# Compatible Video Coding of Stereoscopic Sequences using MPEG-2's Scalability and Interlaced Structure

**Belle L. Tseng and Dimitris Anastassiou**

Columbia University, Dept. of Electrical Engineering & Center for Telecommunications Research, Room 801 Schapiro Research Building, Columbia University, New York, NY 10027, U.S.A.
TEL: +1 (212) 854-6481 FAX: +1 (212) 316-9068

## Abstract

Three approaches of an MPEG-2 compatible coding technique are presented for stereoscopic sequences. The first method utilizes the spatial scalability structure and the second employs the temporal scalability syntax. The scalability extensions of the video coding standard make the processing easier to accommodate the transmission of a stereoscopic video stream. The left and right channels required for a stereo sequence are correspondingly supported in the base and enhancement layers of the scalability structure. The enhancement layer selects the best prediction combination of the spatial and temporal information.

In the third technique, the left and right stereoscopic images are represented and coded as an interlaced video. For interlaced sequences, transmission in either the field or frame picture structure can be chosen for each frame to maximize compression, and therefore improve image reconstruction. Selection of picture structures depends on the temporal changes of the stereoscopic scenery. Field picture structure is chosen for scenes with high stereoscopic activities, whereas frame structure is preferred for stereo images with little or no disparity variations.

Experimental results are presented for the proposed approaches and performance comparisons between them are analyzed and interpreted. Simulations on several stereoscopic sequences illustrate that the proposed spatial and temporal scalability methods always achieve better SNR performances than the simulcast method. Experimental data on progressive stereo sequences with the left-right interlaced structure approach show improving performances, as compared to the simulcast and both scalability techniques, when transmitted with increasing bit rate. As a result, the interlaced field structure approach, supported in the main profile of MPEG-2, is the recommended choice for high bit rate coding of progressive stereo sequences.

## 1 Introduction

Motivated by the idea of a combination 3D system and HDTV, we study the codec of 3DTV compatible with MPEG-2 standardization. Benefits and preferences for stereoscopic 3DTV have been shown in many applications, including medical imaging, remote handling, and quality control, where depth impression enhances the viewing experience and improves brilliance and fidelity. The future possibly stereoscopic wide-screen TV now imposes on the idea of a combination 3DTV and HDTV. It is for this goal that we study the codec of the 3DTV compatible with MPEG-2 standardization.

Basic transmission of a stereoscopic video sequence consists of a left and right channel, each carrying images captured from the corresponding views. With the objective of compression and transmission on a bandwidth-limited channel, efficient coding of the two video signals is achieved by exploiting their interrelationships. Compression can be obtained by studying possible reductions in the spatial redundancy between the left and right images of a stereo pair and analyzing motion compensation of individual video channels.

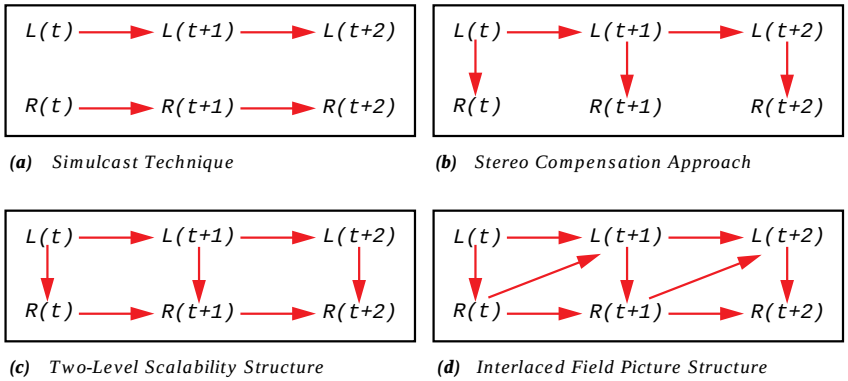


Figure 1: Picture Predictions with Different Coding Approaches

One simple solution to stereoscopic video coding is the "simulcast" technique depicted in Figure 1a, which is based on transmission and reproduction of independently coded channels, The left stereoscopic sequence is encoded and transmitted separately from the right sequence. In accordance, predictions for the right channel are made on the basis of motion in the right channel itself, and similarly for those of the left channel.

