STEREOSCOPIC DELIVERY OF 3D CONTENT TO THE HOME

Walt Husak
Dolby Laboratories, Inc.

Abstract

3D is enjoying a renewed revival in the theatrical market due to the commercial success of 3D films released over the last several years. The technology advances in the cinema coupled with similar advances in consumer electronics promise to provide affordable 3D experiences to the home. Of the many ways to experience 3D, stereoscopic delivery is the most viable method due to the availability of displays and known production techniques. The delivery method described in the following paper addresses a cost effective method to provide stereoscopic content to the home using a tiered approach.

INTRODUCTION

Broadcast Distribution of 3D

The high costs of 3D production favor a distribution model where a premium can be charged to consumers which would offset to the increased production and delivery costs. The most obvious initial 3D providers would be satellite and cable companies where 3D could be packaged as a premium service, sold as pay per view, or delivered as video on demand. Service providers subsidize the customer's set top boxes and recoup those costs through monthly subscription fees. Likewise, network infrastructure costs are absorbed by the providers and recouped through monthly subscription fees.

Given the small amount of content that will be produced in 3D, coupled with the high cost of producing that content, an initially very low cost approach to delivery of 3D would be highly desirable to the service providers. Ideally, the upgrade costs should approach zero while at the same time the operators could collect additional revenue.

There are two methods service providers could use to deliver 3D to the home: a frame compatible method or a 2D compatible method (sometimes referred as a service compatible method). The frame compatible method (e.g. side-by-side) has the advantages of being able to use the current network infrastructure including set top boxes, with an acknowledged penalty of reduced resolution in the initial roll out of 3D services. In the future, the lost resolution can be provided to new decoders by means of a parallel enhancement stream. 2D compatible systems (e.g. MVC[1]) offer the advantage that the transmission consists of a 2D version with an enhancement layer to provide the 2nd eye view. However, to receive and decode 3D, new set top boxes have to be deployed.

Frame compatible 3D video signals closely resemble a normal video signal so few changes are necessary to accept and retransmit the signal from a network perspective. Similarly, current set top boxes can pass the frame compatible signal along to a 3D display for viewing by the subscriber. In the future, the operator can decide to upgrade the plant and set top boxes to pass full resolution signals as part of a larger upgrade cycle.

2D compatible systems offer full resolution upon deployment but require substantial changes to both the network and the set top boxes. Current networks are not designed to accept and process full bandwidth 3D signals so new production and processing equipment is necessary. Likewise, existing set top boxes can only understand the 2D version of the signal and therefore new set top boxes would need to be deployed to receive

3D. Any additional revenue will be absorbed by deploying additional individual customer's set top boxes.

Survey of frame packing methods

The coding performance and image processing considerations of the various decimation and frame packing approaches are important considerations in the selection of the most appropriate method. The order of operations to create a frame compatible image is to take the left and right pair and decimate those images so that each image contains half the samples of the original image. The sub-sampled images are then packed together to form a frame compatible image that is the same size as the original left or right image.

Table 1 lists the various sampling methods commonly used to decimate stereo images in preparation for frame compatible formatting. The first column shows the sampling method and the second column shows the direction the pixels are decimated.

Table 1
Frame Compatible Sampling Methods

Sampling Methods	Decimation Direction
Column decimation	Horizontal
Line decimation	Vertical
Quincunx	Diagonal

Table 2 lists the various packing methods commonly used to create frame compatible images. There are several combinations that are illogical such as column decimation and line interleave packing or line decimation and side-by-side packing. Most frame compatible systems make use of sampling and packing in the same direction. For instance, one could use column decimation with side-by-side packing or line decimation with over/under packing. Quincunx sampling can be used with several packing methods including side-by-side, over/under or checkerboard.

Table 2
Frame Compatible Packing Methods

Packing Methods	Description
Column interleave	Every other pixel
Line interleave	Every other line
Checkerboard	Pixel & line interleave
Side-by-Side	Horizontal half image
Over/Under	Vertical half image

Analysis of frame packing methods

Beyond the obvious combinations of sampling and packing, there are operational and performance issues that need to be considered when deciding which methods and combinations should be used. This section will discuss the various performance and operational considerations. The packing method is most sensitive to operational issues such as video preprocessing in the video encoder and image post-processing in the set top box and display.

