An Extended H.264 CODEC for Stereoscopic Video Coding

Balamuralii Balasubramaniyam, Eran Edirisinghe, Helmut Bez
Department of Computer Science, Loughborough University, UK.

ABSTRACT

We propose an extension to the H.264 video coding standard, which is capable of efficiently coding stereoscopic video sequences. In contrast to previous techniques, the proposed Stereoscopic Video CODEC uses a single modified H.264 encoder and a single modified H.264 decoder in its design. The left (reference) and right (predicted) sequences are fed alternatively to the encoder. The modified H.264 encoder uses a Decoded Picture Buffer Store (DPBS) in addition to the regular DPB of the original H.264 encoder. An effective buffer management strategy between DPBS and DPB is used so that the left sequence frames are coded only based on its previously coded frames while the right frames are coded based on previously coded frames from both left and right sequences. We show that the proposed CODEC has the capability of exploiting *worldline* correlation present in stereo video sequences, in addition to the exploitation of joint spatial-temporal-binocular correlation. Further we show that the coded bit stream fully conforms to a standard H.264 bit-stream and a standard H.264 decoder will be able to effectively decode the left video stream ignoring the right. We provide experimental results on two popular test stereoscopic video sequences to prove the efficiency of the proposed CODEC.

Keywords: Stereoscopic video, H.264, stereo imaging, joint motion-disparity estimation, worldline correlation.

1. INTRODUCTION

Led by the ever demanding nature of viewer requirements, digital television has progressed during the recent past to provide better quality, more efficient, functional and interactive services. This effort has been well supported by the international research activities in video coding, particularly by the MPEG and JVT standardization activities. New standards such as MPEG-4 Visual Texture Coding, MPEG-4 Scalable Video Coding and H.264 [1] (also known as MPEG-4-Advanced Video Coding), have been developed to cater for the above needs. Applications of these video CODECs have extended from digital television to video content provision on handheld devices, DVD recordings, surveillance, security systems etc.

The term 'video' is generally associated with 'monoscopic video', which is a collection of single camera snapshots (images) of a scene, taken at regular time intervals. Alternatively one could use two camera's or a so called stereoscopic camera to capture video, seen by the two eyes of a viewer. If these views are provided separately to the eyes, the human brain is able to fuse the two views to produce a scene description with depth sensation. This technology has been in use in comparatively primitive ways for over 4 decades. With the recent advances in auto-stereoscopic display devices and fully immersive virtual environments, the use of stereoscopic image and video technology is tipped to be the future of digital television and entertainment industry such as digital cinema, computer games, immersive experiences etc. However, the large amount of data that has to be coded and transmitted in stereo video (typically, twice compared to mono-video), provides a challenge for the international research community. This has been well acknowledged within recent research attempts and international standardization activities. Within MPEG-2 (e.g. multi-view profile), MPEG-4 VTC (e.g. depth coding in I3D) numerous attempts have been made to extend the monoscopic video coding technology to its stereo equivalent. The latest such attempt is within H.264/MPEG-4 AVC, which proposes the use of interlaced video for stereoscopic video coding. Unfortunately none of these schemes addresses stereo video coding in terms of exploiting all four forms of redundancy that is most likely to be present within a stereo video sequence, namely, spatial (within frames), temporal (within individual streams), binocular (between streams) and worldline (between current and previous reference frame from main stream) [7]. Nevertheless combined exploitation of all of these correlations would lead to better coding performance. In this paper we propose a stereoscopic video coding extension to H.264 which is capable of achieving efficient stereoscopic video coding with added functionality to decode both monoscopically and stereoscopically. We propose the effective use of existing H.264 technology such as variable block size prediction, block skipping with non-zero motion vectors, and decoupling of parameter sets from video sample data in the above attempt.

Stereoscopic Displays and Virtual Reality Systems XII, edited by Andrew J. Woods, Mark T. Bolas, John O. Merritt, Ian E. McDowall, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5664 © 2005 SPIE and IS&T · 0277-786X/05/$15

The rest of the paper is organized as follows. In section II a brief introduction to H.264 is given, in which we highlight the salient features of the standard, which we subsequently use in stereo video coding. Section III introduces the proposed CODEC in detail. In section IV we critically analyze the performance of the proposed algorithm. Finally Section V concludes with an insight to possible future improvements/extensions to the proposed CODEC.

