MASTER'S THESIS

3D Wavelet-Based Video Codec with Human Perceptual Model

by Junfeng Gu
Advisor: John S. Baras

CSHCN M.S. 99-3
(ISR M.S. 99-3)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Supported by: NASA and Texas Instruments

ABSTRACT

Title of Thesis: 3D WAVELET BASED VIDEO CODEC WITH HUMAN PERCEPTUAL MODEL

Degree Candidate: Junfeng Gu

Degree and Year: Master of Science, 1999

Thesis Directed by: Professor John S. Baras

Institute of Systems Research

This thesis explores the utilization of a human perceptual model in video compression, channel coding, error concealment and subjective image quality measurement. The perceptual distortion model just-noticeable-distortion (JND) is investigated. A video encoding/decoding scheme based on 3D wavelet decomposition and the human perceptual model is implemented. It provides a prior compression quality control which is distinct from the conventional video coding system. JND is applied in quantizer design to improve the subjective quality of compressed video. The 3D wavelet decomposition helps to remove spatial and temporal redundancy and provides scalability of video quality. In order to conceal the errors that may occur under bad wireless channel conditions, a slicing method and a joint source channel coding scenario, that combines RCPC with CRC and utilizes the distortion information to allocate convolutional coding rates are proposed. A new subjective quality index based

on JND is proposed and used to evaluate the overall performance at different signal to noise ratios (SNR) and at different compression ratios.

Due to the wide use of arithmetic coding (AC) in data compression, we consider it as a readily available unit in the video codec system for broadcasting. A new scheme for conditional access (CA) sub-system is designed based on the cryptographic property of arithmetic coding. Its performance is analyzed along with its application in a multi-resolution video compression system. This scheme simplifies the conditional access sub-system and provides satisfactory system reliability.

3D WAVELET BASED VIDEO CODEC WITH HUMAN PERCEPTUAL MODEL

by

Junfeng Gu

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland at College Park in partial fulfillment
Of the requirements for the degree of
Master of Science
1999

Advisory Committee:

Professor John S. Baras, Chair/Advisor
Professor Rama Chellappa
Professor Nariman Farvardin
Professor Haralabos Papadopoulos

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and thanks to my advisor Dr. John S. Baras for his guidance, patience, and support throughout my study and thesis research at University of Maryland at College Park for these two years.

I would thank Mr. Yimin Jiang and Mr. Li Liu. I got a lot from the discussion with them. Several times I came to the “no outlet” of research, their brilliant suggestions helped me find the right ways. I also owe gratitude to Professor Farvardin, Professor Chellappa, Professor Papadopoulos and other faculties at Department of Electrical and Computer Engineering, for their excellent lectures, instructions, helpful advice, and for serving on my committee.

I would like to give me special thanks to my wife. Her support and love accompany me all the time. Thanks for the understandings.

This work was supported by NASA contracts under the cooperative agreement NASA-NCC5227 and NASA-NCC3528, and by a contract from Texas Instruments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Backgrounds

The scale and power of computing and communication systems is keeping on increasing. Meanwhile, the data required to represent the image and video signal in digital form would continue to overwhelm the capacity of many communication and storage systems. In particular, the growth of data-intensive digital video and image applications and the increasing use of bandwidth-limited media such as radio and satellite links have not only sustained the need for more efficient ways to encode analog signals, but have made signal compression central to digital communication and signal-storage technology. Furthermore, for some applications such as image/video database browsing and multipoint video distribution over heterogeneous networks and display devices, there is a growing need for other useful features such as video scalability. Highly scalable video compression schemes [5][14][15] allow selective transmission of different sub-bitstreams to different destinations, depending on their respective needs. In this manner, each receiver can have the best possible quality session according to its bandwidth.

The ultimate object of a video/image compression system is to minimize the average number of bits used to represent the digital video/image signal while maintaining subjective video/image quality as good as possible. Since the traditional measures of image signal quality, mean square error (MSE), and signal-to-noise ratio (PSNR), do not provide a satisfactory reflection of the human's subjective perception on the video/image quality, more satisfactory metrics should take advantage of the

human perception model implicitly or explicitly. They give the encoder certain fidelity to allocate more bits to signals that are more meaningful to the human visual system. And they lead to better quantitative evaluation methods. A variety of schemes have been proposed to incorporate certain psychovisual properties of the human visual system (HVS) into image/video coding algorithms [16][17][18][20][42]. The frequency sensitivity, brightness sensitivity, texture sensitivity and color sensitivity are considered for the distortion sensitivity profiles, which leads to modern human visual models, such as just-noticeable-distortion (JND) [16], visible differences predictor (VDP) [19] and Ran's perceptually motivated three-component image model [20]. Jayant's JND model provides each signal being coded with a threshold level of error visibility, below which reconstruction errors are rendered imperceptible. The JND profile of a video sequence is a function of local signal properties, such as brightness, background texture, luminance changes between two frames, and frequency distribution.

To approach the goal of high compression ratio and scalability, subband coding is widely explored in recent years. The signal is decomposed into frequency subbands and these subbands are encoded independently or dependently. The structures in the high frequency subbands usually appear as sparse edges and impulses corresponding to the localized discontinuities in spatial or temporal domains. After carefully designed quantization, these subbands lead to a large quantity of compression, provide spatial and bitstream scalability naturally, and require less error protection in channel coding. Earlier work on subband coding for image compression applies to the still image [1][2]. Tanabe and Farvardin [8] suggest a scheme using entropy-coded quantization to get the optimal quantizer performance. Kwon and Chellappa [9] use adaptive entropy-

constrained quantizers for different regions in images. The early work on subband coding of video includes [3] and [4]. In recent years, the subband decomposition has been extended to three dimensions. The codec of Taubman and Zakhor [5], which employs a global motion compensation scheme accounting for camera panning motion, generates a single embedded bit stream supporting a wide range of bit rates. Podichuk, Jayant and Farvardin [6] combine a 3-D subband coder with geometric vector quantization and obtain good compression performance at low bit rates. There are novel wavelet-based video coding systems that take advantage of good features of other components such as overlapped motion estimation. The video codec developed by David Sarnoff Research Center [10] uses the technology of overlapped block motion compensation and zero-tree entropy coding (ZTE), which is the extension of Shapiro's EZW [11] and Said & Pearlman's SPIHT [12]. It outperforms the VM of MPEG-4 and provides scalability. Cinkler [13] uses an edge-sensitive subband coding (ESSBC) method and overlapped motion compensation. From an edge map combined with motion vectors, the ESSBC technology generates areas of significance. These areas are processed by a modified wavelet transform to concentrate the energy.

Even if a very high percentage of total signal energy is contained in the lowest frequency subband, the truncation or undercoding of high-band signals will result in the perception of distortion due to aliasing effects. On the other hand, unless the significant signals are cautiously encoded, the overcoding of high-band signals is the price to pay for gaining good image quality. Consequently, the problem to be solved for optimizing the subband coding scheme is how to locate perceptually important signals in each

frequency subband, and how to encode these signals with the lowest possible bit rate without exceeding the error visibility threshold.

As a critical (and often controlling) technology in the video broadcasting industry, a conditional access sub-system comprises a combination of scrambling and encryption to prevent unauthorized reception. Encryption is the process of protecting the secret keys that are transmitted with a scrambled signal to enable the descrambler to work. Way back in 1988, an attempt was made by France Telecom and others to develop a standard encryption system for Europe. The result was Eurocrypt. Unfortunately, in its early manifestations it was not particularly secure and multiplex operators went their own way. Thus, in 1992 when the DVB started their consideration of CA systems, they recognized that the time had passed when a single standard could realistically be agreed upon and settled for the still difficult task of seeking a common framework within which different systems could exist and compete. They therefore defined an interface structure, the Common Interface, which would allow the set top box (STB) to receive signals from several service providers operating different CA systems. The common interface module contains the CA system, rather than the STB itself, if necessary allowing multiple modules to be plugged into a single STB. However, there were serious objections to the common interface module from many CA suppliers on the grounds that the extra cost would be unacceptable. As a result, the DVB stopped short of mandating the Common Interface, instead recommending it, along with simulcrypt, which is one of the DVB recommended approaches for conditional access. These all bring up a diversified market of conditional access system, which makes the exploration in this field so valuable.

1.2 Contribution of This Thesis

We have implemented a video encoding/decoding scheme based on 3-D wavelet decomposition and a human perceptual model. Jayant's just-noticeable-distortion (JND) model is adopted. The quantizers in different subbands are designed to approach perceptual optimum. The source encoder has the global control on subjective distortion of the compressed video quality ahead of time, which is distinct from the conventional compression schemes. A new subjective distortion index for video is proposed and used to evaluate the overall performance. Its fidelity is compared with the traditional quality metric PSNR. From the result of simulation, we conclude that our distortion index based on JND profile is more accurate than PSNR in the sense of measuring the human subjective distortion. Using the new distortion index, the performance of our video codec is compared with the coding of I frames in MPEG. At the same bit rate, our encoder has performance comparable with MPEG encoder for I frames. Our simulation shows that our encoder assigns more error energy to the perceptually less important pixels in the frames. But due to the lack of motion estimation and run-length coding technologies, the overall compression performance of our encoder is worse than MPEG. To present its application more concretely, a practical transmission system over a satellite channel using unequal error protection is discussed. Since in the satellite broadcasting case, a feedback channel is not available, the transmitter has no information about the receivers and their channel environments. It is difficult to guarantee the average video qualities under diversified channel conditions without large channel coding overhead. We derive a new slicing method to truncate the data from each subband into small slices before arithmetic coding to confine the propagation of bit

errors. Rate compatible punctured convolutional (RCPC) codes [27] are adopted in our system to provide unequal error protection for different subbands. The bit rates of RCPC for these subbands are finely chosen following the JND model to make the unequal error protection perceptually sub-optimal. Simulations are done for different combinations of RCPC coding and channel SNR, showing some characteristics of our coding scheme.

Following the rapid expansion of the commercial broadcasting industry, a conditional access sub-system is always included in the broadcasting system. It is used to control which customer can get particular program services. Particular programs are only accessible to customers who have satisfied the required payments. In this paper, a brand new conditional access sub-system which takes advantage of the cryptographic property of arithmetic coding is suggested. And a video broadcasting system based on subband coding is described to present the application of this new condition access sub-system. The performance analysis is provided. Compared to the traditional structures, our scheme is quite simple and of low cost while provides reliable security.

Chapter 2 The Human Perceptual Model of JND

2.1 Human Perceptual Models and Perceptual Coding

2.1.1 Human Perceptual Models

A common model of vision incorporates a low-pass filter, a logarithmic nonlinear transformer, and a multi-channel signal-sharpening high-pass filter [26]. A biologically correct and complete model of the human perceptual system would incorporate descriptions of several physical phenomena including peripheral as well as higher level effects, feedback from higher to lower levels in perception, interactions between audio and visual channels, as well as elaborate descriptions of time-frequency processing and nonlinear behavior. Some of the above effects are reflected in existing coder algorithms, either by design or by accident. For example, certain forms of adaptive quantization and prediction provide efficient performance in spite of inadequate response time because of temporal noise masking. The basic time-frequency analyzers in the human perceptual chain are described as bandpass filters. Bandpass filters in perception are sometimes reflected in coder design and telecommunication practice in the forms of “rules of thumb”.

A particularly interesting aspect of the signal processing model of the human system is non-uniform frequency processing. The critical bands in vision are non-uniform. It is necessary to use masking models with a non-uniform frequency support to incorporate this in coder design. Here masking refers to the ability of one signal to hinder the perception of another within a certain time or frequency range. It is also necessary to recognize that high-frequency signals in visual information tend to have a

short time or space support, while low-frequency signals tend to last longer. An efficient perceptual coder therefore needs to not only exploit properties of distortion masking in time and frequency, but also have a time-frequency analysis module that is sufficiently flexible to incorporate the complex phenomena of distortion masking by non-stationary input signals. All of this is in contrast to the classical redundancy-removing coder, driven purely by considerations of minimum mean square error (MMSE), MMSE bit allocation, or MMSE noise shaping matched to the input spectrum.

Distortion sensitivity profiles of human perception are driven as functions of frequency, brightness, texture, and temporal parameters. These four kinds of sensitivity are under consideration for gray scale video/image [16][18][26]:

(1) Brightness sensitivity:

It was found that human visual perception is sensitive to luminance contrast rather than absolute luminance values. The ability of human eyes to detect the magnitude difference between an object and its background is dependent on the average value of background luminance. According to Web's Law [28], if the luminance of a test stimulus is just noticeable from the surrounding luminance, the ratio of just noticeable luminance difference to stimulus's luminance (Weber fraction) is almost constant. However, due to the ambient illumination falling on the display, the noise in dark areas tends to be less perceptible than that occurring in regions of high luminance. In general, high visibility thresholds will occur in either very dark or very bright regions, and lower thresholds will occur in regions of gray levels close to the mid-gray luminance, which is 127 for 8 bit sampling.

(2) Texture sensitivity:

The reduction in the visibility of stimuli due to the increase in spatial nonuniformity of the background luminance is known as texture masking. Several efforts have been made to utilize some forms of texture making to improve coding efficiency. In many approaches, visibility thresholds are defined as functions of the amplitude of luminance edge in which perturbation is increased until it becomes just discernible. The visibility threshold in this approach is associated with the masking function defined at each pixel as the maximum prediction error from the four neighboring pixels.

