RBF based Spatio-Temporal Representation Technique for Video Compression

Santanu Chaudhury
Indian Institute of Technology Delhi
EED, Hauz Khas
New Delhi
schaudhury@gmail.com

Brejesh Lall*
Indian Institute of Technology Delhi
EED, Hauz Khas
New Delhi
brejesh@ee.iitd.ac.in

Mona Mathur
ST Microelectronics
Plot #1, Knowledge Park III
Greater Noida – 201308
Uttar Pradesh
mona.mathur@st.com

Kartik Mehta
Indian Institute of Technology Delhi
IIT Delhi, Hauz Khas
New Delhi
kartikmehta.iitd@gmail.com

ABSTRACT
Parametric coding is a technique in which data is processed to extract meaningful information and then representing it compactly using appropriate parameters. Parametric Coding exploits redundancy in information to provide a very compact representation and thus achieves very high compression ratios. However, this is achieved at the cost of higher computation complexity. This disadvantage is now being offset by the availability of high speed processors, thus making it possible to exploit the high compression ratios of the parametric video coding techniques. In this paper a novel idea for efficient parametric representation of video is proposed. We perform Oct-Tree Decomposition on a video stack, followed by parameter extraction using Radial Basis Function Networks (RBFN) to achieve exceptionally high compression ratios, even higher than the state of art H.264 codec. The proposed technique exploits spatial-temporal redundancy and therefore inherently achieves multi-frame prediction.

Keywords
Spatio-temporal modeling, Radial Basis Functions, Oct-tree decomposition.

1. INTRODUCTION
The paper presents a parametric video compression scheme using Radial Basis Function Networks (RBFN) to model homogeneous spatio-temporal video segments. The scheme exploits the spatio-temporal redundancy of large homogeneous video segments to achieve high compression ratios. A Parametric Model can be constructed by using any of the following two methods: (i) Model Learned/Formulated apriori & (ii) Models Generated On-the-Fly [[1]], [[2]]. Irrespective of the method used, a model can either be a Geometric/Structural model like Mesh based model, or an Appearance model like Eigen-space or Probabilistic models [[3]],[[4]]. In comparison to these techniques, we propose a parametric model which is generated on the fly, does not employ any region based segmentation (which is prone to error), or makes any assumptions about the camera position or the scene geometry.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICVGIP '10, December 12–15, 2010, Chennai, India.

Copyright 2010 ACM 978-1-4503-0060-5/10/12 …$10.00.

The technique can be applied to any kind of video and automatically selects the regions to be learnt by the parametric model. In a video shot, the technique identifies large regions, homogenous in space and time and models them using the RBF learning functions. Modeling of large homogenous spatio-temporal volumes through small set of the RBFN parameters and minimal overheads allows for much higher compression ratios than those achieved by frame based techniques. Besides, the architecture of the scheme has been so defined that it fits naturally into a very popular current video coding standard H.264.

This paper is organized as follows: Section 1 is the introduction. Section 2 introduces the proposed hybrid coding technique. Section 3 describes the spatio-temporal learning model . Section 4 contains simulation results, and Section 5 is the conclusion.

2. HYBRID CODING MODEL
In this section we propose a hybrid coding model which fits very well with H.264 video coding scheme. The proposed coding model is based on an on the fly learning model which is based on a very intuitive way of representing video shots allowing for an efficient compression.

In any videos shot, there are more or less gradual changes in the scene and the objects over time. The pixel intensities in the sequential frames of a video shot are inter-related to form generally smoothly varying intensity surfaces. The idea is to use the gradual changes in the image intensity over space and time to learn a model of the surface as a combination of several functions. Since the whole video shot would be too complex to be represented as a combination of functions , a video is subdivided into many spatio-temporal segments and each of the segment is then modeled using function approximation learning algorithm.