2. RELEVANT CONCEPTS OF H.264 VIDEO CODING

H.264 (also called as MPEG-4 Advanced Video Coding) is the latest video coding standard, developed by the Joint Video Team [JVT] of, ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. It achieves a significant improvement in rate-distortion efficiency relative to the previous standards such as MPEG-2, H.263, MPEG-4 VTC, MPEG-4 SVC [2]. In contrast to MPEG-4 VTC/SVC standards and similar to the much older MPEG-1, MPEG-2, H.261 and H.263 standards, this standard specifies a non-object based coding strategy for image sequences. For ease of reference, few key features/concepts of H.264 standard, which are effectively used in the design of the proposed stereo video CODEC are described bellow.

2.1 Network Abstraction Layer [NAL]:

This is the top most layer of a H.264 bit-stream. It is designed in order to provide "network friendliness" to enable simple and effective customization of the use of Video Coding Layer (VCL) data for a broad variety of systems such as RTP/IP, File formats [E.g. ISO MP4] etc. The NAL consists of so-called ***NAL units*** and ***Parameter Sets***. The coded video data is organized into *NAL units*, each of which is effectively a packet that contains an integer number of bytes. The first byte of each NAL unit is a header byte that contains an indication of the type of data in the NAL unit, and the remaining bytes contain payload data of the type indicated by the header.

Further NAL units are classified into VCL NAL units that contain the data that represents the values of the samples in the video pictures, and the non-VCL NAL units that contain any associated additional information such as *parameter sets* (important header data that can apply to a large number of VCL NAL units) and *supplemental enhancement information* (timing information and other supplemental data that may enhance usability of the decoded video signal but are not necessary for decoding the values of the samples in the video pictures). A parameter set contains information that is expected to rarely change and helps in the decoding of a large number of VCL NAL units. There are two types of parameter sets:

- Sequence parameter sets [SPS], which apply to a series of consecutively coded video pictures called a *coded video sequence (CVS)*;
- Picture parameter sets [PPS], which apply to the decoding of one or more individual pictures within a CVS.

The SPS and PPS mechanism decouples the transmission of infrequently changing information from the transmission of coded representations of the values of the samples in the video pictures. Each VCL NAL unit contains an identifier that refers to the content of the relevant PPS and each PPS contains an identifier that refers to the content of the relevant SPS. In this manner, a small amount of data can be used to refer to a larger amount of information without repeating that information within each VCL NAL unit. Sequence and picture parameter sets can be sent well ahead of the VCL NAL units that they apply to, and can be repeated to provide robustness against data loss. In some applications, parameter sets may be sent within the channel that carries the VCL NAL units ["in-band" transmission]. In other applications it can be advantageous to convey the parameter sets "out-of-band" using a more reliable transport mechanism than the video channel itself or could be hard-wired so that no parameter sets need to be transmitted, for specific well defined applications.

2.2 Variable Block Size Based Prediction and R-D Optimization:

H.264/AVC allows variable block size motion estimation/compensation. When temporal (rather than spatial) prediction is used, i.e. for P or B frames, motion can be estimated at a 16x16 macroblock level or by partitioning the macroblock into smaller regions of luma size 16x8, 8x16, 8x8, 8x4, 4x8, or 4x4 (see Fig. 6). A distinction is made between a *macroblock partition*, which corresponds to a luma region of size 16x16, 16x8, 8x16, or 8x8, and *sub-macroblock partition*, which is a region of size 8x8, 8x4, 4x8, or 4x4. When (and only when) the macroblock partition size is 8x8, each macroblock partition can be divided into sub-macroblock partitions. A distinct motion vector can be sent for each sub-macroblock partition. The motion can be estimated from multiple pictures that lie either in the past or in the future in display order. However the selection of which reference picture is used is done on the *macroblock partition level* [16x16,

16x8, 8x16, 8x8] (so different *sub-macroblock partitions* [8x8, 8x4, 4x8, 4x4] within the same macroblock partition will use the same reference picture). A limit on number of pictures used for the motion estimation is specified for each Level.

2.3 Supplemental Information:
In addition to basic coding tools, the H.264 standard enables sending extra supplemental information along with the compressed video data. Within the standard it is handled by "supplemental enhancement information" (SEI) or "video usability information" (VUI).

