A Fully SNR, Spatial and Temporal Scalable 3DSPIHT-Based Video Coding Algorithm for Video Streaming Over Heterogeneous Networks

Habibollah Danyali[1] and Alfred Mertins[2]
[1]School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong, NSW 2522, Australia
habib@titr.uow.edu.au
[2]School of Mathematics and Natural Sciences, Institute of Physics
University of Oldenburg, 26111 Oldenburg, Germany
alfred.mertins@uni-oldenburg.de

Abstract—Video streaming over heterogenous networks (e.g. the Internet) requires high degree of scalability from the video coding. To achieve all types of scalability required for a fully scalable video coding, a modification of the 3DSPIHT algorithm is presented in this paper. The proposed algorithm, called Fully Scalable 3DSPIHT (FS-3DSPIHT), adds spatial and temporal scalability features to the 3DSPIHT algorithm, through the introduction of resolution dependent lists and a resolution-dependent sorting pass. The important features of the original 3DSPIHT coder such as compression efficiency and full embeddedness are kept. The output bitstream of the FS-3DSPIHT encoder consists of a set of embedded parts related to the various quality, spatial and temporal resolutions. It can be easily reordered and truncated by a simple parser in order to adapt various clients' requirements (e.g. quality, spatial resolution and temporal resolution) as well as bandwidth variation in a heterogeneous network.

1. INTRODUCTION

For delivering video information over a heterogenous network such as the Internet, a scalable video coding system is needed to provide a bitstream that supports various levels of quality and spatiotemporal resolution. A scalable bitstream consists of a set of embedded parts that offer increasingly better signal-to-noise ratio (SNR) or/and greater spatial resolution or/and higher frame rate. These features are referred to as SNR scalability, spatial scalability and temporal scalability respectively. Different parts of a scalable bitstream can be selected and decoded by a scalable decoder to meet certain requirements. Moreover, different types of decoders with different complexity and access bandwidth can coexist.

Traditional video coding standards such as MPEG-2 and H.263 provide some sort of scalability through a layer-based approach comprised of a base layer and one or more enhancement layers. However, this approach provides only coarse and very limited means of scalability and results in a serious quality drop in comparison to the non-scalable single layer mode at the same rate [1]. A fine granularity scalability (FGS) option adapted in MPEG-4 only includes SNR and temporal scalability and only for the enhancement layer [2].

On the other hand, 3D wavelet video coding schemes, e.g. [3–7], have a great potential to support all types of scalability. This is due to the multiresolution signal representation offered by the 3D wavelet transform. However, the transform itself is not the only issue. It is particularly important how the various spatiotemporal subbands in a 3D wavelet decomposition of a group of frame (GOF) are encoded. The excellent rate-distortion performance and scalable nature of the Set Partitioning in Hierarchical Trees (SPIHT) algorithm [8] for still images make it an attractive coding strategy also for video coding. A 3D extension of SPIHT for video coding has been proposed by Kim et al. in [4, 6]. They applied a 3D wavelet transform to a group of video frames (GOF) and coded the wavelet coefficients by the 3DSPIHT algorithm. As reported in [4], even without motion estimation and compensation this method performs (in test with SIF (352×240) monochrome 30 HZ sequences) measurably and visually better than MPEG-2, which employs complicated means of motion estimation and compensation. Although the 3DSPIHT bitstream is tailored for full SNR scalability and provides a progressive (by quality) bitstream, it does not support spatial and temporal scalabilities and does not provide a parsable bitstream that could be reordered according to desired spatiotemporal resolutions and fidelities.

In this paper, a fully scalable video coding scheme based on the 3DSPIHT algorithm is presented. The proposed algorithm is an extension of our previous works [9,10] on scalable SPIHT-based still image coding and supports all spatial, temporal and SNR scalability features together. The rest of this paper is organized as follow. Section 2 gives an overview of the video codec. Section 3 describes our modified algorithm, called FS-3DSPIHT. The bitstream structure and parsing process is explained in Section 4. In Section 5, some experimental results are presented, and finally, Section 6 concludes the paper.