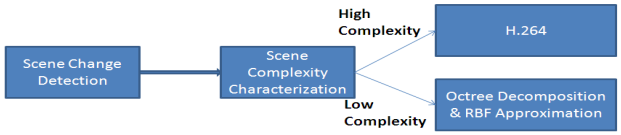

Figure 1 The Hybrid Coding Model

The complete hybrid video coding scheme is shown in Figure 1. In the first stage video sequence is partitioned into shots/scenes to achieve the goal of video-content analysis and content-based video coding. The second stage is classification of the shots thus obtained into complex and simple shots based on their motion and image complexity. This is done as the learning model is amenable to modeling smoothly changing variation in space and time. Relatively simpler shots are decomposed into homogenous blocks using the octree decomposition approach. The larger 3-D volume

segments obtained after decomposition are modeled using RBF parameters, while 3D-DCT is used for representing smaller volumes. Sequences classified as high complexity are encoded using H.264.

Local chi-square scheme which takes into account the degree of brightness and spatial information within a frame is used to successfully determine scene changes [5]. The distance measure between two frames i and j is given by:

$$d(f_i, f_j) = \sum_{bl=1}^{m} d_{x^2}(f_i, f_j, bl)$$

$$d_{x^2}(f_i, f_j, bl) = \sum_{k=1}^{N-1} \left(\frac{(H_i^r(k) - H_j^r(k))^2}{\max(H_i^r(k), H_j^r(k))} \times \alpha + \frac{(H_i^g(k) - H_j^g(k))^2}{\max(H_i^g(k), H_j^g(k))} \times \beta + \frac{(H_i^b(k) - H_j^b(k))^2}{\max(H_i^b(k), H_j^b(k))} \times \gamma \right)$$

where $H_i^r(k)$, $H_i^g(k)$ and $H_i^b(k)$ refer to the histogram distribution of the i-th frame (f_i) for each color space. N refers to the number of bins (k), and m refers to the total number of blocks (bl). In addition α is defined as 0.299, β as 0.587, and γ as 0.114. This method allows for detection of abrupt as well as gradual scene changes.

Scene Complexity can be divided into two parts: Motion Complexity & Image Complexity. Motion complexity is determined by the edge change ratio (ECR) proposed as a characteristic feature by Zabih et al [[6]]. ECR registers structural changes in the scene such as the entry, exit, and motion of objects as well as fast camera operations and is somewhat independent of variations in color and intensity since it relies only on sharp edges. The edge change ratio ecr_n between the frames n-1 and n can be defined as:

$$ecr_n = max\left(\frac{X_n^{in}}{\sigma_n}, \frac{X_{n-1}^{out}}{\sigma_{n-1}}\right)$$

where σ_n be the number of edge pixels in frame n and X_n^{in} and X_{n-1}^{out} the number of entering and exiting edge pixels in frame n and n-1, respectively. Edges are also representative of amount of detail in the image and hence can be used a measure for image (frame) complexity. In this work we have obtained the measure of frame complexity using the following steps:

- Apply edge detector on frame. (with proper threshold so that faint edges are not detected)
- Uniformly divide the frame into sub-blocks. (appropriate size depending on the size of frame)
- Count the number of blocks which contain edge pixels.
- Divide the value obtained in above step by total number of blocks. This ratio is a measure of frame complexity.

3. SPATIO-TEMPORAL MODELING OF VIDEO SHOTS

In this section we detail the proposed parametric model which has the capability to provide very high compression gains by modeling videos in the spatio-temporal domain. The technique

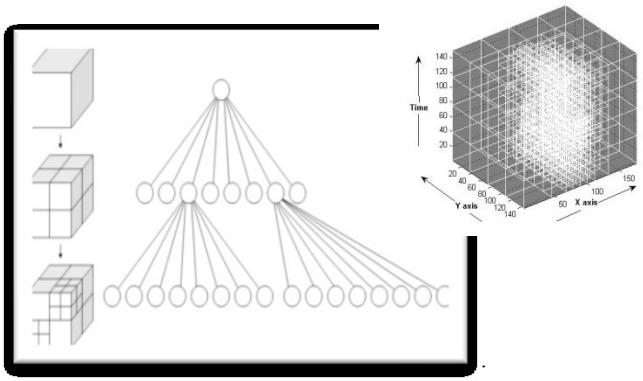

Figure 2 (Left) Octree Decomposition technique and (Right) octree decomposition of the video hybridtest.

is based on a simple yet novel idea of representing video as a function of spatial (x, y) and temporal (t) dimensions. Authors in [[7]] discuss a technique of modeling images using parametric model based on RBFN. We use a similar technique for learning the video content, along both space and time, in homogenous segments and modeling the segments using RBF networks.

