UNIVERSITY OF CALIFORNIA, SAN DIEGO

Packet Loss Visibility and Packet Prioritization in Digital Videos

A Dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy

in

Electrical and Computer Engineering
(Communication Theory and Systems)

by

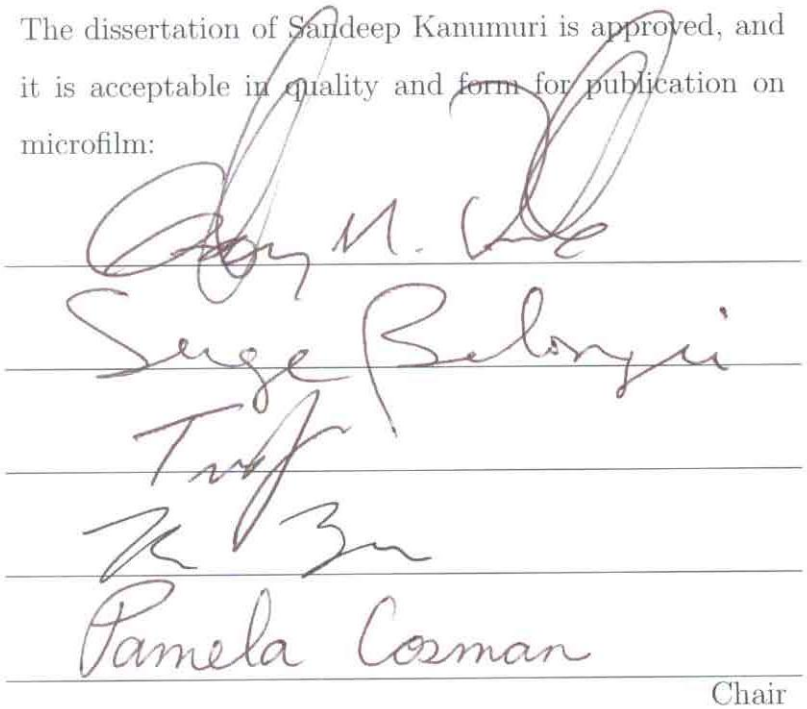
Sandeep Kanumuri

Committee in charge:

Professor Pamela C. Cosman, Chair
Professor Serge J. Belongie
Professor Truong Q. Nguyen
Professor Geoffrey M. Voelker
Professor Kenneth Zeger

2006

The dissertation of Sandeep Kanumuri is approved, and it is acceptable in quality and form for publication on microfilm:



Chair

University of California, San Diego

2006

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the people who have stood by me during the past four years and have helped me accomplish this work.

First, I would like to thank my advisor, Prof. Pamela Cosman, for her constant support and guidance through out these years. She is a great source of motivation and has helped me achieve goals that, I thought, are impossible. She is an excellent teacher and researcher and I thank her for all our discussions which shaped my research and helped me immensely. In spite of her busy schedule, she was very patient and always had the time to meet with me whenever I wanted to. I truly admire the way she balances work and life.

I would like to thank my unofficial advisor, Dr. Amy Reibman from AT&T Labs, for being an excellent mentor in guiding my research. The credit for my research topic goes to her, who introduced me to this topic while I was interning at AT&T Labs. I would like to specially thank her for writing numerous long and detailed e-mails over all these years, which was instrumental in making our long distance collaboration a success.

I would like to thank my dissertation committee members, Prof. Truong Nguyen, Prof. Kenneth Zeger, Prof. Serge Belongie and Prof. Geoffrey Voelker, for their invaluable time and advice. I should also thank Prof. Nguyen for his excellent treatment of wavelets and filter banks, Prof. Zeger for a rigorous foundation in source coding and Prof. Belongie for a comprehensive course in computer vision. I would like to thank Prof. Charles Berry for his guidance with statistical methods.

I am also grateful to Dr. Vinay Vaishampayan, Dr. Svetislav Maric and Ms. Maria Marshall for their advice and support during my internships.

I am also thankful to my fellow researchers in the Information and Coding Laboratory for creating a very friendly atmosphere. Special thanks to Thanos, Yushi, Ramesh, Sitaraman, Solmaz and Mayank for being great colleagues and friends. I am also thankful to all my friends in San Diego for a pleasant experi-

ence. I would like to thank my undergraduate and junior college friends for the encouragement they have provided over the years.

Finally, I would like to thank the people who had the greatest impact on my life. I express my deepest gratitude to my mother Ramadevi Kanumuri and my father Venugopal Naidu Kanumuri for the sacrifices they have made so that my sister and I can get the best education possible. I am very lucky to have Swetha as my sister whose unwavering confidence in me keeps me going. I am highly thankful to my brother-in-law Ravi for his encouragement and support during difficult times.

Chapter II of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, P. C. Cosman, A. R. Reibman and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video", *IEEE Trans. Multimedia*, vol. 8, pp. 341-355, April 2006. I was the primary author and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter II. The co-author Dr. Vaishampayan also contributed to the ideas in this work.

Chapter III of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, P. C. Cosman, A. R. Reibman and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video", *IEEE Trans. Multimedia*, vol. 8, pp. 341-355, April 2006. I was the primary author and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter III. The co-author Dr. Vaishampayan also contributed to the ideas in this work.

Chapter IV of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, S. G. Subramanian, P. C. Cosman and A. R. Reibman, "Packet Loss Visibility and Packet Prioritization in H.264/AVC Videos", *IEEE Trans. Image Processing* (submitted). Co-author Subramanian and I contributed equally towards this publication. Co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter IV.

Chapter V of this dissertation, in part, is a reprint of the material as it appears in S. Kanumuri, P. C. Cosman and A. R. Reibman, "Source-dependent Video Packet Prioritization based on a Visibility Model", *IEEE ICASSP*, April 2007 (submitted). I was the primary author and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter V.

VITA

1981	Born, Tirupati, Andhra Pradesh, India
June 2002	B.Tech., Electrical Engineering Indian Institute of Technology, Madras, India
Jan 2003–Dec 2006	Research Assistant, University of California, San Diego
Summer 2003	Intern, AT&T Labs-Research, Florham Park, New Jersey
Apr 2004	M.S., Electrical and Computer Engineering (Communication Theory and Systems) University of California, San Diego
Summer 2005	Intern, Qualcomm Inc, San Diego, California
Dec 2006	Ph.D., Electrical and Computer Engineering (Communication Theory and Systems) University of California, San Diego

PUBLICATIONS

S. Kanumuri and A.N. Rajagopalan, "Human face detection in cluttered color images using skin color and edge information," *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'02)*, Ahmedabad (India), Dec. 2002.

A. R. Reibman, S. Kanumuri, V. Vaishampayan and P. C. Cosman, "Visibility of individual packet losses in MPEG-2 video," *IEEE International Conference in Image Processing (ICIP)*, pp. 171-174, Singapore, October 2004.

S. Kanumuri, P. C. Cosman and A. R. Reibman, "A generalized linear model for MPEG-2 packet-loss visibility," *Packet Video Workshop*, Irvine, December 2004.

S. Kanumuri, P. C. Cosman, A. R. Reibman and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video," *IEEE Transactions on Multimedia*, vol. 8, pp. 341-355, April 2006.

S. Kanumuri, S. G. Subramanian, P. C. Cosman and A. R. Reibman, "Predicting H.264 Packet Loss Visibility using a Generalized Linear Model," *IEEE International Conference in Image Processing (ICIP)*, pp. 2245-2248, Atlanta, Georgia, October 2006.

S. Kanumuri, P. C. Cosman and A. R. Reibman, "Source-dependent video packet prioritization based on a visibility model," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, April 2007 (submitted).

S. Kanumuri, S. G. Subramanian, P. C. Cosman and A. R. Reibman, "Packet Loss Visibility and Packet Prioritization in H.264/AVC Videos," *IEEE Transactions on Image Processing* (submitted).

FIELDS OF STUDY

Major Field: Electrical Engineering

Studies in Communication Theory and Systems.
Professor Pamela C. Cosman.

Studies in Signal Processing.
Professor A. N. Rajagopalan, Indian Institute of Technology, Madras, India.

ABSTRACT OF THE DISSERTATION

Packet Loss Visibility and Packet Prioritization in Digital Videos

by

Sandeep Kanumuri
Doctor of Philosophy in Electrical and Computer Engineering (Communication Theory and Systems)
University of California, San Diego, 2006
Professor Pamela C. Cosman, Chair

Traditional approaches to video quality assume that all packet losses affect quality equally. In reality, different packet losses have different visual impact and not all packet losses are visible to the average human viewer. The problem of evaluating video quality given packet losses is quite challenging due to the varying impact of packet losses. As a first step towards developing a quality metric for video affected by packet losses, we address the problem of predicting packet loss visibility. Visibility of a packet loss refers to the visibility of artifacts in the video caused by that packet loss.

We consider the problem of predicting packet loss visibility as a regression as well as a classification problem. In the regression problem, the goal is to predict the probability that the packet loss causes a visible artifact. In the classification problem, the goal is to classify each packet loss as visible (invisible) if a visible artifact occurred (did not occur) in the video due to that packet loss. A subjective test is conducted to gain ground truth on visibility of 1080 individual packet losses in MPEG-2 videos. We explore the usefulness of various factors in predicting visibility, which are extracted using measurements from either the entire encoded video, the decoded video pixels, or just the received lossy bitstream. We model the probability of visibility of a packet loss using a Generalized Linear Model

(GLM). We design classifiers to solve the classification problem using a well-known statistical tool called Classification and Regression Trees (CART).

Video transmission over internet or wireless links is typically characterized by bursty packet losses and not individual packet losses. So we generalize the concept of visibility to multiple losses. A multiple loss is defined as a set of L individual packet losses occurring in close temporal proximity. To obtain ground truth, a new subjective test is conducted using H.264/AVC bitstreams instead of MPEG-2 bitstreams. Further, motion-compensated error concealment (MCEC) is used to conceal the packet losses instead of the zero-motion error concealment (ZMEC) which is used earlier. Because of these differences, 2160 individual packet losses are also introduced along with 3240 multiple losses in this subjective test and the visibility of both types of losses is modeled. We introduce new factors that are useful in predicting visibility. The relative importance of these factors is ascertained through statistical modeling. The effect of different factors on packet loss visibility is also analyzed. A new model framework is introduced for predicting the visibility of multiple packet losses and its performance is demonstrated on dual losses (two packet losses occurring together).

One of the applications for the knowledge of packet loss visibility is packet prioritization. We demonstrate the effectiveness of our visibility model for this application. We consider a transmission scenario where packets are dropped at a congested node in the network. A new packet prioritization method is proposed that assigns a priority level to each packet using our visibility model. We show that a priority-based packet drop policy outperforms a conventional DropTail policy in terms of received video quality.

I

Introduction

When sending compressed digital video across today's communication networks, packet losses may occur. Network service providers would like to provision their network to keep the packet loss rate below an acceptable level, and monitor the traffic on their network to assure continued acceptable video quality. Traditional approaches to video quality assume that all packet losses affect quality equally. In reality, different packet losses have different visual impact. For example, one may last for a single frame while another may last for many; one may occur in the midst of an active scene while another is in a motionless area. Not all such packet losses are visible to the average human viewer. Thus, the problem of evaluating video quality given packet losses is challenging. As a first step towards developing a quality metric for video affected by packet losses, we address the problem of predicting packet loss visibility. Visibility of a packet loss refers to the visibility of artifacts in the video caused by that packet loss.

Predicting the visibility of a packet loss can be looked at as either a regression or classification problem. In the regression problem, the goal is to predict the probability that the packet loss causes a visible artifact. In the classification problem, the goal is to classify each packet loss as visible (invisible) if a visible artifact occurred (did not occur) in the video due to that packet loss. Predicting the visibility of a packet loss serves as a useful tool in the following ways:

- Packet Prioritization: When the buffer in a network node becomes full in the event of congestion, the node discards packets. If one assigns priorities to each of the packets at the encoder based on the packet's loss visibility, the packets with lower priority can be discarded. This is expected to improve the quality of the video perceived at the decoder. We will explore this application in Chapter V.

- Perceptual error control using unequal error protection: Packets which are perceptually important (i.e. packets which, when lost, cause significant visual impact on the viewer) can be given more error protection than ones which are perceptually less important. This way, if a perceptually significant packet is corrupted, it is more likely to get corrected by the channel decoder than a perceptually insignificant packet.

- Network Quality Monitor: Knowledge of packet loss visibility will be very useful in building an accurate, real-time network quality monitor. A network quality monitor is a necessary tool for network service providers in implementing applications such as video services with cost-based quality.

I.A Classification of Quality Assessment Methods

From a Quality Monitor's perspective, there are four different types of quality assessment methods, depending on the amount of information available. These methods are illustrated in Figure I.1.

- If the quality monitor has access to the decoder's reconstructed video (with losses) as well as the encoder's reconstructed video, then it is called a Full-Reference (FR) Method.

- If the quality monitor has access to the decoder's reconstructed video (with losses) as well as some factors extracted from the encoder's reconstructed video, then it is called a Reduced-Reference (RR) Method.

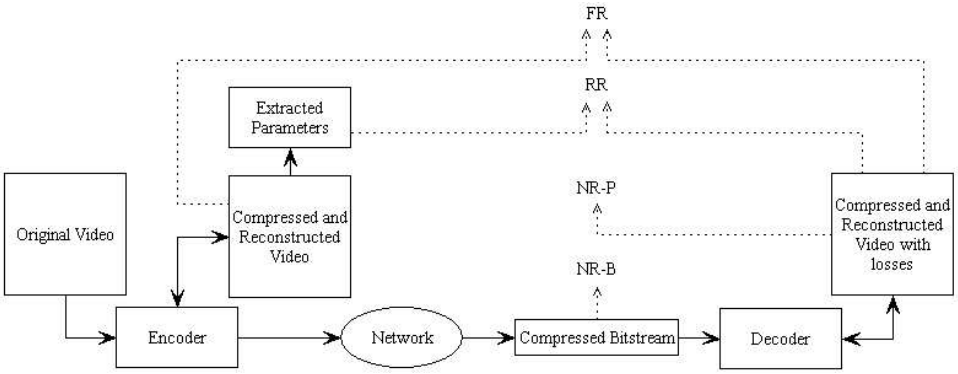


Figure I.1: FR, RR, NR-P and NR-B methods

- If the quality monitor has access to only the decoder's reconstructed video (with losses), then it is called a No-Reference Pixel-Based (NR-P) Method.
- If the quality monitor has access to only the compressed bitstream information (with losses), then it is called a No-Reference Bitstream-Based (NR-B) Method.

In our research, we explore RR, NR-P and NR-B methods for predicting the visibility of packet losses.

I.B Literature Survey

Considerable research has been done on developing objective perceptual metrics for characterizing compression artifacts [1]-[12]. For example, a metric based on a multichannel model of the human visual system was proposed in [1] that generates continuous estimates of perceived quality for low bit rate compressed video. Gastaldo et al. [2] used a Neural Network approach to design an objective quality assessment algorithm for MPEG-2 video streams without decoding. While these metrics are designed to predict the quality degradation caused by compression artifacts, they are not tailored to handle the degradation caused by network impairments.

The joint impact of encoding rate and ATM cell losses on MPEG-2 video quality was studied in [13, 14]. They show the existence of an optimal coding rate for a given loss ratio. A similar study [15] on QoS performance of end-to-end video transmission also showed that an increase in video bit-rate may improve video quality only when cell loss ratio is below a certain level. In both cases, the quality of video is judged based on an existing picture quality model and not based on subjective tests. A framework for employing objective perceptual quality assessment methods, evaluating the quality of audio, video and multimedia signals, to model network performance is demonstrated in [16]. In their paper, they focus on modeling network performance of multiplexed VOIP calls using the perceptual analysis approach.

Much of the effort to understand the visual impact of packet losses [17]-[20] has focused on the average quality of videos subjected to average packet loss rates (PLR). Video conferencing was studied in [17] using the average judgment of consumer observers on the relative importance of bandwidth, latency and packet loss. The impact of packet loss on the Mean Opinion Score (MOS) of real-time streaming media was studied in [18] for Microsoft Windows Media encoder 9 (beta version) video. A neural network was trained in [19] to viewer responses on the ITU-R 9-point quality scale, when a single 10-second sequence was subjected to different bandwidth, frame rate, packet loss rate, and I-block refresh rate.

Hughes et al. [20] used MOS to evaluate the subjective quality of VBR video subjected to ATM cell loss over a 10-second period. They showed that performance is sensitive not only to the magnitude of the bursts, but also to their frequency. "Very different" results were obtained for different sequences. Other challenges identified by these authors [20] were: (a) many different realizations of both packet loss and video content are necessary to reduce the variability of viewer responses; (b) very low PLRs are difficult to explore because the typical test period (10 seconds) is so short that typical realizations may have no packet losses; (c) the "forgiveness effect" causes viewers to rank a long video based on more recently

viewed information.

In [21], two different subjective testing procedures, namely Single Stimulus Continuous Quality Evaluation (SSCQE) and Double Stimulus Impairment Scale (DSIS), were compared. The first procedure shows one stimulus to the subjects, the second two. The data obtained with these procedures was found to be highly correlated and of comparable prediction accuracy. Further, blockiness, blurriness, and jerkiness metrics were not able to accurately predict viewers' subjective responses to packet losses.

In [22], subjective tests were conducted to validate the usefulness of an existing spatio-temporal model for predicting quality in the presence of packet losses. Both one- and two-layer encodings were studied. According to the study, the model examined did not have a significant advantage over PSNR.

Typically, these studies [17]-[22] used subjective tests to evaluate quality using MOS. However, the MOS quality rating methodology has a number of difficulties, as detailed in [23]. First, the impairment (or quality) scales are generally not interpreted by subjects as having equal step-size, and labels in different languages are interpreted differently. Second, subjects tend to avoid the end-points of the scales. Third, the term "quality" itself is actually not a single variable, but has many dimensions.

Instead of asking viewers to respond with a scaled rating (ie, MOS), viewers in recent subjective tests have been asked simpler questions. For example, in [24, 25], viewers were asked to indicate when an artifact was visible. In [26] and [27], viewers were asked to adjust the artifact strength until it became just visible. In these studies, the artifacts were imposed on natural, not synthetic, images and videos. Answers from the subjective viewers were then analyzed to obtain a deeper understanding of the factors that affect visual quality [24]-[27]. The subjective tests in our research are designed with similar motivation.

I.C Thesis Outline

In Chapter II, the visibility of individual packet losses in MPEG-2 videos is considered as a regression problem. The subjective test, conducted to obtain ground truth on visibility of packet losses, is described. The factors that might be useful in modeling the visibility are also described. A generalized linear model (GLM) is used to predict the probability that a packet loss causes a visible artifact.

In Chapter III, the visibility of individual packet losses in MPEG-2 videos is considered as a classification problem. The Classification and Regression Trees (CART) algorithm is used to classify packets as visible or invisible. A method for using the GLM models, developed in Chapter II, to design a classifier is also described and its performance is compared to the CART-based classifier.

In Chapter IV, the visibility of individual and multiple packet losses in H.264/AVC videos is modeled. New factors that exploit the advanced features of H.264/AVC are introduced. In addition, new factors based on spatial and temporal coherence of motion, spatial clutter, contrast and side match distortion are introduced. The relative importance of these factors is ascertained through statistical modeling. The effect of different factors on packet loss visibility is also analyzed. A new model framework is introduced for predicting the visibility of multiple packet losses and its performance is demonstrated on dual losses (two packet losses occurring together).

In Chapter V, a new packet prioritization method is proposed that assigns a priority level to each packet using a visibility model developed in Chapter IV. A transmission scenario where packets are dropped at a congested node in the network is considered and a priority-based packet drop policy is compared to a conventional DropTail policy in terms of received video quality.

In the Conclusions section, we enumerate our contributions in this dissertation for each individual chapter, and discuss future work. We note that partial conclusions are also given at the end of each individual chapter.

II

Regression Problem

In this chapter, we address the regression problem wherein we predict the probability of visibility of individual packet losses in MPEG-2 videos. Ground truth data is gathered using subjective tests for a total of 1080 packet losses over 72 minutes of video. Visibility is modeled using a generalized linear model (GLM), whose input consists of factors that can be easily extracted from the video near the location of the loss.

This chapter is organized as follows: Section II.A gives an overview of MPEG-2 packet losses and their impact. Section II.B describes our subjective test. Section II.C describes the logistic regression model, the GLM which suits our purpose. Section II.D describes the objective factors that we believe should be included in our models. Section II.E describes our modeling results and Section II.F concludes.

II.A Effect of a Packet Loss

Video is typically packetized in one of two ways: it can be segmented and packetized into small fixed-size packets (such as ATM cells or MPEG-2 Transport Stream packets), or a variable-sized packet can contain one or more slices. In both cases, a packet loss will cause the loss of one or more slices. Typical scenarios for fixed-size packetization are (a) a packet contains part of one slice, (b) a packet

Figure II.1: Single slice, Double slice and Frame losses

contains the end of one slice and the beginning of another, and (c) a packet contains a frame header. These will cause the loss of (a) one slice, (b) two slices, and (c) an entire frame as shown in Figure II.1. Therefore, we focus on exploring the impact of these three situations.

