Adaptive Format Conversion Information as Enhancement Data for the High-definition Television Migration Path

by

James R. Thornbrue

B.S. Electrical Engineering
Brigham Young University, 2001

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Electrical Engineering

at the

Massachusetts Institute of Technology

June 2003

Signature of Author........ ....
Department of Electrical Engineering and Computer Science
April 14, 2003

Certified by.............. ....
Jae S. Lim
Professor of Electrical Engineering
Thesis Supervisor

Accepted by.. ....
Arthur C. Smith
Chairman, Committee for Graduate Students

Adaptive Format Conversion Information as Enhancement Data for the High-definition Television Migration Path

by

James R. Thornbrue

Submitted to the Department of Electrical Engineering and Computer Science on April 23, 2003 in partial fulfillment of the requirements for the Degree of Master of Science in Electrical Engineering

ABSTRACT

Prior research indicates that a scalable video codec based on adaptive format conversion (AFC) information may be ideally suited to meet the demands of the migration path for high-definition television. Most scalable coding schemes use a single format conversion technique and encode residual information in the enhancement layer. Adaptive format conversion is different in that it employs more than one conversion technique. AFC partitions a video sequence into small blocks and selects the format conversion filter with the best performance in each block. Research shows that the bandwidth required for this type of enhancement information is small, yet the improvement in video quality is significant.

This thesis focuses on the migration from 1080I to 1080P using adaptive deinterlacing. Two main questions are answered. First, how does adaptive format conversion perform when the base layer video is compressed in a manner typical to high-definition television? It was found that when the interlaced base layer was compressed to 0.3 bpp, the mean base layer PSNR was 32 dB and the PSNR improvement due to the enhancement layer was as high as 4 dB. Second, what is the optimal tradeoff between base layer and enhancement layer bandwidth? With the total bandwidth fixed at 0.32 bpp, it was found that the optimal bandwidth allocation was about 96% base layer, 4% enhancement layer using fixed, 16x16 pixel partitions. The base and enhancement layer video at this point were compared to 100% base layer allocation and the best nonadaptive format conversion. While there was usually no visible difference in base layer quality, the adaptively deinterlaced enhancement layer was generally sharper, with cleaner edges, less flickering, and fewer aliasing artifacts than the best nonadaptive method. Although further research is needed, the results of these experiments support the idea of using adaptive deinterlacing in the HDTV migration path.

Thesis Supervisor: Jae S. Lim
Title: Professor of Electrical Engineering

Dedication

To my three beautiful girls

Acknowledgements

I extend gratitude toward Professor Jae Lim for providing the opportunity to work in his group, for guidance, and for financial support. As a student in *Two Dimensional Signal and Image Processing*, I was impressed by his teaching. I continue to be motivated by his example.

I also recognize Wade Wan, who helped me get started. His experience and contribution were invaluable. I would like to thank the other members of the Advanced Television Research Program: Brian Heng, Ken Schutte, and especially Cindy LeBlanc. In the largeness that is MIT, their friendship was an oasis.

Thanks to my beautiful wife, Allie, for her patience and support. To my daughter, Marie, and another on the way—you are my inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The HDTV Migration Path

1.1.1 History of Terrestrial Television Standards

The NTSC (National Television Systems Committee) standard for terrestrial television broadcasting in the United States was established in 1941, with color added in 1953. NTSC is the analog television standard in North America and Japan. Other conventional television systems used throughout the world include PAL (phase-alternating line) and SECAM (Sequential Couleur a Memoire). These three systems all have similar video, audio, and transmission quality. For example, NTSC delivers approximately 480 lines of video, where each line contains 420 picture elements (pixels or pels). The spatial resolution is described by the number of lines in the vertical dimension and the number of pixels per line in the horizontal dimension. When snapshots of a scene are refreshed at a sufficiently high rate, the human visual system perceives continuous motion. In a tradeoff between spatial and temporal resolution, NTSC uses interlaced scanning at approximately 60 fields/second. Interlaced scanning (IS) means that every alternate snapshot contains only the even or the odd lines. In interlaced scanning, these snapshots are called *fields*.

In 1987, the United States Federal Communication Commission (FCC) established an Advisory Committee on Advanced Television Service (ACATS) consisting of 25 leaders from the television industry with the purpose of recommending an advanced television system to replace the NTSC standard. Initially, ACATS received 23 proposals, ranging from improved forms of NTSC to completely new high-definition television (HDTV) systems. By 1991, the number of competing proposals had been reduced to six, including four all-digital HDTV systems. After years of extensive testing and deliberation, the advisory comittee determined that analog technology would no longer be considered. It would not, however, recommend one of the four remaining systems above another because each had different strengths. As a result, ACATS recommended that the individual companies developing these systems be allowed to implement certain improvements that they had proposed. In addition, the advisory committee expressed enthusiasm for a single, joint proposal made of the best elements from each system.

In response to this incitement, the companies representing the four all-digital systems formed the Digital HDTV Grand Alliance in May, 1993. The members of the Grand Alliance were AT&T, General Instrument, North America Phillips, Massachusetts Institute of Technology, Thomson Consumer Electronics, the David Sarnoff Research Center, and Zenith Electronics Corporation. Another important organization at this time was the Advanced Television Systems Committee (ATSC), a private sector group representing all segments of the television industry. The ATSC took responsibility for documenting the specifications for the HDTV standard based on the Grand Alliance system. On December 24, 1996, the FCC adopted the major elements of the ATSC digital television standard [1, 2].

The ATSC standard has many improvements over its analog predecessor. The primary goal of HDTV is to provide increased spatial resolution (or "definition") compared to conventional analog television systems. One example of a high-definition video format has a spatial resolution of 1080 lines and 1920 pixels per line with interlaced scanning at 60 fields/sec—over ten times the resolution of NTSC. The ATSC standard also supports progressive scanning (PS). Unlike interlaced scanning, where a field contains only the even or odd lines, progressive scanning retains all the lines in each snapshot. A progressively scanned snapshot is called a *frame*. Another improvement in HDTV is the aspect ratio (width-to-height) of the display, where a larger aspect ratio in conjunction with high definition leads to an increased viewing angle and more realistic viewing experience. HDTV has an aspect ratio of 16:9 with square pixels, compared to 4:3 for NTSC. The HDTV standard includes CD quality surround sound as well as the ability to transmit data channels and interact with computers. It also supports different high-definition video formats. For instance, a program originally captured on film can be transmitted in its native frame rate of 24 frames/sec. Other programs such as sporting events may choose to trade off spatial resolution for increased temporal resolution, yielding smoother perceived motion. In addition, multiple programs can be sent on the same HDTV channel. The NTSC standard requires all programs to be converted to the same format before transmission and is limited to one program per channel. Because of digital transmission technology, HDTV receivers are able to reconstruct a "perfect" picture without multipath effects, noise, or interference common to analog television. [3, 4]

The addition of color to the NTSC standard in 1953 was done in a backward-compatible manner so that black-and-white television sets were not made obsolete by the broadcast of color programs. This was done by adding a small amount of color information where it would not significantly interfere with the black-and-white (luminance) part of the signal. However, when high-definition television was being developed, it was decided that a non-compatible standard was necessary to achieve the desired resolution and quality. This means that current analog television receivers will *not* be able to decode an HDTV signal. Transmission of NTSC television is scheduled to be phased out in 2006. By this time, all broadcast television programs will be transmitted in the HDTV format, and consumers will be required to purchase an HDTV receiver in order to watch broadcast television.

The transition to HDTV faces a formidable economic hurdle. At the time of writing, there are an estimated 105 million households in the United States with an average of 2.4 NTSC television sets per household. The average price of an HDTV set is $1,500, and a set-top HDTV converter is roughly $500. Needless to say, the high cost of HDTV technology is discouraging to the individual consumer. Broadcasting equipment must also be replaced. This equipment is estimated at several million dollars per station, and there are 1,650 high-power television stations in the United States. The total economic impact of this transition is on the order of $200 billion.

In order to avoid transition problems like this in the future, the HDTV standard was designed to allow additional features in a backward-compatible manner. In a high-definition television set, data that is not understood is simply ignored. In this way, information may be added to the signal without interfering with functions that are presently defined. It is this flexibility that will allow the migration path to higher resolution video formats that is considered in this thesis.

1.1.2 The Migration Path to Higher Resolutions

Despite the significant improvements in the high-definition standard, a primary target of HDTV has still not been met—the ability to transmit 1080x1920 progressively scanned video at 60 frames/sec (1080P) in a single 6-MHz channel. 1080P requires a sample rate of approximately 125 Mpixels/sec, which exceeds the maximum rate of 62.7 Mpixels/sec allowed by the MPEG-2 video compression portion of the HDTV standard. MPEG-2 video compression will be discussed in detail in chapter 2.

In addition to exceeding the sample rate limit, 1080P cannot be compressed into a single HDTV channel without significant loss of picture quality for difficult scenes. The transmission technology used for HDTV transmission allows a bandwidth of approximately 19 Mbits/sec for the video portion of the signal. This means that 1080P must be encoded at 0.16 bits per pixel (bpp). Raw video is 24 bpp, meaning that the required compression is approximately 150 to one. In comparison, MPEG-2 can achieve a compression factor of about 70–80 while maintaining an adequate level of picture quality for most programs. Even if the sample rate required for 1080P were allowed, the high level of compression would limit the viability of this format.

A list of high-definition video formats is provided in Table 1.1. The first six formats are supported in the digital television standard and are commonly used by the television industry. For each video format, the table includes the spatial resolution, frame (or field) rate, scanning method, and pixel rate. Because the transmission bandwidth is limited to 20 Mbps, the pixel rates shown in the table illustrate the need for high amounts of compression. The last video format shown in the table, 1080P, exceeds the sample rate limit for MPEG-2. For this reason, 1080P is not an allowable format.

Format Name	Spatial Resolution	Frame/Field Rate	Scanning	Pixel Rate
720P	720 x 1280	60 frames/sec	PS	55.3 Mpixels/sec
720P@30fps	720 x 1280	30 frames/sec	PS	27.6 Mpixels/sec
720P@24fps	720 x 1280	24 frames/sec	PS	22.1 Mpixels/sec
1080I	1080 x 1920	60 fields/sec	IS	62.2 Mpixels/sec
1080P@30fps	1080 x 1920	30 frames/sec	PS	62.2 Mpixels/sec
1080P@24fps	1080 x 1920	24 frames/sec	PS	49.8 Mpixels/sec
1080P	1080x1920	60 frames/sec	PS	124.4 Mpixels/sec

Table 1.1: High-definition Video Formats

The first six high-definition formats shown in the table are permitted in the U.S. HDTV standard and are commonly used in the television industry. Spatial resolution of VxH means V lines of vertical resolution and H pixels of horizontal resolution. The frame/field rate is the number of frames per second for progressive scanning and the number of fields per second for interlaced scanning. Scanning is either interlaced or progressive. The pixel rate, in pixels per second, demonstrates the need for video compression because the bandwidth for video in an HDTV channel is approximately 19 Mbits per second. The last format, 1080P, exceeds the sample rate constraint of MPEG-2 and usually cannot be compressed to 19 Mbps with a sufficient level of picture quality.

The need for resolutions even higher than 1080P, such as 1440x2560 PS at 60 fps or 1080x1920 PS at 72 fps, has already been predicted. However, in order for any higher-resolution formats to be broadcast, two things must happen: first, the HDTV standard must be evolved to accept higher sample rates; second, more bandwidth must become available. The sample rate problem can be accommodated by using a scalable coding scheme that is backward compatible with current HDTV transmissions—such a solution is presented in this thesis. Additional bandwidth may become available in several ways. Once analog television is phased out, the FCC may allocate more bandwidth for HDTV channels; alternatively, improvements may be made in video compression techniques that free up existing bandwidth. In either case, the increased bandwidth is expected to be small in the near future. How to add support for 1080P and other higher-resolution video formats while dealing with these two issues—backward compatibility and limited bandwidth—is what is known as the *migration path problem*, or simply the *migration path*, for high-definition television.

1.2 Scalable Video Coding

This thesis presents a migration path based on scalable video coding. With the proliferation of video content on the internet, the problem of scalable video coding has received considerable attention. Because of different connection speeds, video on the internet is often available with several levels of quality or resolution, and scalable techniques are used to minimize the total amount of information that must be stored or transmitted. Instead of encoding each different format independently, a scalable scheme encodes a single, independent *base layer*, and one or more dependent *enhancement layers*, where each enhancement layer is used to increase the resolution or quality of the previous layer.

For each enhancement, the higher resolution is first provided by a conversion from the base layer format to the enhancement layer format (note that for quality scalability, no format conversion is necessary). Then, *residual* information is added to improve the picture quality. The residual, or error, is defined as the difference between the interpolated base layer video and the original video sequence. Residual coding is well understood and is used in popular scalable coding schemes such as the MPEG-2 and MPEG-4 multimedia standards. Another type of enhancement is *adaptive format conversion* (AFC). Because the encoder has access to the original video sequence, the conversion to the enhancement layer format can be done adaptively. This is accomplished by breaking the video into small blocks and deciding which of several predefined interpolation methods best reconstructs the original sequence on a block-by-block basis. Adaptive format conversion is a fairly new area of research. The next two sections discuss residual and AFC coding in more detail.

1.2.1 Residual Coding

As defined above, the residual is the difference between the original video sequence and an interpolated version of the decoded base layer video. Figure 1.1 shows a block diagram of a spatially scalable video codec (encoder/decoder) using residual information in an enhancement layer. In this example, the base video format is 720P, and the enhanced video format is 1080P. The encoder downsamples the 1080P video sequence to 720x1280 pixels, and the lower-resolution video is encoded as the base layer bitstream. The encoding process is generally lossy, meaning that the reconstructed video will not be the same as the original video sequence. After the base layer is reconstructed, it is upsampled to 1080P, and the residual is encoded as an enhancement layer. At the decoder, the base layer bitstream is decoded to produce 720P video. The decoder interpolates this video to 1080x1920 pixels, then adds the residual information to create the enhanced-resolution video, which is 1080P. Note that both the encoder and decoder must decode the base layer bitstream and perform the format conversion in exactly the same way.

A scalable coding scheme such as this, applied to the migration path, is backward compatible. The base layer video format is allowed in the current HDTV standard and is independent of the enhancement layer. Given a large enough enhancement layer bandwidth, residual coding has the ability to generate enhanced video of arbitrarily good quality. Unfortunately, even for low-quality enhancements, the bitrate required for residual data is typically higher than that foreseen in the near future for high-definition television.

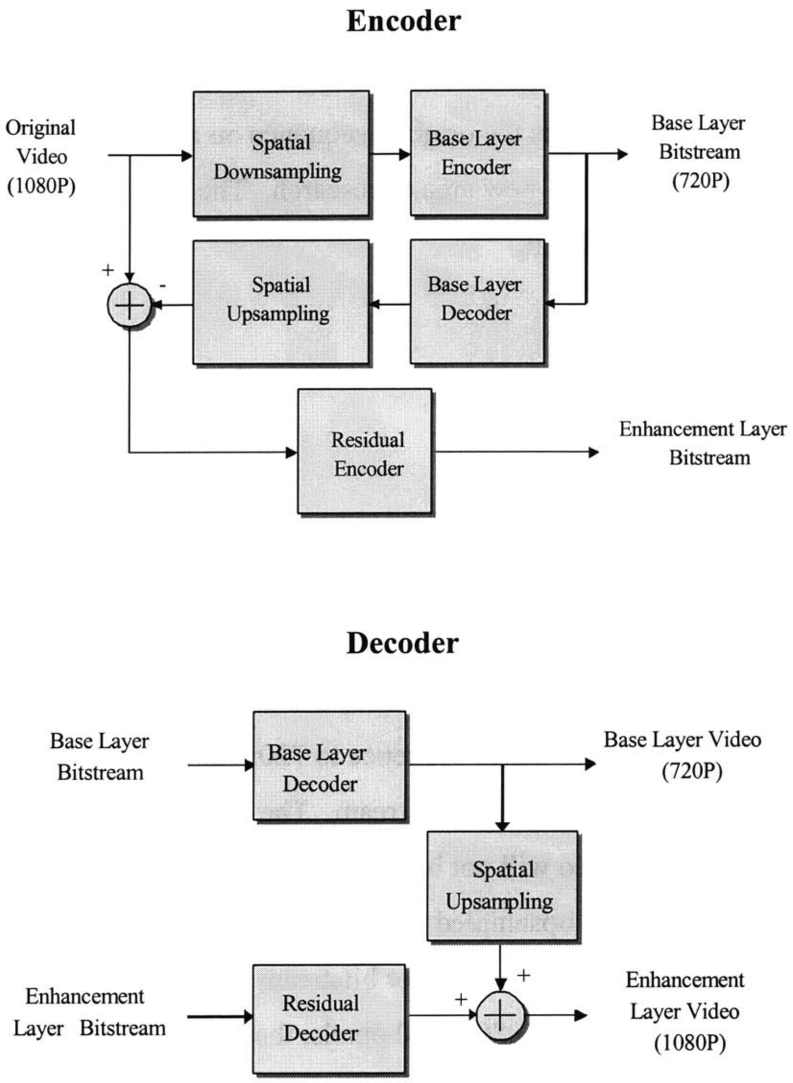


Figure 1.1: A Spatially Scalable Video Codec Based on Residual Coding

The original 1080P video is downsampled to 720x1280 pixels and encoded as the base layer. After it is reconstructed, the base layer video is upsampled to the original format, and the residual, or error, is encoded as the enhancement layer. The decoder uses the base layer bitstream to reconstruct the 720P video. The 1080P enhancement layer video is created by spatial upsampling and the addition of the residual, which improves the picture quality. Note that this scalable codec is backward compatible with the current HDTV standard.

1.2.2 Adaptive Format Conversion

Adaptive format conversion is an alternative, or addition, to residual coding proposed by Sunshine [5, 6] and Wan [7, 8, 9]. In the residual coding example of the previous section, a single format conversion method was used to create the enhanced-resolution video. However, the format conversion can be made even better if more than one technique is used adaptively. Sunshine and Wan look specifically at adaptive deinterlacing by selecting four deinterlacing methods: linear interpolation, Martinez-Lim deinterlacing [10], forward field repetition, and backward field repetition. The encoder partitions the video sequence into nonoverlapping blocks and selects the best deinterlacing method for each block. Information about how the blocks are partitioned and which deinterlacing method is used in each block becomes enhancement data that is sent to the decoder. In addition to the AFC enhancement data, the encoder may also send residual information in a second enhancement layer.

An example video codec with both AFC and residual enhancement is shown in Figure 1.2. The base layer format is 1080I, and the enhancement layer format is 1080P. The original 1080P video sequence is interlaced by discarding the even or odd lines in alternate frames, encoded, then decoded again. The resulting video is partitioned into blocks and the best deinterlacing method is chosen for each block—the partitioning and format conversion information becomes the first enhancement layer. The resulting video quality will be better than if a single deinterlacing technique were used for the entire sequence. Next, the residual is encoded as the second enhancement layer. At the decoder, the base layer bitstream is decoded to reconstruct the interlaced video sequence. Partitioning and deinterlacing information in the AFC enhancement layer is used to create progressive video, and the residual enhancement layer is used to further improve the quality of the 1080P video sequence.

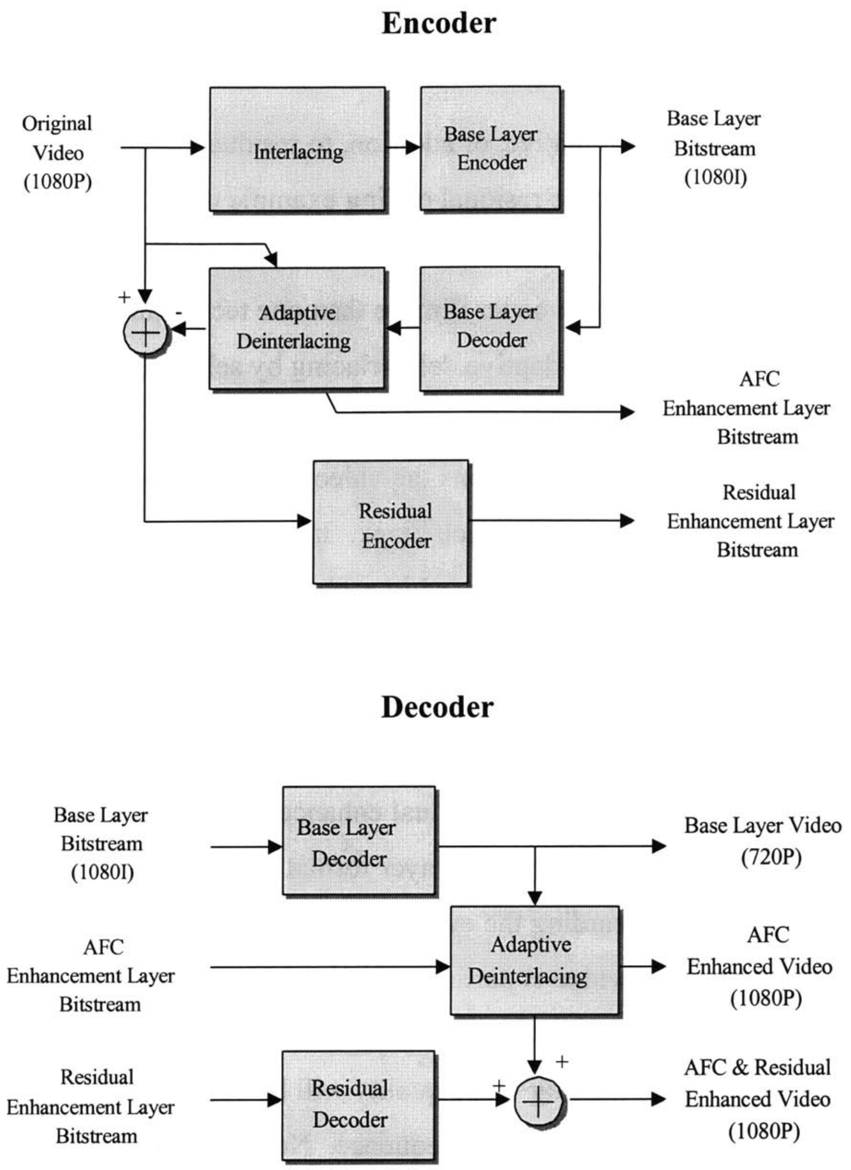

Encoder
Original Video (1080P)
Interlacing
Base Layer Encoder
Base Layer Bitstream (1080I)
+
-
Adaptive Deinterlacing
Base Layer Decoder
AFC Enhancement Layer Bitstream
Residual Encoder
Residual Enhancement Layer Bitstream
Decoder
Base Layer Bitstream (1080I)
Base Layer Decoder
Base Layer Video (720P)
AFC Enhancement Layer Bitstream
Adaptive Deinterlacing
AFC Enhanced Video (1080P)
Residual Enhancement Layer Bitstream
Residual Decoder
+
+
AFC & Residual Enhanced Video (1080P)

Figure 1.2: A Scalable Codec Based on Adaptive Deinterlacing and Residual Coding

The original 1080P video sequence is interlaced by discarding the even or odd lines in each frame, and the resulting video is encoded as the 1080I base layer. The base layer is reconstructed and compared to the original video sequence in order to find the best deinterlacing method for each partition. Partitioning and deinterlacing information is sent to the decoder as the first enhancement layer. The residual is encoded as the second enhancement layer. The decoder uses the base layer bitstream to create 1080I video. The partitioning and deinterlacing information in the first enhancement layer is used to convert the format to 1080P. Finally, the residual information in the second enhancement layer is added to improve the quality of the 1080P video sequence.

Wan shows that adaptive format conversion provides substantial improvement over the best nonadaptive method and requires lower bitrates than residual coding. However, unlike residual coding, the ability of adaptive format conversion to exactly reproduce the original video sequence is bound by the performance of the individual format conversion methods. In other words, using adaptive format conversion information alone cannot achieve an arbitrarily high quality reproduction of the original video sequence, no matter how much bandwidth is allocated to the enhancement layer bitstream.

Despite the above limitation, but because of the low bitrate requirement, adaptive format conversion is a natural choice in the migration to higher-resolution HDTV. For instance, the target format of 1080P can be reached in several ways by adding only a low bitrate AFC enhancement layer to one of the common HDTV formats in table 1.1. The example described above shows how a 1080I base layer and adaptive deinterlacing information is used to produce 1080P video. This type of enhancement is illustrated in figure 1.3, which shows how the information in the base layer fields is used create the progressive enhancement layer frames. The use of adaptive deinterlacing in the migration to 1080P is the focus of this thesis.

Another common base layer format is 720P. In the migration to 1080P, a codec that uses adaptive format conversion would select the best spatial interpolation technique for each section of video. There are many different spatial interpolation filters that could be used, including nearest neighbor, bilinear interpolation, bicubic interpolation, and two dimensional transform techniques. An illustration of spatial scalability for the HDTV migration path is provided in figure 1.4.

Finally, the base layer format could be 1080P@30 fps, in which case the AFC enhancement layer would contain information about temporal upsampling techniques. These techniques could include forward frame repetition, backward frame repetition, linear interpolation, or motion compensated linear interpolation. Figure 1.5 shows how the information in a 1080P@30fps base layer is used to create a 1080P enhancement layer in this example of adaptive temporal upsampling.

The three base layer formats discussed above (1080I, 720P, and 1080P@30fps) are permitted in the current HDTV standard; therefore, HDTV receivers that do not support the enhancement layer would still be able to decode the base layer video in a backward-compatible manner. Next generation HDTV receivers, on the other hand, would exploit the information in the AFC enhancement layer bitstream to increase the video resolution to 1080P.

Using adaptive format conversion information in the migration to 1080P would be the initial step in the migration path. In the near future, when the available bandwidth is small, adaptive format conversion would provide video scalability at low enhancement layer bitrates. As more bandwidth becomes available, the migration path could be extended by adding a residual enhancement layer to improve the video quality or by further increasing the resolution to formats like 1440P or 1080P@72fps. Previous research has shown that when the base layer is encoded well, the use of adaptive format conversion in conjunction with residual coding is more efficient than residual encoding alone.

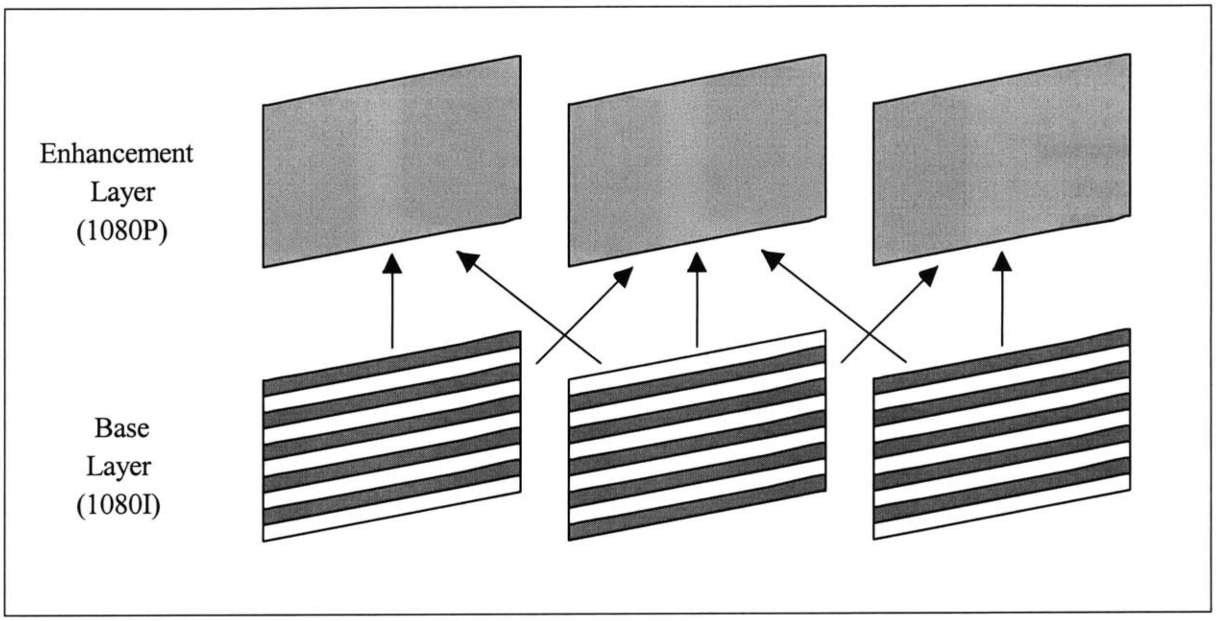


Figure 1.3: Adaptive Deinterlacing for the HDTV Migration Path

This figure illustrates how a 1080I base layer is used to create a 1080P enhancement layer. The arrows indicate the base layer fields that are used to create the progressive enhancement layer frames.