Both checkerboard and line interleave frame packing suffer from processing techniques such as filtering and resizing. In both cases, the processing causes inter-pixel contamination resulting in ghosting at best and complete loss of the stereoscopic effect at worst. Since it is difficult to predict and control image processing throughout the video path, these methods are poor choices for the frame packing method. Side-by-side and over/under are less sensitive to these processing techniques.

The next issue to consider is interlaced video and its impact on the sub-sampling process. Interlaced video by its nature is vertically decimated. The two vertical decimation methods applied consecutively (interlacing and line interleave) compound the problem by doubly decimating the video. As an example, a 1080i60 signal has 540 lines per field. Decimating the image further would reduce the vertical resolution to 270 lines – equivalent to QVGA. By its nature, interlacing introduces vertical aliasing making reconstruction from images that have been

vertically decimated much more difficult. Side-by-side is not affected by interlacing and checkerboard falls in between.

Coding performance is another consideration when selecting the most appropriate frame packing and method. Figure 1 shows the relative coding efficiency using MPEG-4 AVC[2] of several sampling systems. Column interleave sampling with side-by-side packing and line interleave sampling with over/under packing have the same coding efficiency relative to each other across a variety of bit rates. However, quincunx decimation with checkerboard packing requires more than twice the bit rate compared to side-by-side or over/under for the same quality (PSNR).

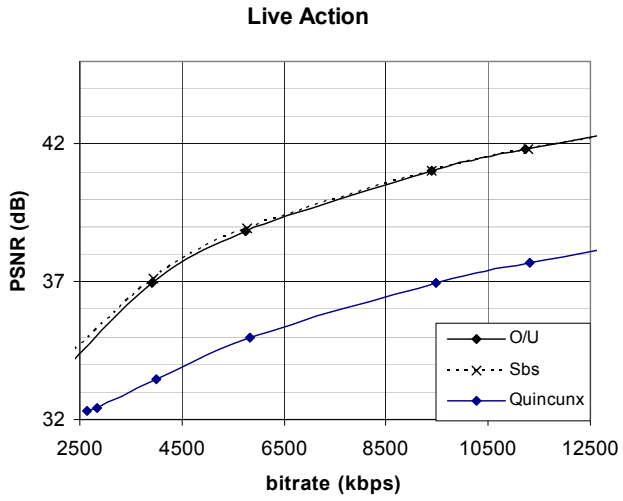

Figure 1
Relative coding efficiency of column interleave, line interleave and quincunx sampling.

It is clear from the analysis that side-by-side packing is the most appropriate base layer packing method for use with both progressive and interlaced formats. This is due to side-by-side being robust to interlacing, image processing and having superior coding performance to quincunx and the same (actually slightly better) performance than over/under.

One can also decimate in one format and pack in another. A popular method is to decimate in quincunx and pack in side-by-side. Figure 2 shows the performance comparison between side-by-side decimation and quincunx (checkerboard) decimation, both packed into the side-by-side format. For reference, normal AVC coding of 2D is also shown. The coding performance of quincunx decimation with side-by-side packing is lower than column decimation with side-by-side packing. At 10 Mbps, side-by-side decimation has a 2.5 dB performance advantage over quincunx decimation. When using the data from Figure 5, one sees that quincunx decimation with side-by-side packing has a coding efficiency that is superior to using quincunx decimation and checkerboard packing but inferior to side-by-side decimation and packing. In short, the coding efficiency of the combined method is roughly in between the efficiency of each method on its own. The mediocre efficiency is due to vertical and horizontal edges are no longer straight and require extra bits to code; also quincunx sampling with any sort of packing is sensitive to vertical resampling and color processing.

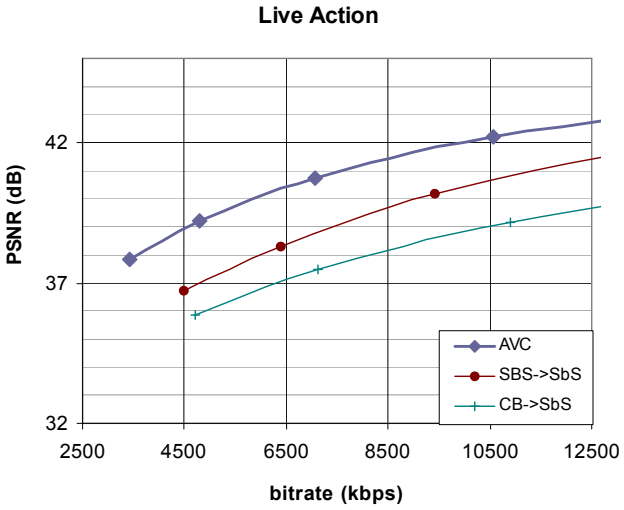

Figure 2
Relative coding efficiency of different sampling and packing formats

It should be noted that any frame compatible format has a decreased coding efficiency when compared to 2D video. This can be seen in Figure 2 where side-by-side sampling and packing (the most efficient frame packing method) and quincunx sampling with side-by-side packing both require a higher bitrate relative to the 2D AVC coding. This is due to the increased high spatial frequency image energy resulting from squeezing two images into the space of one image.