The initial error induced by a packet loss depends on the error concealment strategy used by the decoder. A typical concealment strategy, used in this first experiment, is zero-motion error concealment (ZMEC) shown in Figure II.2, in which a lost macroblock is concealed using the macroblock in the same spatial location from the closest reference frame. In this case, the initial error is simply the difference between the current encoded frame and the closest reference frame for the affected macroblocks.

The initial error caused by a packet loss propagates in space and time as a result of the video decoding algorithm. An example of error propagation is shown in Figure II.3, where an error in Frame N+1 propagates to Frame N+2. The exact error due to the packet loss can be completely described by (a) the initial error for each macroblock in the lost packet, (b) the macroblock type, and (c) motion

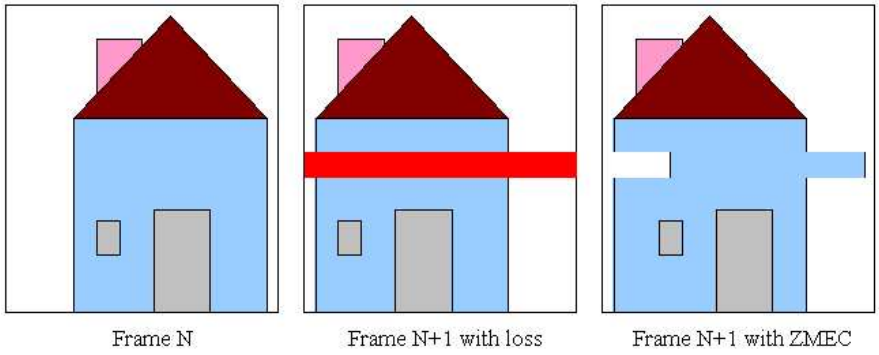


Figure II.2: Zero-motion error concealment (ZMEC)

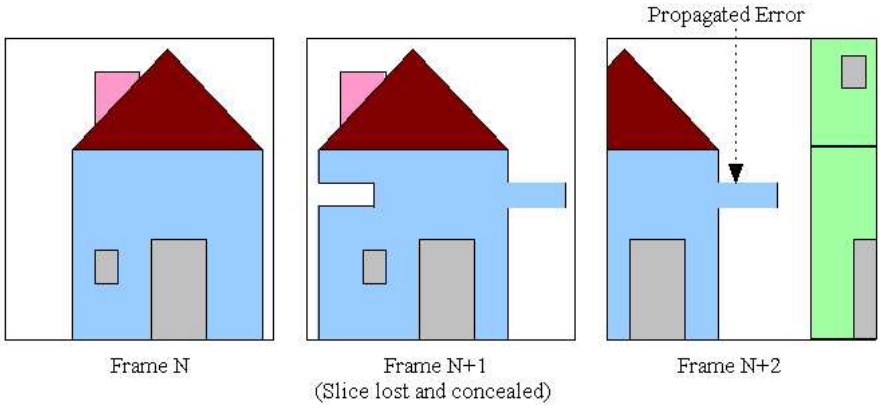


Figure II.3: Error Propagation

information for subsequently received macroblocks [28]. The latter two control the temporal duration and spatial spread of the error.

We expect the visibility of a loss to depend on a complex interaction of its location, the video encoding parameters, and the underlying characteristics of the video signal itself. For example, the texture and motion of the underlying signal may potentially mask the error. To isolate the impact of the various parameters, one approach could be to inject different error amplitudes against an identical signal background, as was done in [25] for blocky, blurry and noisy artifacts. However, for packet losses, the error itself is highly correlated with the underlying signal and so we do not have control over the amplitude of the error. Therefore, we must take a different approach.

When choosing the packet losses to inject for our subjective tests, we have independent control over the location, initial spatial extent and temporal duration of each loss we inject. The other factors depend on the signal. Thus, we choose whether to lose a single slice, double slice or an entire frame. We also choose the loss to be in a B-frame (which would last a single frame) or in a reference frame (which will last until the next I-frame). In choosing the vertical location of the loss, we uniformly distribute the losses within the frame.

II.B Subjective Tests

For the subjective tests, we can conduct either a single-stimulus test or a double-stimulus test. In a single-stimulus test, only the video being evaluated (here, video with packet losses) is shown. The reference or original video is not shown. In a double-stimulus test, both videos are shown. We conducted a single-stimulus test because the test mimics the perceptual response of a viewer who does not have access to the original video, which is a natural setting for most applications. The viewer bases his/her judgment on the lossy video only.

In the test, the viewers' task is to indicate when they saw an artifact,

where an artifact is defined simply as a glitch or abnormality. We wanted viewers to be immersed in the viewing process and not scrutinizing the video for any possible impairment. Thus we chose DVD-quality MPEG-2 video from travel documentaries. Audio was not presented. The video sequences had a resolution of 720×480 pixels and had 30 frames per second. The average bitrate for the sequences varied from 3.5 Mbps to 4.4 Mbps. In our encoding structure, we had two consecutive B-frames between reference frames and we had an I-frame every 13 frames. ZMEC using the closest reference frame was used whenever there was a packet loss. This presumes a minimum amount of intelligence on the part of the decoder. Decoders that use sophisticated error concealment methods may have fewer visible packet losses. However, since we would like to predict the visibility of packet losses in the network, without necessarily knowing which decoder the viewer is using, we assume only this minimal error concealment strategy.

The video sequences we chose contain a wide variety of scenes with different types of camera motion (panning, zooming) and different types of object motion. The high motion scenes included bike racing, bull fighting, dancing and flowing water. The low motion scenes included showing maps, historic buildings and structures. The videos also had scenes with varying spatial content such as a bird's eye view of a city, a crowded market, portraits, sky and still water, etc.

We chose twelve 6-minute video sequences, for a combined length of 72 minutes. We grouped the sequences into 4 sets, each consisting of three sequences. This limited a viewing session to 18 minutes so as not to tire or bore the viewers. During each session, a viewer evaluated a set of video sequences with a short break after each sequence. Some viewers participated in more than one viewing session, although never on the same day. Each set of video sequences (and hence each packet loss) was evaluated by 12 viewers.

The age of the viewers varied from 25 to 60 years. All the viewers had either normal or corrected-to-normal vision. None of the subjects had previous experience in video quality except for one expert subject, who evaluated all the

four sets of video sequences. The profession of the viewers was either technical or secretarial.

Viewers were told that the videos they were watching would have impairments caused by packet losses, and that when they saw something unexpected in the video like a glitch, they should respond by pressing the space bar. They were asked to keep their finger on the space bar so they would not be distracted by that task. All the tests were conducted in a well lit room using the same monitor and settings. Viewers were positioned approximately six picture heights from the screen. We observed that the viewers were able to perform the task without any difficulty although they were untrained.

A total of 1080 packet losses were randomly injected into these videos. We are not trying to simulate a typical packet loss scenario, which may include bursty losses, but are instead trying to answer the question "What causes a single packet loss to result in a visible artifact?". Therefore, we introduce isolated losses randomly into each non-overlapping four-second interval. To ensure viewers have adequate time for responding, we randomly inject a packet loss in the first three seconds of each interval and allow a one-second guard interval during which the decoder can recover and the user can respond.

We distributed the losses such that 30% affected an entire frame, 10% affected two adjacent slices, and 60% affected a single slice. Here we consider a slice to be one horizontal row of macroblocks. Further, we chose to have 30% of the losses be in B-frames (and hence have a temporal duration of one frame), and the remaining 70% evenly distributed across the available P- and I-frames in the 3-second interval.

The output of the subjective test was a set of files containing the times that the viewer pressed the space bar relative to the start of the video. We processed these to create a matrix with 1080 rows and 12 columns, whose entries indicate whether a viewer responded to a packet loss or not. If a viewer pressed the space bar within two seconds after a packet loss occurred and before the next

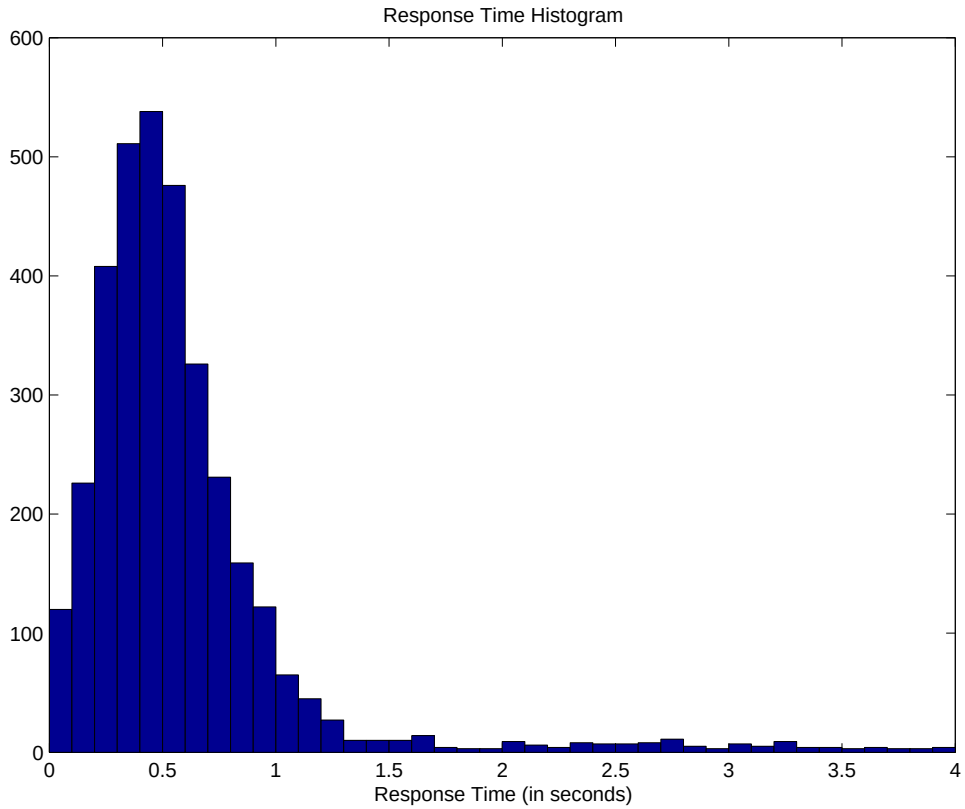


Figure II.4: Histogram of response times

packet loss occurs, he/she is considered to have responded to that packet loss. Otherwise, he/she is considered not to have responded to the packet loss. Figure II.4 shows a histogram of the response times (time difference between the packet loss and the key press). From the histogram, we can see that 91% of the responses occur within one second of the packet loss, and 97% of them occur within two seconds. The average response time is 0.6 seconds. We believe that a viewer who saw a packet loss should be able to respond within two seconds and we consider the responses that come after two seconds to be false alarms. The ground truth for the probabilities of visibility of a packet loss was defined from these viewer responses. The probabilities were calculated as the number of viewers who saw the packet loss divided by 12.

Viewers were not told the pattern of injected packet losses. There is a

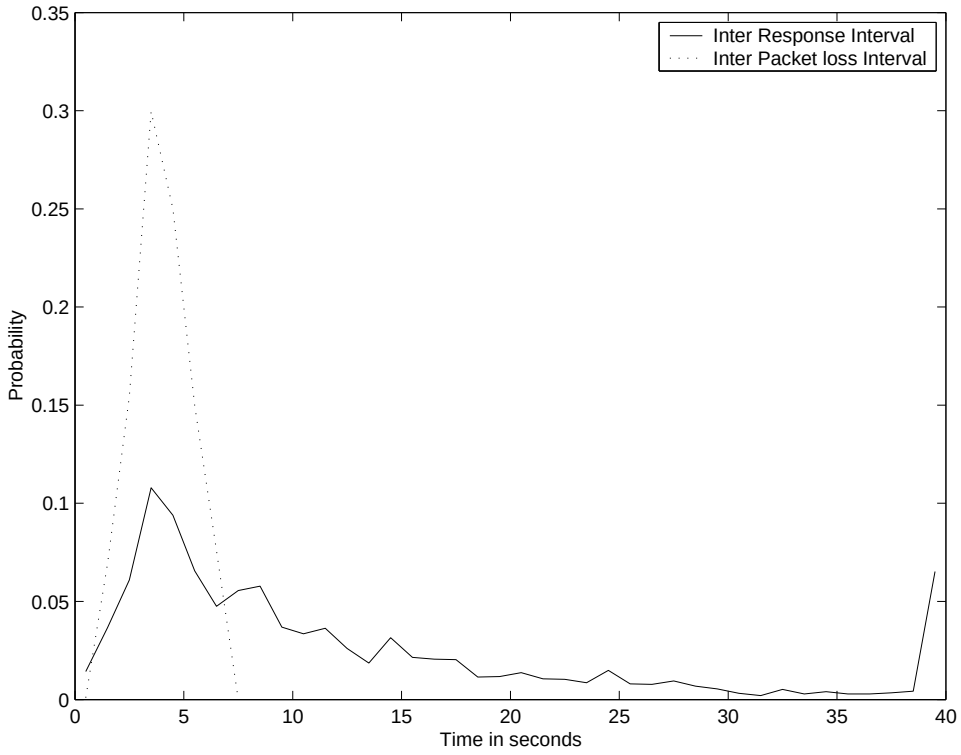


Figure II.5: Histogram of times between adjacent losses

concern, however, that while viewing the video they might infer that a packet loss occurs in every 4-second interval. If viewers were able to predict this, it might bias their responses. To analyze this, we examined the time between adjacent packet losses (Inter Packet loss Interval), and the time between adjacent responses of a viewer (Inter Response Interval). Figure II.5 shows that the density of *Inter Packet loss Interval* (IPI) is triangular with a minimum, mean, and maximum of one, four, and seven seconds, as expected. Also shown in Figure II.5 is the density of *Inter Response Interval* (IRI). Its long tail out to 150 seconds is not shown; instead all the samples larger than 40 seconds have been assigned to the last bin of the histogram which explains the spike in the tail. This density has a peak near four seconds. However, only a small percentage (11.3%) of the IRI samples are between 3.5 and 4.5 seconds, which strongly suggests that viewers would not have been able to infer that a packet loss occurs in every four-second interval and begin

to anticipate an artifact.

II.C GLM - Logistic Regression

We model the probability of visibility using a Generalized Linear Model (GLM). Logistic Regression is a type of GLM which models the parameter p of a binomial distribution. Generalized linear models are an extension of classical linear models [29]. First we will give a brief overview of the classical regression problem and then explain the generalized linear model and logistic regression.

Let $y_1, y_2, ..., y_N$ be a realization of independent random variables $Y_1, Y_2, ..., Y_N$ such that Y_i has binomial distribution with index m_i and parameter p_i. Let $\mathbf{y}$, $\mathbf{Y}$ and $\mathbf{p}$ denote the N-dimensional vectors represented by y_i, Y_i and p_i respectively. We are trying to model the parameter p as a function of P factors. Let $\mathbf{X}$ represent a $N \times P$ matrix, where each row i contains the P factors influencing the corresponding parameter p_i. An ordinary linear model between $\mathbf{p}$ and $\mathbf{X}$ can be written as

$$\mathbf{p} = \gamma + \sum_{j=1}^{P} \mathbf{x_j} \beta_j \tag{II.1}$$

where $\mathbf{x_j}$ is the j^{th} column of $\mathbf{X}$ and $\beta_1, \beta_2, ..., \beta_P$ are the coefficients of the factors. Coefficients β and the constant term γ are usually unknown and need to be estimated from the data. A simple linear regression model is incapable of estimating the parameter p of a binomial model because the output of a linear model typically has the range $(-\infty, \infty)$ while we know $p \in [0, 1]$.

A generalized linear model can be represented as

$$g(\mathbf{p}) = \gamma + \sum_{j=1}^{P} \mathbf{x_j} \beta_j \tag{II.2}$$

where $g(.)$ is called the link function, which is typically non-linear. Classical regression is a special case of GLM where the link function $g(.)$ is an identity. For logistic regression, the link function is the logit function, which is the canonical (therefore default) link function for the binomial distribution. The purpose of the

link function here is to map $p \in [0,1]$ onto the entire real line $(-\infty, \infty)$. The logit function is defined as

$$g(p) = log(\frac{p}{1-p}). \tag{II.3}$$

Given N observations, we can fit models using up to N parameters. The simplest model, also called the Null model, has only one parameter: the constant γ. At the other extreme, it is possible to have a model with as many parameters as there are observations, called the Full Model; however, this is not practically useful. The goodness of fit for generalized linear models can be characterized by the deviance value, which is formed as the logarithm of a ratio of likelihoods.

If we denote the log-likelihood function for model $\mathbf{p}$ (which is a function of β), and the observations $\mathbf{y}$ as $l(\mathbf{p};\mathbf{y})$, then for the binomial distribution we can write the log-likelihood function as

$$l(\mathbf{p};\mathbf{y}) = \sum_{i=1}^{N}[y_i log(\frac{p_i}{1-p_i}) + m_i log(1-p_i) + log(\begin{pmatrix} m_i \\ y_i \end{pmatrix})] \tag{II.4}$$

where m_i represents the number of trials made for observation i. The log-likelihood function $l(\mathbf{p};\mathbf{y})$ is maximized for the full model. Let the full model and the current model be represented by $\tilde{\mathbf{p}}$ and $\hat{\mathbf{p}}$ respectively. Then, we can write $\tilde{p_i} = \frac{y_i}{m_i}$. Further, the deviance for the model represented by $\hat{\mathbf{p}}$ is defined as

$$D(\mathbf{y};\hat{\mathbf{p}}) = 2[l(\tilde{\mathbf{p}};\mathbf{y}) - l(\hat{\mathbf{p}};\mathbf{y})]. \tag{II.5}$$

From the definition, we can see that the deviance for the Full model is zero and the deviance for all other models is positive. A smaller deviance implies that there is a better model fit. The deviance for the null model is also called the null deviance. The deviance is often used as a goodness-of-fit statistic for testing the adequacy of a fitted model. Under the assumptions of independence and $p \in (0,1)$, the deviance can be shown to be asymptotically distributed as $\chi^2_{n-(P+1)}$, where (P+1) is the total number of parameters fitted for the model [29]. Furthermore, the difference in deviance between two models is also known to be approximately distributed as χ^2_k under the assumption of independence alone for

large values of N, where k is the difference in the number of parameters estimated for each model. This is very useful in determining the significance of different factors.

We use the statistical software R [30] for our model fitting and analysis. To obtain the model parameters, R uses an iteratively re-weighted least-squares technique to generate a maximum-likelihood estimate. After fitting a particular model, the importance of each factor in the model can be evaluated by the resultant increase in deviance when we remove that factor from the model. This increase can be compared with the appropriate χ^2 statistic to compute the p-value for this factor. If the p-value is less than 0.05, then the factor is significant at the 95% level. This is known as the likelihood ratio test (LRT). We represent the observed probability of visibility as $\tilde{p}$ and the predicted probability of visibility as $\hat{p}$.

II.D Factors affecting Visibility

In this section, we describe the objective factors that we believe will be useful in modeling the visibility of a packet loss. These objective factors can be classified into two types: content-independent factors and content-specific factors. Content-independent factors depend on the location of the packet loss in the MPEG-2 bitstream, but do not depend on the content of the video. Content-independent factors can therefore be calculated exactly from the lossy bitstream itself. Content-specific factors depend on the content of the video at the location of the packet loss. Content-specific factors can be calculated exactly at the encoder side, by using the original bitstream without losses. However, these content-specific factors cannot be exactly obtained from a bitstream in which packets are already lost.

The first content-independent factor we consider characterizes the duration of time an error persists. We start by using the temporal duration (TMDR), which represents the maximum number of frames that may be affected by the

packet loss. In our data, this varies from one to thirteen because of the encoder's prediction structure. An error in a B-frame lasts a single frame. An error in a reference frame may propagate in time but will always be removed by the next I-frame. A preliminary analysis showed that if TMDR= 1, the packet loss is almost always invisible. However, the correlation coefficient between the number of viewers who saw a packet loss and TMDR is only 0.051, which is very low.

Figure II.6: FRAMETYPE value for different frames in a GOP

Thus, for GLM, we also explore two alternate ways to represent the temporal duration. The first is the boolean variable, BFRAME, which is set whenever the packet loss occurs in a B-frame. The second is the categorical variable FRAMETYPE, which has 6 levels depending on the type of frame in which the packet loss occurs. These 6 levels correspond to a B-frame, four P-frames with a different distance to the next I-frame, and an I-frame. We call these levels B,P1,P2,P3,P4 and I. Figure II.6 illustrates how these frames occur in the GOP structure of our videos. FRAMETYPE captures all the information in the temporal duration of a packet loss. For example, a packet loss in a P3 frame will have a temporal duration of 9. In the GLM, a categorical variable with N levels is treated as a vector of $N-1$ boolean variables. (The N-th level is represented by setting all $N-1$ boolean variables to zero.) Thus for FRAMETYPE, we considered five boolean variables: FRAMETYPE-P1, FRAMETYPE-P2, FRAMETYPE-P3, FRAMETYPE-P4 and FRAMETYPE-I. FRAMETYPE-B is considered default and its effect is included in the constant term.

