Predicting Optimal Operation of MC-3DSBC Multi-Dimensional Scalable Video Coding Using Subjective Quality Measurement

Yong Wang[*], Tian-Tsong Ng[*], Mihaela van der Schaar[#], Shih-Fu Chang[*]
*Dept. of Elec. Engr., Columbia Univ., 1312 Mudd, 500 W120 St. New York, NY, 10027
[#]Dept. of Elec. & Comp. Engr., UC Davis, One Shield Ave. 3129 Kemper Hall, Davis, CA 95616

ABSTRACT

Recently we have witnessed a growing interest in the development of the subband/wavelet coding (SBC) technology, partly due to the superior scalability of SBC. Scalable coding provides great synergy with the universal media access applications, where media content is delivered to client devices of diverse types through heterogeneous channels. In this respect, SBC system provides flexibility in realizing different ways of media scaling, including scaling dimensions of SNR, spatial, and temporal. However, the selection of specific scalability operations given the bit rate constraint has always been ad hoc – a systematic methodology is missing. In this paper, we address this open issue by applying our content-based optimal scalability selection framework and adopting subjective quality evaluation. For this purpose we firstly explore the behavior of SNR-Spatial-Temporal scalability using Motion Compensated (MC) SBC systems. Based on the system behavior, we propose an efficient method for the optimal selection of scalability operator through content-based prediction. Our experiment results demonstrate that the proposed method can efficiently predict the optimal scalability operation with an excellent accuracy.

Keywords: MC-3DSBC, SNR-Spatial-Temporal Scalability, Content Based Adaptation, Perceptual Quality

1. INTRODUCTION

Recently we have witnessed the rapid development of the subband/wavelet image/video coding technology. The success of subband coding has been shown in static image coding such as JPEG2000[21]. The extension of subband coding to motion pictures relies on the efficient processing of motion compensation along the temporal dimension. Approaches based on MC-3D subband coding have been proposed to address this issue. One early important progress in 3D subband coding of video is the work by Ohm[1], where a non-recursive MC-3DSBC scheme is proposed. Choi and Woods[2] extended Ohm's work by proposing a 3DSB-FSSQ coding system for video. Both techniques employ unidirectional motion compensation. Chen and Woods[3] further advanced the work by using bidirectional motion compensation temporal filtering (MCTF). The promising performance of the subband coding system also comes from the usage of embedded coding architecture. Since the nature of subband filtering provides a layered hierarchical representation, it can efficiently organize the transform coefficients based on their energy distribution and correlations in different layers. The embedded coding explores two basic properties of subband coding coefficients: First, coefficients in the same spatial position from different subbands share similar energy behaviors; second, the coefficient magnitude of a higher level subband is statistically smaller than the magnitude of the ones at the same spatial location from a lower level subband. From EZW[18] to 3D-EZBC[9], embedded coding systems have made significant progress. Interested readers are directed to references[3, 5, 6, 7, 8, 9, 18] for the details.

Scalable coding mentioned above like MC-3DSBC presents great synergy with emerging applications in universal media access, where the media content is delivered through various channels to diverse client devices. Due to such heterogeneity, various constraints in resource and preference are imposed. To meet such diverse constraints, MC-3DSBC can be used to facilitate multiple degrees of freedom in adapting a video stream. Based on the spatio-temporal decomposition and layered embedded coding structure, MC-3DSBC provides efficient methods allowing dropping some of the bit planes of wavelet coefficients (thus lowering the SNR), dropping some of the spatial subbands (thus reducing the spatial size), or skipping some of the temporal subbands (thus reducing the frame rate). Such flexibility in multi-dimensional scalability is very useful for meeting the diverse conditions in UMA applications[22]. In contrast, conventional coding techniques such as MPEG-2, -4 or H.26x do not provide equivalent flexibility in so many dimensions at the same time.

One critical issue arises when we need to compare and select various options provided by multi-dimensional scalable coding like MC-3DSBC to meet specific resource constraints. For example, given a target bit rate, how do we select and combine scalable operations from SNR, spatial, and temporal dimensions to meet the rate constraint and achieve the highest video quality? There is no systematic solution that exists today.

In this paper, we address this open issue by applying our content-based optimal scalability selection framework[13] and conducting subjective evaluation to compare perceptual quality of different spatio-temporal scalability combinations. We first explore the behavior of SNR-Spatial-Temporal scalability in MC-3DSBC systems. Specifically, for the SNR scalability, we investigate the optimal rate allocation among multiple spatio-temporal subbands by comparing uniform bit plane truncation, Gaussian shaping truncation, and Lagrange-based optimal truncation. Afterwards, we propose an efficient method for optimal scalability operator selection using content-based prediction. The optimality is measured based on the perceptual quality measurements obtained through subjective evaluation of adapted video streams. The basic idea is to predict the scaling operation achieving the optimal perceptual video quality based on the content features that are computable from the encoded streams. This idea originates from our previous work on utility function based MPEG-4 optimal transcoding[13], where the content features are used to predict the MPEG-4 transcoding operation achieving the optimal quality. Different from our prior work, the video quality measurement is based on subjective evaluation, which is essential for comparing video streams encoded with different spatio-temporal resolutions. Moreover, we use principle component analysis (PCA) and mutual information (MI) to systematically select the dominant content features.

Our experiment results demonstrate three points. First, for rate control of SNR scalability, the uniform truncation scheme, though extremely simple, can achieve close to optimal performance. Second, for optimal multi-dimensional operation selection, our proposed content-based framework can predict the optimal scalability operation efficiently with very high accuracy. Last, for automatic feature selection, our PCA-MI combined method is able to find salient subsets of features that are required for developing accurate predictor mentioned above.

The rest of this paper is organized as follows. In Section 2, the MC-3DSBC coding systems is briefly reviewed; Section 3 illustrates the scalability options supported by MC-3DSBC; Section 4 presents the content-based prediction framework of optimal scaling operation; the experiment result is presented in both Section 3 and section 4; Section 5 concludes the paper and discusses future work.