Another conventional coding approach involves stereo compensation between the stereo pairs as illustrated in Figure 1b, where each right image is predicted from its corresponding left signal [1]. The left video sequence is encoded and transmitted, whereas only residual information is send for the construction of the right sequence. Predictions for the right channel are made on the basis of disparity information between the corresponding left and right stereo images.

We propose a more efficient alternative to stereo reproduction based upon the application of scalable video coding, as diagrammed in Figure 1c, compatible with the MPEG-2 specification [2]. Utilizing the high profile double layer structure, the base layer can support the left stereoscopic sequence while the enhancement layer manages the prediction of the right sequence.

We introduce two scalability structures for the transmission of stereoscopic video. One is the spatial scalability syntax, whose main purpose is for the transmission of different spatial resolutions. For stereo transmission, spatial scalability takes advantage of the adaptive selection and combination of spatial and temporal predictions. The second is the temporal scalability option, which can produce variable temporal resolutions of the input video. In application to stereo signals, the main feature of temporal scalability is the direct and easy implementation.

Another novel approach presents itself by combining the two stereo channels into one interlaced video as shown in Figure 1d with field picture structure. Handling the stereo signals as one interlaced video offers a mixture of predictions between the two sequences. Because interlaced video coding allows adaptive selection of two picture structures, field picture structure benefits for those sceneries with "high stereoscopic activity" whereas frame picture structure takes advantage of those scenes with less activity. However this approach can not accommodate the transmission of an interlaced stereoscopic signal, only progressive stereo sources.

# 2 MPEG-2's Scalability Profile

MPEG-2 specifies the addition of the scalability extensions to the video coding process. The scalability tools are incorporated to handle processing of various multiple resolution reproductions of a single video source. The coding of one video can be reutilized in the coding of other resolutions of the same video bitstream, thus conforming to the multiple complexities of the individual decoding systems. As a result, scalable video processing is the more efficient alternative to independent transmission/storage of multiple resolution videos, otherwise referred to as the simulcast technique.

In the scalability syntax, two layers of video structure referred to as the base layer and the enhancement layer are allowed. The base layer is coded by itself to provide for the more important basic video stream, whereas the enhancement layer carries the residual information for full reproduction of the original video signal. The lower base layer performs the same process as the main nonscalable video coding, and the higher enhancement layer performs the resolution scalable extension, whether it be data partitioning, SNR scalability, spatial scalability, or temporal scalability.

In addition to the multi-layer processing advantage, the scalability extension provides resilience to transmission errors as the more important data of the base layer can be send over the channel with better performance, while the less critical bitstream of the enhancement layer can be send over the channel with poorer performance. Consequently, graceful degradation of the overall reconstruction of the original video signal is achieved in the presence of channel errors. As a result, robustness of one channel is consistently satisfied, providing the receiver with at least the reconstruction of the basic sequence.

# 3 Spatial Scalability Video Coding

## 3.1 Spatial Scalability Profile

The main feature of the spatial scalable coding is its capability to accommodate multiple spatial resolutions of an input. Spatial scalability divides one video source into at least two signals, one at a low spatial resolution, and another usually at the full spatial resolution. The downsampled low resolution sequence is coded in the base layer, and this information is partially utilized in the higher resolution coding, which takes place in the enhancement layer.

The base layer codec adapts the same syntax as specified by the non-scalable video coding of MPEG-2. In the enhancement scalable layer, two forms of predictions are available for each macroblock. One is the motion compensated "temporal" predictions, made with respect to pictures in the same enhancement layer. The second is the interpolated "spatial" predictions, formed by up/down sampling the decoded signals of the base layer. Furthermore, the temporal prediction, the spatial prediction, or their weighted combinations, can be objectively chosen as the best selection for each enhancement macroblock.

## 3.2 Stereo Compatible Spatial Scalable Coding

For stereo sequences where two distinct video streams are involved, the key advantage of the scalable syntax is to provide the non-stereoscopic bitstream from the standalone base layer and adding the residual "stereoscopic" signal to the enhancement layer.