The second content-independent factor we consider is spatial extent (SPTXNT) which represents the number of slices affected by the packet loss. In our case, it is either 1, 2 or 30 corresponding to single slice, double slice or frame loss

respectively. SPTXNT can be treated as an ordinal variable, taking on values 1, 2, and 30, or as a categorical variable with three levels to distinguish the cases of single slice, double slice, and frame loss errors. In the remainder of the thesis, SPTXNT refers to the ordinal variable except where categorical is stated. For SPTXNT (categorical), in the context of GLM, we consider two boolean variables: SPTXNT-2 and SPTXNT-30. SPTXNT-1 is considered default.

The third content-independent factor we consider is the vertical position (HGT) of the error induced by the packet loss. HGT is the number of the topmost slice affected by the packet loss, where the slices are numbered from 0 to 29 from top to bottom. This factor captures the varying attention viewers have on different regions in the frame. In our study, the values of each of the content-independent factors can be controlled at the time of choosing which losses to introduce. Since the content-independent factors can be extracted exactly from the lossy bitstream, they are identical across our RR, NR-P, and NR-B models.

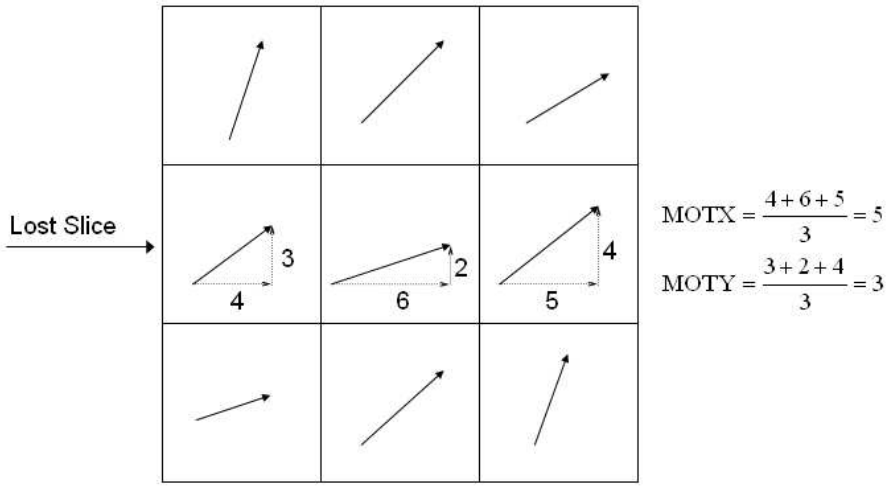


Figure II.7: Computation of MOTX and MOTY using motion vectors of macroblocks in lost slice

Content-specific factors should include some description of motion. We use the MPEG-2 bitstream to extract the received motion vectors corresponding to each frame, so no motion estimation step is needed by the quality monitor. These motion vectors are linearly scaled first such that they represent the motion between the current frame and the past frame in display order. For example, if a motion vector has a value of (2,6) for a MB in a P frame referring back to the previous P frame (2 frames ago) then it would be scaled to (1,3). Likewise a motion vector of (2,6) for a MB in a B frame referring to the subsequent P frame (1 frame later) would be scaled to (-2,-6). Since MPEG-2 uses x and y directional motion vectors, the most immediate way to express motion for purposes of predicting visibility is to use these vector components. MOTX and MOTY represent the scaled motion vector components in x and y directions respectively, averaged across all macroblocks initially affected by the loss. The computation of MOTX and MOTY from the scaled motion vectors is illustrated in Figure II.7. The motion vector variance is denoted by VARMX and VARMY respectively.

To explore whether visibility might be governed more by total motion than by x and y directional motion, we introduce MOTM to represent the magnitude, MOTA to represent the angle, and VARM to represent the variance in overall motion. We calculate them as follows:

$$MOTM = \sqrt{MOTX^2 + MOTY^2} \tag{II.6}$$

$$MOTA = \arctan(\frac{MOTY}{MOTX}) \tag{II.7}$$

$$VARM = VARMX + VARMY \tag{II.8}$$

We also define HIGHMOT, a boolean variable, to be set when $MOTM > 0.707$. This threshold was set to correspond to motion that is greater than half a pixel per frame in both x and y directions.

Content-specific factors also include Residual Energy (RSENGY) and Initial Mean Square Error (IMSE). RSENGY denotes the average residual energy per

Table II.1: Description of factors affecting visibility

Factor Name	Description
TMDR	Time Duration: Maximum number of frames affected by the packet loss
BFRAME	Boolean variable set if packet loss occurs in a B-frame
FRAMETYPE	Type of frame in which packet loss occurred
SPTXNT	Spatial Extent: Number of slices lost
SPTXNT (categorical)	Number of slices lost, categorized as single, double or frame loss
HGT	Height: Number of the topmost slice lost
MOTX	Average motion in x-direction
MOTY	Average motion in y-direction
VARMX	Variance of motion in x-direction
VARMY	Variance of motion in y-direction
MOTM	Magnitude of overall motion
MOTA	Angle of overall motion
VARM	Variance in overall motion
HIGHMOT	Boolean variable set when MOTM > 0.707
RSENGY	Average residual energy per pixel after motion compensation
IMSE	Mean square error per lost pixel

pixel after motion compensation for the lost slices. IMSE is the mean squared error per pixel, after loss and error concealment, in the frame affected by the packet loss averaged over only those pixels in the lost slices. Table II.1 summarizes the descriptions of all the factors.

The content-specific factors described above can be extracted exactly using both the complete bitstream (available at the encoder) and the decoded pixels. For the RR method, the content-specific factors can be extracted at the encoder for all slices, and this information can be made available to the quality monitor via reliable means. This information is then combined with the knowledge of which slices are lost to generate the content-specific factors for lost slices. These factors can be exactly obtained only for FR and RR methods. NR-P and NR-B methods must estimate these factors for the missing slices. Further, to compute IMSE,

decoded pixels are necessary; however, these are unavailable to the NR-B method since an NR-B method has access only to the compressed bitstream and not the decoded pixels.

For the NR-P and NR-B methods, the parameters MOTX, MOTY, VARMX, VARMY (and thereby MOTM, MOTA, VARM and HIGHMOT) and RSENGY are extracted directly from the bitstream for all *received* slices. Parameters for the missing slices are then estimated using one of two approaches. The first approach estimates the parameter using co-located slices in the previous frame. The second approach estimates the factor using spatially neighboring slices in the same frame. We tried each approach on one video sequence and found that the first approach performed better for all the above mentioned parameters.

For the NR-P case, IMSE is computed for all received slices, where IMSE for received slices is defined to be the IMSE that would have resulted if the slice had been lost. The second approach above was found to be more effective for estimating the IMSE of the missing slices. For the NR-B method, neither of the above two approaches can be used to estimate IMSE since the decoded pixels are not available. Thus, for the NR-B case, we use the approach described in [28], which extracts and estimates additional parameters (such as mean, spatial correlation, spatial variance) using the DCT coefficients from the received slices, to estimate IMSE for the missing slices.

II.E Results

In this section, we apply logistic regression to the problem of estimating the probability that a packet loss is visible to an average viewer. We use the factors extracted from our RR, NR-P, and NR-B methods to derive a separate model for each case.

We use the word "model" to characterize the set of factors which comprise the matrix $\mathbf{X}$, introduced in Section II.C. We note that for each "model", we

actually consider three: one for each of the RR, NR-P, and NR-B cases. The distinction among the three lies in whether the content-specific factors are extracted exactly, or estimated as described in Section II.D.

We explored a number of models with different sets of factors, to determine the best way to characterize sequence motion and loss-duration for our objective. Our final model, denoted Model 3, uses the factors FRAMETYPE, SPTXNT (categorical), MOTM, HIGHMOT, VARM, RSENGY, IMSE and HGT to predict the probability of visibility of a packet loss.

The deviances obtained with this model for the RR, NR-P and NR-B cases are 4797.6, 5106.7 and 5115.7 respectively with 1066 degrees of freedom for the χ^2 distribution, while the deviance for the null model (Model 0) is 9254.8 with 1079 degrees of freedom. The MSE obtained between actual probability $\tilde{p}$ and predicted probability $\hat{p}$ is 0.0565 for RR, 0.0608 for NR-P, and 0.0611 for the NR-B case.

To verify the applicability of this model to new data, we perform a 4-fold cross-validation procedure. For this, we use the data from three out of the four sets of video as a training set. The data from the remaining set is used for testing. We repeat this process four times, each time choosing a different set for the testing set. Thus we have a predicted probability, $\hat{p}$, for each packet loss obtained when the packet loss was not used for training. The MSE obtained between $\tilde{p}$ and $\hat{p}$ during cross-validation for Model 3 is 0.0627 for RR, 0.065 for NR-P and 0.0647 for NR-B case. This shows that the model continues to perform well when encountering new data.

The coefficients (γ and βs) for the final model (Model 3) in the NR-B case are tabulated in Table II.2. The values of the coefficients do not necessarily convey the importance of corresponding factors because these factors have different variances and ranges. However, the sign of the coefficients is important and informs whether a packet loss is more visible with a high or low value for a factor. We can make the following conclusions based on the coefficient values:

Table II.2: Coefficients for Model 3 in NR-B

factor	coefficient
constant γ	-4.53
FRAMETYPE-P1	2.116
FRAMETYPE-P2	2.104
FRAMETYPE-P3	2.117
FRAMETYPE-P4	2.188
FRAMETYPE-I	5.326e-01
SPTXNT-2	7.161e-01
SPTXNT-30	1.54
MOTM	4.212e-01
HIGHMOT	1.398
VARM	-1.144e-02
RSENGY	-6.902e-03
IMSE	9.890e-04
HGT	-2.797e-02

High values of MOTM and IMSE cause a packet loss to be more visible. Visibility of losses in I, P and B frames decreases in that order. A large spatial extent (SPTXNT) increases the visibility of a packet loss. High values of VARM and RSENGY cause a packet loss to be less visible. As the physical location of the packet loss is shifted from the top to the bottom of a frame, the visibility of a packet loss decreases.

The significance of different factors in the model can be understood by the increase in the deviance that results if each factor is individually removed from the model. Figure II.8 shows the increase in deviance for each factor, for the RR, NR-P and NR-B cases. From the figure, we see that FRAMETYPE, SPTXNT (categorical), MOTM, HIGHMOT and IMSE are very significant factors affecting visibility. Since HIGHMOT depends completely on MOTM, we can attribute its importance also to MOTM. Considered this way, MOTM becomes the most significant factor affecting visibility.

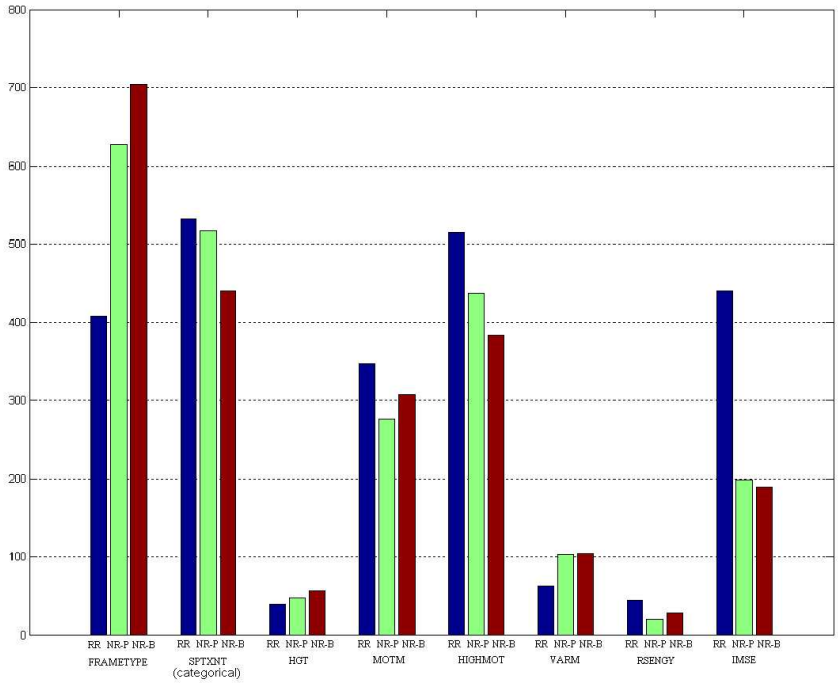


Figure II.8: Factor Significance: Plot showing increase in deviance that results if each factor is individually removed from the model

II.E.1 Other Models

Before arriving at the final model, we explored different models to find the one with the best performance (smallest deviance). We began with our initial model (Model 1) using factors TMDR, SPTXNT (categorical), MOTX, MOTY, VARMX, VARMY, RSENGY, IMSE and HGT. Model 2a drops the four factors related to directional motion, and adds the three factors for overall motion, and it consists of factors TMDR, SPTXNT (categorical), MOTM, MOTA, VARM, RSENGY, IMSE and HGT. MOTA, which was insignificant (95% level), was dropped (Model 2b) and HIGHMOT, a significant factor, was added (Model 2). Model 3a uses BFRAME instead of TMDR. Our final model (Model 3) uses FRAMETYPE instead of BFRAME.

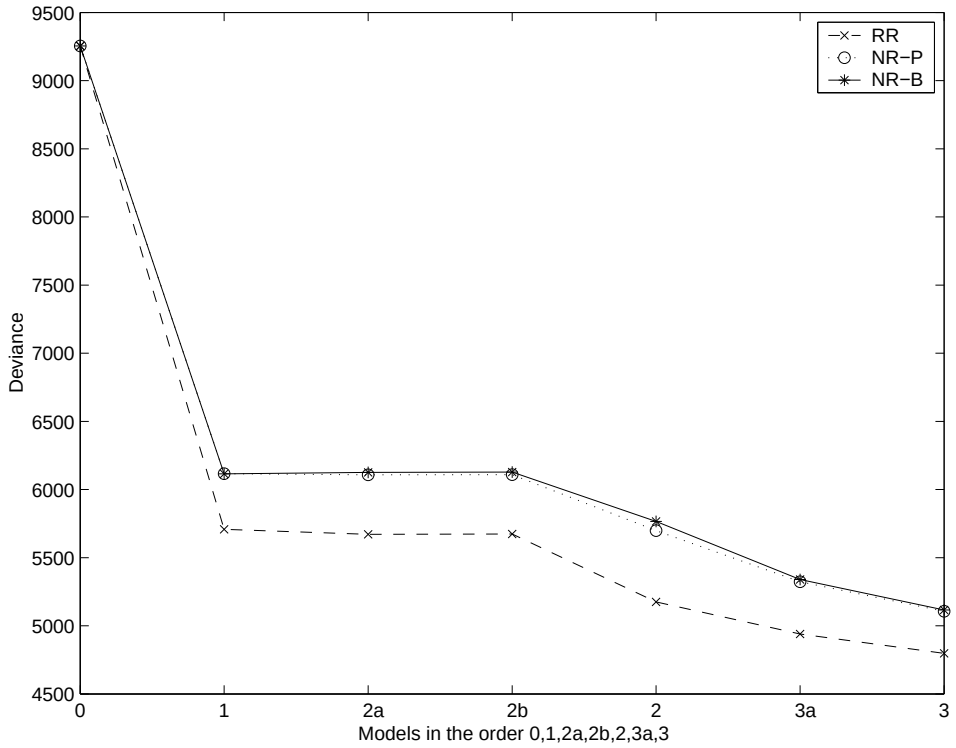


Figure II.9: Plot of Deviance for models considered

The improvement in models from the null model (Model 0) to the final model (Model 3) can be summarized by the plot of deviance, shown in Figure II.9, for all three cases (RR, NR-P and NR-B). There is a huge drop in deviance from the null model to the initial model (Model 1), which is expected. When we improve the treatment of the motion variables and also reduce the model order (Model 2), we see a decrease in deviance indicating a better fit. Also, we see a further decrease in deviance from Model 2 to Model 3 when we treat the time-duration information using a Boolean structure.

II.F Conclusion

In this chapter, we considered the problem of predicting the probability that a packet loss causes visible artifacts, using measurements from either the

entire encoded video, the decoded video pixels, or just the received lossy bitstream. We used a generalized linear model (GLM) to fit the data from subjective tests using these measurements. We examined how to describe pertinent factors such as motion to best predict visibility. As a result, we use MOTM instead of MOTX and MOTY and FRAMETYPE instead of TMDR, and we dropped insignificant factors such as MOTA. We achieved a cross-validated MSE of 0.0627 between the actual and predicted probabilities in the RR case.

II.G Acknowledgements

This work was supported in part by the National Science Foundation, the Center for Wireless Communications at UCSD and the UC Discovery Grant Program of the State of California.

The authors would like to thank Prof. Charles Berry at UCSD for his support and guidance with statistical methods and Dr. Yegnaswamy Sermadevi for software, instrumental in conducting the subjective test.

Chapter II of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, P. C. Cosman, A. R. Reibman and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video", *IEEE Trans. Multimedia*, vol. 8, pp. 341-355, April 2006. I was the primary author and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter II. The co-author Dr. Vaishampayan also contributed to the ideas in this work.

III

Classification Problem

In this chapter, we address the problem of classifying individual packet losses as visible or invisible. Traditionally, thresholds on the average packet loss rate (PLR) have been used to monitor and guarantee a certain level of video quality. However, all packet losses do not cause degradation in video quality and so PLR is not an accurate indicator of video quality. If we know which packet losses are visible, we can use the Visible Packet Loss Rate (VPLR), i.e. the rate of losses causing visible errors, to better predict video quality.

For classification purposes, we define a packet loss to be visible if 75% or more viewers responded to it which is midway between chance (50%) and certainty (100%). This 75% threshold is a typical psychometric threshold used to calculate the Just Noticeable Difference (JND) ([31] is an example). Similarly, a packet loss is invisible if 25% or fewer viewers responded to it. The remaining losses are indeterminate and not used in classification data. We use the subjective test data and the associated factors described in Chapter II. Of the 1080 total packet losses shown to viewers, 732 were invisible, 195 were visible and 153 were indeterminate. For the classification problem, we do not concentrate on the 14% of losses that were indeterminate, but instead focus on understanding the 927 visible and invisible losses.

We use a well-known statistical tool called Classification and Regression

Trees (CART) to design the classifiers for the RR, NR-P and NR-B cases. We will also show how our GLM models from Chapter II can be used to classify each loss as visible or invisible and compare their performance to that of CART classifiers.

This chapter is organized as follows: Section III.A gives a brief introduction to CART. Sections III.B and III.C describe the classification results obtained by using CART and GLM respectively. Section III.D compares the classification performance of CART with that of GLM. Section III.E concludes.

III.A Introduction to CART

Classification and Regression Trees (CART) is a tool for tree structured data analysis introduced by Breiman et al. [32]. CART generates its results in the form of decision trees. This allows CART to handle massively complex data while producing classifiers that are easy to understand. The decision criteria give us insight into what causes a packet loss to be visible, and can be compared with intuition. Other approaches such as Artificial Neural Networks tend to be harder to interpret, since they may involve weighted sums of large numbers of input parameters, whose individual effects cannot be discerned.

CART uses binary recursive partitioning. The process is binary because parent nodes are split into exactly two child nodes. It is recursive because the process can be repeated by treating each child node as a parent. The key elements of a CART analysis are a set of rules for (1) splitting each node in a tree, (2) deciding when a tree is complete and (3) assigning each terminal node to a class outcome (or predicted value for regression).

We wish to select each split of a subset so that the data in each of the descendant subsets are "more pure" than the data in the parent subset. CART usually splits data based on a threshold applied to the value of a variable. At each node, CART searches through all possible thresholds for all variables and picks the variable and the threshold that give the best split for that node. The best split is

based on a purity criterion, such as the Gini index of diversity [32].

CART continues to split until all the elements in a node belong to the same class or the number of elements in a node is less than a predetermined threshold. Using this process, CART forms the largest possible tree which is later pruned to get the final tree, the one that gives the best cross-validation accuracy among all the pruned trees.

For each terminal node, CART assigns a class that minimizes the misclassification cost incurred on account of this assignment. If the misclassification costs for all the classes are the same, then the class with highest representation in the terminal node will be assigned as the class for the terminal node.

III.B Classification using CART

We compare two objective classifiers that classify each packet loss to be visible or invisible to an "average" human observer. Each classifier is a decision tree; the classifier traverses a tree where the path at each node depends on a binary decision using one of the factors discussed in Section II.D. During the formation of the tree, a node is split to minimize the probability of misclassification.

