

Information Theoretic Measures for Encoding Video

by

Giridharan Iyengar

M. A. Sc, Electrical and Computer Engineering
University of Ottawa, Ottawa, Canada
(1993)

SUBMITTED TO THE PROGRAM IN
MEDIA ARTS AND SCIENCES,
SCHOOL OF ARCHITECTURE AND PLANNING
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1995

©1995 Massachusetts Institute of Technology.
All rights reserved.

Author:

Program in Media Arts and Sciences
July 31, 1995

Certified by:

Andrew B. Lippman
Associate Director, MIT Media Laboratory
Thesis Supervisor

Accepted by:

Stephen A. Benton
Chairperson, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

OCT 26 1995

Rotch

LIBRARIES

Title: Information Theoretic Measures for Encoding Video

Author: Giridharan Iyengar

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science

July 31, 1995

Abstract

A method for analyzing a volume of video is presented and is used to design a hierarchical class of image coders. This thesis is based on the observation that no single image coder is optimal for all image sequences. The performance of any single coder varies over different regions of the same image as well as over sequence of images. Hence a class of hierarchically structured image coders is presented. The task is to describe each region with as few parameters as possible.

The thesis work is split into two major parts. First, we analyze the input video to measure the spatio-temporal predictability of different regions in this volume. This analysis forms the basis for the design of the hierarchical class of coders. The volume of video thus analyzed is then encoded efficiently using multi-resolution approaches. Associated research also indicates that this analysis enables content information to be embedded in the bitstream with marginal cost.

Ultimately, these coders form a hierarchical representation language that can be described as part of the coding process, transmitted with the pictures and reconstructed at the receiver.

Advisor: Andrew B. Lippman
Associate Director, MIT Media Laboratory.

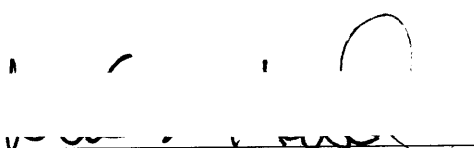
This research was supported by contracts from the Television of Tomorrow consortium.

**Information Theoretic Measures
for Encoding Video**

by


Giridharan Iyengar

Reader:



Michael Hawley
Assistant Professor
Program in Media Arts and Sciences

Reader:



Jules Bellisio
Executive Director, Video Systems and Signal Processing Research
Bellcore

To my parents and my teachers

Contents

1	Introduction and Motivation	12
1.1	Context	12
1.2	Motivation	13
1.3	Approach	17
2	Information Theory Review	19
2.1	Entropy and Mutual Information	20
2.1.1	Entropy	20
2.1.2	Mutual Information	22
2.2	Prediction and Information Theory	24
3	Image Coding Review	28
3.1	The JPEG Standard	28
3.1.1	Transform coding block	29
3.1.2	Quantization block	30
3.1.3	Entropy Coding Block	31

3.2	Wavelets and Filter Banks	32
3.3	MPEG and Motion Compensation	34
3.3.1	Motion Estimation and Compensation	34
3.4	Recent Advances in Image Representation and Coding	36
3.4.1	Salient Stills	37
3.4.2	Structured Video Coding	37
3.4.3	Layered Coding	38
3.4.4	Enhanced Resolution Coding	39
4	Mutual Information and Characterization of Video	43
4.1	Mutual Information Metric for Characterization	45
4.1.1	Selection of Spatial and Temporal Activity Measures	45
4.1.2	Design of the Mutual Information Metric	48
4.2	Experiments Performed on Test Sequences	50
4.3	Use of the Characterization Measure	52
5	Controlled Coding	55
5.1	Hierarchical Image Coder	57
5.2	Performance Evaluation of the Controlled Coder	61
5.2.1	Experiment	61
5.2.2	Discussion of results	62
6	Summary and Future Work	71

6.1	Results of the research work	71
6.2	Discussion	72
6.3	Future Work	73
	Bibliography	75
	Acknowledgments	79

List of Figures

1.1	One frame from the “Sharky” video sequence	15
1.2	Bits per region image for the MPEG encoder for the given frame . . .	16
1.3	Bits per region image for the LVQ encoder for the same frame	17
3.1	The 64 DCT basis functions	40
3.2	The classic Mallat tree-structured filter bank in wavelet decomposition	41
3.3	Time frequency decomposition by a Wavelet transform	42
4.1	One frame from the plain two-color cross image	51
4.2	One frame from the “sharky’s machine” sequence	52
4.3	Pictorial result of the analysis	53
5.1	Time domain response of Adelson’s 9 tap QMF	59
5.2	The Haar scaling and wavelet functions	60
5.3	Quantization table of the mother algorithm	64
5.4	Frame 0 of original test sequence 1	65

5.5	Same frame reconstructed using the baseline coder. Compression ratio	
	100:1	66
5.6	Same frame reconstructed using the hierarchical coder. Compression	
	ratio 100:1	67
5.7	Frame 32 of original test sequence 2	68
5.8	Same frame reconstructed using the baseline coder. Compression ratio	
	20:1	69
5.9	Same frame reconstructed using the hierarchical coder. Compression	
	ratio 20:1	70

Chapter 1

Introduction and Motivation

1.1 Context

We are making a transition from analog television to digital distribution. The impact of this is a major technological change in the most pervasive medium of all time. Past work at the Media Laboratory and MIT has concentrated on exploiting the transition to digital high definition so that new generations of TV can be scalable, international and universal. The point has been amply demonstrated and is starting to become accepted.

On the other hand, the first digital standards fail to reach these goals. They are narrow, restrictive and cumbersome. Their most important feature is that they will

get the revolution in image representation started. MPEG is not the best that can be done, but its worldwide acceptance is in large measure responsible for digital television; and it is becoming used in commercial systems (Hughes DirecTV for example).

An underexplored benefit of digital television is that the signal can simultaneously carry the raw image reconstruction information as well as correlated descriptive data. In particular, the age of standard algorithms may well be past: each image data stream can carry the decoding instructions along with it. This thesis exploits that characteristic to allow a set of individual coders to all cooperatively analyze, compress and represent video sequences.

1.2 Motivation

The current state of the art in video compression provides acceptable consumer TV quality with compression in the order of 25:1. Nevertheless, the level of sophistication of these coders is rather low. These coders reduce the bitrate very efficiently but rarely do anything beyond that. There is no information concerning the content of the video itself. As broadband information channels mature, it will be crucial that new representations are found for video that enable querying, searching and navigating through the infinitude of material in the global database.

These are lofty ideals, given the limitations of present day technology. Nevertheless, steps need to be taken to bridge the gap between the need and the availability.

Popular lossy image coding can be roughly classified into two main subcategories: transform domain techniques (MPEG, Wavelets, and Sub-band coding) and structured coding techniques (Layered coding and Model based coding). Each coding technique performs differently on different image sequences. In other words, there is a strong link between coding method and the image type to be encoded.

For example, we compare two existing coders: MPEG2 and Library-based Vector Quantization (LVQ) [1]. LVQ differs from MPEG2 in that it uses vector quantization in the prediction space as opposed to transmitting motion vectors. Fig. 1.1 shows one frame of the original sequence that was coded using MPEG2 and LVQ. Figs. 1.2 and 1.3 represent the bits per region spent in representing that frame by the MPEG and the LVQ coders, respectively. In these images, the grey level value of each macroblock is a representative of the number of bits required to encode that macroblock using the given coder, with white and black representing the top and bottom of the scale respectively. Both the coders operate at the same target bitrate of $4Mb/s$.

Even a qualitative comparison as the one above using two very similar coders illustrates the relationship between region type and the encoder. Moreover, this relationship depends very much on the bitrate at which the coders operate.



Figure 1.1: One frame from the “Sharky” video sequence

This difference in performance between different coders can be explained in terms of the assumptions that the particular coder makes. For example, the layered coder of Adelson [17] tries to fit an affine model to each one of the layers that it builds. Perspective motion is one clear example of a motion that cannot be accurately modeled by affine parameters. Hence, we can expect that the affine model will fail in regions where there is perspective motion. Almost all motion detection algorithms perform poorly in regions with low textural information (“flat regions”). Hence, any coder based on motion estimation can be expected to perform poorly in such regions. It is probably best to detect such regions, segment them and encode using a very