FULL RESOLUTION 3D

Full Resolution Frame Compatible

The ultimate goal is for content distributors to deliver full resolution stereoscopic signals to the home. As stated earlier, one method is the 2D compatible service which requires replacement of set top boxes in the home and an upgrade of network infrastructure. Frame compatible systems can also support full resolution by sending metadata that can recreate the full resolution using common layering techniques. Dolby has introduced such a system to meet the needs of the broadcasters.

Dolby's 3D system is a two tiered 3D delivery system that allows low cost initial deployment using a frame compatible base layer, with an available enhancement layer allowing a path to full resolution. Side-by-side has been chosen for the reasons discussed above, as well as widespread acceptance by 3D display vendors.

Figure 3 shows a functional overview of Dolby's 3D Full Resolution Frame Compatible delivery solution. A stereo pair is multiplexed into two frame compatible images one using one set of pixels and the other using the complementary pixels. The

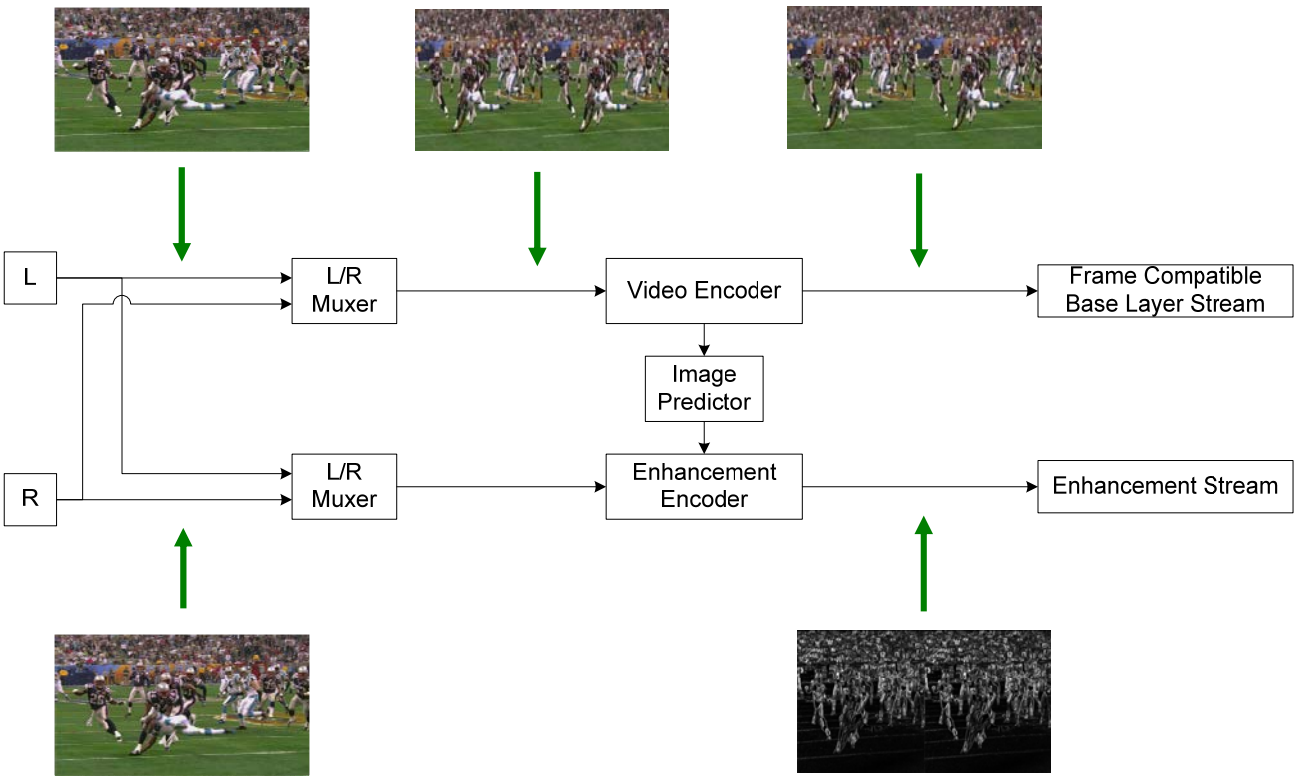

Figure 3
System Overview

first frame compatible image is compressed using MPEG-4 AVC as if it were normal video image set. The complementary image is used as the basis for the enhancement encoder. The enhancement encoder uses information from the base layer encoding process to predict the enhancement layer. By making use of redundant information in the base layer, the total amount of data in the enhancement layer is greatly reduced.

The frame compatible base layer selected by Dolby is the side-by-side method for decimation and the packing. Dolby offers the option to use a variety of pre-decimation low-pass filters in order to provide the optimum performance for a given piece of content and bitrate. The side-by-side packed video stream is compressed and transmitted using the standard service provider work flow. In the case of legacy MPEG-2[3] video delivery, only the base layer would be transmitted allowing the use of legacy MPEG-2 set top boxes. For operators that have enabled MPEG-4 AVC (H.264), the base layer would be encoded using AVC with an option to also encode an enhancement layer.