The first classifier we consider is ***Root-tree*** + *CART*. **Root-tree**, shown in Figure III.1, is based on our observations in the RR case regarding the impact of short temporal duration, small motion and spatial extent. Of all packet losses with TMDR = 1, only one is visible and the remaining 119 are invisible. Only 11 out of 330 packet losses with both MOTX and MOTY less than 0.5 (half a pixel) are visible. Of the full-frame packet losses, 39% are visible, while only 13% of the single- and double-slice losses are visible.

Root-tree consists of the following decisions. First, all packet losses with temporal duration of one frame (TMDR = 1) are classified as invisible. Second, all packet losses with small motion, defined by (MOTX$\leq$0.5 & MOTY$\leq$0.5), are classified as invisible. The application of **Root-tree** results in 12 and 35 misclas-

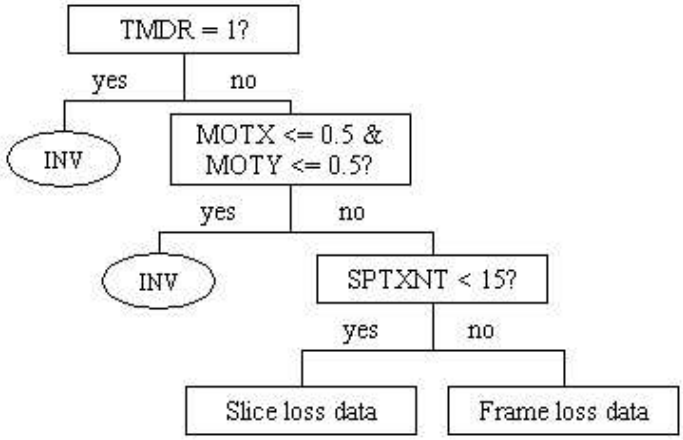


Figure III.1: **Root-tree**

sifications out of 450 and 604 cases for the RR and both NR cases respectively. Both the NR cases have the same result with the **Root-tree** since the variables involved have the same values for both NR-P and NR-B cases. Next, we split the tree based on the initial spatial extent (SPTXNT<15) without making any decision. The threshold for the split on SPTXNT could be any value between 3 and 30; the goal of the split is merely to separate slice losses (single and double) from frame losses. At this stage, we apply CART to classify the data in each of the two nodes.

The second classifier we consider is *CART*. This classifier is designed by applying CART to the entire data set using the factors TMDR, SPTXNT, HGT, MOTX, MOTY, VARMX, VARMY, RSENGY, and IMSE.

We also consider the two classifiers described above with modified motion variables MOTM and VARM instead of MOTX, MOTY, VARMX and VARMY. The **Root-tree** is slightly modified to incorporate the MOTM variable instead of MOTX and MOTY. Now, packet losses with small motion are defined by MOTM<0.707 instead of MOTX$\leq$0.5 & MOTY$\leq$0.5. The threshold for MOTM

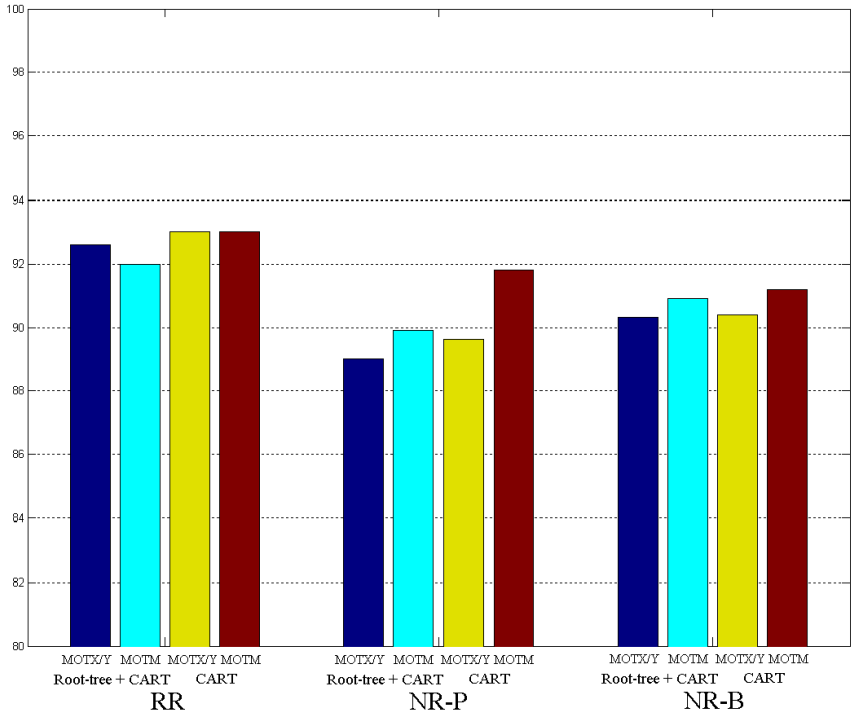


Figure III.2: Comparison of classification accuracy. For example, the first bar on the left shows that the RR method using $Root - tree + CART$, with motion information expressed as x and y directional motion, achieves a cross-validated correct classification of 92.6%

is obtained by using the thresholds for MOTX and MOTY in the equation for MOTM (see Chapter II).

Figure III.2 shows the cross-validated classification accuracy for the two classifiers with and without using modified motion variables under different methods RR, NR-P and NR-B. As we can see, both the classifiers ***Root-tree*** + *CART* and *CART* show an overall improved performance with this treatment of motion variables. *CART* has a slight edge over ***Root-tree*** + *CART*. As expected, the RR method always performs better than the NR methods, but the improvement in performance is not large. The maximum improvement in performance observed is

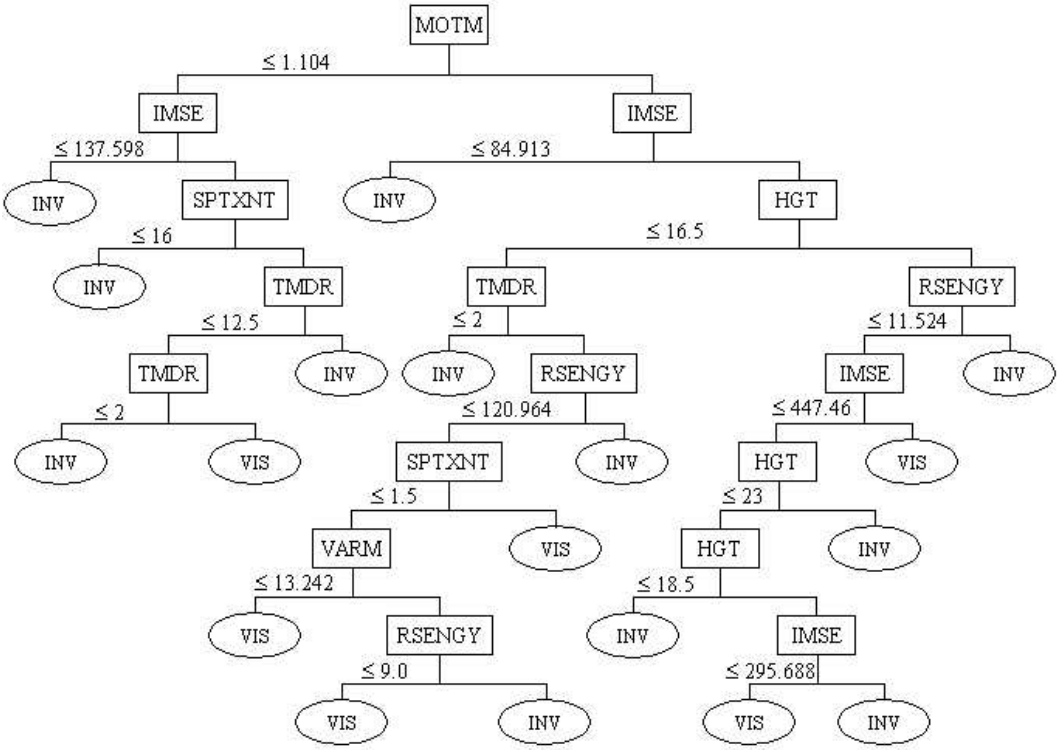


Figure III.3: *CART* classifier tree in the NR-B case

3.6% which occurs over the NR-P method, with the ***Root-tree*** + *CART* classifier when MOTX and MOTY variables are used.

Figure III.3 shows the classification tree obtained using *CART* for the NR-B case. The terminal nodes are represented by ovals and the internal nodes are represented by rectangles. Each internal node is split on the variable shown in its rectangle. If a terminal node is marked VIS, then all packet losses that fall into this node are classified as visible. Similarly, if a terminal node is marked INV, all packet losses that fall into this node are classified as invisible. The tree has splits based on MOTM and IMSE at the top, which shows that they are very important factors. This classifier tree performs the best in the NR-B case with a cross validation accuracy of 91.2% (as opposed to 93.0% in the RR case). Most of the splits of the tree are intuitively reasonable; for example, branches on the left half of the tree with lower IMSE or SPTXNT lead to terminal nodes labeled

INV. However, we find some counter-intuitive splits close to some terminal nodes. One of these is in the left half of the tree, where the data set is split on TMDR with a threshold of 12.5, and then it is immediately split again with a threshold of 2. Both the data less than 2 and the data greater than 12.5 are classified as invisible, which is counter-intuitive. The other occurrence of a counter-intuitive split is in the lower right of the tree, where there is a repeated split on HGT. We will only incur an additional 10 classification errors out of 927 (approximately 1%) during re-substitution if we remove these counter-intuitive splits. Since these splits classify a very small fraction of the data, they do not affect the overall performance of the classifier significantly. We believe that these spurious splits are a result of few available data points to judge the split and can be rectified with a larger data set.

III.C Classification using GLM

Until now, we have used GLM to predict the probability of visibility only. In this section, we describe one way to use the GLM model for classifying packet losses, and we analyze the results.

For this study, we classify a packet loss to be visible, invisible, or indeterminate, based on its probability of visibility. We divide the interval $[0, 1]$ into three regions, using the parameter α:

[0,0.5-α]	Invisible region	
(0.5-α, 0.5+α)	Indeterminate region	(III.1)
[0.5+α, 1]	Visible region	

The only exception is that when $\alpha = 0$, a probability of 0.5 is considered to be indeterminate and the invisible and visible regions are half open intervals. Our classifier takes as input the extracted parameters, and applies the final model (Model 3). If the resulting probability of visibility does not fall in the indeterminate region, we classify the packet loss to be visible or invisible appropriately.

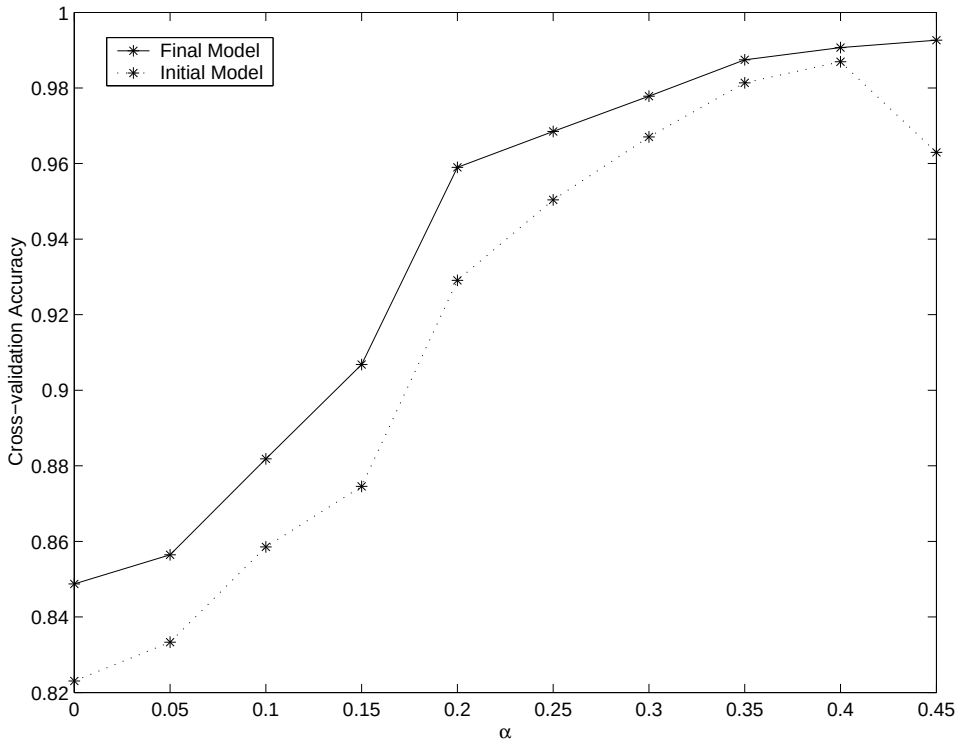


Figure III.4: NR-B: Cross-validation accuracy versus α for GLM classifier

To evaluate the accuracy of the model for classification purposes, we compute the ground truth regarding visibility using the results of the subjective test. Further, for the evaluation process, we only consider those packet losses where the ground truth regarding visibility is not indeterminate. Thus, we only consider those cases where both $\tilde{p}$ and $\hat{p}$ do not fall into the indeterminate region. A decision is correct if $\tilde{p}$ and $\hat{p}$ both fall into the visible region or the invisible region. A decision is wrong if $\tilde{p}$ falls in the invisible region and $\hat{p}$ falls in the visible region or vice-versa. Here, we assign zero cost to classifying an invisible/visible packet loss as an indeterminate packet loss, and unit cost for each wrong decision described above.

We vary α from 0 to 0.45 in steps of 0.05 and calculate the accuracy of the model for each value of α. Figure III.4 shows the variation of cross-validation accuracy with α for the initial (Model 1) and final (Model 3) models using NR-B

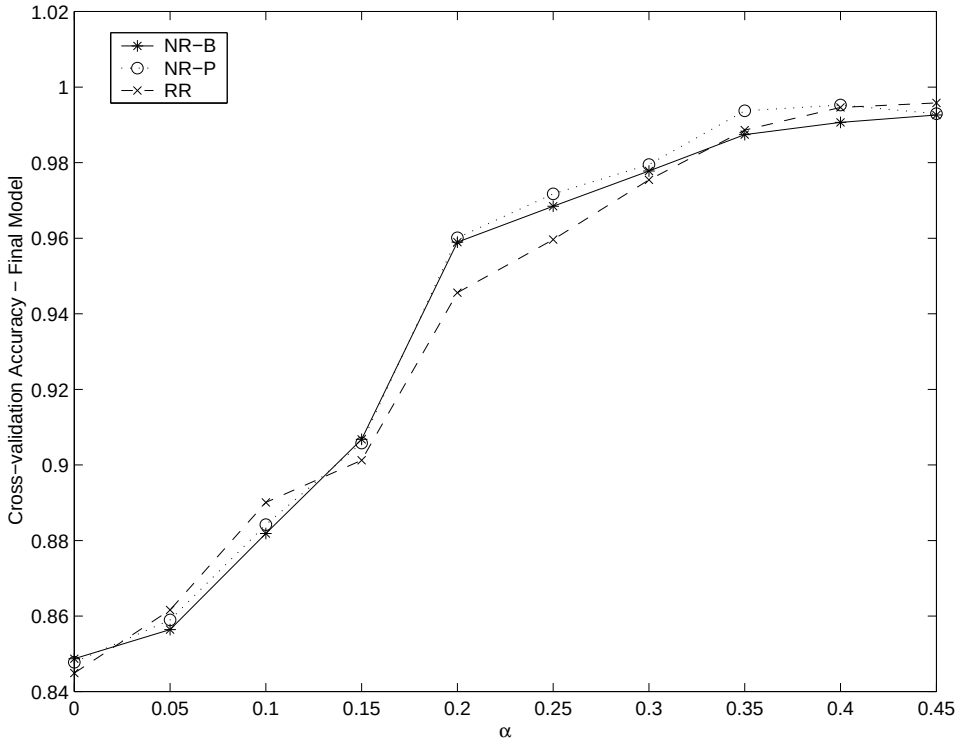


Figure III.5: GLM classifiers: Comparision of RR, NR-P and NR-B methods

method. RR and NR-P methods also exhibit similar variation of accuracy with α. The final model is more accurate than the initial model for all three methods.

Figure III.5 compares the accuracy of the RR, NR-P and NR-B methods using the final model for different values of α, and Figure III.6 shows the corresponding number of decisions in each case. Clearly, all three methods (RR, NR-P and NR-B) perform very similarly for different values of α. In particular, our NR-B method performs almost as well as our RR method. As expected, fewer decisions are made as the size of the indeterminate region (2α) increases, but accuracy of classification increases. If we choose a large value of α, we will obtain high accuracy but fewer decisions. On the other hand, a small value of α allows us to make more decisions, but with lower accuracy.

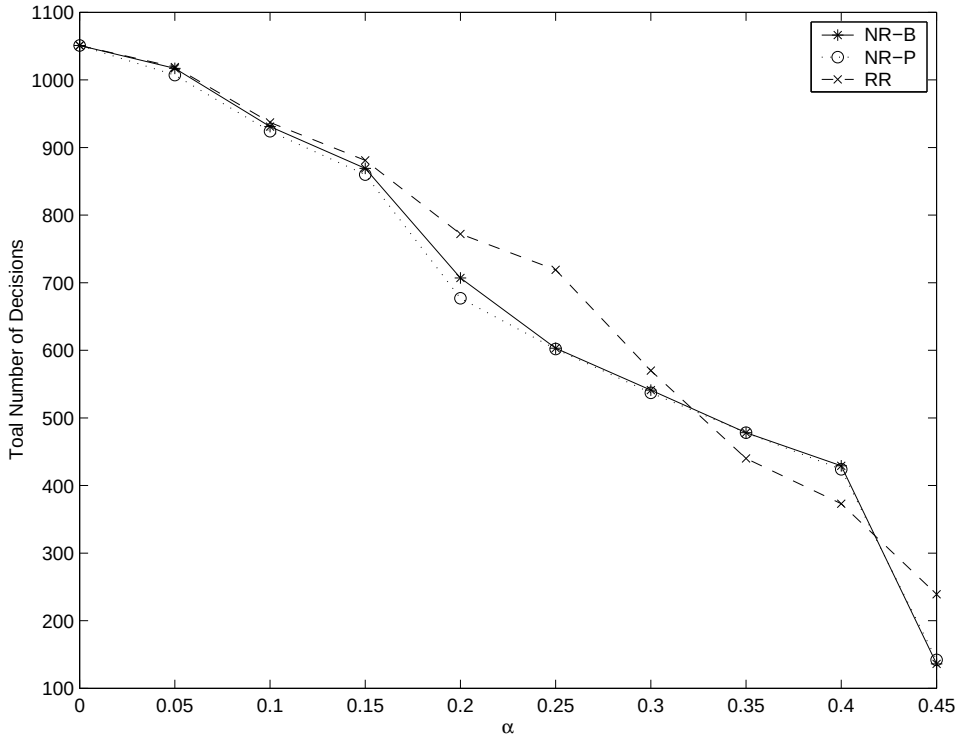


Figure III.6: GLM classifiers: Number of decisions made versus α

III.D Comparison - CART and GLM

We now compare CART and GLM in terms of classification performance. In order to do this and be consistent, we need to compare their classification performance on the same set of packet losses. Using CART, we are able to classify 927 packet losses labeled as visible or invisible based on the ground truth of visibility defined at the beginning of this chapter. When we apply the classification procedure described in Section III.C, we do not classify the same set of packet losses as CART does, even when α is equal to 0.25 (CART classifies packet losses when $\tilde{p}$ is not in the indeterminate region, but GLM classifies only when both $\tilde{p}$ and $\hat{p}$ are not in the indeterminate region).