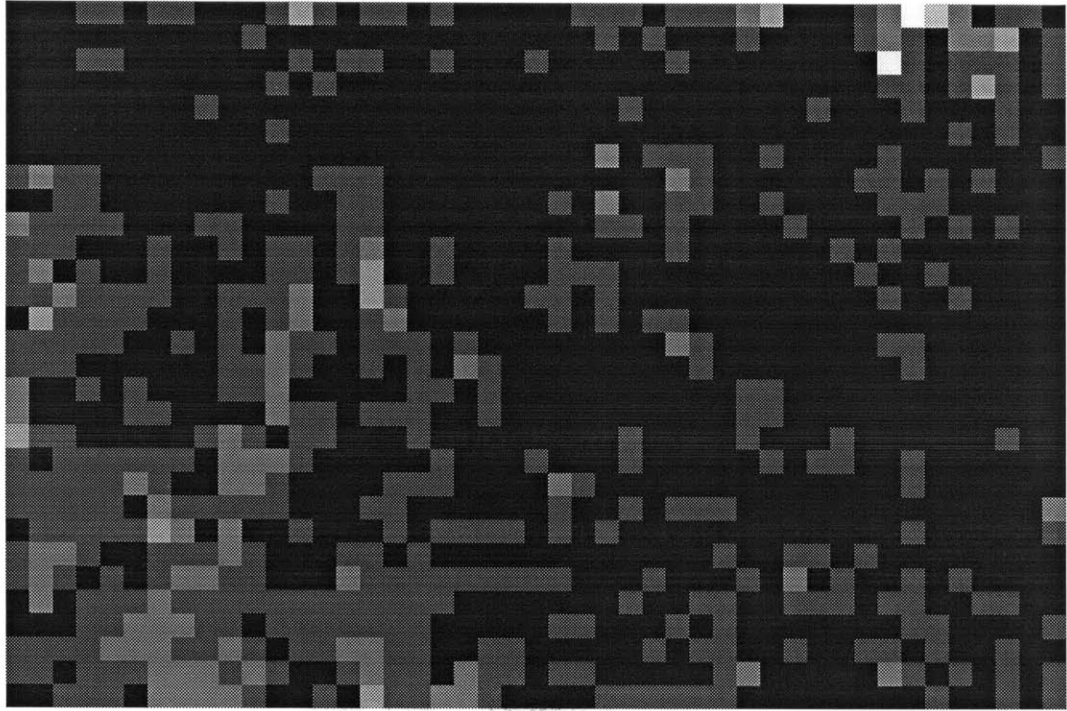


Figure 1.2: Bits per region image for the MPEG encoder for the given frame

simple technique such as transmitting the average value for the region or using DCT or some other similar technique. Many object oriented coding techniques depend on good segmentation. Accurate segmentation is coupled to motion estimation and is, in general, an ill-posed problem.

Hence the work of this thesis. Any single algorithm is good in its own merit but does not work well for all images. Also, its performance varies from region to region within the same image that it encodes. Therefore, it is desirable to design a coder that prunes its spatial and temporal extents to optimally suit the image statistics in

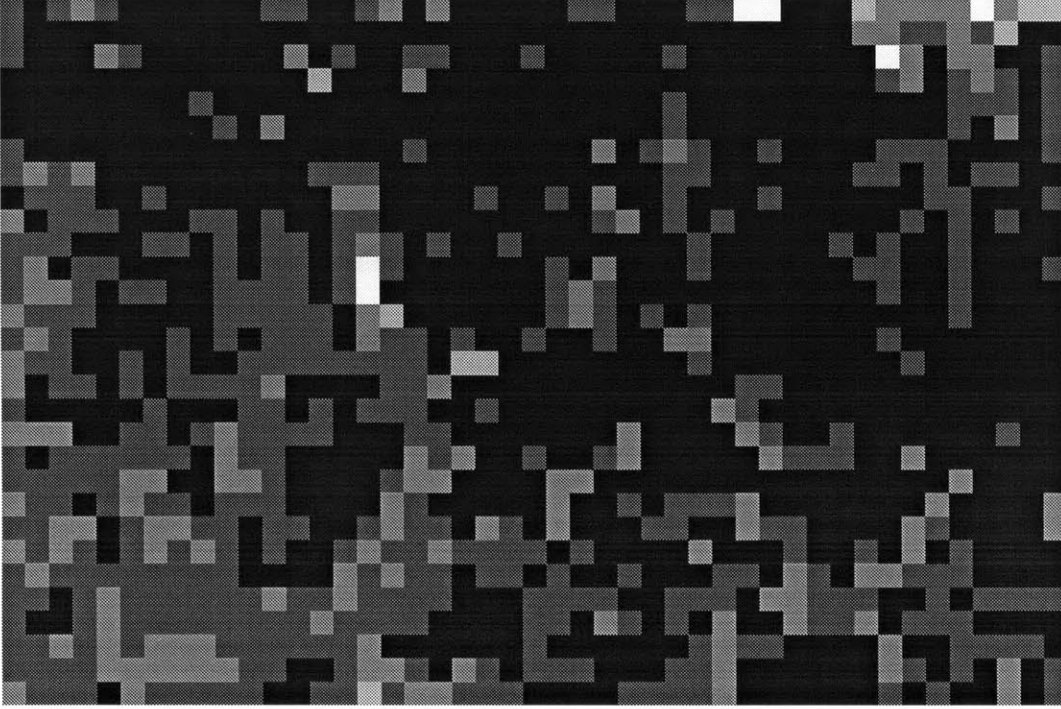


Figure 1.3: Bits per region image for the LVQ encoder for the same frame that region.

1.3 Approach

This thesis proposes a method to analyze video in a conjoint spatio-temporal volume. As a practical application to this analysis, an architectural extension to existing Quadrature Mirror Filter coders is proposed. The approach assumes no apriori knowledge of the characteristics of the video it analyses. The coder is no longer a

simple algorithm operating identically on every region of the image but rather the analogy is that of a class of coders organized in a hierarchical fashion, each tuned to best suit the characteristics of the spatio-temporal region.

Each coder in this hierarchical class is composed of a separable Quadrature Mirror Filter (QMF) in the spatial domain and a Haar basis in time domain. In a sense, this construction can be viewed as a “best basis” decomposition of the volume of video rather than a single image as is done currently. In the traditional “best basis” approach, only the frequency space is tessellated to suit the image characteristics. In this hierarchical approach, we tessellate the conjoint spatio-temporal-frequency space with targeted bases.

The remainder of this thesis is organized as follows. Chapter Two briefly reviews the fundamentals of Information theory. Chapter Three presents some current work in image coding and representation. Chapter Four details the spatio-temporal analysis algorithm along with some sanity experiments. Chapter Five outlines the hierarchical coder. Simulation results are presented in Chapter Six followed by suggestions for future work and conclusions.

Chapter 2

Information Theory Review

Information theory is a discipline centered around the mathematical approach to the study of the collection and manipulation of information. It provides a theoretical basis for data compression, communication, estimation and pattern recognition [2]. The foundation for the subject of information theory was laid by Shannon. By means of an existence proof, he showed that nearly error-free communication is possible over noisy channels, with a suitable *encoder* and a *decoder* under certain constraints [3].

2.1 Entropy and Mutual Information

There are strong links between information theory, statistical pattern recognition, decision and estimation theory and physics and thermodynamics. This probably justifies the choice of the name *entropy* for the fundamental quantity of information theory [2, 4]. There are three fundamental mathematical functions that are used to measure information. We begin by introducing *entropy*.

2.1.1 Entropy

Entropy is the simplest of the three functions. It is probably best introduced in terms of an example news event. “The weather in Boston is rainy and cloudy” is a news event that contains some information, though hardly surprising to a Boston resident. If it were first of February, the news “The expected high of today is 78 degrees” would be a much more surprising event and certainly carries more information. Moreover, exchanges like “It is hot today” and “What happened to Carmen San Diego?” are unrelated and hence must add together. We can conclude that the quantitative definition of information must have some of the intuitive properties such as:

1. Information contained in an event must be related to the uncertainty of the event.
2. Less certain events must contain more information than more certain events.
3. Information of unrelated events taken together must equal the sum of information of these events.

Given these desirable and intuitive properties, Shannon defined $-\log P(\alpha)$ as a measure of information [3] where P is the probability of the event α occurring. This measure satisfies the above three criteria and the specific base to which the logarithm is taken only determines the scaling factor and hence the unit of information. Customarily, the logarithm is taken to base 2 and the unit of information is then given in *bits*. We then define *entropy* as the average amount of information per source (where the source produces one of a finite set of events at any instance).

$$H_1(S) = - \sum_{k=1}^N P(\alpha_k) \log P(\alpha_k) \quad (2.1)$$

where α_k , $k = 1, \dots, N$ are the N different source symbols.

This measure is also called the first-order entropy indicating that the entropy was calculated taking one isolated event at a time. We can also define an N th-order

entropy where N source symbols are taken together.

$$H_N(S) = \sum_{\alpha_1, \alpha_2, \dots, \alpha_N} P(\alpha_1, \alpha_2, \dots, \alpha_N) \log P(\alpha_1, \alpha_2, \dots, \alpha_N) \quad (2.2)$$

If the source symbols are statistically independent, then $H_N(S) = NH_1(S)$, as can be easily verified. Similarly, if the $N - 1$ symbols are completely determined by the first symbol, $H_N(S) = H_1(S)$. Hence, in general $H_N(S) \leq NH_1(S)$. This inequality leads us to the definition of Mutual Information.

2.1.2 Mutual Information

Shannon demonstrated how information can be reliably transmitted over noisy channels [3]. To do this, he constructed a measure of the amount of information about the transmitted message contained in the observed output of a channel. We can do this by defining a notion of mutual information between two events α and β denoted by $I(\alpha : \beta)$. To be consistent with the previous definition of entropy, $I(\alpha : \beta)$ must have the following properties:

1. If α and β are independent events, the occurrence of β would provide no information about α . Hence $I(\alpha : \beta) = 0$.

2. If β implies that α is the only event that occurred, then $I(\alpha : \beta) = I(\alpha) = \log\left(\frac{1}{P(\alpha)}\right)$

These two conditions are satisfied if we define *mutual information* as follows.

$$I(\alpha : \beta) = \log \frac{P(\alpha | \beta)}{P(\alpha)} \quad (2.3)$$

The *average mutual information* is hence

$$I(a : b) = \sum_b \sum_a p(b | a)q(a) \log \frac{p(b | a)}{\sum_{a'} p(b | a')q(a')} \quad (2.4)$$

where $p(b)$ and $q(a)$ denote the probability density functions of the alphabets a and b . We can see that

$$I(a : b) = H(a) + H(b) - H(ab) \quad (2.5)$$

That is, mutual information is the difference between the separate entropies and the joint entropy.