2. OVERVIEW OF THE VIDEO CODEC

Figure 1 shows the block diagram of the video coding system. On the encoder side, the input video sequence is first divided into separate groups of frames (GOF). A 1D temporal filtering is first applied to each GOF, followed by a 2D spatial

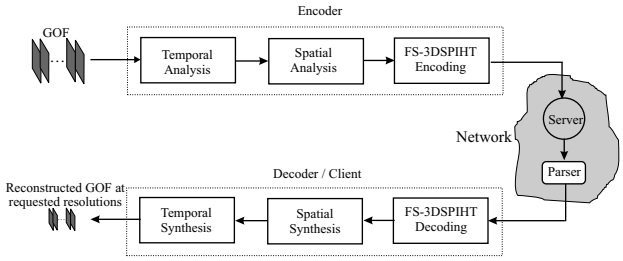

Figure 1. Block diagram of the video coding system.

wavelet transform. The number of temporal decomposition levels can be different from the number of spatial decomposition levels. Figure 2 shows a spatiotemporal multiresolution structure of a decomposed GOF. The FS-3DSPIHT encoder then progressively encodes the decomposed GOF from the lowest (coarsest) spatiotemporal band to the highest band in a way that the bits that belong to the individual spatiotemporal bands are distinguishable in the output bitstream. Details will be given in the next section.

The decoder receives a reordered bitstream for each GOF which is tailored by a parser (transcoder) for the requested spatial resolution, frame rate and bit rate. The FS-3DSPIHT decoder uses this bitstream to decode only the required spatiotemporal bands. The inverse spatial and temporal wavelet decomposition is then applied to the decoded spatiotemporal subbands to create a reconstructed version of the video sequence in the requested resolution.

3. The Fully Scalable 3DSPIHT (FS-3DSPIHT) Algorithm

3DSPIHT

The 3DSPIHT algorithm of [6] considers sets of coefficients that are related through a parent-offspring dependency like the one depicted in Figure 2. In its bitplane coding process, the algorithm deals with the wavelet coefficients as either a root of an insignificant set, an individual insignificant pixel, or a significant pixel. It sorts these coefficients in three ordered lists: the list of insignificant sets (LIS), the list of insignificant pixels (LIP), and the list of significant pixels (LSP). The main concept of the algorithm is managing these lists in order to efficiently extract insignificant sets in a hierarchical structure and identify significant coefficients, which is the core of its high compression performance.

Spatiotemporal Resolutions

The 3D (1D temporal + 2D spatial) wavelet decomposition of a GOF, as illustrated in Figure 2, provides a multiresolution structure that consists of different spatiotemporal subbands. In general, by applying N_t levels of 1D temporal decomposition and N_s levels of 2D spatial decomposition, $N_t + 1$ levels of temporal resolution and $N_s + 1$ levels of spatial resolution are achievable. The total number of spatiotemporal resolutions in this case is $(N_s+1)\times(N_t+1)$. To distinguish between

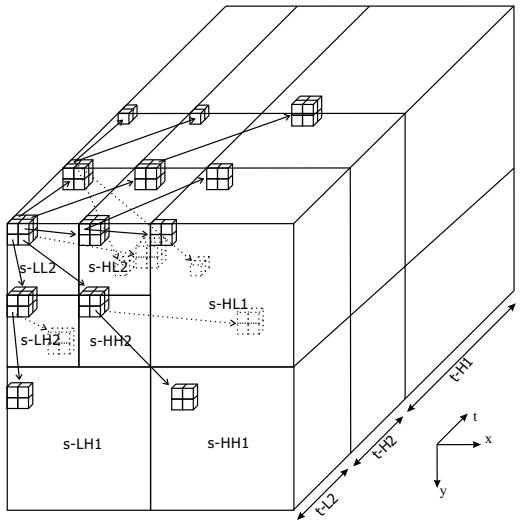

Figure 2. A 3D wavelet decomposition of a GOF with 2 levels of both temporal and spatial decomposition (21 spatiotemporal bands). The parent-offspring relationship between the wavelet coefficients assumed by 3DSPIHT is also shown.

different resolution levels, the spatiotemporal resolution level (k, l) is used to denote the spatial resolution level k and temporal resolution level l. The actual spatial and temporal resolutions related to the level (k, l) are respectively $1/2^{k-1}$ and $1/2^{l-1}$ of the spatial and temporal resolutions of the original sequence. Thus the lowest spatiotemporal resolution (the lowest spatiotemporal frequency subband in the decomposed GOF) is referred to as level $(N_s + 1, N_t + 1)$. The full spatiotemporal resolution (the original sequence) then becomes resolution level $(1, 1)$.