2.4 Block Skipping and MV prediction:
If a macroblock has motion characteristics that allow its motion to be effectively predicted from the motion of neighboring macroblocks, and it contains no non-zero quantized transform coefficients, then it is flagged as skipped. This mode is identified as the Skip mode. Note that, when a block is skipped neither the transformed coefficients nor motion data are transmitted. In contrast to prior standards, non-zero motion vectors can be inferred when using the Skip mode in P slices.

3. PROPOSED H.264 STEREO CODEC

3.1 Monoscopic/Standard H.264 CODEC
Figure 1 illustrates a typical block diagram of the standard H.264 encoder. Note that Coder Control signal flow is denoted by dashed lines whereas video data flow is denoted by solid lines. In this paper we assume that the reader is aware of the basic steps involved in a video coding process. Therefore the basic operational steps of diagram 1 are not discussed. Rather the discussion is limited to those processes within the H.264 encoding that need to be altered in the design of the proposed stereoscopic video CODEC. We refer readers interested in the detail design/techniques of H.264 to [1,2].

Figure 1, illustrates that every coded frame of an input video sequence is locally decoded at the encoder end and is stored in a so-called Decoded Picture Buffer (DPB) after going through a de-blocking process. In contrast to previous standards, in H.264 standard compliance encoders, the DPB can hold multiple frames at a given time (typically 5 frames). These multiple frames of the DPB are subsequently used in the motion estimation and compensation of the next predicted (of type P or B) frame.

The input un-coded (raw) frames to a video CODEC are divided into independently coded groups, named Group of Pictures (GOP). A GOP typically consists of a single Intra (I) coded frame and a number of Predicted (P) and Bi-directionally Predicted (B) frames.

Figure 2 illustrates a group of pictures that consists of a frame type order, IBBP. I frames are intra coded exploiting the spatial redundancy within the frame itself and acts as anchor frames to reduce the possibility of progressive loss accumulation due to repeated predictive coding. P frames are coded in relation to previous I or P frames whereas B frames can be coded in relation to I, P or other B frames located sequentially before or after, within the video sequence. Note that due to the latter case, the coding order of frames would differ in comparison to their original order in the input video sequence (or display). In figure 2, the coding order is given within brackets and the numbering shows a mismatch with the input/display order. For example the P frame is coded soon after the I frame. The second B frame in the GOP is coded as the fourth frame, based on I, P frames and the first B frame. In predicting the P frame, the prediction is done based on the I frame, witch is time sequentially dependent on a frame that is three frames away from itself. However in predicting the second B frame, adjacent frames in the time sequence is selected. This often leads to the B frame being predicted at higher accuracy as compared to the P frames.

3.2 Stereoscopic/Extended H.264 Video CODEC
Stereoscopic video coding attempts made within and outside the standardization bodies [3-6] can be categorized into three different coding architectures as illustrated in figure 3. The horizontally oriented arrows illustrate motion estimation and vertically oriented arrows illustrate disparity estimation, with the arrow-head pointing towards the frame

which is predicted from the other. The joint motion-disparity estimation architecture (fig 3(c)) is preferred over the Simulcast (fig 3(b)) and Compatible (fig 3(c)) methods due to improved coding performance resulting from the joint estimation of spatial, temporal and binocular redundancy.