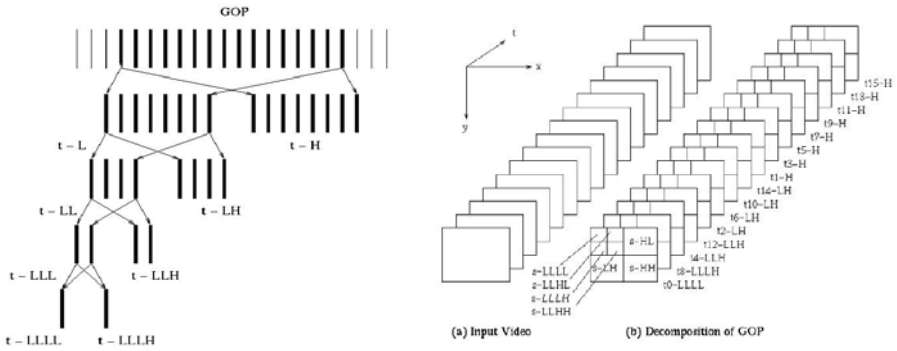
2. MC-3DSBC CODING SYSTEMS

MC-3DSBC coding systems are based on the 3D spatio-temporal decomposition of motion compensated video signals. The source video frames firstly undergo temporal decomposition. Figure 1(a) illustrates this procedure[9] where the GOP size is 16 and 4-level temporal decomposition is applied. The temporal filtering is applied along the motion trajectory. A pair of temporal low- and high-subband layers is generated for each two successive input frames, denoted as *temporal layers*. In the meantime one set of motion vectors is associated with these two frames and coded as the overhead. The temporal decomposition continues in an octave way until reaching a specified level, which is decided by the GOP size. For each temporal layer located in different hierarchical levels, spatial octave decomposition is utilized afterwards and the generated subbands are denoted as *spatial subbands*, which is similar with conventional wavelet decomposition employed in image coding systems, as indicated in Figure 1(b). Due to temporal filtering, each spatial subband contains interleaved spatio-temporal signals.

After the decomposition, the transform coefficients will be further coded by the bitplane-based schema. There are several ways to construct the bitplane. Some typical examples in the literature include EZW[4], SPIHT[5] and SPECK[7]. These methods take advantage of the nature of energy clustering of subband coefficients in space. A hierarchical set partitioning process is applied to split off significant coefficients (with respect to the threshold in the current bitplane coding pass), while maintaining areas of insignificant coefficients in the zero symbols. In this way, a large region of insignificant pixels can be coded into one symbol, thus providing efficient coding.

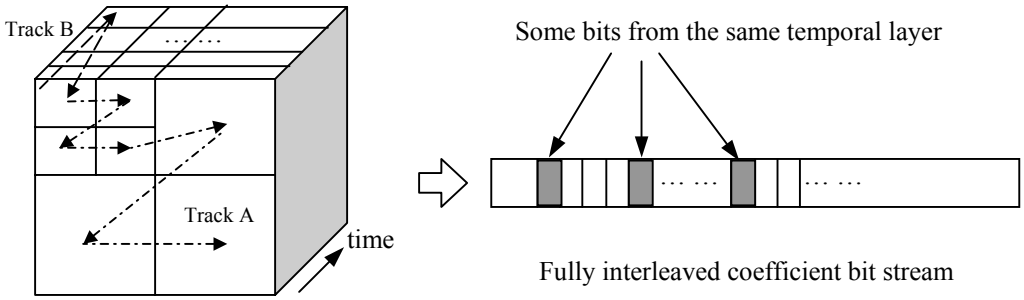
Each spatial subband undergoes the bitplane coding and the output bits are combined together to generate an embedded bit stream. According to the requirement of scalability, the organization of coefficients from different spatial subbands in the bit stream might vary, depending on the scanning order of the subbands. One possible implementation is that the

coefficients from both spatial subbands and temporal layers are fully interleaved together. In each temporal layer, the scanning order is as track A in Figure 2 (a). Before the process moves to next spatial subband, all of the coefficients in previous spatial subband from all temporal layers will be scanned first, shown as the track B in Figure 2(a). This will generate a fully interleaved bit stream of subband coefficients. The advantage of this method is its simple implementation, and spatial scalability is as easy as truncating and throwing away the tail of the bit streams based on the bit rate limitation. The drawback of this method is also obvious: Since the bits belonging to the same temporal layer are scattered at different locations in the stream, it is not convenient to realize the temporal scalability. Considering the shaded bit blocks in Figure 2 (b), removing any of them need realign the whole bit stream, which is very tedious.



(1) (2)

Figure 1: Spatial-temporal decomposition in MC-3DSBC
(1) Octave based five-band temporal decomposition; (2) The 3-D subband structure in a GOP



(a) (b)

Figure 2: Fully interleaved bit stream generation

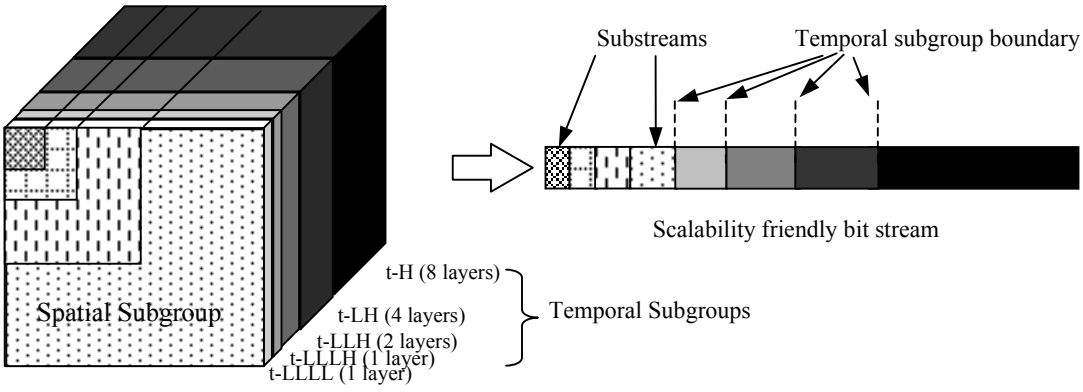


Figure 3: Scalability friendly bit stream generation

An alternative way of coefficient organization is shown in Figure 3. Firstly each temporal layer is clustered into several subgroups according to their temporal decomposition level. For each temporal subgroup, spatial subbands are further clustered in a similar way according to their spatial decomposition level. The coefficients located in a same spatial and temporal subgroup are coded into a fully interleaved substream. All of the substreams are aligned as shown in Figure 3. Note the temporal subgroup boundaries are determined after the spatial scalability operation is chosen, thus avoiding the scattering problem in Figure 2. This implementation is complex compared to the previous one. However, the advantage gained is its considerable freedom in terms of temporal scalability coding.

Before MC-3DSBC, there existed some scalable coding techniques based on traditional DCT-quantization based video coding techniques adopted by MPEG and H.26x, such as requantization, coefficient dropping and frame dropping. Among them, the MPEG-4 FGS[24] provides the capability to distribute bit rates over layered streams with high scalability flexibility by using bitplane truncation, which is similar to the method employed in MC-3DSBC. Compared with these methods, MC-3DSBC coding systems have great advantages in terms of scalability performance. Prior experiments[3] indicated that MC-3DSBC appeared to achieve excellent performance close to the optimal rate-distortion performance of H.26L[3], which is impossible for DCT-based reshaping methods[23, 13]. Moreover, MC-3DSBC provides more convenient ways to realize scalability in multiple dimensions.

3. SNR-SPATIO-TEMPORAL SCALABILITY

MC-3DSBC provides multi-dimensional freedom to realize several different ways of scalability. In this section we provide detailed descriptions of these scalability options.