In Figure 2, a block diagram of the scalable video encoder and decoder is presented for stereo signals with several simplifications for clarity purposes. At the encoder end, the original *left sequence* is coded as per specification of MPEG-2 and transmitted on the *base layer bitstream*. However instead of coding the original *right sequence* similarly, as is done with the simulcast approach,

estimates for the spatial and temporal predictions are derived. *Spatial predictions* are obtained from the disparity estimates between the corresponding left and right stereo pairs, and simultaneously *temporal predictions* are calculated from the motion vector estimates. Following, the spatial and temporal predictions can be selected or combined to form the actual prediction, which achieves the lowest mean squared error with respect to the original right image. The possible prediction weights are specified in the scalability syntax of [2]. Following, the relevant prediction data are transmitted along with the combination weight code in the *enhanced layer bitstream*.

At the decoder, on receiving the *base layer bitstream*, decompression and prediction of the *left sequence* is obtained according to the non-scalable syntax of MPEG-2. On the other hand when the *enhanced layer bitstream* arrives at the receiver, spatial and temporal predictions of the right signal are simultaneously decoded. Using stereo compensation on the spatial information and motion compensation on the temporal signal, the final *right sequence* is reconstructed.

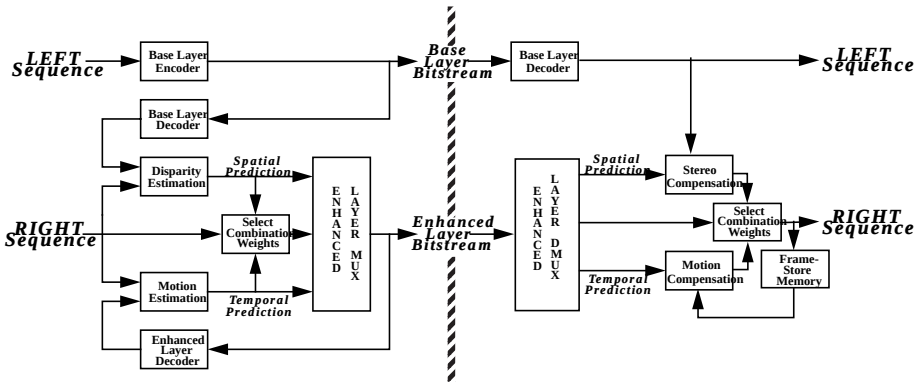


Figure 2: Two-Level Spatial Scalability Encoder/Decoder for Stereoscopic Sequences

Incorporating spatial scalable coding with compatibility for stereoscopic video, graceful degradation of the overall stereo reconstruction is achieved in the presence of channel errors. The "non-stereoscopic" signal, namely one of the two binocular sequences, becomes the base layer bitstream and can be transmitted on the guaranteed channel. As a result, robustness of one channel is consistently satisfied, providing the receiver with at least the construction of a non-stereoscopic sequence.

# 4 Temporal Scalability Video Coding

## 4.1 Temporal Scalability Feature

The primary objective of the temporal scalability feature is to accommodate future sophisticated high temporal resolution systems. In order for compatibility of various temporal resolutions, the video coding standard adapts the multilayer structure of scalable video coding. The source input signal is partitioned into two or more lower temporal resolution bitstreams, each with the same spatial resolution. The primary low temporal resolution signal is coded on the base layer and is the only bitstream decoded by the low temporal resolution systems. For more sophisticated higher temporal resolution systems, the residual video bitstream can be utilized to enhance the primary lower layer signal, by remultiplexing to form the original input video in full frame rate. In the two layer temporal scalability structure, those remaining frames are coded on the enhancement layer.

Temporal scalability, similar to the other scalability profiles, is built upon the main framework of MPEG-2. The base layer codec performs the equivalent operations as the non-scalable video coding. In addition, the codec for the enhancement layer is similar to that for the lower layer. The difference being that the motion compensated predictions are now with respect to frames from either the base or the enhancement layer.