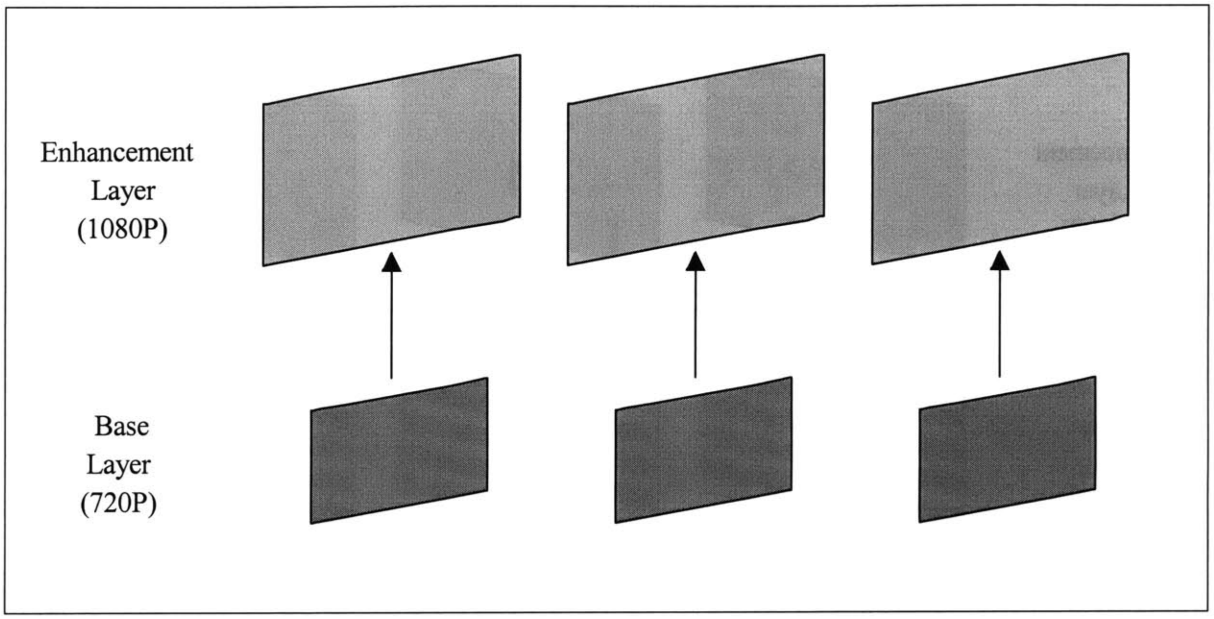


Figure 1.4: Adaptive Spatial Upsampling for the HDTV Migration Path

This figure illustrates how a 720P base layer is used to create a 1080P enhancement layer. The resolution of each frame is increased adaptively using a number of spatial interpolation techniques.

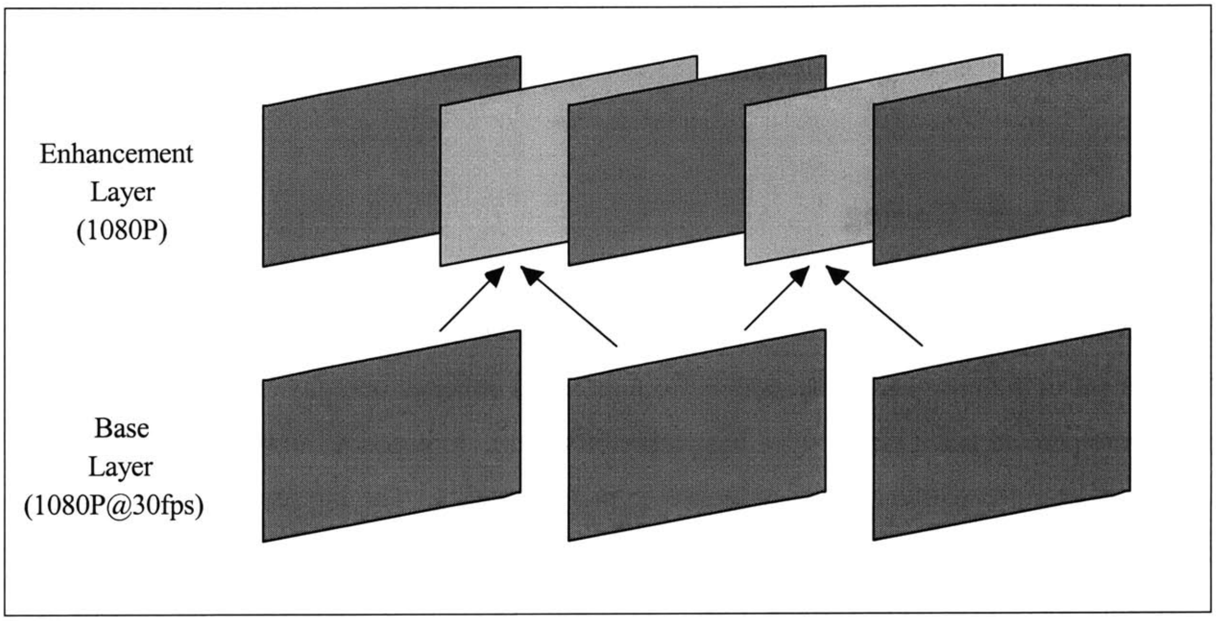


Figure 1.5: Adaptive Temporal Upsampling for the HDTV Migration Path

This figure illustrates how a 1080P@30fps base layer is used to create a 1080P enhancement layer. The arrow indicate the base layer frames that are used to extrapolate the missing enhancement layer frames.

1.3 Motivation for Thesis

Current research into adaptive format conversion has encouraging implications in the migration to higher resolution HDTV. One such proof of concept is that an AFC enhancement layer requires a very small bandwidth. However, there are still many issues specific to the migration path that have not been addressed. These include the coding of the base layer video and the optimal allocation of base and enhancement layer bandwidth in a fixed bandwidth environment. These two issues are introduced in the following sections as the fundamental motivation for this thesis.

1.3.1 Base Layer Coding

When Sunshine introduced the idea of an AFC enhancement layer, he measured the improvement in picture quality that comes from different enhancement layer bandwidths. These results were created using an uncoded base layer bitstream. In practice, however, an HDTV bitstream is encoded with a marked reduction in picture quality. The improvement in video quality derived from an uncoded base layer can be thought of as an empirical upper bound on the performance of AFC as enhancement data.

Wan extended the analysis to include encoded base layers of various qualities. Like Sunshine, he measured the improvement in video quality as a function of enhancement layer bandwidth—not just for an uncoded base layer but for many levels of base layer degredation. Not surprisingly, the quality of the reconstructed video using AFC enhancement data was an increasing function of base layer quality. Furthermore, when the base layer was coded poorly, residual coding outperformed AFC for higher enhancement layer bitrates. Wan concluded that the decision to use AFC, residual coding, or a combination of the two is a function of both base layer quality and enhancement layer bandwidth and is not always obvious. He also concluded that because it requires such a small bandwidth, the use of adaptive format conversion seems ideally suited for the HDTV migration path.

In the work cited above, no attempt was made to determine what level of base layer quality is typical to the compression of HDTV video sequences. In practice, different video encoders have varying levels of performance (higher performance means an encoder generates better quality video for the same amount of compression). Wan avoided the issue of encoder implementation by using a nonstandard encoding strategy and by using base layer quality (rather than bandwidth) as the independent variable. A side effect of this approach is that the results do not show how AFC enhancement performs in the specific context of the HDTV migration path. In order to fit into a 19 Mbps bandwidth, for example, 1080I is encoded at approximately 0.3 bits per pixel—a compression factor of about 80 to 1. Depending on the encoder implementation, the resulting video quality may vary considerably. It is uncertain from previous results what kind of quality can be expected from this level of compression and whether adaptive format conversion will improve the video quality in any significant way. This leads to the first motivation for this thesis:

Motivation #1: *How does adaptive format conversion perform when the base layer video is compressed in a manner typical to high-definition television?*

In order to answer this question, the implementation presented in this thesis uses a standard HDTV encoding strategy. With this encoder, the video quality is determined by the base layer bandwidth, and it is possible to determine where HDTV falls in the spectrum of previous results.

1.3.2 Tradeoff Between Base Layer and Enhancement Layer Bandwidth

In a fixed bandwidth environment such as HDTV, it is important to discuss the tradeoff between the base layer and enhancement layer. For example, if the total bandwidth (base plus enhancement) is fixed at 0.32 bits per pixel, how much should be allocated to the base layer and how much to the enhancement layer? The answer to this question depends on two factors:

1. How much is the base layer degraded?
2. How much is the enhancement layer improved?

When part of the total bandwidth is allocated to the enhancement layer, the base layer video will be degraded to some degree. The integrity of the base layer is important because not all receivers will be equipped to use the enhancement layer information. If the base layer degradation is severe, then no amount of improvement in the enhancement layer is worth the cost. On the other hand, a small enhancement layer bandwidth may be viable if it does not significantly affect the base video quality but provides substantial improvement to the enhanced resolution video. This question is the second motivation for this thesis:

Motivation #2: *What is the optimal tradeoff between base layer and enhancement layer bandwidth?*

1.4 Summary

Previous research suggests that a scalable video codec based on adaptive format conversion information may be an ideal solution to the migration path for high-definition television. The first section of this chapter introduced the migration path problem. It began with a brief history of terrestrial television standards in the United States, highlighting the advantages of the high-definition standard over the current analog system. Even with these improvements, however, there are still limitations on the transmittable video resolution, and the need to transmit higher resolution formats in the future has already been recognized. In particular, 1080P is a desirable format that is not permitted in the current U.S. HDTV standard. The question of how to add support for 1080P and other, higher-resolution formats in a way that is backward compatible with the current HDTV standard is known as the migration path problem.

Section 1.2 introduced the concept of scalable video coding as a solution to the HDTV migration path. Residual and adaptive format conversion information are two types of enhancement information that can be added on top of a compatible base layer. It was shown that 1080I, 720P, and 1080P@30fps are all compatible base layer formats that can be used in the migration to 1080P. The use of adaptive deinterlacing information in the migration from 1080I to 1080P is the focus of this thesis.

Adaptive format conversion has been studied before in the general context of multicast video coding. The results prove the concept of using an AFC enhancement layer in many scalable coding scenarios, but the implementation ignores two issues specific to the HDTV migration path: the coding of the base layer video and the optimal allocation of base and enhancement layer bandwidth. These limitations are restated below as the fundamental motivation for this thesis:

Motivation #1: *How does adaptive format conversion perform when the base layer video is compressed in a manner typical to high-definition television?*

Motivation #2: *What is the optimal tradeoff between base layer and enhancement layer bandwidth?*

1.5 Thesis Overview

The next chapter gives an introduction to MPEG-2 and video coding terminology that is used throughout the thesis. MPEG-2 is the video compression algorithm used in the U.S. HDTV standard. The main difference between this and previous work is the base layer codec. In this thesis, motion compensation and rate control are used to encode the base layer video in a way that is typical to HDTV.

Chapter three provides a detailed description of the adaptive format conversion enhancement layer, including the four deinterlacing techniques, fixed and adaptive frame partitioning schemes, optimal parameter selection, and parameter coding. The implementation of adaptive deinterlacing that is used in this thesis has been studied previously and is reviewed here for convenience.

The results of two main experiments are found in chapter four. In the first experiment, the base layer video is encoded like HDTV, and the quality of the enhancement layer video is measured as a function of enhancement layer bandwidth. It is determined how these results compare to previous work. In the second experiment, the total bandwidth is fixed and divided between the base and enhancement layer in order to discover the optimal tradeoff. The experiment is performed for ten video sequences with different characteristics.

The final chapter draws conclusions and indicates directions for future research.

Chapter 2
Base Layer Video Coding

MPEG-2 is the video compression algorithm used in the U.S. HDTV standard. This chapter provides an introduction to MPEG, its structure, and the techniques that are used to achieve high levels of compression with relatively little loss in picture quality. MPEG gives an encoder a great deal of flexibility, and an encoding strategy for constant bitrate applications such as HDTV is explained. With this background in mind, the first motivating question for this thesis is revisited: *how does adaptive format conversion perform when the base layer is compressed in a manner typical to HDTV?* It is noted that the previous implementation used a base layer coding strategy that cannot answer this question. The final section of this chapter introduces the MPEG-2 encoder that is used in the current work, along with its performance characteristics.

2.1 MPEG-2 Video Compression

2.1.1 Introduction to MPEG

The Moving Picture Experts Group (MPEG) was formed in 1988 in response to a growing need for audio and video compression standards across many diverse industries. MPEG is formally called ISO/IEC JTC1/SC29/WG11 within the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC). Their efforts soon produced the highly popular MPEG-1 multimedia standard. MPEG-1 is intended for bitrates around 1.5 Mbps, such as CD-ROMs and digital video recorders. Even before MPEG-1 was complete, work began on MPEG-2, which is designed for larger picture sizes and higher bandwidth applications such as HDTV and DVD (Digital Versatile Disk). MPEG-2 also includes support for interlaced scanning and video scalability that was not part of MPEG-1. Another standard, MPEG-4, is currently being developed. Originally intended for very low bitrate applications such as video telephones, it has also come to include support for irregular shaped video objects, content-based manipulation, and editing. Widely accepted video compression standards such as MPEG-1 and MPEG-2 have reduced the cost and risk involved with deploying new technology.

MPEG-2 contains a large set of video compression tools, including different color sampling formats, frame prediction types, and scalability options. Because not all applications require the complete set of tools, MPEG-2 is divided into profiles that contain various subsets of the video compression techniques. The five MPEG-2 profiles are simple, main, SNR scalable, spatially scalable, and high. MPEG-2 is also divided into levels that constrain the maximum spatial resolution, frame rate, sample rate, and bitrate for each profile. The four levels are low, main, high-1440, and high. Table 2.1 describes the level definitions for the main profile of MPEG-2. The video compression portion of the ACATS standard for high-definition television uses the MPEG-2 main profile at high level (MP@HL). Allowable video formats for HDTV are limited by the maximum values shown in the table.

The remainder of this chapter introduces the MPEG-2 main profile. Some excellent references are provided for a more general discussion of video compression [11], an overview of MPEG [12, 13, 14], and the MPEG standard documents [15, 16, 17].

Level	Parameter	Maximum Value
High (MP@HL)	spatial resolution frame rate sample rate bitrate	1152x1920 pixels 60 fps 62,668,800 samples/sec 80 Mbps
High-1440 (MP@H-14)	spatial resolution frame rate sample rate bitrate	1152x1440 pixels 60 fps 47,001,600 samples/sec 60 Mbps
Main (MP@ML)	spatial resolution frame rate sample rate bitrate	576x720 pixels 30 fps 10,368,000 samples/sec 15 Mbps
Low (MP@LL)	spatial resolution frame rate sample rate bitrate	288x352 pixels 30 fps 3,041,280 samples/sec 4 Mbps

Table 2.1: Level Definitions for the MPEG-2 Main Profile

The video portion of the ACATS high-definition television standard uses the MPEG-2 main profile at high level (MP@HL). Permissible video formats for HDTV are limited by the MP@HL spatial resolution, frame rate, sample rate, and bitrate constraints. The video coding techniques used in the MPEG-2 main profile are introduced in this chapter.

2.1.2 Motivation for Video Compression

Video is a sequence of rectangular frames of the same size, displayed at a fixed rate. Each frame is represented by an array of pixels where the number of pixels determines the resolution (or definition) of the frame. The individual pixels are comprised of three color components—for raw video, each color is represented digitally by 8 bits. For example, when the 1080P format is uncompressed, it requires a bitrate of

$$1080\times 1920\,\frac{\text{pixels}}{\text{frame}}\cdot 60\,\frac{\text{frames}}{\text{sec}}\cdot 8\,\frac{\text{bits}}{\text{color}}\cdot 3\,\frac{\text{colors}}{\text{pixel}}\cong 2{,}848\,\text{Mbps}\ . \tag{2.1}$$

In comparison, the bandwidth available for video transmission in an HDTV channel is only about 19 Mbps—it is impossible to transmit raw, high-definition video with this technology.

Fortunately, a typical video sequences has a large amount of redundant information. Within a single frame, pixels tend to be similar to those that surround them; between frames, pictures are similar to the frames that precede and follow them. Video compression techniques take advantage of these redundancies to reduce the amount of information necessary to describe a complete video sequence. Further reduction is possible because of characteristics of human perception. The human visual system is less sensitive to detail in moving objects. It is also less sensitive to detail in certain color characteristics. As much as possible, the redundant and irrelevant information in a video sequence are eliminated in order to compress raw video to more manageable bitrates.

2.1.3 Lossy Compression and PSNR

There are two types of compression: *lossless* and *lossy.* Lossless compression results in perfect picture reconstruction, yet there is a limit on the amount of data reduction that is possible using lossless techniques. Lossy compression, on the other hand, is generally able to achieve much higher levels of compression. The tradeoff is that the reconstructed picture is not exactly the same as the original, but is degraded to some degree. MPEG is an example of a lossy compression algorithm where the lost information is often undetectable or does not significantly detract from the picture quality.

With lossy compression, or whenever there is picture degradation, it is useful to have a quantitative measure of video quality. One such measure is the *peak signal-to-noise ratio* (PSNR), defined here:

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\mathrm{MSE}}, \tag{2.2}$$

where 255 is the peak pixel value and MSE is the mean squared error of the luminance pixels. It is important to point out that no single measurement can encompass all characteristics of the video degradation; however, PSNR is widely used in the field and is useful as a rough indicator of video quality. In this thesis, PSNR will be used to plot results and to recognize general trends (i.e. the video quality is getting better or worse). However, the final test will always be visual inspection. Where it is instructive, the measured PSNR will be accompanied by a written description of the compression artifacts.

2.1.4 Color Representation and Chrominance Subsampling

The human eye detects different amounts of red, green, and blue light, and combines them to form the various colors. For this reason, video capture and display devices typically use the RGB (Red, Green, Blue) color space. Another way to represent color information is the YUV color space. YUV is related to RGB by the following matrix equation:

$$\begin{bmatrix} \mathrm{Y} \\ \mathrm{U} \\ \mathrm{V} \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ -0.1687 & -0.3313 & 0.5000 \\ 0.5000 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{bmatrix}. \tag{2.3}$$

The Y component represents the *luminance* (brightness) of a color. The U and V components represent the *hue* and *saturation*, respectively, and together are the two *chrominance* components. The advantage of separating the luminance and chrominance is that they can be processed independently. The human visual system is more sensitive to high frequency variation in the luminance component, and insensitive to high frequency variation in the two chrominance components. This characteristic is exploited by subsampling the chrominance information. MPEG defines the following chroma sampling formats: when the full chrominance components are used it is called 4:4:4; when the chrominance components are subsampled by two in the vertical dimension it is called 4:2:2, and when the chrominance components are subsampled by two in the vertical and horizontal dimensions it is called 4:2:0. The 4:2:0 chroma format reduces the total amount of information by 50%, and the loss is largely imperceptible. For this reason, the MPEG-2 main profile uses 4:2:0 chroma sampling.

2.1.5 MPEG Layers

MPEG divides the video sequence into various layers: from the smallest to the largest, these layers are block, macroblock, picture, group of pictures, and video sequence. The individual layers are described in the remainder of this section, including the techniques that allow MPEG to exploit redundant and irrelevant information and achieve high levels of compression.

On the smallest scale, MPEG divides each picture into non-overlapping 8x8 pixel squares, called blocks, which form the basic building block of an MPEG video sequence. Spatial redundancy is exploited on the block level by transforming the pixel information to frequency information using the *discrete cosine transform* (DCT). The DCT transforms the 64 pixel intensities into 64 coefficients which, when multiplied by 64 orthonormal basis functions, sum to form the original picture. For typical blocks, the DCT concentrates most of the energy into a relatively small number of the lower-frequency coefficients. After the DCT operation, the coefficients are *quantized* by dividing by a nonzero positive integer (called a *quantization value*, or *quantizer*) and discarding the remainder. Quantization is the lossy part of MPEG video compression. The human visual system is more sensitive to low frequency variations, meaning that the higher-frequency DCT coefficients can be represented with less precision without perceptible loss in image quality. As a result of the DCT and quantization, many of the coefficients will have a value of zero. The coefficients are encoded in order from low to high frequency, and runs of zero coefficients are grouped together with the subsequent nonzero value in a process called *run-level coding*. Finally, the distinct run-level events—different length strings of zeros followed by a nonzero value—are each assigned a unique, variable-length code word (Huffman code) depending on the probability of the event occurring. For example, events that are most likely to occur are given the shortest code words, while less likely events are given longer code words. Huffman codes are an example of *entropy coding*. Together, the discrete cosine transform, quantization, run-length coding, and entropy coding greatly reduce the amount of information required to represent the picture in a single, 8x8 pixel block.

A macroblock is a group of four blocks from the luminance component (a 16x16 pixel square) and one corresponding block from each of the two chrominance components. Temporal redundancy is exploited on the macroblock level through motion compensation. Because adjacent pictures are often similar, an estimate can be made by finding a similar 16x16 pixel region in one or two nearby pictures. Instead of encoding all the information in the macroblock, only the difference between the macroblock and its estimate is encoded. If the estimate is good, then the individual blocks that make up the macroblock will be encoded with very few bits. A macroblock that uses motion compensation is called *inter-coded*. Every inter-coded macroblock is accompanied by one or two motion vectors that indicate the location of the estimate in the neighboring reference pictures. A macroblock that does not use motion compensation is called *intra-coded*.

MPEG defines three different kinds of pictures: I, P, and B. A picture where every macroblock is intra-coded is called an I-picture. Predictive pictures (P-pictures) are encoded using motion compensation from a prior I- or P-picture. In a P-picture, a single motion vector accompanies each inter-coded macroblock; however, individual macroblocks may instead be intra-coded. Finally, bidirectionally-predictive pictures (B-pictures) are encoded using motion compensation from one prior and one subsequent I- or P-picture. A macroblock in a B-picture may be intra-coded, or reference one or two other pictures. B-pictures are never used to predict other pictures.

A group of consecutive pictures in MPEG is called a *group of pictures* (GOP). The first picture in a GOP is always an I-picture. The rest of the GOP structure is defined by the two parameters N and M: N is the total number of pictures in the GOP, and M is the distance between the I-picture and the first P-picture and between two successive P-pictures. An example GOP structure with N = 9 and M = 3 is shown in figure 2.1. The arrows between pictures in the figure indicate the reference pictures that are used for motion compensation. The individual GOP structure is repeated to make up the entire video sequence.

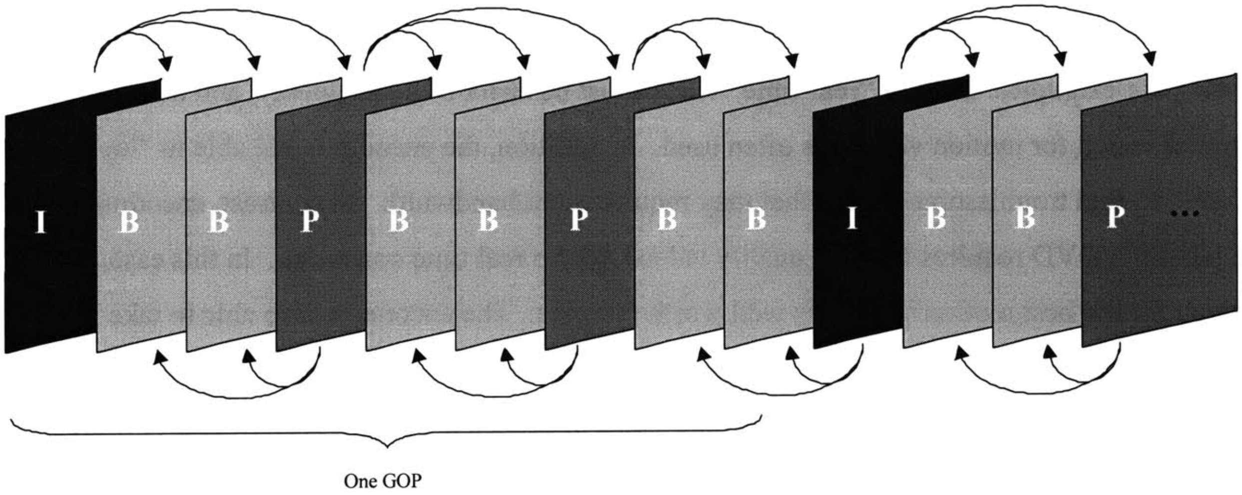


Figure 2.1: Example MPEG GOP Structure With N = 9 and M = 3

The first picture in a GOP is an I-picture. The parameter N indicates the total number of pictures in the GOP; M is the distance between the I-picture and the first P-picture and between consecutive P-pictures. The arrows indicate the reference pictures that are used for motion compensation.

2.2 Rate Control Strategy for MPEG Encoders

Encoding and decoding an MPEG bitstream is highly asymmetric in terms of complexity and computation. For example, the encoder must find motion vectors and decide what quantization values to use—the decoder simply does what it is told to do. The MPEG standard does not specify how an encoder should carry out these tasks; instead, it specifies the syntax that a compliant bitstream must follow. As a result, there is a great deal of flexibility in how a particular video sequence is encoded.

The optimal encoding strategy often depends on the target application and even the video content itself. For example, "live" or "real-time" video must be encoded very quickly, and a narrow, limited search for motion vectors is often used. In addition, the encoder is not able to "look ahead" to find troublesome scenes that may require more bandwidth. In contrast, encoding a movie for a DVD requires the best quality video with no real time constraint. In this case, a full search for the best motion vectors is well worth the time. The encoder is also able to take bandwidth from easily encoded scenes and use it during more difficult scenes. In order to find where these tradeoffs should occur, the entire video sequence is often processed two or more times.

Constant bitrate applications such as HDTV are encoded using the rate control strategy shown in figure 2.2. The transmission channel is preceded by an encoder buffer and followed by a decoder buffer; the purpose of the buffers is to absorb instantaneous variations in the bitrate. Data is transferred between the two buffers at a constant bitrate over the transmission channel, but added and removed at a variable bitrate by the encoder and decoder. The encoder is responsible for protecting against buffer overflow and underflow, and it does so primarily by changing the *quantizer scale factor*. As mentioned previously, quantization allows the DCT coefficients to be represented with various levels of precision. MPEG defines a quantizer scale factor, Q, in the range $\{1, 2, 3, \ldots, 31\}$ that may be changed on the macroblock level. Increasing Q tends to decrease the instantaneous encoder output, while decreasing Q has the converse effect. The encoder monitors the buffer fullness and changes Q appropriately when the buffer levels are too high or too low. The strategy for choosing the appropriate quantizer scale factor may be a simple function of buffer fullness, or it may also involve looking ahead at the video content to see how many bits will be actually required (note that for "live" or "real-time" encoding, the encoder cannot look ahead). In either case, a good encoder will attempt to maximize the video quality while providing a constant stream of data to the transmission channel.