3.1 SNR scalability

For MC-3DSBC, SNR scalability is used to generate a degraded version of bit stream from the original one, while keeping the spatial and temporal resolution unchanged. This is realized by truncating the bitplane based on a target bit rate. The consequence for SNR scalability is the increased distortion measured in MSE or PSNR. By SNR scalability, for a given bit rate, several different bit steams can be generated with various distortions if different bitplane truncation methods are adopted. One goal for SNR scalability is to find an optimal bit stream that causes the least distortion; i.e., minimum PSNR loss.

There might be concern about the relationship between the SNR and perceptual quality. It is widely acknowledged that SNR may not be an adequate measurement of image perceptual quality[25]. However, if we consider the scenario without additional noise sources (such as white Gaussian noise) and artificial processing (such as shifting image by certain pixels), given a video codec, SNR can be acceptable to subjective evaluation. In a recent study[26] by the Video Quality Experts Group (VQEG), PSNR was shown to be a well approximation to DMOS, as well as several other objective models. Therefore, by achieving optimal SNR scalability, we can reasonably expect an ensured perceptual quality.

For better explanation, we illustrate the bitplane truncation with an example in Figure 4, where only one temporal subgroup is considered. Within the temporal subgroup, as introduced in Section 2, the coefficients are coded into different substreams. During the SNR-scalability, all substreams are aligned based on their most significant bitplane (MSB). The truncation begins from the least significant bitplane (LSB). Theoretically, each substream has its own bitplane truncation boundary denoted as the bitplane index, say K_{ij}, where i is the temporal subgroup index and j is the spatial subgroup index. The bit rate associated with K_{ij} is $R_{ij} = R(K_{ij})$. For a given vector $\vec{K} = (K_{00}, K_{01}, ..., K_{ij}, ...)$, we can get a target rate $R = R(\vec{K})$. Obviously, there might be a set of different vectors $V=\{\vec{K}\}$ reaching the same bit rate, SNR scalability operation is responsible for selecting the right vector $\vec{K}_{op}$ that yields minimum quality loss. Namely, $\vec{K}_{op} = \arg\min D(\vec{K}) \left| \vec{K} \in V, R(\vec{K}) = R \right.$, where $D(\vec{K})$ is the distortion associated with $\vec{K}$. In the remaining part of this subsection, we will explore several possible ways of truncation.

3.1.1 Uniform Bitplane Truncation (UBT)

As implied by its name, in UBT all of the substreams will have the same truncation boundary $K_{uniform}$, which is a function of target bit rate: $K_{uniform} = g(R)$. This is illustrated in Figure 5 (a). The term "uniform" is in the sense of the

number of bitplanes, instead of bit amount. The reason for this is non-trivial. In MC-3DSBC systems, the bitplane index is a natural indicator about the importance of the coefficients in terms of the perceptual quality. On the other hand, due to the spatio-temporal decomposition structure, each subband layer has a different size, as is indicated in Figure 4. Therefore, cutting the same number of bits from each layer may not produce an adequate perceptual quality. As shown later in experiments, the *UBT method*, though simple, can be used to achieve excellent video quality.

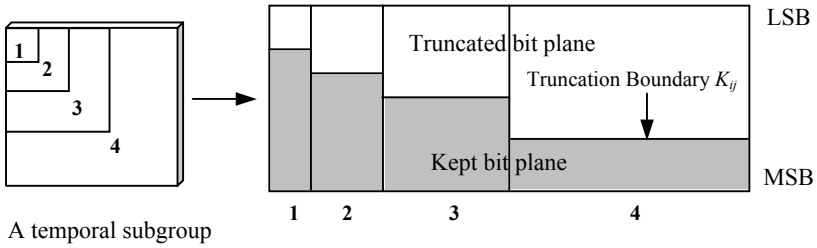
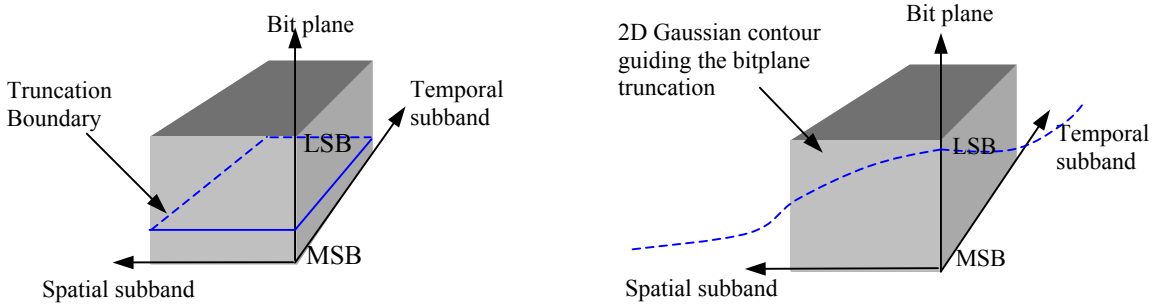


Figure 4: Bitplane truncation to realize the SNR-Scalability

3.1.2 Gaussian Shaped Truncation (GST)

The motivation of performing GST originates from the conjecture that the quality could be improved if a relatively bigger bit budget could be assigned to the low frequency subbands. This conjecture is based on the characteristic that human vision is less sensitive to the higher spatial frequency than the lower one[19]. GST is shown in Figure 5(b). Instead of a uniform value, we select the truncation boundary based on a 2-D Gaussian based contour. In fact, the curve we used is the amplitude adjusted Gaussian, which is given by the following equation:

$$f(x) = A\exp\{\frac{x^2}{2\sigma^2}\} \tag{1}$$



(a) Uniform Bitplane Trancation (b) Gaussian Shaped Truncation

Figure 5: Bitplane Truncation

Equation (1) outlines a shrinking Gaussian curve used to allocate bits to each subband on the temporal or spatial dimension. A Gaussian is represented in a hypothetical 2-D space (*x*-*f*(*x*)), where each subband is represented by a value on the x axis within the normalized range [0,1] and the value of the function $f(x)$ represents the ratio of total bit planes to be kept. The Gaussian curve is shrunk until the amount of bits under the curve is no greater than the bit budget. Given an overall bit budget, we first perform such bit allocation along the temporal dimension. Once the amount of bits allocated to each temporal subgroup is determined, we then perform the same bit allocation procedure along the spatial dimension for each of the temporal subgroup. In implementation, the shrinking Gaussian function is fully specified by a parameter known as the shaping angle θ, which is defined as below:

$$\theta = \tan^{-1}(\frac{\Delta_r A}{\Delta_r \sigma}), \quad 0^o < \theta < 90^o \tag{2}$$

where $\Delta_r A = \frac{A_o - A}{A_o}$ and $\Delta_r \sigma = \frac{\sigma_o - \sigma}{\sigma_o}$ with A_o and σ_o being the amplitude and standard deviation of the pre-defined initial Gaussian curve. As we set A_o and σ_o to be 4 and 1.4 respectively, it is possible that the value $f(x)$

exceeds 1, and therefore the ratio of total bit plane is given by min $\{f(x),1\}$. Figure 6 shows some examples of the shrinking Gaussian curves for different θ. As we can see from them, the shrinking curves of 80 degree are very close to horizontal lines. Thus, we can emulate UBT by the shrinking curves with 80 degrees (or higher).