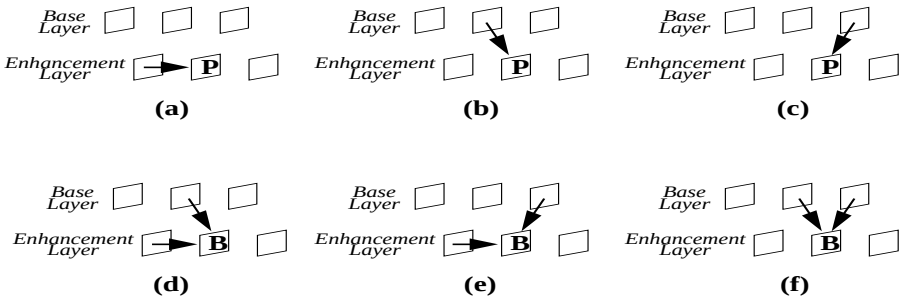


Figure 3: Motion Compensated Prediction Reference Selection for P-Pictures and B-Pictures

Each frame from the enhancement layer can be designated to be either an I, P, or B-picture. The prediction reference selection differs for the I, P, and B-pictures and are fully specified in the MPEG-2 standard [2]. In brief, the I-pictures of the enhancement layer are similarly intraframe coded as in the base layer. In P-pictures, the forward prediction is selected from three reference frames as illustrated in Figures 3(a-c). For the B-pictures, the forward and backward predictions are chosen from three sets of forward and backward reference frames as shown in Figures 3(d-f). Two possible configurations of picture structure assignments for temporal scalability video coding are presented and recommended in [3]. After specifying the picture structure, the enhancement encoder finds the appropriate prediction motion vectors and the rest is identical to the base layer processing.

## 4.2 Stereo Compatible Temporal Scalable Coding

For stereoscopic video transmission, the temporal scalability feature of MPEG-2 offers two interdependent layers of coding. The left stereo sequence is coded on the lower layer and offers the basic non-stereoscopic signal. The right stereo bitstream is then transmitted on the enhancement layer and when combined with the left view results in the full stereoscopic video.

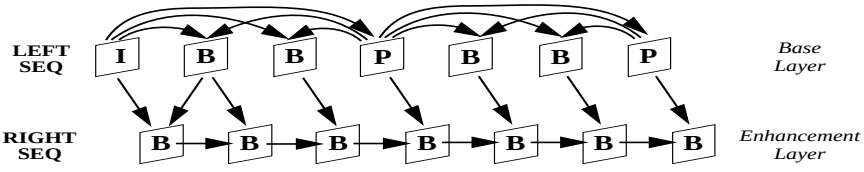


Figure 4: Selected Picture Configuration for Temporal Scalability Coding of Stereoscopic Sequences

Processing of the enhancement layer is similar to that of the base layer except for the selection of picture structures and their prediction reference frames. Assuming the base layer is coded using the M=3 structure as describe in the MPEG-2 documentation. We have chosen the configuration shown in Figure 4, where all the enhancement frames are B-pictures. For the enhanced B-pictures corresponding to the I-pictures of the base layer, both predictions are derived from the surrounding decoded frames in the lower layer. For the other enhanced B-pictures, the forward prediction is

the motion compensated prediction from the decoded previous enhanced frame. The backward prediction is obtained from the disparity estimated prediction of the most recently decoded lower layer frame. Thus these two predictions represent the spatial and the temporal information.

Implementing temporal scalable coding with compatibility for stereoscopic video is a much easier addition when compared with spatial scalability. In the temporal scalable case, the disparity estimates are transmitted as prediction vectors which are embedded in the basic motion compensated framework of MPEG-2. In spatial scalability, no motion vectors are required in the blowup of spatial resolutions; consequently, disparity estimates are externally processed in the user data section.

Another advantage offered by temporal scalable coding is graceful degradation of the overall stereo reconstruction in the presence of transmission errors. By coding one binocular sequence on the standalone base layer and transmitting it on a reliable channel, all decoding systems receive at least the non-stereoscopic sequence. Although the enhancement bitstream is sent over the worse channel, decoding of this stereo signal is relative to the decoded base sequence and thus performs better than if coded and transmitted independently.