Low image and motion complexity video shots are modeled using a spatio-temporal on the fly learning model where in the video shot is represented as a function of the spatial (x, y) and temporal (t) dimensions. Representing an entire video shot as a function of (x, y, t) is sub-optimal since the volume corresponding to the entire video may not be homogenous, whereas, sub-sets of it might be. So the video shot is decomposed into homogenous 3D blocks and each spatio-temporal volume is modeled independent of the others. The shot is decomposed into homogenous 3D volumes using octree decomposition technique. A video segment thus obtained represents a relatively smooth volume.

Octree Decomposition of the video shot

An octree is the three-dimensional analog of quadtree where a three-dimensional space is recursively divided into eight octants based on a predefined homogeneity criterion (e.g., if all the pixels in the block are within a specific dynamic range). To obtain octree decomposition, first the whole video shot is considered as one region. If the variations in the intensity are more than the defined homogeneity criteria then the shot is divided into eight equal segments (Figure 2). This regular subdivision continues till each of the 3D segments obtained satisfy the homogeneity criteria. The homogeneity criterion used in this work is the variance of luminance values in the region of interest. A key advantage of the oct-tree decomposition is that the overheads in octree representation are small as it has a very regular structure. Figure 2 shows an octree decomposition of a video sequence. It was found that for very small sized homogeneous regions the compression offered by modeling using RBFN was poor, hence the oct-tree decomposition is not performed for block sizes smaller than a chosen minimum size.

Learning spatio-temporal blocks using RBFN

After a video is segmented into 3D homogenous blocks as described above, each segment represents a smoothly varying function in space and time. RBFN are known to provide universal approximations on a compact subset of $\mathbb{R}^n$ [[8]]. This means that RBF network with enough hidden neurons can approximate any

continuous function with arbitrary precision. We employ these functions to model each of the homogenous segments as a finite sum of Radial Basis Functions. Each spatio-temporal block is modeled using a three-layered RBF neural network that uses Gaussian activation functions and a Euclidean norm. The output of the network is represented as:

$$f(x_i) = \sum_{j=1}^{N} a_j \exp\left[-\beta_j \|x_i - c_j\|^2\right]$$

where N is the number of neurons in the hidden layer, c_j is the centre vector for neuron j, and a_j are the weights of the linear output neuron. In the basic form all inputs are connected to each hidden neuron.

Given that x_i is the n-dimensional input vector, if y_i be the desired value of the function at point x_i, then we need to find the function $f(x_i)$ such that $f(x_i) = y_i$. The optimum values of a_j, c_j and β_j are to be determined such that the following sum of squared errors is minimized.

$$E = \sum_{i=1}^{P} (f(x_i) - y_i)^2$$

The number of functions required to represent a region is determined such that the error E above is kept below a threshold. The error equation above is a regression equation which can be solved using various regression methods (for example normal equations). For the results presented in this paper we have used 1% of the total data as training data and Bayesian Information Criterion (BIC) has been used as the model selection criterion. The data in a homogenous volume is modeled using a regression tree. The nodes of this tree are then used for determining the centres and radii. A subset of these is selected as the final RBFs using Plain Forward Selection.

The number of RBFs chosen for modeling a given spatio-temporal segment depends on the variation in pixel intensities and the smoothness of the surface being learnt. For the case where the 3D volume being learnt is small the benefit gained by modeling using the basis function is offset by the number of parameters required to be optimized, i.e. the centers, weights and widths of the functions. So the smaller spatio-temporal regions below a minimum volume threshold were represented using a 3D-DCT. RBFN based parametric representation is applied to the large volume spatio-temporal segments of the video, while smaller volumes are represented using 3D DCT. The proposed scheme is easily integrable into the existing H.264 video codec implementation. A mechanism for integrating the proposed encoding scheme with H.264 is shown in Figure 1.

4. EXPERIMENTAL RESULTS

The compression results of the proposed scheme for the following three videos are presented here: Bridge far, Highway (CIF, 100 frames each), which are MPEG-4 Standardization sequences, and a hybrid test (non standard) video with 400 frames composed of multiple video shots.