In a video compression system, most of the coding efficiency is realized by using prediction techniques to recreate pixels. Using a split filter system (complementary filtering of the high frequency component from the base layer) or a difference signal from the base layer and compressing it using standard coding techniques suffers from several fundamental weaknesses. A simple high frequency split system suffers from the two video codecs (high and low frequency respectively) operating open loop relative to each other. Unless a very high bitrate is used for the enhancement layer, the recreated pixels will not be phase coherent with the source nor will the bit depth be adequate for the high frequency information. A simple differencing system - where the enhancement layer is subtracted from the base layer - is limited in performance gain due the simplicity of the prediction. In addition, production tools such as keystone, floating windows and occlusions cause a significant amount of energy to be coded as residual data.

Dolby's full resolution method predicts the enhancement layer from the base layer and makes use of redundant information between

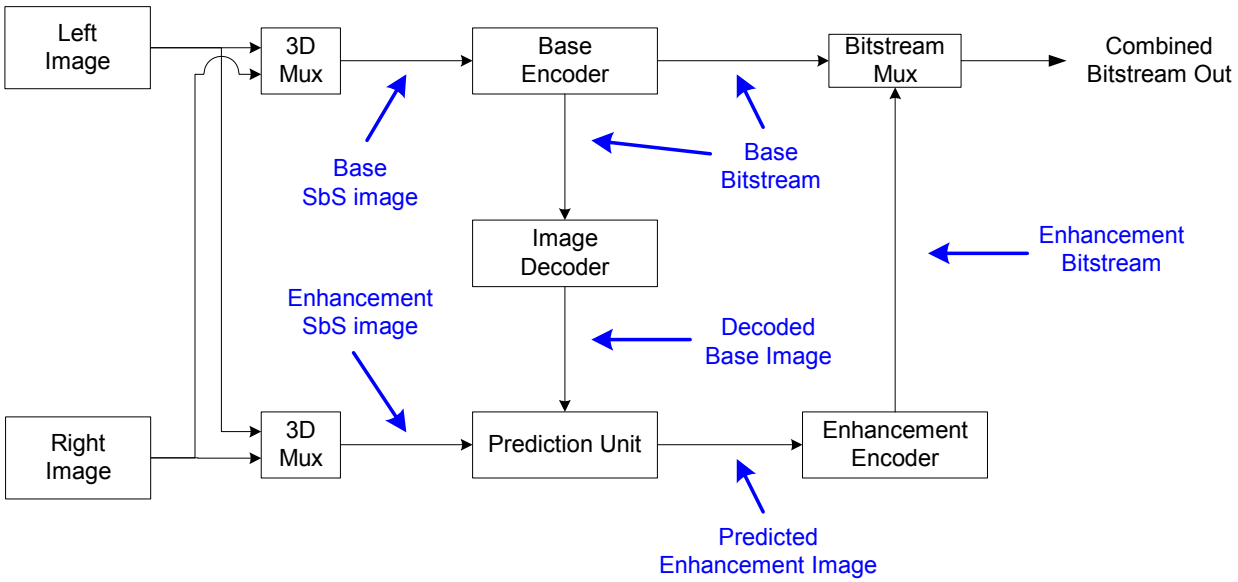

Figure 4
Dolby prediction system

layers to reduce the resulting bit stream. Figure 4 illustrates how the prediction between layers is accomplished. A stereo pair is presented to a pair of 3D multiplexers. The multiplexers filter, decimate and format the stereo images into the side-by-side format. The base layer uses one set of pixels from the original image set and the enhancement layer uses the complementary pixels. The base layer is then encoded using a standard video encoder. The resulting base bit stream is applied to the bit stream multiplexer and is also decoded locally.