# 5 Interlaced Video Coding

Each frame of an interlaced video sequence consists of two fields. MPEG-2 specification allows for two picture structures, either field picture or frame picture. With field picture selection, the two field pictures are coded independently of each other, whereas in frame picture coding, the two fields are interleaved to form a frame picture and coded as one image. Furthermore, interlaced sequence coding allows dynamic switching between the two picture structures on a frame-by-frame basis.

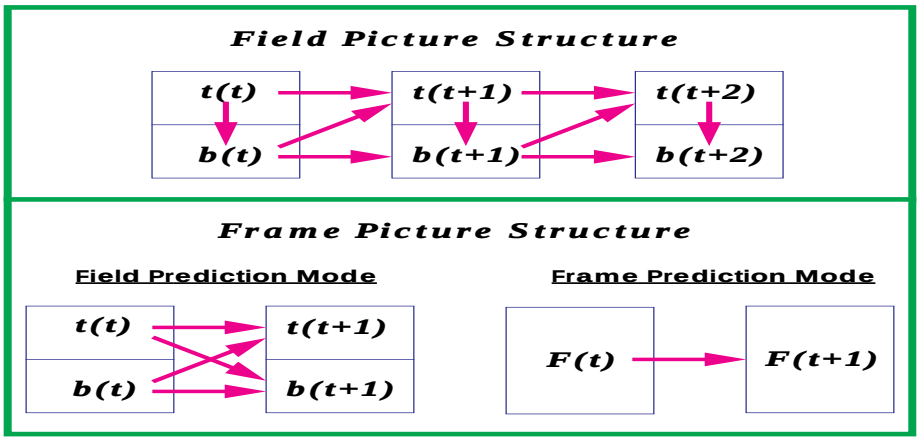


Figure 5: Field and Frame Picture Structures in Interlaced Video Coding

Figure 5 shows the two picture structures used in interlaced video coding. Field picture structure allows field predictions and even between fields of the same frame, but does not permit frame predictions. Frame picture structure allows either field prediction or frame prediction, selected on the macroblock level, but field predictions do not exist between fields of the same frame.

Converting two progressive stereo sequences into an interlaced video signal, the left images represent the top fields while the right images represent the bottom ones. Adaptive selection, on a per frame basis, for the appropriate picture structure in a stereoscopic sequence depends mainly on the disparity attributes between picture pairs, whether the images contain stereoscopic components or not. If the scene contains mostly non-stereoscopic objects, hence the left and right field images are essentially the same with low disparity values, then the preferred coding structure is frame picture.

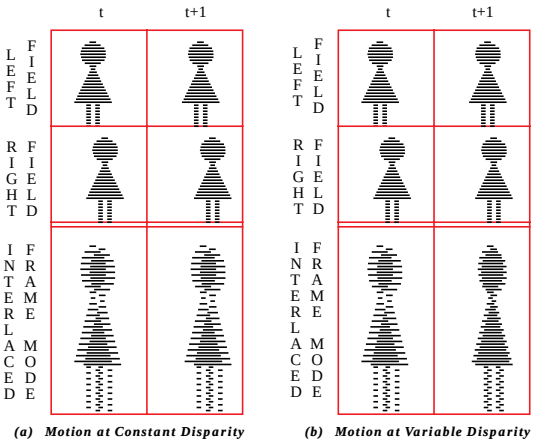


Figure 6: Dynamic Field or Frame Prediction Mode Selection of Interlaced Video Coding

For scenes containing stereoscopic objects, temporal disparity changes are examined, whether objects move at variable or constant depth. For images with limited object motion, frame picture structure is selected. At relatively constant disparity values where objects move in the direction of constant depth, the interlaced frame maintains the same composition in the temporal sequence, therefore predictions with the frame picture structure works better, as shown in Figure 6a. For objects changing in depth position, the disparity between the left and right images have changed as well. As a result, the better selection is the field picture structure, as demonstrated by Figure 6b.

Utilizing frame picture structure eliminates extra bit transmission of similar data from the left and right channels, thus contributing more bits to correct prediction errors. To our advantage, frame picture structure is selected most of the time because temporal scene movements tend to have continuous and small changes in disparity values.