A detailed study of the hybrid test stream video containing multiple shots was carried out. The different shots, detected using the local-chi square technique, are shown in Figure 3(a). The obtained video shots are characterized for their motion and image complexity. As can be seen form Figure 3 (a) and (b) the first and fourth shot have low scene and motion complexity and were coded using our scheme, whereas the second and third shots had high complexity and were encoded using H.264 codec.

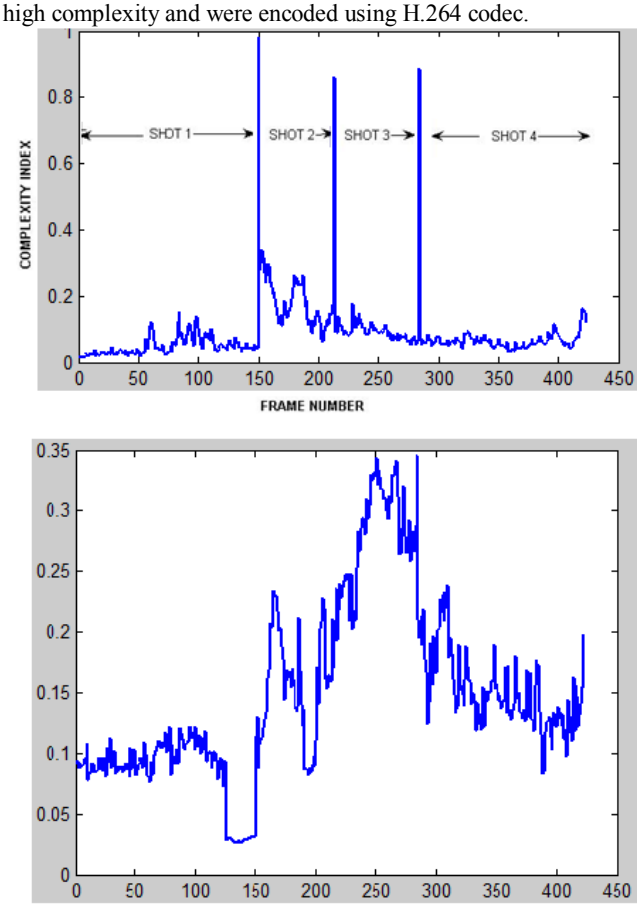

Figure 3 (a) Motion Complexity (b) Image / Frame Complexity

Table 1 gives a summary of the compression gains achieved in the 4 shots. The table clearly illustrates that the proposed scheme performs well for low complexity shots / scenes and overall the hybrid coder provides higher compression gain of 25.92 as compared to an overall compression gain of 7.6 achieved with standard H.264 video codec.

Table 1 Gains for different shots of hybrid encoder

Shots	Hybrid Scheme Gain	H.264 gain
1	53.30 (our scheme)	22.45
2	10.35 (H.264)	10.35
3	9.19 (H.264)	9.19
4	12.20 (our scheme)	8.99

The compression results obtained for the three sequences are compared against the H.264 codec in Tables 2 to 4. The input streams to both the proposed codec and H.264 are in the YUV 4:4:4 format. For obtaining the comparison, RBF parameters were coded using 8 bits per parameter and for the remaining region was coded using H.264 with the RBF regions replaced by black. The efficacy of RBF as tool for modeling homogeneous volumes is clearly illustrated by the compression gains achieved in the

regions where the RBF has been applied versus the gain that is achievable on the entire sequence using the state of the art H.264 codec. These compression results along with PSNR values for the regions encoded using RBF are summarized in Table 5.

The gain obtained by the proposed scheme is achieved because of the suitability of the spatio-temporal model i.e. the RBF learning model as a video compression tool. The proposed mechanism is intuitively appealing because it inherently achieves multi-frame prediction and this is clearly borne out by the performance figures obtained for the various sequences. To highlight the quality performance of the proposed scheme the MOS scores of the reconstructed videos were obtained. The MOS scores, listed in Table 6, show that the reconstruction quality of proposed scheme is similar to that of the standard H.264 encoder with QP=21. QP is a critical parameter in H.264 to map a signal to a reduced range of values to represent it with fewer bits than the original. Higher the value of QP, higher is the compression at the cost of loss in quality.

Coding for H.264 has been performed using the JM encoder [[9]].

