Video content Analysis based on Full-Range Autoregressive Model using Multi-resolution Features

***M. Kamarasan**
Dept. of Computer and Information Science,
Annamalai University, Annamalainagar-608002, Tamilnadu, India

ABSTRACT -Over the past decade, the video content analysis is attracted by many scholars. In this paper, we present Multiresolution based Full Range Auto-Regressive model (FRAR) for video content analysis. The video signal is assumed to be a linear combination of a spatial-temporal independent component which is temporally approximated by FRAR parameters. The FRAR model yields a linear approximation to the temporal evolution of a stationary stochastic process. The 2D wavelet transform is incorporated to decompose the video frame into various subbands which represent the spatial-temporal multiresolution features. The spatial-temporal features are extracted from each subbands which is efficiently resemble the Human Visual System (HVS) characteristics. The FRAR model describes each spatial-temporal feature vector as a linear combination of the previous vectors within a reasonable time interval. Shot boundaries are well detected based on the FRAR prediction errors, and then at least one keyframe is extracted from each shot for further analysis. The experimental result shows that the proposed method obtains higher precision when compared to that of existing methods.

KEYWORDS: FRAR, wavelet transform, multiresolution, spatial-temporal, keyframe, subbands.

I. INTRODUCTION

Visual signal analysis (VSA) has been an emerging research area in the past two decades. The video signal can be represented as a sequence of an image which has both spatial and temporal features [1]. Spatial-temporal feature plays a vital role in many applications of video content analysis and produces successful results [2]. One of the crucial problems occurs in spatial-temporal is how to choose the proper spatial-temporal features according to their specific applications. Many combined spatial-temporal features based works are reported in [3][4]. For spatial features, the Color histogram has been frequently used and it is invariant to image rotation and translation, but it ignores the spatial organization of colors. Subsequently, Color correlogram describes the probability of finding color pairs at a fixed pixel distance and it provides an efficient spatial information. Color correlogram obtained better retrieval accuracy compared with color histogram. Huang et al [5] have proposed that the autocorrelogram is a statistical method which captures the spatial correlation between identical colors at a fixed distance and it improves the significant computational benefits over color histogram and color correlogram methods[6]. The color space generally plays a vital role in video content analysis, and particularly the selection of color space affects to the effectiveness of video content.

Partitioning a video sequence into a shot is the first steps in video content analyze [7]. Shots can be viewed as a sequence of interrelated consecutive frames in video and represent continuous action in temporal space. Shots features considered to be the vital role for higher level content analysis, shot classification, video indexing, video retrieval, scene change detection[10][11][12], etc. To associate the spatial-temporal relation, several time series modeling algorithm such as hidden Markov model (HMM)[8], Markov chain Monte Carlo (MCMC)[9], probability models with Bayes approach are taken into account for video content analysis on hand. Parametric model space such as Auto-regressive (AR) model, Moving-average (MA) model and auto-regressive-moving-average (ARMA) model are another way of approach for analyzing the spatial-temporal features. But the Autoregressive Moving Average (ARMA) models reveals in literature all of the finite order type and so contains only a finite number of parameters considered. As result, the future value would be influenced only by a limited number of past values and not have the ability to measure long range temporal relations and the relations in such features are hard to quantify. Since AR method analysis only short range relation of sequences of video and it is failed to capture the motion pattern in video frames. Further, most of the work in time series analysis are concerned with series having the property that the degree of dependence between observations, separated by a long time span, is zero or highly negligible In this paper, we employ a framework of Full-range auto-regressive model that exhibit the temporal characteristics of the video content in long range relation under analysis. This time-domain models used for the representation of discrete-time signals especially as parametric methods to estimate the covariance and the power spectral density of the stochastic processes. With the great ability to present the temporal motion pattern relation in the frame's spatial feature sequence, this framework is applied to several applications of video content analysis. The proposed work is mainly on the motivation of shot boundary detection, while it can also be applied to key-frame extraction and shot classification. The rest of the paper is organized as follows: Section –II, describes FRAR model for video content analysis The FRAR parametric distance method is discussed in section III. Spatial features extractions are discussed in in section IV. Experimental result is illustrated in section-V and concludes with conclusion in section VI.