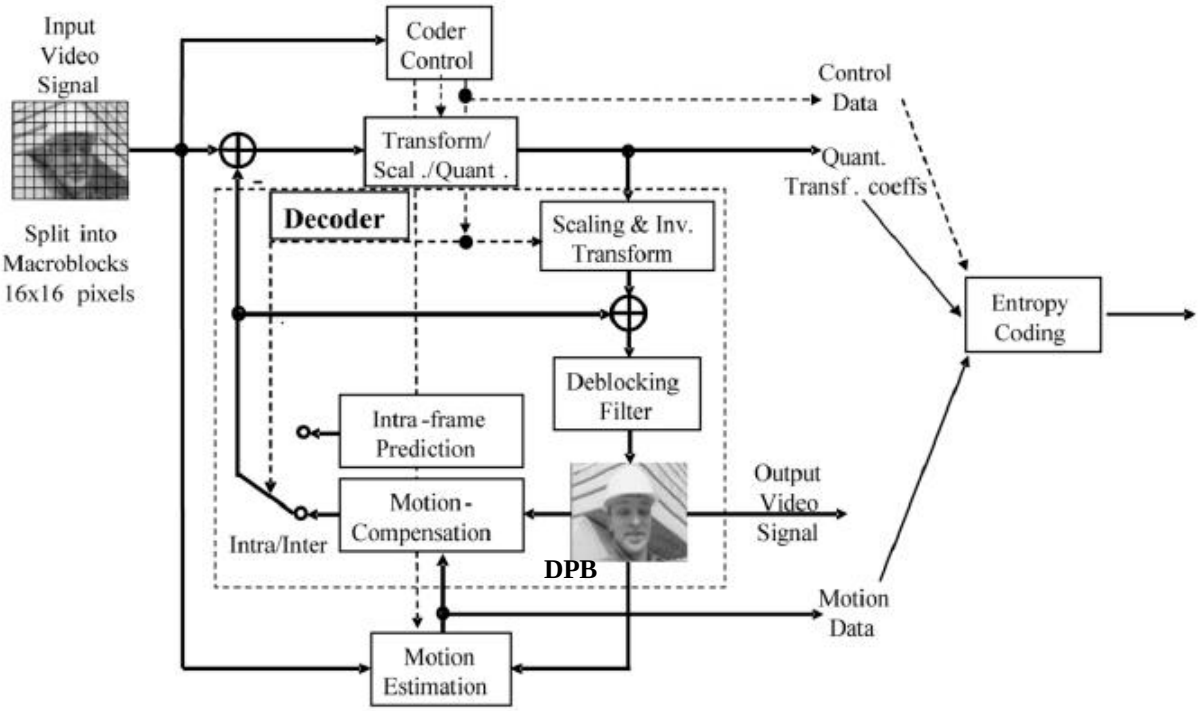


Figure 1. H.264 Encoder

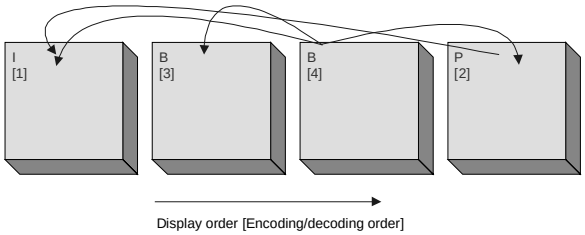


Figure 2. GOP in Video Coding

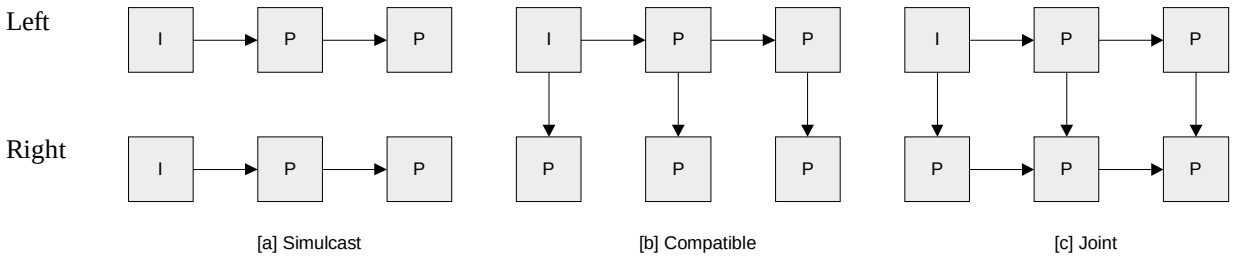


Figure 3. Stereoscopic Video: Typical Coding Architectures

However there is a fourth form of redundancy, i.e. *worldline* correlation that is likely to be present within a stereoscopic video sequence. As depicted in figure 4, worldline correlation corresponds to the similarity of the frame t = T of the right (predicted) sequence to that of frame t = T – dT in the left (reference) sequence. Note that under joint motion-disparity estimation (fig 3(c)) the frame t = T of the right sequence is coded only in relation to the frame t = T in the left sequence and/or frame t = T – dT in the right sequence. Therefore the joint motion-disparity estimation architecture of fig 3(c) does not exploit worldline correlation. Within the design of the proposed stereoscopic extension to H.264, we aim to exploit worldline correlation in addition to jointly exploiting spatial-temporal and binocular correlation.