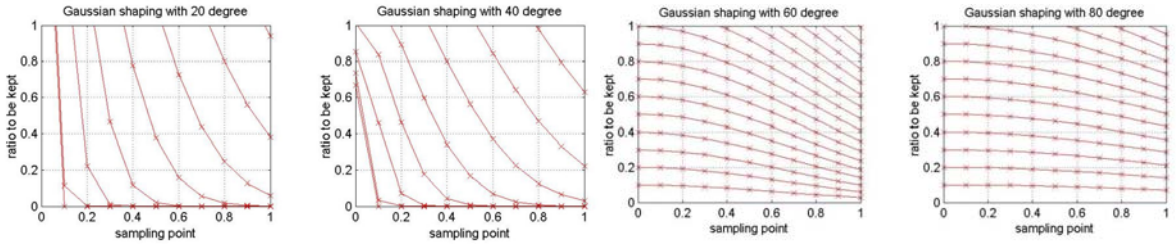


Figure 8: Shrinking Gaussian with different shaping angles

3.1.3 Lagrange Searching Truncation (*LST*)

UBT can be considered as a specific example of GST with a very flat Gaussian curve. GST searches for possible truncation boundaries in the subspace defined by a Gaussian function. If we consider an arbitrary truncation boundary $\vec{K} = (K_{00}, K_{01}, ..., K_{ij}, ...)$, the problem of optimal operation selection can be modeled as constrained optimization:

$$minimize \sum_{i,j} D(K_{ij}) \text{ , } such\ that \sum_{i,j} R(K_{ij}) \le R \tag{3}$$

This problem can be solved by using classic Lagrange Optimization method. The key issue is to properly find the mapping function $D(K_{ij})$ and $R(K_{ij})$. Conceptually, the distortion can be modeled as:

$$D = \sum_{i,j} \sum_{k \in S_{ij}} (P_k - \widetilde{P}_k)^2 \tag{4}$$

where S_{ij} is the substream defined by i, j, and $P_k, \widetilde{P}_k$ denote the original and the reconstructed coefficient respectively. Some existing embedded coding such as EBCOT[15] employs this model to select the optimal truncation point. One potential drawback is it needs the computation of square values for each coefficient and is not convenient for bitplane-based embedded coding system. Thus we try to find an approximate solution. For this purpose, we instantiate the analysis based on the Embedded Zeroblock Coding (EZBC)[11] coding implementation.

EZBC is one of the latest MC-3DSBC systems. It constructs the bit streams using the methods illustrated in Figure 3 and 4. Within each substream, the coefficients are coded by using bitplane-based scanning. During each scanning, compared with a decaying threshold n, the coefficients are assigned into three lists: list of insignificant pixels (LIP), list of insignificant sets (LIS) and list of significant pixels (LSP) according to their magnitudes; and then some corresponding bits are output. The purpose of the coefficient scanning is to locate the significant coefficients while processing areas of insignificant coefficients into zero symbol. It provides an efficient method to compactly represent a group of leading zeros of subband coefficients[9, 10]. For each given threshold n, the coefficients will undergo several rounds of scanning, each generating a sub bitplane with some corresponding output bits. The scanning in LIP and LIS might add some pixels into LSP, while the scanning in LSP refines the magnitude of the pixels in LSP. Figure 7 illustrates this procedure. Each sub bitplane will output some bits $r_{i,j,b,s}$ with corresponding distortion contribution $d_{i,j,b,s}$, where i, j denote the temporal subgroup index and spatial substream index respectively, and b, s denote bitplane index and sub bitplane index respectively. $r_{i,j,b,s}$ can be obtained from the bit stream. $d_{i,j,b,s}$ can be modeled as a function of the effective number of pixels in LSP $m_{i,j,b,s}$. Specifically,

$$d_{i,j,b,s} = w_s \cdot m_{i,j,b,s} \cdot 2^{2b} \tag{5}$$

For LIP and LIS sub bitplanes, $m_{i,j,b,s}$ is defined as the number of pixels added into LSP after the scanning. For LSP, $m_{i,j,b,s}$ is defined as the number of pixels, which are affected by the refinement step and generate nonzero bits. w_s is a weight scalar related to the sub bitplane and defined by Islam, *et al.*[7]

Given $r_{i,j,b,s}$ and $d_{i,j,b,s}$, the optimization problem is modeled as:

$$minimize \sum_{i,j} \sum_{0 \le b \le K_{ij}} \sum_{s} d_{i,j,b,s} \ , \ such\ that \quad \sum_{i,j} \sum_{0 \le b \le K_{ij}} \sum_{s} r_{i,j,b,s} \le R \tag{6}$$

By using the Lagrange multiplier method, we can find the optimal truncation boundary $\vec{K} = (K_{00}, K_{01}, ..., K_{ij}, ...)$ for the above equation.

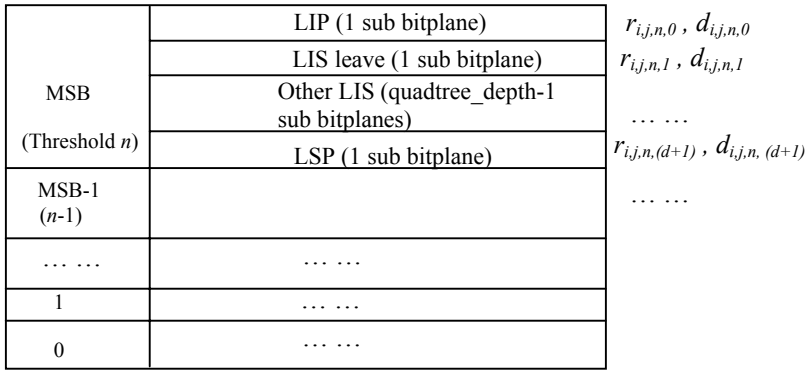


Figure 7: Bitplane scanning procedure in EZBC

3.1.4 Performance Analysis

In order to evaluate the performance of different rate controls methods for SNR scalability, we perform an experiment on 4 typical sequences of about 1 second long that (i.e., 2 GOP): *Foreman*, *Mobile*, *Stefan* and *Akiyo*. All of them are in the format of CIF resolution and 30 frames per second. We adopt the coding implementation in MC-EZBC[3], whose specification is summarized in Table 1.