(3) Temporal sensitivity:

The masking of temporally changing stimuli is extremely important in interframe coding. However, temporal masking is complicated by many factors, and its application to video coding is still in its infancy. Many researches have attempted to evaluate the losses of spatial resolution and magnitude resolution as an object moves in a scene. If movement is drastic, such as scene change, the perceived spatial and intensity resolution is significantly reduced immediately after the scene change. It was found that the eye is noticeably more sensitive to flicker at high luminance than at low luminance [28].

(4) Frequency sensitivity:

Many psychovisual studies have shown that the human perception of distortion depends on its frequency distribution. The response of the HVS to sinewave gratings of different frequencies has been experimentally measured as the so-called contrast sensitivity function (CSF). Many models of spatial-domain CSF have been proposed,

which indicate general bandpass characteristics. The spatial-domain CSF has been widely used to improve the quality of the coded still images. There are only a few models of spatio-temporal CSF reported in the literature, among which the most well known model is proposed by Kelly [22]. The spatio-temporal CSF provides relative sensitivities of the HVS to different spatio-temporal frequencies, or relative tolerance of noises at different spatio-temporal frequencies. It can be used to allocate coding bits, or distortion, by adjusting the quantizer stepsize of the target signal as inversely proportional to the sensitivity of the corresponding frequency.

2.1.2 Perceptual Coding

There are two intrinsic operations to signal coding: removal of redundancy and reduction of irrelevancy [16]. The removal of redundancy is the effect of predictive coding or transform coding. Almost all sampled signals in coding are redundant because Nyquist sampling typically tends to preserve some degree of inter-sample correlation. This is reflected in the form of a nonflat power spectrum. Greater degrees of nonflatness, as resulting from a low-pass function for signal energy versus frequency, or from periodicities, lead to greater gains in redundancy removal. These gains are also referred to as prediction gains or transform coding gains, depending on whether the redundancy is processed in the spatial domain or in the frequency (or transform) domain.

The reduction of irrelevancy is the result of amplitude quantization. In a signal compression algorithm, the inputs of the quantizing system are typically sequences of prediction errors or transform coefficients. The idea is to quantize the prediction error, or the transform coefficients, just finely enough to render the resulting distortion

imperceptible, although not mathematically zero. If the available bit rate is not sufficient to realize this kind of perceptual transparency, the intent is to minimize the perceptibility of the distortion by shaping it advantageously in space or frequency, so that as many of its components as possible are masked by the input signal itself. The term perceptual coding is used to signify the matching of the quantizer to the human visual system, with the goal of either minimizing perceived distortion, or driving it to zero where possible. These goals do not correspond to the maximization of signal-to-noise ratio or the minimization of mean square error.

2.2 Just-noticeable-distortion (JND) Profile

To remove the redundancy due to spatial and temporal correlation and the irrelevancy of perceptually insignificant components from video signals, the concept of just-noticeable distortion profile introduced by Jayant [21] has been successfully applied to perceptual coding of video and image. JND provides each signal to be coded with a visibility threshold of distortion, below which reconstruction errors are rendered imperceptible. The JND profile of a still image is a function of local signal properties, such as the background luminance and the activity of luminance changes in the spatial domain. For video sequences, the derivation of JND profiles must take both spatial and temporal masking effects into consideration. For a successful estimation of JND profiles, the subject should not be able to discern the difference between a video sequence and its JND-contaminated version. Once JND profiles of video signals are obtained, the perceptual redundancy can be quantitatively measured, and the perceptual significance of each target signal can be evaluated.

Whenever transparent coding cannot be attained due to a tight bit-rate budget, the minimally-noticeable-distortion (MND) profile rather than the JND profile is required. In MND the increased distortion can be uniformly distributed over the reconstructed video signals and thus there is minimally perceptible. The perceptual quality of the reconstructed video signals is expected to degrade gracefully if the available bit rate is reduced.

The generation of a JND model consists of several steps [18]. First, the perceptual redundancy inherent in the spatial domain is quantitatively measured as a 2D profile by a perceptual model that incorporates the visibility thresholds due to average background luminance and texture masking [17]. It is described by the following expression [17]:

$$JND_S(x,y) = \max\{f_1(bg(x,y), mg(x,y)),\ f_2(bg(x,y))\}, \quad for \quad 0 \le x < W, 0 \le y < H \tag{1}$$

where f_1 represents the error visibility threshold due to texture masking, f_2 the visibility threshold due to average background luminance; H and W denote respectively the height and width of the image; $mg(x,y)$ denotes the maximal weighted average of luminance gradients around the pixel at (x,y); $bg(x,y)$ is the average background luminance.

$$f_1(bg(x,y), mg(x,y)) = mg(x,y)\alpha(bg(x,y)) + \beta(bg(x,y)) \tag{2}$$

$$f_2(bg(x,y)) = \begin{cases} T_0 \cdot (1 - (bg(x,y)/127)^{1/2}) + 3 \\ \qquad for \quad bg(x,y) \le 127 \\ \gamma \cdot (bf(x,y) - 127) + 3 \\ \qquad for \quad bg(x,y) > 127 \end{cases} \tag{3}$$

$$\alpha(bg(x,y)) = bg(x,y)\cdot 0.0001 + 0.115$$

$$\beta(bg(x,y)) = \lambda - bg(x,y)\cdot 0.01 \qquad for \quad 0 \le x < H \quad 0 \le y < W$$

where T_0, γ and λ are found to be 17, 3/128 and ½ through experiments [17].

The value of *mg(x,y)* across the pixel at *(x,y)* is determined by calculating the weighted average of luminance changes around the pixel in four directions. Four operators, $G_k(i,j)$, for *k*=1,…4, and *i, j*=1,…5, are employed to perform the calculation, where the weighting coefficient decreases as the distance away from the central pixel increases.

$$mg(,x,y) = \max_{k=1,2,3,4}\{|grad_k(x,y)|\} \tag{4}$$

$$grad_k(x,y) = \frac{1}{16}\sum_{i=1}^{5}\sum_{j=1}^{5} p(x-3+i, y-3+j)\cdot G_k(i,j)$$
$$for \quad 0 \le x < H, \quad 0 \le y < W,$$

where *p(x,y)* denotes the pixel at *(x,y)*. The operators $G_k(i,j)$ are

$$G_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 8 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -3 & -8 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \; G_2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 8 & 3 & 0 & 0 \\ 1 & 3 & 0 & -3 & -1 \\ 0 & 0 & -3 & -8 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix},$$

$$G_3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 8 & 0 \\ -1 & -3 & 0 & 3 & 1 \\ 0 & -8 & -3 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}, \; G_4 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 8 & 0 & -8 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

The average background luminance, *bg(x,y)*, is calculated by a weighted lowpass operator, *B(i,j), i, j*=1,…,5.

$$bg(x,y) = \frac{1}{32}\sum_{i=1}^{5}\sum_{j=1}^{5} p(x-3+i, y-3+j)\cdot B(i,j), \qquad (5)$$

where $B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$.

At the second step of JND model generation, the JND profile representing the error visibility threshold in the spatio-temporal domain is expressed as

$$JND_{S-T}(x,y,n) = f_3(ild(x,y,n))\cdot JND_S(x,y,n) \qquad (6)$$

where $ild(x,y,n)$ denotes the average interframe luminance difference between the nth and $(n-1)$th frame. Thus, to calculate the spatio-temporal JND profile for each frame in a video sequence, the spatio JND profile of itself and its previous reference frame are required.

$$ild(x,y,n) = 0.5\cdot[p(x,y,n) - p(x,y,n-1) + bg(x,y,n) - bg(x,y,n-1)]$$

f_3 represents the error visibility threshold due to motion. The empirical results of measuring f_3 for all possible interframe luminance differences are shown in **Figure 1**. To purposely minimize the allowable distortion in the nonmoving area, the scale factor is switched to 0.8 as $|ild(x,y,n)| < 5$. It can be noted that the error visibility threshold is increased with the increasing interframe luminance difference. This conforms the research findings, that after a rapid scene change or large temporal difference, the

sensitivity of the HVS to spatial details is decreased. Moreover, it can be found that temporal masking due to high-to-low luminance changes is more prominent than that due to low-to-high luminance changes.

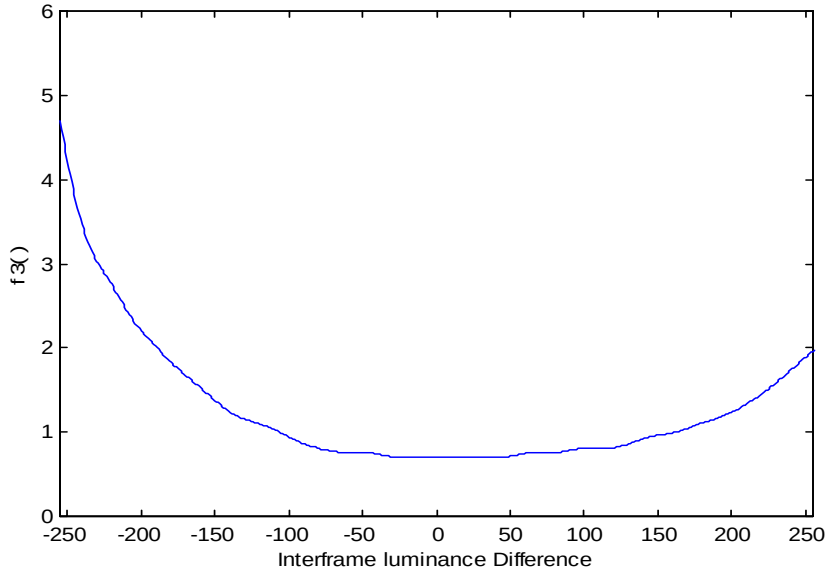

Figure 1 Error visibility threshold in the spatio-temporal domain

From the spatio-temporal JND profile that quantitatively measures the perceptual redundancy inherent in full-band video signals, the JND profiles for each subband are set up with certain distortion allocation [22] for different subbands. The JND for subband q is a function of spatio-temporal JND values at corresponding locations multiplied by a weight that indicates the perceptual importance of this subband. When each pair of video frames is decomposed into 11 spatio-temporal subbands as in **Figure 2** (the details of such a 3D wavelet decomposition is described in Chapter 3), the relationship between the full-band JND profile and the component JND profiles can then be obtained by the following equations:

$$JND_q(x,y) = \left[\left[\sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{t=0}^{1} JND_{S-T}^2(i + x \cdot 4, j + y \cdot 4, t) \right] \cdot \omega_q \right]^{1/2}, \quad (7)$$

$$for \quad 0 \le q \le 3, \quad 0 \le x < W/4, \quad 0 \le y < H/4$$

and

$$JND_q(x,y) = \left[\left[\sum_{i=0}^{2} \sum_{j=0}^{1} \sum_{t=0}^{1} JND^2_{S-T}(i + x \cdot 2, j + y \cdot 2, t) \right] \cdot \omega_q \right]^{1/2} \quad (8)$$
$$for \quad 4 \le q \le 10, \quad 0 \le x < W/2, \quad 0 \le y < H/2$$

The weighting function for distributing the full-band JND energy to a subband can be derived as the relative sensitivity of the HVS to the frequency subband. For 11 spatio-temporal subbands, the weighting function of the *q*th subband is obtained as

$$\omega_q = \frac{S_q^{-1}}{\sum_{i=0}^{10} S_i^{-1}}, \qquad for \quad 0 \le q \le 10 \quad (9)$$

where S_q represents the average sensitivity of the HVS to the *q*th subband. S_q is obtained from the spatio-temporal CSF presented by Kelly [22].

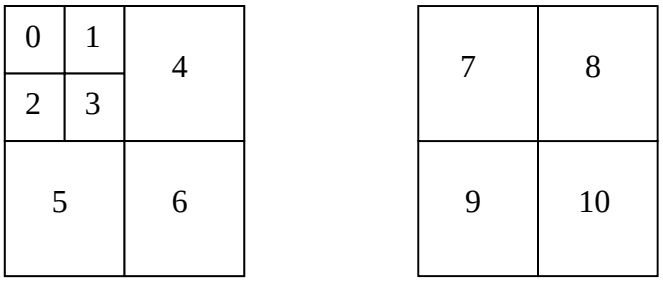

Figure 2 Subbands after 3D Wavelet Decomposition

2.3 A Novel Human Perceptual Distortion Measure Based on JND

Based on the basic concept of the JND, the idea of minimally-noticeable-distortion (MND) is developed for situations where the bit-rate budget is tight and the distortion in the reconstructed image is perceptually minimal at the available bit-rate and uniformly distributed over the whole image[26]. The perceptual quality of the reconstructed image is accordingly expected to degrade evenly if bit-rate is reduced. MND is expressed as:

$$MND(x, y) \equiv JND(x, y) \cdot d(x, y) \tag{10}$$

where $0 \le x < W$, $0 \le y < H$, W and H are the width and height of an image respectively, $d(x, y)$ is the distortion index at point (x, y).