# 6 Simulations & Results

## 6.1 Experimental Setups

Three stereoscopic video sequences courtesy of CCETT[4], *Discussion, Manege,* and *Train*, are used in our performance evaluations. Each stereoscopic signal consists of a left and a right interlaced sequence, at European CCIR 601 resolution (720x576 at 25 frames/sec) with 422 chroma format.

Our simulations are run using Columbia's MPEG-2 software package, which includes spatial and temporal scalability features. In addition, the existing scalability formats of MPEG-2 have been extended for stereoscopic sequences. Performance measures are based on the luminance signal to noise ratio (SNR) of the decoded reconstruction with respect to the original image.

In our simulations, equal bandwidths are allocated for the left and right stereo coding, even though existing studies have shown that the bandwidth of one stereo signal may be reduced by half with no visual degradation in the overall stereo perception [5]. Some efficient parameters in the basic MPEG-2 coder are chosen for all our experiments, and are fixed for comparison reasons. *M=3* and *N=15*. Motion estimation is found using the "brute force" full search method on the search area [-15,+15] horizontally and [-15,+15] vertically, with the criterion of minimizing block luminance mean squared error (MSE). Disparity estimation is similarly found utilizing the full search method limited to the rectangular search area [-30,+30] horizontally and [-3,+3] vertically, based on minimizing block luminance and chrominance weighted MSE.

## 6.2 Scalability Experiments

Our first simulation is the simulcast base case in which the scalability techniques are to be compared against. In the simulcast approach, the left and the right stereo sequences are coded independently of each other, and conforms to the non-scalable structure of the MPEG-2 standard.

In the experimental setups of spatial and temporal scalabilities, two layers of codings are performed, with the base layer coded first followed by the enhancement layer. In the base layer, the left stereoscopic video is encoded like the simulcast encoding. Afterwards, the right sequence is coded on the enhancement layer and utilizes the left sequence information from the lower layer.

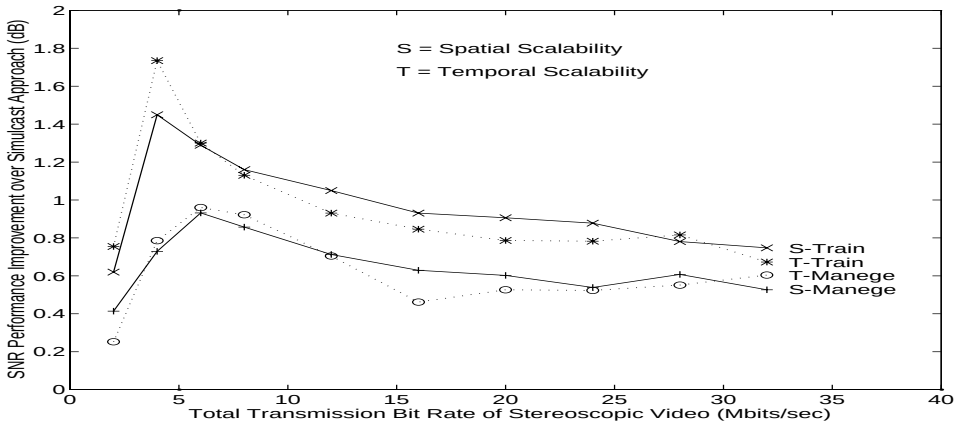


Figure 7: Average Luminance SNR Performance Improvement of the Proposed Scalability Approaches with respect to the Simulcast Method as a function of Varying Bandwidth

In Figure 7, performance comparisons between the simulcast, spatial scalability, and temporal scalability approaches are given for the *Manege* and *Train* sequence. With respect to the performance of the simulcast technique, spatial and temporal scalability offer significant SNR improvements. The average SNR performance differences between the scalability approaches and the simulcast method are plotted for a set of common transmission bit rates. From the diagram, it can be concluded that independent of the sequence and the transmission bandwidth, the spatial scalability approach achieves about the same performance improvement as the temporal scalability feature.