Table 1: MC-EZBC system specification

GOP Size	Decomposition Level	Temporal Filter	Spatial Filter	Motion Compensation	Bitplane coding
16 frames	5 temporal levels 6 (5) spatial levels for CIF (QCIF)	2-tap Haar filter	Daubechies 9/7 filter	Bidirectional with Hierarchical Variable Size Block Matching (HVSBM)	Zeroblock coding with context modeling

For GST, we quantize the shaping angle θ in both spatial and temporal dimensions into four bins starting from 20 degrees with an incremental of 20 degrees. Table 2 shows the average PSNR for all the sequences shaped to a rate 200*kbps* except for Stefan sequence where 400*kbps* is used. From the table, we can observe that the rate shaping results are always the best or among the best when both the temporal and spatial shaping angles are 80 degrees. As mentioned earlier, a shrinking curve of 80 degrees is very close to uniform shaping. Actually, we observe that results from either method indeed are very close in terms of subjective quality and average PSNR. This is an interesting result, implying that uniform bitplane truncation will yield close to optimal quality. Another observation from Table 2 is that the change of the quality (average PSNR) is more drastic for the spatial dimension for different shaping angles than the temporal dimension. In the extreme case of *Mobile* sequence, the average PSNR is unchanged for different temporal shaping angles.

Figure 8 shows the comparison between UBT and LST. The curves show the ratio of bit allocation (i.e., the percentage of bits kept) for each substream. The rate operation is run for a target rate of 200kbps on *Foreman* sequence with a format of QCIF, 1 second and 30fps. The PSNR of the target bit streams are 34.42dB and 34.23dB for UBT and LST respectively. We can see that the ratios of bit allocation for both approaches are very similar. For the *Foreman* sequence, there are 5 temporal subgroups, each having 5 spatial substreams. There are 25 substreams in total. In Figure 8 the substream ID is obtained by: SubstremID = i*5 + j, where i, j are the index of the temporal subgroup and the spatial substream respectively. From the bit allocation ratio, we can justify our assumption about the bit allocation policy: In

order to get an optimal quality, (no matter spatial or temporal) low frequency components need be assigned more proportions of bits. Besides, the observation in Figure 8 enhances the conclusion drawn from the result of GTS: LST achieves a close to optimal performance not only in the subspace defined by Gaussian shaping, but also in the larger space searched by the Lagrange optimization. The slightly degradation of LST compared with UBT in terms of PSNR is because of two reasons: Firstly, the assumption of uniform quality contribution from different subbands might be invalid; secondly, the bit rate adjustment of LST can only be achieved at the boundary of the sub bitplane, which will miss some truncation operations at a finer level.. In contrast, UBT can access these finer points and realize a more precise reshaping[11].

Table 2: Average PSNR for GST results

Gaussian curve shaping		Temporal shaping angle							
		Akiyo (200*kbps*)				Stefan (400*kbps*)			
		20	40	60	80	20	40	60	80
Spatial shaping angle	20	33.81	33.81	33.52	33.46	23.74	23.74	23.56	23.25
	40	35.83	35.83	35.08	35.05	24.45	24.45	24.36	24.31
	60	40.46	40.46	41.62	41.59	26.23	26.23	26.30	26.28
	80	40.55	40.55	41.89	41.94	26.34	26.34	26.42	26.41
		Foreman (200*kbps*)				Mobile (200*kbps*)			
	20	27.06	27.06	27.08	27.08	18.40	18.40	18.40	18.40
	40	27.82	27.82	27.53	27.54	18.83	18.83	18.83	18.83
	60	30.62	30.62	30.71	30.71	20.97	20.97	20.97	20.97
	80	30.69	30.69	30.78	30.80	21.05	21.05	21.05	21.05

It is interesting to observe that UBT, though extremely simple, can achieve excellent results. This observation is reasonable when we consider two factors of MC-EZBC: the embedded zeroblock coding method can effectively compress the coefficients based on their magnitude correlations and thus reduce the distortion dramaticly; the context modeling can efficiently reduce the entropy of the bit stream and thus further improve the coding efficiency. In our remaining experiments, we will simply apply UBT.

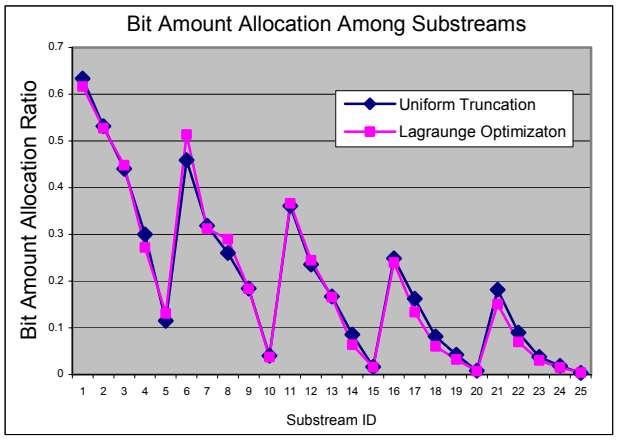


Figure 8: The ratio of bit allocation (the percentage of bits kept) for each substream

3.2 Spatial scalability

Spatial scalability provides us another freedom in shaping the bit stream. It reshapes the original bit stream by spatial resolution reduction. This is implemented by removing the whole substreams exceeding some specified spatial decomposition levels. Each spatial level dropping leads to a reduction of the image size by half in both width and height. Spatial scalability is useful when the bandwidth is very low and SNR scalability degrades the video quality excessively. By removing some high frequency spatial subbands, more bits can be assigned to the low frequency components, instead of spreading bit budget among all subbands.

Spatial scalability can be considered as a special case of SNR scalability where the truncation boundaries for the dropped spatial subbands are zeroes. In the meatime, the performance of spatial scalability may be poor due to impact of spatially

reduced reference frames on motion compensation – issues of changed references in motion prediction and error drifting. This problem degrades the applicability of spatial scalability. However, recent work represented by Andreopoulos, etc.[17] proposes so call In-Band MCTF (IBMCTF) that utilizes the motion compensation within each subband and reported some promising perceptual quality improvement for spatial scalability. This is beyond the discussion scope of this paper. On the other hand, our experiment result indicates that at the bandwidth range where the spatial scalability is superior to SNR scalability in terms of perceptual quality, the temporal scalability usually turns out to be the best option. Based on the facts above, we therefore did not include the spatial scalability in the our subsequent study of optimal operation selection. Note in some specific scenario such as delivering videos to the devices with limited display resolution, spatial scalability should be a viable option and be considered separately.

3.3 Temporal scalability

Similar to the spatial scalability, temporal scalability reshapes the bit stream by temporal resolution reduction; i.e., the temporal subgroups beyond some specified decomposition level are truncated. The consequence of temporal scalability is a reduced frame rate. Temporal scalability is a fairly course way to realize rate reshape, which saves the bit budget by dropping some temporal layers and enhances the quality of the remaining layers. For some types of video, it can also achieve acceptable perceptual video quality.