The energy of JND can be understood as the minimum energy of quantization error that will cause conceivable loss to the human visual system. So if the error is so small in a small area that the error energy here is less than the JND energy, the compression will be perceptually lossless. We define the energy of the MND of a small area indexed by *(i,j)* as:

$$\sum_{(x, y) \in r_{ij}} MND^2(x, y) \equiv \sum_{(x, y) \in r_{ij}} JND^2(x, y) \cdot \delta(i, j) \tag{11}$$

where r_{ij} is a small block (typically 8 by 8 pixels), $\delta(i, j)$ is the distortion index for this block. We can define our global human perceptual distortion measure based on evaluating $\delta(i, j)$ as follows:

$$\Delta_G \equiv \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \varepsilon(k, l) \tag{12}$$

where $\varepsilon(k,l)$ is the distortion measure of a medium block indexed by (k,l). We decompose the whole image into *K* by *L* non-overlapped medium blocks (R_{kl}); each medium block is divided into *M* by *N* small blocks ($r_{ij}(k,l)$), i.e.,

$$R_{kl} = \bigcup_{i=1,M;\ j=1,N} r_{ij}(k,l).$$

$\varepsilon(k,l)$ is defined as:

$$\varepsilon(k,l) \equiv \text{median}\left(\delta(i,j) \middle| r_{ij}(k,l) \in R_{kl}, 1 \le i \le M, 1 \le j \le N \right) \qquad (13)$$

The larger Δ_G is, the larger the subjective perceptual distortion is. Compared with PSNR or MSE, Δ_G has the same convenience to describe the picture quality with one quantitative value. However, Δ_G takes the human visual model into account, therefore it can reflect subjective visual quality better than PSNR or MSE. It is well accepted that the value of MSE or PSNR is meaningless to video/image subjective evaluation and when two images are compressed, the comparison of their MSE or PSNR values cannot give a creditable psychovisual conclusion. On the other hand, the distortion Δ_G can be explained as " this image is compressed at a scale of Δ_G times the perceptually noticeable distortion". Generally speaking, if one image is coded with Δ_G larger than another one, the former's subjective quality is higher. Due to these considerations, we will use Δ_G as our index of performance for our video compression system. PSNR will be calculated at the same time as reference.

Chapter 3 Video Coding/Decoding System

Figure 3 and **Figure 4** show the JND model based video encoder and decoder respectively. In the video encoder, the input video sequence is decomposed into eleven spatio-temporal frequency subbands in the 3-D wavelet analysis module. The Frame Counter & Motion Detector controls the renewing of the JND profiles from frame counter and drastic movement detection. The JND Model Generators estimate the spatio-temporal JND profile from analyzing local video signals and the distortion allocation algorithm that determines the JND profile for each subband. The Perceptually Tuned Quantizer implements quantization for the wavelet coefficients in each subbands according to their JND profiles. The spatial LLLL temporal L subband will be encoded by DPCM. Then the data from all subbands goes through the Slicer and Arithmetic Coding part to do slicing and entropy coding. Afterward we get compressed video signals in 11 bit streams. These bit streams are fed into the Unequal Error Protection Channel Coder and an error protection indication from JND Model Generator is also given to the channel coder. These modules will be discussed subsequently.

In the decoder, there are Arithmetic Decoding and Merging, Error Detection, Error Concealment, Inverse Quantization and 3D Wavelet Synthesis modules. The Arithmetic Decoding and Merging part decodes the bit streams back to slices of wavelet coefficients, and gives them appropriate arrangements in the corresponding subbands. CRC coding alarms the Error Detector for bit errors. The latter's output information about the location of error helps the Error Concealment Module to diminish the effects of bad data from fault-causing noisy channel. Inverse Quantization and 3D Wavelet

Synthesis parts implement the inverse operation of the quantization and 3D wavelet decomposition in the video encoder. Finally the video signal is restored.

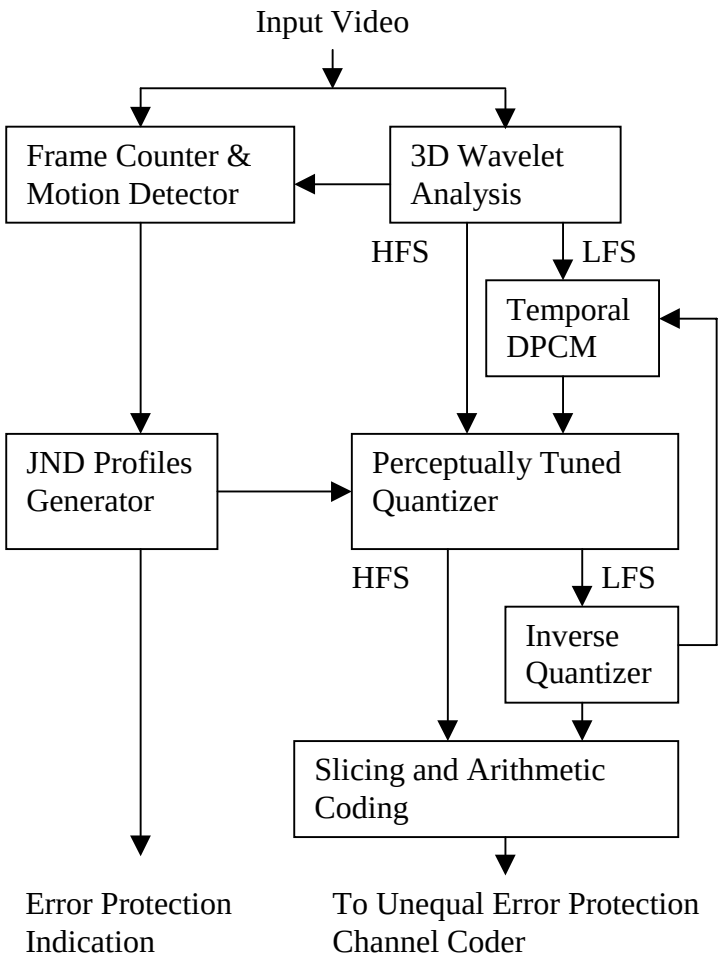

Figure 3 JND Based Video Encoder

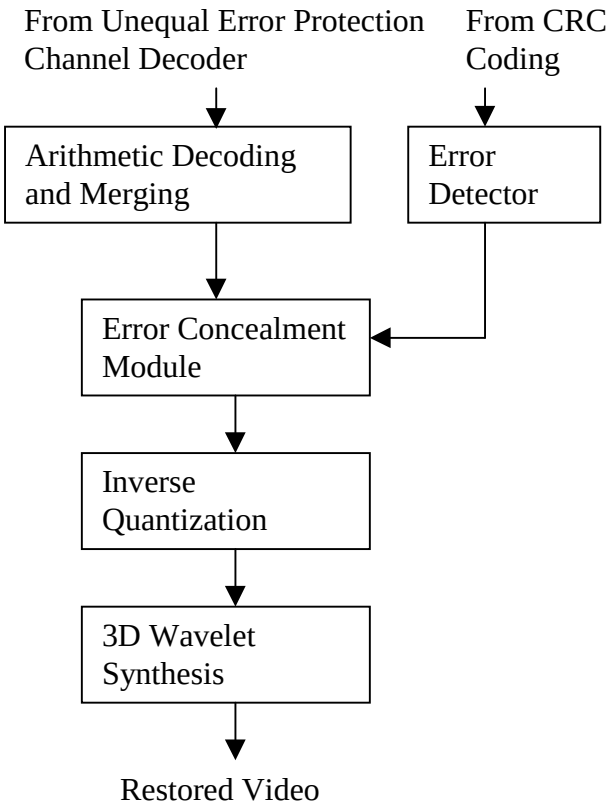

Figure 4 JND Based Video Decoder

3.1 3-D Wavelet Analysis

Wavelet multiresolution analysis techniques have been applied primarily to 1D and 2D signals. These techniques project the signal onto a chain of embedded approximation and detail spaces designed to represent the signal and its details at various levels of resolution [29]. For practical purposes, the projection coefficients are obtained using a discrete subband transform that employs a quadrature mirror filter pair related to the type of wavelet used in the analysis. In conventional 2D wavelet multiresolution analysis, the separable 2D approximation spaces are formed from the tensor product of identical 1D approximation spaces [30]. This restriction generates analyzing filters with homogeneous spectral characteristics in 2D frequency space.

When extended to three dimensions, the multiresolution analysis is constructed from a separable 3D analyzing, or "scaling", function formed from the product of three nonidentical 1D scaling functions, or two identical 1D spatial scaling functions and one different 1D temporal scaling function. This brings a much richer set of orthonormal basis vectors with which to represent 3D signals, and it produces filters that can be easily tailored to more closely match the spatial and temporal frequency characteristics of the 3D signal.

An $L_2(\Re)$ multiresolution analysis consists of a chain of closed, linear "approximation" spaces V_j and a scaling function ϕ which satisfy the following properties for all $f \in L_2(\Re)$.

1) $...V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset ...$

2) $\overline{\bigcup_{j \in Z} V_j} = L_2(\Re); \quad \bigcap_{j \in Z} V_j = \{0\}$

3) $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}; \quad j \in Z$
$f(x) \in V_j \Rightarrow f(x + n/2^j) \in V_j; \quad n \in Z$

4) The set of functions $\{2^{j/2}\phi(2^j x - n) \,|\, j \in Z, n \in Z\}$ forms an orthonormal basis for the approximation space V_j.

As presented by Mallat, the purpose of multiresolution analysis is to create a mathematical framework that facilitates the construction of a wavelet orthonormal basis for the space for all finite energy signals $L_2(\Re)$. To this end, denote the orthogonal complement of V_j in V_{j+1} by W_j where

$$V_{j+1} = V_j \oplus W_j$$

and the symbol $\oplus$ indicates the direct sum. W_j is typically referred to as the *j*th detail space, because it captures the difference in signal information between the approximation spaces V_{j+1} and V_j.

Mallat has shown that one can create a mother wavelet $\psi(x)$ such that the set of functions $\{2^{j/2}\psi(2^j x - n) | n \in Z\}$ forms an orthonormal basis for W_j. The spaces W_j, where $j \in Z$, are mutually orthogonal; thus, by the denseness property of the multiresolution analysis, the set of scaled and dilated wavelets $\{2^{j/2}\psi(2^j x - n) | j \in Z, n \in Z\}$ forms an orthonormal basis for $L_2(\Re)$. The scaling functions and the mother wavelet are related by the "two-scale" recursion relations

$$\begin{aligned} \phi(x) &= \sum_{n=-\infty}^{\infty} h_n \sqrt{2}\phi(2x-n) \\ \psi(x) &= \sum_{n=-\infty}^{\infty} g_n \sqrt{2}\phi(2x-n) \end{aligned}, \qquad (14)$$

where h_n and g_n are the coefficients of the QMF pair which is used to compute the approximation and detail projection associated with V_j and W_j from the approximation at the next higher scale V_{j+1}.

Approximation and detail signals are created by orthogonally projecting the input signal f onto the appropriate approximation or detail space. Since each space is spanned by an orthonormal basis set, the signal projection onto a given approximation or detail space at the *j*th resolution, is equivalent to the sequence of projection coefficients obtained by the inner product operations

$$a_{j,n} = \int_{-\infty}^{\infty} f(x) 2^{j/2} \phi(2^j - n) dx$$
$$d_{j,n} = \int_{-\infty}^{\infty} f(x) 2^{j/2} \psi(2^j - n) dx \quad (15)$$

where $a_{j,n}$ and $d_{j,n}$ represent the *j*th approximation and detail coefficients respectively.

Figure 5 shows the block diagram of the heterogeneous 3D wavelet decomposition.

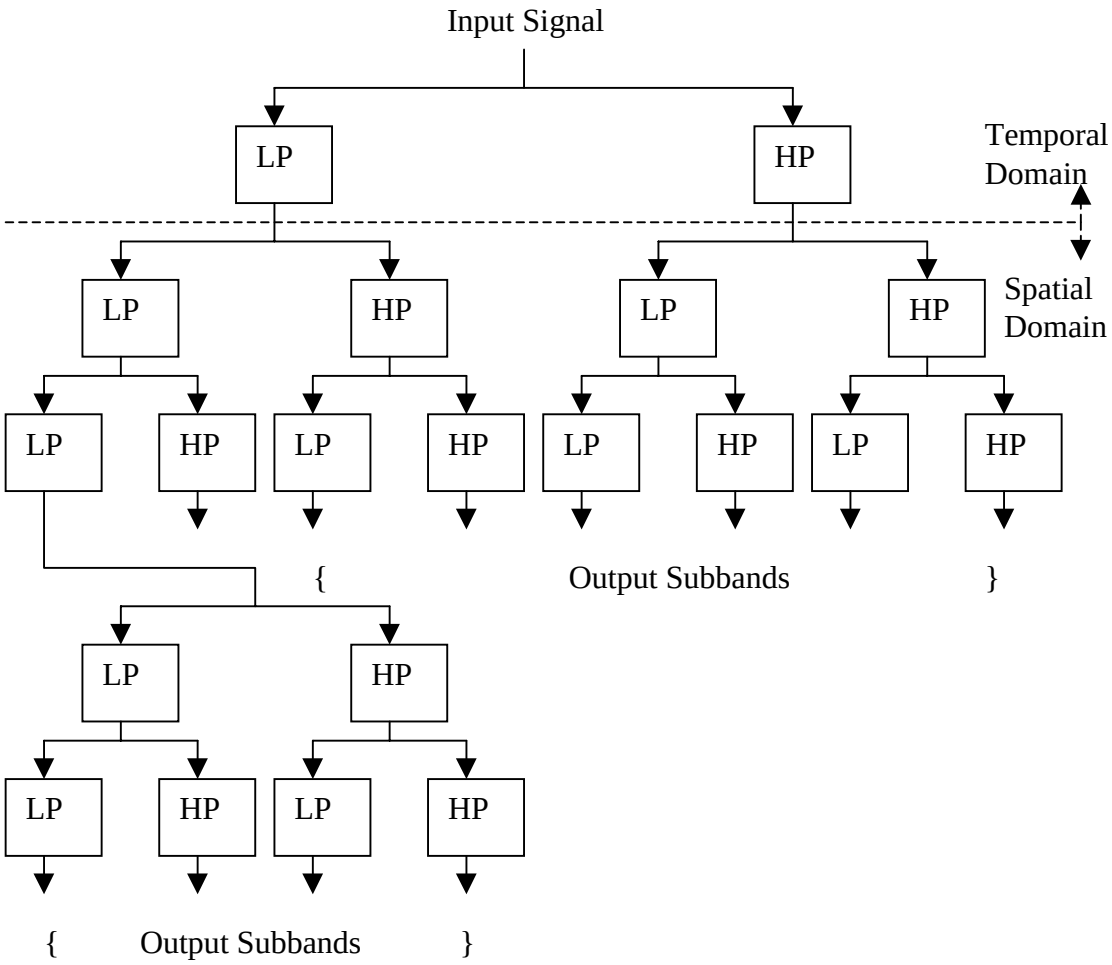

Figure 5 Heterogeneous 3D Wavelet Decomposition

In our video codec system, the wavelet transform performs decomposition of video frames into a multi-resolution subband representation. Each pair of incoming video frames is decomposed into 11 subbands as shown in **Figure 6**.