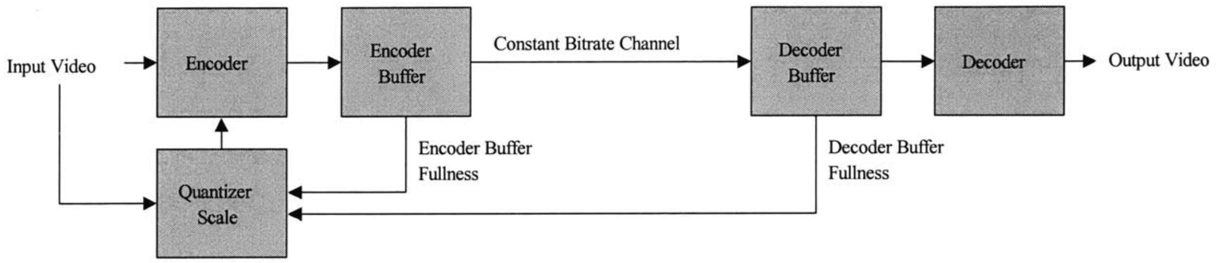


Figure 2.2: Rate Control Strategy for MPEG Encoders

The transmission channel is surrounded by an encoder buffer and decoder buffer whose purpose is to absorb instantaneous variation in the bitrate. The encoder protects against buffer overflow or underflow by changing the quantizer scale factor, Q. Using a larger value of Q results in smaller instantaneous bandwidth because the DCT coefficients are represented with less precision, and vice versa. The strategy for choosing the quantizer scale factor may be a simple function of buffer fullness or a more complicated decision involving the video content.

2.3 Base Layer Coding for Adaptive Format Conversion

With the understanding of MPEG-2 video compression developed in this chapter, it is possible to revisit the first motivation for this thesis in more detail—namely, understanding how adaptive format conversion performs when the base layer video is compressed in a manner typical to HDTV. This section begins by describing how different quality base layers were created in the previous implementation and points out that it not the strategy used to encode HDTV. In order to relate the current results to the migration path, a video codec is used that implements a practical encoding approach. The encoder that is used in this thesis (Test Model 5) is described, including its performance characteristics. The last section of this chapter describes how the base layer video is encoded for the experiments in this thesis.

2.3.1 Previous Implementation

As formerly mentioned, Wan's objective was to determine how an encoded base layer influenced the performance of an AFC enhancement layer. The results were given as a function of base layer quality. In this previous implementation, the base layer was encoded using all I-pictures, and the quantizer scale factor, Q, was a single, fixed value. As a result, no motion compensation was used, nor was the bitrate constant. The various base layer qualities were created by changing the quantizer scale factor. The motivation behind this encoding strategy was that it resulted in equal distortion within each picture and from picture to picture—one drawback was that the base layer bandwidth had no meaningful value. In HDTV, video quality is determined by the channel bandwidth, not artificially controlled in this way. Although the domain of previous experiments certainly includes the level of video quality that is expected for HDTV, it is impossible to determine from this work alone exactly how AFC performs for typical HDTV bitstreams.

2.3.2 Test Model 5 Video Codec

The MPEG-2 codec that is used in the current implementation is the MPEG Software Simulation Group's Test Model 5 Video Codec (TM5) [18, 19]. The codec was developed during the collaboration phase of MPEG-2 in order to determine the merit of proposed compression techniques and resolve ambiguities in what became the official MPEG-2 standard. For this thesis, TM5 was chosen for its numerous advantages. First, the codec is free and open-source. It is also well documented, easy to configure, performs full, half-pel accuracy motion vector searches, and uses the buffered rate control strategy discussed previously (the quantizer scale factor is a function of buffer fullness only). Finally, TM5 has relatively good performance for the compression levels of interest. Figure 2.3 shows the PSNR of compressed bitstreams as a function of the bitrate (measured in bits per pixel) for the TM5 codec. The 1080I format, for example, is normally encoded at approximately 0.3 bpp, and at this bitrate TM5 yields an average video quality between 31 and 32 dB. Visual inspection of the compressed video sequence reveals a small amount of ringing or blurring in detailed areas such as text and sharp edges. Although these compression artifacts are noticeable, their level seems consistent with actual broadcast HDTV. Notice that the compression of progressive video is more efficient than interlaced video (this is generally true for any MPEG encoder). Figure 2.3 was created by averaging the results for ten different high-definition video sequences.

Using an encoder such as TM5 means that the results of this thesis are tied to a specific encoder technology; in other words, the generality of previous work is lost. However, because the TM5 encoder has a reasonable level of performance, the results are useful in evaluating the proposed migration path. Prior research shows that the quality of AFC enhanced video improves with increasing base layer quality, so that when video compression technology improves in the future, AFC enhancement information will be even more beneficial than reported in this thesis.

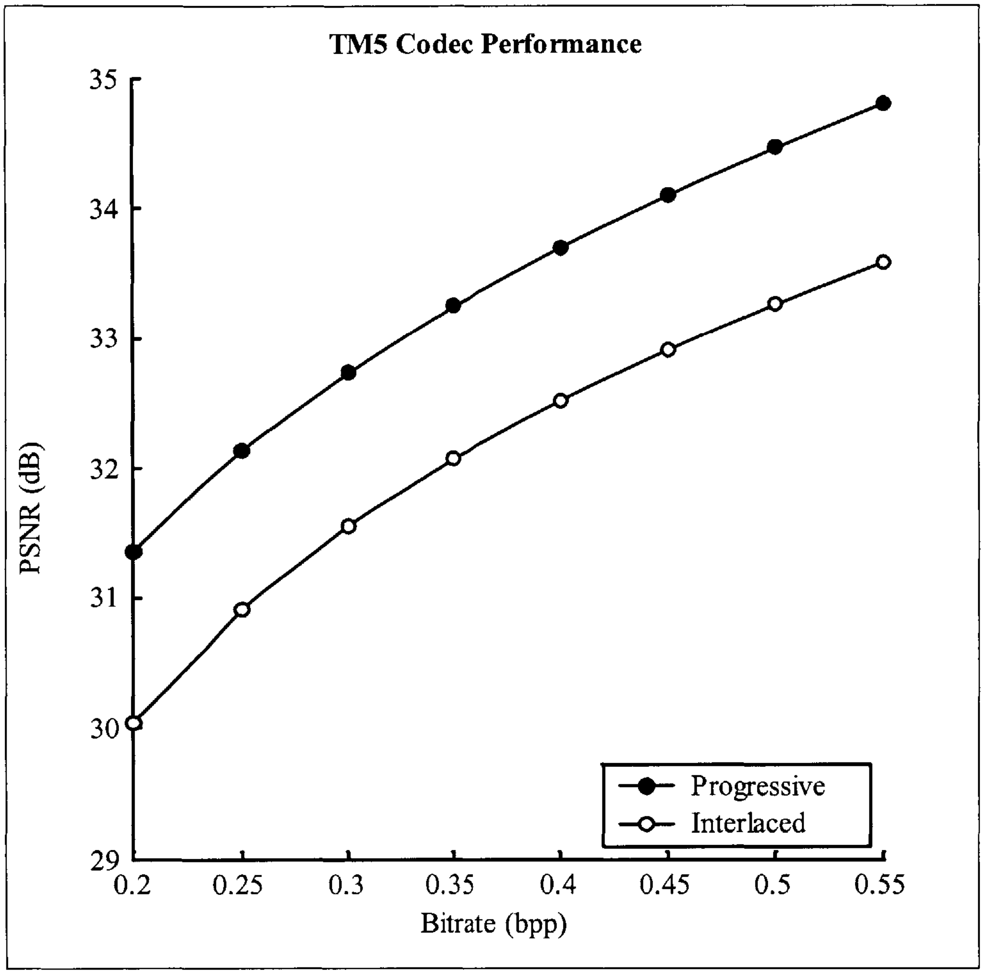


Figure 2.3: TM5 Codec—PSNR Versus Bitrate

In this figure, the PSNR of progressive and interlaced video is measured as a function of bitrate for the TM5 video codec. The standard HDTV format 1080I is encoded at approximately 0.3 bpp. At this bitrate, TM5 yields a compressed video quality (PSNR) between 31 and 32 dB. Visually, the compression artifacts from TM5 seem consistent with actual broadcast HDTV.

2.3.3 Codec Configuration for Adaptive Deinterlacing

For the adaptive deinterlacing experiments in this thesis, the TM5 codec is configured to encode the base layer video as follows using the MPEG-2 main profile at high level. The GOP structure is characterized by the parameters N=15 and M=3. For macroblock prediction, a full, half-pixel accuracy search is made for the best estimate. The search range is limited to 15 pixels in each direction for P-pictures, and 7 pixels in each direction for B pictures. Interlaced fields are encoded separately. The encoder is given a target bitrate and uses the encoding strategy for constant rate bitstreams described earlier in this chapter.

Because many of the video test sequences used in this thesis have different spatial resolutions, the target bitrate is normalized by the number of pixels. For example, the bandwidth available for video in an HDTV channel is approximately 19 Mbps. When compressed to this size, the 1080I format has a normalized bitrate of

$$\frac{19\,\text{Mbits}}{\text{sec}} \cdot \frac{\text{sec}}{60\,\text{fields}} \cdot \frac{2\,\text{fields}}{\text{frame}} \cdot \frac{\text{frame}}{1080\text{x}1920\,\text{pixels}} \approx 0.3\,\text{bpp} \ . \tag{2.4}$$

Using a normalized measurement allows different spatial formats to be compressed by the same amount and allows results to be compared from sequence to sequence.

2.4 Summary

This chapter introduced video coding terminology and concepts that are used throughout the remainder of the thesis. The MPEG-2 main profile at high level is the video compression algorithm used in the U.S. HDTV standard. A migration path is necessary because the 1080P video format exceeds the sample rate constraint in the MP@HL definition.

To encode video for HDTV, MPEG uses a large set of compression tools, including different picture types for motion compensation. In addition, the quantizer scale factor is controlled in order to create constant bitrate video streams. This is in contrast to the base layer coding technique used in previous research on AFC enhancement. The prior implementation used a fixed quantizer scale factor to create video with constant distortion; it did not use motion compensation or attempt to optimize the bitstream for any particular bandwidth. A fundamental motivation for this thesis is to encode the base layer as it is typically encoded for HDTV. The MPEG Software Simulation Group's Test Model 5 codec is used for this purpose. TM5 is openly available with relatively good performance for the type of compression found in high-definition television. The last section of this chapter explained how the TM5 codec is configured to encode the base layer video for the experiments in this thesis.

Chapter 3

Adaptive Format Conversion

The previous chapter described base layer video coding. This chapter explains how an adaptive format conversion enhancement layer is coded. The migration path proposed in this thesis uses a base layer video format that conforms to the current HDTV standard, ensuring backward compatibility. The AFC enhancement layer is information that may be used to increase the video resolution beyond what is allowed by MPEG-2.

In a scalable coding scenario, there are two types of information that can be included in an enhancement layer: residual and adaptive format conversion. Examples of residual coding are common. For instance, the spatial scalability profiles in MPEG-2 and MPEG-4 specify a fixed format conversion method, and the residual is encoded in an enhancement layer. Adaptive format conversion information is a different type of enhancement data. Instead of using a single format conversion technique, several different methods are used. Because the encoder has access to the original video sequence, before information is lost through downsampling and base layer coding, the conversion back to the original resolution can be done intelligently. This is accomplished by partitioning the video sequence into nonoverlapping blocks and choosing the best interpolation method for each block. In a codec based on adaptive format conversion, the encoder and decoder both have knowledge of the format conversion methods that are used; the enhancement layer simply describes how the video sequence is partitioned and which conversion method is used in each block. Compared to residual coding, the bandwidth required for an AFC enhancement layer is small. The advantage of adaptive format conversion is that it can use interpolation techniques that may not work well overall, but have superior performance in certain situations. Compare this to nonadaptive format conversion, which requires a single conversion technique for the entire video sequence—one that gives adequate performance in all situations.

This thesis considers a particular example of adaptive format conversion: adaptive deinterlacing. Interlacing was introduced as part of the NTSC standard as a tradeoff between spatial and temporal resolution. Although it fulfills its intended purpose well, interlacing also creates undesirable artifacts such as flickering and distortion of moving objects. In contrast, current technology such as computer screens and HDTV use progressive video, which is not so afflicted. As shown earlier in table 1.1, five of the six common HDTV video formats are progressive, and the remaining format, 1080I, is interlaced. Because NTSC has been the dominant technology for over fifty years, most video capture and display devices are designed to handle interlaced video. Support for interlaced video in the new HDTV standard was arguably included for compelling economic reasons rather than technological necessity.

Adaptive deinterlacing is the same example studied by Sunshine and Wan. The implementation used in this thesis is the same as that developed by Wan, and is repeated in this chapter for convenience. The chapter begins by describing the four deinterlacing techniques that are used, followed by an explanation of the frame partitioning and parameter coding.

3.1 Deinterlacing Methods

Interlaced video is created by discarding either the even or odd lines in alternating frames; deinterlacing is the process of replacing the information that was lost. Deinterlacing techniques rely on temporal and spatial similarities in order to estimate the missing pixel values. For example, inter-frame techniques use information at the same spatial location in a previous or subsequent field. These methods work well for relatively stationary regions. Intra-frame techniques are better for regions in motion, and use information from surrounding pixels in the same frame. The AFC implementation described in this chapter uses four different deinterlacing techniques: forward field repetition (FFR), backward field repetition (BFR), linear interpolation, and Martinez-Lim deinterlacing. These methods have different advantages and are described in detail below.

Forward field repetition is an inter-frame deinterlacing technique that replaces the missing lines in a field with the corresponding line from the previous field. This is illustrated in figure 3.1 (a), where each missing pixel in frame N is replaced by the pixel in the same spatial location in frame N-1. Adaptive format conversion does not use FFR to deinterlaced the first frame in a video sequence because there is no previous frame.

Backward field repetition is another inter-frame deinterlacing technique. Figure 3.1 (b) shows that for BFR, each missing pixel in frame N is replaced by the pixel in the same spatial location in the next frame, frame N+1. In this case, BFR is not used to deinterlace the last frame of the video sequence. Together, FFR and BFR are simple to implement and sufficiently replace the missing information in areas where there is little motion from picture to picture. It is easily observed that if there is no motion, these techniques preserve all the detail in the picture.

The first intra-frame deinterlacing technique is linear interpolation. For linear interpolation, each missing pixel is replaced by an average of the pixels immediately above and below it in the same frame, as illustrated in figure 3.1 (c). When the top line is missing, it is replaced by the line below it; the bottom line is replaced by the line above it. Because linear interpolation uses average values, this method tends to blur the detail in the picture; however, if there is motion between frames, linear interpolation usually gives a better estimate than either of the two inter-frame techniques. The human visual system is not as sensitive to detail in moving objects, so the smearing effect often goes unnoticed. Computationally, linear interpolation is easy to implement.

Martinez-Lim deinterlacing is the most complex of the four deinterlacing techniques. The basic idea is that for small regions, adjacent lines are related by a simple horizontal shift of Δx pixels/line. A missing pixel is the average of the two pixels that are located Δx pixels to the left on the line below it and Δx pixels to the right on the line above it in the same frame. The parameter for the best line shift, Δx, is determined using the algorithm in [10], which fits the surrounding region (in this case the closest five pixels in each adjacent line) to a polynomial model and uses the model to find a least squares estimate of the horizontal shift, Δx, rounded to the nearest pixel. Lines on the top and bottom are replaced by their nearest neighbor, and pixels on the left and right that cannot be modeled with a line shift are interpolated linearly. Figure 3.1 (d) illustrates this technique when the horizontal shift is estimated to be one pixel per line ($\Delta x=1$). Martinez-Lim deinterlacing generally produces sharper images than linear interpolation, but the tradeoff is added complexity.

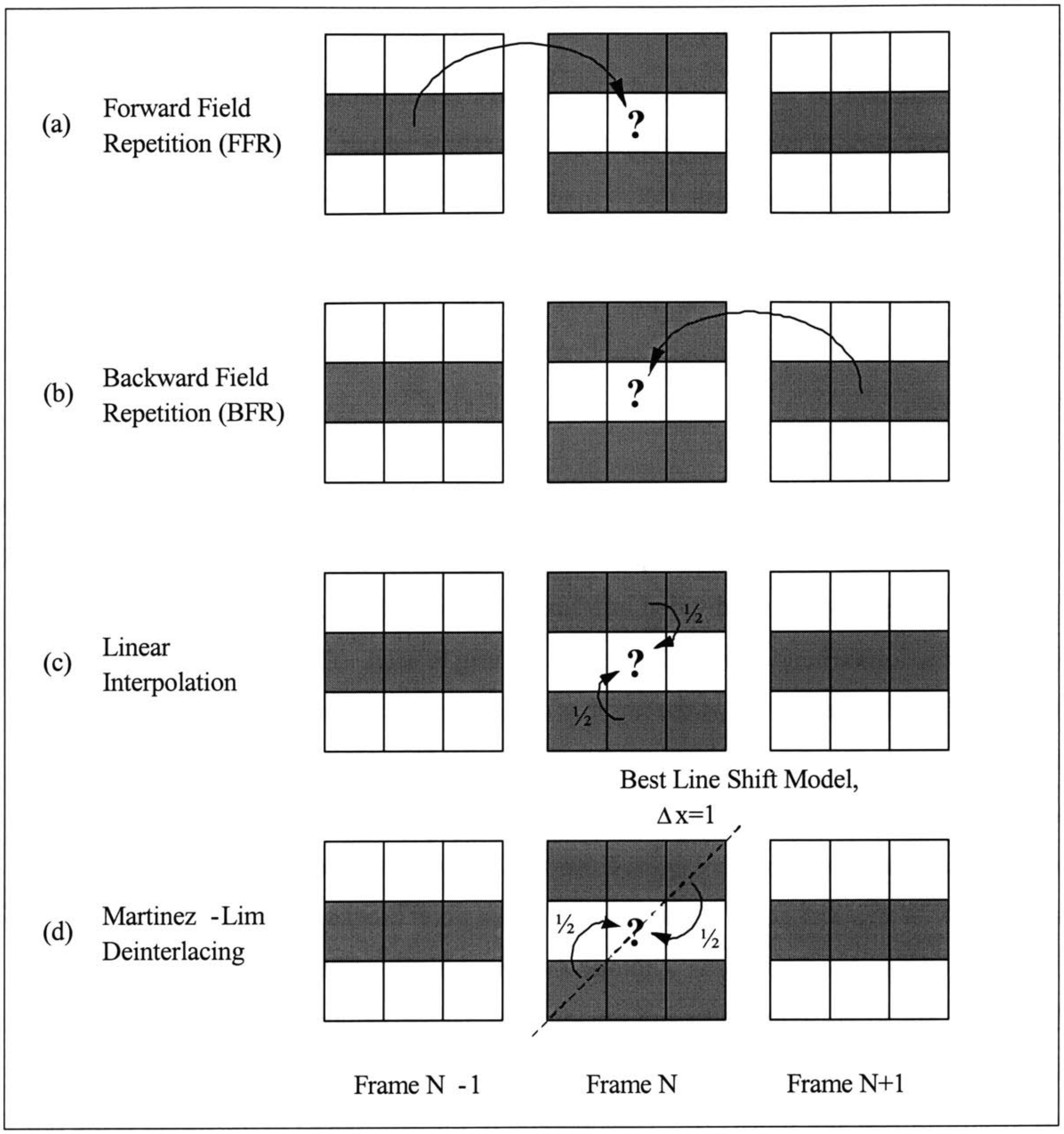


Figure 3.1: Deinterlacing Methods

The four deinterlacing methods used in this thesis are forward field repetition (FFR), backward field repetition (BFR), linear interpolation, and Martinez-Lim deinterlacing. This figure illustrates how a missing pixel is estimated from its surroundings in each of these techniques. FFR and BFR are inter-frame techniques that are generally good for stationary regions. Linear interpolation and Martinez-Lim deinterlacing are intra-frame techniques that are better for regions in motion. Martinez-Lim deinterlacing uses a model of the surrounding region to determine the best integer-valued line shift and is the most complex of the four techniques.

3.2 Frame Partitioning

This AFC implementation features two different frame partitioning schemes: fixed and adaptive. Fixed partitioning divides each frame into nonoverlapping, square blocks of equal size. Three different block sizes are considered: 16x16 pixels, 8x8 pixels, and 4x4 pixels. The format conversion technique that is selected for each block is simply the method that results in the highest PSNR for that block. The enhancement layer is a list of the methods used, one parameter per block, in a set order. The enhancement layer bandwidth is determined primarily by the total number of blocks.

For adaptive frame partitioning, each frame is initially divided into nonoverlapping, 16x16 pixel blocks. If needed, each of these 16x16 pixel blocks can be subdivided into four 8x8 pixel blocks, each of which can be further subdivided into four 4x4 pixel blocks. Figure 3.2 shows all the possible permutations when adaptive frame partitioning is used. The rate given in the figure is the number of sub-blocks, which is the number of parameters that must be encoded if that partitioning is used. There is only one way to subdivide the 16x16 pixel block into one, four, and sixteen sub-blocks. For rate equals seven, there are four ways to subdivide the block into three 8x8 pixel blocks and four 4x4 pixel blocks. Likewise for a rate of ten, there are six permutations with two 8x8 pixel blocks and eight 4x4 pixel blocks. Finally for a rate equal to thirteen, there are four permutations with one 8x8 pixel block and twelve 4x4 pixel blocks. All together, there are six different rates and seventeen possible permutations. Like before, the format conversion technique that is chosen for each sub-block is the one that results in the highest PSNR for that sub-block. The enumeration of the sub-blocks in the figure is the order that the parameters are encoded. Not only does adaptive frame partitioning concentrate the available bandwidth where it is needed the most, but it also allows finer control over the enhancement layer bandwidth.

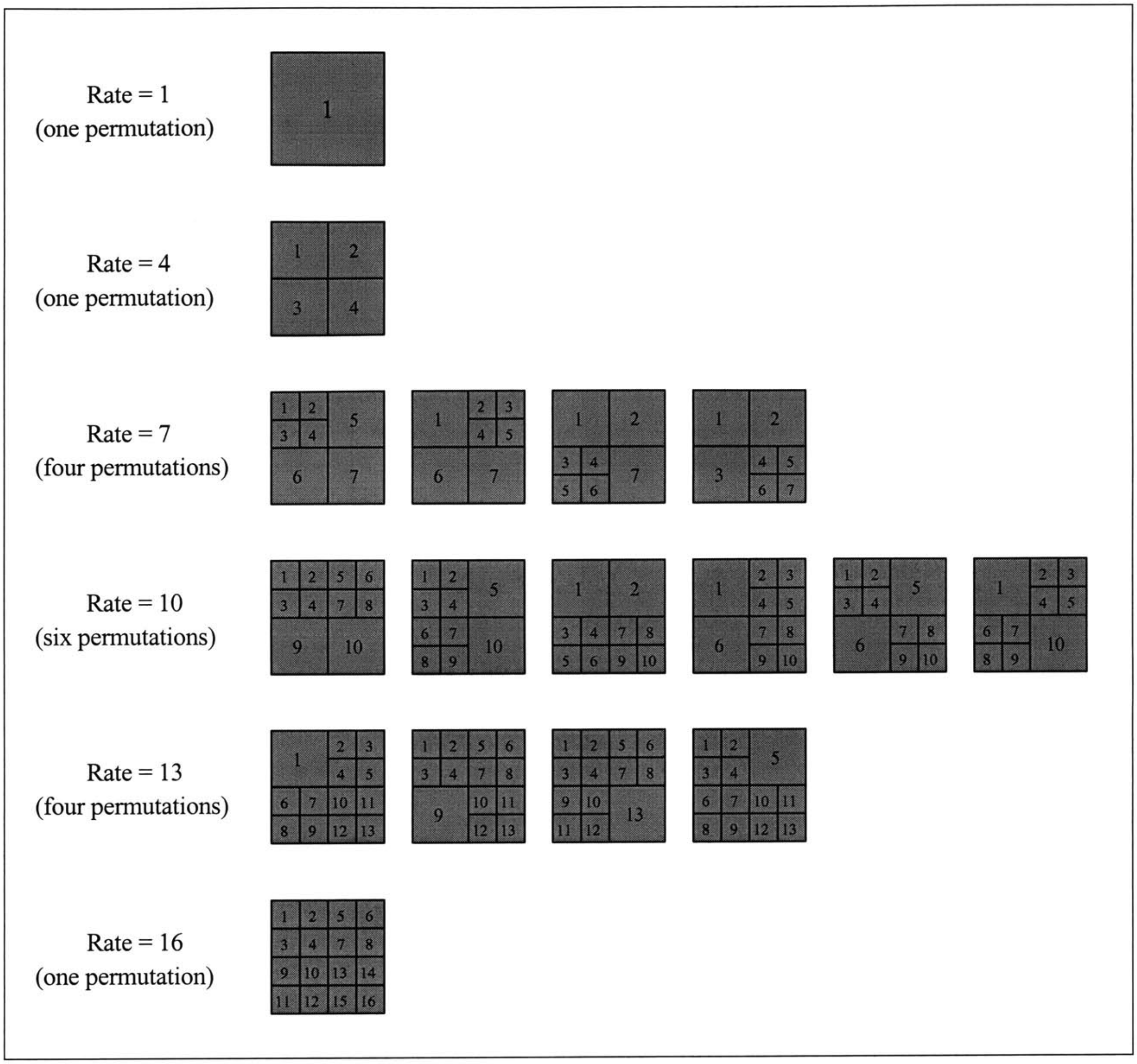


Figure 3.2: Seventeen Permutations for Adaptive Frame Partitioning

Adaptive frame partitioning initially divides a frame into nonoverlapping 16x16 pixel blocks. Each block is further subdivided into one of the permutations shown here. The format conversion method that is selected for each sub-block is the one which results in the highest PSNR for that sub-block. The rate indicates the number of sub-blocks in a permutation, and is also the number of parameters that must be encoded when that permutation is used. The sub-block enumeration indicates the order in which the parameters are encoded. There are six different rates and seventeen possible permutations. Using adaptive block sizes concentrates the available bandwidth on areas that need it the most and allows finer control over the enhancement layer bandwidth than fixed block sizes.