The incorporation of temporal scalability into the scalability framework makes the selection of optimal quality measurement non-trivial. How to evaluate the perceptual quality of two video clips with different frame rates is still not well addressed in the literature. In human vision modeling research, the mechanism how temporally accumulated distortion affects an instantaneously perceived distortion is still unknown[12]. Recognizing this difficult issue, we adopt a subjective quality evaluation method as the criterion to compare the performance of our content-based optimal scalability selection algorithm.

4. CONTENT BASED OPTIMAL SCALABILITY PREDICTION

So far we have discussed three types of scalability: SNR, spatial and temporal scalability. The combination of these three scalability operations yield quite different bit streams with various perceptual quality. If we ignore the spatial scalability as mentioned in Section 3.2, we can define arbitrary scalability operation as $O_s = (O_{SNR}, O_{Temp})$, where O_{SNR}, O_{Temp} represent SNR and temporal scalability respectively. For SNR scalability, itself has ability of generating various bit streams, we select the uniform bitplane truncation method as discussed in Section 3.

Usually there are a set of O_s satisfying a given target bit rate *R*, while giving different perceptual quality. Finding the optimal scalability operation is an interesting while challenging task. The general rule[20] is that when the SNR quality of the video is beyond certain tolerance threshold, the users will prefer higher frame rate for more temporal details. Otherwise a lower frame rate is desired so that more bits are used to imporve the SNR quality. However, this tolerance threshold is dependent on the type of content. In our experiment, for videos with different content characteristic, user preference varies considerably. This observation motivates our content based solution: we try to address this issue by using a statistical pattern recognition model. The basic idea is to automatically discover distinctive classes of video – the preferred tempral resolution is consistent among videos in the same class but vary for videos across classes. A pattern classification model can be developed based on the content features computed from the video in the compressed domain. For an incoming video, its optimal operation can thus be predicted based on its content features. Specifically, the purpose of this method is to find a model C such that $c_l = C(\vec{F})$, where c_l is the labeled class from a set $\{c_1, c_2, ...\}$, whose values are taken from a set of pre-generated classes, and $\vec{F}$ is the content feature vector extracted from the video. For each class of videos c_l, there is a corresponding optimal operation recommendation $O_s(c_l)$ shared within this class, which is denoted by a suggested temporal scalability operation and the SNR scalability can be deduced afterwards by considering the target bit rate constraint.

Figure 9 includes an illustrative diagram of our proposed system. The upper part is the overall structure. For each video stream, the content features are extracted and the optimal operator prediction is applied. The adaptation engine, which is located in either the server or a standalone third part proxy, reshapes the stream according to the predicted optimal operator. The lower part provides more details of the scalability operator prediction model. It can be roughly categorized

to an offline training route and an online processing route, while the former is in charge of the class definiation and classifier learning, and the later undergoes online classification decision. For the class definition part, firstly we build a media pool using training video clips. The perceptual quality data and content features of each clip in the pool are collected in advance. Then the videos are clustered based on their perceptual quality behavior. Given the labeled instances in the pool, the classifier can be trained using the content features. For the online processing routing, the content features of the incoming video will be used by the classification function. According to the classification result, corresponding optimal operator will then be activated during the bit rate adaptation process.

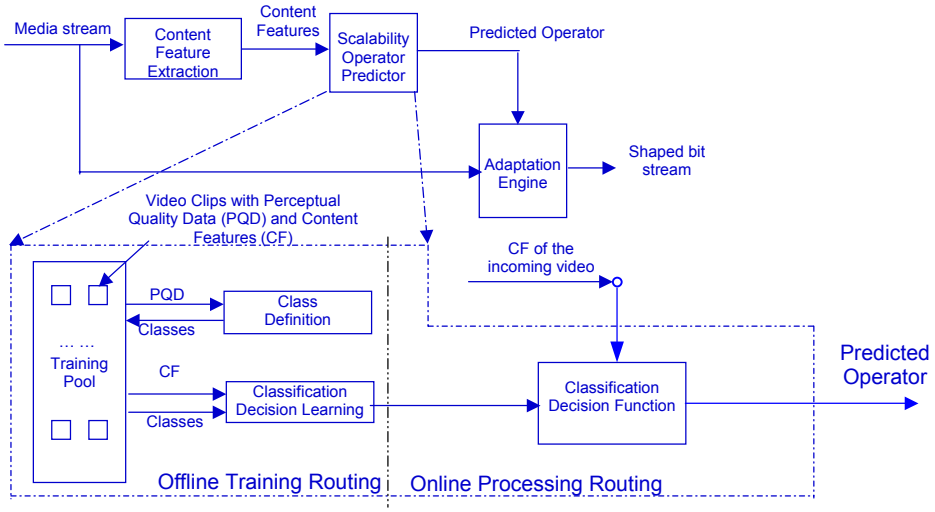


Figure 9: Proposed optimal scalability prediction system architecture

4.1 Optimal Scalability Operation Based On Subjective Experiment

The first step of our statistical model is the classification of the optimal operations for a video set with different content types. We collect some typical video sequences and pre-generate a set of bit streams using different operations for several representative bit rates. We then setup a subjective experiment and ask the users to estimate the perceptual quality of some video clips in the same bit rate and select the optimal one based on their preference. The subjective evaluation results are collected and further averaged over different users. At last, the video clips are manually classified into several classes based on their optimal operation. Note the distinctive classes are formed based on the required temporal resolution of the video instead of the content features.

4.2 Content Feature Selection

Content feature selection is an important step in classification application. The relevant content feature set depends on the implementation of the systems. Empirically we found for MC-3DSBC systems, the optimal operation selection is highly related to the complexity of texture and motion. In our experiments, we use MC-EZBC[3] codec and include the following features: texture energy (the summation of the squared coefficient magnitude) for each temporal layer; motion magnitude for each temporal layer; histogram of the variable block size during motion compensation. 20 features are totally extracted from each video. These content features are extracted during the coding process and detailed in Section 5. In general these features reflect the texture energy and motion magnitude information of the video. These raw features need further analysis and careful selection in order to simplify the classification. Basically we adopt the mutual information feature selection (MIFS) algorithm[14] to help us pick a smaller set of salient content features. MIFS select out the salient feature one by one from the pool by evaluating the mutual information between the class label and each feature. In order to determinate the number of selected features, we analyze the eigenvalues by using principle component analysis (PCA) and choose the minimum number of features when a specified energy ratio is reached.

4.3 Prediction Model Training

We employ the multiple layer perceptron (MLP) as our classification method, or classificatin decision function. MLP is known as a neural network that can present arbitrary logical expression[16]. Other classification models may also be used. Our goal is simply to discover the relationship between the low level content feature and the optimal scalability operation.