The locally decoded base image is then used to predict the enhancement image. The base and the enhancement side-by-side images are very similar to each other due to both being taken from the same original images but merely offset by one pixel. The predicted enhanced image largely contains the differences between the base side-by-side image and the enhanced side-by-side image. The enhanced side-by-side image is then coded and the resulting bit stream is combined with the base layer bit stream for delivery to the decoder.

An important point in using the locally decoded image is the results of coding decisions made by the base encoder are automatically applied to the enhancement layer. This overcomes the weaknesses of using two separate open loop codecs for the base and enhancement layers. Figure 5 shows the performance gain using a predicted resolution enhancement system. At 7.5 Mbps, the enhancement system adds 0.4 Mbps while increasing the quality by 3.25 dB. As a percentage of the original bit rate, 5.4% overhead yields a doubling in video quality.

The enhancement layer can be delivered as a compressed stream along with the base layer by including MPEG-4 structures called Network Abstraction Layer (NAL) units that are specific to the enhancement layer. Legacy decoders should ignore the enhancement layer NAL units and decode the base layer as if it were a standard 2D video stream, and output the side-by-side image. Another means to deliver the enhancement layer is to use a secondary video stream with its own PID within the MPEG-2 transport stream[4]. Decoders that are enabled to decode the Dolby solution will extract the enhancement layer and decode the data to recreate the original full resolution video.

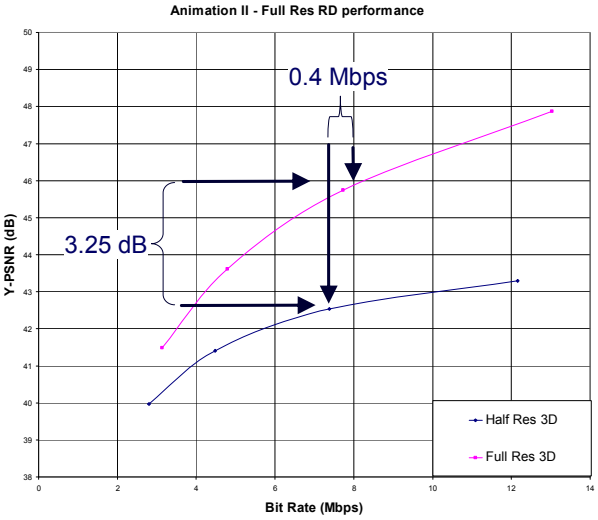

Figure 5
Enhancement layer performance

TESTING

Objective Performance

An important feature of the Dolby 3D system is the compression efficiency. A more efficient video compression system allows content to occupy a smaller part of the service multiplex than a compression system that is less efficient while still maintaining the same quality. Another way to consider the effects of a more efficient encoder is a higher quality image can be transmitted within the same bitrate.

This section will summarize tests that were performed. The following tests were conducted using 23,000 frames from four

sequences. Three sequences were live action and the fourth was an animation. All tests were performed using identical quality for both eyes in order to have an accurate comparison between methods.

Figure 6 shows the relative performance of the full resolution system using Peak Signal to Noise Ratio (PSNR). PSNR is a method of testing widely used for comparisons between different bitrates or toolsets that takes the Mean Squared Error (MSE) of each pixel and averages the information across the image as a root mean square.