When using adaptive block sizes, it is desirable to maximize the PSNR given a particular enhancement layer bandwidth. This is done by choosing a partitioning for each block such that the sum of the rates equals a desired value, and the sum of the distortion is minimized. To state this another way, let us define $D_{i,p}$ as the distortion in the ith 16x16 pixel block when the pth partitioning is used and $R_{i,p}$ as the rate of the corresponding permutation. Then for every block, i, we choose a partitioning, p, that minimizes

$$\mathrm{D}_{\mathrm{TOTAL}} = \sum_i \mathrm{D}_{i,p} \tag{3.1}$$

subject to the constraint that

$$\mathrm{R}_{\mathrm{TOTAL}} = \sum_i \mathrm{R}_{i,p} \leq \mathrm{R}_{\mathrm{DESIRED}} \, . \tag{3.2}$$

The distortion metric, D, is the mean squared error (MSE), since

$$\begin{aligned} \mathit{arg\ max}\ \mathrm{PSNR}_{\mathrm{TOTAL}} &= \mathit{arg\ max}\ 10\log_{10}\frac{255^2}{\mathrm{MSE}_{\mathrm{TOTAL}}} && (3.3) \\ &= \mathit{arg\ max}\ 10\log_{10}\frac{255^2}{\sum_i \mathrm{MSE}_i} && (3.4) \\ &= \mathit{arg\ min}\ \sum_i \mathrm{MSE}_i \ . && (3.5) \end{aligned}$$

The classic solution to this type of budget constrained allocation problem is Lagrangian optimization [20, 21]. A Lagrange cost function, J, and multiplier, λ, are defined as follows:

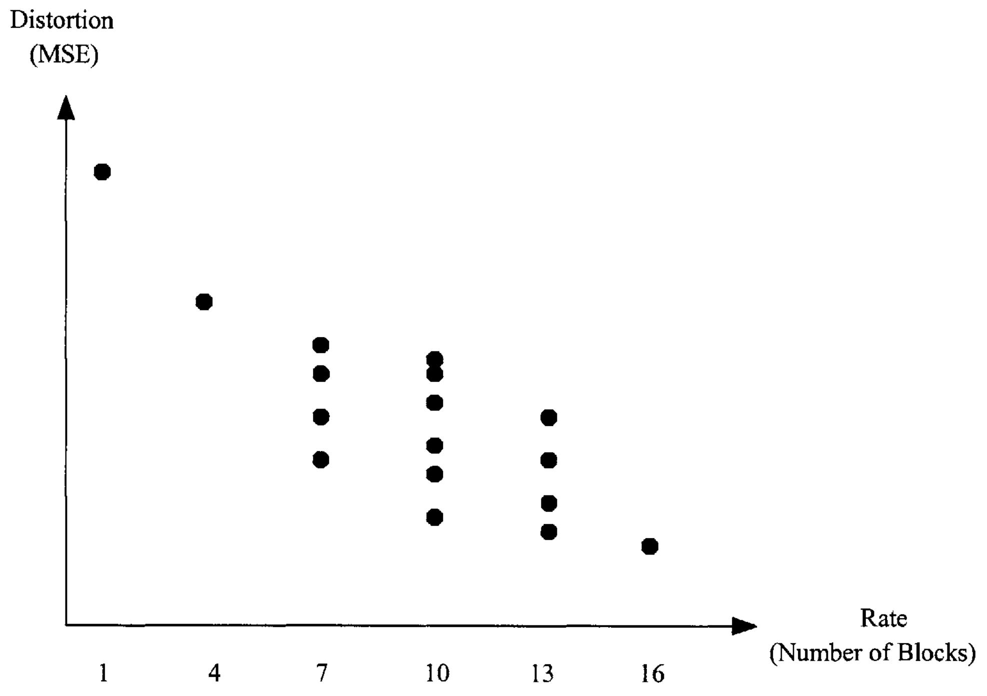
$$\mathrm{J}(\lambda) = \sum_i \left(\mathrm{D}_{i,p} + \lambda \mathrm{R}_{i,p}\right) \, , \tag{3.6}$$

where $\lambda \geq 0$. For a particular value of λ, the R, D pairs that minimize the Lagrange cost function are also the optimal solutions to the budget constrained allocation problem. This minimization can be computed independently for each block as follows:

$$\mathit{min}\,\mathrm{J}(\lambda) = \mathit{min}\left[\sum_i \left(\mathrm{D}_{i,p} + \lambda \mathrm{R}_{i,p}\right)\right] \tag{3.7}$$

$$= \sum_i \mathit{min}\left(\mathrm{D}_{i,p} + \lambda \mathrm{R}_{i,p}\right) . \tag{3.8}$$

To illustrate how this technique is used, a rate-distortion plot for a single 16x16 pixel block is shown in figure 3.3, where each of the seventeen permutations corresponds to a single point in the plane. For every block, *i*, the permutation, *p*, with the smallest value of $\mathrm{D}_{i,p} + \lambda\ \mathrm{R}_{i,p}$ is selected. Note that only the permutation with the lowest distortion for each rate needs to be considered in the minimization. When λ equals zero, this technique is equivalent to minimizing the total distortion; when λ is arbitrarily large, the result is the smallest total rate. An intermediate value of λ can be found that corresponds to a desired enhancement layer bandwidth.



Wan, 2000, used with permission

Figure 3.3: Example Rate-Distortion Plot for Adaptive Frame Partitioning

The seventeen different permutations of a single 16x16 pixel block are plotted according to their rate and distortion. The method of Lagrangian optimization says that for each 16x16 pixel block, the permutation with the lowest value of D + λ R should be chosen in order to minimize the total distortion given a fixed bandwidth. Only the permutation with the lowest distortion for each rate needs to be considered in the minimization. For $\lambda = 0$ and λ large, this technique is the same as minimizing the total distortion and total rate, respectively. An intermediate value of λ yields the optimal tradeoff between rate and distortion.

3.3 Enhancement Layer Coding

The information in an AFC enhancement layer describes the way a frame is partitioned and the format conversion techniques that are used in each of the blocks. For fixed block sizes, the partitioning information is trivial—one number that denotes the block size. For adaptive block sizes, partitioning information is required for every 16x16 pixel block. The format conversion methods are assigned unique, variable-length code words (Huffman codes) based on the *a posteriori* probability that each method is used. For ease of discussion, these code words will be called method codes. In this implementation, new method codes are calculated for every frame, and a *code book* that describes the code word assignment is given at the beginning of each frame. For adaptive block sizes, the seventeen permutations for each 16x16 pixel block are also assigned Huffman codes based on *a posteriori* probability. These code words will be called partition codes and may be the same as the method codes. For adaptive block sizes, a code book with the partition *and* method codes is provided at the beginning of every frame. The use of Huffman codes significantly reduces the enhancement layer bandwidth, and a similar compression technique would be employed in any practical encoder implementation.

For both fixed and adaptive block sizes, the frames are initially divided into 16x16 pixel blocks which are encoded in a set order from top to bottom, left to right. For adaptive block sizes, each 16x16 pixel block is first described by a partition code. In both cases, the method codes for the 16x16 pixel sub-blocks are listed in the order that is enumerated in figure 3.2. The additional overhead needed to represent partitioning information makes the enhancement layer for adaptive block sizes slightly larger than that of fixed block sizes when the same number of blocks is encoded. The tradeoff is better allocation of bits and a more finely scalable enhancement layer bandwidth.

Recall that the method of Lagrangian optimization for selecting adaptive block sizes is optimal in the rate-distortion sense where the rate is the number of sub-blocks, or method codes, that must be listed. This method may no longer be optimal after entropy coding is introduced for the partition and method parameters.

3.4 Summary

This chapter explained how the enhancement layer is encoded in this implementation of adaptive format conversion. The focus of this thesis is adaptive deinterlacing using four different deinterlacing techniques: forward field repetition, backward field repetition, linear interpolation, and Martinez-Lim deinterlacing. The enhancement layer describes how a video sequence is partitioned and which deinterlacing method is used in each block. Two frame partitioning schemes are employed: fixed and adaptive. Fixed partitioning divides the frame into 4x4, 8x8, or 16x16 pixel blocks. The enhancement layer bandwidth for fixed partitioning is determined primarily by the total number of blocks. Adaptive partitioning allows seventeen different subdivisions of a single, 16x16 pixel block where the optimal subdivision is selected using the method of Lagrangian optimization. Adaptive partitioning allows the enhancement layer bandwidth to be more finely scalable by concentrating available bits where they are most needed. Finally, the enhancement layer is encoded using Huffman code words for the partitioning and conversion method paramenters.

Chapter 4
Results

The first experiment in this thesis answers the motivating question, *"How does adaptive format conversion perform when the base layer video is compressed in a manner typical to high-definition television?"* Recall that in previous implementations, the base layer was either not encoded or encoded in a fashion that was not meaningful to HDTV. In this implementation, the base layer video is encoded in a way that is consistent with the common HDTV format, 1080I. The base layer video is encoded at 0.3 bpp, and the PSNR of the adaptively deinterlaced video is given as a function of enhancement layer bandwidth. The results of this experiment are contained in section 4.1.

The second experiment answers the question, *"What is the optimal tradeoff between base layer and enhancement layer bandwidth?"* For this experiment, the base layer is interlaced, and the total bandwidth is shared between the base and enhancement layers. Ten video sequences with different characteristics are used in order to see if an optimal tradeoff exists—one that consistently provides the highest quality enhancement layer video without sacrificing base layer video quality. The results of this experiment are found in section 4.2.

4.1 PSNR Versus Enhancement Layer Bandwidth

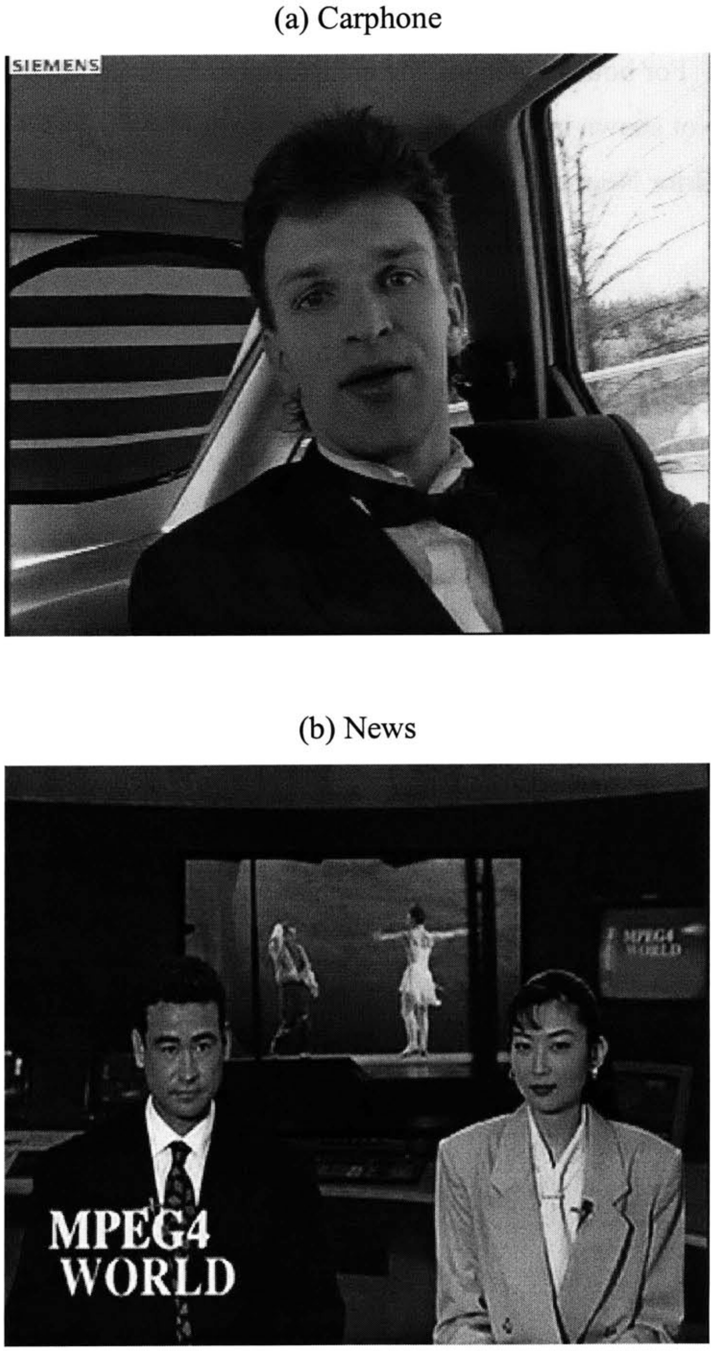
In the first experiment, an interlaced base layer is encoded at 0.3 bpp in the same way that video is encoded for HDTV. The PSNR of the adaptively deinterlaced video is measured as a function of enhancement layer bandwidth using both fixed and adaptive block sizes. When the block size is fixed, the enhancement layer bandwidth is determined primarily by the number of 4x4, 8x8, and 16x16 pixel blocks. For adaptive block sizes, the enhancement layer bandwidth is more finely scalable and is determined by the Lagrange multiplier, λ. Many different values of λ are used in order to show the behavior across the entire domain. Measurements of the enhancement layer bandwidth are normalized by the number of pixels in the base layer (*not* the enhancement layer—this is different from previous implementations).

Sunshine performed a similar experiment using original (not encoded) base layer video. The result may be considered an empirical upper bound on the performance of AFC enhancement information since the encoder had access to perfect information. Wan extended the analysis by using encoded base layers of variable quality, which were created using all I-pictures and a fixed quantizer scale factor. These results were indicative of the kind of performance that can be expected for a broad range of scalable coding applications. The current implementation uses the TM5 codec configuration described in section 2.3.2. The results of this experiment are useful when evaluating adaptive format conversion as a proposed migration path for high-definition television.

The two video test sequences used by Wan—Carphone and News—are also used in this experiment. With the same test sequences, the base layer encoder is the only difference between this and past work. Figure 4.1 shows the first frame of the Carphone and News sequences, which are CIF (Common Intermediate Format) resolution of 288x352 pixels and 60 frames long. The Carphone sequence is characterized by a man speaking to the camera in an animated fashion while scenery passes quickly in the car window. In the News sequence, two news anchors speak in the foreground while dancers perform ballet on a large television screen in the background.

The results are given in figure 4.2, which shows the PSNR of the adaptively deinterlaced video as a function of enhancement layer bandwidth for the Carphone (a) and News (b) video test sequences. Results are shown for adaptive block sizes (circles), fixed block sizes (squares), and the best nonadaptive deinterlacing method (diamonds). The best nonadaptive deinterlacing method is Martinez-Lim deinterlacing for Carphone and linear interpolation for News. Notice that the overhead required to transmit partitioning information for adaptive block sizes is about 0.01 bpp, but that the use of adaptive block sizes results in a higher PSNR for most enhancement layer bandwidths. The PSNR improvement for the Carphone sequence, compared to the best nonadaptive method, is between 3 and 4 dB; for the News sequence, the PSNR improvement is between 3 and 3.5 dB. For both sequences, the enhancement layer bandwidth is between about 0.015 and 0.25 bpp. Not shown in the figure is the base layer PSNR, which is 36.1 dB for Carphone, and 34.9 dB for News.

Despite the different encoder implementation, the PSNR improvement is almost exactly the same as reported by Wan for these values of base layer PSNR. Another important result from previous work is that for these base layer qualities (above 30 dB) and enhancement layer bandwidths (less than 0.25 bpp), adaptive format conversion is a better choice than residual coding.

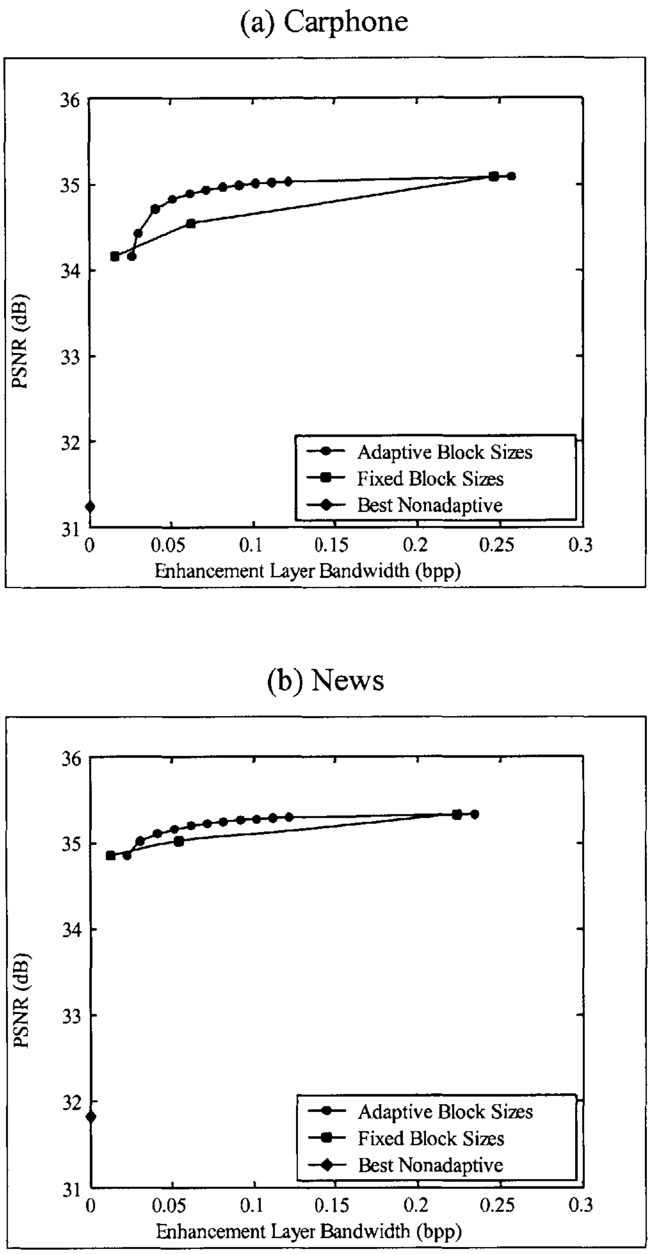


(a) Carphone

(b) News

Figure 4.1: First Frame of the Carphone and News Sequences

This figure shows the first frame of the Carphone (a) and News (b) video test sequences. These sequences are CIF resolution and 60 frames long. In the Carphone sequence, a man speaks animatedly towards the camera while scenery passes by out the window. The News sequence shows two news anchors speaking in the foreground and two dancers performing ballet on a large television screen in the background.


(a) Carphone

(b) News

Figure 4.2: PSNR Versus Enhancement Layer Bandwidth for Carphone and News

These graphs show the enhanced video PSNR versus enhancement layer bandwidth for the Carphone (a) and News (b) video test sequences, which were interlaced and encoded at 0.3 bpp. Results are for adaptive and fixed block sizes as well as the best nonadaptive deinterlacing method. Notice that the overhead required to transmit partitioning information is about 0.01 bpp, but that using adaptive block sizes results in higher PSNR for most enhancement layer bandwidths. The PSNR improvement (compared to the best nonadaptive method) is between 3 and 4 dB for the Carphone sequence and between 3 and 3.5 dB for the News sequence. The enhancement layer bandwidth is between about 0.015 and 0.25 bpp for both sequences.

4.2 Tradeoff Between Base and Enhancement Layer Bandwidth

In this experiment, the total bandwidth (base plus enhancement layer) is held constant at 0.32 bpp. This number was chosen because it represents a small increase in the video bandwidth that will likely become available before a migration path is implemented. The base layer is allocated a percentage of the total bandwidth (from 75% to 100%) with the remainder allocated to the AFC enhancement layer. Results are calculated for fixed, 16x16 pixel blocks and for adaptive block sizes, which allow finer bandwidth scalability.

Measurements of the base layer and enhancement layer bandwidth are both normalized by the number of pixels in the base layer. This normalization allows the tradeoff between base and enhancement layer bandwidth to be discussed more easily. Normalized measurements also allow results to be compared for video sequences with different spatial resolutions.

The tradeoff experiment is performed using each of the high-definition video test sequences in table 4.1. The table gives the name of the sequence, the spatial resolution, number of frames, and a brief description of the video content. Each video sequence is discussed individually in the sub-sections below.

The purpose of these tests is to determine if there exists an optimal tradeoff between base layer and enhancement layer bandwidth for many different types of video. With the total bandwidth constant, allocating more bandwidth to the enhancement layer will necessarily cause the base layer quality to diminish. In a proposed migration path, the integrity of the base layer video is important in order to ensure backward compatibility. An optimal tradeoff would consistently yield the highest quality enhancement layer video without significantly disturbing the base layer video quality.

Name	Rows	Columns	Frames	Description
Car	480	720	16	Remote control car and bouncing soccer balls.
Football	720	1024	60	A play in a football game.
Football Still	720	1024	60	The first frame of the football sequence, repeated.
Marcie	880	1200	16	Head and shoulders of a woman.
Girl	512	512	60	Panning still picture of girl (with test graphics).
Toy Train	720	1280	60	Moving toys on a table with passing train.
Tulips Scroll	720	1280	60	Horizontal scrolling of tulips image.
Tulips Zoom	720	1024	60	Zoom in on tulips image.
Picnic	720	1024	60	Zoom in on still picture of a picnic.
Traffic	880	1200	30	Traffic near a mall parking lot.

Table 4.1: List of High-definition Video Test Sequences

This is a list of the high-definition test sequences that are used to determine the optimal tradeoff between base and enhancement layer bandwidth. The table gives the name of the sequence, the spatial resolution, number of frames, and a brief description of the video content.

4.2.1 Car

The first frame of the Car sequence is shown in figure 4.3. This video depicts a remote control car spinning on a tabletop with miniature soccer balls bouncing around it. The background is stationary.

The results of the tradeoff experiment are shown in figure 4.4. As expected, the base layer PSNR (white circles) improves continuously with increasing base layer allocation. The enhancement layer PSNR is shown in black. When adaptive block sizes are used (black circles), the enhancement layer PSNR reaches a peak at 91% base layer allocation (point A). As the base layer allocation approaches 100%, the remaining bandwidth is too small for an enhancement layer. The result for fixed, 16x16 pixel blocks is indicated by the black square (point B) at 96% base layer allocation. Notice that the overhead needed to transmit partitioning information for adaptive block sizes, about 0.01 bpp (or 3.25% of 0.32 bpp), is significant in this context. The black diamond (point C) indicates the best nonadaptive format conversion (NFC) method in which 100% of the bandwidth is allocated to the base layer. For this sequence, the best NFC is Martinez-Lim deinterlacing. The dashed line at the level of the black diamond is provided as a reference—for the car video sequence, the PSNR of the AFC enhancement layer video surpasses the NFC baseline.

From the figure, it seems that the optimal tradeoff between base and enhancement layer bandwidth occurs when fixed, 16x16 pixel blocks are used. At point B, the enhancement layer PSNR is 39.5 dB, which is 0.21 dB higher than the best NFC at point C. However, the improvement in enhancement layer quality comes at a cost in base layer PSNR—the difference between points D and E in the figure is 0.15 dB. Points A and B have almost the same PSNR, but for the optimal tradeoff we choose point B, which minimizes the corresponding base layer degradation.

Visual inspection reveals that the background region at point B is more stationary than at point C, where certain pixels tend to flutter between frames. Motion masks small fluctuations, so there is no improvement around the moving car and soccer balls. A similar comparison between the base layer video at points D and E shows no noticeable difference between these two video sequences. As a result, it appears that the AFC enhanced video is, in fact, a better quality reconstruction of the original video sequence, and that the cost in base layer quality is insignificant. The improvement can be attributed to the adaptive format conversion selecting inter-frame deinterlacing techniques in the stationary background regions, and intra-frame deinterlacing techniques where there is motion between frames.

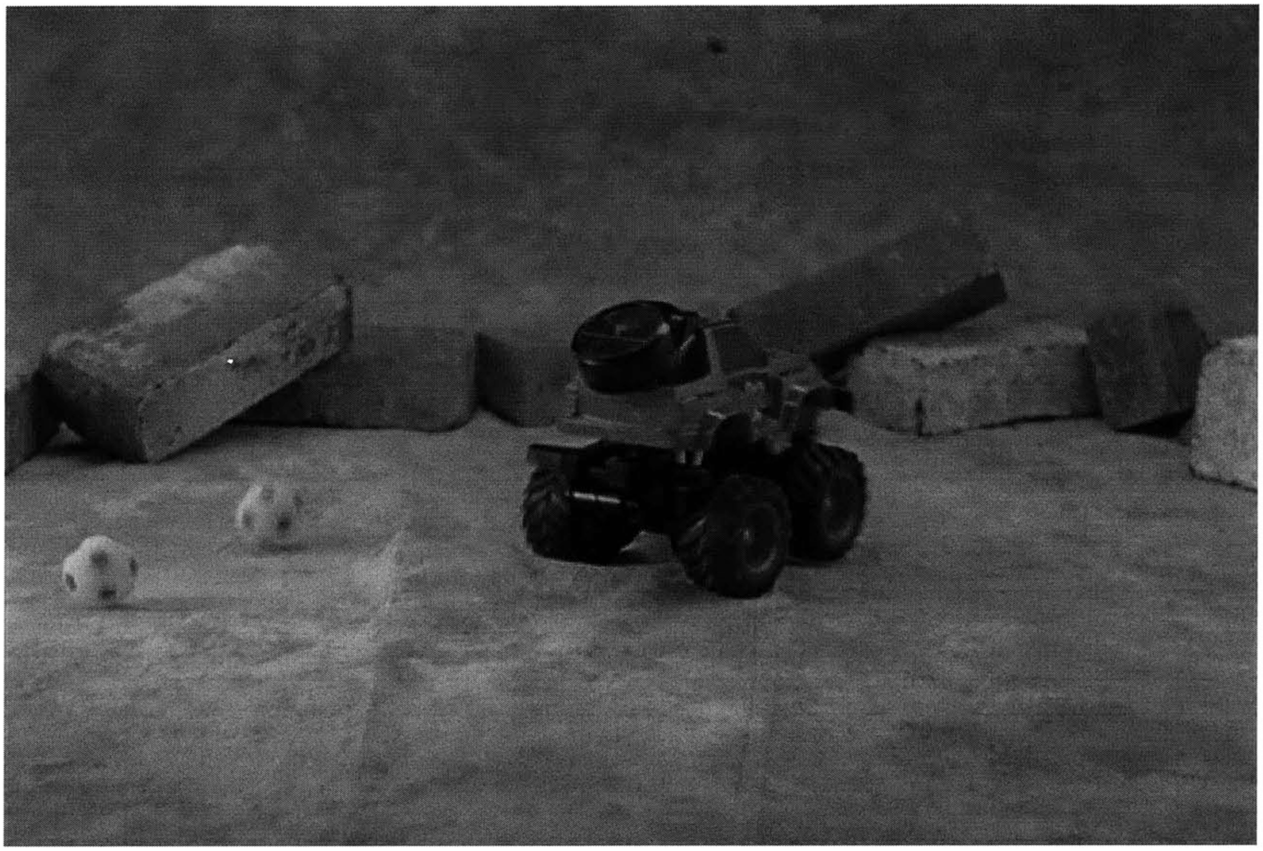


Figure 4.3: First Frame of the Car Sequence

In the Car sequence, a remote control car spins while miniature soccer balls bounce on a tabletop. The background is stationary.