We can now define the *self mutual information*, which is of interest to us in the context of this thesis. It is the amount of information shared between d symbols taken together. Hence, the d -dimensional *self mutual information* for a source S is defined as:

$$I_d(S) = dH_1(S) - H_d(S) \quad (2.6)$$

Similarly, we can define *incremental self mutual information* or *redundancy* as:

$$R_d(S) = H_1(S) + H_{d-1}(S) - H_d(S) \quad (2.7)$$

This essentially tells us the amount of information that a new observation adds, given that we have $d - 1$ observations already.

These two measures play a very important role in this thesis. They will be further elaborated in section 2.2.

2.2 Prediction and Information Theory

So far, we have provided definitions of various measures of information theory. Information theory gets introduced in a totally different light in the domain of time series prediction.

Time series problems arise in many disciplines such as predicting the stock prices, economic models, predicting patterns and orbits of stellar bodies, modeling heartbeat to attack arrhythmia. Time Series analysis concerns itself with Forecasting, Modeling and Characterization. *Forecasting* attempts to accurately predict the short term evolution of any system. The goal of *modeling* is to find a description of the system that effectively captures the long term behavior of the system. *Characterization*

attempts to determine the fundamental properties of the system such as the degrees of freedom, amount of information and randomness etc. In this thesis, we are interested in characterization of video. Information theory plays a very important role in the characterization effort [5].

We recall from Eq. 2.6, that the *self mutual information* indicates the amount of information we can get about $d - 1$ from one observation. So far, we have ignored time in our formulation. We redefine the terms *entropy* and *mutual information* with the time series context. Let $x(t)$ be a time series that has been digitized to integer values between 1 and N . This is very relevant to our context of digital video since video is normally viewed as a discrete time varying series. The 1-dimensional *entropy* of this time series is given by:

$$H_1(N) = - \sum_{x=1}^N p_1(x) \log_2 p_1(x) \quad (2.8)$$

where $p_1(x)$ is the first order probability density function of $x(t)$.

We define a d -dimensional *lag vector*, $X_t = (x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau})$. This vector is essentially the time series at t taken along with $d - 1$ observations, each separated by a fixed time delay τ . Essentially, we are interested in the *mutual information* that these d observations taken together share. It makes a powerful statement about the predictability of these d observations. The *joint entropy* is then defined on this lag

vector by:

$$H_d(\tau, N) = - \sum_{x_t=1}^N \cdots \sum_{x_{t-(d-1)\tau}=1}^N p_d(x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau}) \log_2 p_d(x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau}) \quad (2.9)$$

This is the average number of bits required to describe a sequence. In the limit of small lags, $\tau \rightarrow 0$, we obtain:

$$\lim_{\tau \rightarrow 0} p_d(x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau}) = p_1(x) \quad (2.10)$$

and hence $H_d(0, N) = H_1(N)$. In the opposite case, when $\tau \rightarrow \infty$,

$$\lim_{\tau \rightarrow \infty} p_d(x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau}) = p_1(x_t)p_1(x_{t-\tau}), \dots, p_1(x_{t-(d-1)\tau}) \quad (2.11)$$

which implies $\lim_{\tau \rightarrow \infty} H_d(\tau, N) = dH_1(N)$. These two equations make intuitive sense. That is if our observations are very close to each other, they are very redundant. If they are far apart, they have no common relationship. We can similarly define *mutual information* and *redundancy*. The importance of the additional variable (time) introduced in this context is that it allows us to characterize and analyze time series. For example, we can ask questions such as

1. What is the ideal embedding dimension? That is the number of observations that need to be considered together.

2. What is the ideal temporal depth? We can quantitatively evaluate the ideal depth at which we must observe a time series in order to represent it compactly. We ask this question extensively in this thesis.

The concepts of *entropy* and *mutual information* are directly linked to *generalized dimensions*, *Lyapunov exponents* etc. For the sake of brevity and focus, we will not address these aspects in this thesis. Interested readers are encouraged to refer to [5] for an excellent overview of the subject. The reference [5] also contains some examples of usage of these concepts in real life problems.

Chapter 3

Image Coding Review

Currently all displayed pictures, whether static or moving, are composed of still pictures. The present day coding techniques for movies hence tend to be based on approaches similar to those used for encoding still pictures. In the following section, we will examine some of these issues as handled in the JPEG and MPEG coders [6, 7]. In the next section, we will study some of the recent advances in image representation.

3.1 The JPEG Standard

The Joint Photographic Experts Group (JPEG) standard for still image compression is a fairly simple combination of three main functional blocks: a Transform Coding

Block, a Quantization Block and an Entropy Coding block. Almost all of the present day encoders fit within such a framework of three main functional units operating sequentially. These functional units are typically partially independently designed.

3.1.1 Transform coding block

The transform coding block transforms the spatial data in a still image into an equivalent representation in the transform domain. The major advantage of such a transformation is the compaction in energy that it achieves. Typically, a good transformation will concentrate the energy of the image into a few narrow bands in the transformed space [8]. One measure of performance of the goodness of transformation is the *Transform Coding Gain* which is a quantitative measure for the energy compaction achieved by the transformation [8]. JPEG uses the Discrete Cosine Transform (DCT) in its transformation block. In order to keep the computation complexity within bounds, JPEG divides the image into 8×8 blocks and independently transforms each block. This 8×8 block is represented by a set of 64 coefficients, organized approximately in increasing frequency bands. The formulae for the DCT and its inverse are given below:

$$F(u, v) = \frac{2}{N} C(u) C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos\left(\frac{(2x+1)u\pi}{N}\right) \cos\left(\frac{(2y+1)v\pi}{N}\right) \quad (3.1)$$

$$f(x, y) = \frac{2}{N} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} C(u)C(v)F(u, v)\cos\left(\frac{(2x+1)u\pi}{N}\right)\cos\left(\frac{(2y+1)v\pi}{N}\right) \quad (3.2)$$

where

$$C(u), C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } u, v = 0 \\ 1 & \text{otherwise} \end{cases} \quad (3.3)$$

where x, y are the coordinates in the pixel domain and u, v are the coordinates in the transform domain.

The 64 basis functions of DCT are illustrated in figure 3.1. As we can see, they are arranged in increasing order of frequencies in horizontal and in vertical direction. The DCT basis functions are very close to the optimal basis functions, Karhunen Loeve Transform (KLT, which are the eigenvectors of the correlation matrix), for first order Markov sources [8].

3.1.2 Quantization block

The second block in the JPEG encoder is the quantization block. This is where maximum compression is achieved in the JPEG algorithm and is also the part where information loss occurs.

The coefficients corresponding to lower frequency bands of the DCT basis functions contain most of the energy in the image. Moreover, the human visual system is per-

ceptually less sensitive to the distortions in the higher frequency coefficients. Hence, in JPEG, the coefficients are quantized with varying severity. The DC or the zero-frequency coefficient is particularly important since it controls the average intensity of the entire block. Hence, it is treated with a much finer quantization.

A similar strategy is employed for color images. An RGB color image is first transformed into the YUV space where the image is represented by one luminance and two color differences. These two color differences are typically subsampled by half spatially, throwing away half the samples. This is followed by a usual transform and quantization as mentioned above.

3.1.3 Entropy Coding Block

Entropy coding block is the last stage of the JPEG coder. This is a lossless stage of the algorithm. This stage exploits the redundancy in the representation and thereby reducing the number of bits required.

For example, consider that a data set is completely random. That is, if each symbol is equally likely, from the definition of entropy in the previous chapter, we see that the entropy of such a data set would be the maximum possible value. There will be no redundancy in the representation and the mutual information between the symbols will be zero. The best we can possibly do is to code each symbol with the number of

bits specified by the entropy of the data set.

On the other hand, let the data set be such that a few symbols occur with very high probability and the others are more rare. In such a case, we can represent the frequently occurring symbols with fewer bits and others with more number of bits and on the average reduce the number of bits required for the representation. Similarly, if the data set is redundant, that is we can make a very good guess of the following few symbols given one symbol, we can represent this entire block more efficiently.

After the quantization, many of the coefficients are zeroed out in every block. JPEG tries to exploit this by a zig-zag scanning of coefficients followed by run length encoding and Huffman coding. The zig-zag scanning arranges the coefficients in an approximately increasing order of frequencies. The run length encoder represents a run of zeros by a value rather than having a symbol for every zero. This is then followed by a Huffman coder which tries to reduce the redundancy further.

3.2 Wavelets and Filter Banks

Wavelets and filter banks are two related ideas that enhance the theory and understanding of the subject of coding and representation in a very big way [9, 10, 11].

A *filter bank* is a set of filters separating the input signal into frequency bands (by

what is called the *analysis* bank). Often times, there are two filters, one low pass and one high pass. These sub-signals can then be compressed more efficiently and recombined by the *synthesis* bank. Figure 3.2 illustrates a typical recursive Wavelet decomposition tree expressed as a series of cascaded filter banks. The subject of Multiresolution Analysis is intricately tied with the theory of Wavelets [12].