The data has been normalized for easy comparison between delivery methods. The first bar is the 2D equivalent delivery bitrate and has been fixed at 100%. Not surprisingly, a 3D simulcast (one channel per eye) is twice the data rate of the 2D signal. The side-by-side coded data is 35% greater than the 2D bit rate due to the increased high frequency content resulting from squeezing two images into the space of one image. The added enhancement data is 6% more bitrate for a total of 41%. The overhead for the MVC signal is 81% more than the 2D signal. The MVC bitrate overhead is highly dependant on content and can range from 40% for animations to as much as 90% for live action material. The ability of MVC to use inter-view prediction is based on how well the views in the stereo image pair are correlated.[5]

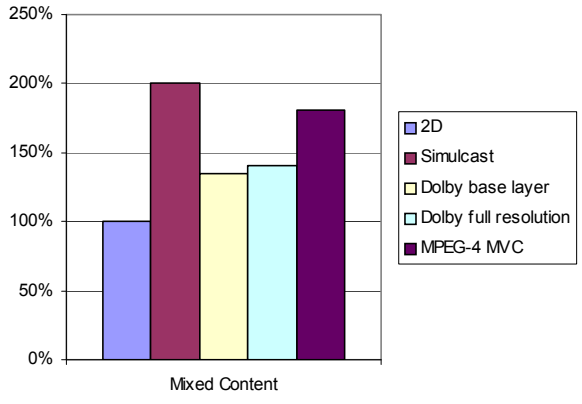

Figure 6
Relative performance of the Dolby full resolution 3D system

Subjective Performance

PSNR provides the engineer a simple and rapid test for comparing similar codecs, tools and content. It is difficult to use PSNR across substantially different content or coding systems due to the different artifacts that may manifest themselves specific to those codecs. For instance, it is valid to use PSNR to compare a number of AVC based codecs but mixing a wavelet codec and a block based codec such as AVC would limit the functionality of the PSNR metric. An additional shortcoming of PSNR is the inability of PSNR to consistently track a real viewer's Mean Opinion Score (MOS) when they are rating the quality of a subjective test across a variety of content. Nevertheless, PSNR is a simple test that is widely understood in the image processing community without having to run complicated subjective tests for each codec, bitrate and piece of content.

Dolby performed a series of subjective tests to understand the real world performance of stereoscopic delivery systems. The test used ITU-R Rec. 500[6] as a reference for designing the test. Modifications were made to the procedures since Rec. 500 did not thoroughly address stereoscopic subjective testing. The tests were conducted as a double blind quality rating test using MOS values obtained from both expert and non-expert viewers. The test was broken down into two stages. The first stage was a ranging exercise that compared 2D broadcast bitrates with 3D broadcast bitrates. The second stage used the results of the first stage as a baseline and compared several different coding techniques to understand what broadcasters may expect when deploying stereoscopic delivery systems.

The first stage was performed by presenting the viewers with a clip coded at multiple bitrates and the viewer was told to select the quality that most closely represented the target quality for their delivery system. The test was conducted twice – once for 2D content and once for 3D content. The data was coded as a standard AVC 3D simulcast (one channel per eye) with the 2D content represented by the left eye view. Each viewer's MOS was tabulated and the results were normalized to the 2D MOS values.

Figure 7 is a chart of the results from the 2D to 3D comparison. Due to the normalization, the 2D data is fixed at 100% of the bitrate. Intuitively, one expects the 3D results to be approximately twice the 2D results. The results from the subjective tests show that viewers did not find coding artifacts in 3D as objectionable as coding artifacts in 2D

In some cases – such as the animation – the 3D simulcast actually required fewer bits for the simulcast transmission than the 2D transmission. The movie sequence and the concert sequence showed slightly higher bitrates for 3D on the order of 40% and 25% respectively. The football sequence is an anomaly but is included in the chart for completeness. The right eye contained a source artifact that was not seen during the 2D presentation due to using the left eye as the reference. Subsequent testing has shown the Football sequence behaving similarly to the other sequences. In addition, several other clips were tested that also showed less than 50% overhead for 3D simulcast over 2D.

The effect of 3D content scoring higher for a given bitrate is commonly referred as "stereo masking". This phenomenon can be seen in Digital Cinema[7] where the maximum bitrate for 2D delivery to the theater is the same as the maximum bitrate for 3D delivery even though the two views are sent as two completely separate image streams.

