گرفتن شرایط مشابه برای تابع خطی غیرنزولی مثبت $g(.)$ به طوری که رابطه $6l(TX^2)+g_2(TX^2)\leq g(TX^2)$ برقرار باشد، پیچیدگی محاسباتی کلی می‌تواند با $g(TX^2)$ بیان شود.

برای مرحله کلاس‌بندی همان طور که توضیح داده شد، از طبقه‌بند SVM استفاده شده است که پیچیدگی محاسباتی‌اش از مرتبه $O(kn_{fv}n_{tr}{}^2)$ برای فاز آموزش و $O(kn_{fv}n_{tr})$ برای فاز آزمون است. k بیان‌گر تعداد کلاس‌ها، n_{fv} تعداد ویژگی‌ها و n_{tr} تعداد نمونه‌های آموزش است[124,125]. به بیان دیگر، پیچیدگی مرحله آزمون تابعی خطی از تعداد بردارهای پشتیبان[17] است، که باند بالایی آن تعداد نمونه‌های آموزش است و همچنین این پیچیدگی به طور خطی به طول بردارهای ویژگی و تعداد کلاس‌ها وابسته است [124,125] . بنابراین پیچیدگی الگوریتم در مراحل آموزش و آزمون می‌توانند به توابع خطی مثبت غیر نزولی $h_1(.)$ و $h_2(.)$ بیان شوند که به ترتیب توابعی خطی از k, n_{fv}, n_{tr} و k, n_{fv}, n_{tr}^2 هستند.

در الگوریتم ارائه‌شده، $14lev$ تعداد ویژگی‌های استخراج‌شده از خواص آماری حاشیه‌ای است که levتعداد سطوح تجزیه موجک است و معمولاً برابر ٤ است؛ و ٣٤ ویژگی از خواص آماری توأم استخراج می‌گردند. برای مرحله کلاس‌بندی، دو سناریو را در نظر می‌گیریم:

[17] Support Vectors

ابتدا، طبقه‌بند SVM برای تشخیص ٦ کلاس رفتار انسان در یک مرحله استفاده می‌شود. اینجا n_{fv}=90، $n_{tr} = 1536$ و $k = 6$ است و پیچیدگی با $h_1(k, n_{fv}, n_{tr}^2)$ برای مرحله آموزش و $h_2(k, n_{fv}, n_{tr})$ برای مرحله آزمون بیان می‌گردد. در سناریوی دوم، طبقه‌بند چند مرحله‌ای ارائه‌شده را در نظر بگیرید که ابتدا در مرحله اول رفتارها به دو کلاس دست و پا با کمک ویژگی‌های حاشیه‌ای تقسیم می‌شوند. در مرحله بعد رفتارهای با دست به سه کلاس با کمک همان ویژگی‌های حاشیه‌ای و با تعداد نصف نمونه‌های آموزش کلاس‌بندی می‌گردند. سرانجام رفتارهای پا با کمک ویژگی‌های حاشیه‌ای و توأم و تعداد نمونه‌های آموزش برابر نصف تعداد نمونه‌های اولیه تقسیم می‌شوند. بنابراین پیچیدگی محاسباتی مرحله آموزش این سناریو برابر می‌شود با :

$$h_1\left(\frac{k}{3}, \frac{56}{90}n_{fv}, n_{tr}^2\right) + h_1\left(\frac{k}{2}, \frac{56}{90}n_{fv}, \left(\frac{n_{tr}}{2}\right)^2\right) + h_1\left(\frac{k}{2}, n_{fv}, \left(\frac{n_{tr}}{2}\right)^2\right) = 0.21 \times h_1(k, n_{fv}, n_{tr}^2) + 0.078 \times h_1(k, n_{fv}, n_{tr}^2) + 0.125 \times h_1(k, n_{fv}, n_{tr}^2) = 0.413 \times h_1(k, n_{fv}, n_{tr}^2) < h_1(k, n_{fv}, n_{tr}^2) \qquad (٤-٦)$$

که همان طور که دیده می‌شود کمتر از نصف پیچیدگی محاسباتی سناریوی اول است. این پیچیدگی محاسباتی برای مرحله آزمون با روابط زیر بیان می‌شود:

$$h_2\left(\frac{k}{3}, \frac{56}{90}n_{fv}, n_{tr}\right) + h_2\left(\frac{k}{2}, \frac{56}{90}n_{fv}, \frac{n_{tr}}{2}\right) + h_2\left(\frac{k}{2}, n_{fv}, \frac{n_{tr}}{2}\right) = 0.21 \times h_2(k, n_{fv}, n_{tr}) + 0.156 \times h_2(k, n_{fv}, n_{tr}) + 0.25 \times h_2(k, n_{fv}, n_{tr}) = 0.62 \times h_2(k, n_{fv}, n_{tr}) < h_2(k, n_{fv}, n_{tr}) \qquad (٥-٦)$$

که باز کمتر از حالت مشابه در سناریوی اول است. در حالی که نتایج بهتری در کلاس‌بندی داریم. این محاسبات نشان می‌دهد که با کمک روش چند مرحله‌ای کلاس‌بندی SVM به نتایج بهتر و در عین حال پیچیدگی کمتر می‌رسیم.

برای مقایسه پیچیدگی محاسباتی روش ارائه‌شده با روش [72]، مرحله کلاس‌بندی را در نظر می‌گیریم، که هر دو روش از کلاس‌بندی SVM استفاده می‌نمایند. تعداد ویژگی‌های استفاده‌شده به طور خطی بر پیچیدگی محاسباتی تاثیر می‌گذارد. در [72] ٤٠٠٠ کلمه به‌صورت تجربی برای تشکیل بردارهای ویژگی درنظر گرفته شده‌اند در حالی که برای ما در بدترین حالت (سناریوی ١ که پیچیدگی محاسباتی بیشتری نسبت به سناریوی ٢ دارد) حدود ١٠٠ ویژگی انتخاب می‌شود. بنابراین پیچیدگی محاسباتی روش ما حداقل ٠,١ کمتر از روش [72] است.

ما پیچیدگی محاسباتی مرحله استخراج ویژگی را در نظر نگرفتیم. زیرا روش [72] نقاط مورد را برای استخراج ویژگی استفاده می‌نماید. برای انتخاب این نقاط، باید یک نمایش مکانی-زمانی با کمک کانولوشن بلوک ویدئو با کرنل‌های گوسی برای پارامترهای مقیاس متفاوت زمانی و مکانی محاسبه شود

[18] Interest points

که هر کانولوشن حداقل XYT ضرب در ابعاد کرنل دارد. بنابراین پیچیدگی حداقل از مرتبه‌ای بزرگتر از $O(XYT)$ خواهد بود که این کران پایین با پیچیدگی محاسباتی روش ما برابر است.