II. FRAR MODEL FOR VIDEO CONTENT ANALYSIS

In this study, a Full-range autoregressive model [17] has been presented for shot boundary detection, while it can also be applied to key-frame extraction and shot classification. FRAM is a time-series model that effectively predicts granularity in full range spatial-temporal of sequence of frames in video analysis. Let *X* be a random variable that represents the intensity value of a pixel at location t of two dimensional image (frame). Then the full range autoregressive model is presented by differential equation

$$X_t = \sum_{r=1}^{\infty} a_r X_{t-r} + e_t \qquad (1)$$

where, $a_r = \dfrac{K\sin(r\theta)\cos(r\varphi)}{\alpha^r}$, r=1, 2, 3...

and K, α, θ and ϕ are real parameters. The $a_r s$ are the model coefficients which are computed by substituting the model parameters K, α, θ and ϕ in equation (1). The model parameters are interrelated closely to each other. It is assume that X_t will influence r for all positive values and $e1, e2, e3, \ldots$ independent and identical distribution (iid) normal random variables with mean 0 and variance σ^2. The presented model is employed to analyse video frame with size L × L. The frame(image) is partitioned into various sub image (subframes) with size M × M (M < L), to spatially characterize the nature and structure of the frame. With the Markovian assumption, the conditional probability of *X*(*s*) given all other values only depends upon the nearest neighbourhood values. It is assumed that the parameters are K ∈ R, α > 1, and θ, ϕ ∈ [0, 2π]. Thus the presented FRAR model incorporates the full range dependency and it involves four parameter and totally evacuates the problem of order determination. As result, FRAR influence only a few parameters with full range of values so it captures minute's discontinuity between the frames.

In order to estimate model parameter to analysis spatial-temporal relation, we decomposed frame i into various sub components using wavelet transform [18] and extract colour and texture on each component which is discussed in section 4. Suppose $\mathrm{fv_i} = \left[\mathrm{fv_{i,1}\ fv_{i,2}\ fv_{i,3}..fv_{i,n}}\right]$ is feature vector extracted from ith frame. The FRAR model is used analysis $a_{i,j}$ parameter, such as

$$fv_{ij} = \sum_{l=1}^{p} a_{i,j}\, fv_{i-l,j} + e_{i,j} \qquad (2)$$

Where $e_{i,j}$ is residual error and full range autoregressive prediction error (FAPE) value, which is used detect short boundary in a sequence of frames and then select the keyframe, is computed as

$$FAPE = \sum_{l=1}^{n} \left|e_{i,l}\right| w_j \qquad (3)$$

where w_j is weight of spatial related features of the current frame. We used Recursive Least square (RLS) algorithm[19] to predict FAPE value to detect the granularity changes in sequence of frames over a time space.