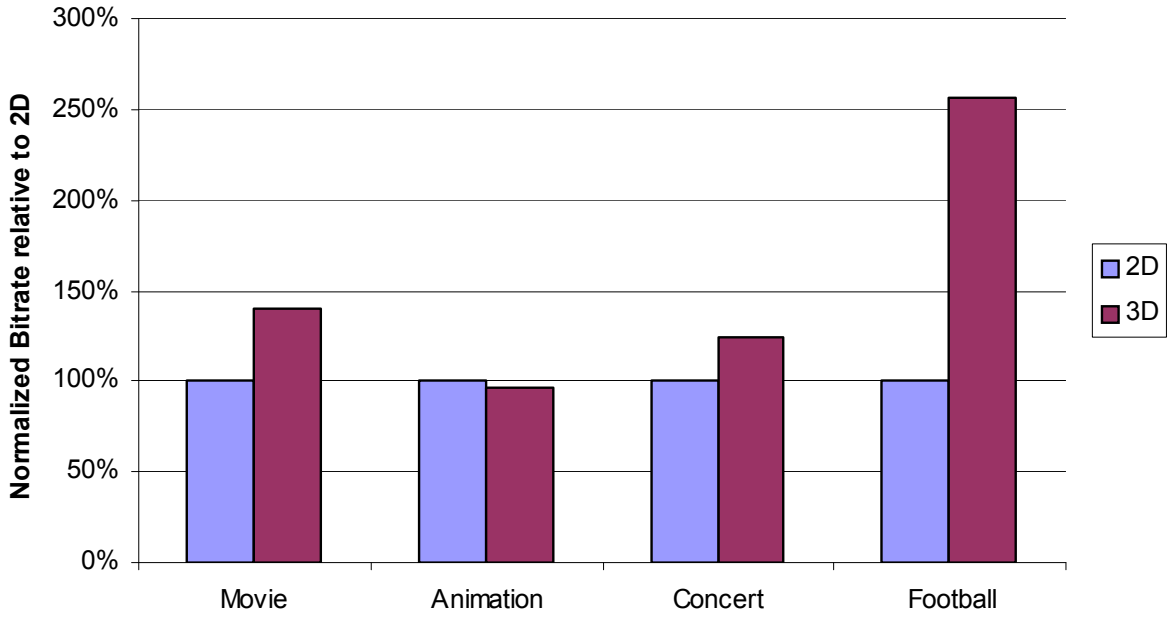

Figure 7
Subjective comparison of 2D and 3D codings

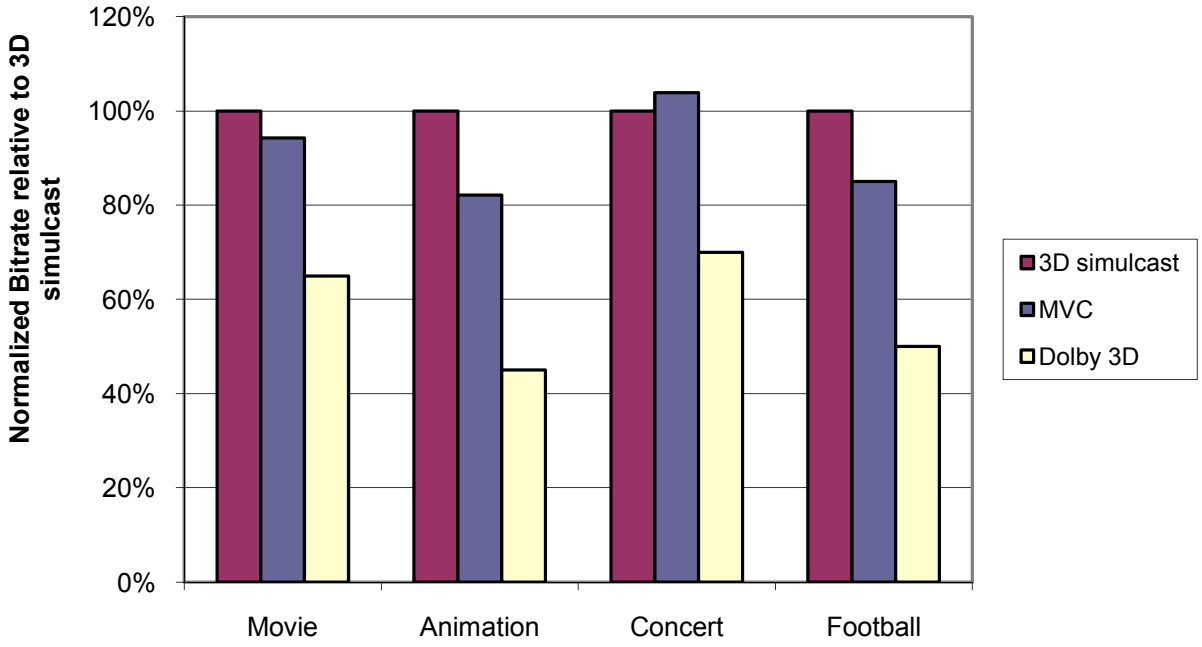

Figure 8
Subjective comparison between 3D coding techniques

Stereo masking creates an interesting dilemma for 2D (or service) compatible systems. The most critical aspect of the 2D compatible systems such as MVC or 2D+Delta is the ability to extract a 2D signal from the service. Forgetting for the moment that most content producers have stated 2D productions will be completely separate from 3D productions; the coding strategy for the entire stereo signal must use a 2D bitrate that meets the quality needs of current broadcast. This means one cannot use stereo masking to tailor their 2D compatible bitrate to minimize the impact to their service channel.