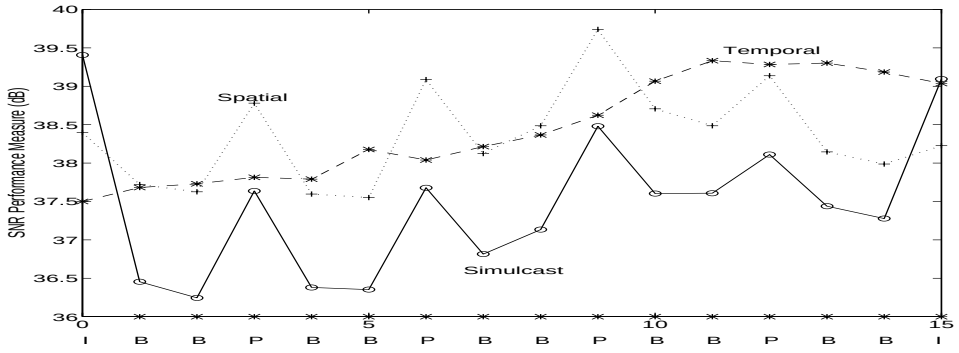


Figure 8: Luminance SNR Frame Performance Comparison of Three Approaches for one Group of Picture N=15 with Total Transmission Bit Rate=12 Mbits/sec using the *Train* Stereo Sequence

One important feature of MPEG-2 coding is the distribution of I, P, and B-picture types, and the corresponding bandwidth allocated to each type. In spatial scalability, the picture types of the enhancement frames follow the same types as the base layer sequence. On the other hand, in temporal scalability the frames in the enhancement layer are all B-picture types. Consequently, individual frame performance may not resemble the average sequence SNR performance presented above. Figure 8 illustrates this situation precisely. As expected, the performance curve using the spatial scalable approach follows the same pattern as the simulcast method, whereas the performance path using the temporal scalable approach is more linear.

## 6.3 Interlaced Picture Experiments

The proposed interlaced structure experiments operate on progressive stereoscopic sequences. Consequently, our original interlaced stereoscopic video sequences are converted to progressive formats by utilizing only the even fields of each frame. For the base experiment, the derived left and right progressive stereo sequences are coded using the simulcast approach. Afterwards, each left picture is interleaved with the corresponding temporal right picture to obtain an interlaced frame. The resulting interlaced sequence is coded by the main profile MPEG-2 encoder, taking advantage of the adaptive field and frame picture structure selection mentioned earlier. The crucial alteration required for this stereo experiment is to replace the interfield motion estimation routine of field prediction mode with a disparity estimation module, thus only involve changing the search area.

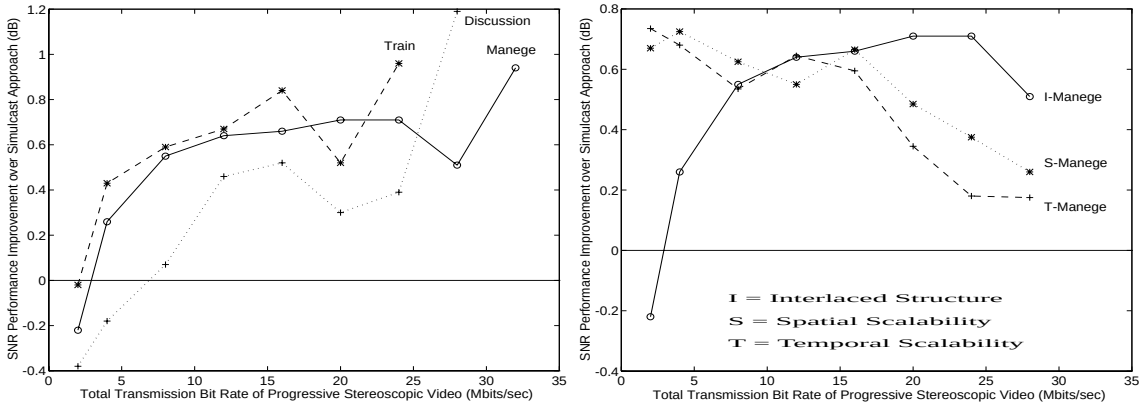


Figure 9: (1) Performance Improvements of the Proposed Interlaced Structure Approach with Different Sequences [left] & (2) Performance Comparison of the Three Approaches with the *Manege* Sequence [right] both with respect to the Simulcast Method as a function of Varying Bandwidth

From our simulations, it is found that the average luminance SNR performance using interlaced structure coding depends on the bandwidth transmission and the stereoscopic sequences. At low bit rates, all sequences perform worse as compared to the simulcast approach, but gradually improve as the bit rate is increased. Furthermore at high bit rates, performances simulated with the interlaced structure technique also surpass those with either the spatial and temporal scalability methods. These results were derived after analyzing experimental performance improvements with variable bandwidth transmissions, including those illustrated in Figure 9[left] & [right].