٦-٦- جمع‌بندی

در این فصل به مطالعه خواص آماری تبدیل سه‌بعدی موجک پرداختیم و روشی نوین برای تشخیص رفتار انسان و تحلیل سطح فعالیت در ویدئو بر اساس این خواص ارائه دادیم. با کمک مقادیر kurtosis هیستوگرام‌های حاشیه‌ای نشان دادیم که این توزیع‌ها دارای خواص غیرگوسی هستند. همچنین، هیستوگرام‌های حاشیه‌ای با توزیع‌های GGD بخوبی تخمین زده شدند و به این ترتیب این توزیع‌های حاشیه‌ای با دو پارامتر توصیف می‌شوند. بررسی خواص آماری توأم نشان داد که ضرایب، مشروط به والد، همسایه و عموزاده‌هایشان غیر همبسته[19] و در عین حال وابسته[20] هستند. قطع‌های عمودی این توزیع‌ها نشان می‌دهد که توزیع‌های شرطی نیز غیرگوسی هستند. MI به عنوان یک تخمین کمی از وابستگی نشان می‌دهد که با کاهش سطح فعالیت در ویدئو، وابستگی بین ضرایب زیرباندهای مختلف افزایش می‌یابد. علاوه بر این، منحنی‌های kurtosis تعریف شده، به چهار گروه بر اساس درجه فعالیت در ویدئو تقسیم شدند. نتایج بیان‌گر توانایی این منحنی‌ها در بیان سطح فعالیت در ویدئو است. در نهایت پارامترهای مکانی-زمانی استخراج‌شده بر اساس خواص آماری حاشیه‌ای و توأم تبدیل سه‌بعدی موجک به کمک طبقه‌بند SVM برای تشخیص رفتار انسان با دقت ۹۳,٤٪ استفاده شده‌اند که از نتایج روش‌های موجود بالاتر است.

[19] Uncorrelated
[20] Dependent

فصل هفتم

نتیجه‌گیری و پیشنهادات

۷-۱- مقدمه

اهمیت سیستم‌های اطلاعات چندرسانه‌ای با پیشرفت شبکه‌های باند وسیع و پردازشگرهای پرقدرت هر روز بیشتر می‌شود. در حالی که سیگنال ویدئو جزء مهمی از این سیستم‌ها است. منظور از تحلیل و مدل‌سازی پارامتریک سیگنال ویدئو، ارائه روشی ریاضی برای توصیف سیگنال است، که مجموعه پارامترهای استخراج‌شده در این روند در تحلیل ویدئو به کار برده می‌شوند. لذا بحث و بررسی در زمینه مدل‌سازی سیگنال ویدئو از مباحث مهم و قابل‌توجه در دنیای تحقیقاتی امروز است.

در این حوزه مساله ایده‌آل، یافتن مدلی پارامتریک برای بیان بی‌تلف، بدون افزونگی، بازگشت‌پذیر، با پیچیدگی کم و منطبق با سیستم بینایی انسان، برای ویدئو است. هرگونه حرکت در جهت نزدیک شدن به این مدل اصلی، در جایگاه خود از اهمیت ویژه‌ای برخوردار است. خروجی سیستم آنالیز آماری محتوای ویدئو، مجموعه‌ای از پارامترها است که می‌تواند در کاربردهای مختلف پردازش ویدئو استفاده شود. در این فصل ابتدا مباحث ارائه‌شده در رساله را به طور کلی مرور می‌نماییم و در پایان محورهای تحقیقاتی که در ادامه این کار می‌توانند مورد بررسی قرار گیرند را مطرح می‌کنیم.

۷-۲- خلاصه‌ای از مباحث مطرح‌شده

تمرکز اصلی این رساله بر روی تحلیل پارامتریک سیگنال ویدئو است. در فصل اول رساله به طرح موضوع پرداخته، انگیزه‌های تحقیق را بیان نمودیم و چالش‌ها را بررسی کردیم. از آنجا که سیگنال ویدئو دارای ابعاد زمانی و مکانی است، یک مدل خوب، مدلی است که محتوای زمانی- مکانی آن را در نظر بگیرد، در عین حال تاخیر قابل قبولی داشته باشد، قابلیت سرویس دهی در نرخ بیت‌های مختلف را دارا باشد، خصوصیات مختلف ابعاد زمانی و مکانی آن را مورد توجه قرار دهد، پیچیدگی محاسباتی معقولی داشته باشد و از همه مهم‌تر با خصوصیات سیستم بینایی انسان نیز سازگار باشد. در انتها نقاط نوآوری رساله را به اختصار بیان کردیم.

در فصل دوم به بیان تفاوت‌ها و مزایای روش‌های تحلیل و مدل‌سازی پارامتریک و غیرپارامتریک سیگنال پرداختیم و دیدیم که روش پارامتریک با کمک تعداد کمی پارامتر سیگنال و توزیع آن را توصیف می‌کنند و پیچیدگی محاسباتی پایین‌تری نسبت به روش‌های غیرپارامتریک دارد. در ادامه، به بررسی و توضیح روش‌های ترکیبی گوسی، مدل AR و HMM به عنوان مدل‌های مطرح در زمینه مدل-سازی و تحلیل ویدئو، که در صدد بیان ارتباطات زمانی فریم‌های ویدئویی هستند، پرداختیم و معایب و مزایای هر روش را شرح دادیم. سپس مختصری درباره کارهای مرتبط با کاربردهای چکیده‌سازی ویدئو و تشخیص رفتار انسان توضیح دادیم. این دو کاربرد برای ارزیابی روش‌های تحلیل و مدل‌سازی ارائه شده در این رساله استفاده شده‌اند. در ضمن، دادگان‌های ویدئویی که برای ارزیابی‌ها استفاده می‌شوند، در این فصل معرفی شدند.

در ادامه رساله، با توجه به مسائل مطرح شده، چهار روش برای تحلیل پارامتریک سیگنال ویدئو در چهار فصل متوالی ارائه شد:

- در فصل‌های سوم تا پنجم تحولات زمانی پارامترهای مکانی مستخرج از فریم‌های ویدئویی، با سه روش مورد تحلیل و بررسی قرار گرفتند. پارامترهای مکانی مورد استفاده در این روش‌ها بر اساس خواص آماری حاشیه‌ای زیرباندهای تبدیل موجک استخراج شده‌اند. به این ترتیب که ابتدا تبدیل موجک به فریم‌های ویدئویی اعمال می‌گردد. سپس هیستوگرام‌های حاشیه‌ای زیرباندهای این تبدیل با مدل تعمیم‌یافته گوسی تقریب زده می‌شوند و پارامترهای این مدل به عنوان مشخصات مکانی مستخرج از حوزه موجک در نظر گرفته می‌شوند و از این مشخصات به عنوان پارامترهای مکانی استفاده می‌شود. دلیل استفاده از این پارامترها این است که سیستم بینایی انسان به لبه‌ها و بافت تصویر حساس است و تبدیل موجک به صورت ساختاری با این خصوصیت سیستم بینایی انسان هماهنگ بوده، بخوبی این تغییرات را مد نظر می‌گیرد [96, 97]. در ضمن مدل گوسی تعمیم‌یافته تقریب مناسبی برای توزیع حاشیه‌ای زیرباندهای این تبدیل برای فیلترهای مختلف موجک و سطوح تبدیل متفاوت باشد [93-95,11,147]. برای بررسی روابط زمانی در سیگنال ویدئو از سه روش زیر استفاده شده است:
  - ✓ در روش اول که در فصل سوم مطرح شد، سیر تحول زمانی پارامترهای مکانی سیگنال ویدئو با کمک معیار فاصله مناسب KL برای بیان تشابه و تفاوت بین پارامترهای مکانی، بیان شده است. پارامترهای مکانی مناسبی که بر پایه خواص سیستم بینایی انسان انتخاب می‌شوند، در کنار معیار فاصله مناسب که روند تغییرات زمانی بین فریم‌های ویدئو را بیان می‌کند، بخوبی بیان‌گر تحولات محتوایی سیگنال ویدئو در سیر زمان می-