DCT, the work-horse of JPEG can also be formulated as a filter bank. This filter bank decomposes the frequency spectrum into 64 bins. These bins are not exactly equal sized and overlap extensively. The main idea of a Wavelet is that of a multiresolution decomposition. This is probably best illustrated by a figure (3.3). Whereas a windowed transform analyses all frequency bands with equal resolution, a wavelet decomposition treats different frequency ranges differently. The lower frequencies are analyzed for a longer time period and the higher frequencies are analyzed over a shorter time period. The salient point of wavelets over JPEG is the reduction in block artifacts. However, the block artifacts of a JPEG coder are replaced by an overall fuzziness and ringing effects in a wavelet coder.

The two other blocks of a wavelet coder (quantization and entropy coding) are remarkably similar to the JPEG coder. We have similar perceptually weighted quantization schemes followed by run length and huffman / arithmetic coding.

3.3 MPEG and Motion Compensation

We have so far looked at still picture coding strategies. We will present methods to encode moving pictures. These methods are quite similar to the still picture coding techniques. We begin with presenting the Moving Pictures Experts Group (MPEG) standard. The MPEG standards are based extensively on JPEG. In fact, they use the same transform, quantization and entropy coding blocks, but with different parameters. The main difference lies in the fact that the MPEG coders exploit temporal redundancy in moving pictures through motion estimation and compensation.

3.3.1 Motion Estimation and Compensation

MPEG does not specify any particular motion estimation algorithm for generating motion estimates. However, to keep the decoder complexity under control, a decision was made that the motion vectors must represent constant motion for small regions of the pictures. The boundaries of these regions are chosen to contain 4 blocks used by the DCT stage. Thus, the decoder could operate with a minimal amount of memory. The most often used approach to estimate motion in MPEG is a block based motion estimator.

A block based motion estimator is a simple motion estimator that generates a single

motion vector for a block of pixels. The block based motion estimator searches for the best matching block within a neighborhood in the preceding picture for a given block in the current picture. It achieves this by minimizing an error criterion such as Sum of Absolute Differences (SAD), Mean Squared Error (MSE) etc. There are several advantages and disadvantages of using such an approach for motion estimation. The main advantage is that the resulting motion vectors do in fact reduce the entropy of the transform coded error signal. However, the block based approach does not guarantee that the resulting motion vectors are true representatives of the motion in that region of the image. Moreover, blockwise motion estimates assumes purely translational motion model which is not entirely correct.

Optical flow [13] is another approach to estimate the motion occurring in image sequences. It is the apparent motion of brightness pattern in images. Ideally the optical flow will correspond to the motion field but it is not necessarily so [13]. However, in many cases, optical flow is a reasonable estimate of the motion field.

Optical flow is usually estimated by solving a constraint equation given below.

Let $E(x, y, t)$ be the intensity value at time t and at position (x, y) .

If $u(x, y)$ and $v(x, y)$ are the optical flow components. At time $t + \delta t$, we expect this intensity to be at $(x + \delta x, y + \delta y)$ for small δt . Hence:

$$E(x + u\delta x, y + v\delta y, t + \delta t) = E(x, y, t) \quad (3.4)$$

This single constraint is not sufficient to determine u and v uniquely. If we impose the constraint that optical flow is continuous everywhere, we get

$$E_x u + E_y v + E_t = 0 \quad (3.5)$$

where E_x is the partial derivative in x and so forth. This equation is the optical flow constraint equation.

All a local measurement of flow can do is identify this constraint line and we cannot estimate the flow in the direction perpendicular to the gradient [13]. This is known as the aperture problem. There are many techniques to impose additional constraints and estimate the normal flow [14].

3.4 Recent Advances in Image Representation and Coding

In this section we describe some of the recent advances in image representation undertaken at the Media Laboratory.

3.4.1 Salient Stills

Salient Stills developed by Teodosio [15] is a good example of what can be done given memory and processing power. Salient Stills constructs a single image from a sequence of images. This image is salient in terms of the information content. Moreover, it extracts resolution out of time. The main ideas of the salient still algorithm are:

- A single image is created by warping many frames into a common space.
- The value of a pixel at a particular point is the result of a temporally weighted median of all warped pixels at that point.
- The motion model used in this algorithm is a 6 parameter affine model fitted to optical flow.

3.4.2 Structured Video Coding

Patrick McLean's thesis "Structured Video Coding" [16] explores what is possible if one already has an accurate representation of the background. He used footage from the 1950's sitcom "I Love Lucy" to illustrate the advantage of a structured approach to video coding. Having extracted and separately sent the background, McLean showed that it was possible to encode video with reasonable quality at 64 kbits/s using Motion JPEG. Moreover, this thesis also illustrated that it is important

to encode on content and not on statistical information.

3.4.3 Layered Coding

More recently, work in the area of layered coding has been done by Wang, Adelson and Desai [17]. Along the same lines as the Salient Still and the Structured Coder, the Layered Coder addresses the idea of modeling not just the background by creating a small number of regions whose motions can be defined by a 6 parameter affine model. These regions are transmitted once at the beginning of the sequence with subsequent frames generated by compositing the regions together after they have been warped using the 6 parameter models.

This technique works best for scenes where rigid body non-3D motion takes place. Another problem is that the inaccuracies in segmentation cause distortion in region boundaries when the regions are composited together. Finally, since only one image per region is transmitted, closure is not guaranteed. That is, in the composited image, it is not guaranteed that each pixel will have a non-null value.

3.4.4 Enhanced Resolution Coding

Enhanced resolution coding [19] can be viewed as an extension to the Salient Stills coder. It employs the same idea of Salient Stills of obtaining additional resolution from time. It begins by decomposing each frame into background and foreground objects. It then enhances the resolution of the background by looking over several frames and compositing a high resolution background. The foreground object is then composited back into the background and thus completing the frame.

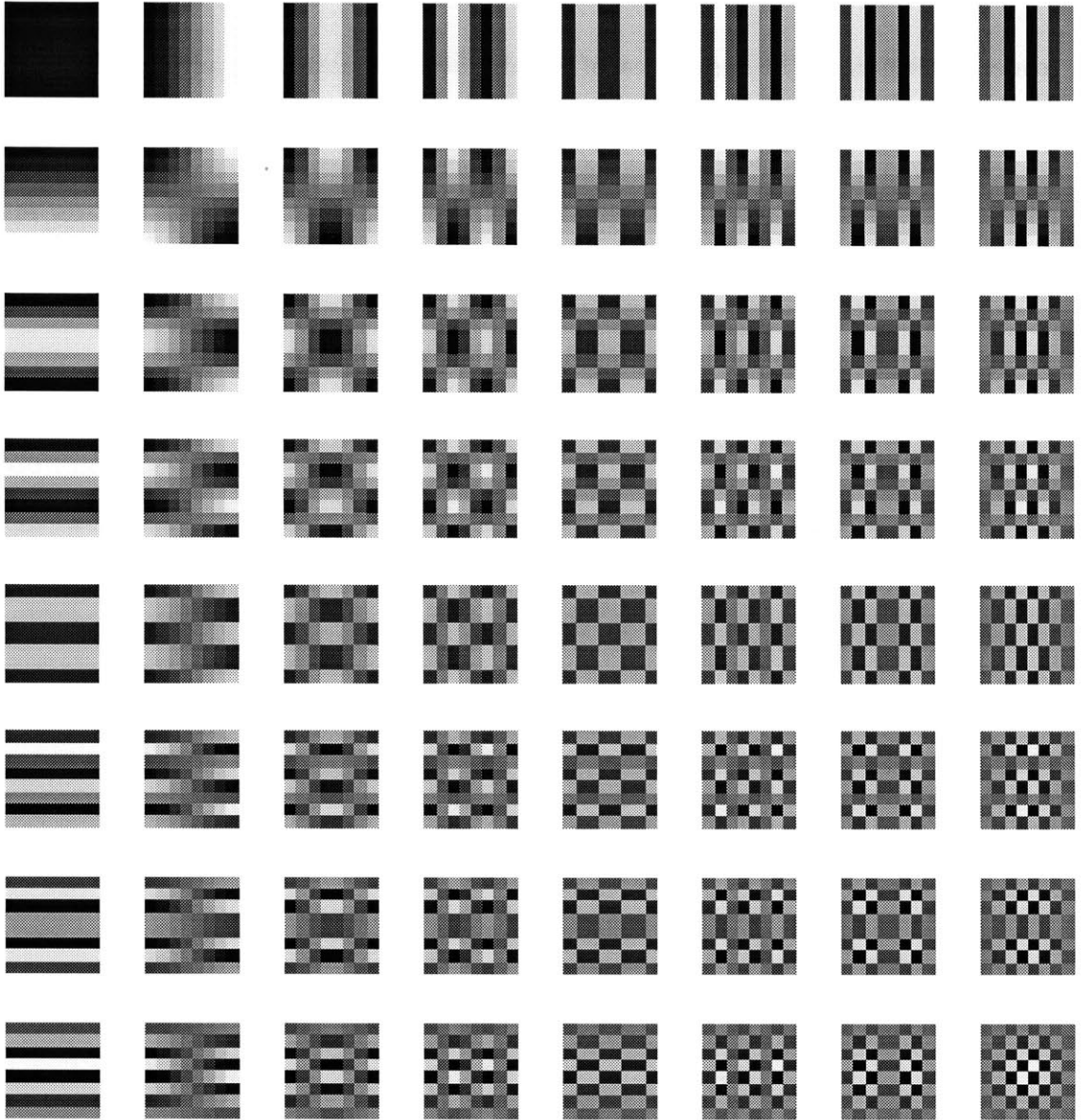


Figure 3.1: The 64 DCT basis functions

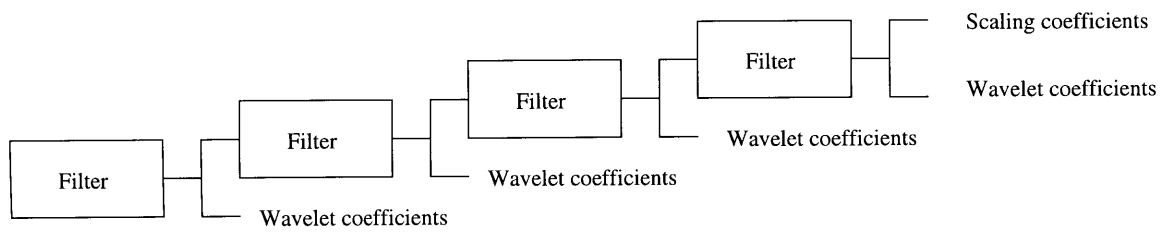


Figure 3.2: The classic Mallat tree-structured filter bank in wavelet decomposition