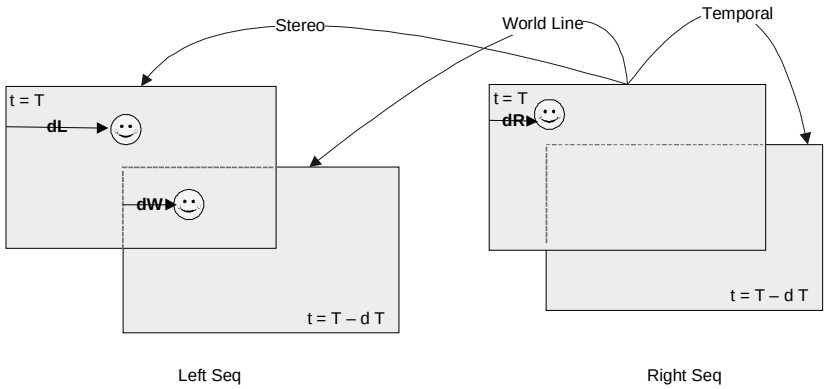


Figure 4. *Worldline* correlation in stereo video

3.2.1 Design considerations:
The proposed stereoscopic video coding strategy is designed based on four main requirements namely:

1. A receiver with a standard H.264 decoder should be able to decode the left sequence.
2. *Worldline* correlation should be exploited jointly with spatio-temporal-binocular redundancy.
3. A single, extended H.264 encoder/decoder should be used for encoding/decoding both left and right sequences.
4. The coded output bit stream should conform to H.264 standard.

Within our proposed design, we meet all four of the above requirements by altering the DPB management strategy of the standard H.264 compliance encoder/decoder described in section 3.1. In addition we effectively make use of the Video Usability Information (VUI) of Supplementary Enhancement Information (SEI) of the resulting H.264 compliance bit stream to indicate the presence of stereoscopic data as against monoscopic video data and to indicate ownership of specific frames to left/right sequences.

3.2.2 Frame organization:
Figure 5 illustrates an example of a typical frame organization/layout of the proposed stereoscopic video CODEC. The superscript ‘s’ (for *stereo*) in the B and P frames of the right sequence is used to differentiate them from corresponding frames of the reference sequence in subsequent discussions. The arrows indicate predictive coding with the arrow-head pointing towards the reference frame. However to maintain clarity of the diagram, the reference frames are illustrated only for some selected frames. Note that, I frames are not used in the coding of the right sequence. Instead the first frame of the right sequence is a P frame, predicted based on the first frame (an I frame) of the left sequence.

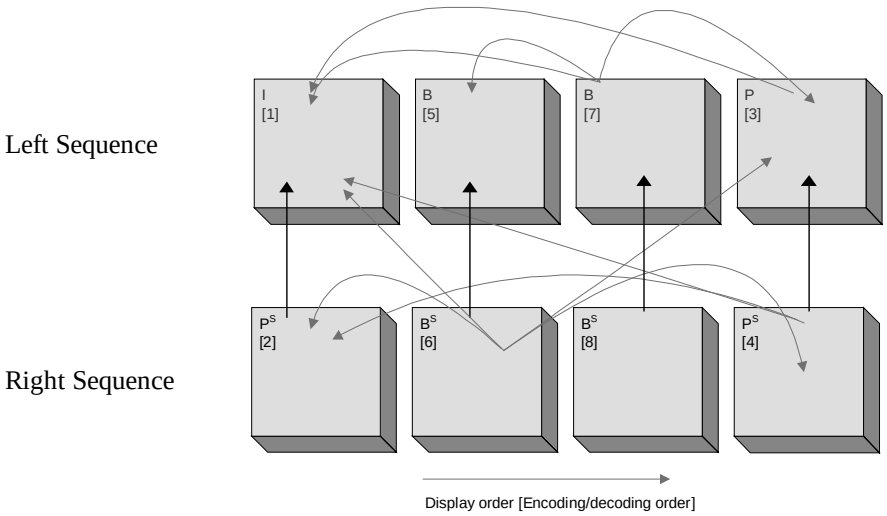


Figure 5. Frame Organization of Proposed Stereo CODEC

In addition to the above the frame organization in figure 5 leads to encoding the left sequence frames at odd occurrences whereas encoding the right sequence frames at even occurrences. The VUI is used to indicate this to the decoder. Therefore a decoder with monoscopic video decoding/display capability can ignore the decoding of right sequence frames based on notification by the VUI. In contrast a decoder with stereoscopic decoding capability (proposed) would decode the complete incoming sequences and will use the VUI to separate the frames to the two sequences. Thus by using the regular frame organization illustrated in figure 5 and appropriate VUI provision we meet the first design requirement stated in section 3.2.1.