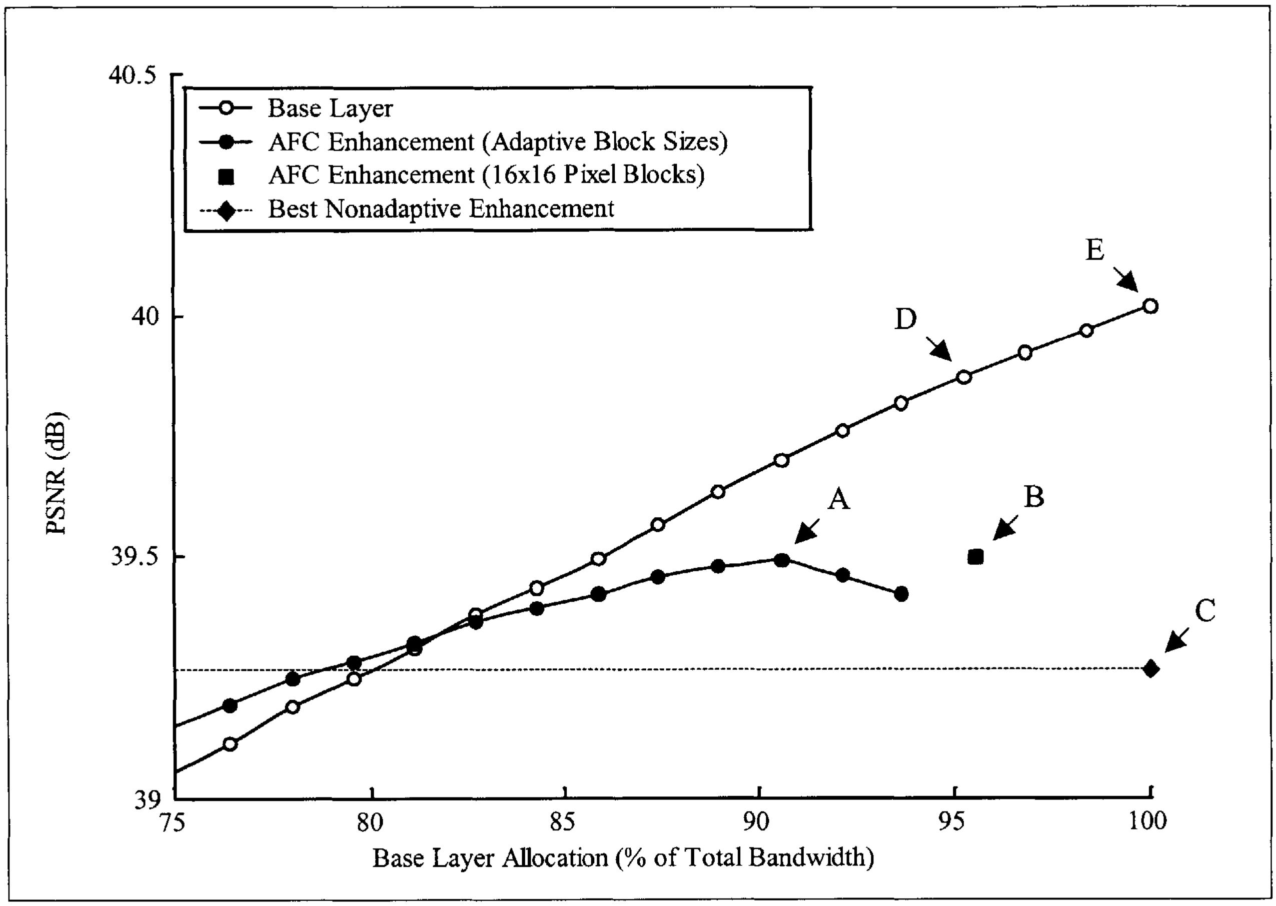


Figure 4.4: PSNR Versus Base Layer Allocation for the Car Sequence

The optimal tradeoff between base and enhancement layer bandwidth occurs when fixed, 16x16 pixel blocks are used and 96% of the total bandwidth is allocated to the base layer. At point B, the background is more stationary than at point C, which represents Martinez-Lim deinterlacing and 100% base layer allocation. The difference in PSNR between points B and C is 0.21 dB. On the other hand, there is no visible difference between the base layer video at points D and E, even though the difference in PSNR is 0.15 dB. Points A and B are nearly identical in quality, but point B corresponds to less distortion in the base layer.

4.2.2 Football

Figure 4.5 shows the first frame of the Football sequence. In this scene, the crowd in the background pans slowly to the left while the football players move quickly in different directions. The original image appears somewhat grainy.

A graph of PSNR versus base layer allocation is given in figure 4.6. Like before, the base layer PSNR (white circles) increases with increasing base layer bandwidth. For adaptive block sizes (black circles), the enhancement layer PSNR peaks at about 89% base layer allocation (point A). For fixed, 16x16 pixel blocks (black square), the base layer allocation is 96%. The best nonadaptive format conversion (black diamond) is linear interpolation.

Notice that the values of PSNR for this sequence are in the range of 33 dB, compared to 39-40 dB for the Car sequence. More complicated motion and detail make this video sequence more difficult to compress.

For the Football video sequence, it seems that the optimal tradeoff is again using fixed block sizes. However, the peak enhancement layer PSNR at point B is only 0.04 dB above the best nonadaptive conversion at point C. This improvement is offset by a 0.07 dB decrease in base layer quality from point E to point D.

By visual inspection, the background region at point B appears slightly better than at point C. Although the background is not stationary as it was in the Car sequence, any non-uniform motion is easy to detect. The football players appear approximately the same at both points. In contrast, the football players at point A (adaptive block sizes) show more disturbing block effects than at points B and C, which may be attributed to an increase in the base layer distortion. Once again, there is no visible difference between the base layer video sequences at points D and E. This examination confirms what is suggested in the graph, that AFC enhancement using fixed, 16x16 pixel blocks is the best choice.

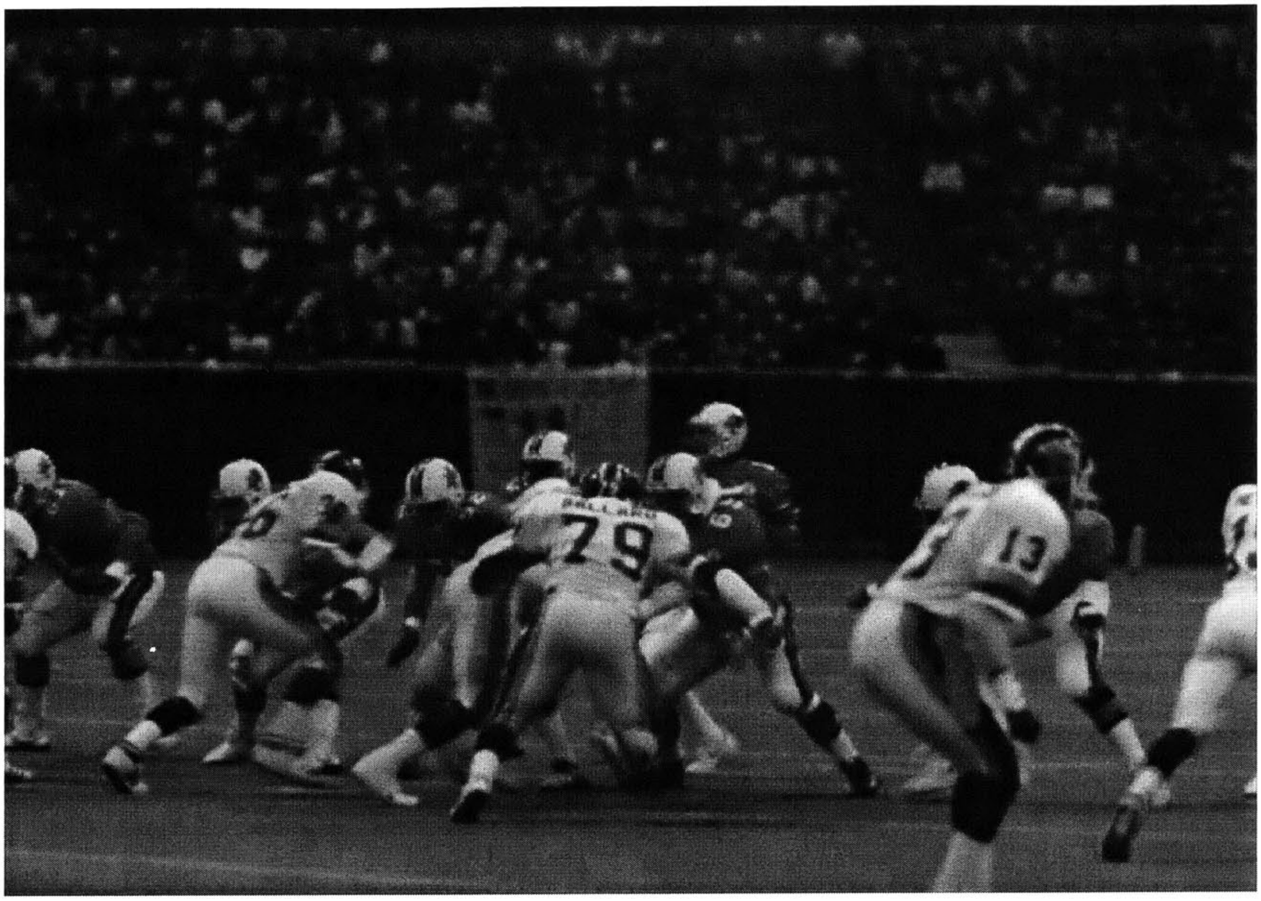


Figure 4.5: First Frame of the Football Sequence

The football players move quickly in different directions while the crowd in the background pans slowly to the left.

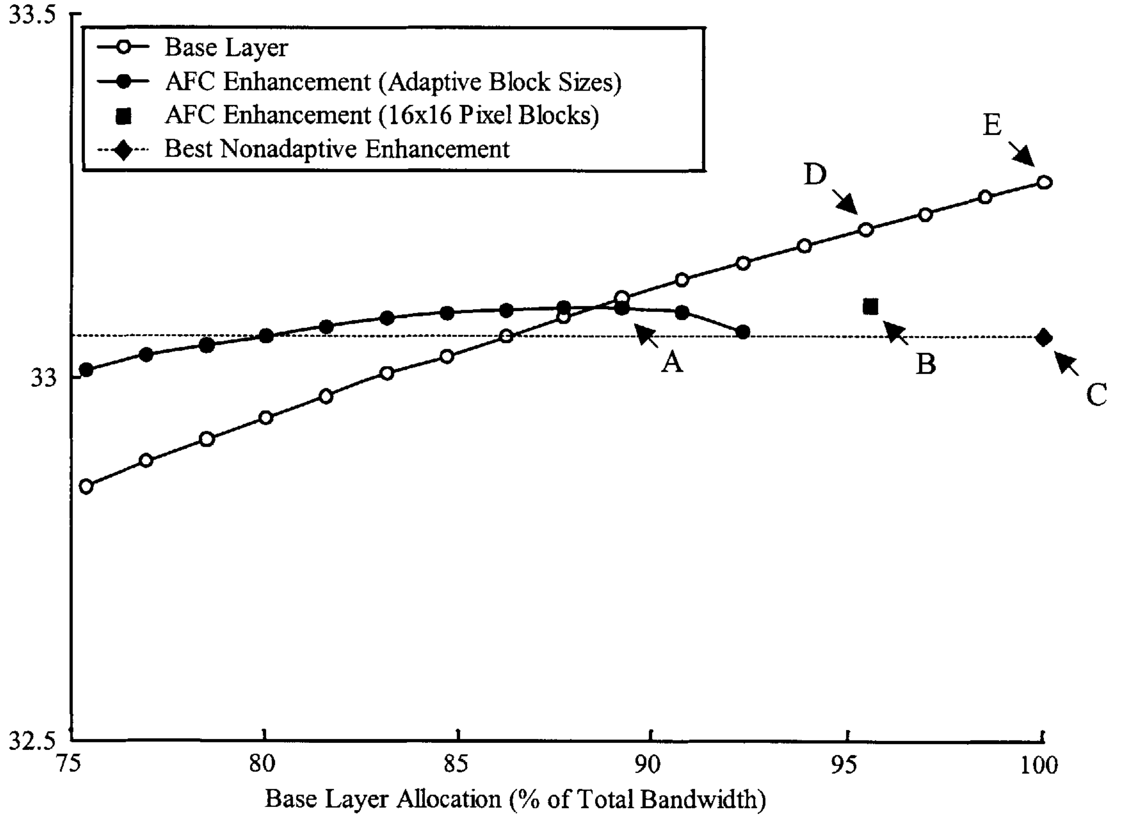


Figure 4.6: PSNR Versus Base Layer Allocation for the Football Sequence

In this graph, the optimal tradeoff between base and enhancement layer bandwidth appears to be at 96% base layer, 4% enhancement layer using fixed block sizes. However, the overall video quality is quite poor. A visual comparison of points B and C shows that AFC improves some areas like the background but does not improve the football players, which are distorted by block effects. Points B and C differ by only 0.04 dB. Even though points A and B share the same PSNR, the video at A visibly reflects the increase in base layer distortion. The difference in PSNR in the base layer video is 0.07 dB from point D to point E.

4.2.3 Football Still

In order to see how the results change when there is no motion, the Football Still sequence was created by repeating the first frame of the Football sequence (figure 4.5). The outcome of this test is provided in figure 4.7, which shows the PSNR of the AFC enhanced video (black) and the base layer video (white) increasing continuously with greater base layer allocation. Unlike the previous two cases, however, the best nonadaptive format conversion (black diamond) results in the highest enhancement layer PSNR. The best nonadaptive deinterlacing technique is either FFR or BFR (they are the same). Adaptive format conversion also uses FFR or BFR exclusively, so the information in the enhancement layer is not useful and only takes bits away from the base layer.

The range of PSNR for the Football Still sequence is between 35.5 and 36.5 dB—considerably higher than for the moving Football sequence. When encoding a still sequence, MPEG is able to allocate more bits to the I-pictures and very few to the P- and B-pictures, thus improving the picture quality.

This result suggests a practical approach in which the AFC enhancement layer could be "turned off" and the best nonadaptive method used instead. An encoder would create two alternative base layer bitstreams: one that uses the total available bandwidth and another that preserves a fraction of the bandwidth for the AFC enhancement layer. Then, by comparing the enhanced resolution video, the encoder would determine which option to use. The decision would be indicated in the encoded enhancement layer by a flag (one bit), followed by a number indicating the best nonadaptive method in one case or by the AFC enhancement information in the other. Because of the coding dependence between frames in an MPEG bitstream, the decision to turn the enhancement layer on or off would be made most easily on the video sequence or perhaps even the GOP level. Instead of turning the enhancement layer off completely, the best format conversion method could be selected on a frame-by-frame basis using negligible bandwidth. The complexity introduced in this approach is contained almost entirely in the encoder, and would not increase the complexity (cost) of the decoder.

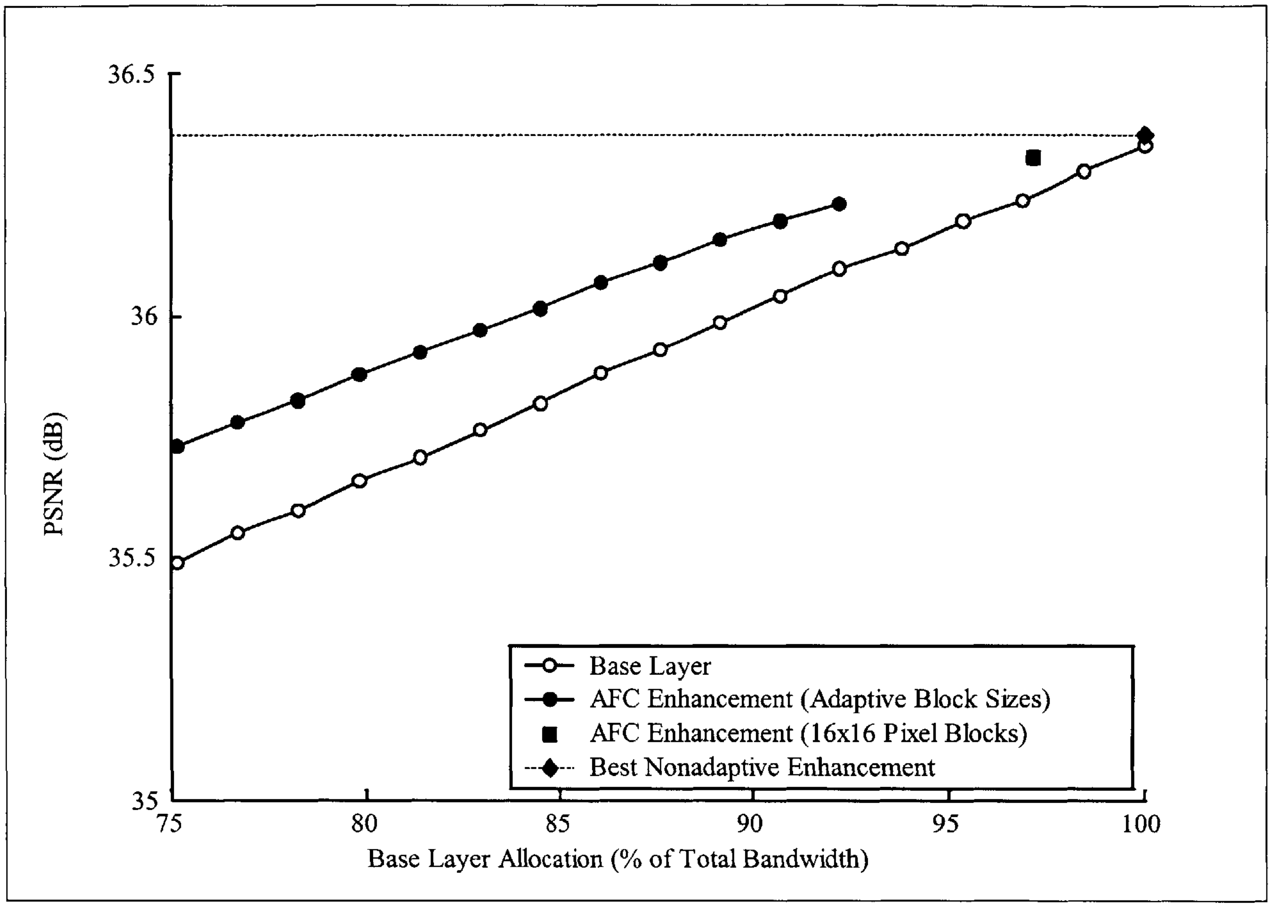


Figure 4.7: PSNR Versus Base Layer Allocation for the Football Still Sequence.

The PSNR of the AFC enhanced video and base layer video both improve continuously with increasing base layer allocation. In this case, the best nonadaptive format conversion is either FFR or BFR. Because there is no motion, adaptive format conversion also chooses FFR or BFR exclusively, so nothing is gained when bandwidth is allocated to the enhancement layer.

4.2.4 Marcie

The first frame of the Marcie sequence is illustrated in figure 4.8. The focus is on Marcie's head, which tilts slowly to the left as the video progresses. The hand in the foreground moves upward as sand trickles through the fingers, and a small breeze blows the leaves in the background.

The results of the tradeoff experiment are given in Figure 4.9. From the graph, it appears that the optimal tradeoff occurs when adaptive block sizes are used and the base layer receives 92% of the total bandwidth. The PSNR of the AFC enhanced video at point A is 0.06 dB higher than point B and 0.12 dB higher than the best NFC enhanced video (Martinez-Lim deinterlacing) at point C. For fixed block sizes, the base layer allocation is 97%. The PSNR of the base layer decreases by 0.31 dB from point D to point E.

Visually, the AFC enhanced video at point A is approximately the same as point B, and both are slightly better than C in the background region. The movement of the leaves masks much of the inter-frame flickering that would otherwise appear more prominently in the background. At points A, B, and C, the images are sharp and nearly identical in the foreground. The quality of the base layer video is excellent at points D and E, and leads to the high quality enhancement layer video. Because the base layer quality is good overall and point A has the highest PSNR, we say the optimal tradeoff is at point A for the Marcie video sequence.

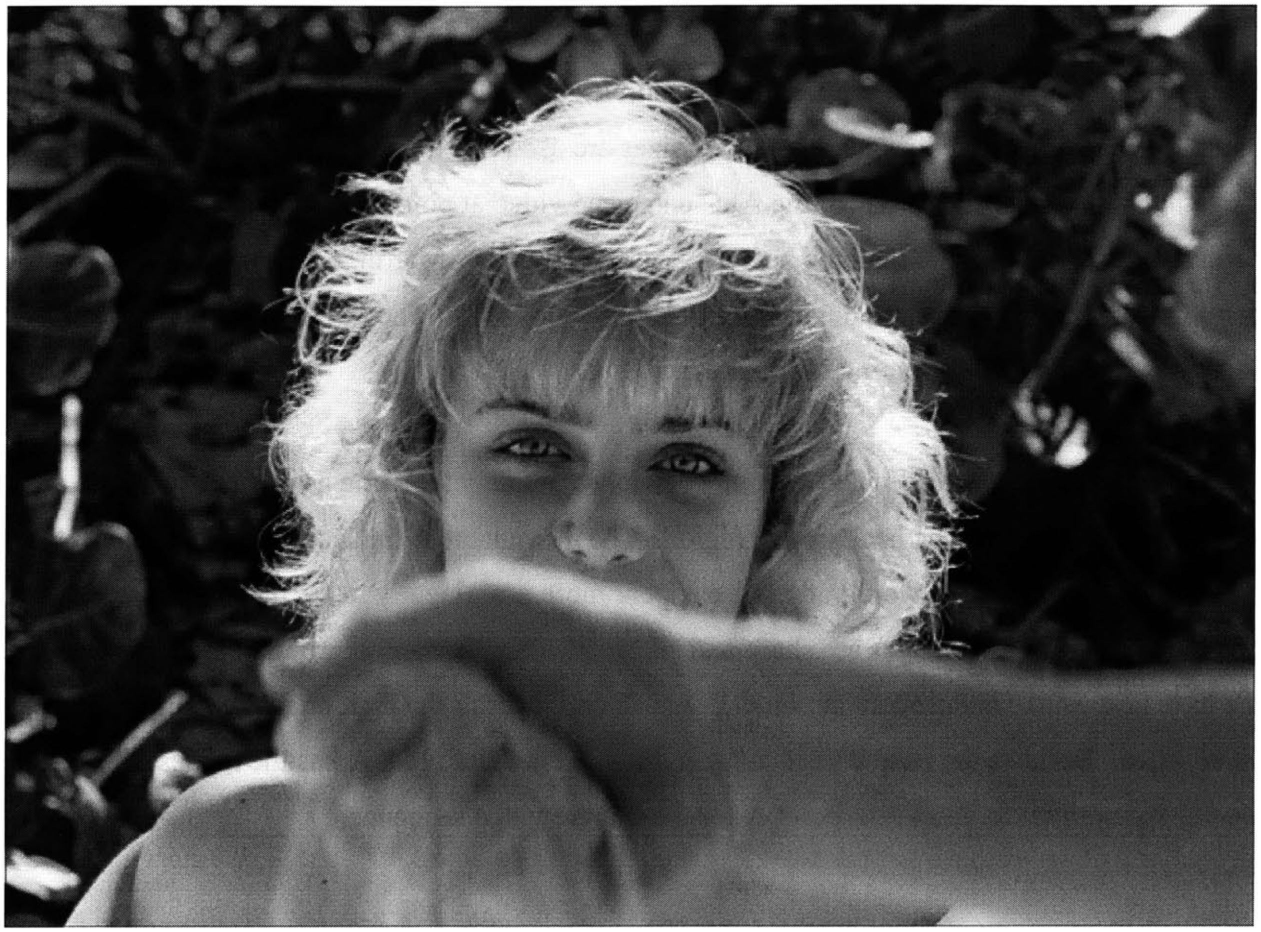


Figure 4.8: First Frame of the Marcie Sequence

The hand in the foreground moves upward as sand trickles through the fingers, Marcie's head tilts slowly to the left, and a breeze moves the leaves in the background.

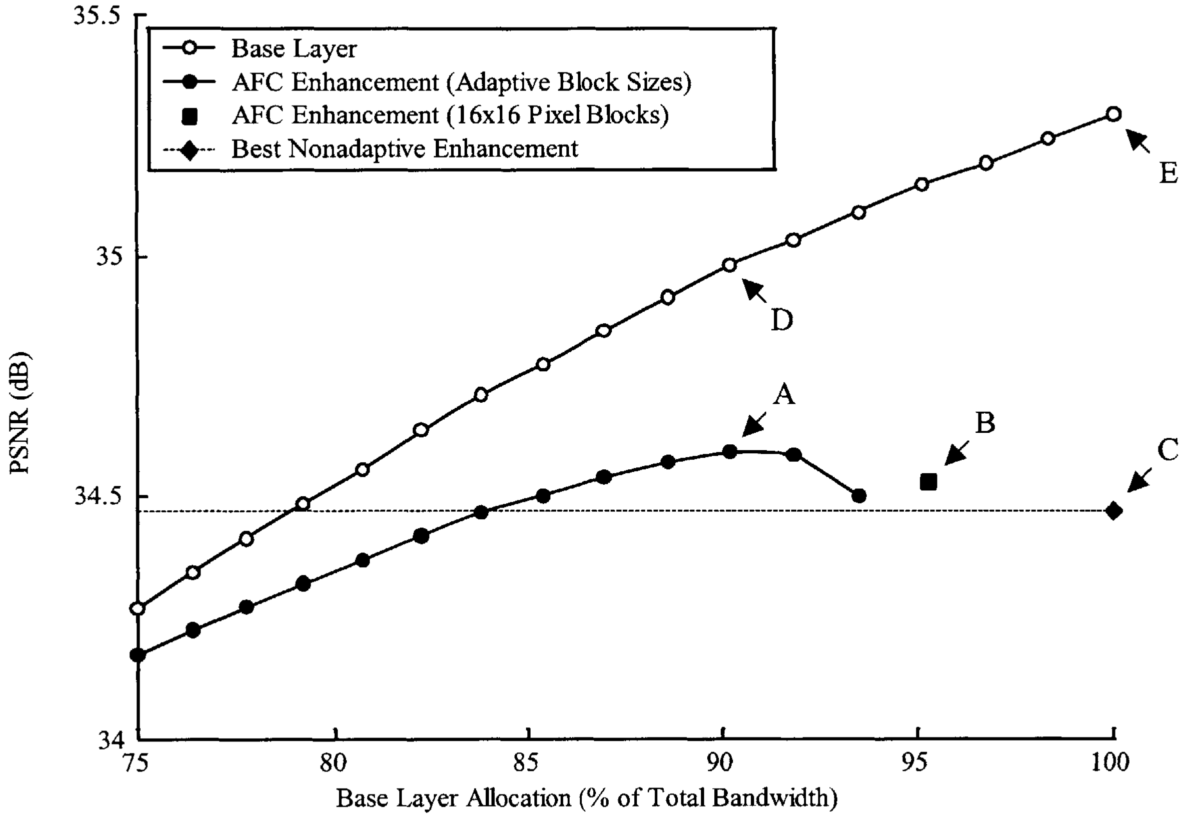


Figure 4.9: PSNR Versus Base Layer Allocation for the Marcie Sequence

In this graph, the best tradeoff occurs at 90% base layer, 10% enhancement layer allocation using adaptive block sizes. The best nonadaptive technique is Martinez-Lim deinterlacing. The quality of the compressed base layer is high overall and leads to good quality enhanced resolution video. The difference in PSNR between points A and C is 0.12 dB, between points B and C it is 0.06 dB, and between D and E it is 0.31 dB. Visually, the AFC enhanced video at points A and B are about the same and are slightly better than the NFC video (point C) in the background regions; however, the motion of the leaves hides much of the inter-frame flickering that would otherwise be more noticeable. Points A and B are both good candidates for the optimal tradeoff point, but A has a higher PSNR.

4.2.5 Girl

The first frame of the Girl sequence is shown in figure 4.10. Occupying most of the frame is a still picture of a girl and surrounding objects. During the first 30 frames, this image pans slowly to the right; during the last 30 frames, the camera zooms in on the wine bottle on the table. The smaller picture in the top, left corner is a moving video sequence of a woman walking. The other objects, "Super 8" and "high-resolution," are test graphics that spin and scroll, respectively. This video sequence contains a large amount of detail that is difficult to compress with high quality.