باشند. با درنظر گرفتن معیارهای مناسبی برای تعبیر تغییرات سیگنال ویدئو، این روش با دقت بسیار بالایی برای انتخاب مرز شات‌های تدریجی و آنی بین شات‌ها، خوشه-بندی شات‌ها و انتخاب فریم‌های کلیدی بکار برده شده است. دلیل موفقیت این روش در استفاده از پارامترهای مکانی مناسب، انتخاب معیار فاصله مقتضی و در نظر گرفتن ضوابط صحیح برای انتخاب فریم کلیدی – با توجه به شباهت و تفاوت بین فریم کلیدی و سایر فریم‌های داخل و خارج خوشه ویدئو– است.

- ✓ در فصل چهارم به ارائه روش دوم پرداختیم که در آن تحلیل سیگنال ویدئو بر پایه رخدادهای زمانی– مکانی[1] انجام می‌گیرد. در این روش برای تحلیل زمانی–مکانی رخدادهای بصری سیگنال ویدئو، به تجزیه زمانی پارامترهای مکانی پرداخته شده است. سیگنال ویدئو به عنوان مجموعه‌ای از مولفه‌های بصری مستقل همپوشان، رخداد، در نظر گرفته می‌شود. رخدادها همان توابع فشرده معمول دارای همپوشانی هستند که سیر تحول زمانی مجموعه‌ای از پارامترهای مکانی سیگنال ویدئو را توصیف می‌کنند. ما از تجزیه زمانی برای حل ساختار همپوشان رخدادها، که از مهم‌ترین مسائل موجود در تحلیل ویدئو است، استفاده می‌کنیم. ابتدا مجموعه پارامترهای مکانی، از سیگنال ویدئو استخراج می‌شوند و به صورت ترکیب خطی از مجموعه‌ای از تابع‌های فشرده همپوشان زمانی به نام رخداد، طی یک مرحله بهینه‌سازی، بیان می‌گردند. این روش سریع و دقیق برای تحلیل و مدل‌سازی سیگنال ویدئو استفاده شده است و قابلیت و کارایی بالای این رویکرد در کاربرد چکیده سازی[2] ویدئو ارائه شده است. مزیت اصلی این روش، عدم نیاز به تعیین مرز شات و خوشه‌بندی ویدئو برای انتخاب فریم کلیدی و در نتیجه سرعت بالای آن است. بنابراین روش مناسب تحلیل زمانی به کار رفته برای استخراج رخدادهای بصری همپوشان علاوه بر پارامترهای مکانی مقتضی نقش موثری در کارایی بالای این روش دارد.
- ✓ در روش سوم که در فصل پنجم مطرح شد، با کمک مدل‌سازی AR و استفاده از پارامترهای بهینه مکانی مستخرج از فریم‌های ویدئو سیر تغییرات و تحولات زمانی فریم‌ها مورد مطالعه قرار گرفته است. استفاده از مدل AR گزینه مناسبی برای مدل-کردن روابط زمانی دنباله ویژگی فریم‌ها است. انتخاب ویژگی برای هر فریم با در نظرگرفتن خواص سیستم بینایی انسان از روی تبدیل‌یافته فریم در حوزه تبدیل موجک،

[1] Spatio-temporal event-based approach
[2] Video abstraction

استخراج می‌گردد. انتخاب پارامترهای مدل AR مساله مهمی است. از روش آموزش RLS در این تحلیل استفاده شده است و نتایج این تحلیل پارامتریک برای انتخاب مرز شات و تعیین فریم کلیدی شات به کار گرفته شده است. باز هم انتخاب پارامترهای مکانی مناسب در کنار مدل مقتضی که برای بیان روابط زمانی فریم‌ها استفاده شده است، فاکتورهای اساسی در موفقیت این روش هستند.

- در روش تحلیلی چهارم که در فصل ششم بیان شده است، به تحلیل پارامترهای مکانی-زمانی ویدئو در حوزه تبدیل سه‌بعدی موجک پرداخته شده است. خواص آماری تبدیل سه‌بعدی موجک سیگنال‌های ویدئوی طبیعی مورد مطالعه قرار گرفته و این خواص آماری در مدل‌سازی و تحلیل ویدئو بکار رفته‌اند. استفاده از تبدیل موجک که نمایش تنکی از سیگنال ویدئو ارائه می‌دهد و با ساختار بینایی انسان مطابقت دارد [96,97] و استخراج ویژگی‌های مناسب بر اساس خواص آماری این تبدیل که حامل اطلاعات حرکت بر مبنای HVS هستند[11]، دو فاکتور مهم در کارایی بالای روش ارائه شده است. خواص آماری حاشیه‌ای[1] و مشترک[2] و تخمین اطلاعات متقابل[3] تبدیل سه‌بعدی موجک استخراج شده است و رابطه بین سطح فعالیت در ویدئو و توزیع شرطی بدست آمده است. نشان داده شده است که هیستوگرام‌های حاشیه‌ای بخوبی با تابع توزیع گوسی تعمیم‌یافته تقریب زده می‌شوند و MI بین ضرایب با افزایش سطح فعالیت در ویدئو کاهش می‌یابد. پارامترهای استخراج‌شده از توزیع‌های توأم ضرایب برای تحلیل سطح فعالیت ویدئو استفاده شده‌اند. همچنین، پارامترهای زمانی-مکانی ویدئو مستخرج از خواص آماری حاشیه‌ای و توأم برای تشخیص رفتار انسان بکار رفته‌اند.

مباحث و تحلیل‌های فوق در یک یا برخی از کاربردهای واسط زیر مورد استفاده قرار گرفته اند: تشخیص مرز شات، تشخیص نوع تغییر شات، بررسی سطح فعالیت در ویدئو، تشخیص وقایع تصویری، انتخاب فریم‌های کلیدی، تشخیص رفتار انسان.