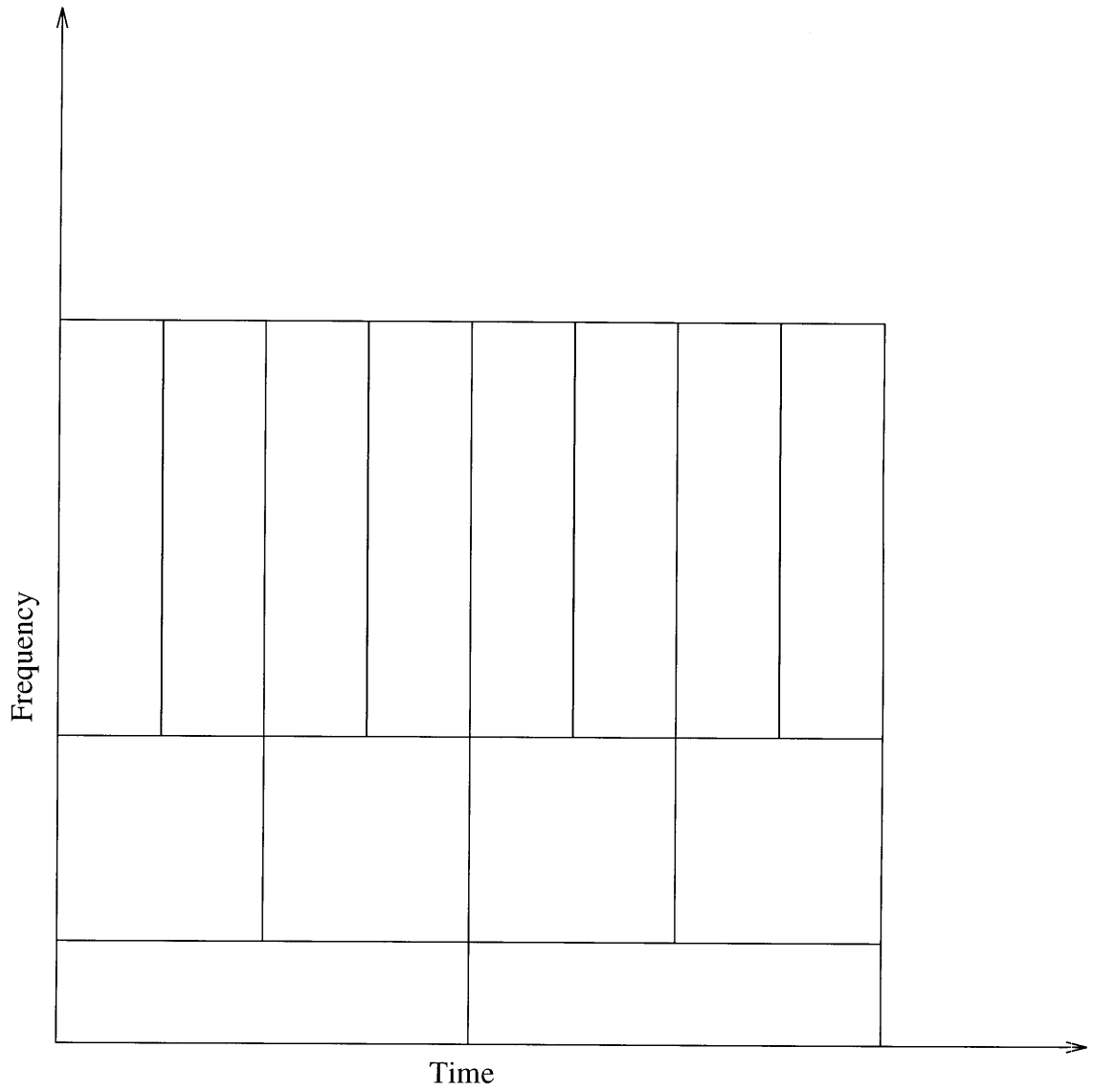


Figure 3.3: Time frequency decomposition by a Wavelet transform

Chapter 4

Mutual Information and Characterization of Video

In chapter 2, we reviewed some of the basic notions of information theory. In section 2.2, we alluded to some of the ideas used in this thesis. The addition of time in the formulation of entropy neatly ties the concept of entropy with the problem of time series analysis and we now have a powerful tool to analyze and characterize time series, including digital video [5].

In this chapter, we will expound the questions introduced in section 2.2 and present the characterization algorithm that provides the basis for doing controlled encoding of video, querying and many such interesting and challenging problems. We recall that

entropy and *mutual information* permit us to characterize and analyze time series.

Specifically, we can ask questions such as:

1. What is the ideal embedding dimension? That is the number of observations that need to be considered together.
2. What is the ideal temporal depth? We can quantitatively evaluate the ideal depth at which we must observe a time series in order to represent it compactly.
3. For a 2D series, such as video, we have the additional questions of ideal spatial depth and the selection of spatial versus temporal embedding.

The last two questions are very relevant in the context of controlled encoding. They would enable us to choose and design suitable algorithms for compact and efficient encoding, the fundamental idea of this thesis. Given a video sequence, the task of estimating suitable spatial and temporal depths and the selection of suitable encoders requires measures for temporal and spatial activity. These activity measures can then be used to analyze and characterize the video sequence. We study two such measures in this chapter.

4.1 Mutual Information Metric for Characterization

We recall from chapter 2 that *mutual information* is the amount of information shared between a few measurements taken together (see equation 2.6). We saw importantly, in the context of a time series, the measurement of mutual information of a quantity tells us about the predictability of the time series along that measurement (see equation 2.9).

Hence, given a basic measurement such as temporal activity, we can then characterize the time series in terms of the predictability along that measurement. Characterization buys us the power to design good algorithms, choose the appropriate models, be it for encoding or for querying or any other analogous problem. The problem of querying, encoding and the like get translated into the problem of finding good measures of activity along the dimension of interest. Moreover, it permits us to research and experiment with different measures and models in an objective manner.

4.1.1 Selection of Spatial and Temporal Activity Measures

Given the context of intelligent encoding of video, we now propose spatial and temporal measures that can be used for this purpose. Any measure that can be used to

compute *mutual information* has to meet the following criterion:

- It must be a point measurement. That is, the measurement must be centered on one point and must be repeatable for every point in the time series. This ensures that we can then use this measure to estimate a probability density function and hence compute *mutual information* and predictability etc.

An image sequence can be visualized as a spatial pattern changing in time. Hence, motion is a good measure of activity in the temporal direction. We recall from section 3.3.1 that optical flow is a good local measurement of motion occurring in the image sequence. Optical flow, being local in scope, is well suited for estimating predictability in the temporal direction. However, it is not a very accurate model of motion for all cases [13]. We choose to use optical flow because of its simplicity, locality and ease of computation. We note here that this approach is not tied to optical flow as the motion measure. We can adopt a better algorithm for measurement and the characterization algorithm would work just as perfectly. In that sense, the characterization algorithm is scalable and modular.

For predictability in the spatial direction, we can similarly choose an activity measure. We choose a measure derived from a Laplacian operator for the experiments detailed

in this thesis. The Laplacian operator can be defined as:

$$L(i, j) = \nabla^2 f_{ij} = \frac{\partial^2 f_{ij}}{\partial x^2} + \frac{\partial^2 f_{ij}}{\partial y^2}. \quad (4.1)$$

The Laplacian operator is a local texture measure and hence can be used to estimate the probability density function for determining the predictability. In this thesis, we implement a modified Laplacian operator which emphasizes small differences and scales down large differences. The rationale behind such an implementation is that large differences are typically because of object boundaries and sharp features in the region. While sharp features contribute to the texture, the smaller variations probably play an important role in texture but often get missed and classified in the same category during quantization. Hence, we emphasize small variations and de-emphasize larger variations prior to quantization. Thus, we hope to capture both local and larger variations in texture.

The point measures that we defined are not the only possible ones. They have been selected because of their analytical tractability, relative ease of computation and locality of scope. All these three factors are essential for a good, pragmatic measure. We can now proceed with the implementation of the Mutual Information Metric. Based on the previous discussion in chapter 2, we can formulate the metric rather simply.