3.2.3 **Alteration to the DPB and its management**:

To enable stereo video coding/decoding conforming to the requirements set out in section 3.2.1, we alter the standard DPB management strategy of H.264. We introduce a second buffer, named Decoded Picture Buffer Store (DPBS) to temporary store contents of the DPB.

The modified buffer management strategy can be explained with the use of the data flow diagram of figure 6. First, the incoming frame F, is checked to see whether it belongs to a left or right sequence. If the frame belongs to the left sequence, it is checked to find out whether it is an I frame. This is the case if previously a GOP has been completely coded and the system has been initialized. If the frame F is an I frame, both buffers, i.e. DPB and DPBS are cleared and the locally decoded F, is copied to both buffers. The buffers are then ready to handle the next frame. If F is a left frame, but is not classified as type I, the locally decoded F is copied to both buffers and subsequently the contents are swapped. The buffers are then ready to handle the next frame. If the frame F belongs to the right sequence, the locally decoded F is only copied to the DPB and the contents of the two buffers are swapped subsequently. The buffers are then ready to handle the next frame.

The frames to be encoded (i.e. F in the above discussion) are predicted based on the contents of the DPB. Typically the DPB contains a copy of up to five previously (an immediately) coded (and subsequently locally decoded) frames. Therefore in the above flow chart, as soon as the size of the DPB increases above 5 frames, the frame furthest in decoding order to the current frame is dropped. The algorithm described above ensures that when coding P or B frames belonging to the left (reference) sequence, only locally decoded copies of previous frames belonging to the left sequence are used as reference. This enables independent decoding of the left sequence at the decoder, if required, and satisfies the first design criterion of section 3.2.1. However in coding P and B frames of the right sequence, reference frames are obtained from both sequences. For example, in figure 5, in coding the first B frame of the right sequence all five previously coded frames are used as references, regardless of whether they belong to left or right sequences. Note that in coding the above B frame, the reference to the I frame (i.e. the first frame) of the left sequence represents exploiting possible *worldline* correlation. Thus the second design consideration of section 3.2.1 is met.