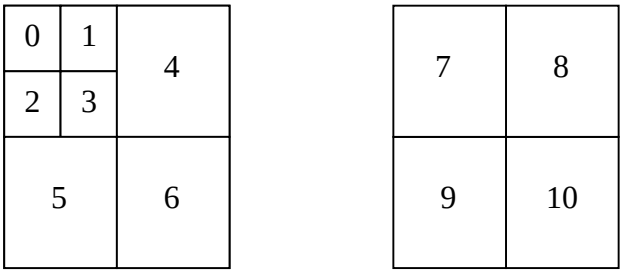

Figure 6 Subbands after 3D Wavelet Decomposition

We use the simple two-tap Haar filter, which essentially separates the signal into temporal low frequency and high frequency parts. Then, the low frequency part is decomposed for two levels with the spatial 2-D wavelet decomposition, while the high frequency part is decomposed for one level. The Antonini (7,9)-filter [21] is used here. The Haar and Antonini filters work together to achieve a satisfying balance between complexity and performance.

The coefficients for the Antonini (7,9) wavelet filter are:

For the LPF, the h_n are given by {0.0378284555, 0.0238494650, 0.1106244044, 0.3774028556, 0.852698679, 0.3774028556, 0.1106244044, 0.0238494650, 0.0378284555}.

For the HPF, the g_n are given by {0.0645388826, 0.0406894176, 0.4180922732, 0.788485616, 0.4180922732, 0.0406894176, 0.0645388826}.

3.2 Frame Counting & Motion Detection

The computation of the JND model is resource consuming for the real-time video system. Even if it is implemented with dedicated hardware, the reconstruction of the JND profile for each pair of incoming video frames is expensive and unnecessary. The Frame Counter & Motion Detector is designed to control the renew process of the JND. The frame counter is used to count the number of frames being coded. After a certain number of frames (typically 10 or 20 frames) have been coded with the current JND model, this model is refreshed with the updated signal. Our assumption is that the scene in frames remains almost the same so the original JND model becomes effective before the update.

If, however, drastic movement or scene cut happens, the scene changes greatly. So the original JND has to be refreshed right away to follow the change. This is why a motion detector is necessary.

In our system, a simple motion detection scheme is adopted at the current time and stage of development. It considers two factors relative to picture contents change, which are scene cut and drastic movement. Initially, after the 3D wavelet decomposition, the energy in the spatial-LLLL temporal-L subband (i.e. subband 0 in **Figure 6**) is calculated for each group of two frames and stored as E_{old}. It is then compared with the new generated energy E_{new} for subband 0. If their difference exceeds the threshold, we assume a scene cut has occurred. That threshold can be obtained from experiments. As for detecting drastic movement, the energy in the spatial-LL temporal-H subband (i.e. subband 7 in **Figure 6**) is also calculated after the wavelet

decomposition. If it exceeds a given threshold, a drastic motion is assumed to be happening.

With each detection event from the Frame Counter & Motion Detector, the JND model is reconstructed.

3.3 JND Model Generation

The JND provides each signal a threshold of visible distortion, below which reconstruction errors are rendered imperceptible. In this part, the spatial-temporal JND profile for each group of two frames in a video sequence and the JND profiles for subbands are generated sequentially. The details are as described in Chapter 2.

Since the generation of the spatial-temporal JND profile for each frame requires one previous frame as reference, when a video sequence is being coded, the encoder has to assign the first frame a reference frame. In our implementation, the first frame in the video sequence will use itself as reference to generate each renewed JND profile.

3.4 Perceptually Tuned Quantization

The perceptually tuned quantizer is the core of the source encoder. With the JND profiles for each subband at hand, the most important task is to allocate the available bits to certain coefficients obtained from the wavelet decomposition as efficiently as possible.

The beauty of JND is that it provides a quantitative measure of the error sensitivity threshold with spatial and frequency localization. In schemes using DCT like

MPEG-2, the DCT coefficients are quantized using a quantization table that assigns more bits to the more important lower frequency coefficients in one 8 by 8 block. Although such a table is designed according to the HVS response based on psychovisual theory and experiments, its wide use all over the whole picture brings shortcomings, because different parts in a picture can have different visual importance at different scenes. On the other hand of the quantization table based on JND, which can adapt itself to local scenes results in less perceptual distortion.

At the beginning of our video coding, a parameter is requested for by the encoder. It is the object global distortion index Δ_G that will control the overall video quality during the coding procedure. The same symbol Δ_G is used here as the symbol we used in Chapter 2 to indicate our new human perceptual distortion measure based on JND, since these two have the same psychological meaning. Ideally, the coding procedure controlled by the object global distortion index $\Delta_G = d$ will produce a compressed video sequence whose average perceptual distortion index Δ_G value is $d' = f(d)$. This powerful functionality to control the compressed video quality ahead of time makes our scheme so distinct from conventional video coding systems. In addition the coding distortion is perceptually evenly distributed across the whole image, as our scheme will show.

3.4.1 Uniform Quantization

A mid-rising uniform quantizer is used as our basic quantizer due to its simplicity of error analysis and its sound performance under certain conditions for

optimality [24]. First, a global distortion index Δ_G is given for the quantization procedure. It usually ranges from 1.0 to 10.0, where 1.0 stands for just noticeable distortion. Second, each subband is partitioned into non-overlapped blocks ($r_{ij}(k,l)$). These blocks are set up with the size of 8x8 or 16x16. For each block $r_{ij}(k,l)$, the step size of the mid-rising uniform quantizer is maximized under the condition that quantization error energy is less than or equal to the MND energy in this block that has the distortion index $\delta(i,j)$ equal to Δ_G, i.e.,

$$\max \arg_{\tau_{ij}(k,l)} \left| \sum_{(x,y)\in r_{ij}(k,l)} MND^2(x,y) \geq \sum_{(x,y)\in r_{ij}(k,l),|\, w(x,y)\,|<\tau_{ij}(k,l)} w(x,y)^2 + \sum_{(x,y)\in r_{ij}(k,l),|\, w(x,y)\,|\geq\tau_{ij}(k,l)} [w(x,y)-\hat{w}(x,y)]^2 \right. \tag{16}$$

where the energy of MND defined as (11), $w(x,y)$ is the wavelet coefficient, $\hat{w}(x,y)$ is the quantized wavelet coefficient, $\tau_{ij}(k,l)$ is the quantization step size of $r_{ij}(k,l)$.

Therefore, one quantization table that leads to the uniform error energy distribution over all subbands is set up for each subband. It is transmitted in the header of the bit stream for this subband. If after quantization the proportion of zero signals in a block is larger than 7/8 or 15/16, this block is assumed to be unimportant. Its stepsize is recorded as 0, and in the decoder, all values in this block are recovered as 0's. So the coefficients in such an unimportant block need not be transmitted.

When the bandwidth is dynamically assigned to this source encoder, Δ_G can be kept constant to maintain the video quality at the same level. If the bandwidth is fixed, i.e., the bit rate is limited for this source encoder, a bunch of Δ_G values should be tried and the corresponding bit rates are compared with the available channel capacity. One Δ_G value that provides proper bit rate will be chosen finally. And this procedure of refreshing of the quantization table and Δ_G value choice is repeated when the Frame Counter reaches a certain number or a drastic movement happens resulting to the need to update the JND model.

3.4.2 Lloyd-Max Quantizer

An optimum mean square error quantizer is also tried in our scheme. This quantizer minimizes the mean square error for a given number of quantization levels. For a random variable *u*, the reconstruction levels a_k are calculated with [31]

$$a_k = \frac{\int_{t_k}^{t_{k+1}} u p_u(u) du}{\int_{t_k}^{t_{k+1}} p_u(u) du} \qquad (17)$$

where $p_u(u)$ is the continuous probability density function of the random variable *u*, and t_k is the transition level

$$t_k = \frac{(a_k + a_{k+1})}{2}. \qquad (18)$$

To design an appropriate quantizer, the distribution of the subband samples needs to be known. It has been suggested that the probability density function of the

subband values and their prediction errors is Laplacian [32]. Although more accurate probability distribution of the subband values, i.e., generalized Gaussian distribution, has been suggested and its shape parameters estimation method is explored [33], the resulting overhead for the calculation and transmission of the parameters is too large. So we still use the Laplacian distribution as a reasonable approximation.

In the procedure of quantization, first, a global object distortion index Δ_G is selected. Second, the variance of the wavelet coefficients for each subband is calculated. And all the coefficients are normalized. Third, each subband is partitioned into blocks ($r_{ij}(k,l)$) with size of 8 by 8 or 16 by 16. For each block $r_{ij}(k,l)$, the levels of Lloyd-Max quantizer are minimized under the condition that the quantization error energy is less than or equal to the MND energy in this block that has the distortion index $\delta(i,j)$ equal to Δ_G, i.e.,

$$\min_{n} \arg \left| \sum_{(x,y)\in r_{ij}(k,l)} MND^2(x,y) \geq \sum_{(x,y)\in r_{ij}(k,l),|\, w(x,y)\,|< t_1^{(n)}} w(x,y)^2 + \sum_{(x,y)\in r_{ij}(k,l), t_{m+1}^{(n)} > |\, w(x,y)\,| \geq t_m^{(n)}, m=1,2,\ldots,N} [w(x,y) - a_m^{(n)}(x,y)]^2 \right. \tag{19}$$

where the energy of MND is defined as in equation (11), $w(x,y)$ is the wavelet coefficient, n is the number of quantizer levels with $n=3,5,\ldots 2N+1$, $2N+1$ is the maximal number of quantization levels, m indicates the interval in which $w(x,y)$ is located, and $a_m^{(n)}(x,y)$ is the quantized wavelet coefficient for a quantizer with n levels. Here a look-up table is set up for the $t_m^{(n)}$ and $a_m^{(n)}$, since the wavelet coefficients for

quantization have been normalized. The index of levels for the optimum quantizer is transmitted in the header of the bit stream of this subband.

3.5 Arithmetic Coding and Slicing

The arithmetic coding scheme developed by Witten, Neal and Cleary [25] is widely used in video codecs due to its superior performance. It easily accommodates adaptive models and is computationally very efficient. The arithmetic coding scheme provides efficient compression, however the decoding result of one coefficient depends on the decoding result of the previous one because of the adaptive coding procedure employed.

In the environment of a noisy channel, in order to prevent decoding errors from spreading, a slicing algorithm is derived to truncate the whole subband into short bit streams before arithmetic coding. The idea is to make each such small bit stream carry the same amount of "distortion sensitivity ". If we want to segment the subband S_l into *I* short bit streams, we can define *I* sets G_i of the point *(x,y)*, such that for each set G_i $(i = 1,...,I): S_l = \bigcup_i G_i$ and

$$G_i = \left\{ (x,y) \left| \begin{array}{l} \sum_{(x,y) \in G_i} \frac{1}{JND^2(x,y)} = \frac{1}{I} \sum_{(x,y) \in S_l} \frac{1}{JND^2(x,y)}, \\ (x,y) \in S_l, (x,y) \notin G_j, j \neq i \end{array} \right. \right\} \quad (20)$$

So every time when such a short bit stream is being encoded, a new adaptive statistical model is set up for it. Before the arithmetic encoded output data of these short bit streams are merged into one bit stream, header and trailer symbols are added so they

can be selected out from the received bit stream at the decoder side. The slicing information is transmitted along with the data stream.

3.6 Perceptual Channel Coding

In our practical video transmission system over a satellite channel, rate compatible punctured convolutional (RCPC) codes [27] are adopted. The advantage of using RCPC codes is that the high rate codes are embedded into the lower rate codes of the family and the same Viterbi decoder can be used for all codes of a family. Reed-Solomon code and Ramsay interleaver plus RCPC is used to protect the data from spatial LLLL temporal L subband. Cyclic redundancy check (CRC) codes are combined with RCPC for other less significant subbands to assure acceptable video quality even under bad channel conditions. [34]

In order to optimize the overall subjective video quality at a reasonable coding cost, a rate allocation scheme based on JND distortion is proposed. We define the average JND distortion of subband l (l=0,…,10) as follows:

$$D_l = \frac{1}{H_l W_l} \sum_{(x,y) \in S_l} JND(x,y)^2 \qquad (21)$$

where S_l is the set of pixels of subband l, H_l and W_l are height and weight of it separately. D_l is an indication of the robustness of S_l to errors. The larger D_l is, the more robust it is to errors, the higher coding rate we choose. Table 1 shows the D_l for the video sequence " Calendar-Train".