4.1.2 Design of the Mutual Information Metric

Once we have defined the activity measures, we can then design the mutual information metric. For the purposes of intelligent encoding, we would like to split the image sequence into regions that have the optimal spatial and temporal depth and then encode these regions using a suitably designed class of encoders. We will study one such implementation of a hierarchical class of encoders in the next chapter. Every member of this class of encoders looks alike and is derived from the same mother algorithm. They only differ in their spatial and temporal extents. Given a video sequence split into chunks based on their spatial and temporal predictability, this class of encoders can then be commanded to compactly represent these chunks. Thus we not only have the possibility of encoding video efficiently, but also we have split it up into meaningful chunks based on their spatial and temporal activity signature. This would enable querying and indexing as we shall see later.

In the context of a time series, we recall from section 2.2 that the optimal embedding depth or the dimensionality of data depends upon the number of observations and the physical separation between the observations. That is, observations that are far apart do not share any information and hence if our model employs these observations, it will gain nothing. If the observations are close together, then they are highly redundant and we can predict the rest of the observations from just one observation. So, we would like to find the ideal embedding depth where there is sufficient redundancy to

exploit and yet the observations are as far apart as possible. If we take d observations together, then $H_1 \leq H_d(\tau) \leq dH_1$. Hence, if we now compute the quantity

$$NMI = \frac{dH_1 - H_d}{dH_1} \quad (4.2)$$

NMI is the *Normalized Mutual Information* which is a quantity between 0 and 1. To find the suitable window depth in the spatial and temporal directions for encoding, we now compute the NMI of the the two activity measures mentioned earlier, starting with a large window size and reducing the window sizes till NMI exceeds a preset threshold or we reach the lower limit on window sizes. Both the preset threshold and the lower limit on window sizes can be determined from the compression ratio and the quality desired. Other criteria such as bit rate requirement, buffer control etc can also be factored into the computation of these parameters.

Finding the dimensionality of data is a problem of interest to people that seek to model the system. Of the three goals of characterization (Prediction, Forecasting and Modeling), in the context of image sequence coding, we are interested in prediction the most. Hence, the degrees of freedom in the system and the dimensionality of data are not of immediate relevance to us. Hence, in this thesis, we limit the scope of our analysis to finding the optimal embedding depths spatially and temporally. With this end in mind, we choose the parameter d , the number of observations taken together in the estimation of NMI as a parameter to the system and not a quantity

that we will analytically compute. In these experiments, we choose $d = 4$ for spatial activity. This enables to measure horizontal, vertical and diagonal similarities. In the temporal direction, we choose $d = 2$. Since there are two temporal components, this choice enables us to measure similarities along x and y . We see that this level of complexity is within computable limits and also gives us enough information for us to conduct meaningful experiments.

4.2 Experiments Performed on Test Sequences

We will evaluate the performance of these measures on a few test sequences. These will serve as a sanity check on the analysis algorithm.

We start with a simple 256×256 pixel image which remains unchanged in time. This image is shown in figure 4.1. This is the trivial case where we expect the predictability to be very high and the entire volume being classified as one region. The algorithm performs as expected and classifies the entire image sequence as one region. The metric has withstood the first sanity check.

We now take one frame of a video sequence and replicate it in time. We expect it to break up this image sequence in spatial direction alone. The test image is as shown in figure 4.2. The result of the analysis is shown pictorially in figure 4.3. Figure

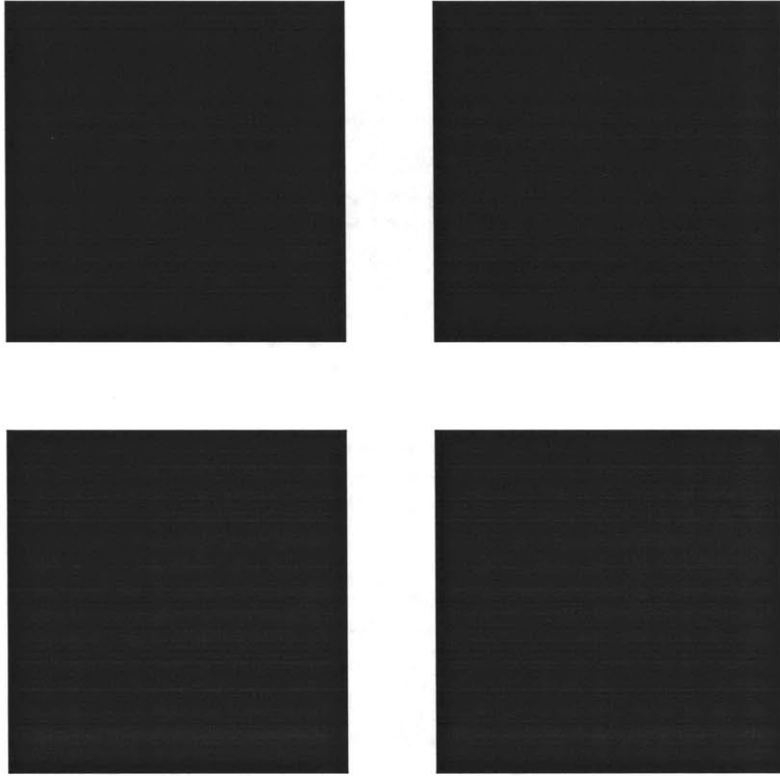


Figure 4.1: One frame from the plain two-color cross image

4.3 indicates that the sequence was analyzed into 16 regions; each region has the same temporal depth as expected. In other words, the algorithm concludes that the image sequence has high temporal predictability and has to be split only in the spatial direction. This is well within our expectations.

We can try the measure on a normal sequence. Unfortunately, we no longer have a quantifiable means of measuring whether the algorithm performs well. We can indirectly measure the success of the algorithm by testing the performance of the



Figure 4.2: One frame from the “sharky’s machine” sequence

class of encoders against a baseline class. We can also build querying applications and see if the characterization has been successful. That will be our next step.

4.3 Use of the Characterization Measure

We now examine the use of this characterization technique in applications such as coding, querying etc. For the purposes of splitting video into regions in x, y and t and



Figure 4.3: Pictorial result of the analysis

encoding, the approach lends itself automatically. The task that remains to be done is to use this information in the design of intelligent coders, better tuned filters etc. We will see one implementation of coder design in the next chapter.

In order to use such a characterization in a useful manner to do querying is a much more interesting question. Querying is a hard problem. It is much more open-ended and nascent in scope. Nevertheless, we can proceed in two possible directions, both potentially fruitful. We can study different basic measures that can be combined with

this characterization to produce meaningful queries. Or, we can use the spatial and temporal activity measures designed for the purposes of coding intelligently to do querying as well. The first approach, although very promising, is difficult. It implies that the measures that need to be designed and studied can embody some of the common knowledge that humans know about video. Such measures can be complex and computationally expensive. This is a top-down approach. The second approach is a bottom-up approach in which we start with simple measures and see if they hold meaning in a broader context. Gaining from what we learn, we can then design better aggregate measures that embody some of what humans know about video and use it for querying and indexing applications. We have attempted to use these characterization techniques for doing indexing and have obtained some interesting results [29]. Since this discussion about querying is out of the scope of this thesis, the reader is encouraged to test the strength of the characterization by accessing the URL <http://tvot.www.media.mit.edu/projects/tvot/Agenda95/VideoBook.html> [29].

In the next chapter, we study the design of a hierarchical coder and examine its performance. We also discuss the possibility of building a universal “society of coders” which is a coder made up of multiple coders.

Chapter 5

Controlled Coding

In chapters 2 and 4, we studied the use of information theory for characterization of digital video. We closed the discussion in chapter 4 with an illustration of the spatio-temporal analysis. This analysis splits the video volume into chunks based on their spatial and temporal predictability. Such an analysis lends itself to design of a class of encoders; each of which can be used to encode a different spatio-temporal region. In addition, this analysis can also be used to tradeoff in the continuum between spatial and temporal encoding methods. That is, in video sometimes texture reveals structure and sometimes motion does it. This analysis helps us objectively measure the amount of structure revealed by texture and motion. The task that remains at hand is to design a class of coders that can exploit the structure thus revealed.

In our analysis, we split the video into chunks which maximize the predictability along two dimensions: spatial and temporal. One possible way to encode these regions would be to choose appropriate texture and motion models for regions thus analyzed. This is in general a difficult problem. As such, the modeling approaches have to mature further before they can be employed effectively in such a coder. Another approach would be to choose appropriate encoders from an existing class of encoders such as MPEG, JPEG, Wavelets etc. This is a practical approach given the present level of technology and can be easily implemented. However, this approach takes us further away from our goal of a family of encoders since each of these coders operate differently and hence would require building an encoder/decoder that can handle all these various types of algorithms. This approach may not be illustrative enough for our purposes.

Hence, for this thesis, we choose yet another approach: a class of encoders, all based on the same mother algorithm, that operate on different spatial and temporal extents. Hence, the encoder now is a class of encoders working in unison but differently on each region. Thus, we achieve controlled coding without undue increase in the encoder/decoder complexity. Moreover, since the encoded bit stream is structured into separate regions, it can be used for querying purposes with very minor additional overhead (estimated at one kilobyte per half a second of video [29]).

5.1 Hierarchical Image Coder

There are many ways of building a class of encoders, each with its own advantages. We choose to build a hierarchically structured class of encoders because of the simplicity and elegance of the approach. In such a class, not only are all the coders related to each other but each encoder is a sub-class of the encoder directly above it. This gives us the bonus advantage of computational scalability. A simple wristwatch decoder can operate at the coarsest level to reproduce the image sequence at a low resolution and a more complex decoder can be built into a living room decoder that has sufficient computational power to reconstruct at high resolution with high fidelity.