For comparison purposes, we restrict our data set to the 927 visible and invisible packet losses. We use this restricted set with their actual probabilities of visibility, $\tilde{p}$, to train the GLM. When the predicted probability, $\hat{p}$, is greater than

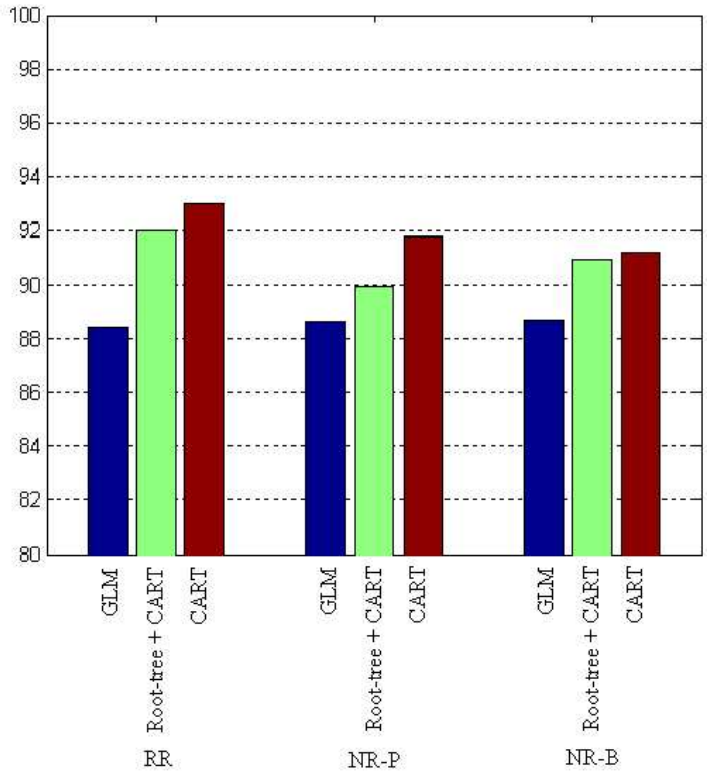


Figure III.7: Comparison: Classification accuracy of GLM and CART based classifiers. For example, the first bar on the left shows that the RR method using GLM achieves a cross-validated correct classification of 88.4%

0.5, we classify the packet loss as visible. Otherwise, we classify the packet loss as invisible. Figure III.7 shows a bar plot comparing the cross validation classification accuracy of GLM to that of the two classifiers based on CART.

As we can see, the classifiers based on CART outperform the GLM based classifiers for all of the three cases. From these observations, we conclude that CART is still a better model for classification purposes, while GLM gives more information in the form of probability of visibility of a packet loss.

III.E Conclusion

In this chapter, we considered the problem of classifying each packet loss as visible or invisible, using measurements from either the entire encoded video, the decoded video pixels, or just the received lossy bitstream. We used CART to design classifiers using these measurements and achieved a cross-validation classification accuracy of 93% in the RR case. We also showed how to design classifiers using our GLM models from Chapter II. We observed that the CART-based classifiers outperform the GLM-based classifiers.

III.F Acknowledgements

This work was supported in part by the National Science Foundation, the Center for Wireless Communications at UCSD and the UC Discovery Grant Program of the State of California.

Chapter III of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, P. C. Cosman, A. R. Reibman and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video", *IEEE Trans. Multimedia*, vol. 8, pp. 341-355, April 2006. I was the primary author and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter III. The co-author Dr. Vaishampayan also contributed to the ideas in this work.

IV

Visibility of Multiple Losses in H.264/AVC

The problem of predicting the visibility of individual packet losses in MPEG-2 bitstreams was studied in Chapters II and III. The losses introduced were individual (isolated) packet losses and the subjective test was conducted with MPEG-2 bitstreams. However, video transmission over internet or wireless links is typically characterized by bursty packet losses. Stuhlmuller et al. [33] analyzed and modeled the distortion (MSE) of isolated packet losses. They also used a linear additive model to quantify the distortion of multiple packet losses from the individual losses involved. This model is accurate only if the losses are spaced sufficiently far apart with respect to the intra refresh period [34]. In [34], the authors compared bursty losses with isolated losses of equal combined length. They concluded that: (a) the loss pattern has significant impact on resulting distortion, and (b) bursty loss produces larger distortion than an equal number of isolated losses. Chakareski et al. [35] proposed a scheme to predict the distortion incurred due to bursty packet losses. In these papers [33, 34, 35], a packet is assumed to be of variable length comprising a single frame.

To generalize the concept of visibility to practical scenarios, we conducted a new subjective test involving multiple losses. A multiple loss is defined as a

set of L individual packet losses occurring in close temporal proximity. For this subjective test, we used H.264/AVC [36] bitstreams instead of MPEG-2 bitstreams. Further, motion-compensated error concealment (MCEC) is used to conceal the packet losses instead of the zero-motion error concealment (ZMEC) which was used in our previous work. Because of these differences, we also introduced individual packet losses to model their visibility.

In this chapter, we have two goals:

1. The first goal is to model the visibility of individual packet losses. We introduced new factors that exploit the advanced features of H.264/AVC. New factors based on spatial and temporal coherence of motion, spatial clutter, contrast and side match distortion are also introduced. The relative importance of these factors is ascertained through statistical modeling. The effect of different factors on packet loss visibility is also analyzed.

2. The second goal is to model the visibility of multiple packet losses. A new model framework is introduced to predict the visibility of a multiple packet loss and its performance is examined for the case of $L = 2$ (dual loss).

This chapter is organized as follows: Section IV.A describes the design of the subjective experiment. Section IV.B discusses the various factors that are used to predict the visibility of a loss. Section IV.C describes our approaches for modeling visibility. Sections IV.D and IV.E provide the results and conclusion.

IV.A Subjective Tests

We conducted subjective tests in order to obtain ground truth on the visibility of packet losses. In the test, the viewers' task is to indicate when they saw an artifact, where an artifact is defined simply as a glitch or abnormality. The subjective tests were single stimulus tests, which means that the viewers were only shown the videos with packet losses and not the original videos.

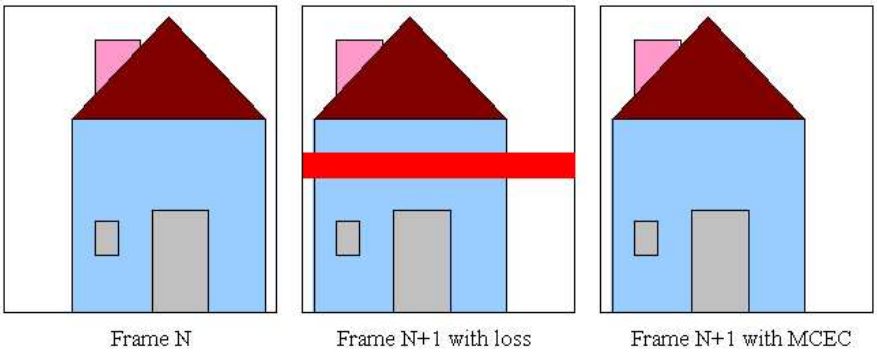


Figure IV.1: Motion-compensated error concealment (MCEC)

The packetization strategy is such that a packet loss entails the loss of a single slice. A slice is assumed to contain one horizontal row of macroblocks[1]. The initial error induced by a packet loss depends on the error concealment strategy used by the decoder. The MCEC algorithm used here and shown in Figure IV.1, incurs a lower initial error compared to the ZMEC method used in our previous work (Chapters II and III). The MCEC algorithm estimates the motion vector (MV) and the reference frame for the lost macroblock and conceals it with the macroblock predicted using the estimated motion vector. Motion compensation in H.264/AVC can occur at different levels from the macroblock level to the smallest block level (4×4 pixel block). Accordingly, each macroblock can have a different number of motion vectors ranging from 1 to 16. These motion vectors can reference different reference frames because of multiple frame prediction. A set of motion vectors is formed from motion vectors of blocks around the lost macroblock. The frame that is referenced the most number of times in the set among all the reference

[1]The Flexible Macroblock Ordering (FMO) option available in H.264/AVC is not enabled.

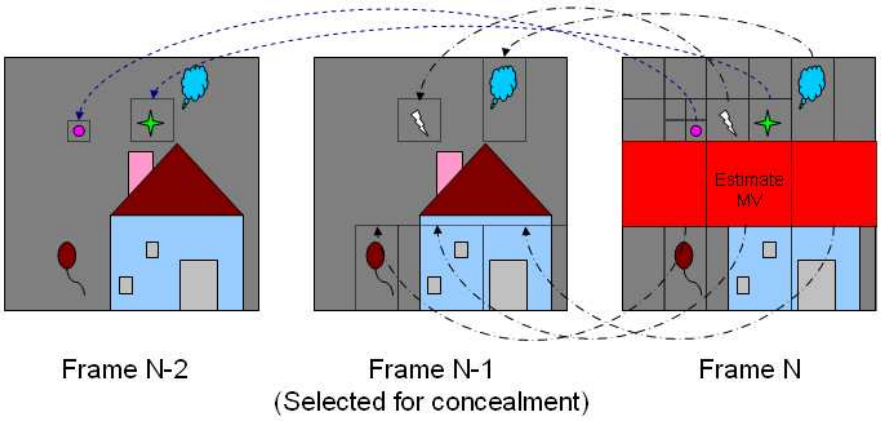


Figure IV.2: Motion vector estimation for a macroblock

frames is selected for concealment. This is illustrated in Figure IV.2. The estimated motion vector is the median of all the motion vectors in the set that refer to this selected frame.

The video sequences chosen for the subjective tests are muted travel documentaries of SIF resolution (352×240) at 30 fps. They are encoded and decoded using the extended profile of H.264/AVC JM Version 9.1 Codec. The encoding structure is I B P B P B...P B with a GOP size of 20 frames. For P frames, two reference frames are used for motion compensation - a long-term reference frame and a short-term reference frame. The long-term reference frame is always the I frame of the current GOP and the short-term reference frame is the previous P frame. B frames use the future P frame and either the long-term or short-term reference frame for bidirectional prediction. The quantization parameter is set at a constant low value of 28 without any rate control so that no compression artifacts are introduced. The bit-rate for these videos varied from 230 to 350 Kbps. The only artifacts that are present in the lossy videos are artifacts caused by packet

losses. Whenever there is a packet loss, the decoder conceals the lost slice using the MCEC scheme. The video sequences used contain the same wide variety of scenes described in Chapter II.

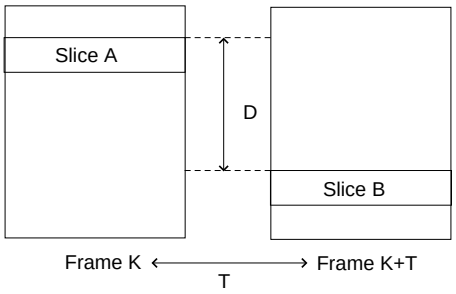


Figure IV.3: Example of a dual loss

We are interested in the visibility of individual packet losses and multiple packet losses. In practice, there are many scenarios in which losses are bursty leading to a multiple packet loss. For example, when there is buffer overflow in a network, multiple packets may be lost; similarly in a wireless link, packets transmitted during a deep fade are lost. Different packet losses within a multiple packet loss interact with each other and so the overall effect is not a summation of individual loss effects. This interaction can be either physical or perceptual. Physical interaction is caused by inter-frame prediction. For example, if the topmost slice is lost in frame 10, and also in frame 11, then the error concealment in frame 11 may work poorly, since it relies on an already compromised frame 10. Perceptual interaction is due to the close proximity (spatially or temporally or both) of the packet losses. In this chapter, the overall effect of two packet losses (dual loss) occurring together is studied. An example of a dual loss is shown in Figure IV.3. Packet losses A and B cause the loss of slices A and B respectively. D represents the spatial separation between the two losses in macroblock units, i.e., $D = 1$ implies the two packet losses affect adjacent slices. T represents the temporal separation between the two losses in number of frames. D has a range of 0 to 14 (15 slices in

each frame). The temporal separation T between the two losses is varied from 0 to 5 frames, which corresponds to a maximum separation of one-sixth of a second with the frame rate set to 30 fps.

Six video sequences of duration 6 minutes each are used for the subjective test. Each video is divided into 4-second intervals called slots resulting in 90 slots for each video and 540 slots in total. A loss (individual/dual) is introduced in the first three seconds of the slot, while the last second is reserved as a guard interval. The guard interval prevents interaction between losses across slots and provides the viewer time to respond to the current loss before the next loss occurs.

We wanted to distribute the dual losses uniformly over possible values of D and T. For each slot, 4 individual losses are chosen and all possible combinations (of pairs of individual losses) are used to get 6 dual losses. The location of the individual losses within the slot was randomly chosen using an adaptive distribution so that the resulting set of dual losses are distributed approximately uniformly over D and T. There are 10 different losses (4 individual and 6 dual) that can be introduced in a slot and so 10 different lossy versions were created from each source video. This resulted in 60 lossy videos from 6 source videos. Each lossy video has both individual and dual losses, but in different slots. The lossy videos are grouped into 10 sets; the first set contains the first lossy version of the 6 source videos, the second set contains the second lossy version and so on. The lossy videos contain 2160 individual losses and 3240 dual losses.

A total of 120 viewers were recruited for the subjective tests and each viewer participated in the subjective tests only once. During the subjective test, a viewer was shown one set of lossy videos. Each set of lossy videos (and hence each loss) was evaluated by 12 viewers. A 1-minute pilot training video is shown to viewers, before the actual test, to help them understand the task and attain a basic level of expertise. Viewers were told that they will watch videos which are affected by packet losses. Whenever they see a visible artifact or a glitch, they should respond by pressing the space bar. They were asked to keep their finger

on the space bar to minimize response time and ensure that this task did not take their attention away from the monitor. The exact instructions, given to the viewers, are listed in Appendix A. The viewers also signed a consent form shown in Appendix B. All tests were conducted in a well lit room using the same monitor and settings. Viewers were positioned approximately six picture heights from the screen.

The age of the viewers varied from 19 to 34. All the viewers had either normal or corrected-to-normal vision. The viewers were mainly students and staff at the University of California, San Diego, and they were not experts in evaluating video quality.

The output of the subjective test was a set of files containing the times that the viewer pressed the space bar relative to the start of the video. These files are processed to create the viewers' boolean responses corresponding to whether they saw a loss or not. If a viewer pressed the space bar within two seconds after a loss occurred and before the next loss occurs, he/she is considered to have responded to that loss. We believe that a viewer who saw a packet loss should be able to respond within two seconds (see Chapter II) and so the responses that come after two seconds are ignored. The ground truth for the probability of visibility of a loss was defined from these viewers' responses. The probabilities were calculated as the number of viewers who saw the loss divided by 12.

Viewers were not told the pattern of injected packet losses. As in our previous subjective experiment (Chapter II), there is a concern that while viewing the video they might infer that packet losses occur somewhere in every 4-second interval. So we examined the densities of *Inter Loss Interval* (ILI) and *Inter Response Interval* (IRI), as done previously. The density of ILI is triangular with a minimum, mean, and maximum of one, four, and seven seconds, as expected. For the density of IRI, the long tail out to 320 seconds is not shown; instead all the samples larger than 40 seconds have been assigned to the last bin of the histogram which explains the spike in the tail. This density has a peak near four seconds.

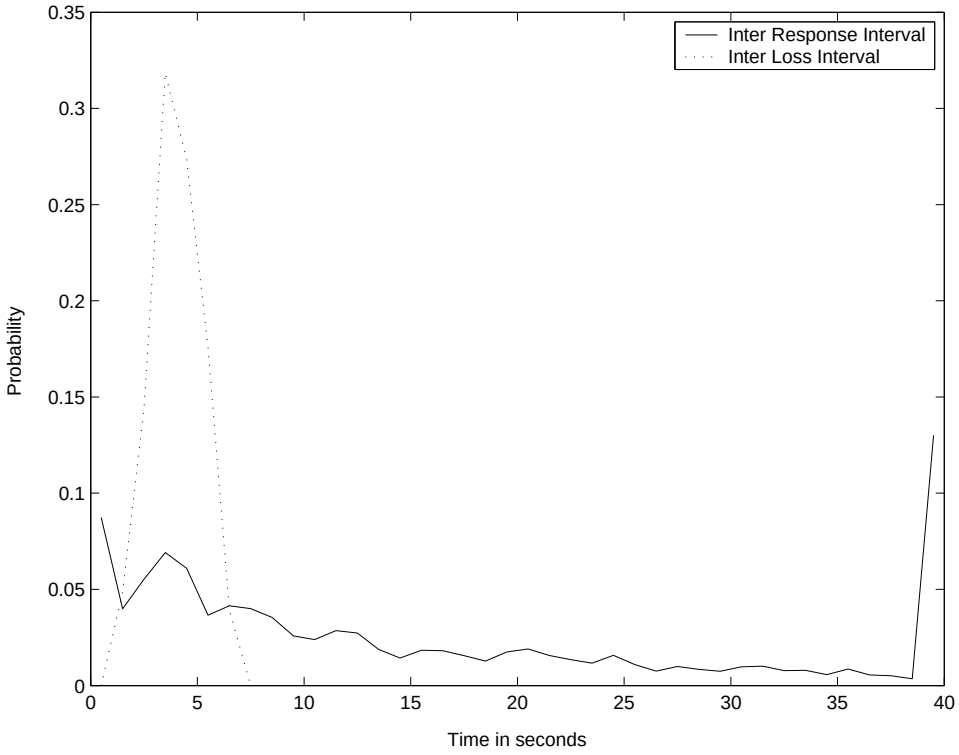


Figure IV.4: Histogram of times between adjacent losses

However, only a small percentage (7.2%) of the IRI samples are between 3.5 and 4.5 seconds, which strongly suggests that viewers would not have been able to infer that a packet loss occurs in every four-second interval and begin to anticipate an artifact.

For classification purposes, we follow our earlier definition that a loss is visible if 75% or more viewers responded to it and invisible if 25% or fewer viewers responded to it. Among the 2160 individual losses, 1893 are invisible, 82 are visible and the rest are indeterminate. This implies that 96% of the non-indeterminate losses are invisible as opposed to 79% in our previous subjective test. As discussed later, this sharp increase in the percentage of invisible losses is attributed mainly to the use of MCEC in this work as opposed to ZMEC. Of the 3240 dual losses, 2549 are invisible, 260 are visible and the rest are indeterminate.

IV.B Factors affecting Visibility

Several factors determine packet loss visibility. We describe the content independent factors first and then describe the content dependent factors.

Content Independent Factors: Content independent factors depend only on the location of the loss and not on the actual content of the loss. Two such factors are considered in this experiment. HGT is defined as the location of the lost slice in a given frame, where slices are numbered from top to bottom. $FRAMETYPE$ is the type of frame (B/P/I) affected by the packet loss and is treated as a categorical factor. $FRAMETYPE\_ML$ is the counterpart of $FRAMETYPE$ for the dual (multiple) loss case. Between the two frames affected, $FRAMETYPE\_ML$ represents the type that belongs to a higher category. Category I is higher than category P, which is higher than category B.

Content Dependent Factors: Content dependent factors, on the other hand, depend on the actual video content at the location of the loss, such as motion information, contrast etc. Many content dependent factors based on initial error, residual energy, scene motion, spatial clutter, contrast and side match distortion are considered.

1. **Initial Mean Squared Error (IMSE):** IMSE is defined as the mean squared error between the error-free reconstructed MB and the concealed MB (affected by packet losses). IMSE is calculated on a macroblock basis. The factor $AVGIMSE$ is computed by averaging the IMSE values of all the macroblocks in a given slice. Similarly, the factor $MAXIMSE$ is defined as the maximum IMSE value among all macroblocks in a given slice.

2. **Residual Energy:** Residual energy is defined as the energy (sum of squares of all DCT coefficients) of the residual obtained after motion compensation. If a slice is lost, and if the MCEC should happen to do a perfect job of recreating the lost motion vectors, the resultant slice would of course still differ from the original slice, and the residual energy is one way of assessing the

magnitude of that difference. Residual energy is calculated on a macroblock basis. The factor $AVGRSENGY$ is computed by averaging the residual energy values of all the macroblocks in a given slice. Similarly, the factor $MAXRSENGY$ is defined as the maximum residual energy value among all macroblocks in a given slice.

3. **Motion-related Factors:** For computing motion-related factors, the motion vectors in each partition of a MB are linearly scaled first, as in Chapter II, such that they represent the motion between the current frame and the past frame in display order. Then a single motion vector is assigned to each macroblock, which is a weighted average of the motion vectors in all the partitions of the macroblock, the weights being directly proportional to the size of partition.

   $MEANMAG$ and $MAXMAG$ are defined as the mean and maximum of magnitudes of all motion vectors of the macroblocks in a given slice. Motion in the x and y directions is averaged over all the macroblocks in a given slice and is named $MEANMOTX$ and $MEANMOTY$. $VARMOTX$ and $VARMOTY$ are defined as the variance of motion in x and y directions for all macroblocks in a given slice.