در تمامی تحلیل‌های انجام گرفته، از آزمون‌های تحلیلی و ادراکی برای ارزیابی کارایی تحلیل‌ها در کاربردهای مختلف پردازش ویدئو استفاده شده است. آزمون‌های تحلیلی آزمون‌هایی تکرارپذیر، کم‌خرج و دقیق هستند؛ در مقابل، آزمون‌های ادراکی قرار دارند که گران و تکرارناپذیر هستند و نتایجی مطابق با سیستم بینایی انسان ارائه می‌دهند. علاوه بر این، از مجموعه دادگان‌های متداول و با اهمیت ویدئو برای این ارزیابی‌ها و مقایسه روش‌های تحلیلی ارائه‌شده با روش‌های موجود استفاده شده است. دادگان

[1] Marginal Statistics
[2] Joint Statistics
[3] Mutual Information Estimate

تشخیص رفتار انسان KTH، مجموعه ویدئویی Hollywood2، دادگان TRECVID و مجموعه‌های 'Simon Fraser University Video Library'، و دادگان 'The Open Video Project' انتخاب شده‌اند؛ که مجموعاً بیش از ۲۰ ساعت ویدئو مورد آزمون قرار گرفته است.

جدول ۷-۱- مقایسه کلی روشهای ارائه شده.

تحلیل ارائه شده	روش اول	روش دوم	روش سوم	روش چهارم
تحلیل زمانی	KLD	TD	AR	پارامترهای آماری حاشیه‌ای و مشترک تبدیل سه-بعدی ویولت
تحلیل مکانی	پارامترهای آماری حاشیه‌ای مستخرج از زیرباندهای تبدیل دوبعدی ویولت			
پیچیدگی زمانی[1]	$O(p.T)$ $\cong O(10T)$	$O(n_{ev}.p.l_{sb}^2)$ $\cong O(10^2T)$	$O(p.P^2T)$ $\cong O(10^2T)$	$O(XYT)$
پیچیدگی مکانی	$O(XY)$			
ارزیابی	ادراکی & تحلیلی			
دادگان	TRECVID Hollywood2 Open Video Project SFU Video database			KTH TRECVID Hollywood2 Open Video Project SFU Video database
کاربرد	▪ انتخاب مرز شات ▪ تعیین نوع مرز شات ▪ انتخاب فریم کلیدی	▪ انتخاب فریم کلیدی ▪ چکیده‌سازی ویدئو	▪ انتخاب مرز شات ▪ انتخاب فریم کلیدی	▪ تعیین میزان فعالیت در ویدئو ▪ تشخیص فعالیت انسان

نتایج تحلیل‌ها و ارزیابی‌های انجام شده، توانایی بالای هر کدام از روش‌های تحلیلی را در بیان محتوای سیگنال ویدئو نشان می‌دهد. در سه روش اول، سیر تحول زمانی پارامترهای مکانی سیگنال ویدئو بخوبی بر اساس روش‌های ارائه‌شده، بیان می‌شوند. انتخاب ویژگی برای هر فریم در این سه روش با در نظرگرفتن خواص سیستم بینایی انسان از روی تبدیل‌یافته فریم در حوزه تبدیل موجک، استخراج می-

[1] p تعداد پارامترهای مکانی و کمتر از ۲۵، n_{ev} تعداد رخدادهای استخراج شده و حدود $0.15T$ و l_{sb} دقت زمانی روش و تقریبا برابر ۲۵ است و P مرتبه مدل AR است که حدود ۲۰ انتخاب می شود.

گردد. این ویژگی‌ها مجموعه‌ای از پارامترهای آماری خروجی آنالیز مکانی فریم‌ها هستند؛ پارامترهای گوسی تعمیم‌یافته به عنوان ویژگی فریم‌ها انتخاب می‌شوند. پارامترهای مکانی-زمانی استخراج‌شده از روش چهارم نیز با دقت بالایی خواص و تغییرات زمانی و مکانی ویدئو را بیان می‌نمایند. مشخصات روش‌های ارائه‌شده با جزئیات بیشتری در جدول ۷-۱ بیان شده‌اند.

۷-۳- پیشنهادات

با توجه به تحقیقات انجام‌شده، زمینه‌های تحقیقی زیر به عنوان کارهای آینده پیشنهاد می‌شوند:

- پارامترهای استخراجی از مدل‌های ارائه‌شده می‌توانند برای بهبود کارایی استانداردهای فشرده-سازی ویدئو به کار روند. به طور مثال با انتخاب فریم‌های کلیدی مناسب شات‌ها در فرآیند فشرده‌سازی MPEG و H.264، فریم‌های مناسب‌تری برای کد کردن I، انتخاب می‌شوند. علاوه بر این، بررسی نوع و میزان فعالیت در ویدئو است، که با تشخیص نوع فعالیت و میزان آن، تعداد بیت‌های بهینه به هر فریم اختصاص داده می‌شود؛ همچنین ضرایب بردار حرکت دقیق‌تر محاسبه می‌شوند که تخمین حرکت نیز بهبود می‌یابد و یا امکانات کدینگ شئ‌گرا فراهم می‌گردد. هر یک از موارد ذکرشده ممکن است به عنوان یک زمینه تحقیقاتی برای ارتقاء کارایی سیستم‌های فشرده‌سازی مورد توجه قرار گیرد.
- در بازیابی و شاخص‌گذاری ویدئو، بردارهای ویژگی شامل پارامترهای مکانی-زمانی استخراجی در بهبود کارایی و کاهش حجم محاسبات و زمان دسترسی نقش بسزایی خواهند داشت. در این حالت استفاده از پارامترهای زمانی- مکانی استخراج شده از ویدئو بر پایه خواص آماری تبدیل موجک سه‌بعدی می‌توانند به عنوان ویژگی به کار روند. با توجه به توانایی این ویژگی‌ها در ارائه اطلاعات مناسب از محتوای لبه‌ها و بافت ویدئو و همچنین سطح فعالیت در آن، استفاده از آن‌ها باعث پیشرفت سیستم‌های مزبور خواهد شد. علاوه بر این فریم‌های کلیدی مستخرج از ویدئو با کمک تحلیل‌های معیار فاصله، تجزیه زمانی و مدل AR، می توانند به عنوان نمایندگان مناسبی برای بیان شات‌های ویدئو به حساب آیند و پارامترهای استخراج‌شده از این فریم‌ها به عنوان ویژگی‌های شات مربوطه استفاده گردد. از آنجا که در بازیابی ویدئو پارامترهای رنگ نیز پارامترهای مهمی به حساب می‌آیند، این پارامترها در کنار پارامترهای تبدیل موجک به بهبود عملکرد سیستم‌های مزبور کمک خواهند نمود. استفاده از معیار فاصله KL نیز برای بیان تفاوت میان پارامترهای انتخابی از زیرباندهای تبدیل موجک مناسب است.