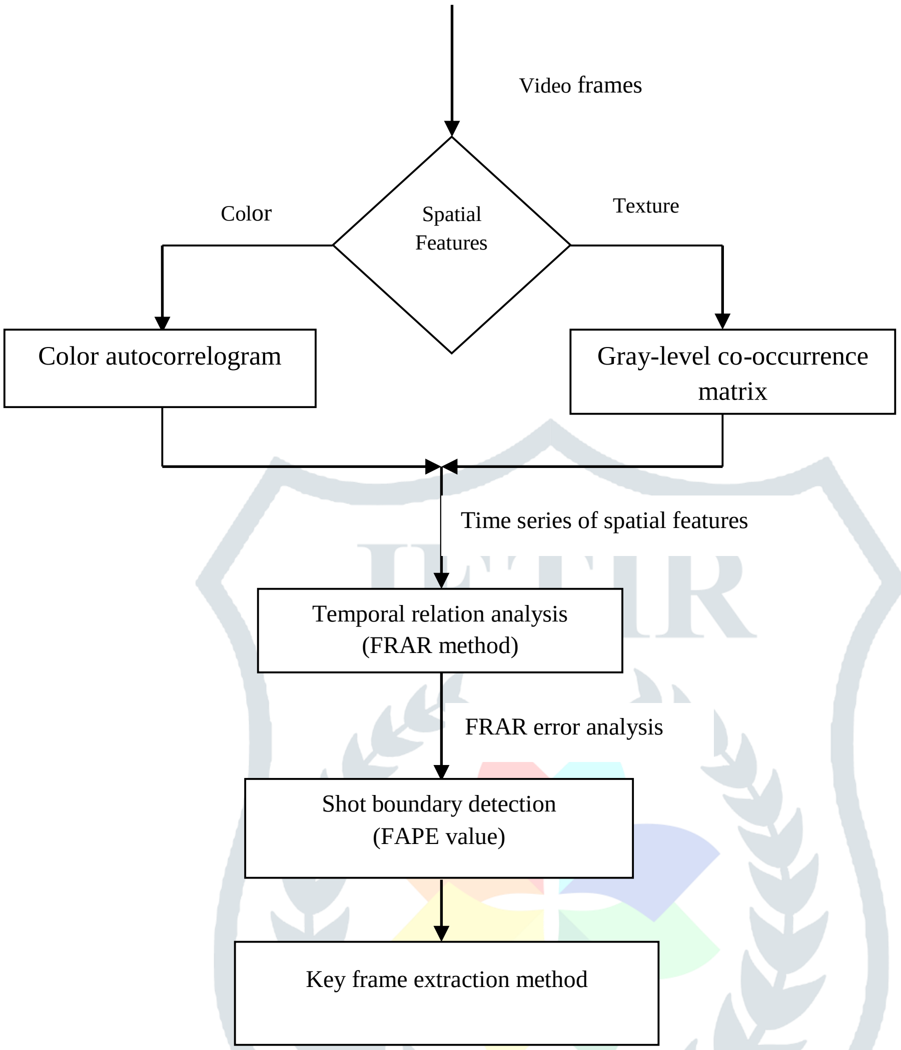

Video frames
Color
Spatial Features
Texture
Color autocorrelogram
Gray-level co-occurrence matrix
Time series of spatial features
Temporal relation analysis (FRAR method)
FRAR error analysis
Shot boundary detection (FAPE value)
Key frame extraction method

Fig.1 FRAR key frame extraction system

III. DISTANCES BETWEEN FRAR PARAMETRIC MODELS

In order to evaluate the temporal relation between video frames that can represent underlying visual features, distance method Minkowski used for spatial features between the frames. It is a simple method leading to very efficient computation, which in turn makes image ranking scalable (a quality that greatly benefits real-world applications) and is given in equation

$$D(f^q, f^t) = \sum_{i=1}^{N} |f^q(i) - f^t(i)| \quad (4)$$

where $D(f^q, f^t)$ represents distance value between frames $f^q(i), f^t(i)$ represent subband of the frames.

Wavelet Transformation

The Daubechies wavelet[18] transforms are defined in the same way as the Haar wavelet transforms by computing running averages and differences via scalar products with scaling signals and wavelets, the only differences between them consists in how these scaling signals and wavelets are defined. The Daub-4 wavelet transforms is defined in essentially the same way as the Haar wavelet transform.

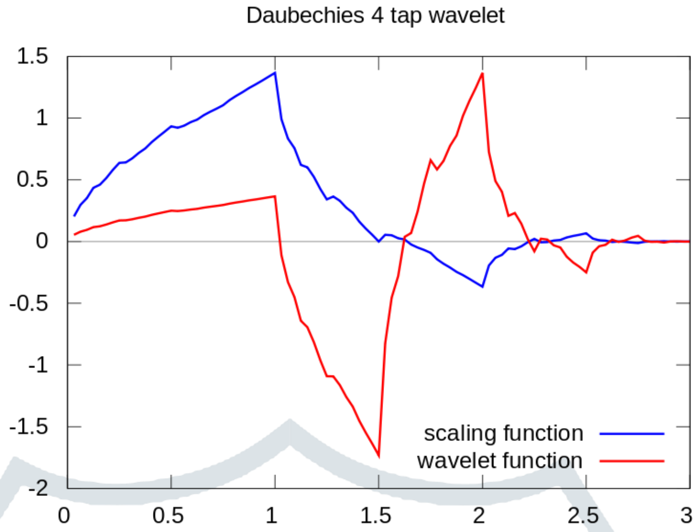


Fig.2 Daub-4 Wavelet Transform

The Daub-4 wavelet transform are computed a set of four approximation coefficients h_0, h_1, h_2, h_3 and g_0, g_1, g_2, g_3 detailed coefficients respectively. The approximation function coefficients are

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}} \quad h_1 = \frac{3+\sqrt{2}}{4\sqrt{2}} \; h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}} \quad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}} \tag{5}$$

The detail function coefficients are

$$g_0 = h_3; g_1 = -h_2; \; g_2 = h_1; g_3 = -h_0 \tag{6}$$

The approximation (scaling) value, a_i and detail (wavelet) value c_i are computed by taking the inner product of the h_i and g_i coefficients and signal s_i. The inner product equations are defined as

$$a_i = h_0 s_{2i} + h_1 s_{2i+1} + \; h_2 s_{2i+2} + h_3 s_{2i+3}$$
$$a[i] = h_0 s[2i] + h_1 s[2i+1] + h_2 s[2i+2] + h_3 s[2i+3] \tag{7}$$

Daub-4 wavelet function coefficients are

$$c_i = g_0 s_{2i} + g_1 s_{2i+1} + \; g_2 s_{2i+2} + g_3 s_{2i+3}$$
$$c[i] = g_0 s[2i] + g_1 s[2i+1] + g_2 s[2i+2] + g_3 s[2i+3] \tag{8}$$

For computational convenience, the given input color image I with size $N \times N$ is converted to a sequence $s(n)$ where $n = N \times N$. Then the input image $s_1, \ldots, s_{n-1}$ contains s_N coefficients, where there will be $N/2$ approximations and $N/2$ wavelet coefficients. The approximations are stored in half of the higher order array [a₀...a₃] and wavelet coefficients are stored in another half of the lower order array [c₀…c₃]. Subsequently, half of the higher order array which is approximation coefficient becomes the input for the next step in the wavelet computation until the optimum results obtained . Based on these concepts, we have presented a sample 2-level wavelet decomposition image in fig.4.