   In addition to the magnitude of motion, it is possible that the direction of motion (vector angle or phase information) might be an important predictor of visibility. Phase is defined for all vectors except the (0,0) motion vector which has a zero magnitude and no direction. For computing motion phase related factors, if less than half of the MBs in the slice have their phase defined (i.e. their motion vectors are non-zero), the phase information is considered undefined and a boolean factor called PH_UDEF is set. If PH_UDEF is not set, $MEANPHASE$ and $MAXPHASE$ are the mean and maximum of all the defined phases of the macroblocks in the slice. In the dual loss case, there are two boolean factors PH_UDEF1 and PH_UDEF2. PH_UDEF1 is set

if at least one of the losses has an undefined phase and PH_UDEF2 is set if both losses have undefined phase. Since a value cannot be assigned when the phase is undefined, variants of $MEANPHASE$ and $MAXPHASE$ are used as factors instead. These variants take on the original values incremented by 1 when the phase is defined, and the value 0 when it is undefined.

$$MEANPHASE\_V = (1 - PH\_UDEF) * (MEANPHASE + 1) \quad \text{(IV.1)}$$

$$MAXPHASE\_V = (1 - PH\_UDEF) * (MAXPHASE + 1) \quad \text{(IV.2)}$$

H.264/AVC has significant differences from the earlier standards including variable block sizes. Also, MCEC will be quite successful in concealing areas where there is uniform motion. These suggest the use of a measure for *non-uniformity of motion* around the lost slice. The following factors were computed for this purpose:

- $AVGINTERPARTS$: Average number of Inter Macroblock partitions in a given slice
- $MAXINTERPARTS$: Maximum number of Inter Macroblock partitions in a given slice
- $INTRASLICE$: Boolean variable indicating whether the lost slice is coded as an intra slice
- SP_PH_ENT: Spatial Coherence of Phase of the motion [37]
- TP_PH_ENT: Temporal Coherence of Phase of the motion [37]
- SP_MAG_ENT: Spatial Coherence of Magnitude of the motion
- TP_MAG_ENT: Temporal Coherence of Magnitude of the motion

$AVGINTERPARTS$ and $MAXINTERPARTS$ directly relate to the predictability of the motion vector for a lost macroblock. A macroblock which has 16 Inter partitions will be less predictable than one with just one partition. The number of macroblock partitions in H.264/AVC can range from one for the coarsest partition to sixteen for the finest partition.

$INTRASLICE$ is set when the lost slice is coded as an intra slice. In the case of dual losses, there are two boolean factors $INTRASLICE1$ and $INTRASLICE2$. $INTRASLICE1$ is set if at least one of the lost slices is coded as an intra slice and $INTRASLICE2$ is set if both the lost slices are coded as intra slices.

Spatial Coherence of Phase is defined as the entropy of the phase information of motion vectors for all macroblocks in the lost slice and its neighboring slices. The candidate motion vectors for the calculation of entropy are taken from the current slice as well as the ones above and below it. The phase of the candidate motion vectors is computed and their histogram is obtained. The probability of a particular bin k in the histogram is

$$P_{sph}(k) = \frac{N_k}{L} \tag{IV.3}$$

where L denotes the total number of MBs considered and N_k is the total number of MBs whose phase lies in the kth bin. In this experiment, $L = 66$ (i.e. 3×Number of MBs per slice). Note that for the coherence factors, only those MBs whose phase is defined (i.e. the MBs which have non-zero motion vector) are considered. The Spatial coherence of phase is defined as:

$$SP\_PH\_ENT = \sum_{j=1}^{n} P_{sph}(j) \times log\Big(\frac{1}{P_{sph}(j)}\Big) \tag{IV.4}$$

where n denotes the total number of bins in the histogram. From the definition, it is clear that when motion around the lost slice is uniform, SP_PH_ENT is low, and when the motion is non-uniform, SP_PH_ENT is high.

The temporal coherence of phase is computed in a similar way by considering the current slice and the co-located slices in the future and past frames. Spatial and Temporal Coherence of magnitude are obtained by finding the entropy of the magnitude information of the candidate slices.

4. **Spatial Clutter:** It relates to the amount of high-frequency energy present in a particular region of the video. When there is a lot of high-frequency energy in a portion of an image, it might lead to lower packet loss visibility in that region due to masking effects. For example, a packet loss in a crowded scene might go unnoticed when compared to an error in a moving object in a clear blue sky. The following factors characterize the spatial clutter:

   - SD_Y_NORM, SD_U_NORM, SD_V_NORM: Normalized Standard Deviation of the luminance value (Y) and the two chrominance values (U and V) for pixels in a particular slice (normalization done with respect to the mean of the slice).
   - SD_YUV: Root Mean Square (RMS) value of the above three factors.

5. **Contrast:** We define a contrast measure using the Ratios (and squared differences) of the Variances (and Mean) of the current slice and a slice that is offset by 8 pixels above and below. These factors measure the variability in scene statistics across slice boundaries and may be useful in predicting visibility. If the variability is high, concealment of lost slices may not be very effective since a small error in motion estimation will cause a large error in concealment.

   The following 12 factors are used in the study:

   - $RATIO\_VAR\_Y$, $RATIO\_VAR\_U$, $RATIO\_VAR\_V$ - Maximum of ratio of variances of the current slice to offset slices.
   - $RATIO\_MEAN\_Y$, $RATIO\_MEAN\_U$, $RATIO\_MEAN\_V$ - Maximum of ratio of means of the current slice to offset slices.
   - $DIFFSQ\_VAR\_Y$, $DIFFSQ\_VAR\_U$, $DIFFSQ\_VAR\_V$ - Maximum of squared difference of variances of the current slice and offset slices.
   - $DIFFSQ\_MEAN\_Y$, $DIFFSQ\_MEAN\_U$, $DIFFSQ\_MEAN\_V$ - Maximum of squared difference of means of the current slice and offset slices.

6. **Side Match Distortion (SMD):** This metric is used to evaluate the estimated motion vector in the non-normative error concealment algorithm [38] of H.264/AVC. SMD is measured as the 1-norm distance between the border pixels of the current macroblock and its neighboring macroblock(s). SMD is a measure of blockiness across the macroblock boundary. Blocking artifacts, which occur due to erroneous motion vector recovery, can cause considerable visibility of packet loss. The average and maximum of SMD of the MBs in a given slice, denoted by $AVGSMD$ and $MAXSMD$ are used as factors.

We want to predict the visibility of individual and multiple packet losses. For the individual loss case, the factors listed above are used for modeling visibility. In the multiple loss case, our goal is to develop a generic model for visibility irrespective of the number (L) of individual packet losses involved in the multiple loss. However, there are L values now, for each of the factors, corresponding to each of the L packet losses involved. If K factors are available for each individual loss, there are LK available factors for the multiple loss. However, for ease of algorithm implementation and to reduce computational complexity, a generic model with the same set of factors irrespective of the value of L is necessary.

To satisfy the above requirement, a total of $2K + 5$ factors representing a multiple loss is derived as follows. The highest and lowest of the L values for each of the K factors give $2K$ new factors. They are named by attaching "HIGH_" or "LOW_" appropriately as prefix to the factor name. Similarly, 4 new factors are formed using the highest and lowest values of D_i and T_i, the spatial and temporal separation between each pair of packet losses. The number of packet losses in the multiple loss, L, is another factor.

In this chapter, we demonstrate the effectiveness of this framework in predicting visibility of dual losses ($L = 2$). In the dual loss case, there is only one pair of losses and so D and T are used directly as factors. Also, L is not used as a factor since it is kept constant.

IV.C Modeling Approaches

We solve both the regression and classification problems associated with predicting the visibility of a loss. For the regression problem, we use logistic regression described in Chapter II. In the case of GLMs, the set of factors chosen has a significant impact on the model's performance. To identify the factors that are important and to build a good model, we use a six-stage model building approach described in Chapter 4 of Hosmer and Lemeshow [39]. This is briefly summarized in the Subsection IV.C.1. CART, described in Chapter III, is used for the classification problem. Since CART automatically chooses which factors to split on, we do not apply the six-stage model building approach.

IV.C.1 Six-stage Model Building Approach

The six-stage approach to model building is summarized here.

1. The first stage involves a univariable analysis of each factor and is useful in identifying factors that show little association with the predicted variable (visibility, in our case). These factors do not contribute to the prediction process and are not considered further for multivariable analysis.

2. The second stage involves building a multivariable model using a stepwise approach for adding factors, one at a time. The word "model" is used to characterize the set of factors which comprise the matrix $\mathbf{X}$. The starting model is the Null model and at every step, the factor that causes the maximum decrease in deviance per degree of freedom is added. In order to calculate the decrease in deviance per degree of freedom, the decrease in deviance is divided by the number of degrees of freedom that a factor has. At each step, the MSE between actual and predicted probabilities for both training and 4-fold cross validation (CV) is recorded. The MSE (CV) decreases first and then starts increasing or remains the same. Thus a list of models with increasing numbers of factors and their corresponding MSE (CV) is formed.

3. In the third stage, the model in the list that corresponds to the first local minimum of cross-validated MSE is selected. The importance of each factor in the selected model is verified using the Likelihood Ratio Test (LRT). If a factor is not significant, it is dropped from the model. This process is continued until all the factors in the model are significant. This model is called the preliminary effects model.

4. In the fourth stage, the correct parametric representation for each factor in the model is verified. For example, a factor F might be better represented by F^2 instead of F. For this, a categorical factor with 4 levels is created from the ordinal factor using its three quartiles as the cutpoints. Then, the model is fitted with the categorical factor to obtain the coefficients for each level and a plot is made of the estimated coefficients versus the midpoints (obtained after applying the parametrization of interest) of the region for each category. The parametrization that results in the best linear plot and also reduces deviance is the correct parametrization. The model at this stage is called the main effects model.

5. In the fifth stage, any interaction factors that make intuitive sense are tried to check if they improve the prediction capability. Interaction factors are created as the product of pairs of main effect factors. This model is called the preliminary final model.

6. In the sixth stage, the importance of each factor in the preliminary final model is verified using the Likelihood Ratio Test and any insignificant factors are dropped from the model. This marks the completion of the model building process. The model at this stage is called the final model.

Table IV.1: Description of factors

Factor	Description
$MAXIMSE$	Maximum initial MSE of an MB in the lost slice
PH_UDEF	Boolean variable set when motion vector phase is undefined
$INTRASLICE$	Boolean variable set when lost slice is coded as an intra slice
$MAXPHASE\_V$	Phase (direction) of motion
$MAXRSENGY$	Maximum residual energy of an MB in the lost slice
$FRAMETYPE$	Type of frame in which packet loss occurred - B/P/I
$AVGINTERPARTS$	Average number of Inter partitions in lost slice
$VARMOTX$	Variance of motion in horizontal direction
HGT	Location of lost slice in the frame numbered from top to bottom

IV.D Results

This section is subdivided into two parts. In the first part, the results obtained using GLM for the regression problem are discussed. In the second part, the results obtained using CART for the classification problem are discussed.

IV.D.1 GLM Results

The model building process described in Section IV.C.1 is followed to get the final models in the individual and dual loss cases. In the first stage (univariable analysis), most factors had a strong association with visibility. A few factors (mostly spatial clutter and contrast-based factors) showed little association and were dropped from further consideration. In the individual loss case, the factors included in the preliminary effects model are $MAXIMSE$, PH_UDEF, $INTRASLICE$, $MAXPHASE\_V$, $MAXRSENGY$, $FRAMETYPE$, $AVGINTERPARTS$ and $VARMOTX$. A brief description of these factors is given in Table IV.1 for quick reference. The corresponding model in the dual loss case includes factors $HIGH\_MAXIMSE$, PH_UDEF2, $HIGH\_AVGINTERPARTS$, $INTRASLICE1$, $HIGH\_MAXPHASE\_V$,

Table IV.2: Factors and their coefficients in the final model (Individual Losses)

Factor	Coefficient
Constant γ	-2.750e+00
$MAXIMSE\_S$	5.141e-01
PH_UDEF	-1.419e+00
$MAXPHASE\_V$	-5.566e-01
$MAXRSENGY$	-4.481e-04
$FRAMETYPE-P$	8.333e-01
$FRAMETYPE-I$	9.778e-01
$AVGINTERPARTS$	-3.298e-01
$VARMOTX$	-1.861e-03
$INTER\_IL$	7.346e-04

$HIGH\_MAXRSENGY$, PH_UDEF1, $FRAMETYPE\_ML$ and $HIGH\_HGT$. In the fourth stage, $MAXIMSE\_S$ (the fourth root of $MAXIMSE$) is found to be a better representation for $MAXIMSE$. In the case of dual losses, the same scaling gives a better representation for $HIGH\_MAXIMSE$ denoted by $HIGH\_MAXIMSE\_S$. For all other factors, the linear treatment was found to be adequate. In the fifth stage, the interaction between $MAXRSENGY$ and $INTRASLICE$ (a boolean variable) is found to be useful. This interaction factor is named $INTER\_IL$ in the individual loss case and $INTER\_ML$ in the dual loss case. In the final stage, the factor $INTRASLICE$ is dropped in the individual loss case as it turned out to be insignificant when $INTER\_IL$ is included. Similarly $INTRASLICE1$ is dropped in the dual loss case.

$$INTER\_IL = MAXRSENGY * INTRASLICE \quad \text{(IV.5)}$$

$$INTER\_ML = HIGH\_MAXRSENGY * INTRASLICE1 \quad \text{(IV.6)}$$

The final model for the individual loss case has 8 factors. Its residual deviance is 5237.7 whereas the null deviance is 8597.5 and the MSE obtained during cross-validation is 0.0253. Similarly, in the dual loss case, the final model has 9 factors. Its residual deviance is 10402.2 whereas the null deviance is 17802.3 and the MSE obtained during cross-validation is 0.0398.

Table IV.3: Factors and their coefficients in the final model (Dual Losses)

Factor	Coefficient
Constant γ	-1.769e+00
$HIGH\_MAXIMSE\_S$	5.139e-01
PH_UDEF2	-1.813e+00
$HIGH\_AVGINTERPARTS$	-3.273e-01
$HIGH\_MAXPHASE\_V$	-6.127e-01
$HIGH\_MAXRSENGY$	-5.315e-04
PH_UDEF1	-1.767e-01
$FRAMETYPE\_ML - P$	8.862e-01
$FRAMETYPE\_ML - I$	1.249e+00
$HIGH\_HGT$	-5.114e-02
$INTER\_ML$	4.700e-04

The factors in the final models and their coefficients are listed in Tables IV.2 and IV.3 for the individual and dual loss cases, in the order in which they are added during the second stage of the model building process. The values of the coefficients do not necessarily convey the importance of corresponding factors because these factors have different variances and ranges. However, the sign of the coefficients is important and informs whether a packet loss is more visible with a high or low value for a factor. Most of the factors in the final models for individual losses and dual losses have one-to-one correspondence, and corresponding coefficients have the same sign.

The significance of different factors in a model can be understood by the increase in the deviance that results if each factor is individually removed from that model. Figure IV.5 shows the increase in deviance corresponding to each factor in the final model for the individual loss case. Figure IV.6 shows a similar bar graph for the dual loss case. From the figures, one can see that $MAXIMSE\_S$ is the most significant factor in predicting visibility, followed by $FRAMETYPE$ and $AVGINTERPARTS$.

The following conclusions can be made about the effect of factors on visibility based on the final models for both individual and dual loss visibility.

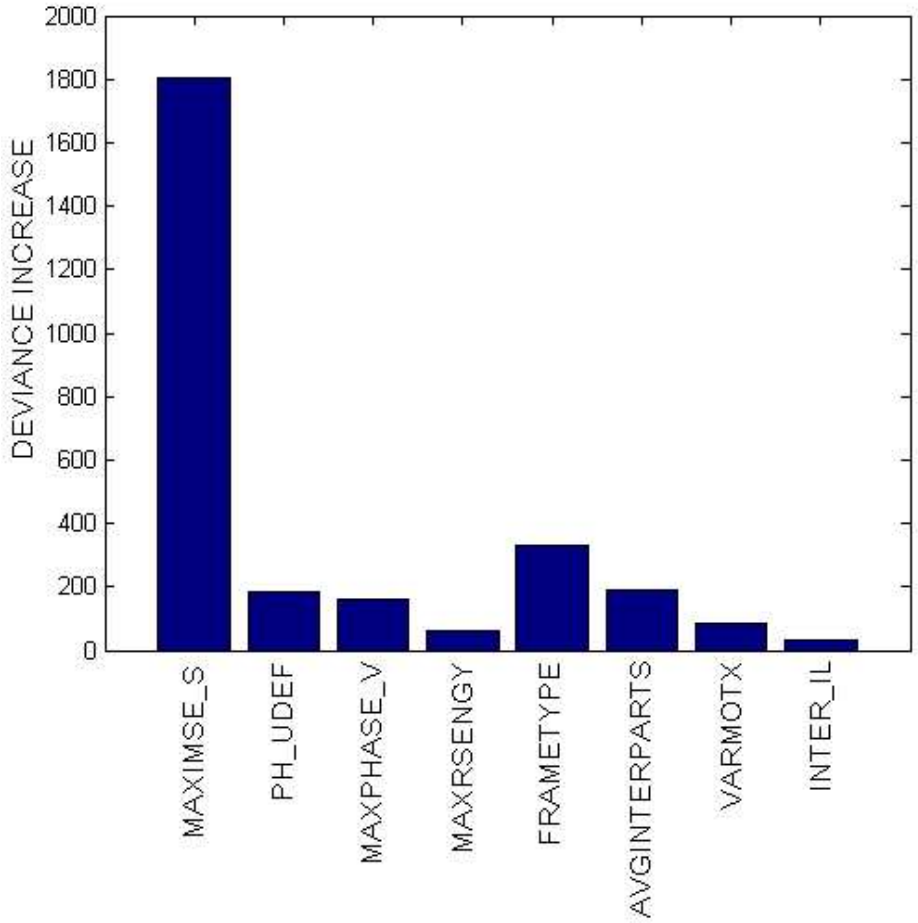


Figure IV.5: Factor Significance: Plot showing the increase in deviance when a factor is dropped (Individual Loss)

- Factor $MAXIMSE\_S$ is directly proportional to visibility. This makes intuitive sense. If the initial MSE of a loss is high, one would expect the loss to be more visible. Visibility increases, as expected, in the order B, P and I for $FRAMETYPE$.

- Factors $MAXRSENGY$, $AVGINTERPARTS$, $VARMOTX$ and $HIGH\_HGT$ are inversely related to visibility. A high value for residual energy can occur when the signal has a lot of high frequency content and the motion is inconsistent (for example, a market crowded with peo-

ple). In such a case, visibility is reduced due to masking effects. When $AVGINTERPARTS$ is large, it means the macroblock needs to be subdivided in order to achieve good motion compensation; one motion vector for the whole macroblock does not suffice. This means the motion is somewhat complex and occurring on a fine scale; this complexity may be providing a spatial and temporal masking which masks the loss and makes it less visible. Similarly, a large $VARMOTX$ means that the motion is not smooth and is highly variable across the slice. The negative coefficient for $HIGH\_HGT$ indicates that viewers' sensitivity to losses goes down as we move from the top to the bottom of the frame. When PH_UDEF is set (i.e., when the majority of the motion vectors in the slice are (0,0)), visibility decreases since concealment works well.

- The effect of factors $MAXIMSE\_S$, $FRAMETYPE$, $MAXRSENGY$, $VARMOTX$, $HIGH\_HGT$ is consistent with our earlier findings from Chapter II. However, in our earlier work, the magnitude of underlying motion was a highly significant factor for predicting visibility. ZMEC was used then, as opposed to MCEC now. Due to this, motion is no longer a significant factor. Losses with non-zero motion, which used to be visible with ZMEC, are no longer visible when MCEC is used. This explains the sharp increase in the percentage of invisible losses compared to our previous subjective test.

- Horizontal motion causes losses to be more visible than vertical motion. This is due to the fact that $MAXPHASE\_V$ (ranges from 0 to $\pi/2$) is a significant factor and has a negative coefficient. One explanation for this is based on the slice structure used. A slice is assumed to be a horizontal row of macroblocks, and a packet loss causes the loss of a slice. When there is horizontal motion, vertical edges longer than a macroblock cause discontinuous edges when concealed, and horizontal edges do not cause any new artifacts since the motion is horizontal. On the other hand, when there is vertical mo-

tion, vertical edges do not cause new artifacts, and horizontal edges either appear twice or disappear completely depending on whether the concealing slice contains the horizontal edge or not. However, they will not cause discontinuous edges. If the slice structure was a vertical row of macroblocks, vertical motion might cause losses to be more visible than horizontal motion.

In our previous experiment from Chapter II, the factor MOTA representing the direction of motion was found to be insignificant. However, the losses were concealed then using ZMEC and so motion caused losses to be visible irrespective of direction. When losses are concealed using MCEC, many losses becomes invisible and the amount of motion is no longer a significant factor. Since MCEC is not perfect due to the difference between the actual and estimated motion vector, visible artifacts are caused based on direction of motion.

Figure IV.7 illustrates both the concepts in a graphical manner. As we can see from frames N-1 and N, the upper square is moving vertically upward and the lower square is moving horizontally to the right. On the decoder side, frame N is affected by two packet losses and so two slices are lost. As we can see, visual artifacts are caused for both the squares with ZMEC irrespective of the direction of motion. However, with MCEC, the upper square (vertical motion) is concealed with out any visual artifacts (even with imperfect motion estimation), while the lower square (horizontal motion) has some discontinuous edges.

To implement the final model either inside the network or at the decoder, one needs access to the factors involved in the model. In our current study, these factors are transmitted for each slice from the encoder along with the compressed bitstream. For example, in the individual loss case, among the 8 distinct factors in the final model, $FRAMETYPE$ is a content independent factor. So only 7 factors are transmitted. If 1 byte for each factor is used and a typical video compression

ratio of 100 : 1 is assumed, an overhead of approximately 8% is incurred. Obviously, for applications at the encoder such as perceptual error control using unequal error protection, or for packet prioritization, this overhead for transmitting the factors is not required.