L	D_l	l	D_l	L	D_l
0	4.5	4	7.5	8	27.6
1	7.6	5	7.5	9	27.6
2	7.6	6	8.7	10	42.8
3	8.8	7	7.4		

Table 1 Average Distortion D_l for Each Subband

From simulations we can see that D_l divides S_l into four categories, $\{S_0\}$, $\{S_1, S_2, S_3, S_4, S_5, S_7\}$, $\{S_8, S_9\}$, $\{S_{10}\}$, which is intuitive for RCPC unequal error protection. According to their different importance, the subbands are assigned different RCPC coding rates correspondingly.

3.7 Video Decoder and Error Concealment

In the video decoder, the functionality of most modules is straightforward. In the error concealment, when errors are detected via CRC decoding in the slices in subband S_l, these slices are discarded. If no error is detected, but there are some errors in the received slice, the arithmetic decoder can detect some conflicts during decoding sometimes, therefore can find some errors and discard this slice. In this case, coefficients in this slice are retrieved from its DPCM reference if it belongs to the spatio-LLLL temporal-L subband. And the coefficients are set to zero's if the slice belongs to other subbands of higher frequency. The error effect is trivial and will be confined within the slice. In the worst case where errors are not detected, they will not spread to the whole subband due to slicing.

In the MPEG system, decoding error is also confined within slices. But the corruption of data will destroy the whole slice thoroughly. In our wavelet-based system, even if slices in one subband are corrupted, slices in other subbands will contribute to the same area of the frame. And more flexible concealment schemes can be implemented to improve the quality.

Chapter 4 Simulation Results

A C/C++ program was written to simulate the performance of our video codec system. For simulation of a video system, C code is better than Matlab due to the large volume of data and the consideration of speed. It is easy to be transplanted between PCs and Sun Workstations to match the requirements of other application software.

After the input video is decomposed into 11 subbands, each of them is treated as one object variable of class Subband. So they can be operated separately for fine adjustment of parameters if necessary. The class Subband provides corresponding methods to deal with the data in subbands, such as:

```
void Quantize( );              //uniform quantization

void adjustStepSize( );        //stepsize adjustment for each block from JND profile

void LQuantize( );             //optimum quantization for Laplacian distribution

void adjustLapQIndex( );       //index of levels adjustment for each block from JND

void AC_put_Qvalue( );         //arithmetic encoding

void AC_get_Qvalue( );         //arithmetic decoding
```

We use the video sequences of “Calendar-Train” and “Claire” downloaded from http://www.ipl.rpi.edu. They are resized to 512x384 and 352x288 for easy manipulation.

4.1 Spatio-temporal JND Profiles and JND Profiles for Subbands

Original frame #1 of “Calendar-Train” and its JND subband profiles are shown in **Figure 7** and **Figure 8**. This JND model is calculated from the signal of this frame and its reference frame. It reflects the visual just-noticeable distortion for video at a speed of 30 frames/sec.

Figure 7 Frame #1 of "Calendar-Train"

From the JND subband profile we can tell the difference of the perceptual importance of different subbands. The JND subband profile for spatial LLLL temporal L subband (subband 0) has smaller values of JND (displayed as smaller gray scale values in **Figure 8**) than other subbands. It means that the threshold of just noticeable

distortion is lower, i.e., a small distortion in this subband will be captured by the human visual system while it won't be noticed if it happens in other subbands. We can also see the perceptual importance of different parts in this frame. For example, the calendar part, while comprised of many fine lines and digits, has smaller values of JND, therefore the distortion there is more sensible to human eyes. However, human eyes have more tolerance of the distortion at the moving train, whose body is almost evenly dark.

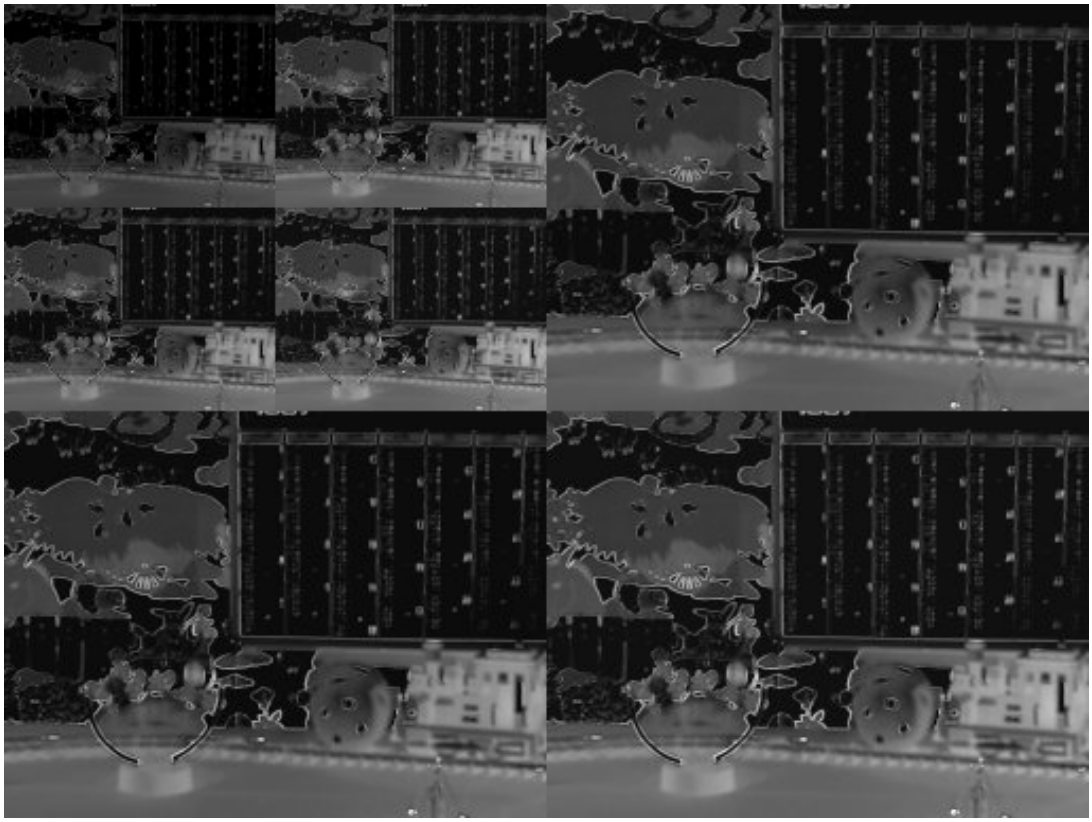

Figure 8 JND Subband Profile for Subband 0 to 6

4.2 Human Perceptual Distortion Measure vs. PSNR

Figure 9 Decoded Frame of "Claire", Δ_G =2.38, PSNR=30.80dB

Figure 10 Decoded Frame of "Claire", Δ_G =3.07, PSNR=30.15dB

Figure 9 and **Figure 10** show that our distortion measure (equation (12)) is better than PSNR in the sense that it reflects the subjective visual quality of image/video better. **Figure 9** and **Figure 10** show frame #1 in the decoded sequence of "Claire". The PSNR of these two frames are almost the same, but the Δ_G values indicate that the distortion of **Figure 9** is smaller than that of **Figure 10** as we can tell from direct observation (e.g. shoulder, hair and cheek).

4.3 Video Transmission over Satellite Channels

In Chapter 3 we discussed the design of a channel coding scheme using RCPC code, Reed-Solomn Coding, CRC Code and Ramsay interleaver. OQPSK modulation is in our simulation. The human vision model based rate allocation for RCPC is also described. The average JND energy for each subband is calculated as shown in **Table 2**. Because the difference between S_{10} and S_8, S_9 is relatively small**,** we finally assign 11 subbands into three error protection groups: $\{S_0\}$, $\{S_1, S_2, S_3, S_4, S_5, S_7\}$, $\{S_8, S_9, S_{10}\}$.

Index	4	5	6	7	8	9
Rate	8/18	8/16	8/14	8/12	8/10	8/9

Table 2 Rate Index of RCPC

Table 2 shows the coding rate index. The sequence of "Calendar-Train" is coded and transmitted over AWGN channel at different SNR**. Figure 11** and **Figure 12** show the distortion index of the first 10 decoded frames with different protection

schemes at different SNR (In **Figure 11** and **Figure 12**, the legend 3dB (4,7,8) means that the E_b / N_o is 3dB and we use channel coding rate 8/18 for subband 0, rate 8/12 for subband 1 to 7, and rate 8/10 for subband 8 to 10).

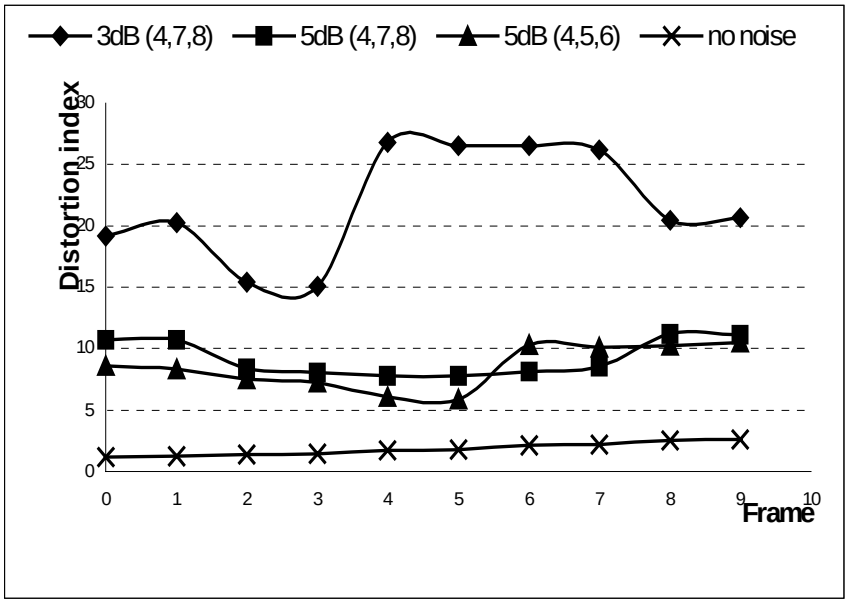

Figure 11 Distortion of the Decoded Frames over Noisy Channel with Object Distortion Index=1

In **Figure 11** the original frames are encoded with the object distortion measure Δ_G =1, which means the compression brings just noticeable distortion in the pictures. Even if the channel is ideal, the distortion Δ_G goes larger gradually (Δ_G =1.19 for Frame 0 and 1.77 for Frame 5) in the following frames. The reason is that the JND model was initially constructed from the first two frames and is not renewed for the subsequent frames, which brings bias in the encoding. The larger the frame number is, the more the bias accumulates.

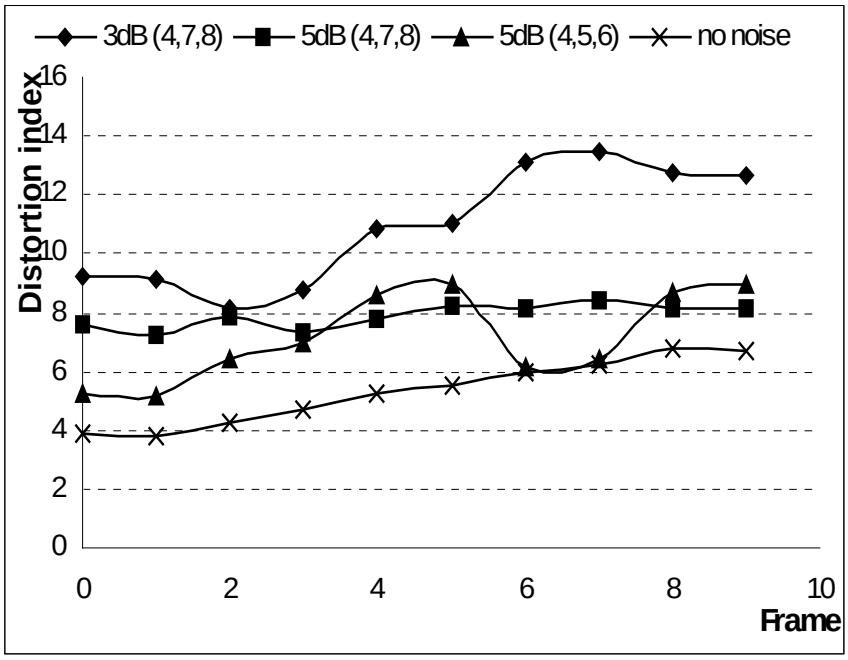

Figure 12 Distortion of the Decoded Frames over Noisy Channel with Object Distortion Index=5

In **Figure 12** the original frames are encoded with the object distortion measure Δ_G =5, which means the perceptual distortion is 5 times of the just noticeable distortion. Intuitively, one would accept that the distortion in **Figure 11** should be less than that in **Figure 12**; however this is not correct. The reason is that the slices generated by the source encoder are longer when Δ_G =1 (finer step size is chosen and more data is transmitted), and therefore the probability that a slice is corrupted by errors increases as the number of bits in this slice increases at the same bit error rate (BER). So if a video service requires better quality, the corresponding better channel protection scheme should be chosen.

Figure 13 shows frame No. 3 in a recovered sequence of "Calendar-Train". Some areas corrupted by the channel noise can be observed.