Figure 4.11 gives the result of the tradeoff experiment. For this video, the output PSNR is relatively low—between 25 and 27 dB. The best AFC enhanced video (point A) occurs when adaptive block sizes are used and the base layer is given 87% of the total bandwidth. The PSNR at point A is 0.91 dB greater than at point C, which corresponds to nonadaptive, linear interpolation. The difference in PSNR for the base layer is 0.65 dB between the points labeled D and E. Point B, at 96% base layer allocation, is 0.68 dB above the NFC baseline.

Visual inspection reveals that the AFC enhanced video at both A and B are significantly better than the NFC enhanced video. Because of the high amount of detail, the linear-interpolated video at point C suffers from a large amount of flickering between frames and from aliasing. Even though distortion is still visible, adaptive format conversion eliminates a great deal of these effects, with adaptive block sizes being slightly more effective than fixed block sizes. The improvement in flickering can only be seen when the video is played back and cannot be determined by looking at a single frame. However, the elimination of aliasing can be seen by looking at the "Super 8" test graphic shown in figure 4.12. The original video is provided in panel (a) as a reference. The snapshot in panel (b) is from point A (adaptive block sizes). By taking advantage of both inter and intra-frame deinterlacing techniques, the AFC enhanced video in panel (b) is closer to the original than the linear-interpolated video in panel (c), which suffers from aliasing.

The base layer is noticeably degraded at points D and E, with little visible difference between the two. Because the AFC enhancement layer results in a substantial improvement, a small decrease in base layer quality is an acceptable tradeoff for the Girl video sequence.

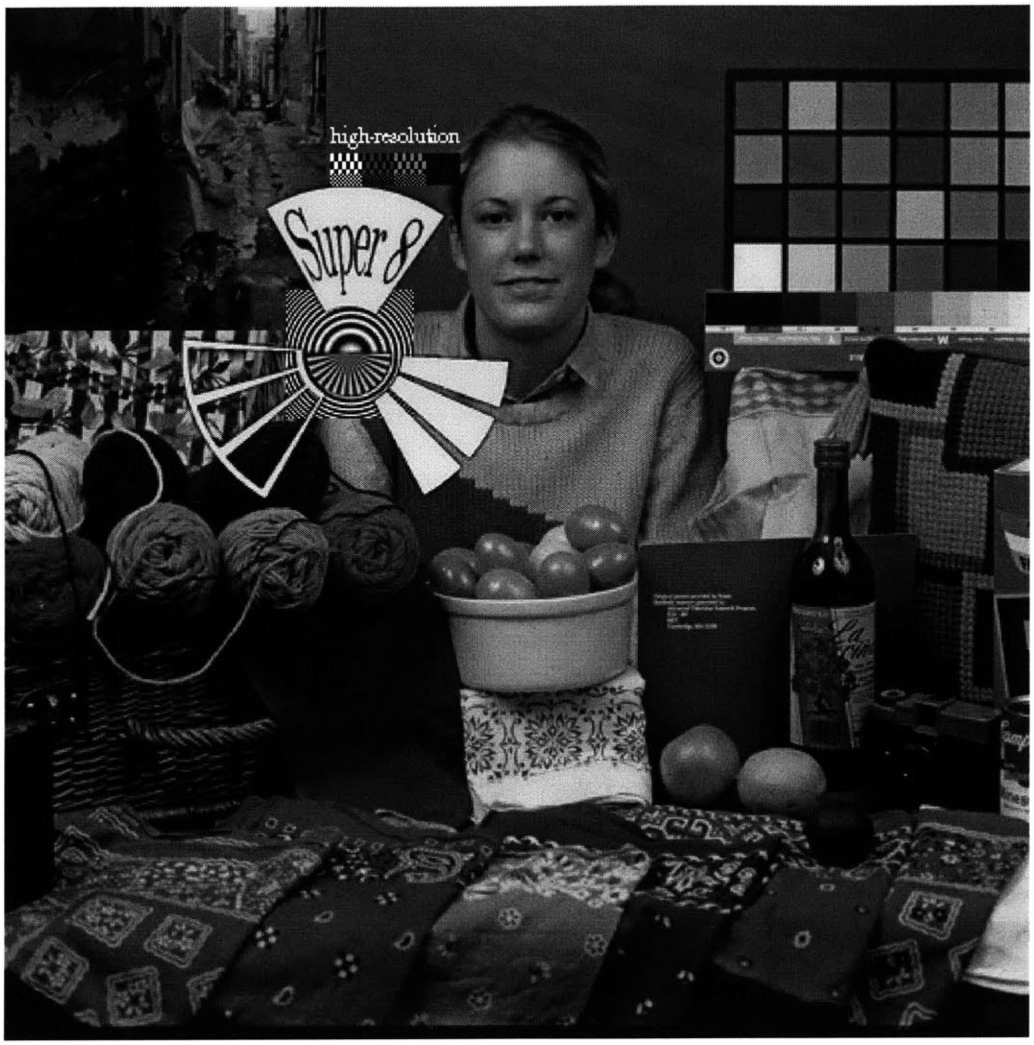


Figure 4.10: First Frame of the Girl Sequence

A still picture of a girl and surrounding objects occupies the majority of this video sequence. The picture pans slowly to the right for the first 30 frames, then zooms in on the wine bottle for the last 30 frames. The small picture in the top left corner is a moving video sequence of a woman walking. The "Super 8" and "high-resolution" test graphics spin and scroll, respectively, during playback. The detail in this video sequence makes it difficult to compress with high quality.

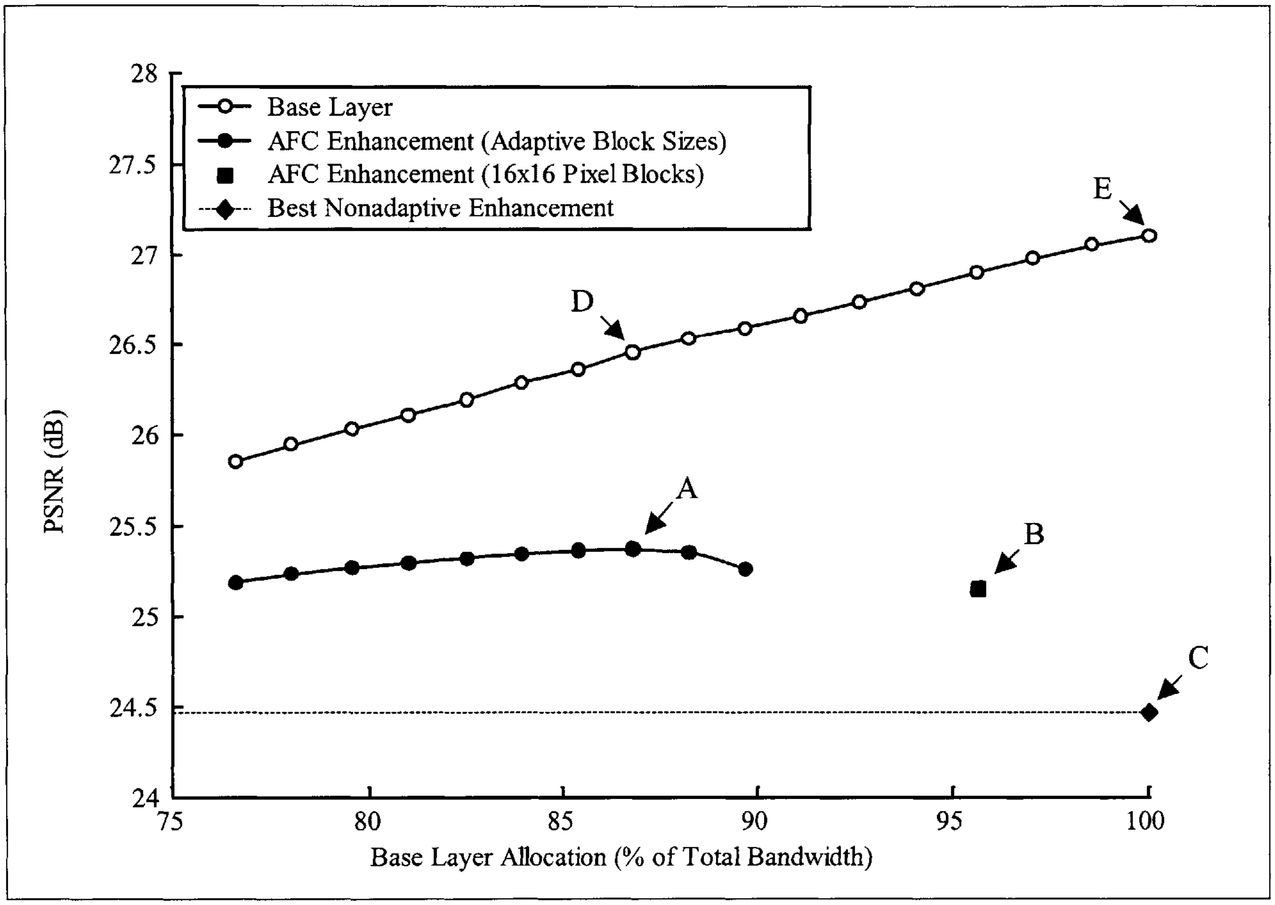
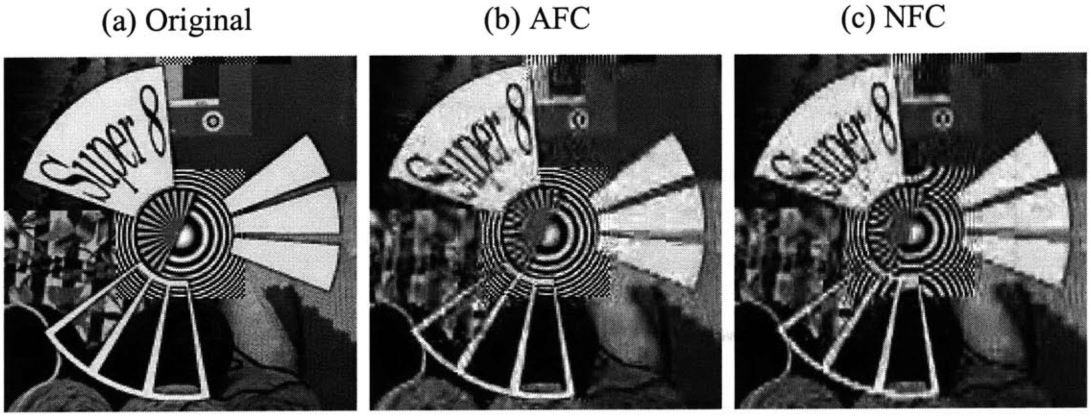


Figure 4.11: PSNR Versus Base Layer Allocation for the Girl Sequence

The optimal tradeoff for the Girl sequence occurs at 87% base layer allocation using adaptive block sizes. The PSNR of the AFC enhanced video at point A is 0.91 dB greater than the linear-interpolated video at point C. With fixed block sizes (point B), the base layer allocation is 96% and the PSNR is 0.68 dB higher than point C. The difference in base layer PSNR is 0.65 dB between points D and E. Visually, the video sequences at points A and B do not suffer as much from aliasing and flickering effects that are common at point C, and adaptive block sizes are slightly more effective in eliminating the distortion. The base layer video at points D and E are visibly equal.


(a) Original

(b) AFC

(c) NFC

Figure 4.12: Aliasing in the Girl Sequence

The AFC and NFC enhanced video in panels (b) and (c) are compared to the original video sequence in panel (a). Compression and linear interpolation introduce aliasing in the NFC video that is largely eliminated by adaptive format conversion.

4.2.6 Toy Train

The Toy Train video sequence features a passing train and some moving toys (figure 4.13). As the video is played, the background and many other objects are stationary.

A plot of PSNR versus base layer allocation in figure 4.14 exhibits familiar characteristics. At point A, the PSNR of the adaptive block size, AFC enhanced video reaches a peak at 89% base layer allocation that is 0.58 dB greater than the best nonadaptive format conversion (Martinez-Lim deinterlacing) at point C. Point B, which corresponds to fixed block sizes, is at 95% base layer allocation and is 0.57 dB greater than point C. The base layer PSNR experiences a loss of 0.26 dB from point E to point D.

Examining the video sequences reveals that the AFC enhanced video at point B is superior to the NFC video at point C *and* the AFC video that uses adaptive block sizes at point A. The most noticeable improvement is in stationary regions like the background and in slow moving objects. Once again, adaptive format conversion eliminates much of the flickering noise that is introduced by intra-frame deinterlacing. A particularly good example of the improvement is the word "DISNEYLAND" on the side of one of the toys. In the original and AFC enhanced video the word is stationary, but in the NFC enhanced video it wiggles from frame to frame. The video at point A reflects increased distortion in the underlying base layer video and is visibly inferior to the video at point B, even though the PSNR is slightly higher.

The base layer quality is clearly reduced from the original at points D and E, but it is difficult to tell if one video sequence is worse than the other. As a result, a small increase in base layer distortion is tolerable when weighed against the larger improvement in the AFC enhancement layer.

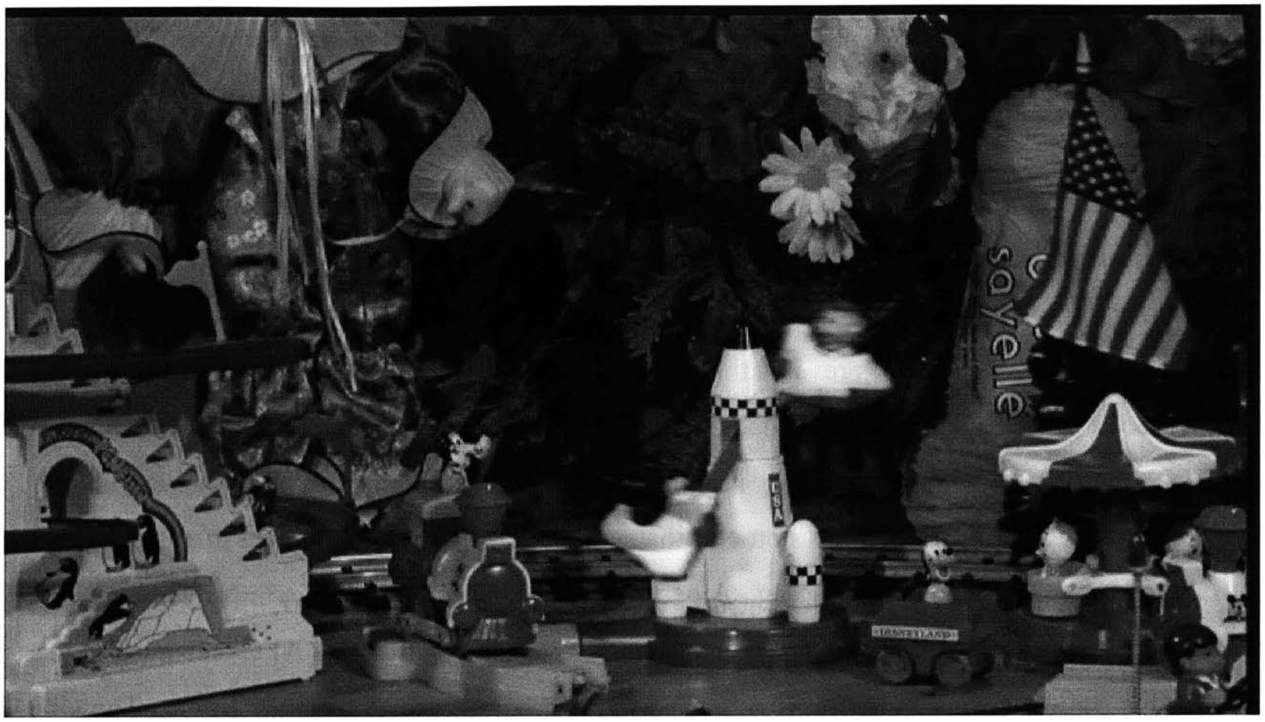


Figure 4.13: First Frame of the Toy Train Sequence

A passing train and moving toys characterize the Toy Train sequence. The background and some of the other objects are stationary.

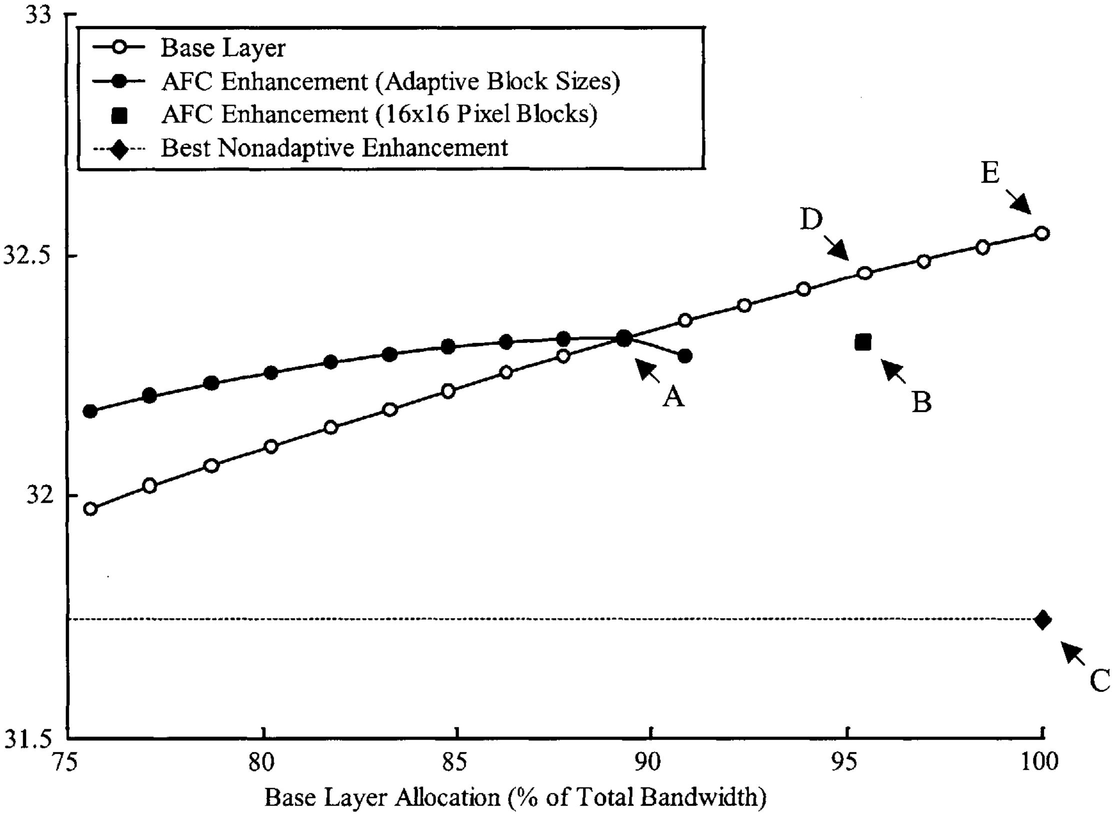


Figure 4.14: PSNR Versus Base Layer Allocation for the Toy Train Sequence

At point B, the tradeoff between base and enhancement layer bandwidth is optimal. Point B is at 95% base layer allocation and is 0.57 dB higher than point C, which represents nonadaptive, Martinez-Lim deinterlacing. The difference in base layer PSNR between points D and E is 0.12 dB. Visual inspection confirms that a little loss in the base layer quality leads to a large improvement in the enhancement layer. Using adaptive block sizes actually yields the highest PSNR, but the video at point A reflects an increase in base layer distortion and is visibly inferior to the video at point B.

4.2.7 Tulips Scroll

Tulips Scroll is a still picture (shown in figure 4.15) that moves to the left at a rate of one pixel per frame. The test graphic in the bottom, left corner is stationary—i.e. it does not move with the rest of the picture or have internal motion.

Results of the tradeoff experiment for the Tulips Scroll sequence are shown in figure 4.16. The PSNR of the AFC enhanced video at point A is 0.83 dB higher than point C (linear interpolation). The corresponding change in base layer PSNR is 1.44 dB between points D and E. This tradeoff occurs when the base layer is allocated approximately 85% of the total bandwidth. Point B is at 96% base layer allocation and is 0.61 dB higher than point C.

Despite the relatively large depreciation in base layer PSNR, the video at points D are E are approximately the same, visually, and exhibit the same overall sharpness as the original picture. Much of the distortion is hidden in the natural textures of the road, grass, flowers, and trees. In the enhancement layer, the video at point A is noticeably better than at point B, which is itself better than C. Both B and C exhibit flickering between frames, especially near horizontal lines on the buildings and other man-made objects. It is clear that for this video sequence, 16x16 pixel partitions help improve the quality, but that the picture is too coarsely divided. Furthermore, in the nonadaptive video, diagonal lines on the road, benches, and buildings look stair-stepped instead of straight, and closely spaced lines on the buildings and test graphic are aliased. At points A and B, adaptive deinterlacing eliminates most of these artifacts and produces a sharper image than linear interpolation alone. Focusing on the edge of the curb in figure 4.17 illustrates how a combination of deinterlacing techniques produces a straight line rather than a jagged edge. The three panels in the figure compare a snapshot from the original video , AFC video using adaptive block sizes, and NFC video.

For the Tulips Scroll video sequence, the optimal tradeoff is clearly in favor of the AFC enhancement layer, with adaptive block sizes yielding better results than fixed block sizes.

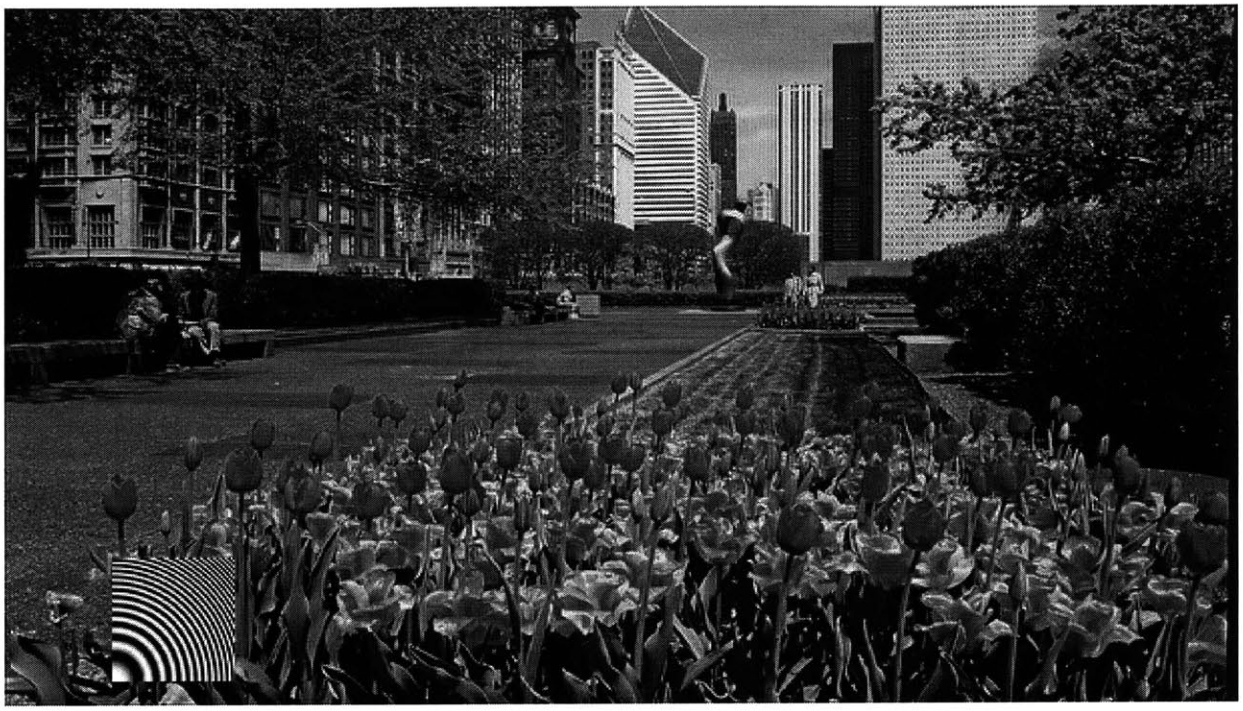


Figure 4.15: First Frame of the Tulips Scroll Sequence

Tulips Scroll is a still picture that moves one pixel to the left in every frame. The test graphic in the bottom, left corner is stationary.

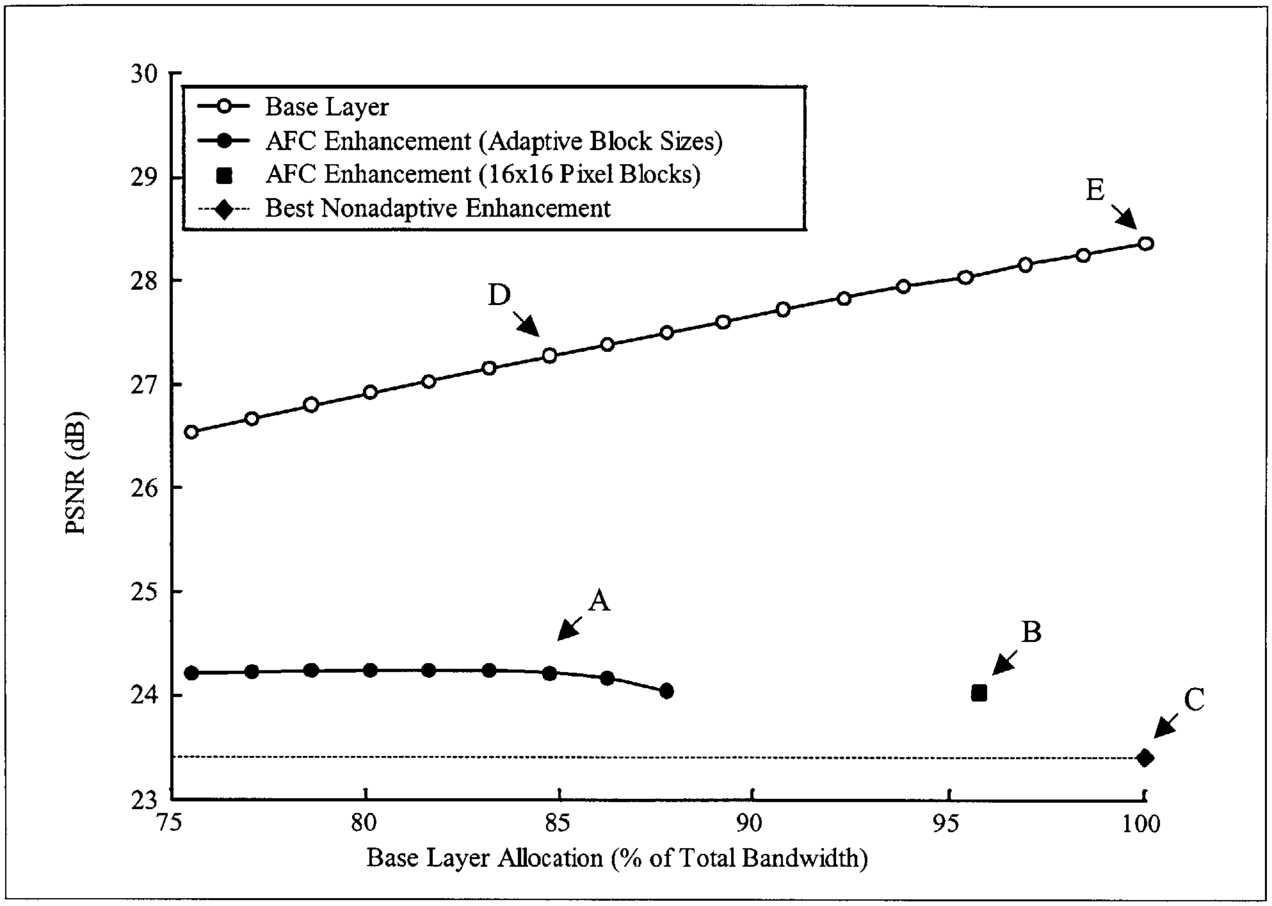


Figure 4.16: PSNR Versus Base Layer Allocation for the Tulips Scroll Sequence

For the Tulips Scroll sequence, the optimal tradeoff occurs at about 85% base layer allocation. The PSNR at point A is 0.83 dB higher than at point C (linear interpolation), and visual inspection confirms that the AFC enhanced video is significantly better than the nonadaptive enhancement. Even though the difference between points D and E is 1.44 dB, the difference in base layer quality is indistinguishable. Fixed block size AFC enhancement at point B (96% base layer allocation) appears to be too coarsely partitioned to eliminate as much flickering as point A.