In this thesis, we chose to implement a wavelet based hierarchical class of encoders for the following reasons:

- Wavelets are compact and highly efficient to compute.
- They are multi-resolutional by construction and hence offer the potential to design different wavelet bases that have different spatial and temporal extents.
- Unlike MPEG, where motion estimation is difficult computationally, a temporal wavelet giving similar performance is very simple to implement.

Filter bank theory shows that 1D wavelets can either be orthogonal or symmetric. We prefer symmetry over orthogonality because symmetry results linear phase filters.

The human eye is very sensitive to phase distortions [27] and hence linear phase filters are desirable. However, non-orthogonal filter banks cannot satisfy the perfect reconstruction property. That is, in theory they cannot perfectly reconstruct a decomposed signal. However, in the case of lossy compression where quantization is inevitable, perfect reconstruction is never possible. Therefore, the lack of perfect reconstruction may not be a serious handicap.

Adelson and Simoncelli have demonstrated that it is possible to design almost orthogonal symmetric filters [26, 28]. These filters deviate from orthogonality almost at the numerical precision limit of the computer. Hence, for all practical purposes, we have an orthogonal filter bank. For this thesis, we chose the 9-tap filter [28]. This filter has an error in the range of e^{-17} and is also very compact. The filter has a complexity that is lower than that of DCT that is used in MPEG. Figure 5.1 shows the time domain response of the 9 tap filter. The high pass filter is related to the low pass filter by a simple shift and multiplication by $(-1)^n$. The remarkable feature of these filters are that they are very close to being orthogonal and symmetric. By design, both the analysis and the synthesis filters are identical.

We decided to use different filters for spatial and temporal directions because of the vast difference in statistics in these two dimensions. The Haar wavelet, though does not possess good filter properties, performs exceedingly well as a temporal filter. The Haar wavelet is exceedingly simple to implement (it requires no multiplies) and has

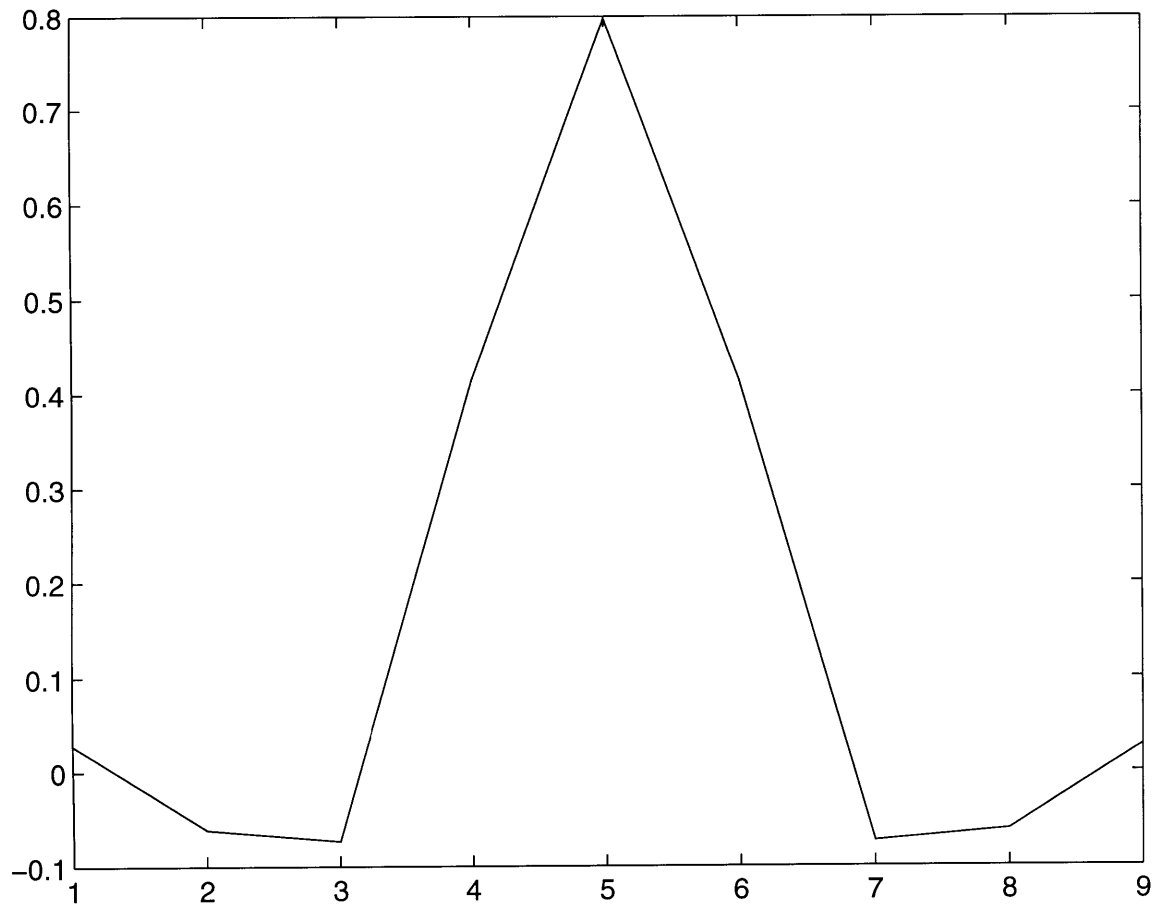


Figure 5.1: Time domain response of Adelson's 9 tap QMF

negligible complexity compared to that of the different motion estimation algorithms used in MPEG. Figure 5.2 shows the low pass and the high pass Haar filters. They are just simple sum and differences.

These two filters form the foundation of the mother algorithm for our hierarchically structured class of encoders. Every encoder in this class differs from one another only in the number of iterations that it performs in the spatial and temporal directions.

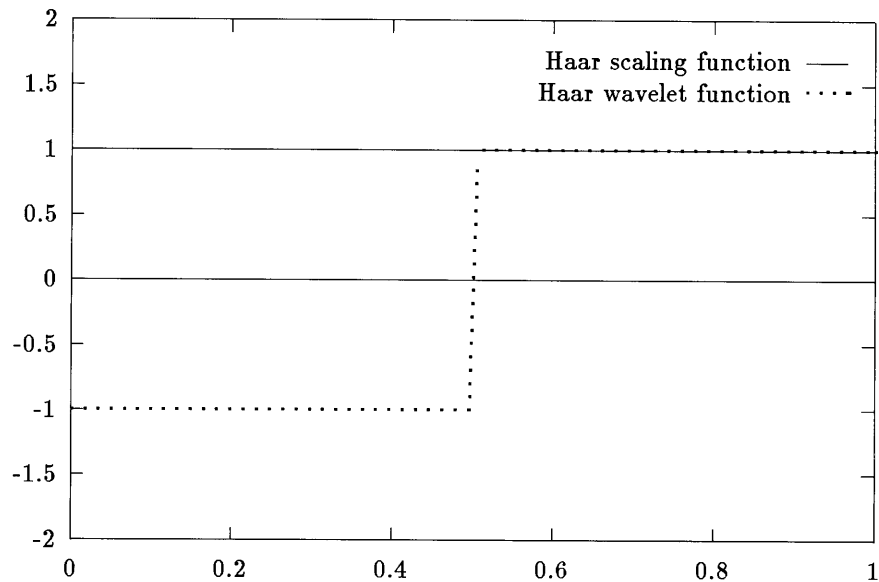


Figure 5.2: The Haar scaling and wavelet functions

These iterations determine the effective lengths of the resulting basis functions. We use larger basis functions for large regions and small basis functions for small regions.

For quantization, one of the crucial steps in the design of any encoder, we used a non uniform quantization scheme. As in DCT, the higher frequency coefficients are quantized more than the low frequency coefficients. This non-uniform quantization results from perceptual studies that the human eye is less sensitive to high frequency information and hence the high frequency information can be quantized more severely. The quantization table is illustrated in figure 5.3. Each encoder run length encodes the quantized data followed by an adaptive Huffman encoding to further reduce the redundancy in the bitstream.

5.2 Performance Evaluation of the Controlled Coder

We used approximately 1 second worth of video data in each of our controlled coding experiments. There are 4 main steps in this experiment. First we compute the spatial and temporal activity in the video sequence. These are computed using the optical flow and the non-linearized Laplacian operator as explained in chapter 4. Once the activities are measured, we compute the optimal spatial extent and then the temporal extent. We note that this is just one of the many approaches in which the spatio-temporal predictability analysis can be conducted. Once the optimal spatial and temporal extents are computed as described in chapter 4, we determine the set of coders that will be employed in encoding this chunk of video data. The video data is then split into chunks and encoded using specifically designed coders. For example, for small regions, we use the encoder that has shorter basis functions whereas the larger regions are encoded with the coder with both large and small basis functions.

5.2.1 Experiment

For comparison, we use a baseline coder that is just one coder derived from the mother algorithm. This coder spans the entire spatial and temporal extent of the data. That is, the entire 1 second worth of video data is treated as one big chunk. In all other aspects, the baseline coder and the hierarchical class of coders are similar. They are

both derived from the same mother algorithm. The hierarchical class of coders have a higher overhead since they have to transmit additional region information for proper decoding. However, if the ultimate aim is to not just efficiently represent but also use the characterized information as an indexing aid, this additional overhead is justified.