Although dynamic selection of field or frame picture structure is limited to a frame-by-frame basis, frame picture coding allows either field or frame prediction to be chosen on the macroblock level. For comparison, in the proposed scalability approaches the selection of spatial or temporal prediction is also made on the macroblock level, providing similar flexibility in prediction structure.

# 7 Conclusions

The spatial and the temporal scalability extensions of the MPEG-2 video coding standard and the field and frame picture structures of interlaced video are presented as compatible coding approaches for stereoscopic video sequences. Each coding technique combines concepts from simulcast and traditional stereo compensation methods, thus incorporating image information from the spatial and temporal domain.

The spatial and temporal scalability performances achieve higher luminance SNR reconstructions as compared to the simulcast approach, with the amount of improvements dependent on the total transmission bandwidth and the test sequences. Furthermore in the presence of channel errors, the two-layer scalability structure provides the receiver with graceful degradation of the stereoscopic signals. Consequently the "non-stereoscopic" view is consistently provided for, thus masking away possible disruptions in the communication channel.

Combining two progressive stereo sequences to form an "interlaced" video input, the resulting experiments demonstrate varying luminance SNR performances as compared to the simulcast technique, depending on the transmission bit rate and the test sequences. At low bandwidth transmission the proposed approach underperforms, but as the bandwidth increases, gradually outpaces the simulcast, spatial, and temporal results. Consequently, the proposed interlaced structure method provides a high performance main profile MPEG-2 coding approach for high bit rate transmission.

Based on the experimental results from the proposed approaches, we recommend coding all interlaced stereoscopic video sequences with the temporal scalability extension of MPEG-2. For stereo, temporal scalable video coding offers MPEG-2 compatibility, implementation convenience, graceful stereo image degradation, and high luminance SNR reconstructions.

A potential area for future stereoscopic video coding research is to improve the perceptual quality of the reconstructed sequence using temporal scalable coding. One method currently being investigated is a variable rate control for the enhanced B-pictures while taking advantage of temporal masking. A second approach of interest may be to allocate different bandwidths for the transmission of the left and right stereo sequences, so as to take advantage of the stereo masking effect of the human visual system. Another idea is to refine the disparity estimation process so as to obtain the true depth value, since we have unofficially demonstrated that a minor difference in disparity value effects the overall depth perception of objects in an image. In conclusion, we offer three MPEG-2 compatible coding techniques for 3D video on which future stereoscopic researches can build upon.

## Acknowledgments

The authors wish to thank Bruno Choquet from Centre Commun D'Etudes de Telediffusion et Telecommunications (CCETT) for providing us with the Digital Stereoscopic Imaging and Applications (DISTIMA) stereoscopic test sequences. It is also our pleasure to acknowledge Manfred Ziegler of Siemens for his friendly discussions and for providing us with numerous helpful references.

## References

1 M. Ziegler, et al. Digital Stereoscopic Television - State of the European Project Distima. 4th European Workshop on Three-Dimensional Television, Rome, Italy, Oct. 1993.

2 MPEG Committee Draft. Coding of Moving Pictures and Associated Audio. Mar. 1994.

3 A. Puri, L. Yan, B.G. Haskell. Temporal Resolution Scalable Video Coding. To appear in International Conference on Image Processing, Austin, Nov. 1994.

4 Stereoscopic Test Sequences DISCUSSION, MANEGE, and TRAIN shot within DISTIMA.

5 S. Pastoor, M. Wopking, J. Fournier, T. Alpert. Human Factors Data. DISTIMA Deliverable 26.