As an example, sports programming would be the most challenging content to deliver. This is because sports programming is by its nature, live content. The second view needs to use as much as 90% of the left eye's data rate to send equal quality video to both eyes. One could send asymmetric quality between the two eyes (e.g. sending higher quality to one eye and lesser quality to the other eye), but the effects of eye dominance between viewers is not well understood. People that are left eye dominant would be well served while people that are right eye dominant would receive a sub-standard image.

Figure 8 shows the results of the second stage of the subjective tests. The second stage was conducted as a Double Stimulus Comparison Scale (DSCS) using the results from the first test as the baseline reference. The three systems compared were a 3D simulcast where each eye is coded separately, a 2D compatible system represented by MVC and the Dolby full resolution frame compatible system. Again, the results were normalized – this time to 3D simulcast.

As expected, the MVC system requires nearly the same bitrate as the simulcast except when coding animations. The concert footage actually scored higher that simulcast due to the content. The lights flashing and the stage background caused significant differences

between the two eyes making the prediction between eyes difficult to achieve. This resulted in most frames being coded as two separate bit streams with little interdependency. While this particular clip was more stressful in that regard than a typical concert, the differences in lighting due to the flashing and spinning lights will limit the amount of prediction between views.

The viewer MOS scores showed the frame compatible system having equivalent quality with substantially lower bitrates than either simulcast or MVC. The bitrates were around 50% of the 3D simulcast which from stage one we know is just slightly higher than bitrates used for 2D services. One point to note is the lower bound of the subjective test did not exercise the video codec. In other words, the values shown in this paper are conservative numbers for delivery of 3D.

SUMMARY

In this paper we examined delivery of 3D content using 2D compatible systems and frame compatible systems. Frame compatible systems allow a broadcaster to deliver 3D using existing set top boxes and network infrastructure. There are several methods of sub-sampling and packing to create the frame compatible image, although several of them suffer from operational issues. Side-by-side offers a simplified approach that codes with the same compression efficiency as over/under albeit without operational limitations imposed by interlaced video. Quincunx sampling offers no additional benefit but adds unneeded complexity.

A means to migrate to full resolution using predicated layering techniques was discussed allowing the operator to deploy a backward compatible system serving existing set top boxes with frame compatible 3D and new set top boxes with full resolution. The method shown allows the operator to upgrade their set top boxes and network infrastructure over time. The use of advanced prediction specialized for frame compatible 3D overcomes weaknesses such as open loop codecs and limitations in complementary filter systems. The relative overhead for the enhancement layer is between 5-10% and also increases the measured by quality over 3 dB.

Finally, objective and subjective test results were discussed. Stereo content was shown to require substantially lower bitrates than intuitively imagined dues to stereo masking. Furthermore, 2D compatible systems were shown to have a significant Achilles Heel in regard to needing to fix one eye to equivalent 2D quality while at the same time requiring high enhancement bitrates for the second eye due to eye dominance. The enhanced frame compatible system required substantially lower bit rates than MVC and 3D simulcast while delivering equivalent quality.

BIBLIOGRAPHY

[1] G.J. Sullivan, et al, "Text of ISO/IEC 14496-10:2009/FDAM 1 Constrained baseline profile, stereo high profile, and frame packing arrangement SEI message," Doc. N10707, London, UK, July 2009

[2] ISO/IEC 14496-10 - MPEG-4 Part 10, "Advanced Video Coding"

[3] ISO/IEC 13818-2 - MPEG-2 Part 2, "Video Coding"

[4] ISO/IEC 13818-1 - MPEG-2 Part 1, "System"

[5] W. Gish & C. Vogt, "MVC compression coding for 3D applications", presented at the 2009 SMPTE technical conference October 28, 2009.

[6] ITU-R Recommendation BT.709-11, "Methodology for the subjective assessment of the quality of television pictures"

[7] SMPTE 429-10-2008, "D-Cinema Packaging — Stereoscopic Picture Track File"