- با استخراج فریم‌های کلیدی بهینه، که کاندیداهای مناسبی از کل ویدئو هستند، کارایی خلاصه‌سازی هم بهبود می‌یابد. در خلاصه‌سازی ویدئو، قسمتهایی از ویدئو به عنوان نمایندگان محتوای ویدئو، انتخاب می‌گردند. این قطعات انتخابی در کنار قطعات صدای انتخاب‌شده، چکیده مطلوب‌تری از دنباله ویدئویی برای بیننده فراهم می‌آورند. در روش تجزیه زمانی، مراکز برخی رخدادهای دیداری به عنوان فریم‌های کلیدی انتخاب می‌شوند. در این حالات می‌توان با کمک تابع رخداد متناظر با فریم کلیدی منتخب، قسمت مربوطه را در دنباله ویدئویی انتخاب نمود و بدین‌سان خلاصه ویدئویی را تشکیل داد.
- استفاده از روش تجزیه زمانی در بیان تحولات زمانی ویدئو می‌تواند در کاربردهای دیگری از پردازش ویدئو نیز به کار برده شود. استفاده از این روش در نهان‌نگاری ویدئو مفید بوده، باعث بهبود کارایی سیستم می‌شود. برای این منظور لازم است در ابتدا دقت این تجزیه بالا رود. بنابراین افزایش دقت این تجزیه با در نظر گرفتن سرعت آن یکی از زمینه‌های موجود تحقیق به شمار می‌رود.
- نتایج بررسی میزان فعالیت در ویدئو می‌تواند در تخمین دقیق‌تر پهنای باند مورد نیاز برای یک سیگنال ویدئو مورد استفاده قرار گیرد. این ایده می‌تواند در کاربردهای کنفرانس ویدئویی به کار رود که نیاز به داشتن ویدئو با کیفیت بالا و در عین حال با توجه به پهنای باند متغیر قابل استفاده است. همچنین مدل AR نیز در این زمینه می‌تواند به پیش‌بینی پهنای باند لحظات آینده و انتخاب معیار کیفیت و پارامتر کوانتیزاسیون کمک نماید.
- در صورتی که مرحله پیش‌پردازشی برای جداسازی پیش‌زمینه به مدل‌های ارائه‌شده اضافه شود، اطلاعات دقیق‌تری از روند تغییرات زمینه و محتوای ویدئو استخراج می‌شود و عملکرد سیستم‌ها بهبود می‌یابد.
- همچنین، استخراج مدلی بر مبنای درک انسان از تغییرات و دنبال کردن اشیاء در ویدئو یکی از زمینه‌های مهمی است که انجام آن باعث بهبود بسیاری از زمینه‌های پردازش ویدئو می‌شود.

[1] G.E. Hicham, F.M. Mohamed, G.A. Walid, “Spatio-temporal histograms,” Lecture Note Computer Science, 3633, pp. 19–36, 2005.
[2] I. Laptev, T. Lindeberg, “Local descriptors for spatio-temporal recognition,” In European Conf. on Computer Vision, pp. 91–103, 2004.
[3] C.W. Ngo, T.C. Pong, H.J. Zhang, “Motion-based video representation for scene change detection,” Computer vision, vol. 50, issue 2, pp. 127–142, 2002.
[4] F. Coudert, J. Benois-Pineau, P.Y. Le Lann, D. Barba, “A system for video content analysis ‘‘on the fly’’,” IEEE Trans. Conf. Multimedia Comput. Syst. 1, 679–684, 1999.
[5] H. Nicolas, A. Manaury, J. Benois-Pineau, W. Dupuy, D. Barba, “Grouping video shots into scenes based on 1D mosaic descriptors,” In Proc. of IEEE Int’l Conf on Image Processing, ICIP, pp. 637–640, 2004.
[6] R. Rajesh, T.O. Michael, “Synthesizing processed video by filtering temporal relationships,” IEEE Trans. Image Proc., vol. 11, issue 1, pp. 26–36, 2002.
[7] L.Y. Duan, M. Xu, Q. Tian, C.S. Xu, S.J. Jesse, “A unified framework for semantic shot classification in sports video,” IEEE Trans. Multimedia, vol. 7, issue 6, pp. 1066–1083, 2005.
[8] W. Cui, H.K. Xiang, Y. Yao, J.W. Liu, Q. Cao, “Research on the temporal model in video compression,” IJCSES Int’l Journal of Computer Sciences and Engineering Systems, vol.1, issue 2, April 2007, pp. 71-76.
[9] G. Xu, Y.F. Ma, H.J. Zhang, S.Q. Yang, “HMM-based framework for video semantic analysis,” IEEE Trans. Circuits Syst. Video Technol., vol. 15, issue 11, pp. 1422–1433, 2005.
[10] Y. Zhai, M. Shah, “Video scene segmentation using Markov chain Monte Carlo,” IEEE Trans. Multimedia, vol. 8, issue 4, pp. 686–697, 2006.
[11] M.N. Do, “Directional multiresolution image representations,” PhD thesis, Department of Communication Systems, Swiss Federal Institute of Technology Lausanne, 2001.
[12] M. Crouse, R. D. Nowak, R. G. Baraniuk, “Wavelet-based signal processing using hidden Markov models,” IEEE Trans. Signal Proc. (Special Issue on Wavelets and Filterbanks), pp. 886–902, 1998.
[13] A. L. da Cunha, M. Do., M. Vetterli, “A stochastic model for video and its information rates full text,” In Proc. of Data Compression Conference, pp. 3-12, 2007.
[14] R.R. Lawrence, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77, issue 2, pp. 257- 286, 1989.
[15] R.R. Lawrence Raniner, J. Biing-Hwang , Fundementals of speech processing, Copyright @ Prentice-Hall Interational, Inc, 1993.
[16] M. Eftekhar, “HMM modeling to track hand movements based on EMG signals,” M.Sc. Thesis, Biomedial Eng., Electrical Eng. Dep., Sharif Univ. of Tech., 2005.
[17] W. Chen, Y.J. Zhang, “Parametric model for video content analysis,” Elsevier B.V., Pattern Recognition Letters 29, pp.181–191, 2008.
[18] W. Chen, Y.J. Zhang, “Video segmentation and key frame extraction with parametric model,” In Proc. on Int’l Symposium on Communications, Control and Signal Processing, pp. 1020-1023, 2008.
[19] S. Gelfand, M. Pinsker, “Coding for channel with random parameter,” Probl. Contr. Inform Theory, vol. 9, issue 1, pp 19-31, 1980.
[20] H. Greenspan, J. Goldberger, L. Ridel, “A continuous probabilistic framework for image matching,” Computer Vision and Image Understanding, vol. 84, pp. 384-406, 2001.
[21] C. Carson, S. Belongie, H. Greenspan, J. Malik, “Blobworld: image segmentation using expectation-maximization and its application to image querying,” IEEE Trans. Pattern Analysis and Machine Intel., vol. 24, issue 8, pp. 1026-1038, 2002.
[22] H. Greenspan, J. Goldberger, A. Mayer, “Probabilistic space-time video modeling via piecewise GMM,” IEEE Trans. on Pattern Analysis and Machine Intel., vol. 26, issue 3, pp. 384-396, 2004.
[23] X. Mo and R. Wilson, “Video modeling and segmentation using Gaussian mixture models,” In Proc. of Int’l Conf. on Pattern Recognition, vol.3, pp. 854-857, 2004.
[24] S. Cheng, L. Xingzhi, S.M. Bhandarkar, “A multiscale parametric background model for stationary foreground object detection,” In Proc. of IEEE Workshop on Motion and Video Computing, pp. 18 – 22, 2007.
[25] H. Greenspan, S. Gordon, J. Goldberger, “Probabilistic models for generating, modeling and matching image categories,” In Proc. of Int’l Conf. on Pattern Recognition, pp. 970-973, 2002.
[26] G. Dvir, H. Greenspan, Y. Rubner, “Context-based image Modeling,” In Proc. of Int’l Conf. on Pattern Recognition, pp. 162-165, 2002.