Figure 13 Decoded Frame #3 of "Calendar-Train" with $E_b / N_o = 3dB$ (4,7, 8) Protection

4.4 Comparison with MPEG

The performance of our JND based video codec is compared with the MPEG-1 system. Due to the fidelity of our JND distortion index Δ_G, it will be used as the major performance indicator. Because the focus of our system is the subjective quality, more powerful compression schemes (e.g. zero-tree, motion estimation and run-length coding) are not applied in our system. Therefore, our system will be compared with the performance of I frames of MPEG.

There are two MPEG-1 encoders used in our experiments. The performance of these two encoders, say, MPEG A and MPEG B, is a little bit different. MPEG-A has a

higher compression ratio and MPEG-B has a higher PSNR at the same quantization scale. But basically, their output bit streams all follow the MPEG-1 standard and the key features of MPEG like DCT, zig-zag scanning, Huffman coding and run-length coding are implemented in the encoder.

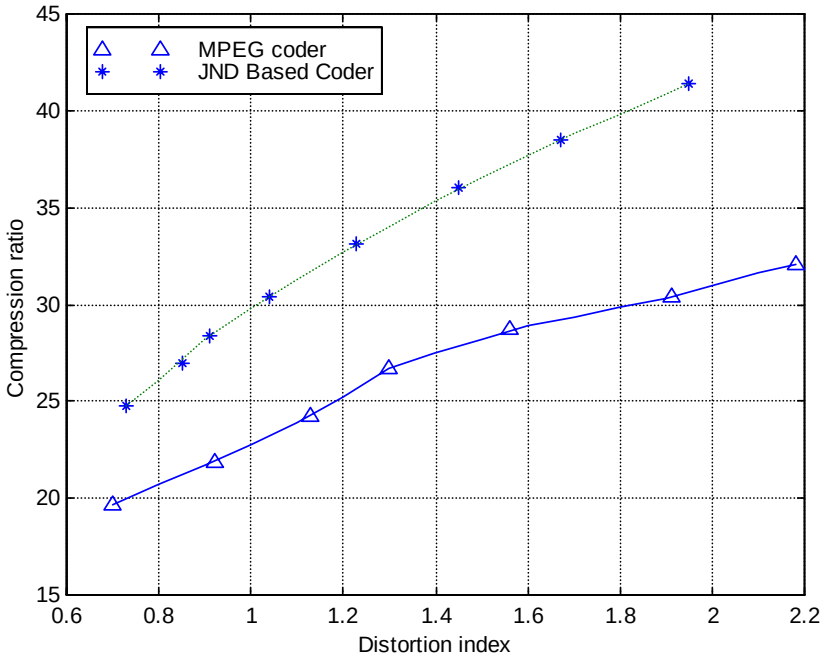

Figure 14 Performance Comparison between the MPEG Coder (MPEG A) and the JND Based Coder

In **Figure 14** the original sequence "Claire" is encoded according to MPEG-1 at different scales. It gives different compression ratios for the I frames. The distortion indexes of these I frames are calculated using spatial-temporal JND profiles. At the same perceptual distortion level, the JND based encoder provides higher compression ratio than the MPEG-1 encoder for I frames. It also turns out that the subjective video quality of JND based compression is better than the subjective quality of I frames from MPEG-1 at the same compression ratio.

Figure 15 "Claire" from the JND based coder, Δ_G =0.85, PSNR=36.3dB, compression ratio=27.0:1

Figure 16 "Claire" I frame from the MPEG-1 coder, Δ_G =1.3, PSNR=37.5dB, compression ratio=26.7:1

Generally, the PSNR of decompressed frames from the MPEG-1 coder is 1-2 dB larger than that from the JND based coder at the same compression ratio. This is a natural result for the JND based compression scheme. Since with the consideration of JND profile, the irrelevant information for perception is removed, it renders more compression at the same perceptual quality. Ant it renders more numerical error which means less PSNR at the same compression ratio, while keeping the subjective quality superior.

Figure 15 and **Figure 16** show the decompressed "Claire" of the JND based coder and the MPEG-1 coder (MPEG A). The compression ratio is almost the same. MPEG-1 provides better PSNR and the JND based coder provides better Δ_G. From direct observation we can tell that the subjective quality of **Figure 15** is better than that of **Figure 16**.

Figure 17 and **Figure 18** show the decompressed "Claire" of the JND base coder and MPEG-1 coder (MPEG B). At a higher compression ratio (35:1), their distortion can be seen clearly. But it is proper for us to describe the characteristics of our JND based encoder. The MPEG-1 encoder keeps the quality of the contour of her body, the texture of her hair, and even the glare on her ear ring, while it brings smear to the lady's face, which is the focus of a viewer. The JND based coder assigns the distortion mainly to the moving edges of the body, the moving mouth, and the texture of hair, while it makes the face seem more comfortable. That is the objective of our perceptual optimal coding.

Figure 17 "Claire" from the JND based coder, PSNR=34.5dB, CR is 35.0:1

Figure 18 "Claire" from the MPEG-1 coder, PSNR=37.6dB, CR is 35.5:1

The video sequence “Calendar-Train” is more difficult to compress due to the complex contents of this sequence. The performance of the JND based coder is very close to that of the MPEG-1 coder on I frames. **Figure 19** and **Figure 20** show the result.

Figure 19 “Calendar-Train” from the JND based encoder, Δ_G =3.03, PSNR=32.6dB, compression ratio 8.93:1

Figure 20 “Calendar-Train” from the MPEG-1 encoder (quantization scale=7), Δ_G =3.11, PSNR=34.0dB, compression ratio 9.06:1

4.5 Quantization Schemes Comparison

The performance of uniform quantization and mixed optimum quantization for Laplacian distribution is shown in **Figure 21**. Here the mixed optimum quantizer implements optimum quantization for subbands 4,5,6,7 and uniform quantization for subbands 0,1,2,3,8,9,10. This scheme is based on experimental adjustment for getting the best performance. But as shown in **Figure 21**, the optimum quantization is not better than the uniform quantization, which is contrary to our expectation. It can be explained by the observation that the application of entropy coding following quantization diminishes the benefits of bit-constrained optimal quantization. Further research leads to the topic of entropy-constrained quantization. In our other simulation experiments, actually uniform quantization is adopted.

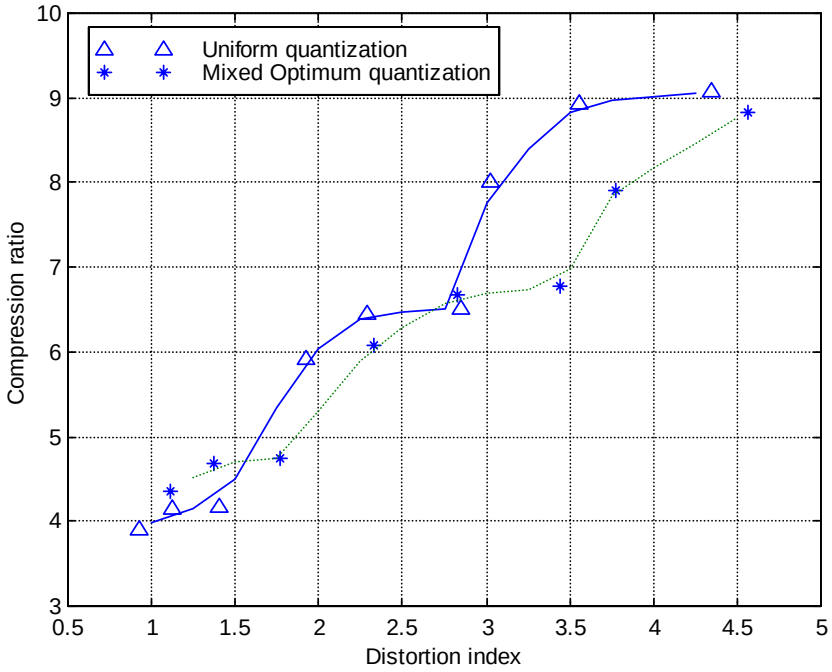

Figure 21 Performance Comparison between Uniform Quantization and Mixed Optimum Quantization

Chapter 5 CONDITIONAL ACCESS WITH ARITHMETIC CODING

The conditional access sub-system is being deployed extensively to follow the rapid expansion of the commercial broadcasting industry. It is used to control which customer can get particular program services. Particular programs are only accessible to customers who have satisfied prepayment required.

Current digital broadcasting systems, e.g., DVB, uses a scrambler unit to implement encryption [35]. The bit stream out of the MPEG-2 encoder/multiplexer is fed into the scrambler unit along with the scrambler key, called the Control Word (CW). The CWs are sent to the receiver in encrypted form as an entitlement control message (ECM). The CA subsystem in the receiver will decrypt the control word only when authorized to do so; that authority is sent to the receiver in the form of an entitlement management message (EMM). The subscribers' information necessary for authority is maintained in the RSMS-TAM (Resource and Subscriber Management System – Transmitter Access Management) unit. Scrambling is performed with an ASIC due to the high bit rate.

The benefits of applying data compression prior to transmission are well known and widely exploited. In situations where both data compression and data encryption are desired, such as in commercial video broadcasting, the system will be quite simplified if we combine these two techniques efficiently. The idea of using data compression schemes for encryption dates back to 13^{th} century. After the suggestion of Huffman coding, Fraenkel and Klein [36] have shown that the problem of finding the encoding rule given both a sample of the source and the corresponding sample of the encoded file is NP-complete. Gillman et al. [37] have discussed the information-theoretic

impossibility of cryptanalyzing a Huffman code. Witten and Cleary [38] initially suggest the idea of combining data compression together with encryption using adaptive arithmetic coding. In their scheme, the key of the cryptosystem is adopted as the starting point of the model, i.e., taking the encryption key as a specific starting probability distribution for the symbols in the adaptive arithmetic coding model. Boyd [40] gives cryptanalysis for this scheme with the acclamation that there exists an order preserving property that allows a known plaintext attack in the case of a binary alphabet.

Since entropy coding is the indispensable part in all practical video compression systems, in this chapter, we investigate the problem of realizing scrambling by utilizing the cryptographic properties of arithmetic coding. A new scheme of low cost and reliable conditional access is the result.

5.1 Introduction to Arithmetic Coding

Theoretically, arithmetic coding encodes a message as a number in the unit interval [0,1]. Unlike most schemes, including Huffman Coding, there is no specific code word defined for any specific symbol, but how each symbol is encoded depends, in general, on the previous symbols of the string. For the source sequence x^n, let us denote its probability mass function as $p(x^n)$ and its cumulative distribution function as $F(x^n)$. We can use a number in the interval $(F(x^n)\text{-}p(x^n), F(x^n)]$ as the code for x^n. For example, expressing $F(x^n)$ to an accuracy of $\lceil -\log p(x^n) \rceil$ will give us a uniquely decodable code for the source [39]. However this is equal to the entropy of the message encoded, so that by Shannon's theory we have achieved the theoretical bound. There are two reasons why the theoretical bound cannot usually be achieved in practice [40]:

(1) For a message of unbounded length, arithmetic of unbounded precision is required for maintaining the value of the current interval. In practice this is overcome by scaling procedures which add to the average encoded word length by decreasing the size of the actual interval calculated.

(2) The decoding is not unique unless the length of the message is known, since a single point may represent any interval of which it is a member. Without knowledge of the message length, decoding can proceed indefinitely. This may be overcome in practice either by sending a length indicator or by using an end-of-message symbol. Both of these add overheads to the encoding.

In the sequential encoding procedure, each symbol in the message is assigned a half-open subinterval of the unit interval of length equal to its probability. We call this the coding interval for that symbol. As encoding proceeds a nesting of subintervals is defined. Each successive subinterval is defined by reducing the previous subinterval in proportion to the length of the current symbol's coding interval. This process continues until all symbols have been encoded.

In some cases that a precise probability distribution for the source is unavailable, a dynamical procedure of updating the symbol frequency model, which renders removal of the redundancy quite efficiently, will be used to adapt to the source. This is the basic idea of adaptive arithmetic coding.

5.2 Dependency of Arithmetic Coding

In arithmetic coding, the source sequence x^n is represented by its cumulative distribution function $F(x^n)$. The encoding output of the current symbol depends on the

previous one due to the tree-structured encoding procedure. So basically, it is a polyalphabetic cryptosystem. With the knowledge of a source symbol frequency model, the decoder restores the source symbol one by one with successive dependency. We will show this dependency with the following examples. Due to the popularity of the arithmetic coding implementation in C of Witten *et al.* [25], we will base our discussion on it. In this implementation, a binary bit stream is obtained as the encoding output.

Example 1:

We assume the source symbol frequency model is $\begin{bmatrix} 0 & 1 & 2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$, which indicates that Prob(symbol=0) is 0.3, Prob(symbol=1) is 0.3, Prob(symbol=2) is 0.4. The source sequence for AC encoding is

1 1 1 2 0 2 0 1 0 1 0 1 2 2 1 0 0 2 1 0

The encoded binary bit stream is

10010000 10010001 01000110 11100100 11 (34 bits)

Note: Here the frequency model doesn't match the distribution of source sequence. But it only decreases the compression ratio and doesn't change the characteristics of AC.

If the third symbol 1 is replaced by 2 and 0 respectively, the encoder will give out the corresponding binary bit stream

10001000 00011000 10011011 01000101 1 (33 bits)

and

10010111 01111010 11010100 01100011 00 (34 bits)

Only a few bits in the front are the same in these three cases, and the numbers of output bits are also different, i.e., the change of the previous symbol changes the encoding of the following symbols completely.