Figures 5.4 - 5.9 illustrate the different original frames and the reconstructed frames. We can see that since the hierarchically structured coder tends to localize the encoding, it results in smaller ringing effects. Thus it results in slightly higher subjective quality. The baseline coder and the hierarchical coders are similar. The baseline coder operates on the entire image and a different hierarchical coder operates on different regions of the image sequence with different spatial and temporal extents of the basis functions.

5.2.2 Discussion of results

The performance evaluation demonstrates that the hierarchically structured class of coders and the baseline case of a single coder operating on the entire image have similar performance in terms of bit rate and output quality. The additional advantage that we gain from using the hierarchical coder is the fine degree of control that it gives us over the bit stream added with the possibility of embedding content information. We can use this coder to prioritize packets in a limited network situation so that

in case the packets are to be dropped the decoder can perform a reasonable job of approximating from statistical information that it knows about the video. In addition, this coder brings us a step closer to the goal set out in the introduction. We now have a more intelligent bit stream that not only represents video compactly but it also permits additional embedding of content information for context based querying [29]. Thus, we have a society of coders that operate in unison. Each of the individual coders may not be an ideal encoder but together they are well suited for a diverse class of video sequences. In addition they offer robustness, scalability and embedding of content information. We could have very well used different algorithms on different regions of the image sequence, each algorithm well suited to exploit the characteristics of the region that it encodes. This would truly be a “society of coders”.

2	4	8	16
4	6		
8		12	
16			24

Figure 5.3: Quantization table of the mother algorithm



Figure 5.4: Frame 0 of original test sequence 1



Figure 5.5: Same frame reconstructed using the baseline coder. Compression ratio 100:1



Figure 5.6: Same frame reconstructed using the hierarchical coder. Compression ratio
100:1



Figure 5.7: Frame 32 of original test sequence 2



Figure 5.8: Same frame reconstructed using the baseline coder. Compression ratio 20:1



Figure 5.9: Same frame reconstructed using the hierarchical coder. Compression ratio 20:1

Chapter 6

Summary and Future Work

6.1 Results of the research work

In the transition to digital video, a new benefit of digital television that has been underexplored surfaces: the ability to add image reconstruction information and correlated descriptive data. This thesis exploited that characteristic to allow a set of individual coders to all cooperatively analyze, compress and represent video sequences.

In order to achieve this goal, a methodology for analyzing and characterizing digital video using information theoretic measures was designed, studied and implemented. This method analyzes video and splits it into optimal spatial and temporal chunks that can then be compressed with a suitable encoder. Thus, we achieve compact

representation and added descriptive information in one shot. This descriptive information has been successfully used in a querying application [29] with encouraging results.

This analysis method was then used to design a hierarchical class of wavelet encoders. These coders are compact and efficient. Their performance was compared with a baseline coder and was found to be satisfactory with the added bonus of descriptive information being sent with encoded bit stream. The hierarchical coders have a superior subjective performance. In addition, by design these coders have a higher tolerance to errors.

The analysis methodology is independent of the underlying encoding algorithm used. Hence, it can be easily used in conjunction with any of the existing coding algorithms to build efficient and content dependent coders.

6.2 Discussion

This work raised several interesting issues. One immediate application is the pursuit of automated querying mechanism in a much smaller parameter space [29]. Preliminary tests provide encouraging results for such low level mechanisms. With minor extensions, the encoded bit stream can piggy back this parameter space information;

which may then be used for later content-based retrieval. This raises interesting possibilities in terms of interactivity. Moreover, the age of standard algorithms may well be past. Such an approach permits the possibility of intelligently choosing multiple algorithms to represent video compactly with the added bonus of being able to use this bit stream for not just compact representation but other potentially richer goals such as querying, selecting etc. The work clearly demonstrates that there is much to be learned from looking at the statistics of an image sequence. It also amply demonstrates the potential benefits of moving from a restrictive single universal coding standard to a society of video representation schemes.

6.3 Future Work

The next logical step to pursue would be to study the power of this approach for content based querying a video database. Some preliminary work has already been done in this area and we report very encouraging progress [29]. Even such a preliminary study indicates that there is lot to be learned. There is sufficient indication that information theoretic measures and time series mechanisms capture some of the intrinsic nature of a system. However, if we have to truly understand time series such as video, we have to develop techniques that allow us to move from the limited horizons that *prediction* offers to the much richer knowledge offered by *forecasting*

and *modeling*. Such studies have barely begun for 1-dimensional complex systems [5]. However, 2-dimensional time series unfold a totally new realm of complexity. Some of the classic techniques that work very well for 1-dimensional systems fail badly for 2-dimensional systems. Therefore, there remains a whole lot of work to be done in this exciting area both in the near term (e.g. better prediction) and in the long term (e.g. forecasting and modeling). Information theory combined with knowledge from vision studies will surely go a long way in setting stage for things to come.

Bibliography

- [1] N. Vasconcelos, “Library based image coding using vector quantization of the prediction space,” *Master’s thesis, Program in Media Arts and Sciences, MIT*, Cambridge MA, September 1993.
- [2] Richard E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, Reading MA, 1987.
- [3] Claude E. Shannon, “The mathematical theory of communication,” *The Mathematical Theory of Communication*, University of Illinois Press, Chicago, 1949.
- [4] Andrew J. Viterbi and Jim K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, New York, 1979.
- [5] A. S. Weigend and N. A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison Wesley, Reading MA, 1993.
- [6] ISO/IEC DIS 10918-1, ITU-T Rec.T.81 (*JPEG*) “Information Technology - Digital compression and coding of continuous tone still images”, 1992.

- [7] ISO/IEC 11172 (1993) “Information technology - Coding of moving picture and associated audio for digital storage media as up to about 1.5 Mbit/s”.
- [8] N. S. Jayant and Peter Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, New Jersey, 1984.
- [9] Ingrid Daubechies, “Orthonormal bases of compactly supported wavelets,” *Comm. Pure and Appl. Math*, Vol. 41, pp909-996, 1988.
- [10] Charles K. Chui, *An Introduction to Wavelets*, Academic Press, San Deigo, 1992.
- [11] Gilbert Strang and Trong Nguyen, *Wavelets and Filter Banks*, Wellesley Press, Wellesley, 1995.
- [12] Stephane Mallat, “Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$,” *Trans. Amer. Math Soc.*, Vol. 315, pp69-87, 1989.
- [13] Berthold K. P. Horn, *Robot Vision*, The MIT Press, Cambridge, 1986.
- [14] J. Barron, D. Fleet, S. Beauchemin, T. Burkit, “Performance of Optical Flow Techniques”, *University of Western Ontario, Dept. of Comp. Sci. TR 299*, 1992.
- [15] L. Teodosio, “Salient Stills”, *Master’s thesis, Program in Media Arts and Sciences, MIT*, Cambridge MA, June 1992.
- [16] P. McLean, “Structured Video Coding”, *Master’s thesis, Program in Media Arts and Sciences, MIT*, Cambridge MA, June 1991.

- [17] Edward H Adelson and John Y A Wang, "Representing moving images with layers," *MIT Media Laboratory, Perceptual Computing Group, Technical Report #228*, Cambridge MA, April 1993.
- [18] H. G. Musmann, M. Hotter, J. Ostermann, "Object-Oriented analysis-synthesis coding of moving images," *Signal Processing: Image Communication*, vol 1, pp 117-138, Elsevier Science Publishers, 1989.
- [19] R. Kermode, "Building the Big Picture: Enhanced Resolution from Coding", *Master's thesis, Program in Media Arts and Sciences, MIT*, Cambridge MA, June 1994.
- [20] A. B. Lippman and R. Kermode, "Generalized Predictive Coding of Movies", *Picture Coding Symposium (PCS'93)*, Lausanne, Switzerland, March 1993.
- [21] E. Chalom, V. Bove, Jr., "Segmentation of frames in a video sequence using motion and other attributes", *To be presented in Digital Video Compression: Algorithms & Technologies 1995 Conference*, San Jose, CA, February 1995.
- [22] M. Minsky, *The Society of Mind*, Simon and Schuster, 1988.
- [23] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston MA, 1991.
- [24] A. N. Netravali and B. G. Haskell, *Digital Pictures*, Plenum Press, New York 1988.

- [25] C. Blakemore, *Vision: Coding and Efficiency*, Cambridge University Press, 1990.
- [26] Edward H. Adelson and Eero P. Simoncelli, "Subband Image Coding with Threetap Pyramids," *IEEE Picture Coding Symposium*, Cambridge MA, 1990.
- [27] A. V. Oppenheim, A. S. Willsky and Ian T. Young, *Signals and Systems*, Prentice-Hall, Englewood Cliffs, 1983.
- [28] Eero P. Simoncelli, "Orthogonal sub-band image transforms," *Master's thesis*, M.I.T. Cambridge MA, May 1988.
- [29] Giridharan Iyengar, "VideoBook: An experiment in characterization of video," *MIT Media Lab Internal Report*, June 1995.

Acknowledgments

So many people, so little space ...

First, thanks to Andy Lippman, my advisor, for the good ideas and direction, and an unique perspective of research and life; to the thesis readers, Michael Hawley and Jules Bellisio, for the patience to read it and the comments they provided; to everybody in the garden for making this a fun place to work; and to Gillian, Lena, Linda, Santana and Celia for all their help and support.

Special thanks to all my family and friends for the love and support that made my stay away from home much easier: For all the long distance morale boosting and support through difficult times. For all their delicately concocted Indian spices to give that unique taste of home.

Finally a very special thanks to Sri, the real strength behind the success of my stay abroad. Thank you, Sri, for being there every single moment. I am truly lucky.