Spatiotemporal Subband Sets

A set of three spatial subbands, $\{\text{s-HL}_k, \text{s-LH}_k, \text{s-HH}_k\}$, in temporal subband level l that improve spatial resolution of that temporal subband from level $k + 1$ to level k is defined as *spatiotemporal subband set* level (k, l) and is referred to as $B_{k,l}$. Note that for the maximum spatial resolution level (i.e. $k = N_s + 1$), $B_{k,l}$ only contains one subband which is the lowest spatial frequency subband (LL_{N_s}) for all temporal subband levels. To improve spatiotemporal resolution from level (k_1, l_1) to level (k_2, l_2), all spatiotemporal subband sets that are located between these two resolutions need to be added. These are $B_{k,l}$ for $k_2 \le k < k_1$ and $l_2 \le l < l_1$.

An algorithm that provides full spatial and temporal scalability would encode the different resolution subbands separately, allowing a parser or a decoder to directly access the data needed to reconstruct a desired spatial or/and temporal resolution. The original 3DSPIHT algorithm of [6], however, sorts the wavelet coefficients in such a way that the output bitstream contains mixed information of all subbands in no particular order, making it impossible to transcode the bitstream without decoding it.

Coding Description

To add full spatial and temporal scalability features to the 3DSPIHT, the spatiotemporal subband sets in the decomposed GOF need to be coded separately. The fully scalable 3DSPIHT (FS-3DSPIHT) proposed here, fulfills this requirement through the introduction of multiple resolution-dependent lists and a resolution-dependent sorting pass. It defines a set of LIP, LSP and LIS lists for each $B_{k,l}$, therefore there will be $\text{LIP}_{k,l}$, $\text{LSP}_{k,l}$, and $\text{LIS}_{k,l}$ for $k = k_{max}, k_{max} - 1, \ldots, 1$ and $l = l_{max}, l_{max} - 1, \ldots, 1$ where k_{max} and l_{max} are the maximum number of spatial and temporal resolution levels respectively, supported by the encoder.

The FS-3DSPIHT encoder transmits bitplane by bitplane and defines the same parent-offspring dependency between the coefficients in the 3-D wavelet pyramid as 3DSPIHT (see Figure 2). In each bitplane, the coder starts encoding from $B_{k_{max},l_{max}}$ and proceeds to $B_{1,1,}$. In the resolution-dependent sorting pass of the lists that belong to the $B_{k,l}$, the algorithm first does the sorting pass for the coefficients in the $\text{LIP}_{k,l}$ to find and output significance bits for all list entries and then processes the $\text{LIS}_{k,l}$. During processing the $\text{LIS}_{k,l}$, sets that lie outside the spatiotemporal resolution level (k, l) are moved to their appropriate LIS related to the next higher level of spatiotemporal subband sets (i.e. $\text{LIS}_{k-1,l}$ or $\text{LIS}_{k,l-1}$ or $\text{LIS}_{k-1,l-1}$). After the algorithm finishes the sorting and refinement passes for resolution level (k,l) it will do the same for all other remaining lists related to other spatiotemporal subband set levels. The resolution-dependent lists can be scanned in such away that first all temporal resolution be completed, then spatial resolutions, or vice versa. According to the magnitude of the coefficients in the decomposed GOF, coding of the spatiotemporal subband sets related to the the higher spatiotemporal resolution usually starts from lower bitplanes. The FS-3DSPIHT manages its multiple lists during its resolution-dependent sorting pass in an efficient way such that the total number of bits spent in a particular bitplane is the same as for 3DSPIHT, but FS-3DSPIHT arrange them according to their spatiotemporal resolution dependency in the bitstream.

Note that the total storage requirement for the $\text{LIP}_{k,l}$, $\text{LSP}_{k,l}$, and $\text{LIS}_{k,l}$ for all resolutions is the same as for the LIS, LIP, and LSP used by the 3DSPIHT algorithm.