The loss of source symbol also diversifies the encoding of the following symbols. If the sequence of source symbols is

1 1 2 0 2 0 1 0 1 0 1 2 2 1 0 0 2 1 0

(with the first symbol lost), it's encoded to

10001100 10010000 10000101 01101111 (32 bits)

Example 2:

We assume the same symbol frequency model as that in the previous case. The sequence of symbols for encoding is

1 2 2 0 2 0 1 0 1 0 1 2 2 1 0 0 2 1 0 1

The encoding output is

01110000 01010011 10110001 11110010 01 (34 bits)

Suppose there is one bit (7^{th} bit) error in the bit stream, which turns it into

01110010 01010011 10110001 11110010 01.

Then the decoded symbol sequence is

1 2 2 0 0 1 0 0 2 2 1 0 2 0 2 2 2 0 2 0

The symbols after the 3^{rd} one are totally different from the original symbols.

If the decoder erroneously locates the start bit position in the bit stream, the following symbols cannot be decoded correctly also, since in both cases the decoding path in the tree structure is misled.

Input symbol sequence:

1 2 2 0 2 0 1 0 1 0 1 2 2 1 0 0 2 1 0 1

Output bit stream after encoding:

01110000 01010011 10110001 11110010 01 (34 bits)

1 bit left shift:

11100000 10100111 01100011 11100100 1

Output symbol sequence after decoding:

0 1 1 0 2 0 1 2 2 2 2 1 1 2 0 1 2 1 2 1

1 bit right shift:

00111000 00101001 11011000 11111001 001

Output symbol sequence after decoding:

2 1 1 2 0 2 0 2 1 2 0 0 1 0 2 1 2 1 1 2

From these examples we can see that even with a fixed frequency model, the correctness of encoding/decoding of the previous symbol dominantly decides the correctness of the following one.

5.3 Conditional Access with Arithmetic Coding

Because the precise location of the start bit of an AC encoded bit stream is uniquely important, we propose a conditional access sub-system based on this property and we will basically identify the start bit locations of the bit streams to our scrambler key κ.

First, the data frame which will be transmitted is broken into slices $\{\mathbf{s_1}, \mathbf{s_2}, \ldots \mathbf{s_M}\}$. The locations of breaking points are decided by a randomly generated vector **v**. This vector **v** is updated after a short time interval. Each slice is AC encoded respectively into bit stream $\{\mathbf{b_1}, \mathbf{b_2}, \ldots \mathbf{b_M}\}$. Then, these bit streams are concatenated into one bit stream $\mathbf{b_{total}}$. We assign function $l(\mathbf{b})$ to represent the length of bit stream **b**.

So the value $\sum_{k=1}^{i-1} l(\mathbf{b_k})$ determines the start bit positions of $\mathbf{b_i}$, i=1, 2, …M, in $\mathbf{b_{total}}$. Only the values of $l(\mathbf{b_i})$ (i.e. the scrambler key κ) are encrypted into ECM using any available encryption algorithm, and inserted into the header of this data frame and transmitted. This is shown in **Figure 22**.

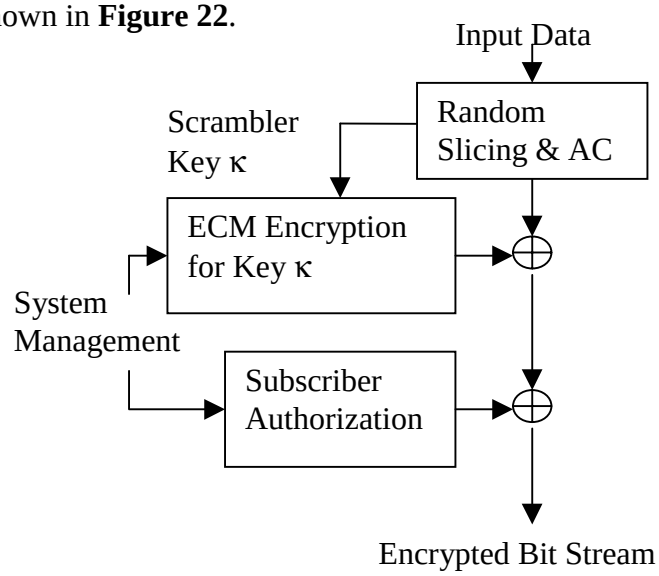

Figure 22 Arithmetic Coding Based Conditional Access Sub-system

As an example of its application, let us implement this scheme in a video broadcasting system adopting subband coding. The system is shown in **Figure 23**. The input video is decomposed into several subbands using wavelets. The wavelet coefficients in the lowest frequency subband (LFS) are encoded with DPCM before quantization. Since this subband contains the most important information, the encryption of this part is enough for video broadcasting (only the sparse edges and impulses in the high frequency subbands have no meaning to the program viewers). So LFS is fed into our CA sub-system. And the other subbands of high frequency (HFS) are simply sliced at fixed locations and arithmetic encoded. Slicing here has two functions: (1) randomly breaking the LFS into slices to generate the scrambler key κ;

(2) determinedly slicing the HFS and doing arithmetic coding for each slice to confine the propagation of error due to the noisy channel within short data units, i.e., slices.

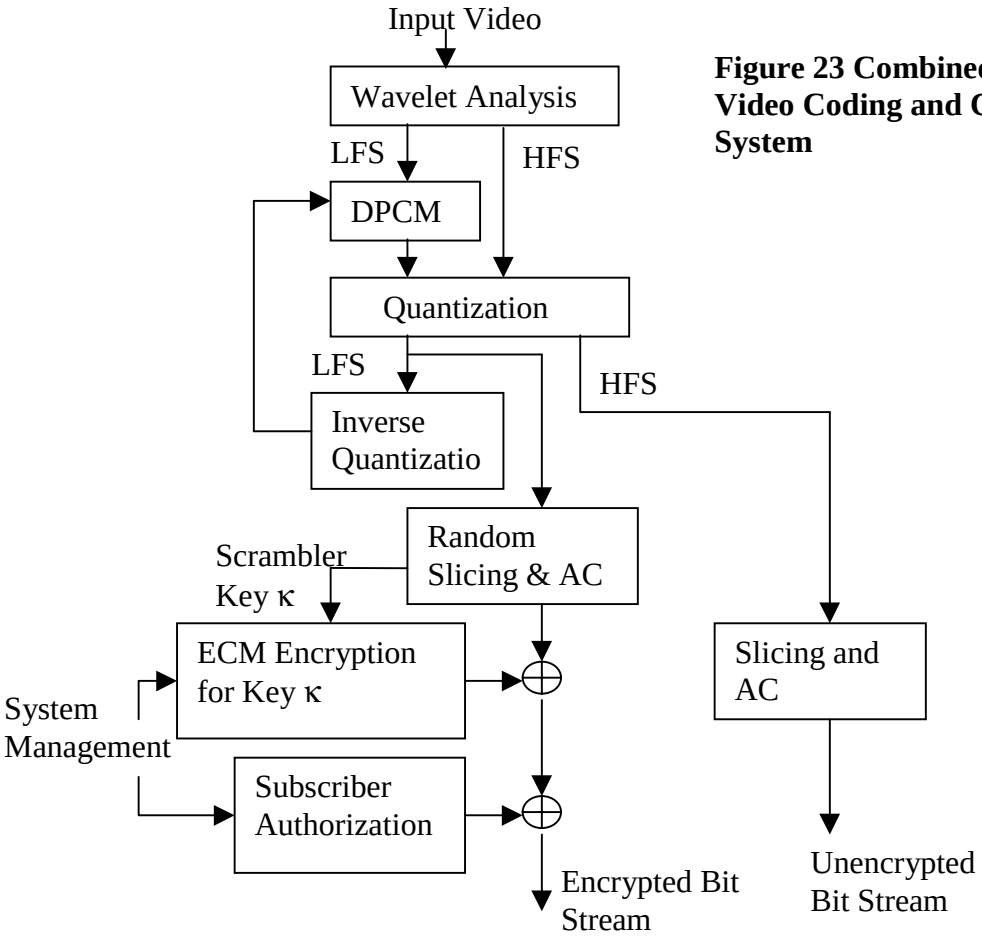

Figure 23 Combined Video Coding and CA System

For the illegal user, cryptanalysis of such a bit stream will be almost computationally impossible in the broadcasting case, even if the only thing unknown to him is the $l(\mathbf{b}_i)$ values. Let's assume that the frame size as 352x288. If each frame is wavelet decomposed in 2 levels, the most important subband LLLL (LFS), which actually needs encryption, has the size of 88x72. These coefficients are quantized and organized into 8x8 non-overlapped blocks. So we have 11x9 blocks totally. Suppose these 99 blocks are broken into 11 slices for AC randomly as shown in **Figure 24**.

There will be $\binom{98}{10}=1.4\times10^{13}$ schemes to choose. Only if the illegal user decides these slice lengths in the procedure of decoding, he can know when he has got enough symbols for the current slice, and he should stop the current AC decoding process and start a new procedure for the next slice. But the large amount of possibilities in deciding slice locations eliminates the feasibility of the method of "making mistakes and trying", i.e., the key space is large enough to stop the real-time decryption of the data in commercial broadcasting.

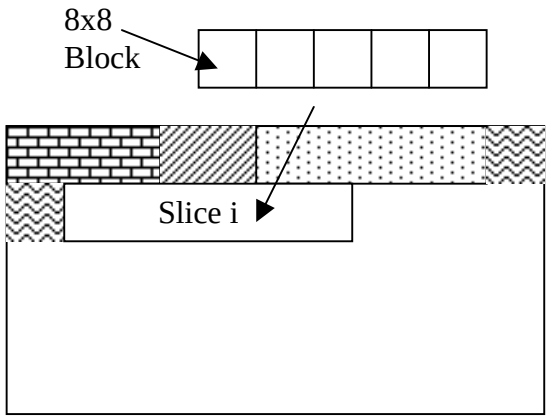

Figure 24 Slicing of Subband

The real conditional access scheme will certainly be made more complicated for the illegal user. First and most important, an adaptive source symbol frequency model will be used. No fixed source symbol frequency model is available and the frequency model is dynamically updated along with the coding procedure. Thus the AC encoding / decoding of the symbol depends more tightly on the previous ones. Even if the illegal user gets the correct start position of $\mathbf{b_i}$, i=1,2,…,M, he cannot decode it without knowing the **NoS** (number of symbols), which is vital for setting up the adaptive model and variable in different slices. And this **NoS** will also be encrypted and transmitted in the frame header. For the wavelet coefficients in the video coding system, the data

range differs frame by frame, so the illegal user cannot know it a priori. Example 3 shows the effect of **NoS**.

Example 3:

The symbol sequence for adaptive AC encoding is

1 2 2 0 2 0 1 0 1 0 1 2 2 1 0 0 2 1 0 1

If we adaptively encode with **NoS** of 3, and decode with **NoS** of 4, the decoded output symbols are

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

which turns out to be a total loss.

Second, the start position of $\mathbf{b_1}$ in $\mathbf{b_{total}}$ is variable. Stuffing bits with random length is added ahead of it, and the length of stuffing bits is encrypted along with the scrambler key κ.

The third trick played in a practical CA sub-system is that some slices will be picked out randomly for secondary encryption. The mingling of encrypted and unencrypted slices brings more difficulties to the breaking of the arithmetic code.

Fourth, because of the low computation load for scrambling, we can adapt more complicated encryption algorithms for scrambler keys at will and alternate them at different time intervals (with index given before the change).

Taking advantage of the cryptographic properties of arithmetic coding, we considerably decrease the amount of data that need to be processed for scrambling. To compare with the current CA sub-systems, we list the advantages of our scheme as follows: With relative small amount of information in the frame headers for encryption, the computational load of the encryption unit is light. A low cost processor can afford

the task. And the FIFO, timing unit, transfer stream mux/demux unit in the traditional conditional access system can be simplified considerably. The encryption algorithm used to protect the scrambler key information is independent from this scheme. So this system is flexible and more sophisticated algorithms can be easily designed and implemented. Furthermore, since arithmetic coding can be looked as the existing module for compression in a digital video broadcasting system, and breaking the data into slices is naturally necessary to constrain the propagation of error from noisy channels, there is no requirement for the cost of extra devices.

5.4 Summary

Arithmetic coding is a powerful entropy coding scheme for data compression. It is a promising replacement to Huffman coding in video broadcasting systems because of its superior performance. Our scheme combines it with the conditional access sub-system to provide encryption solution at low cost. The illegal user who tries to break that AC based conditional access system faces these problems: (1) The positions of slices $\{\mathbf{b_1}, \mathbf{b_2}, \ldots \mathbf{b_M}\}$ in the concatenated bit stream are unknown. So they cannot be picked out for decoding. (2) The number of symbols in the adaptive source symbol frequency model is unknown. So it brings ambiguity in decoding. (3) Certain $\mathbf{b_i}$, i=1,2,…,M, in the concatenated bit stream are encrypted with probability $p<1$. Such a mixture of cipherdata and plaindata brings more difficulties to cryptanalysis.

According to Shannon's theory on secrecy systems [41], the strength of a cryptosystem is inversely proportional to the source redundancy. Arithmetic coding

diminishes the information redundancy efficiently, so the random appearance of its output bit stream brings a good performance for encryption purposes.

Chapter 6 CONCLUSIONS AND OPEN ISSUES

6.1 Video Codec with JND

In this thesis, we explored the application of a human visual model in video compression, channel coding, error concealment and subjective quality measurement. The 3-D subband video coding scheme is optimized by exploiting the psychovisual properties of the human visual system, so that high perceptual quality of the compressed video can be maintained at a reasonable bit rate. In order to reach the optimality where the coding error is invisible or minimally noticeable at a given viewing distance, the human visual model of JND/MND profiles in subbands are calculated. This model renders a perceptually optimal procedure to quantize the wavelet coefficients in these subbands. The estimation of JND profiles is performed in the spatial domain through analyzing local properties of the video signals. Then in the analysis, in the frequency domain, the JND profile is accordingly decomposed by the modulation transfer function of the HVS into subband profiles.