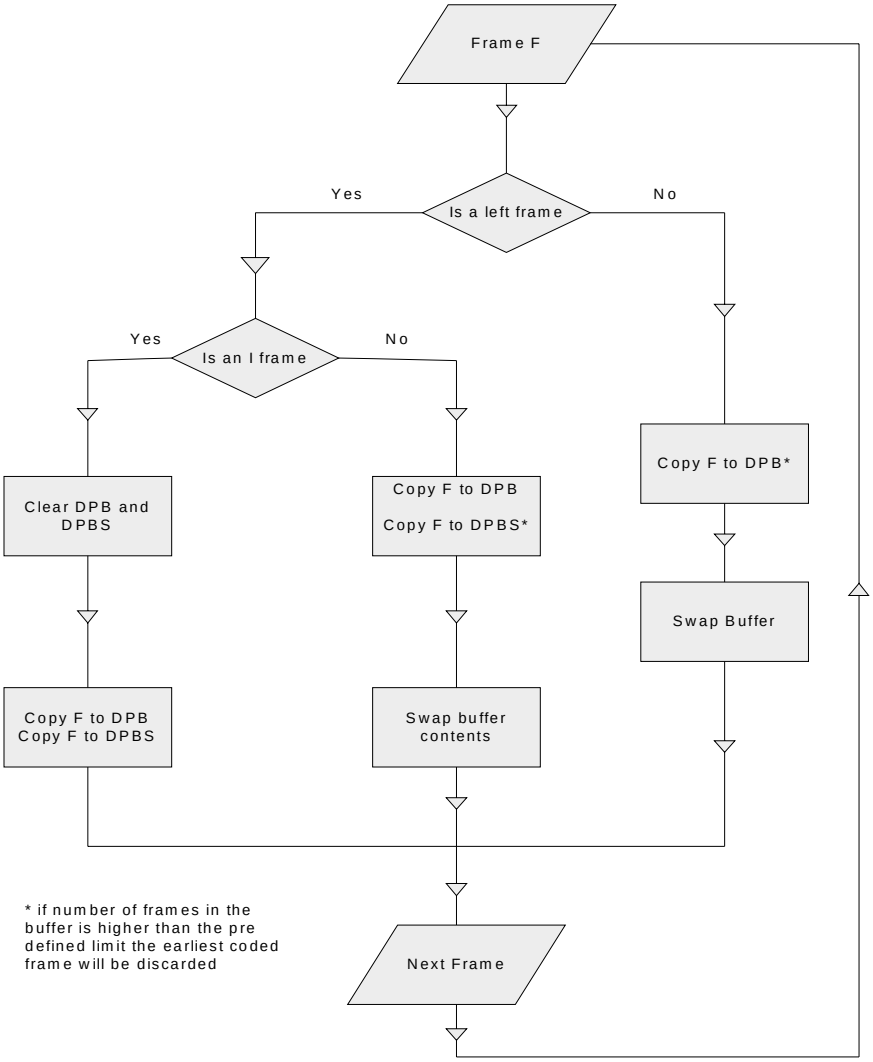


Figure 6. Buffer Management Strategy

The above alteration to the DPB and its proposed management maintains compatibility of the resulting encoded bit stream to H.264. It results in the possibility of using a single encoder to code both left and right sequences (see figure 7). The decoding of the resulting bit stream is possible with the use of a single decoder with similar modifications to that done at the encoder side. Therefore the proposed design meets the third and fourth design considerations set out in section 3.2.1.

To evaluate the performance of the proposed CODEC experiments were carried out on popular, colour test stereo video sequences, 'booksale' and 'crowd'. The frame size was 240 ×640 pixels. Figures 8 (a) - (d) illustrate a part (240 × 240) of the booksale scene. Figure 8(b) represents the frame 15 of the right video sequence, which is the frame to be predicted. Figures 8(a), (c) and (d) represent the three mostly correlated frames to that of figure 8(b), which are used as reference frames within the proposed coding process. Note that frame 14 is disregarded here as it is coded as a B frame. A visual comparison of the pictures illustrate the possibility of finding the best match for a given part of the frame in figure 8 (b) from any of the frames in figure 8 (a), (c) or (d). This justifies the importance of the proposed joint spatial-temporal-binocular-worldline redundancy exploitation as compared to the approaches used in previous literature, illustrated in figure 3.