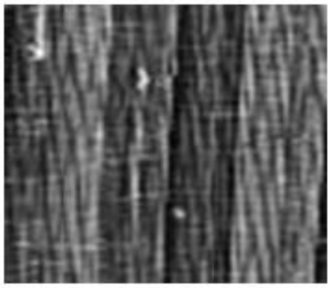


Fig.4(a)Original image

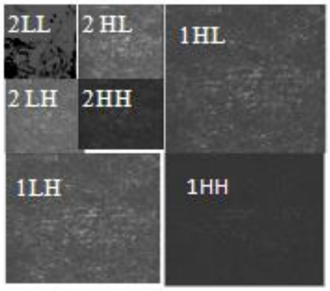


Fig.4.(b)2-level decomposition of original image using Daub4

IV. SPATIAL FEATURE EXTRACTION

In this section, we presented two kinds of spatial features which are incorporated in the underlying work: Color scalable descriptor (SCD) and spatial dependency matrix which are strappingly relay on spatial information in video content.

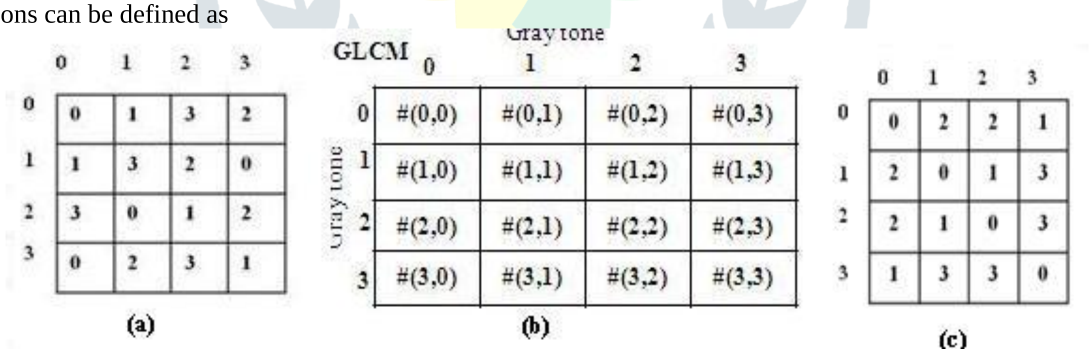
Scalable color descriptor (SCD)

Scalable Color Descriptor (SCD) is in the form of a Color Histogram in the HSV Color space encoded using a Haar transform. The generic color histogram method is a compound descriptor in MPEG-7 which consists of color space and quantization, and histogram descriptors. This would allow the specification of color histograms with varying numbers of bins and nonuniform quantization of different color spaces. However, it is not desirable to provide too much flexibility in such a specification, as it would limit interoperability between different descriptions based on MPEG-7. Meanwhile SCD addresses the interoperability issue by fixing the color space to the HSV which is uniformly quantized with 256 bins. This descriptor uses Haar transform and its representation is scalable in terms of bin numbers.

In this work, SCD consider to extract feature extracted at the upper co-efficient of the wavelet transform in sequence of frames. It applies discrete wavelet transform on the each frame as a result the discrete wavelet transform reveals in three detailed components and one approximation [18].

Gray-level spatial dependence matrix

In the proposed work used the Haralick[10] method which is used to examine the texture that contains spatial relationship of pixel element in form of gray-level cooccurrence matrix(GLCM). These texture features lead to achieve better results in video content analysis. The idea of the method is to consider the relative frequencies for which two neighboring pixels are separated by a distance on the frame and it collect information about the pair of pixels instead of a pixel which is also known as second order statistics[1]. But, it measures the relationship between groups of two pixels (usually neighbors), whereas first order statistics considers single pixel. Let us suppose an image I to be analyzed in rectangular and has N_x resolution cells in the horizontal direction and N_y resolution cells in the vertical direction and the gray tone appearing in each resolution cell is quantized to N_g levels. Let $L_x = \{1, 2, ..., N_x\}$ be the horizontal spatial domain, $L_y = \{1, 2, ..., N_y\}$ be the vertical spatial domain and $G = \{1, 2, ..., N_g\}$ be the set of N_g quantized gray tones. The set $(L_x \times L_y)$ is the set of resolution cells of the image I ordered by their row-column designations can be defined as