[27] J. Goldberger, S. Gordon, H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," In Proc. of Int'l Conf. on Computer Vision, pp. 487-493, 2003.
[28] H. Greenspan, J. Goldberger, L. Ridel, "A continuous probabilistic framework for image matching," Computer Vision and Image Understanding, vol. 84, issue 3, pp. 384-406, 2001.
[29] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," IEEE Trans. on Image Processing, vol. 9, issue 1, pp. 3-19, 2000.
[30] E. Iain, G. Richardson, Video codec design, developing image and video compression systems. Copyright © 2002 by John Wiley & Sons Ltd.
[31] J. Watkinson, The MPEG handbook, MPEG-1, MPEG-2, MPEG-4. Copyright © 2001 by Focal Press.
[32] E. Iain, G. Richardson, H.264 and MPEG-4 video compression. Copyright © 2003 by John Wiley & Sons Ltd.
[33] Y. Wang, J. Ostermann, Y.Q. Zhang, Video processing and communications. Copyright © 2002 by Prectice-Hall.
[34] C. Wootton, A practical guide to video and audio compression, from Sprocket and Rasters to Macro blocks. Copyright © 2005, Elsevier Inc.
[35] D. Salomon, Data compression. Copyright © 2004, 2000, 1998 Springer-Verlag New York, Inc.
[36] G.J. Sullivan, P. Topiwala, A. Luth, "The H.264/AVC advanced video coding standard: overview and introduction to the fidelity range extensions," In Proc. of SPIE Conference on Applications of Digital Image Processing, Vol. 5558, pp. 454-474, 2004.
[37] A. Fort, A. Ismaelli, C. Manfredi, P. Bruscaglioni, **"**Parametric and non-parametric estimation of speech formants: application to infant cry," Med. Eng. Phys., vol. 18 (6), pp. 677-691, 1996.
[38] J. Goldberger, H Greenspan, S. Gordon, "Unsupervised image clustering using the information bottleneck method," In Proc of DAGSM-Symposium, pp. 158-165, 2002.
[39] X. Song, G. Fan, "Selecting salient frames for spatiotemporal video modeling and segmentation," IEEE Trans. on Image Processing, Vol. 16, No. 12, pp. 3035-3046, 2007.
[40] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statistical Soc. B, vol. 39, no. 1, pp. 1-38, 1997.
[41] R.O. Duda, P.E. Hart, Pattern classification and scene Analysis. John Wiley and Sons, 1973.
[42] G. Caccia, R. Lancini, S. Russo, "Key frames extraction in athletic video," SPIE, 5150, pp. 2097–2104, 2003.
[43] C.Y. Lang, D. Xu, W.G. Cheng, Y.W. Jiang, "Efficient key-frame extraction using density-estimation-based non-parametric clustering," In Proc. of IEEE Region 10 Conf. vol. A, pp. 21–24, 2004.
[44] Z. Cernekova,, C. Nikou, I. Pitas, "Shot detection in video sequences using entropy-based metrics," In Proc. of IEEE Int'l Conf. on Image Processing, vol. 3, pp. 421-424, 2002.
[45] B. Furht, S.W. Smoliar, H.J. Zhang, "Video and image processing in multimedia systems," Kluwer Academic Publishers, Norwell, MA, 1995.
[46] H. Lu, Y.P. Tan, "An effective post-refinement method for shot boundary detection," IEEE Trans. CSVT, vol. 15, issue 11, pp. 1407–1421, 2005.
[47] A. Said, W.A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," IEEE Trans. on Circuits and Systems for Video Tech., vol. 6, issue.3, pp. 243-250, 1996.
[48] D. D.-Y. Po, M. N. Do, "Directional multiscale modeling of images using the Contourlet transform," IEEE Trans. on Image Processing, vol. 15, no. 6, pp. 1610-1620, 2006.
[49] I. Cohen, A. Garg, T.S. Huang, "Emotion recognition from facial expressions using multilevel HMM," In Proc. of NIPS Workshop on Affective Computing, 2000.
[50] T. Otsuka, J. Ohya., "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," In Proc. Int. Conf. on Image Processing, pp. 546–549, 1997.
[51] J. Huang, Z. Liu, Y. Wang, Y. Chen, E. K. Wong, "Integration of multimodal features for video scene classification based on Hmm," In Proc. of the IEEE Workshop on Multimedia Signal Processing, 1999.
[52] A.D. Wilson, A.F. Bobick, "Parametric hidden Markov models for gesture recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, No. 9, pp. 884-900, 1999.
[53] J. Williams, K. Aggelos, "An HMM-based speech-to-video synthesizer," IEEE Transactions on Neural Networks, Vol. 13, No. 4, pp. 900-915, 2002.