4.4 Experiment Validation

To validate the proposed optimal operation prediction framework, we adopt the MC-EZBC system as the scalable coding platform. We select 10 video clips with 288 frames each. These videos are *Akiyo*, *Stefan*, *Mobile*, *Coastguard*, *Container*, *Foreman* and four other clips taken from commercial movies. These four clips are described in Table 3. The format of the video is CIF resolution with 30fps. During the subjective expeirment, each video is coded into several typical bandwidth ranging from 50*kbps* to 600*kbps*. For bit rates higher than this, user perferences have little variation since the SNR quality is good enough and the higher frame rate is always preferred. For each bit rate, three versions of bit streams are generated with different temporal scalability: full frame rate (no temporal scalability), one half frame rate (one-level down temporal scalability) and one fourth frame rate (two-level down temporal scalability). Users are asked to select the preferred frame rate and their opinions are averaged. 10 Ph.D. students participate in the test and the subjective evaluation results are shown in Figure10, where each curve corresponds to one video clip. The average frame rate preference indicates the suggested frame rate (temporal scalability operation) on a given bit rate. Based on the clustering of the temporal rate preferences over different bit rates, we manually divide these videos into 5 classes: 1) *Container*; 2) *Akiyo*, *Mary3*; 3) *Coastguard*, *Mary9*, *Foreman*, *Laurance4*; 4) *Stefan*, *Laurance6*; 5) *Mobile*. For a larger video corpse, we can apply unsupervised clustering to automatically discover these classes. Within each class, we can abstract an averaged curve among all video clips as the representative one (not drawn in Figure 10 for clarity). Once an incoming video is classified into one of these classes, it can use the representative curve to select the optimal scalability operation.

Table 3: Description for video sequence besides the standard test sequences

Clip Name	Content
Mary3	A couple are talking during an interview with no camera motion
Mary9	A woman and a man are eating in a restaurant with camera undergoing translation
Laurence 4	Two men are riding slowly
Laurence 6	Two men are talking in a car with people marching in the background

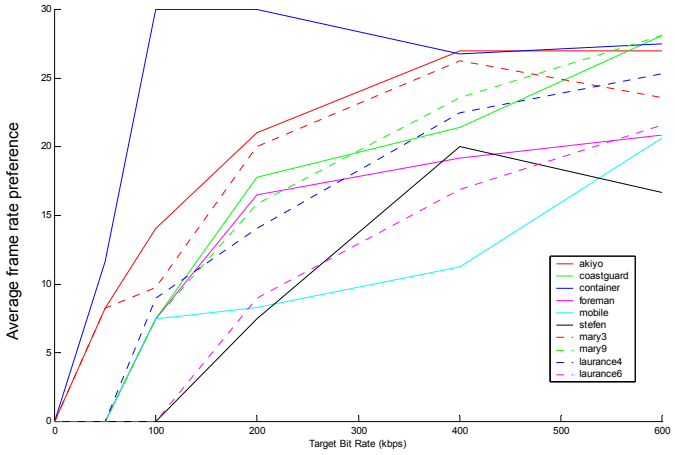
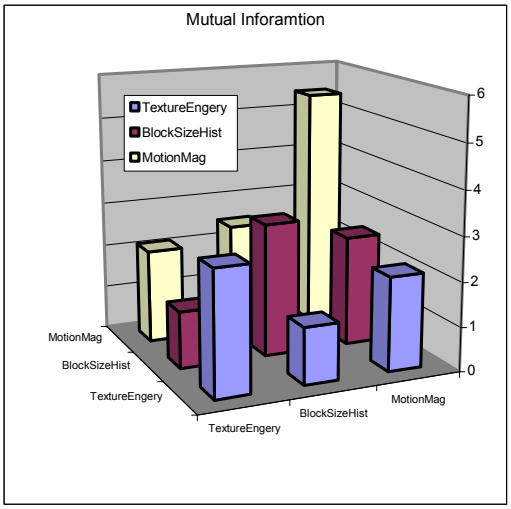
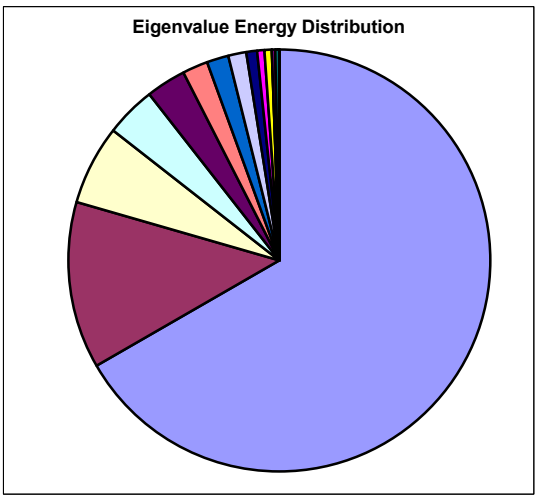


Figure 10: Subjective evaluation result

During the content feature analysis step, based on the characteristic of the MC-EZBC codec, totally 20 features are extracted from each video: 5 texture energy (TextureEnergy) for each temporal layer; 10 motion magnitude (MotionMag) for each temporal layer; 5-bin histogram (BlockSizeHist) of the variable block size during motion compensation. The mutual information among these three sets of features is shown in Figure 11(a). We can see that the motion magnitude and the block size histogram have the maximum correlation. Feature selection is carried out based on MIFS algorithm. MIFS itself can not automatically decide the amount of target features. To address this problem, we apply PCA on our feature set. The PCA result in Figure 11(b) indicates the energy representative from all 20 eigenvalues. We select six features for MIFS because above 95% energy of original space can be reached by 6 dominant eigenvalues. The selected features by MIFS comprise one from texture energy, two from block size histogram and three from motion magnitude. This result indicates that these three sets of features all have their contribution in some degree. We will verify their representativeness during the classification procedure.



(a) Mutual information among the feature sets



(b) Energy distribution after PCA
(Each pie slice stands for one eigenvalue)

Figure 11: Content feature analysis

During the optimal operation prediction step, the video clips are decomposed into units with a length of 16 frames (one GOP) and the content features for each GOP are extracted. Each GOP is considered as a data observation in the pool. Thus, we generate (288/16*10=)180 samples in total. These samples undergo the cross-validation testing and the accuracy of classification is shown in Figure 12. The accuracy is calculated by averaging the result over 10 runs. In the figure, the "full feature set" result is based on the whole feature space; the "MIFS selected 6 features" result is based on the selected 6 features by MIFS. As a comparison, four other results are also listed: three cases using only motion magnitude, block size histogram and texture energy respectively; and one case using arbitrary selected 6 features. It is clear that our proposed content-based optimal operation prediction can yield satisfactory accuracy. Also the content feature selection by PCA and MIFS can successfully find a compact subset of salient features which can be used to achieve performance comparable to that using the full set.