	0	1	2	3
0	0	1	3	2
1	1	3	2	0
2	3	0	1	2
3	0	2	3	1

(a)

GLCM	0	1	2	3
0	#(0,0)	#(0,1)	#(0,2)	#(0,3)
1	#(1,0)	#(1,1)	#(1,2)	#(1,3)
2	#(2,0)	#(2,1)	#(2,2)	#(2,3)
3	#(3,0)	#(3,1)	#(3,2)	#(3,3)

Gray tone
Gray tone

(b)

	0	1	2	3
0	0	2	2	1
1	2	0	1	3
2	2	1	0	3
3	1	3	3	0

(c)

Fig.5. (a) 4x4 image with four gray-tone values 0-3 (b) General form of any gray-tone spatial-dependence matrix for image with gray-tone values 0-3. #(*m*, *n*) stands for number of times gray tones *m* and *n* have been neighbors.(c) Spatial occurrence calculation of Horizontal direction(0°)

The proposed work consider only four texture features in Haralick method which is computed on GLCM of $W_{m,n}^{V}$,

Contrast

$$f_1 = \sum_{i,j} |i-j|^2 \, p(i,j) \tag{9}$$

Correlation

$$f_2 = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \tag{10}$$

Angular second moment

$$f_3 = \sum_{i,j} p(i,j)^2 \tag{11}$$

Diagonal distribution

$$f_4 = \sum_{i,j} \frac{p(i-j)}{1+|i-j|} \tag{12}$$

where μ_i *and* μ_j represent mean value; σ_i*and* σ_j indicates the standard deviation. Feature f_1 measures the intensity of contrast between a pixel and its neighborhood over the $W_{m,n}^{v}$ subbands, and it is used for indentify the change of gray levels in a texture; f_2 measure correlates a pixel to its neighborhood over $W_{m,n}^{v}$, f_3 which denotes the sum of squared elements over the $W_{m,n}^{v}$ and is also recognized as uniformity, non-uniformity of energy and angular to moment, and finally, f_4 measures the nearness of distribution of elements in the $W_{m,n}^{v}$ to the $W_{m,n}^{v}$ diagonal subbands. Based on above texture features which is extracted each frame and texture feature vector f_T is

$$f_T = [f_1, f_2, f_3, f_4] \tag{13}$$

Shot boundary detection

Shot boundary plays an important role in order to analysis content of video which composed of sequence of frames taken by single camera without interruption. Shot boundary is fundamental step to any kind of video analysis and its application is enables the video segmentation into its basic unit. The proposed FRAR model enables shot transitions based on FAPE score, and subsequent task is whether the detected FAPE is abrupt or gradual transition. Comparatively, detecting gradual change is difficult than an abrupt cut. The reason is that, abrupt shot boundary where the change takes place over a single frame and gradual transition where the change of video content occurs over a sequence of frames gradually [13]. Abrupt boundaries between shots generally have sudden intensity changes whereas as graduation transition mainly includes: fade-in, fade-out and dissolve.

To detect the abrupt boundary, we use FAPE as the criteria. In the FRAR model, FAPE can measure the accumulated errors as long as the orders of the model. When the APE grows big, it means the present model's parameter cannot fit the current frame well, and a shot change may occur. In order to detect possible shot cuts, we employ an approach by selecting adaptive threshold on that parameter:

$$T_p = \mu \times \sigma \tag{14}$$

where μ is the local mean value of APE of sequence of frames size p, σ represent standard deviation it takes values in the range of 3 to 5 according to our empirical study. To combine FRAR model with scalable color descriptor and GLCM based Haralick texture features which exhibits the new spatial-temporal relationship to represent the region of interest(ROI) in sequence of frames as result which captures the minutes texture and structures in frames to improves the shot boundary detection.

In the gradual transitions, the frame sequences change gradually, according to progress in the FAPE values and relations among frames. The figure-6 shows the FRAR model detect key frame in shot boundaries based on FAPE score value.