4. Bitstream Structure and Scaling

Bitstream Structure

Figure 3 shows the bitstream structure generated by the FS-3DSPIHT encoder. The bitstream is first divided into different parts according to the different bitplanes. The bitplane codepart P^n denotes all bits obtained at bitplane coding process level n. Inside each bitplane codepart, the bits that belong to different spatial subband sets are separable, and similarly, inside each spatial subbans set codepart, the bits that belong to different temporal subband sets come in order. The smallest codepart unit in the bitstream is referred to as $P^n_{k,l}$ in Figure 3. It is the codepart that belongs to spatiotemporal subband set level (k, l) (i.e. $B_{k,l}$) at bitplane level n. To support bitstream parsing, some markers are required to be put into the bitstream to identify the parts of the bitstream that belong to the different spatial and temporal resolution levels and bitplanes.

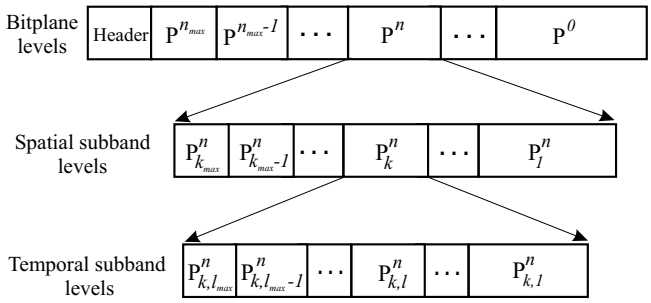

Figure 3. Structure of the FS-3DSPIHT encoder bitstream which is made up of different quality, spatial and temporal resolution parts. $P^n_{k,l}$ denotes the codepart that obtain for coding the spatiotemporal subband set level (k, l) (i.e. $B_{k,l}$) at bitplane level n.

Bitstream Scaling

The flexible hierarchical structure of the FS-3DSPIHT bitstream allows bitstream parsing to obtain sub-bitstreams for different reduced spatiotemporal resolutions and qualities, all from a single high bit rate, full resolution bitstream. The parsing (reordering) process is a simple scale reducing task in which only the related codeparts of the main bitstream that belong to the requested resolution are selected and ordered without need to decode. For example, to provide a bitstream for resolution level (k_0, l_0), in each biplane codepart, only the codeparts of the spatial subband sets that are located inside the spatial resolution level k_0 are kept, and similarly, in each selected spatial part, only the temporal parts that fall inside the requested temporal resolution level (l_0) are kept, and all other parts are removed. Therefore, the selected codeparts in bitpalne level n are $\{P^n_{k,l} | k_0 \leq k \leq k_{max}, l_0 \leq l \leq l_{max}\}$. Note that all marker information for identifying the individual bitplanes and resolution levels are only used by the parser and do not need to be sent to the decoder. A distinct feature is that, after parsing, the reordered bitstreams for all spatiotemporal resolutions are completely fine granular and can be truncated at any point to obtain the best reconstructed sequence at the desired spatiotemporal resolution and bit rate.

The decoder required for decoding the reordered bitstream exactly follows the encoder, similar to the original 3DSPIHT algorithm. It needs to keep track of the various lists only for spatial and temporal resolution levels greater or equal to the required one. Thus, the proposed algorithm naturally provides computational scalability as well.

Figure 4. Foreman and Hall Monitor sequences at different spatial resolutions.

Table 1. Mean Y-PSNR results for 240 frames of 30HZ CIF Foreman and Hall Monitor sequences at different spatial and temporal resolutions and bit rates obtained by a FS-3DSPIHT decoder.

Spatial resolution	Temporal resolution	Rate (kbps)	Mean Y-PSNR (dB)	
			Forman	Hall Monitor
CIF	30	256	29.37	33.82
CIF	15	128	29.55	30.23
CIF	7.5	80	31.26	28.08
QCIF	15	128	31.16	35.55
QCIF	15	80	29.40	31.73
QCIF	7.5	64	31.78	29.89
QCIF	7.5	48	30.20	27.98
QCIF	3.75	48	35.04	29.32
$\frac{1}{4}$ QCIF	7.5	32	30.30	31.10
$\frac{1}{4}$ QCIF	7.5	28	29.33	29.18
$\frac{1}{4}$ QCIF	3.75	28	34.48	31.88

5. Experimental Results

To show the FS-3DSPIHT encoder's full scalability features, tests were performed on two CIF sequences: Foreman and Hall Monitor. The original frame rate of these sequences is 30 frames per second. The GOF size used by the encoder is 16 and for the 3D wavelet decomposition of each GOF, 3 levels of 1D decomposition by Haar filter in temporal direction followed by 4 levels of 2D spatial decomposition by 9/7-tap filters [11] were applied. The FS-3DSPIHT encoder was set to encode 240 frames (15 GOFs) of each sequence at 600kbps and to provide 5 levels of spatial resolution and 4 levels of temporal resolution (total 20 spatiotemporal resolutions), which are the maximum possible resolution levels accessible by the above mentioned 3D decomposition.