IV.D.2 CART Results

CART is used to design a classifier that classifies losses as visible or invisible. The performance of the classifier is evaluated using two metrics - False Alarm rate (FA) and False Dismissal rate (FD), which are defined as follows:

$$\text{FA} = \frac{\text{Number of invisible losses classified as visible losses}}{\text{Total number of invisible losses}} \tag{IV.7}$$

$$\text{FD} = \frac{\text{Number of visible losses classified as invisible losses}}{\text{Total number of visible losses}} \tag{IV.8}$$

. Overall accuracy is used as the metric for classifier performance previously. While this metric is typically used for performance analysis, it has certain drawbacks. It does not inform the individual percentages of false alarms and false dismissals and both errors are treated as having equal misclassification cost. However, in practice, false dismissals are usually more expensive errors compared to false alarms. Second, in the subjective data, there is a large percentage of invisible losses, 96% for individual losses and 91% for dual losses. Therefore one can define a default classifier, which classifies all losses as invisible, and this trivially achieves a fairly high accuracy of 96% or 91%. However, FD for the default classifier is 100% (all visible losses are dismissed as invisible) while FA is 0%. Hence FA and FD are used to characterize the performance of a classifier.

As opposed to GLM, CART is able to automatically choose, from the complete set of factors, useful ones for predicting visibility. At each stage, CART selects the best factor on which to split. CART is deployed with all the factors listed in section IV.B and it designs a classifier that classifies each loss as visible or invisible. CART allows us to give more importance to false dismissals by assigning different misclassification costs for different types of errors. Here, we explore the

effect of varying misclassification costs on FA and FD. The cost associated with false alarms is fixed at unity, while the cost associated with false dismissals (C_{FD}) is varied from 1 to 100.

Figures IV.8 and IV.9 show the performance of CART classifiers with varying misclassification costs (C_{FD}) for individual and dual losses. As C_{FD} increases, it can be observed that the decline in FD is rapid while the rise of FA is only gradual. The operating value of C_{FD} can be chosen based on the type of application for which the classifier is being designed.

For $C_{FD} = 10$, the false dismissal rate is 9.76% and the false alarm rate is 16.96% for individual losses. We can compare this against $C_{FD} = 1$, where the false dismissal rate is 100% and the false alarm rate is 0%. So we see that with a relatively small increase in false alarm rate, the false dismissal rate reduced significantly. Similarly, for dual losses, a significant reduction (59.62% to 13.46%) in false dismissal rate is achieved with a relatively small increase in false alarm rate (1.84% to 17.73%). The classifiers CART designed for $C_{FD} = 10$ are simple, and either $MAXIMSE$ or $AVGIMSE$ is one of the most influential factors. In the individual loss case, the entire classifier is just a threshold on $MAXIMSE$. If $MAXIMSE > 402.6$, the loss is visible; otherwise, the loss is invisible. The classifier for the dual loss case is shown in Figure IV.10. As we can see, the first and second splits are on $HIGH\_AVGIMSE$. From these observations, it can be concluded that an MSE-based factor alone is enough to identify most of the visible losses if some false alarms can be tolerated.

IV.E Conclusion

In this chapter, we considered the problem of modeling the visibility of individual and multiple packet losses in H.264/AVC bitstreams. We proposed a new model framework to predict the visibility of a multiple packet loss and we successfully demonstrated the performance of this framework for the case of dual

losses. We explored the importance of new factors in predicting visibility. We modeled the problem of predicting visibility as a regression as well as a classification problem. The regression was carried out with GLMs using a well-established model building approach. We used CART for the classification problem and demonstrated the effect of varying misclassification costs on false alarm and false dismissal rates.

We now summarize our observations based on this experiment:

1. Of all packet losses, 12% have indeterminate visibility (between 25% and 75% of viewers responded to the loss). Of the 88% of packet losses that are not indeterminate, 96% are invisible, which shows the effectiveness of MCEC in concealing packet losses.

2. As opposed to our previous work using ZMEC, the amount of motion is no longer a significant factor in predicting visibility. However, other factors such as $MAXIMSE\_S$, $FRAMETYPE$ and $MAXRSENGY$ continue to be significant and their effect on visibility is consistent with our earlier findings.

3. Among the new content dependent factors considered, $AVGINTERPARTS$ based on variable block size in H.264/AVC and factors PH_UDEF and $MAXPHASE\_V$ based on motion phase turned out to be significant factors.

4. Horizontal motion causes losses to be more visible than vertical motion. Factors $MAXIMSE\_S$ is directly proportional to visibility. Factors $MAXRSENGY$, $AVGINTERPARTS$, $VARMOTX$ and $HIGH\_HGT$ are inversely related to visibility.

5. An MSE-based factor alone can be used to identify most of the visible losses if false alarms can be tolerated on the order of 15-20%.

IV.F Acknowledgements

This work was supported in part by the Center for Wireless Communications at UCSD and by the UC Discovery Grant Program.

Chapter IV of this dissertation, in part, is a partial reprint of the material as it appears in S. Kanumuri, S. G. Subramanian, P. C. Cosman and A. R. Reibman, "Packet Loss Visibility and Packet Prioritization in H.264/AVC Videos", *IEEE Trans. Image Processing* (submitted). Co-author Subramanian and I contributed equally towards this publication. Co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter IV.

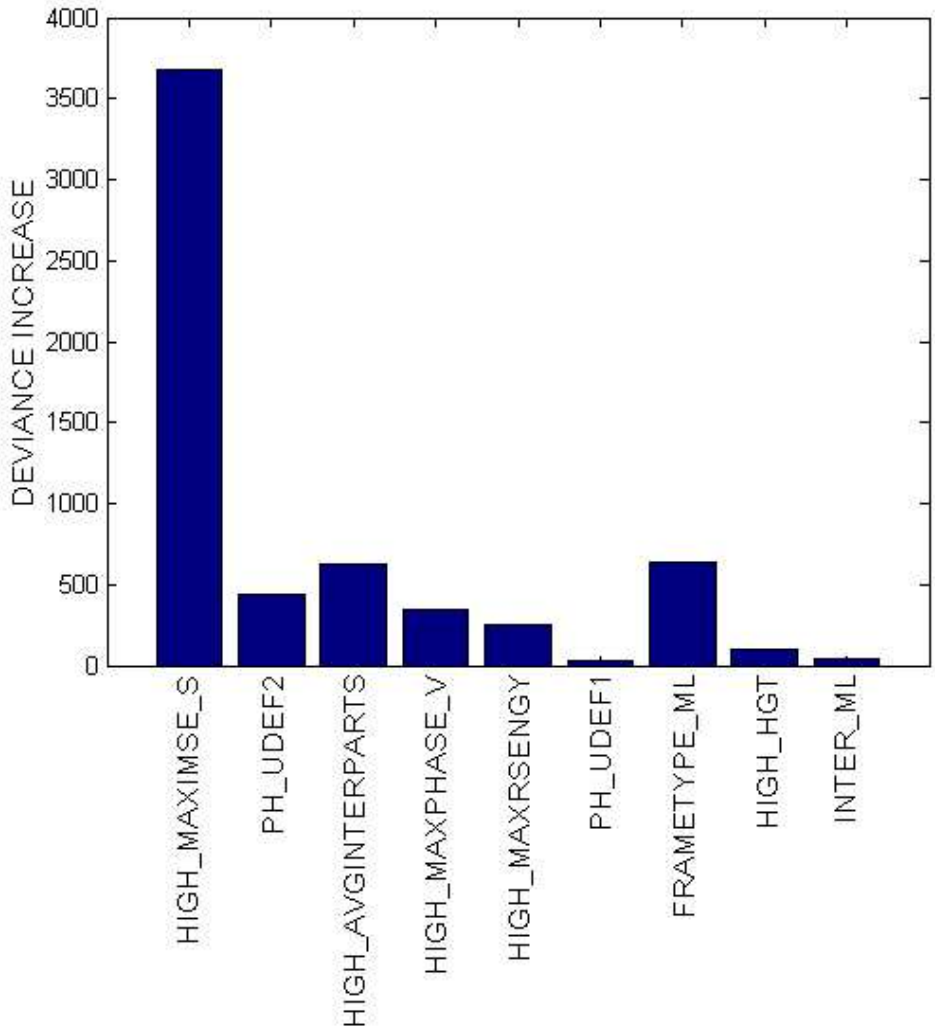


Figure IV.6: Factor Significance: Plot showing the increase in deviance when a factor is dropped (Dual Loss)

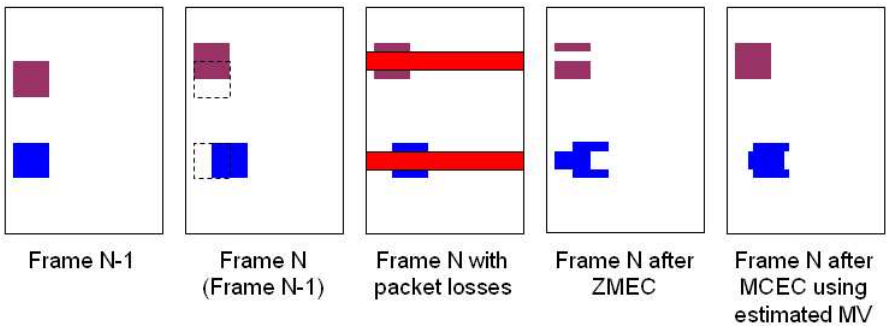


Figure IV.7: A cartoon example to show that horizontal motion causes losses to be more visible than vertical motion with MCEC

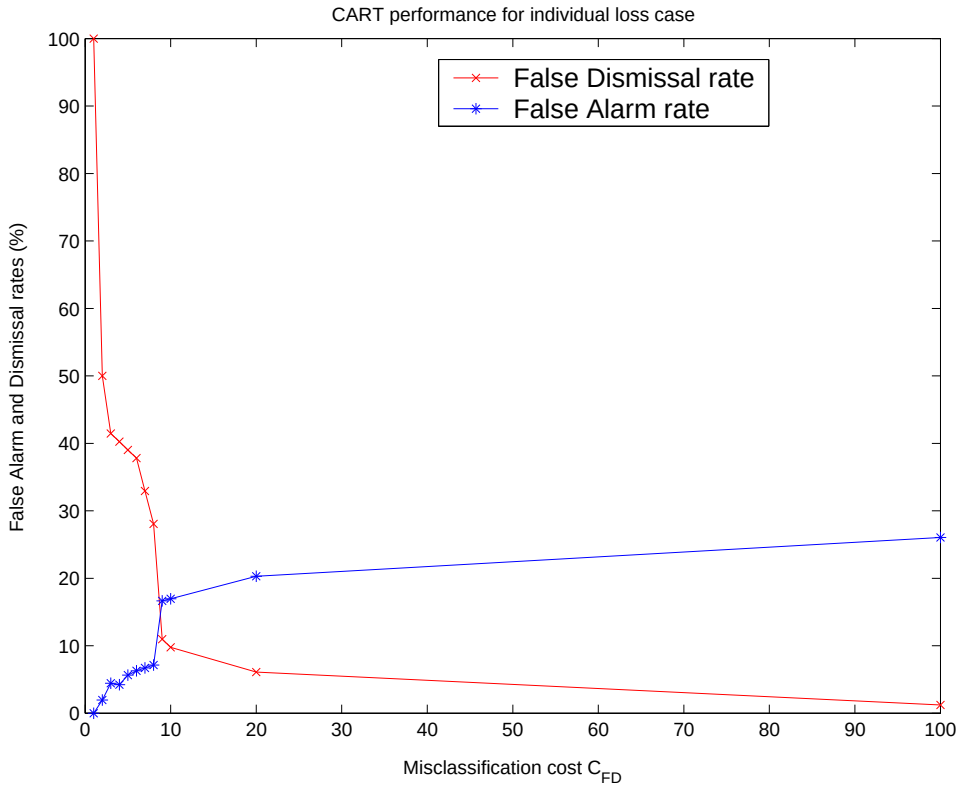


Figure IV.8: Classification performance for different misclassification costs (Individual Loss)

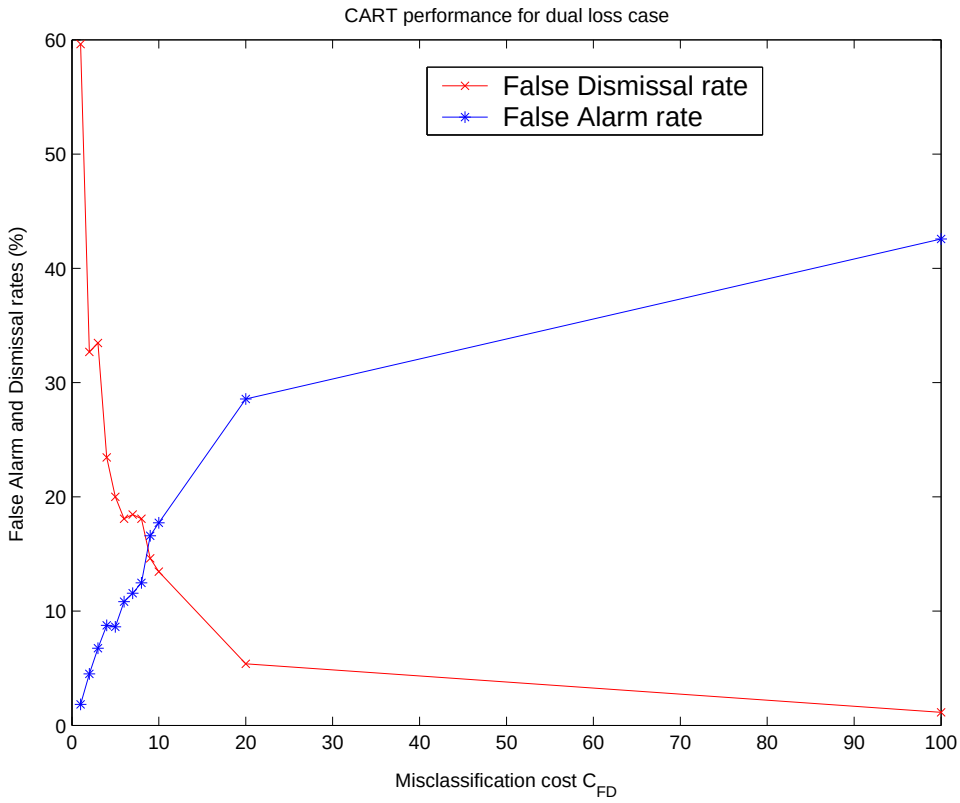


Figure IV.9: Classification performance for different misclassification costs (Dual Loss)

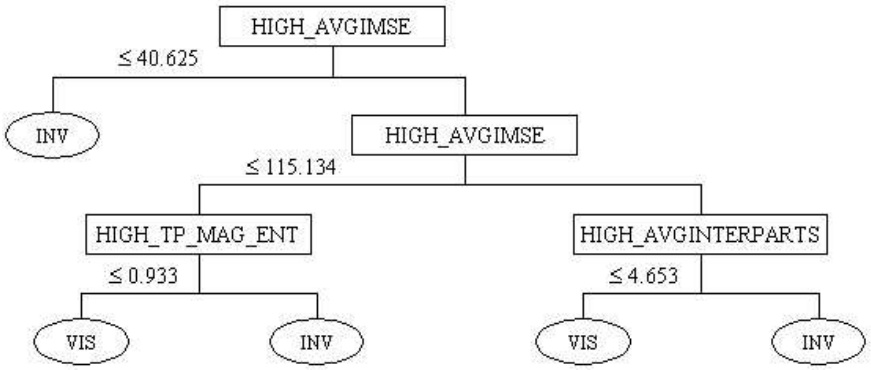


Figure IV.10: CART classifier for dual loss case

V

Packet Prioritization

Prioritization of video data has been proposed as early as [40] which uses ideas of data partitioning for MPEG videos. Traditionally, video packet prioritization methods use information available from the encoding process and are independent of scene statistics of the video source. These methods can be broadly classified into two categories: data partition methods and layered coding methods.

Data partition methods partition compressed data into different types and assign priorities. For example, data representing picture headers, macroblock modes, motion vectors or DC coefficients is given a higher priority, while data representing AC coefficients is given a lower priority [40, 41]. Similarly, data representing intra frames is given a higher priority compared to data representing inter frames [42, 43].

Layered coding methods, for example bit plane coding, code a coarse representation of the video into the base layer and progressively refine it using enhancement layers (bit planes). Due to the nature of the coding, the base layer is the most important as the video cannot be decoded without it. Similarly, the first enhancement layer is more important than the second one and so on. Van der Schaar et al. [44] and Vehkapera et al. [45] show the usefulness of the natural priorities offered by layered coding methods. Layered coding methods have lower compression efficiency and higher computational complexity compared to non-

layered methods and they do not prioritize within the base layer.

Yang et al. [46] introduced a priority assignment method that uses both data partitioning and layered coding. However, all of these methods are dependent only on the encoding process and are not dependent on the scene statistics of the video. For example, packets from moving scenes are more important than packets from still scenes. Shin et al. [47] proposed a heuristic method to assign priorities based on scene statistics such as coding modes and underlying motion. However, it is neither developed nor verified using subjective tests.

In chapter IV, we developed a model, based on subjective test data, that uses various factors extracted from the underlying scene to predict the probability of visibility of a packet loss. Packets can be assigned different priorities at the encoder based on the visibility predicted by this model. In this chapter, we consider a transmission scenario where packets are dropped at a congested node in the network and show that a priority-based packet drop policy outperforms a conventional DropTail policy in terms of received video quality.

This chapter is organized as follows: Section V.A explains the packet priority assignment method. Section V.B describes the design of our experiments while Section V.C discusses the results. Section V.D concludes.

V.A Priority Assignment

The assignment of priorities to packets using the visibility model developed in Chapter IV is explained here. All packets are assigned one of two priority levels at the encoder. A packet is assigned low priority if the probability of visibility for the packet's loss is less than a cutoff. Else, it is assigned a high priority. The cutoff is chosen as 0.25 which is the midpoint between certain invisibility (0) and complete ambiguity (0.5). One can assign more priority levels and/or choose different cutoffs depending on the required distribution of packets over priority levels.

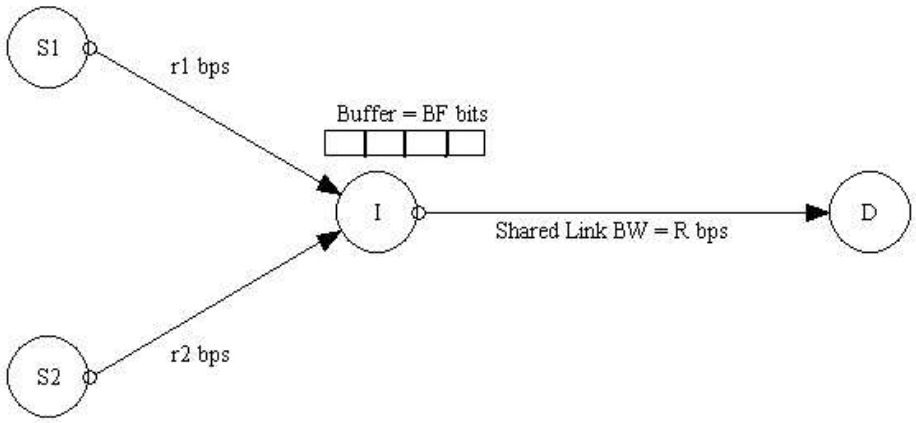


Figure V.1: Topology of experimental network

V.B Experimental Design

When digital video is transmitted over a network using protocols such as UDP [48], packets are dropped at a network node in the event of congestion. Typically, the DropTail [48] policy is employed wherein the last (tail) packets to arrive at the node are dropped. However, if priority information of the packets is known, one can drop packets intelligently to reduce the quality deterioration due to packet losses. To demonstrate this and to quantify the improvement, we constructed an experimental network which has two video sources S1 and S2 and one destination D as shown in Figure V.1. There is an intermediate node I which is the start of the bottleneck link and it has a buffer of size BF bits. The bottleneck link's bit-rate is constant at R bps and each of the two sources produces traffic that occupies half the link bit-rate on the average. However, the instantaneous rates of the sources ($r1$ and $r2$ bps) need not add up to R bps. Whenever $r1 + r2 > R$, packets accumulate in the buffer. If this condition persists, the buffer will eventually overflow and packets must be discarded. Whenever $r1 + r2 < R$ instantaneously, the queue diminishes.