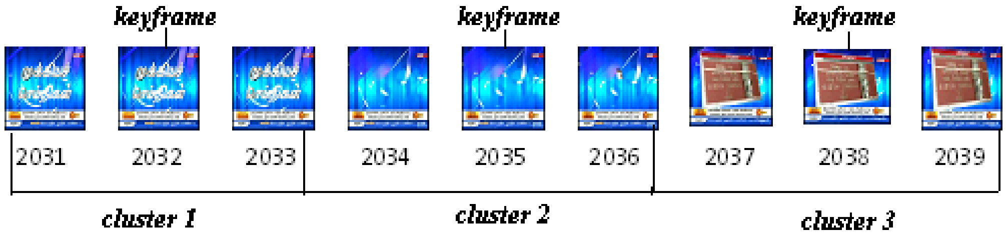


Fig.6 Key frame extraction from shot boundaries

V. EXPERIMENTAL RESULT

The proposed wavelet based FRAR method has been tested and applied numerous video clips of more than 95000 video frames with different characteristics. Most of our test videos are selected randomly from movie human Actions and scenes dataset (CVPR09) [16], which is shown in Table 1. The frame size is set to CIF/QCIF with 324x289 and the sampling rate is taken to be 15 or 24 frames per second. Video tests are captured from different locations – indoor and outdoor- of different characteristics. They display tennis, football, human actions, car racing, news broadcasting, airplane flying, advertisement, etc. Also, they contain camera zooming, panning, translations and scene fade in/outs. Most of the test videos are very dynamic in both temporal and spatial domains, though containing few static samples also. The movies and advertisement contains all kinds of shot boundaries such as cut, fade in/out and dissolve.

In order to evaluate the efficiency of the proposed FRAR method, the performance fair comparisons are made with parametric model for video content analysis [3] and autoregressive video modeling through 2D wavelet statistics [4]. In the first method detect the gradual shot boundaries based on histogram properties with parameter estimation and the second case, they used Auto Regressive (AR) model to detect the shot boundaries. The proposed work incorporated full range autoregressive model based on wavelet subband spatial-temporal features with parameter estimation to detect shot boundaries based on FAPE values. Precision and recall methods [13] are used to evaluate the performance of a shot boundary detection. Recall measure the ratio of correct shot detection over the number of all correct shot detection whereas the precision ratio of correct shot boundaries detection over the number of all detections, they defines as

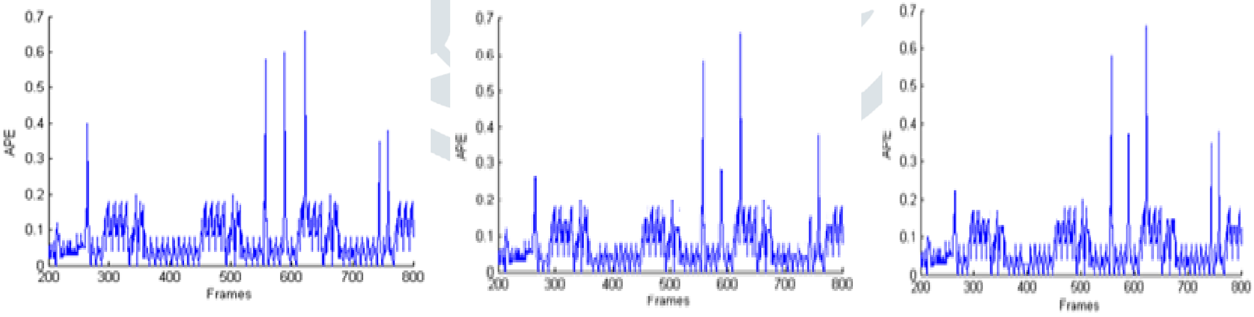
$$recall = \frac{no.of\ hits}{no.of\ hits + no.of\ misses}$$

$$precision = \frac{no.of\ hits}{no.of\ hits + no.of\ \ false\, alarms} \tag{15}$$

Relationship between the presented FRAR model and motion technique, the FRAR model constructed under time-series and motion picture technique also based on temporal non-stationary signal. Features such as color and texture are extracted under the spatial-temporal domain.

Table-1. Numbers of frames, shots, cuts, fade in/outs and dissolves in the test video

Video	Frame	Cut	Shot	Fade in/out	Dissolve
News -1	12409	234	230	23	13
News-2	15345	167	162	19	19
Cartoon	21046	312	310	0	0
Football	46212	432	423	0	0
Advertisement	9000	123	119	7	6



method1(proposed) method2 method3

Fig.7 Comparison measure of FAPE based shot boundaries detection with proposed method with other two methods.