After encoding, the FS-3DSPIHT bitstream was fed into a parser to produce progressive (by quality) bitstreams for different spatial and temporal resolutions. The bitstreams were decoded at various spatial resolutions (see Figure 4), frame rates and bit rates. Table 1 shows the mean PSNR results obtained for the luminance components of the decoded sequences. Reference frames for lower resolutions were defined by taking the lowest frequency subband frames after applying appropriate levels of temporal and spatial wavelet decomposition to the original sequences. For full spatial and temporal resolution these results are comparable with the 3DSPIHT coder of [4, 6], but our coder additionally yields the full scalability features. The PSNR values in Table 1 show that high-SNR bitstreams can be easily generated at low rates with reduced spatio temporal resolution.

6. Conclusions

In this paper we presented a fully scalable 3DSPIHT algorithm (FS-3DSPIHT) for video coding that supports all types of spatial, temporal and SNR scalability together. The interesting features of the original 3DSPIHT algorithm such as high compression efficiency and rate embeddedness of the bitstream are kept. The FS-3DSPIHT encoder bitstream can be easily reordered (parsed) to obtain rate embedded sub-bitstreams for various spatiotemporal resolutions without the need of decoding. The proposed fully scalable video codec is a good candidate for multimedia applications such as video information storage and retrieval systems, and video streaming over heterogenous networks where a wide variety of users needs to be differently serviced according to their network access and data processing capabilities.

7. Acknowledgment

The first author would like to acknowledge the financial support provided for him by the Ministry of Science, Research and Technology (MSRT), Iran and Kurdistan University, Sanandaj, Iran during doing this research as a part of his PhD study at the University of Wollongong, Australia.

References

[1] L. Yang, F.C.M. Martins, and T.R. Gardos, "Improving H.263+ scalability performance for very low bit rate applications," in *Proc. SPIE-VCIP'99*, San Jose, CA, Jan. 1999, vol. 3653, pp. 768–779.

[2] W. Li, F. Ling, and X. Chen, "Fine granularity scalability in MPEG-4 for streaming video," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2000, vol. 1, pp. 299–302.

[3] C. Podichuk, N. Jayant, and N. Farvardin, "Three-dimensional subband coding of video," *IEEE Trans. Image Processing*, vol. 4, no. 2, pp. 125–139, Feb. 1995.

[4] B.-J.Kim and W. A. Pearlman, "An embedded video coder using three-dimensional set partitioning in hierarchical trees (SPIHT)," in *proc. IEEE Data Compression Conf.*, Mar. 1997, pp. 251–260.

[5] S.-J. Choi and J. W. Woods, "Motion compensated 3-d subband coding of video," *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[6] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-d set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. Circ. and Syst. for Video Technology*, vol. 10, no. 8, pp. 1374–1387, Dec. 2000.

[7] S.-T. Hsiang and J. W. Woods, "Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank," *Signal Processing: Image Communication*, vol. 16, no. 8, pp. 705–724, May 2001.

[8] A. Said and W. A. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circ. and Syst. for Video Technology*, vol. 6, pp. 243–250, June 1996.

[9] H. Danyali and A. Mertins, "Highly scalable image compression based on SPIHT for network applications," in *Proc. IEEE Int. Conf. Image Processing*, Rochester, NY, USA, Sept. 2002, vol. 1, pp. 217–220.

[10] H. Danyali and A. Mertins, "Fully scalable wavelet-based image coding for transmission over heterogeneous networks," in *Proc. 1st Workshop on the Internet, Telecommunications and Signal Processing, WITSP'02*, Wollongong, NSW, Australia, Dec. 2002, pp. 173–178.

[11] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, Apr. 1992.