5. CONCLUSION

In this paper a novel parametric video coding scheme is proposed. The scheme is based on exploiting spatio-temporal redundancy of the video sequence. The characterization of video scene in terms of image complexity and motion complexity is performed. This characterization enables easy integration with H.264. Octree decomposition is applied to low complexity shots to segment them into homogeneous blocks. These homogeneous blocks are then modeled using RBF approximation.

Table 2 Comparison results for sequence Bridge far

H.264		Proposed scheme		
QP value	Compression	% of region encoded using RBF	Compression (RBF)	Overall Compression
24	39.7	25	2112.7	40.21
25	47.29	25	2112.7	47.17
26	75.57	25	2112.7	74.68
27	108.39	25	2112.7	108.52

Table 3 Comparison results for sequence Highway

H.264		Proposed scheme		
QP value	Compression	% of region encoded using RBF	Compression (RBF)	Overall Compression
24	36.89	25.6	918	38.71
25	43.04	25.6	918	44.85
26	66.59	25.6	918	67.31
27	93.40	25.6	918	92.45

Table 4 Comparison results for Hybrid sequence

H.264		Proposed scheme		
QP value	Compression	% of region encoded using RBF	Compression (RBF)	Overall Compression
24	328.55	46.7	1091	331.78
25	411.00	46.7	1091	395.39
26	535.48	46.7	1091	487.01
27	632.84	46.7	1091	546.78

Table 5 Summary of compression gains

Stream	Compression Gain		RBF PSNR
	H.264 (QP 24)	RBF encoding	
Bridge far	39.7	2112.7	41.68
Highway	36.89	918	50.93
Hybrid Sequence	328.55	1091	43.24

Table 6 MOS score comparison for the different test sequences

Stream	Original	H.264 (QP=21)	Our reconstruction
Hybrid test	5	4.7	4.6
Highway	5	4.6	4.6
Bridge far	5	4.5	4.2

As can be seen from the simulation results the compression gain achieved in the regions coded using RBF is substantially higher than the overall compression gains achieved using the H.264 video codec. The compression gain results presented in this paper are the gains obtained without applying entropy coding to the model parameters (Hybrid coding flag, Octree indices and RBF and 3D-DCT parameter values). Clearly the proposed scheme will provide even greater compression gains when schemes to reduce statistical redundancy are added to the proposed compression technique.

6. ACKNOWLEDGEMENTS

This work was as part of a joint collaborative research initiative between IIT Delhi and STMicroelectronics. We would like to acknowledge and appreciate the work done by Kartik Wason, as a student at IITD and Preet Kamal Singh and Tanmay Neema two interns at STMicroelectronics for the completion of this work.

7. REFERENCES

[1] Z. Zhu, Hao Tang, Wolberg, G., Layne, J.R.,Content-Based 3D mosaic representation for video of dynamic 3D scenes, IEEE, October 2005.

[2] Zhu Chunbo, Sun Xiaoyan, Wu Feng; Li Houqiang, “Video coding with spatio-temporal texture synthesis”, Multimedia and Expo, 2007 IEEE international conference.

[3] V. Cheung, B. J. Frey, and N. Jojic, “Video Epitomes”, In Proc. IEEE Conf. CVPR, 2005.

[4] M. P. Kumar, P.H.S. Torr, A. Zisserman, “Learning Layered Motion Segmentation of Video”, ICCV, IEEE, Oct 2005.

[5] S. Shin, S. Guo-Rui, K. Park, “A Scene Change Detection Scheme Using Local Chi-square Test on Telematics”, ICHIT, Nov 2006.

[6] Zabih, R., Miller, J., Mai, K., A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. Proc. ACM Multimedia 95, San Francisco, CA, pp. 189-200, Nov. 1995.

[7] H.S. Kim, J.Y. Lee,” Image coding by fitting RBF-surfaces to sub-images”, Pattern Recognition Letters (23), No. 11, September 2002, pp. 1239-1251.

[8] T Poggio and F Girosi, “Networks for approximation and learning”, Proc. IEEE, 78, 1481-1497, 1990.

[9] Advanced video coding for generic audiovisual services, ISO/IEC 14496-10:2005 (E) and ITU-T Rec. H.264 (E).