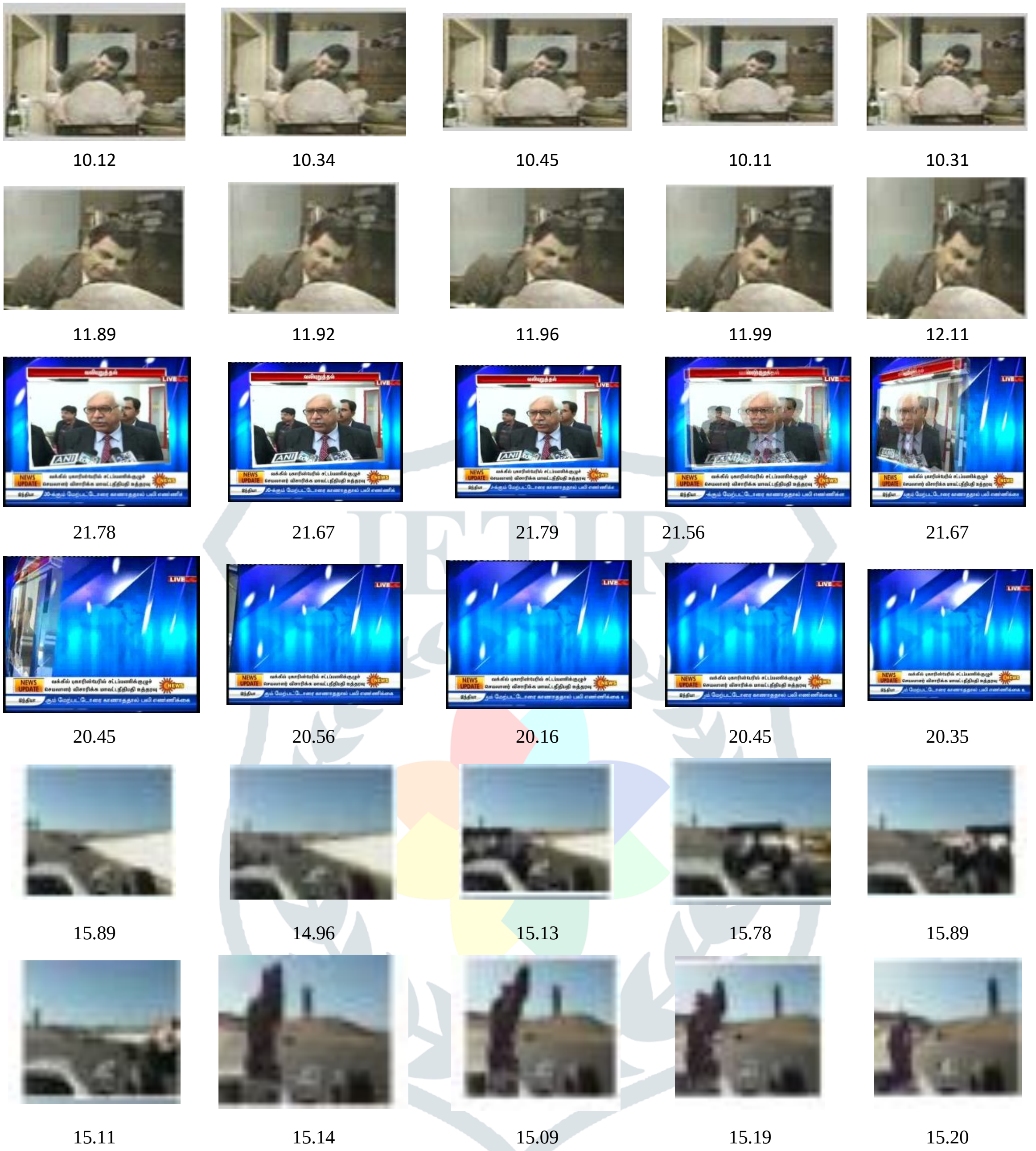


Fig.8 Temporary down sampled video sequence with FAPE values

The Table 2 shows the shot boundary detection obtained by our method and other twos. The proposed method is based on color autocorrelogram and GLCM based texture features.

Table-2. Detection performance of the proposed method with other methods when applied to various test video.

		cut		Fade in/out		dissolve	
		Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
News-1	M1	94.31	96.12	94.43	96.87	89.34	90.99
	M2	91.23	94.89	89.56	93.56	87.87	88.34
	M3	89.67	92.76	92.24	93.23	88.45	89.34
News-2	M1	92.13	94.28	88.41	90.78	76.82	78.83
	M2	89.54	93.02	83.66	84.64	73.61	74.00
	M3	86.98	90.67	82.42	85.67	71.58	72.45
cartoon	M1	72.13	78.28	88.41	90.78	76.82	78.83
	M2	76.54	78.89	N	N	N	N
	M3	73.89	71.32	N	N	N	N
football	M1	91.00	93.90	95.57	96.12	88.65	89.13
	M2	87.34	88.45	91.98	92.76	89.23	88.12
	M3	83.39	84.90	90.17	90.54	91.67	89.10
Advertisement	M1	96.00	97.90	92.75	96.12	89.62	91.63
	M2	94.87	95.19	90.81	93.67	91.63	87.28
	M3	95.93	96.03	95.09	91.54	93.79	88.34