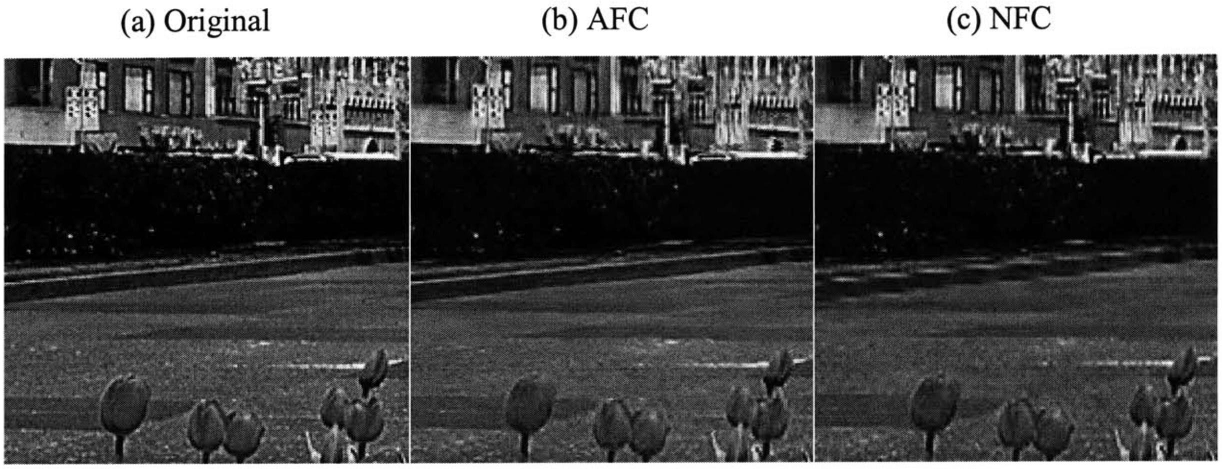


Figure 4.17: Stair-step Discontinuities in the Tulips Scroll Sequence

Linear interpolation creates stair-step discontinuities in diagonal lines, illustrated by the edge of the curb in panel (c). These lines appear straight in the original and AFC enhanced video.

4.2.8 Tulips Zoom

The Tulips Zoom sequence uses the same still picture as Tulips Scroll but starts from a wide angle and slowly zooms in on the area around the statue and buildings as the video progresses.

Figure 4.18 shows that for this sequence, the best AFC enhanced video occurs at 96% base layer allocation (point B) using fixed block sizes. The PSNR at point B is equal to that of the best nonadaptive format conversion at point C (linear interpolation). Point A, using adaptive block sizes, has a lower PSNR than the best NFC enhancement. The base layer PSNR decreases by 0.26 dB from point E to point D.

Comparing points B and C, the AFC video looks better than the NFC video. In the NFC video, the eye is drawn to unnatural movement and flickering where it expects a continuous zoom. Though not eliminated, these effects are reduced by the AFC enhancement layer. The NFC video also suffers from the same stair-step discontinuities that were seen in the Tulips Scroll sequence, and these, too, are removed by adaptive format conversion.

The Tulips Zoom sequence is a case where the measured PSNR does not adequately describe the improvement provided by adaptive format conversion. Trading about 4% of the base layer bandwidth for a small AFC enhancement layer is, in fact, worthwhile.

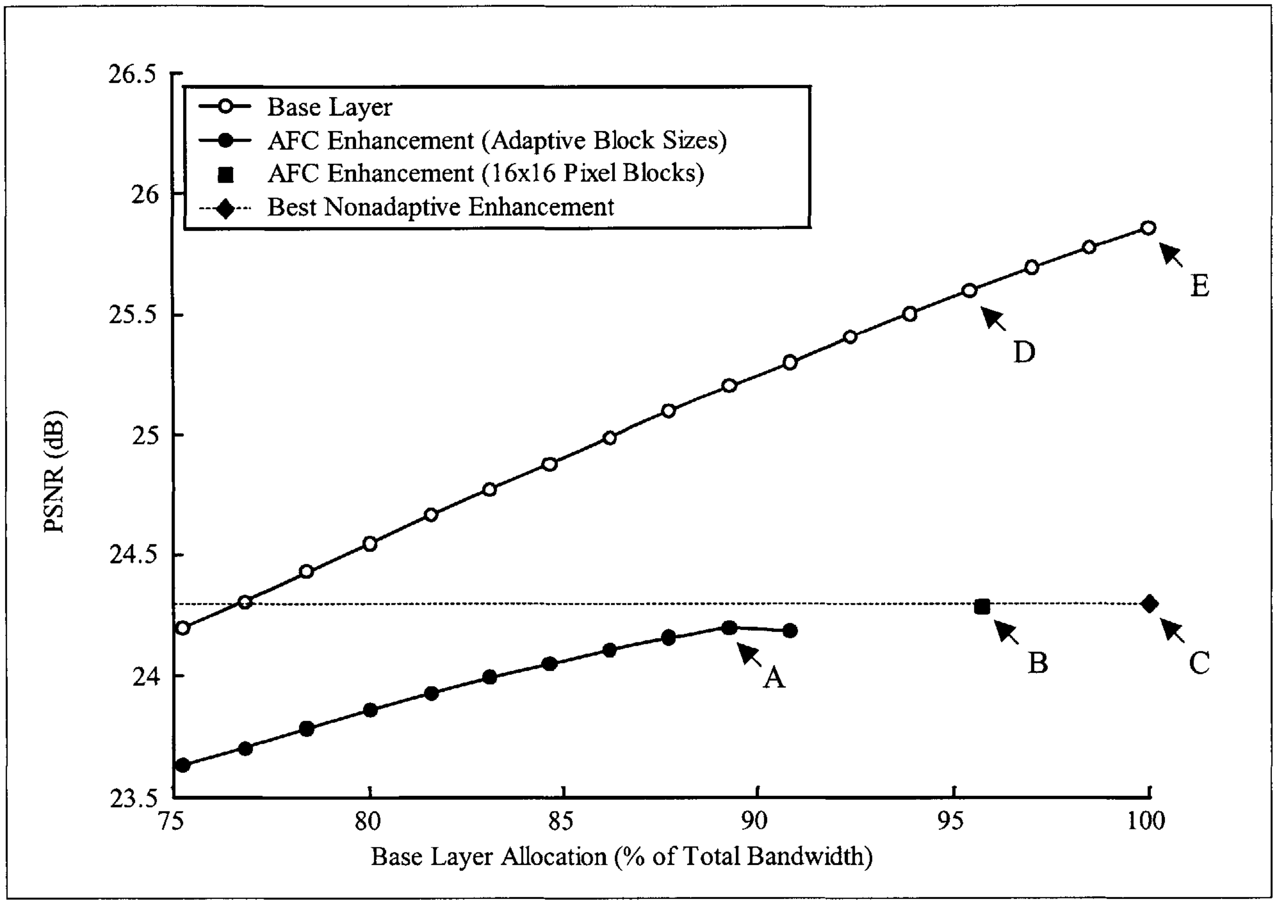


Figure 4.18: PSNR Versus Base Layer Allocation for the Tulips Zoom Sequence

Although the PSNR of the AFC enhanced video at point B is equal to the PSNR of the best nonadaptive conversion, the video at point B looks better than at point C. The change in PSNR between points D and E is 0.26 dB. For the Tulips Zoom sequence, the optimal tradeoff between base and enhancement layer bandwidth is 96% base layer, 4% enhancement layer.

4.2.9 Picnic

The Picnic sequence (figure 4.19) is another still image that slowly zooms in. At the end of the video sequence, the camera is focused on the faces of the man and woman standing at the right side of the picture. A large amount of detail combined with non-translational motion make this video sequence difficult to compress.

In figure 4.20, the PSNR of the AFC enhanced video improves continuously with increasing base layer allocation, but at its highest point (B) it is still 0.07 dB below the PSNR of the NFC enhanced video (point C). The corresponding points in the base layer (D and E) differ by 0.2 dB. The best nonadaptive conversion technique is Martinez-Lim deinterlacing.

Visual inspection reveals that the base layer video sequences at D and E are severely distorted by block effects and that the enhancement layer video at points B and C are relatively equal in quality. Adaptive format conversion does improve certain parts of the picture, but other areas appear to be worse. This is an example of how a poorly encoded base layer neutralizes the positive effect of AFC enhancement. In this situation, allocating the total bandwidth to the base layer is perhaps the best alternative.

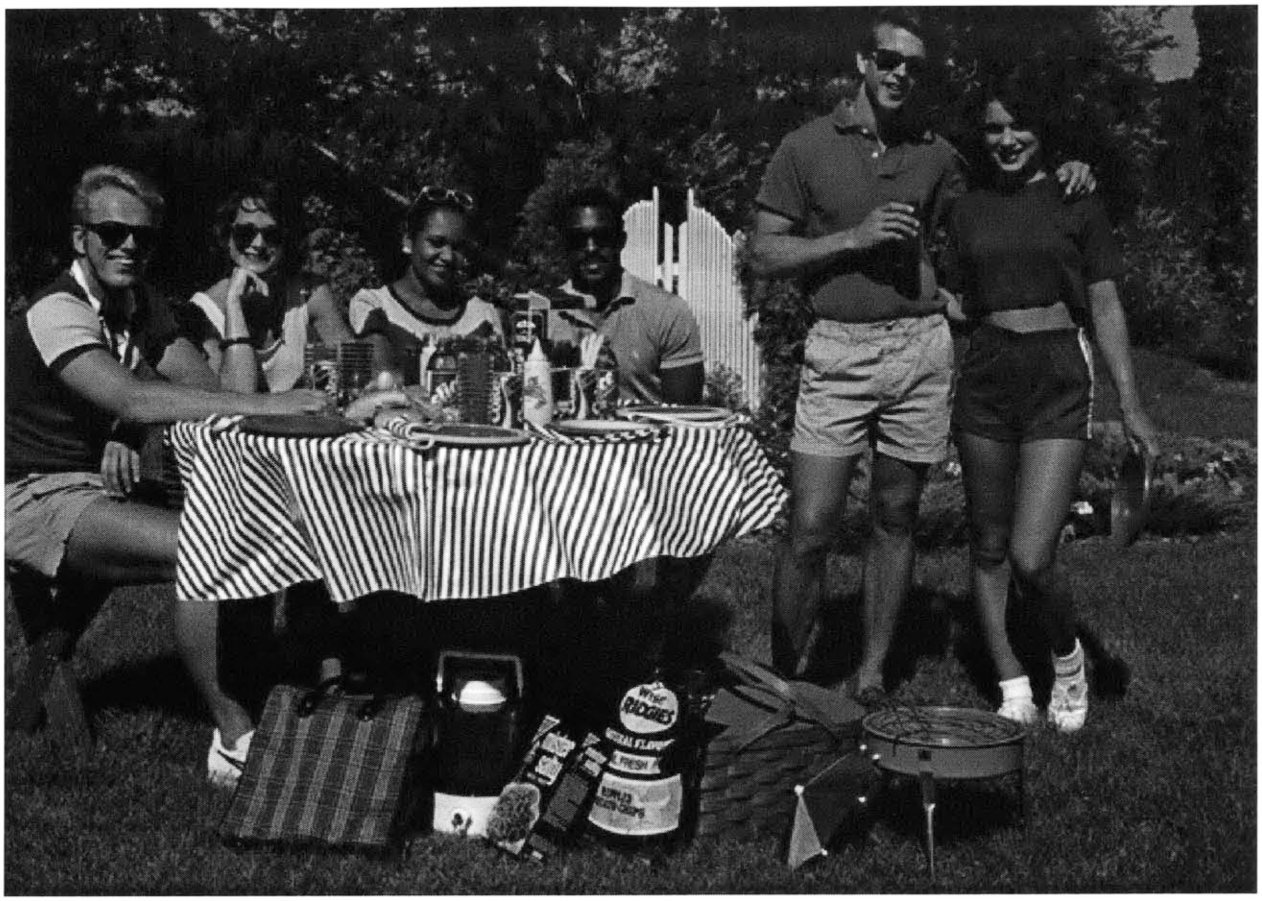


Figure 4.19: First Frame of the Picnic Sequence

The Picnic sequence is a still picture where the camera slowly zooms in on the faces of the couple standing on the right.

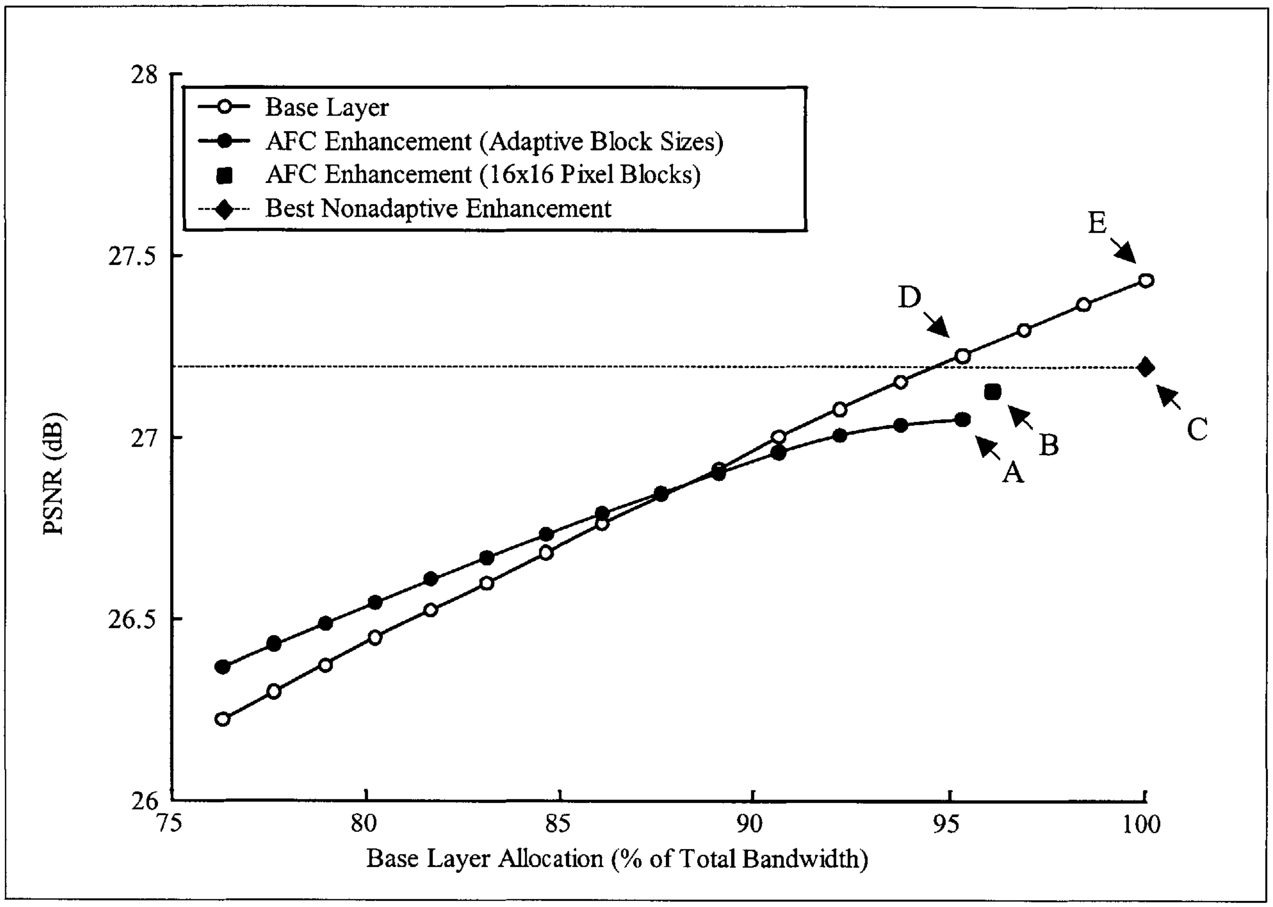


Figure 4.20: PSNR Versus Base Layer Allocation for the Picnic Sequence

The PSNR of the AFC enhanced video improves continuously with base layer allocation, but at its highest point (B) remains 0.07 dB below the best nonadaptive conversion (C). The video at points B and C appear relatively equal in quality. The corresponding difference in base layer PSNR is 0.2 dB between points D and E. Acute block effects in the compressed base layer make adaptive format conversion ineffective in improving the picture quality.

4.2.10 Traffic

The first frame of the Traffic sequence is shown in figure 4.21. This video shows a busy intersection in front of a shopping center parking lot. The cars in the foreground move at various speeds in different directions while the background moves slowly to the left. Because of the amount of detail and motion in this video sequence, certain areas (the cars in the parking lot, for instance) are encoded poorly in select frames.

The result of the tradeoff experiment for the Traffic sequence is illustrated in figure 4.22. For this sequence, the best nonadaptive format conversion yields a 0.16 dB higher PSNR than the best quality adaptive format conversion, which occurs at 96% base layer allocation. In the base layer, the difference between points D and E is 0.22 dB. The best quality nonadaptive format conversion method is Martinez-Lim deinterlacing.

Even though the AFC enhanced video at point B has a lower PSNR than the NFC video at point C, adaptive format conversion helps to eliminate some subtle flickering and distortion. However, the video at point C has less distortion in other areas. In comparing the base layer at points D and E, it is easy to see that the video sequences are distorted, but difficult to say that one is much worse than the other. For the Traffic video sequence, it may be argued that the total bandwidth is best used to improve the base layer quality.

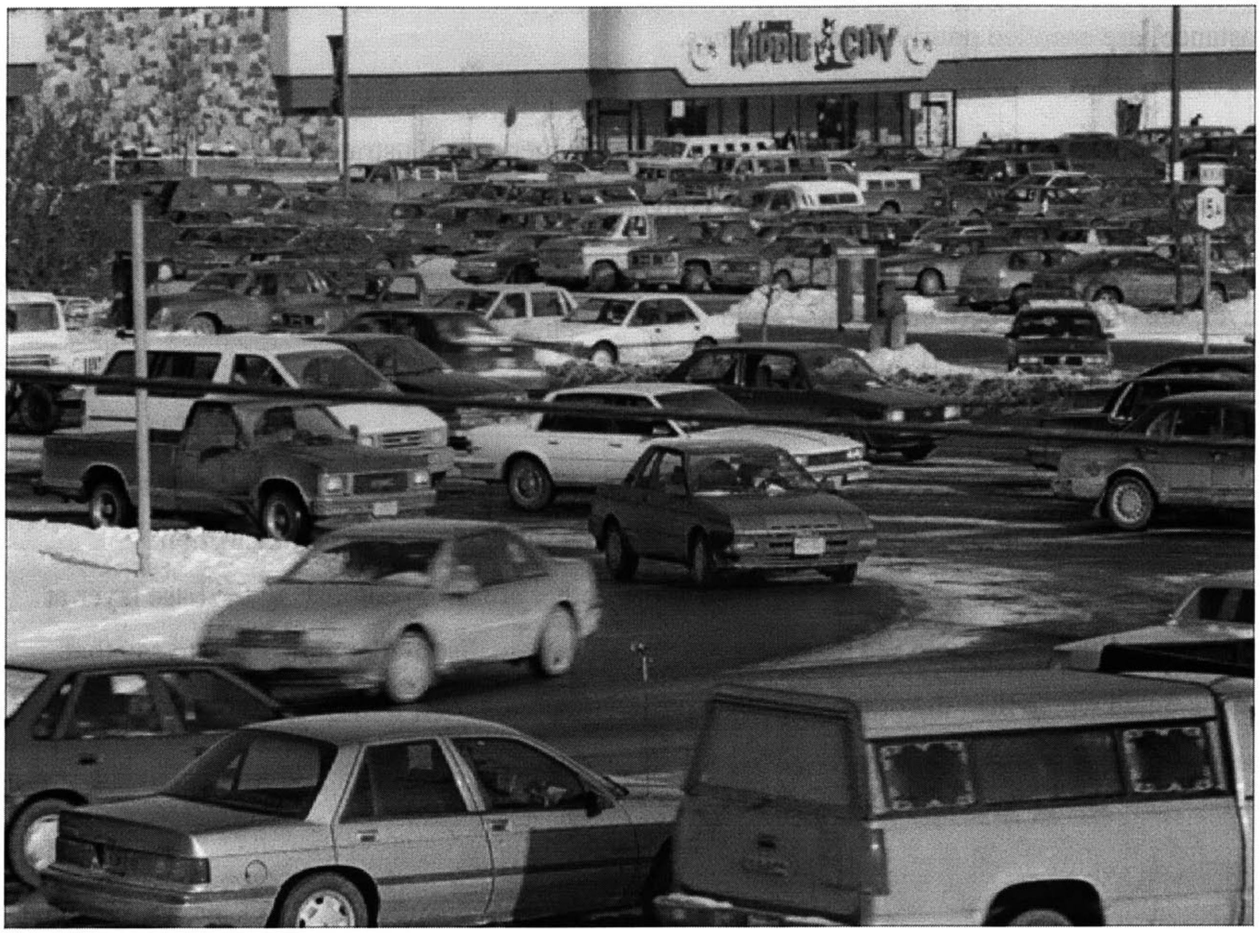


Figure 4.21: First Frame of the Traffic Sequence

The cars in the foreground move at various speeds in different directions while the background pans slowly to the left.

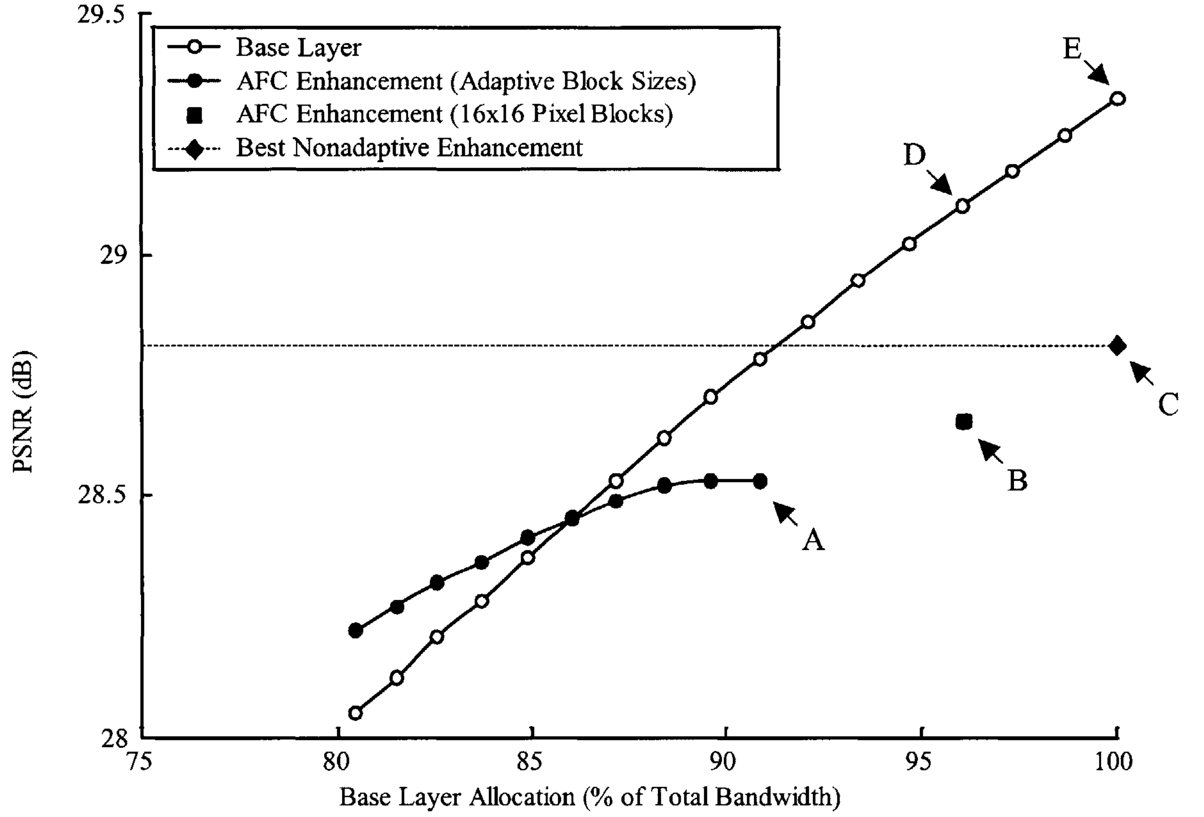


Figure 4.22: PSNR Versus Base Layer Allocation for the Traffic Sequence

Points B and C differ by 0.16 dB; points D and E by 0.22 dB. Adaptive format conversion removes some subtle deinterlacing artifacts, but a better base layer at point E makes the corresponding NFC video superior in certain areas. In this case, the best tradeoff is at point C, with 100% base layer allocation.

4.3 Summary

This chapter contains the results of two main experiments. The purpose of the first was to determine how adaptive format conversion performs when the base layer video is encoded in a manner typical to HDTV. The Carphone and News sequences were interlaced and compressed at 0.3 bpp, and the resulting base layer PSNR was 36.1 dB and 34.9 dB, respectively. Even though the base layer video was encoded using a different approach than in previous work, the PSNR improvement due to adaptive format conversion was almost exactly the same for these values of base layer PSNR. The PSNR improvement was between 3 and 4 dB for Carphone and between 3 and 3.5 dB for News. Usually, a PSNR improvement greater than 1 dB is significant and can be seen quite easily. For both video sequences, the enhancement layer bandwidth ranged from approximately 0.015 to 0.25 bpp.

The second experiment was an attempt to discover the optimal tradeoff between base layer and enhancement layer bandwidth in a fixed bandwidth environment such as HDTV. In this experiment, the total bandwidth (base plus enhancement layer) was fixed at 0.32 bpp, the base layer was interlaced, and the enhancement layer contained adaptive deinterlacing information.

Figure 4.23 summarizes the optimal tradeoff point for each of the ten video sequences that were studied in this experiment. The best enhanced resolution video was determined by visual inspection and was either the NFC video, the AFC video using 16x16 pixel blocks, or the AFC video using adaptive block sizes that had the highest PSNR. For four of the test sequences (Car, Football, Toy Train, and Tulips Zoom), the optimal tradeoff point resulted from fixed block size AFC at about 96% base layer, 4% enhancement layer allocation. Three sequences (Football Still, Picnic, and Traffic) profited the most from 100% base layer allocation and nonadaptive deinterlacing; however, for these three sequences the fixed block size AFC enhancement was nearly identical in quality. The remaining three sequences (Marcie, Girl, and Tulips Scroll) used AFC with adaptive block sizes to achieve the highest enhancement layer quality. For Marcie (90% base layer allocation), the result for fixed and adaptive block sizes was similar, but for Girl (87%) and Tulips Scroll (85%), adaptive block sizes produced superior results.