Taking advantage of the JND profiles, the source encoder has the global control for subjective distortion of the compressed video quality through deciding the coding parameter Δ_G at the beginning of the coding process. It provides a powerful functionality to control the coding quality ahead of time, which is distinct from the conventional compression schemes.

We developed a new perceptual distortion index based on the JND profile to measure the subjective quality of the compressed video. With this numerical metric, we can have a more general and meaningful judgement on the perceptual distortion of

compression. Its performance is compared with the traditional quality measure PSNR. While PSNR can be used to compare the quality of several compression schemes for the same image sequence, heuristically, our distortion index can be used to describe the distortion extent for different image sequences.

In the scenario of a noisy channel, the functionality of error correction, detection and concealment is jointly realized by the channel coding and source coding. A slicing scheme based on JND is suggested to provide a reasonable information packaging according to its importance to human perception, which is the final destination of video information. The objective of the channel coding scheme selection is to approach the compromise of high error protection ability and low bit rate. The energy of JND is computed for each subband as guidance to the choice of RCPC coding rate. Those subbands of more perceptual importance get better error protection.

6.2 Conditional Access using Arithmetic Coding

Arithmetic coding is an efficient entropy coding method. It operates by dispensing with the requirement that each symbol translates into a fixed code. Instead, the code for one symbol in the message merges into that for the next symbol, with no identifiable boundary between. Arithmetic coding is guaranteed to transmit a message in a number of bits that can be made arbitrarily close to its entropy with respect to the model which is used. The probability distribution can change completely from one symbol to another without incurring any penalty in compression efficiency.

The cryptographic property of arithmetic coding is valuable in the design of conditional access sub-systems for commercial broadcasting. Without the knowledge of

start position in the transmitted bit stream for decoding, it is very difficult for the decoder to get synchronization. This contrasts with Huffman coding, where each symbol has its unique corresponding code. Our conditional access sub-system is constructed on the basis of arithmetic coding. The task of the scrambler unit is simplified to break the data into slices. The protection of the information in these slices is provided by the cryptographic feature of the arithmetic coding.

Due to the slicing method also used in our video source coding scheme for error protection, detection and concealment as described previously, to make it consistent with our conditional access sub-system, some variations are necessary. For example, the generation of slicing positions in scrambler cannot be totally random. Taking the JND profile into consideration, the slicing position is basically generated corresponding to the distribution of JND energy as in Chapter 3. Then, for the scrambler, it will modify the position to some extent according to its scrambling key. Since the JND profile is not fixed during the video sequence, it is not easy to figure it out to compute the basis of slicing position (The JND profile data is sliced also). This modification decreases the strength of encryption, but the cryptanalysis for it remains complicated.

6.3 Video Codec with Motion Estimation

Our current video codec is based on subband coding and no motion estimation scheme is used. Actually, lots of explorations have been done in this field. To improve the compression performance, overlapped or multi-scale motion estimation can be adopted along with subband coding. The JND profile can be used for the quantization of the residues after motion compensation. The perceptual distortion energy is still

assigned to proper coefficients in this way. Similar to the technology of P and B frame in MPEG, the distortion energy assigned to these motion predicted frames can be larger while the interframe coded frames have small perceptual distortion to provide a stable basis for motion estimation.

6.4 Joint Source-Channel Coding Based on Human Visual Model

In Chapter 3, in order to constrain the error from noisy channels, we have developed a heuristic way (equation (19)) to segment the data of each subband into slices based on JND profiles. The length of the slice is proportional to its JND energy. Arithmetic coding is applied on each slice with a new adaptive statistical model. As a result, any corrupted slice carries almost the same amount of perceptual importance measured using JND.

Our future work is to develop a route to optimize this slicing process and combine it with the choice of RCPC coding rate for different subbands. Since the statistical expectation of wavelet coefficients in each subbands is close to zero in video coding, whenever a transmission error happens, the corresponding corrupted slice value will be restored as zeros if no other correlative information is available. So the possible error equals the wavelet coefficient value *w(x,y)*. The distortion index for block r_{ij} is defined as:

$$\delta(i,j) \equiv \frac{\sum_{(x,y)\in r_{ij}} w^2(x,y)}{\sum_{(x,y)\in r_{ij}} JND^2(x,y)} \quad , \qquad (22)$$

where *w(x,y)* is the wavelet coefficient at *(x,y)*, and r_{ij} is the block (usually 8 by 8 pixels). The distortion index δ_{ij} is also referred as relative distortion for block r_{ij}.

At the beginning of the slicing procedure for the *k*th subband (*k*=0,1,…,*K*-1, and *K* is the number of subbands), the amount of relative distortion in each slice is determined as ξ_k. The slice is composed of successive blocks, such that in each slice $S_{k,l}$ (*l*=1,2,…, L_k, and L_k is the total number of slices in *k*th subband), heuristically, there holds

$$\sum_{\substack{(i,j)\\ r_{ij}\in S_{k,l}}} \delta(i,j) \cong \xi_k \qquad (23)$$

As a prior, we have the knowledge about the noisy channel, channel coding and modulation in the form of bit error rate $P_{m,n}$ which corresponds to *m*th RCPC coding scheme and channel SNR at *n*dB. We also assume each slice will be coded using arithmetic coding into a bit stream with the length of *c* bits with small variance. Because any slice with one bit error after channel decoding will not be correctly restored in arithmetic decoding, the probability of the corruption of each slice in the *k*th subband is calculated as:

$$P_k = 1-(1-P_{m,n})^c \qquad (24)$$

The relative distortion for the *k*th subband is calculated as

$$\Xi_k = L_k P_k \xi_k \qquad (25)$$

The relative distortion for the whole image is

$$\Xi^* = \sum_{k=0}^{K-1} \omega_k \Xi_k \qquad (26)$$

where K is the number of subbands and ω_k the perceptual weight of the kth subband, as defined in equation (9).

Equations (22) to (26) show the relationship between the global relative distortion of the video and the RCPC coding schemes. Through the adoption of different RCPC coding rate for different subbands, we get channel coding schemes corresponding to certain relative distortion indexes, which reflect the video quality perceptually. This leads to a reasonable control over the channel coding schemes to render the best available video quality, i.e. when the transmission system knows the channel SNR, it can configure the RCPC coding rates for different subbands through a look-up-table, and gets the least relative distortion on video quality. Since the distortion based on the human visual model is assigned to source coding through the selection of quantization stepsize and to the channel coding through the selection of RCPC coding rate, we can adjust the distribution jointly. Essentially, the whole system realizes the joint source-channel coding.

Bibliography

1. J. Woods and S. O'Neil, "Subband coding of images", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp1278-1288, Oct. 1986
2. Y. H. Kim and J. Modestino, "Adaptive entropy-coded subband coding of images", *IEEE Trans. Image Processing*, vol. 1, pp31-48, Jan. 1992
3. G. Karlsson and M. Vetterli, "Three dimensional subband coding of video", in *Proc. ICASSP*, 1988
4. H. Gharavi, "Subband image coding", in *Subband Coding of Video Signals* (J. W. Woods, Ed.). Boston: Kluwer, 1990
5. Taubman and A. Zakhor, "Multirate 3-D subband coding of video", *IEEE Trans. Image Processing*, vol. 3, pp572-588, Sept. 1994
6. C. Podilchuk, N. Jayant and N. Farvardin, "Three-dimensional subband coding of video", *IEEE Trans. Image Processing*, vol. 4, pp125-139, Feb. 1995
7. J. Y. Tham, S. Ranganath and A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment", *IEEE JSAIC*, vol. 16, pp12-27, Jan. 1998
8. N. Tanabe and N. Farvardin, "Subband image coding using entropy-coded quantization over noisy channels", *IEEE JSAIC*, vol. 10, pp926-943, June 1992
9. O. Kwon and R. Chellappa, "Region adaptive subband image coding", *IEEE Trans. Image Processing*, vol. 7, pp632-648, May 1998
10. S. Martucci, I. Sodagar, T. Chiang and Y. Zhang, "A zerotree wavelet video coder", *IEEE Trans. on Circuits and Systems for Video Tech.*, vol.7, pp109-118, Feb. 1997
11. J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients", *IEEE Trans. on Signal Processing*, vol. 41, pp3445-3462, Dec. 1993

12. A. Said and W. Pearlman, “A new fast and efficient image coder based on set partitioning in hierarchical trees”, *IEEE Circuits and Systems for Video Tech.,* vol. 6, pp243-250, June 1996
13. K. Cinkler, “Very low bit-rate wavelet video coding”, *IEEE JSAIC,* vol. 16, pp4-11, Jan. 1998
14. D. Taubman, “Directionality and scalability in image and video compression”, *Ph.D. dissertation,* Univ. California, Berkeley, 1994
15. J. Y. Tham, S. Ranganath and A. Kassim, “Scalable low bit rate video compression using motion compensated 3-D wavelet decomposition”, in *IEEE ICCS/ISPACS 1996,* vol. 3, pp39.7.1-39.7.5, Nov. 1996
16. N. Jayant, J. Johnston and R. Safranek, “Signal compression based on models of human perception”, *Proc. IEEE,* vol. 81, pp1385-1422, Oct. 1993
17. C. H. Chou and Y. C. Li, “A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile”, *IEEE Circuits and Systems for Video Tech.,* vol. 5, pp467-476, Dec. 1995
18. C. H. Chou and C. W. Chen, “A perceptually optimized 3-D subband codec for video communication over wireless channels”, *IEEE Circuits and Systems for Video Tech.,* vol. 6, pp143-156, April 1996
19. S. Daly, “The visible differences predictor: an algorithm for the assessment of image fidelity”, *Measurement and Prediction of Visual Quality,* 1991
20. X. Ran and N. Farvardin, “A perceptually motivated three-component image model - Part I: Description of the model”, *IEEE Trans. Image Processing,* vol. 4, pp401-415, April, 1995

21. M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, “Image coding using wavelet transform”, *IEEE Trans. Image Processing*, vol. 1, pp205-220, April 1992

22. D. H. Kelly, “Motion and vision II. Stabilized spatio-temporal surface*”, J. Opt. Soc. Amer.*, vol. 69, pp1340-1349, Oct. 1979

23. N. Jayant, “Signal compression: Technology targets and research directions*”, IEEE JSAIC,* vol. 10, pp314-323, June 1992

24. N. Farvardin and J. Modestino, “Optimum quantizer performance for a class of non-Gaussian memoryless source”, *IEEE Trans. Information Theory*, vol. 30, pp485-497, May 1984

25. I. Witten, R. M. Neal and J. G. Cleary, “Arithmetic coding for data compression”, *Comm. of the ACM*, Vol. 30, pp520-540, June 1987

26. N. Jayant, “Signal compression: Technology targets and research directions”, *IEEE JSAIC,* vol. 10, pp796-818, June 1992

27. J. Hagenaeur, “Rate-Compatible Punctured Convolutional Codes (RCPC Codes) and Their Application ”, *IEEE Trans. Comm.*, vol. 36, no.4, pp. 389-400, April 1988

28. A. N. Netravali and B. G. Haskell, “Digital Pictures: Representation and Compression”, New York: Plenum, 1988

29. I. Daubechies, “Ten lectures on wavelets”, *Society for Industrial and Applied Mathematics*, Philadelphia, 1992

30. S. G. Mallat, “A theory for multi-frequency signal decomposition”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp674-693, July 1989

31. A. K Jain, “Fundamentals of Digital Image Processing”, Prentice Hall

32. H. Gharavi and A. Tabatabai, “Subband coding of monochrome and color images”, *IEEE Trans. Circuits Sust.*, vol. 35, pp207-214, Feb. 1998

33. K. Sharifi and A. Leon-Garcia, “Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video”, *IEEE Trans. on Circuits and Sys. For Video Tech.*, vol. 5, No. 1, Feb. 1995

34. J. Gu, Y. Jiang and J. Baras, “A Practical Transmission System Based on the Human Visual Model for Satellite Channels”, *ISR Technical Report* TR 99-14, 1999

35. W. Kim, K. Chen and H. Cho, “Design and implementation of MPEG-2/DVB scrambler unit and VLSI chip”, *IEEE Trans. Consumer Electronics*, vol. 43, pp980-985, Aug. 1997

36. A. S. Fraenkel and S. T. Klein, “Complexity aspects of guessing prefix codes”, *Algorithmica*, vol. 12, pp. 409-419, 1994

37. D. W. Gillman, M. Mohtashemi and R; L. Rivest, “On breaking a Huffman code”, *IEEE Trans. Information Theory*, vol. 42, pp972-976, May 1996

38. I. H. Witten and J. G. Cleary, “On the privacy afforded by adaptive text compression”, *Computers and Security*, vol. 7, pp397-408, 1988

39. T. Cover and J. Thomas, “Elements of information theory”, Prentice Hall, 1991

40. C. Boyd, “Applying compression coding in cryptography”, *Cryptography and Coding*, vol.2, Clarendon Press, 1992

41. C. E. Shannon, “Communication theory of secrecy systems”, *Bell Systems Technical Journal*, pp656-715, 1949

42. S. Dalbegue, J. Baras and N. Sidiropoulos, “Compact image coding from multiscale edges”, *ISR Technical Report* TR 98-61, 1998