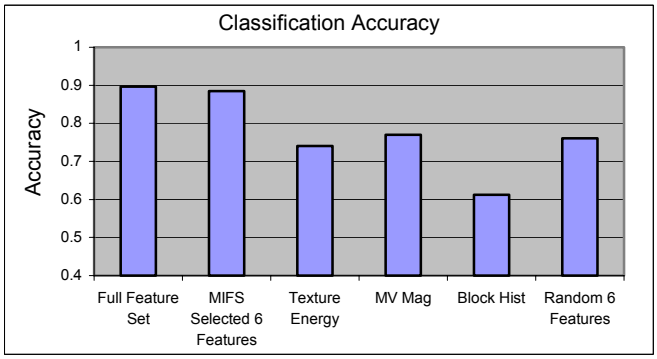


Figure 12: Content based classification accuracy

5. CONCLUSION AND FUTURE WORK

In this paper, we analyze the behavior of SNR-Spatial-Temporal scalability of MC-3DSBC systems. Especially for the SNR scalability, we explore several different rate control methods and find that uniform bitplane truncation, though simple, can yield close to optimal video quality. In order to choose the optimal scalability operation in the multiple scalability dimentions of MC-3DSBC, we propose a content based prediction framework in which computable features

are extracted from video streams and used as input to the prediction function. Also we run PCA-MIFS combined method to guide the feature selection. The experiment yields very promising results: (1) the proposed content-based prediction framework yields very good (up to 90%) accuracy and (2) the content feature selection by PCA and MIFS can well represent the energy of original feature space.

Many open issues still exist. Firstly, we currently simplify the problem by ignoring spatial scalability in our analysis and only predict the SNR-temporal scalability operation. The full SNR-spatial-temporal scalability optimal selection requires further research. Secondly, in SNR scalability, theoretical verification is needed to analyze UBT's superior performance. Lastly, due to the workload of subjective evaluation, the amount of the samples in our video pool is small. We are now expanding our video corpus . In the meantime, development of a spatial-temporal objective quality model that can be used to approximate the subjective quality is also an interesting open issue.

6. ACKNOWLEDGMENT

We thank Peisong Chen for his kind help in providing the codes of MC-EZBC, to Gounyoung Kim and Dr. Deepak S. Turaga for their valuable discussion.

REFERENCES

1. Jens-Rainer Ohm. "Three-Dimension Subband Coding with Motion Compensation". IEEE Transactions on Image Processing. Vol. 3. No. 5. September 1994.
2. Seung-Jong Choi and John W. Woods. "Motion-Compensated 3-D Subband Coding of Video". IEEE Transaction on Image Processing. Vol. 8. No.2. February 1999.
3. Peisong Chen and John W. Woods. "Bidirectional MC-EZBC With Lifting Implementation". August, 2002. Accepted for publish in IEEE trans. on Circuits and Systems for Video Technology, 2003
4. Jerome M. Shaprio. "Embedded Image Coding Using Zerotrees of Wavelet Coefficients". IEEE Transactions on Signal Processing. Vol. 41. No.12 December 1993.
5. Amir Said, William A. Pearlman. "A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees". IEEE Transactions on Circuits and Systems for Video Technology. Vol. 6. June 1996.
6. Andrew, "A simple and efficient hierarchical image coder," ICIP 97, IEEE Int'l Conf. on Image Proc. 3, pp. 658, paper no. 573, 1997.
7. A. Islam and W. A. Pearlman, "An Embedded and Efficient Low-Complexity Hierarchical Image Coder," Visual Communications and Image Processing '99, Proceedings of SPIE Vol. 3653, pp. 294-305, Jan. 1999.
8. Shih-Ta Hsiang and J. W. Woods,"Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," MPEG-4 Workshop and Exhibition at ISCAS 2000, Geneva, Switzerland, May 2000.
9. Shih-Ta Hsiang and J. W. Woods, "Embedded video coding using motion compensated 3-D subband/wavelet filter bank," Packet Video Workshop, Sardinia, Italy, May 2000.
10. D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video", IEEE Trans. On Image Processing, Vol. 3, pp 572-588, Sept. 1994.
11. Peisong Chen and John W. Woods. MC-EZBC video codec. http://mpeg.nist.gov/NIST CVS Repository/AHG-SW
12. Masry, M. and Hemami, S.S., CVQE: A Continuous Video Quality Evaluation Metric for Low Bit Rates, SPIE Conf. on Human Vision and Electronic Imaging, 2002
13. Y. Wang, J.-G. Kim, and S.-F. Chang, Content-based utility function prediction for real-time MPEG-4 transcoding, ICIP 2003, September 14-17, 2003, Barcelona, Spain.
14. Roberto Battiti. "Using Mutual Information for Selecting Features in Supervised Neural Net Learning". IEEE Trans. On Neural Networks, 5(4):537~550, July 1994.
15. D. S. Taubman, "High performance scalable image compression with EBCOT," IEEE Trans. Image Proc., vol. 9, pp. 1158-1170, July 2000.
16. PO. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Wiley, New York, 2001.
17. I. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, "Wavelet-based Fully-Scalable Video Coding with In-band Prediction," Proceedings of IEEE Signal Processing Symposium (SPS), Leuven, Belgium, pp. 217-220, March 21-22, 2002.
18. J.M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients", IEEE Trans. on Signal Processing, v. 41, no. 12, pp. 3445-3463, Dec. 1993

19. CJ van den Branden Lambrecht, Perceptual quality measure using a spatio-temporal model of the human visual system, Proceedings of the SPIE, vol. 2668, pp. 450-461, San Jose, 1996.
20. R. Kumar Rajendran, M. van der Schaar, S.-F. Chang, "FGS+: Optimizing the Joint SpatioTemporal Video Quality in MPEG-4 Fine Grained Scalable Coding," IEEE International Symposium on Circuits and Systems (ISCAS), Phoenix, Arizona, May 2002.
21. Christopoulos, C., Skodras, A. and Ebrahimi, T., The JPEG2000 still image coding system: an overview. Consumer Electronics, IEEE Transactions on, Volume: 46 Issue: 4, Nov. 2000. Page(s): 1103 -1127
22. J.-G. Kim, Y. Wang, S.F. Chang, Content-Adaptive Utility-Based Video Adaptation, IEEE ICME-2003. July 6-9, 2003. Baltimore, Maryland.
23. A. Eleftheriadis, *Dynamic Rate Shaping of Compressed Digital Video*, Doctoral Dissertation, Graduate School of Arts and Sciences, Columbia University, June 1995.
24. W. Li, Overview of fine granularity scalability in MPEG4 video standard, IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 3, pp. 301-317, Mar. 2001.
25. Niranjan Damera-Venkata, Thomas D. Kite, Wilson S. Geisler, Brian L. Evans and Alan C. Bovik, Image Quality Assessment Based on a Degradation Model. IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 636-651, April 2000.
26. A. Rohaly, J. Libert, P. Corriveau, A. Webster, et al., Final report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, ITU-T Standards Contribution COM 9-80-E, Jun. 2000.