For the simulations, six videos of 10s duration each are chosen. The videos are of SIF resolution (352×240) at 30fps. They are encoded and decoded using

the extended profile of H.264/AVC JM Version 9.1 Codec. The encoding structure is I B P B P B...P B with a GOP size of 20 frames. For P frames, two reference frames are used for motion compensation - a long-term reference frame and a short-term reference frame. The long-term reference frame is always the I frame of the current GOP and the short-term reference frame is the previous P frame. B frames use the future P frame and either the long-term or short-term reference frame for bidirectional prediction. As the bottleneck link bit-rate R varies in our experiment from 300Kbps to 1200Kbps, the two source videos are always coded at an average bit-rate of $R/2$ using the codec's built-in rate control algorithm.

When a packet is dropped, the motion-compensated error concealment (MCEC) algorithm used here estimates the motion vector and the reference frame for each of the lost macroblocks and conceals them with the macroblocks predicted using their estimated motion vectors. Motion compensation in H.264/AVC can occur at different levels from the macroblock level to the smallest block level (4×4 pixel block). Accordingly, each macroblock can have a different number of motion vectors ranging from 1 to 16. These motion vectors can use different reference frames because of multiple frame prediction. A set of motion vectors is formed from motion vectors of blocks around the lost macroblock. The frame that is referenced the most number of times in the set is selected for concealment. The estimated motion vector is the median of all the motion vectors in the set that refer to this selected frame.

The network simulator ns-2 [49] is used for simulations. In each simulation, two videos are transmitted simultaneously as sources S1 and S2. Packets belonging to both videos compete for space in the queuing buffer at node I. When the buffer gets full, packets are dropped in accordance with a drop policy. Two received (lossy) videos, one for each source, are assembled from arriving packets at destination D.

Two drop policies, DropTail (DT) and DropLowPri (DL), are compared in terms of received video quality. Both policies drop packets from the two different

sources alternately to ensure fairness among the sources. When a packet has to be dropped because the buffer is full, the DropTail policy drops whichever packet (from the source whose turn it is for dropping) is closest to the tail with no consideration for the priority of the packet. The DropLowPri policy drops whichever *low priority* packet in the buffer (from the appropriate source) is closest to the tail.

The six videos are divided into two sets (set1 and set2) such that each set has three types of videos: one video with 'still'/no motion (collection of still pictures), one video with 'low' motion (slow camera panning) and one video with 'high' motion (sports). Six pairs of videos are formed, three within each set, as follows: (still1,low1), (still1,high1), (low1,high1) from set1 and (still2,low2), (still2,high2), (low2,high2) from set2. Three pairs of videos are formed across the two sets as follows: (still1,still2), (low1,low2) and (high1,high2). Using the nine pairs, we have a balanced representation wherein each type of video is competing twice with the traffic of all the three types (still, low and high). Each simulation of a pair of videos produces two lossy videos, one for each source video. Also, each simulation is run twice, once with each of the two drop policies. The nine pairs of videos result in 18 lossy videos for each drop policy leading to 18 comparisons.

The videos resulting from the two drop policies are evaluated for quality by comparison with their lossless versions using the VQM metric developed by ITS [50]. Loke et al. [51] found that the VQM metric is better correlated to human perception than two competing metrics, DVQ and VSSIM. The VQM score represents the amount of perceptual difference between the reference and test videos and it ranges from 0 (excellent quality) to 100 (poorest possible quality). If the difference between the VQM scores obtained by the two policies is less than 1.0, then it is called a 'Tie'. Else, the policy with the lowest score wins.

Table V.1: Performance comparison for varying values of R with BF = 120Kbits

R	Number of wins			Packet losses (%)	
	DL	Tie	DT	DL	DT
600Kbps	11	5	2	2.41	1.40
800Kbps	10	3	5	3.62	2.84
1000Kbps	11	2	5	5.16	4.18
1200Kbps	11	1	6	6.12	5.48

V.C Results

We examined the relative performance of the two drop policies over various settings for R and BF. Tables V.1, V.2 and V.3 show the number of wins for DropLowPri (DL) and DropTail (DT), the number of ties, and the percentage of packet losses (averaged over all the lossy videos) with DL and DT policies for different values of R and BF. The percentage of packet losses can be increased by increasing R (and thus allowing an increase in source rates $r1$ and $r2$) while holding the buffer size BF constant. It can also be increased by reducing BF while holding R constant.

In Table V.1, we fix the buffer size BF at 120Kbits and vary the bottleneck link bit-rate R. It can be seen that DropLowPri wins substantially more times than DropTail. Also, as R increases, the number of ties decreases. This is because, for a fixed buffer size, increasing R (and therefore increasing the average bit-rate to $R/2$ for each source) will cause more packets to be dropped so there are more opportunities for ties to get resolved.

In Table V.2, we hold the bottleneck link bit-rate R constant at 800Kbps and vary the buffer size BF. Again, DropLowPri performs better than DropTail for all the settings. As BF decreases from 200Kbits to 20Kbits, the number of ties decreases initially but then increases again. This is because when the buffer is relatively large, the percentage of packet losses is low, and there are more ties because the difference in scores will be small as both policies produce high quality. As packet losses increase, the number of ties goes down since there are

Table V.2: Performance comparison for varying values of BF with $R = 800$Kbps

BF	Number of wins			Packet losses (%)	
	DL	Tie	DT	DL	DT
200Kbits	10	3	5	1.83	1.12
120Kbits	10	3	5	3.62	2.84
80Kbits	12	1	5	5.84	4.66
60Kbits	10	2	6	7.88	6.45
40Kbits	9	2	7	16.36	14.69
20Kbits	7	6	5	47.36	44.44

Table V.3: Performance comparison for videos with apparent compression artifacts ($R = 300$Kbps)

BF	Number of wins			Packet losses (%)	
	DL	Tie	DT	DL	DT
60Kbits	8	6	4	1.84	1.19
40Kbits	8	3	7	3.56	2.61

more opportunities for DropLowPri to distinguish itself. However, the trend will reverse after a certain point. The ties increase now because the degradation in quality is so high that all drop policies perform very poorly and the quality is not perceptually different. However, a marginal performance gain for DropLowPri can still be seen for high percentages of packet losses (BF set to 40Kbits and 20Kbits in Table V.2).

The tables show that the DropLowPri policy outperforms the DropTail policy for all the different settings. This shows that source-dependent packet prioritization is useful in improving received video quality. Though our visibility model was designed for videos without apparent compression artifacts, our results show that it is also applicable for videos with such artifacts (see Table V.3).

V.D Conclusion

In this chapter, we considered the problem of packet prioritization in compressed video. Existing methods for prioritizing packets based on data partitioning

and layered coding do not address the importance of a packet from the perspective of scene statistics. We proposed a new source-dependent packet prioritization method based on the importance of packets from a visibility perspective. Although developed using human observer experiments, this packet prioritization approach is fully automated. We demonstrated its success by showing that a priority-based modified DropTail policy outperforms a standard DropTail policy for a wide variety of settings. We can also use this method with a Random Early Drop (RED) queue which foresees a congestion event and drops packets early before the buffer gets full. A RED queue helps to increase the separation between dropped packets, thereby avoiding burst losses, at the cost of causing excessive drops. The proposed packet prioritization method is not an alternative to existing methods but complements them. A comprehensive method that utilizes source scene statistics, data partitioning and layered coding to assign packet priorities is yet to be developed.

V.E Acknowledgements

This work was supported in part by the Center for Wireless Communications at UCSD and by the UC Discovery Grant Program.

Chapter V of this dissertation, in part, is a reprint of the material as it appears in S. Kanumuri, P. C. Cosman and A. R. Reibman, "Source-dependent Video Packet Prioritization based on a Visibility Model", *IEEE ICASSP*, April 2007 (submitted). I was the primary author, and the co-authors Prof. Cosman and Dr. Reibman directed and supervised the research which forms the basis for Chapter V.

VI

Conclusion

In this dissertation, we have proposed models to predict the visibility of packet losses, and demonstrated one application of the knowledge of packet loss visibility. We treated the problem of predicting packet loss visibility as a regression as well as a classification problem. We developed models for packet loss visibility in MPEG-2 and H.264/AVC bitstreams. We also extended the problem to predict the visibility of multiple packet losses. Finally, we used the models developed to prioritize packets and demonstrated the use of priority information in improving received video quality in a network. We now enumerate our contributions, categorized according to the chapter in which they appear.

In Chapter II, we considered the regression problem for visibility of individual packet losses in MPEG-2 videos.

1. A generalized linear model (GLM) is used to model the probability that a packet loss causes visible artifacts.

2. The representation for factors such as motion is examined to best predict visibility. As a result, the overall magnitude of motion is used instead of the horizontal and vertical components of motion; FRAMETYPE is used instead of TMDR. Insignificant factors such as MOTA are dropped.

3. A cross-validated MSE of 0.0627 between the actual and predicted probabil-

ities is achieved.

In Chapter III, we considered the classification problem for visibility of individual packet losses in MPEG-2 videos.

1. The Classification and Regression Trees (CART) algorithm is used to classify packet losses as visible or invisible.
2. A cross-validation classification accuracy of 93% is achieved in the RR case.
3. GLM models from Chapter II are also used for classification. However, the CART-based classifiers outperform the GLM-based classifiers.

In Chapter IV, we considered the problem of modeling the visibility of individual and multiple packet losses in H.264/AVC bitstreams.

1. A new model framework is proposed to predict the visibility of a multiple packet loss and the successful performance of this framework is demonstrated for the case of dual losses.
2. The importance of new factors in predicting visibility is analyzed.
3. The effect of different factors on visibility is explained.
4. GLM is used for the regression problem and CART is used for the classification problem.

In Chapter V, we considered one application of packet prioritization.

1. A new fully automated source-dependent packet prioritization method is proposed based on the importance of packets from a visibility perspective.
2. A priority-based modified DropTail policy outperforms a standard DropTail policy for a wide variety of settings.

VI.A Future Work

Existing methods for prioritizing packets based on data partitioning and layered coding do not address the importance of a packet from the perspective of scene statistics. We proposed a new source-dependent packet prioritization method based on the importance of packets from a visibility perspective. A comprehensive method that utilizes source scene statistics, data partitioning and layered coding to assign packet priorities is yet to be developed.

Other applications of the knowledge of packet loss visibility also need to be explored in the future.

1. A network quality monitor is a necessary tool for network service providers in implementing applications such as video services with cost-based quality. The knowledge of packet loss visibility will be very useful in building an accurate, real-time network quality monitor.

2. An intelligent encoder that gives unequal error protection to packets based on their perceptual importance is yet to be designed.

VII

Appendix

A. Instructions to Viewer

In this subjective test, you will be shown a set of 6 videos of 6 minutes duration each, which are affected by packet losses. These losses may cause some artifacts or glitches in the video, which can be easily recognized not to be part of the normal video. Your role is to press the space bar whenever you see such an artifact or glitch. During the course of the test, sit comfortably and please refrain from changing the position of your chair or leaning forward. In order that you get practice with identifying the artifacts, we will be presenting a pilot video of 1 minute duration.

B. Consent Form

SAMPLE COPY

University of California, San Diego
Consent to Act as a Research Subject

Studying the visibility of packet losses in a video stream.

Pamela Cosman, Ph.D. and her associates are conducting a research study to find out more about factors influencing the visibility of packet losses in a video stream. You have been asked to take part because you are over 18, with normal vision.

If you agree to be in this study, the following will happen to you:

- You will be asked to participate in a session which runs for 45 minutes.
- In a session, you will be asked to look at a computer screen and view several video sequences.
- You will be asked to respond to the visible glitches in the video sequences by pressing a key.

Participation in this study should not involve any risk or discomfort, other than possible fatigue or boredom.

You will receive $10 for a 45 minute session. There will not be any distinct benefit to you from these procedures. The investigator, however, may learn more about the visibility of packet losses in a video.

Participation in research is entirely voluntary. You may refuse to participate or withdraw at any time without jeopardy or loss of any benefits to which you are entitled.

Dr. Pamela Cosman and/or _____ has explained this study to you and has answered your questions.

Research records will be kept confidential to the extent allowed by law.

If you have any other questions, you may reach Pamela Cosman at pcosman@ece.ucsd.edu or (858)-822-0157. You may call the Human Research Protections Program at (858)-455-5050 to inquire about your rights as a research subject or to discuss research-related problems.

You have received a copy of this consent document to keep and the Experimental Subject's Bill of Rights

You agree to participate.

_____ _____ _____

Subject's signature Witness Date

Bibliography

[1] M. Masry and S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions", *Signal Processing: Image Communications*, Special issue on Video Quality, vol. 19, issue 2, February 2004.

[2] P. Gastaldo, S. Rovetta and R. Zunino, "Objective Quality Assessment of MPEG-2 Video Streams by using CBP Neural Networks", *IEEE Trans. on Neural Networks*, vol. 13, pp. 939-947, July 2002.

[3] C. J. Van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza and M. Kunt, "Quality assessment of motion rendition in video coding", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 5, pp. 766-782, August 1999.

[4] M. Miyahara, K. Kotani and V. R. Algazi, "Objective picture quality scale (PQS) for image coding", *IEEE Trans. Communications*, vol. 46, no. 9, pp. 1215-1226, September 1998.

[5] K. T. Tan and M. Ghanbari, "A Multi-Metric Objective Picture-Quality Measurement Model for MPEG Video", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 7, October 2000.

[6] P. Brun, G. Hauske and T. Stockhammer, "Subjective assessment of H.264-AVC video for low-bitrate multimedia messaging services", *IEEE ICIP*, vol. 2, pp. 1145-1148, Oct 2004.

[7] W. Gao, C. Mermer and Y. Kim, "A de-blocking algorithm and a blockiness metric for highly compressed images", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 12, pp. 1150-1159, Dec 2002.

[8] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding", *Signal Processing Letters, IEEE*, vol. 4, no. 11, pp. 317-320, Nov 1997.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600-612, April 2004.

[10] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model", *Proc. SPIE, Human Vision and Electronic Imaging X*, vol. 5666, January 2005.

[11] S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system, *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2005.

[12] P. Marziliano, F. Dufaux, S. Winkler and T. Ebrahimi, "A no-reference perceptual blur metric", *IEEE ICIP*, vol. 3, pp. 57-60, June 2002.

[13] O. Verscheure, P. Frossard and M. Hamdi, "Joint Impact of MPEG-2 Encoding Rate and ATM Cell Losses on Video Quality", *Global Telecommunications Conference (GLOBECOM)*, vol. 1, pp. 71-76, November 1998.

[14] O. Verscheure, P. Frossard and M. Hamdi, "User-oriented QoS analysis in MPEG-2 video delivery", *Journal of Real-Time Imaging*, vol. 5, pp. 305-314, October 1999.

[15] J. Lu, M. Chatterjee, M. D. Schwartz, M. K. Ravel and W. M. Osberger, "Measuring ATM video quality of service using an objective picture quality model", *Proc. SPIE, Multimedia Systems and Applications II*, vol. 3845, pp. 290-297, September 1999.

[16] A. E. Conway and Y. Zhu, "Applying Objective Perceptual Quality Assessment Methods in Network Performance Modeling", *Proc. Eleventh Int'l Conf. on Computer Communications and Networks*, pp. 116-223, October 2002.

[17] Verizon Laboratories (G. W. Cermak), "Videoconferencing Service Quality as a function of bandwidth, latency, and packet loss", T1A1.3/2003-026, May 2003.

[18] B. Chen and J. Francis, "Multimedia Performance Evaluation", AT&T Technical Memorandum, February 2003.

[19] S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks", *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 12, no. 12, pp. 1071-1083, Dec 2002.

[20] C. J. Hughes, M. Ghanbari, D. E. Pearson, V. Seferidis and J. Xiong, "Modeling and subjective assessment of cell discard in ATM video", *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 212-222, April 1993.

[21] S. Winkler and R. Campos, "Video quality evaluation for internet streaming applications", *Proc. SPIE, Human Vision and Electronic Imaging VIII*, vol. 5007, pp. 104-115, January 2003.

[22] K. Brunnstrom and B. N. Schenkman, "Quality of video affected by packet loss distortion, compared to the predictions of a spatio-temporal model", *Proc. SPIE, Human Vision and Electronic Imaging VII*, vol. 4662, pp. 149-158, January 2002

[23] A. Watson and M. A. Sasse, "Measuring perceived quality of speech and video in multimedia conferencing applications", *ACM International Conference on Multimedia*, pp. 55-60, April 1998.

[24] M. S. Moore, J. M. Foley and S. K. Mitra, "Detectability and annoyance value of MPEG-2 artifacts inserted into uncompressed video sequences", *Proc. SPIE, Human Vision and Electronic Imaging V*, vol. 3959, pp. 99-110, San Jose, CA, January 2000.

[25] M. S. Moore, S. K. Mitra and J. M. Foley, "Defect Visibility and content importance implications for the design of an objective video fidelity metric", *IEEE ICIP*, vol. 3, pp. 45-48, June 2002.

[26] M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psycophysical evaluation and analysis", *Journal of the Optical Society of America, A*, vol. 18, no. 10, pp. 2385-2397, October 2001.

[27] D. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortion", *Journal of the Optical Society of America, A*, vol. 20, no. 7, pp. 1164-1180, July 2003.

[28] A. R. Reibman, V. Vaishampayan and Y. Sermadevi, "Quality monitoring of video over a packet network", *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 327-334, April 2004.

[29] P. McCullagh and J. A. Nelder, "Generalized Linear Models", 2^{nd} Edition, Chapman & Hall.

[30] The Website of R Project, http://www.r-project.org/

[31] I. Cheng and P. Boulanger, "A 3D Perceptual Metric using Just-Noticeable-Difference", *Eurographics*, pp. 97-100, August 2005.

[32] L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and Regression Trees.", Wadsworth, Pacific Grove, CA, 1984.

[33] K. Stuhlmuller, N. Farber, M. Link and B. Girod, "Analysis of video transmission over lossy channels", *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 6,pp. 1012-32, June 2000.

[34] Y. J. Liang, J. G. Apostolopoulos and B. Girod, "Analysis of packet loss for compressed video: does burst-length matter?", *IEEE ICASSP*, vol. 5, pp. 684-687, April 2003.

[35] J. Chakareski, J. G. Apostolopoulos, W. T. Tan, S. Wee and B. Girod, "Distortion chains for predicting the video distortion for general packet loss patterns", *IEEE ICASSP*, vol. 5, pp. 1001-1004, May 2004.

[36] I. E. G. Richardson, "H.264 and MPEG-4 Video Compression", John Wiley & Sons, September 2003.

[37] Y. F. Ma, X. S. Hua, L. Lu and H. J. Zhang, "A generic framework of user attention model and its application in video summarization", *Multimedia, IEEE Transactions on*, vol. 7, no. 5, pp. 907-919, Oct 2005.

[38] V. Varsa, M. M. Hannuksela and Y. Wang, "Non-Normative Error Concealment Algorithms", *ITU-T VCEG-N62*, 2001.

[39] D. W. Hosmer and S. Lemeshow, "Applied Logistic Regression", 2^{nd} Edition, Wiley-Interscience.

[40] P. Pancha and M. El Zarki, "Prioritized transmission of variable bit rate MPEG video", *IEEE GLOBECOM*, vol. 2, pp. 1135-1139, Dec 1992.

[41] H. Liu and M. El Zarki, "Transmission of video telephony images over wireless channels", *Springer Journal on Wireless Networks*", vol. 2, no. 3, Sep 1996.

[42] C. Leicher, "Hierarchical encoding of MPEG sequences using priority encoding transmission (PET)", *TR-94-058, ICSI, Berkeley, CA*, Nov 1994.

[43] F. Hartanto and H. R. Sirisena, "Hybrid error control mechanism for video transmission in the wireless IP networks", *IEEE Workshop on Local and Metropolitan Area Networks*, Nov 1999.

[44] M. van der Schaar and H. Radha, "Unequal packet loss resilience for fine-granular-scalability video", *IEEE Trans. Multimedia*, vol. 3, no. 4, Dec 2001.

[45] J. Vehkapera and J. Peltola, "Optimized decoding scheme for erroneous MPEG-4 FGS bitstream", *IEEE ISCAS*, vol. 4, pp. 3423-3426, May 2005.

[46] X. Yang, C. Zhu, Z. G. Li, X. Lin and N. Ling, "An unequal packet loss resilience scheme for video over the internet", *IEEE Trans. Multimedia*, vol. 7, no. 4, Aug 2005.

[47] J. Shin, J. W. Kim and C. C. J. Kuo, "Relative priority based QoS interaction between video applications and differentiated service networks", *IEEE ICIP*, vol. 3, pp. 536-539, Sep 2000.

[48] L. L. Peterson and B. S. Davie, "Computer networks : a systems approach", 3^{rd} Edition, Morgan Kaufmann Publishers.

[49] The website of NS Project, http://www.isi.edu/nsnam/ns/

[50] The website for VQM software,
http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

[51] M. H. Loke, E. P. Ong, W. Lin, Z. Lu and S. Yao, "Comparison of video quality metrics on multimedia videos", *IEEE ICIP*, October 2006.