[54] G. Potamianos, H. P. Graf, E. Cosatto, "An image transform approach for HMM based automatic lip reading," In Proc. of IEEE Conf. on Image Proc. , ICIP, Vol. 3, pp. 173-177, 1998.
[55] C. Yunqiang, R. Yong, S.H. Thomas , "JPDAF based HMM for real-time contour tracking," In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition., pp.543-550, 2001.
[56] N. Dimitrova, L. Agnihotri, G. Wei, "Video classification based on HMM using text and faces," In Proc. of European Signal Processing Conf., 2000.
[57] M. Davis, M. Tuceryan, "Coding of facial image sequences by model-based optical flow," In Proc. of Int'l Workshop on Synthetic-Natural Hybrid Coding and 3D Imaging, pp. 192-194, 1997.
[58] N. Li, J. Bu, C. Chen, R. Liang," 3D facial animation from Chinese text," In Proc. of IEEE Int'l Conf. on Systems, Man and Cybernetics, Vol. 4, pp.3738 – 3743, 2003.
[59] F.I. Parke, K.Waters, Computer facial animation. AK Peters, Wellesley, Massachusetts, 1996.
[60] Z. Liu, Z. Zhang, Chuck Jacobs, M. Cohen, "Rapid modeling of animated faces from video," Technical Report MSR-TR-2000-11, Microsoft Research Microsoft Corporation, Feb. 2000.
[61] R. Heck, M. Gleicher, "Parametric motion graphs," In Proc. of ACM SIGGRAPH Symposium on Interactive 3D Graphics conference proceedings, 2008.
[62] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, "Machine recognition of human activities: a survey," IEEE Trans. on Circuits and Systems for Video Technology, vol. 18, issue 11, pp. 1473-1488, 2008.
[63] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, K.W. Bowyer, "The human ID gait challenge problem: data sets, performance, and analysis," IEEE Trans. on Pattern Analysis and Machine Intel., vol. 27, issue 2, pp. 162-177, 2005.
[64] R. Poppe, "A survey on vision based human action recognition," Image and Vision computing, Elsevier, vol. 28, issue 6, pp. 976-990, 2010.
[65] T.B. Moeslund, A. Hilton, V. Kruger, "A survey of advances in vision based human motion capture and analysis," Computer Vision and Image Understanding, vol. 104, issue 2-3, pp.90-126, 2006.
[66] A.F. Bobick, J.W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, issue 3, pp.257-267, 2001.
[67] A. Fathi, G. Mori, "Action recognition by learning mid-level motion features," In Proc. of Int'l Conf. on Computer Vision and Pattern, pp. 1–8, 2008.
[68] N. Ikizler, R.G. Cinbis, P. Duygulu, "Human action recognition with line and flow histograms," In Proc. of Int'l Conf. on Pattern Recognition, pp. 1–4, 2008.
[69] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, "Actions as space–time shapes," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, issue 12, pp. 2247-2253, 2007.
[70] A. Oikonomopoulos, I. Patras, M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," IEEE Trans. on Systems Man and Cybernetics, Part B: Cybernetics, vol. 36, issue 3, pp.710-719, 2006.
[71] H. Jhuang, T. Serre, L. Wolf, T. Poggio, "A biologically inspired system for action recognition," In Proc. of Int'l Conf. on Computer Vision, pp. 1-8, 2007.
[72] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies," In Proc. of Int'l Conf. on Computer Vision and Pattern Recognition, pp. 1-8, 2008.
[73] J.C. Niebles, H. Wang, L.F. Fei, "Unsupervised learning of human action categories using spatial–temporal words," In Proc. of Int'l Journal of Computer Vision, vol. 79, issue 3, pp. 299-318, 2008.
[74] Y. Song, L. Goncalves, P. Perona, "Unsupervised learning of human motion," IEEE Trans. on Pattern Analysis and Machine Intel., vol. 25, issue 7, pp. 814-827, 2003.
[75] T.K. Kim, R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," IEEE Trans. on Pattern Analysis and Machine Intel., vol. 31, issue 8, pp. 1415-1428, 2009.
[76] A. Oikonomopoulos, M. Pantic, I. Patras, "Sparse B-spline polynomial descriptors for human activity recognition," Image and Vision Computing, vol. 27, issue 12, pp.1814-1825, 2009.
[77] S.F. Wong, T.K. Kim, R. Cipolla, "Learning motion categories using both semantic and structural information," In Proc. of Int'l Conf. on Computer Vision and Pattern Recognition, pp. 1-8, 2007.
[78] S.F. Wong, R. Cipolla, "Extracting spatiotemporal interest points using global information," In Proc. of Int'l Conf. on Computer Vision, pp. 1-8, 2007.
[79] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: A local SVM approach," In Proc. of Int'l Conference on Patter Recognition, pp. 32-36, 2004.

[80] B.T. Truong, S. Venkatesh, "Video abstraction: A systematic review and classification," ACM Trans. Multimedia Comput. Commun. Appl. Vol. 3, pp.1-37, 2007.
[81] Y. Li, S.H. Lee, S.H. Yeh, C.C.J. Kuo, "Techniques for movie content analysis and skimming," IEEE Signal Processing Magzine, vol. 23, pp. 79-89, 2006.
[82] Y. Zhuang, Y. Rui, T.S. Huang, S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," In Proc. of IEEE Int'l Conf. on Image Processing, pp. 283-287, 1998.
[83] A. Nagasaka, Y. Tanaka, "Automatic video indexing and full-video search for object appearances," Visual Database Systems II, Elsevier, vol. 15, issue 2, pp. 113-127, 1992.
[84] L.Z. Manor, M. Irani, "Event-based video analysis," In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Rec., vol. 2, pp. 123-130, 2001.
[85] E. Bulut, T. Capin, "Key frame extraction from motion capture data by curve saliency," In Proc. of Int'l Conf on Computer Animation and Social Agents, CASA, 2007.
[86] L. Shao, L. Ji, "Motion histogram analysis based key frame extraction for human action/activity representation," In Proc. of Canadian Conf. on Computer and Robot Vision, CRV, pp. 88 – 92, 2009.
[87] M.L. Cooper, J. Foote, "Discriminative techniques for keyframe selection," In Proc. of IEEE Int'l Conference on Multimedia and Expo, ICME, pp. 502-505, 2005.
[88] http://www-nlpir.nist.gov/projects/trecvid (2011). National Institute of Standards and Technology (NIST). Accessed 15 aprill 2011
[89] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html (2011). Accessed 15 aprill 2011
[90] K. Rapantzikos, Y.S. Avrithis, S.D. Kollias, "Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition," In Proc. of ACM Int'l Conf. on Image and Video Retrieval, CIVR, pp. 294-301, 2007.
[91] http://nsl.cs.sfu.ca/wiki/index.php/Video_Library_and_Tools (2011). Accessed 15 aprill 2011
[92] http://www.open-video.org (2011). Accessed 15 aprill 2011
[93] E.P. Simoncelli, R.W. Duccigrossi, "Embedded wavelet image compression based on a joint property model," In Proc of IEEE Int'l Conf. Image Processing, vol.1, pp. 640-643, 1997.
[94] K. Sharifi, A. Leon-Garcia. "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," IEEE Trans. Circuits Sys. Video Tech., vol.5, No. 52–56, 1995.
[95] P. Moulin, J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," IEEE Trans. Inform. Th., vol. 45, no. 3, pp. 909–919, 1999.
[96] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans. Patt. Recog. and Mach. Intell., vol. 11, issue 7, pp. 674– 693, 1989.
[97] T.H. Oh, R. Besar, "JPEG2000 and JPEG: image quality measures of compressed medical images," In Proc. of National conf. on Telecom. Tech., pp. 31 – 35, 2003.
[98] C. Cotsaces, N. Nikolaidis, I. Pitas, "Video shot detection and condensed representation: a review," IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 28–37, 2006.
[99] A. Amiri, M. Fathy, "Video shot boundary detection using QR-decomposition and Gaussian transition detection," EURASIP Journal on Advanced Signal Processing, 2009.
[100] A.F. Smeaton, P. Over, A.R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," Elsevier Computer Vision and Image Understanding, vol. 114, issue 14, pp. 411-418, 2010.
[101] M.J. Pickering, S. Ryger, "Evaluation of key frame-based retrieval techniques for video," Elsevier Computer Vision and Image Understanding, CVIU, vol. 92, no. 2-3, pp. 217-235, 2003.
[102] R. Polana, R. Nelson, "Detecting activities," In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition, pp. 2-5, 1993.
[103] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," In Proc. of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, pp. 81-84, 1983.
[104] S. Ghaemmaghami, "Audio segmentation and classification based on a Selective analysis scheme," In Proc. of Int'l Conf. on Multimedia Modeling, MMM, pp. 42-47, 2004.
[105] B. Janvier, E. Bruno, T. Pun, S.M. Maillet, "Information-theoretic temporal segmentation of video and applications: Multiscale keyframes selection and shot boundaries detection," Multimedia tools and application, vol. 3, issue 3, pp. 273-288, 2006.
[106] B.S. Manjunath, S. Chandrasekaran, Y.F. Wang, "An eigenspace update algorithm for image analysis," In Proc. of Int. Symp. on Computer Vision, pp.551-556, 1995.
[107] T. Liu, H.J. Zhang, F. Qi, "A novel video Key-frame extraction algorithm based on perceived motion energy model," IEEE Trans. On Circ. Sys. And Video Tech., vol. 13, issue 10, pp.1006-1013, 2003.