Table-3. Average no.of keyframes are extracted from various kinds of test video

Types of video	News1	News2	cartoon	football	advertisement
No. of key frames	2.9	3.7	1.8	1.5	1.7

VI. CONCLUSION

In this paper, Full range autoregressive model (FRAM) is incorporated to analysis the video content in spatial-temporal domain whereas the traditional methods hardly to find the spatial-temporal relation in video sequences. FRAM is a good predictor over time with limted parameters but not sensitive to motion or noise. The proposed system used FRAM prediction errors with parameter estimation to detect the short boundaries and then classify the scenes based on FAPE value. FRAR method used scalable color descriptor(SCD) and Gray-level co-occurrence matrix(GLCM) based texture features as input to the model which is highly represent the spatial-temporal content with semantic relation. By combining color and texture features which are extracted based on spatial temporal based wavelet transform, the proposed method shows robustness result in key frame selection and then applied to short and scene classification. The proposed method achieves high performance in terms of precision and recall when compared to that of existing methods.

REFERENCES

[1] Li, Y., Lee, S.H., Yeh, S.H., Kuo, C.-C.J.2006, Techniques for Movie Content Analysis and Skimming. IEEE Signal Processing Magzine 23:79-89 doi: 10.1109/MSP.2006.1621451.

[2] Hicham, G.E., Mohamed, F.M., Walid, G.A., 2005. Spatio-temporal histograms. Lect. Note Comput. Sci. 3633, 19–36.

[3] W. Chen and Y.J. Zhang, 2008, "Parametric model for video content analysis", *Elsevier B.V., Pattern Recognition Letters, vol.* 29, pp. 181–191.

[4] *M. Omidyeganeh M, S. Ghaemmaghami, S. Shirmohammadi.2013,* Group-based spatio-temporal video analysis and abstraction using wavelet parameters, journal of signal, image and video processing,vol.7.issue 4,pp. 787-798.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang,2008 "Image retrieval: ideas, influences, and trends of the new age", ACM Computing Surveys 40 (2), pp.1–60.

[6] Huang, S. Kumar, M. Mitra & W. Zhu,1998," Spatial color indexing and applications.", In: Proc. Sixth International conference on Computer Vision (Bombay, India), pp. 602-607.

[7] Alan Hanjalic, 2002, Shot-Boundary Detection: Unraveled and Resolved?, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 2.

[8] Xu, G., Ma, Y.F., Zhang, H.J., Yang, S.Q., 2005. HMM-based framework for video semantic analysis. IEEE Trans. Circuits Syst.Video Technol. 15 (11), pp1422–1433.

[9] Zhai, Y., Shah, M., 2006. Video scene segmentation using Markov chain Monte Carlo. IEEE Trans. Multimedia 8 (4), pp686–697.

[10] C.G.M Snoek and M.Worring,2005 Multimedia event-based video indexing using time interval, IEEE Trans. On Multimedia,vol.7,no.4, pp.638-647.

[11] J.fan, A.K Elmagarmid, X.Zhu,W.G Aref and L.Wu ,2004, Class View:Hierarchical video shot classification, indexing and accessing. IEEE Trans. On multimedia, vol.6,no.1, pp.70-86.

[12] C.W.Ngo,T.C Pong and H J Zhang.2002, On clustering and Retrieval of video shots through temporal slices analysis. IEEE Trans. On multimedia, vol.4, no.4, pp.446-458.

[13] Lu, H., Tan, Y.P., 2005. An effective post-refinement method for shot boundary detection. IEEE Trans. Circuit Syst. Video Technol. 15 (11), 1407–1421.

[14] Cai, C., Kin, M.L., Zheng, T., 2005. A unified shot boundary detection method based on linear prediction with bayesian cost function. IEEE Int. Workshot VLSI Design Video Technol., pp.101–104.

[15] Cernekova, Z., Pitas, I., Nikou, C., 2006. Information theory-based shot cut/fade detection and video summarization. IEEE Trans. Circuit Syst.Video Technol. 16 (1), pp.82–91.

[16] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html.

[17] Karthikeyan, T and R. Krishnamoorthy., 2012. ICTACT Journal on Image and Video Processing, August 2012, Volume: 03, Issue: 01.

[18] Ingrid Daubechies, 1992, Ten Lectures on Wavelets, Society of industrial and applied mathematics, Philadelphia Pennsylvania.

[19] Simon Haykin, 2002, *Adaptive Filter Theory*, Prentice Hall.