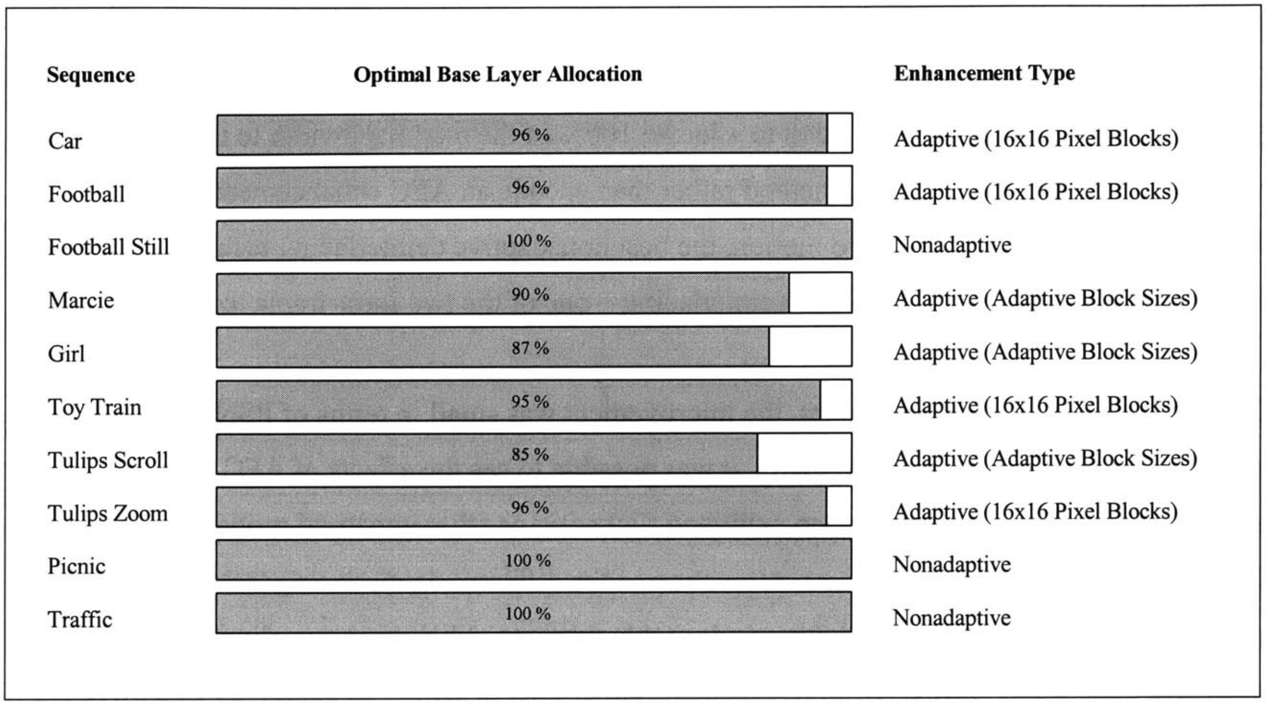


Figure 4.23: Summary of Optimal Tradeoff Between Base and Enhancement Layer Bandwidth for Ten Video Sequences

For the ten video sequences in table 4.1, the optimal tradeoff between base and enhancement layer bandwidth is summarized in this figure. The best quality enhancement layer was determined by visual inspection and the type of enhancement that led to this result is listed—either nonadaptive, adaptive with 16x16 pixel blocks, or adaptive with adaptive block sizes. For eight out of ten sequences, fixed block size AFC produced the best or very close to the best quality enhancement layer video. Only for Girl and Tulips Scroll was adaptive block size AFC visibly superior.

The overhead required to transmit partitioning information for adaptive block sizes is a significant factor. When adaptive block sizes result in the optimal tradeoff, the enhancement layer bandwidth was about fifteen percent of 0.32 bpp, or 0.048 bpp. This value is close to the smallest possible enhancement layer bandwidth, meaning that most of the frame is divided into 16x16 pixel blocks with only a few 8x8 and 4x4 pixel blocks where they are the most beneficial. Fixed block sizes, without the partitioning overhead, use less of the total bandwidth to produce better results in most cases.

Recall that the encoder has the option to allocate 100% of the total bandwidth to the base layer and specify the best nonadaptive method rather than encode an AFC enhancement layer. Ignoring the single case with zero motion, the best nonadaptive deinterlacing technique is either linear interpolation or Martinez-Lim deinterlacing—one of the two intra-frame techniques.

For most of the video test sequences, the improvement was small in terms of PSNR, but by examining the video sequences visually, it was possible to see the effects of AFC and NFC enhancement. Often, the NFC video exhibited flickering or other unnatural motion, aliasing, and jagged edges. In the AFC video, a combination of the different deinterlacing techniques effectively reduced or eliminated these undesirable artifacts. Flickering was diminished by choosing inter-frame deinterlacing techniques—either FFR or BFR—in stationary or slowly moving regions such as the background. A combination of Martinez-Lim deinterlacing, FFR, and BFR was used to minimize the negative effects of aliasing, and jagged edges were almost entirely removed with Martinez-Lim deinterlacing. The AFC enhanced video regularly appeared sharper than when a single deinterlacing technique was used.

Compared to the best nonadaptive format conversion, the biggest improvement (both visually and in PSNR) occurred when part of the video was in motion and the remainder was either stationary or characterized by slow, uniform motion. The improvement can be attributed to the strength of the different deinterlacing techniques in handling these two types of video content. On the other hand, little or no improvement was seen when the video was still or dominated by complex motion. In these situations, a single deinterlacing technique was generally superior for most of the video, so the AFC enhancement layer contained little or no useful information.

Chapter 5

Conclusion

5.1 Summary

This thesis began by introducing the U.S. HDTV standard and its significant advantages over the old NTSC system. With high-resolution video, a wide aspect ratio, CD quality surround sound, digital transmission, and multiple transmission formats, HTDV offers an all-around superior television experience. Despite these improvements, the cost of switching to HDTV is prohibitive. On the other hand, the new HDTV standard allows additional features in a backward-compatible manner, paving the way for future improvements.

Looking ahead, the need is already recognized for video formats that are beyond the scope of current technology. 1080P is one such format that violates the sample rate constraint of MPEG-2, the compression algorithm used for HDTV video coding. 1080P is a logical first step in the migration path to higher-resolutions. As used in this thesis, the term "migration path" refers to the problem of adding support for 1080P in a way that requires little additional bandwidth and is backward compatible with the existing HDTV standard.

The solution presented here is a scalable video codec based on adaptive format conversion information. Traditionally, scalable coding schemes are based on a single format conversion technique and residual coding. Adaptive format conversion is different in that it employs several different format conversion techniques, partitioning the video sequence into small blocks and then choosing the best technique in each block.

Previous research indicates that adaptive format conversion may be an ideal solution to the migration path for HDTV. The base layer video, either 1080I, 720P, or 1080P@30fps, would be a format that is compatible with the current HDTV standard. Furthermore, adaptive format conversion requires a relatively small bandwidth, making AFC uniquely suited to the demands of the migration path. However, prior work failed to address two issues that are important in evaluating AFC in this context. These issues are the coding of the base layer video and the optimal tradeoff between base and enhancement layer bandwidth.

In the previous approach, the base layer was encoded using all I-pictures and a fixed quantizer scale factor; in other words, no motion compensation or rate control strategy was employed. This implementation did not allow meaningful measurements of the base layer bandwidth, so it was impossible to say what quality of video or amount of improvement could be expected for HDTV. In the current work, the TM5 video codec was configured to use motion compensation and rate control and encode the video in a way that is consistent with the common HDTV format 1080I. Under this configuration, the mean base layer PSNR was 32 dB, with individual values ranging anywhere from 25 to 38 dB. For the Carphone and News sequences, in particular, the base layer PSNR was 36.1 dB and 34.9 dB, respectively. The PSNR improvement due to the AFC enhancement layer was between 3 and 4 dB for Carphone and 3 and 3.5 dB for News. The range of enhancement layer bandwidths was between 0.015 and 0.25 bpp. These results are almost exactly the same as reported by Wan, who demonstrated that when the base layer PSNR is above 25 dB and the enhancement layer bandwidth is below 0.25 bpp, this implementation of adaptive format conversion is superior to residual coding. In fact, AFC is the only choice for small enhancement layer bandwidths below about 0.1 bpp. The results of this experiment support the use of AFC in the migration path, where only a small enhancement layer bandwidth is expected in the near future.

The second issue deals with the optimal tradeoff between base and enhancement layer bandwidth when the total bandwidth is fixed. Experiments were performed for ten different video sequences in order to see if an optimal tradeoff exists—one that consistently provides the best enhancement layer video without sacrificing base layer video quality. Again, the framework for these experiments was adaptive deinterlacing.

With the total bandwidth set at 0.32 bpp, the results indicate an optimal tradeoff when 96% of the bandwidth is allocated to the base layer and the remaining 4% is allocated to the AFC enhancement layer, which uses fixed, 16x16 pixel blocks. To reach this conclusion, the AFC enhanced video was compared to the best nonadaptive format conversion when 100% of the bandwidth was allocated to the base layer. Although the improvement was usually small in terms of PSNR, the AFC enhanced video usually appeared sharper, with less flickering, fewer aliasing artifacts, and straighter lines than its NFC counterpart. The biggest improvement was seen in video with a small number of moving objects in front of a mostly stationary background. In this situation, adaptive format conversion was able to exploit the strengths of several different deinterlacing techniques in handling these two types of video—FFR and BFR in stationary regions and linear interpolation and Martinez-Lim deinterlacing in areas with motion. The amount of improvement was also influenced by the quality of the compressed base layer, with better quality generally leading to more improvement in the enhancement layer. There were two cases in which adaptive block sizes and 85-87% base layer allocation was superior to fixed block sizes. On the other hand, fixed 16x16 pixel blocks generally provided substantial improvement over the best nonadaptive conversion and used fewer bits, causing less degradation in the base layer. Going from 100% to 96% base layer allocation *does* result in a reduction in base layer video quality; however, in most cases the distortion is not noticeable. Only when the compressed base layer quality is poor is it better to allocate all the bandwidth to the base layer and use the best nonadaptive interpolation method, and an encoder has the option to do so.

5.2 Directions for Future Research

The conclusions above support the idea of using adaptive deinterlacing in the migration to 1080P, but other questions should be addressed as AFC is considered for the HDTV migration path. Some of these questions are discussed below.

The implementation of adaptive deinterlacing used in this thesis has been studied before. It uses either fixed or adaptive frame partitioning and selects from four different deinterlacing techniques (forward field repetition, backward field repetition, linear interpolation, and

Martinez-Lim deinterlacing). The parameter selection for adaptive block sizes was shown to be optimal (in the rate-distortion sense), and the enhanced video is substantially improved; however, there may be other implementations that are even more efficient. For instance, many of the video test sequences studied in this thesis contained a large background region distinguished by a slow, horizontal pan. The four deinterlacing techniques used here are not particularly suited to this type of motion, but one can imagine a variation of Martinez-Lim deinterlacing that finds the best temporal (rather than spatial) line shift being more successful. Adding more deinterlacing methods or more frame partitioning options would tend to increase the enhancement layer bandwidth. It is presently unclear whether four format conversion methods is the best choice or if increased complexity would lead to a more or less efficient enhancement layer.

As shown previously, a 1080I base layer with adaptive deinterlacing information is not the only way to create a 1080P enhancement layer. The base layer format could also be 720P or 1080P@30fps, and it would be interesting to see how a migration path based on these formats would compare. This question presents some interesting challenges.

First, even though the enhancement layer format (1080P) is the same, it is unclear how to compare the different base layer formats. It may end up that one particular format, say 1080I, consistently leads to the best enhancement layer, but that another format, such as 1080P@30fps, is favored for the base layer. The base layer format is often determined by the source of the video, its content, or broadcaster preference.

Second, deinterlacing, spatial upsampling, and temporal upsampling methods have different characteristics. It is accepted that encoding progressive video is more efficient that encoding interlaced video (see figure 2.3). On the other hand, deinterlacing, which begins with half the information in every frame, has a significant advantage over temporal upsampling, which must reconstruct entire frames. In a temporal upsampling scheme, the error would be concentrated in half the number of frames and may appear quite unpleasant. Any good temporal upsampling technique would use some form of motion estimation, and the computation and complexity required for this approach are prohibitive. Therefore, AFC may not be the best solution for temporal upsampling. A better tactic might be to use motion compensated residual coding, where motion vectors and residual information are encoded in an enhancement layer. Unfortunately, this would require more bandwidth than may be available in the near future.

Spatial upsampling from 720x1280 pixels has unique challenges as well. The number of pixels in 720P is only 44% of those in 1080P, compared to 50% for 1080I and 1080P@30fps. As a result, 720P is encoded at approximately 0.34 bpp. Even though the base layer quality would be higher, format conversion would have to replace more information and could not replace high frequency detail lost in downsampling. In areas with motion, 720P might lead to the best enhancement layer, but spatial upsampling techniques would not be able to match the performance of deinterlacing or temporal upsampling in stationary regions. Instead of selecting from a number of fixed spatial upsampling filters, a better approach might be to create an adaptive filter and transmit the filter coefficients. This can be done with a small bandwidth if the partitions are large, such as an entire frame.

Adaptive spatial upsampling of 720P, temporal upsampling of 1080P@30fps, and deinterlacing of 1080I are compared side by side in figures 5.1 and 5.2 for the Car and Girl sequences, respectively. The two figures show the enhancement layer PSNR as a function of enhancement layer bandwidth. In each case, four different format conversion techniques were used. For spatial upsampling, the methods were nearest neighbor, bilinear, bicubic, and two dimensional DCT (the weighted basis functions were renormalized and resampled at a higher resolution). For temporal upsampling, the methods were forward frame repetition, backward frame repetition, linear interpolation, and a motion compensated technique based on the algorithm in [22]. The deinterlacing methods were the same. The methods themselves are not important, but illustrate how this problem might be approached.

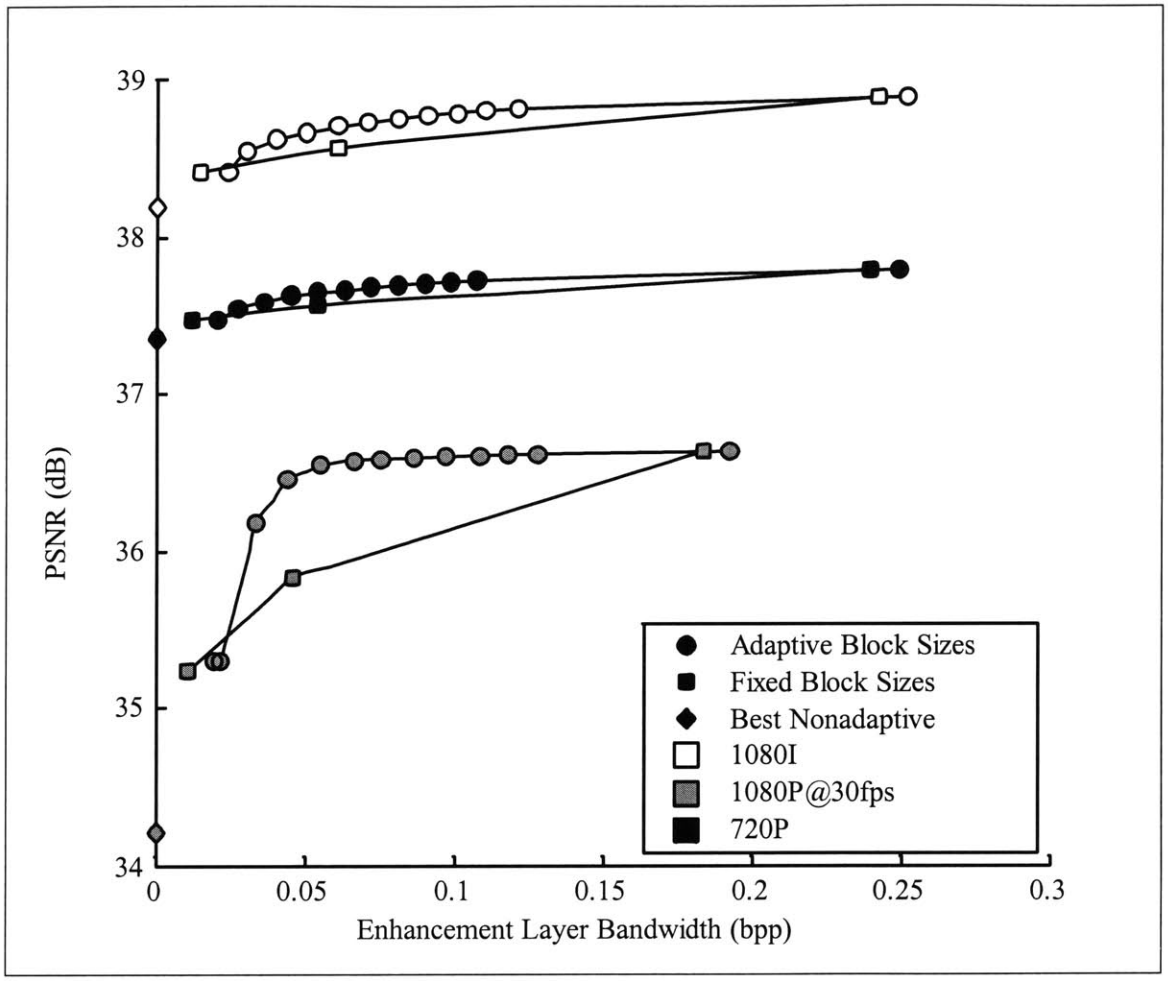


Figure 5.1: Migration to 1080P from 1080I, 1080P@30fps, and 720P for the Car Sequence

For the Car sequence, deinterlacing results in the highest enhancement layer PSNR, followed by spatial interpolation and temporal upsampling. Temporal upsampling gains the most from the AFC enhancement layer—an improvement of 2.5 dB compared to 0.5 dB for deinterlacing or spatial upsampling.

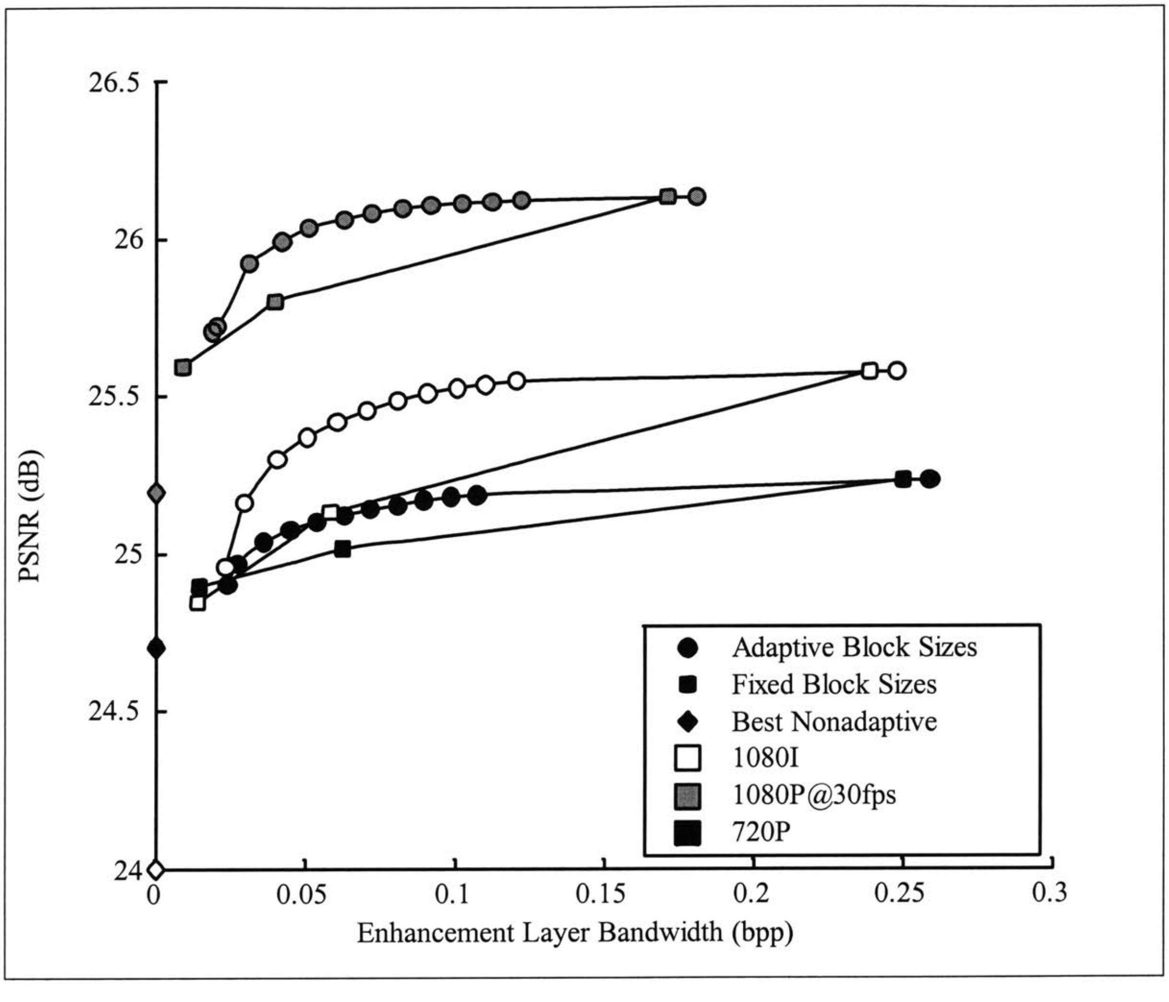


Figure 5.2: Migration to 1080P from 1080I, 1080P@30fps, and 720P for the Girl Sequence

For the Girl sequence, temporal upsampling results in the highest enhancement layer PSNR, followed by deinterlacing, then spatial upsampling. Deinterlacing is most improved by AFC enhancement—by about 1.5 dB. Temporal upsampling of the Girl sequence is superior due to higher quality images and uniform motion that is easy to predict.

For the Car sequence in figure 5.1, deinterlacing produces the highest enhancement layer PSNR, followed by spatial, then temporal upsampling. Temporal upsampling benefits the most from the AFC enhancement layer (up to 2.5 dB, compared to around 0.5 dB for the other two base layer formats). Figure 5.2 has a different order: temporal upsampling has the highest enhancement layer PSNR, followed by deinterlacing and spatial upsampling. The difference may be attributed to uniform motion in the Girl sequence, which is easier to predict than the spinning car. For the Girl sequence, deinterlacing gets the most help from AFC enhancement—about 1.5 dB. In both figures 5.1 and 5.2, we see that adaptive spatial upsampling offers the least amount of improvement and leads to a lower quality enhancement layer than adaptive deinterlacing.

The above analysis is not meant to be complete. Rather, it illustrates that performance depends on the video content. In order to better understand the behavior of adaptive spatial and temporal upsampling, a comprehensive study is recommended similar to the one performed in this thesis for adaptive deinterlacing.

5.3 Recommendation to Broadcasters

Adaptive format conversion is able to improve the quality of deinterlaced video with only a small enhancement layer bandwidth. Deinterlacing techniques are simple to implement, and it was shown that the simplest partitioning scheme (fixed, 16x16 pixel blocks) leads to the best tradeoff between base and enhancement layer bandwidth in most cases. As a result, it is recommended that the migration path from 1080I to 1080P use adaptive filter selection for each 16x16 pixel block.

This migration would not hinder later development of the migration path, nor would the migration from 720P or 1080P@30fps need to be implemented at the same time. It has been shown that when the base layer is coded well, a combination of AFC and residual coding is more efficient than residual coding alone. Therefore, adaptive format conversion may be a starting point for residual enhancement (for quality) or further spatial or temporal upsampling beyond 1080x1920 pixels or 60 frames per second.

References

[1] *ATSC Digital Television Standard*, ATSC Document A/53, September 16, 1995.

[2] *Fourth Report and Order*, MM Docket No. 87-268, FCC 96-493, Adopted December 24, 1996.

[3] R. Graves, *Development of the ATSC DTV Standard*, http://www.atsc.org, December 15, 1999.

[4] J. Lim, "High-Definition Television," *Wiley Encyclopedia of Electrical and Electronics Engineering*, vol. 8, pp. 725-739, 1999.

[5] L. Sunshine, *HDTV Transmission Format Conversion and Migration Path*, Department of Electrical Engineering and computer science, Massachusetts Institute of Technology, Cambridge, MA, September 1997.

[6] J. Lim and L. Sunshine, "HDTV Transmission Formats and Migration Path," *International Journal of Imaging Systems and Technology*, vol. 5, pp. 286-291, Winter 1994.

[7] W. Wan, *Adaptive Format Conversion Information as Enhancement Data for Scalable Video Coding*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, May 2002.

[8] W. Wan and J. Lim, "Adaptive Format Conversion for Video Scalability at Low Enhancement Bitrates," in *MWSCAS 2001: Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems*, vol. 2, (Fairborn, OH), pp. 588-592, IEEE, August, 2001.

[9] W. Wan and J. Lim, "Adaptive Format Conversion for Scalable Video Coding," in *Proceedings of SPIE: Applications of Digital Image Processing XXIV*, vol. 4472, (San Diego, CA), pp. 390-401, SPIE, July 2001.

[10] D. Martinez and J. Lim, "Spatial Interpolation of Interlaced Television Pictures," in *ICASSP-89: 1989 International Conference on Acoustics, Speech and Signal Processing*, vol. 3, (Glasgow, UK), pp. 1886-1889, IEEE, May 1989.

[11] J. Lim, *Two-dimensional Signal and Image Processing*, Prentice Hall Signal Processing Series, Upper Saddle River, NJ, 1990.

[12] J. Mitchell, W. Pennebaker, C. Fogg, and D. LeGall, eds., *MPEG Video Compression Standard*, Digital Multimedia Standard Series, Chapman and Hill, New York, NY, 1997.

[13] B. Haskell, A. Puri, and A. Netravali, eds., *Digital Video: An Introduction to MPEG-2*, Digital Multimedia Standard Series, Chapman and Hill, New York, NY, 1997.

[14] T. Ebrahimi and C. Horne, "MPEG-4 Natural Video Coding – An Overview," *Signal Processing: Image Communication*, vol. 15, pp. 365-385, January 2000.

[15] I.JTC1/SC29/WG11, *11172-2: MPEG-1 Video*, ISO/IEC, 1993.

[16] I.JTC1/SC29/WG11, *13818-2: MPEG-2 Video*, ISO/IEC, 1995.

[17] I.JTC1/SC29/WG11, *14496-2: MPEG-4 Video*, ISO/IEC, 1999.

[18] MPEG-2 Software Simulation Group, *MPEG-2 Encoder/Decoder, Version 1.2*, http://www.mpeg.org/MSSG/, July 1996.

[19] S. Eckart and C. Fogg, ISO-IEC MPEG-2 Software Video Codec, *SPIE Digital Video Compression: Algorithms and Technologies '95*, vol. 2419, pp. 100-109, 1995.

[20] A. Ortega and K. Ramchandran, "Rate-Distortion Methods for Image and Video Compression," *IEEE Signal Processing Magazine*, vol. 15, pp. 23-50, November 1998.

[21] G. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," *IEEE Signal Processing Magazine*, vol. 15, pp. 74-90, November 1998.

[22] C. Cafforio, F. Rocca, and S. Tubaro, "Motion Compensated Image Interpolation," *IEEE Transactions on Communications*, vol. 38, pp. 215-222, February 1990.

3571-87