[108] T.M. Cover, J.A. Thomes, “Elements of Information Theory,” New York, Wiley, 1991.
[109] Z. Li, G. Liu, “Video scene analysis in 3D wavelet transform domain,” Multimedia Tools and Applications, 2011.
[110] B. Boashash, “Time-frequency signal analysis and processing: A comprehensive reference,” Oxford, Elsevier Science, 2003.
[111] R.A. DeVore, B.J. Lucier, Wavelets, In Proc. of Numerica 92, A. Iserles, ed., Cambridge University Press, New York, pp. 1-56, 1992.
[112] Y. Meyer, “Wavelets,” In: Proceedings of Ed. J.M. Combes et al., Springer Verlag, Berlin, pp. 21, 1989.
[113] http://taco.poly.edu/WaveletSoftware/standard3D.html. Accessed 15 April 2011
[114] S. Lian, J. Sun, Z. Wang, “A secure 3D-SPIHT codec,” In Proc. of European Signal Processing Conference, pp. 813–816, 2004.
[115] E. P. Simoncelli, J. Portilla, “Texture characterization via joint statistics of wavelet coefficient magnitudes,” In Proc. of IEEE Int’l Conf. on Image Processing, vol. 2, pp. 62-66, 1998.
[116] J. Liu, P. Moulin, “Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients,” IEEE Trans. on Image Proc., vol. 10, issue 11, pp. 1647–1658, 2001.
[117] R. Moddemeijer, “On estimation of entropy and mutual information of continuous distributions,” Signal Proc., vol. 16, issue 3, pp. 233–246, 1989.
[118] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, 2001.
[119] Recommendation ITU-R BT.500-11 (2002) Methodology for the subjective assessment of the quality of television pictures.
[120] Recommendation ITU-R BT 500-6 (1994) Method for the subjective assessment of the quality of television pictures.
[121] X. Sun, M. Chen, A. Hauptmann, “Action recognition via local descriptors and holistic features,” In Proc. of IEEE Int’l Workshop on Computer Vision and Pattern Recognition, pp. 58-65, 2009.
[122] http://en.wikipedia.org/wiki/Wavelet. Accessed 14 September 2011
[123] M. N. Do, M. Vetterli, “Texture similarity measurement using Kullback-Leibler distance on wavelet subbands,” In Proc. of IEEE Int’l Conf. on Image Processing, vol. 3, pp.730-733, 2000.
[124] V. Kienzle, G.H. Bakir, M.O. Franz, B. Schölkopf, “Efficient approximations for support vector machines in object detection,” Pattern Recognition, Lecture Notes in Computer Science, vol. 3175, pp.54-61, 2004.
[125] F. Lu, X. Yang, W. Lin, R. Zhang, S. Yu, “Image classification with multiple feature channels,” Optical Engineering, vol. 50, page 5, 2011.

Abstract

Video signal has a major role in the transmission of visual information to a wide range of users in various applications. Video modeling and analysis have been of great interest in the video research community, due to their essential contribution to systematic improvements concerned in a wide range of video processing techniques. Parametric modeling and analysis of video provides appropriate means for processing the signal and mining necessary information for efficient representation of the signal. Video comparison, human action recognition, video retrieval, video abstraction, video transmission, video clustering are some of video processing applications that can get certain benefits from video modeling and analysis.

In this thesis, the parametric analysis and modeling of the video signal is studied through two schemes. In the first scheme, spatial parameters are first extracted from video frames and temporal evolution of these spatial parameters is investigated. Spatial parameters are selected based on the statistics of the 2D wavelet transform of the video frames, where wavelet transform provides a sparse representation of the signals and structurally conforms to the frequency sensitivity distribution of the human visual system. To analyze the temporal relations and progress of these spatial parameters, three methods are considered: inter-frame distance measurement, temporal decomposition, and AR (autoregressive) modeling. In the first method, employing the KL[23] distance between spatial parameters as the similarity measure, the temporal evolution of the spatial features is studied. This analysis is used to determine shot boundaries, segment shots into clusters and select keyframes properly based on both similarity and dissimilarity criteria, within and outside the corresponding cluster, respectively. In the second method, the video signal is assumed to be a sequence of overlapping independent visual components called events, which typically are temporally overlapping compact functions that describe temporal evolution of a given set of the spatial parameters of the video signal. This event-based temporal decomposition technique is used for video abstraction, where no shot boundary detection or clustering is required. In the third method, the video signal is assumed to be a combination of spatial feature time series that are temporally approximated by the AR model. The AR model describes each spatial feature vector as a linear combination of the previous vectors within a reasonable time interval. Shot boundaries are well detected based on the AR prediction errors, and then at least one keyframe is extracted from each shot. To evaluate these models, subjective and objective tests, on TRECVID and Hollywood2 datasets, are conducted and simulation results indicate high accuracy and effectiveness of these techniques.

In the second scheme, video spatio-temporal parameters are extracted from 3D wavelet transform of the natural video signal based on the statistical characteristics analysis of this transform. Joint and marginal statistics are studied and the extracted parameters are utilized for human action recognition and video activity level detection. Subjective and objective test results, on the popular Hollywood2 and KTH datasets, confirm high efficiency of this analysis method, as compared to the current techniques.

Keywords: Parametric video analysis and modeling, temporal decomposition, distance measure, AR model, 3D wavelet statistical modeling, keyframe selection, shot boundary detection, human action recognition.

[23] Kullback-Leibler

Sharif University of Technology
Electrical Engineering Department

Ph.D. Thesis in Communication Systems

Parametric Analysis and Modeling of Video Signal

By:
Mona Omidyeganeh

Supervisor:
Prof. Shahrokh Ghaemmaghami

Advisor:
Prof. Shervin Shirmohammadi

January 2012