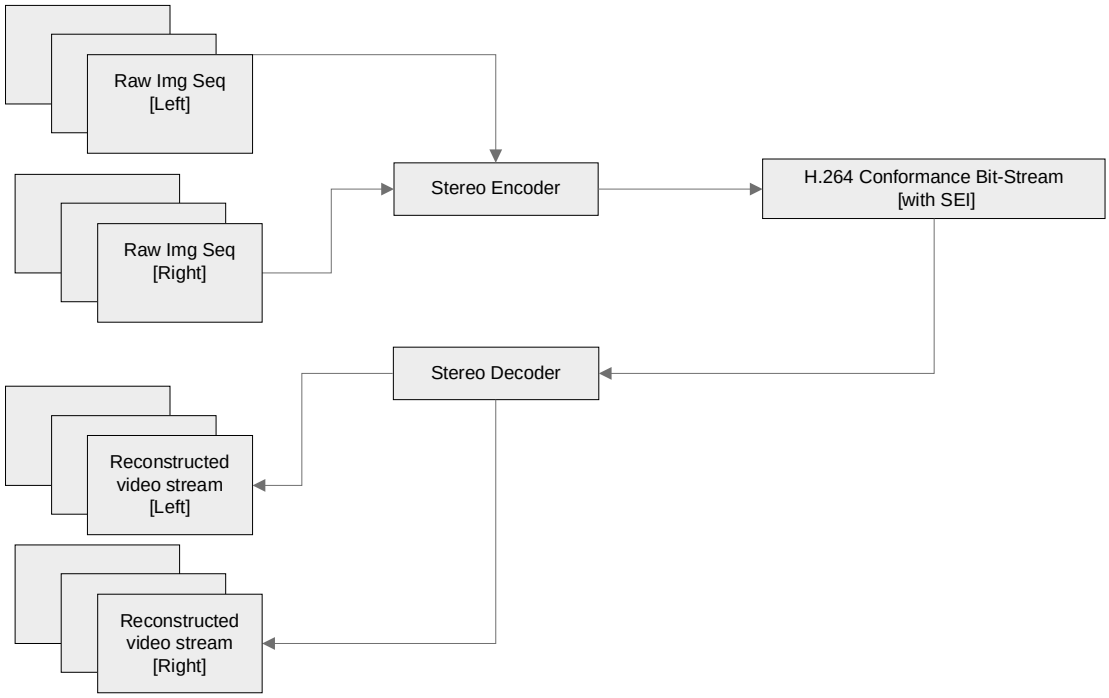
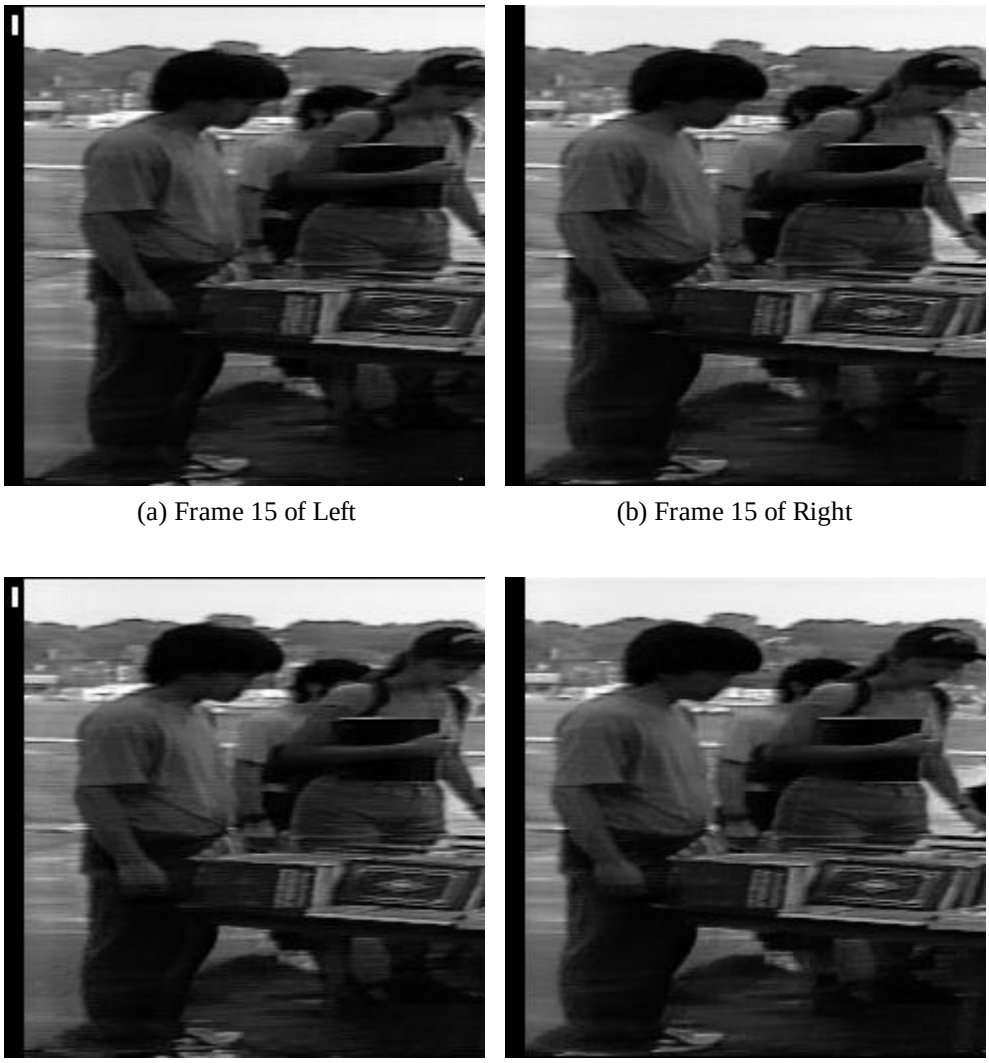


Figure 7. Single Encoder/Decoder Design

4. EXPERIMENTAL RESULTS & ANALYSIS

Figure 9 illustrates macroblock sub-divisions that occur as a result of predictive coding. The transparent sub-macrobocks denotes areas in the predicted frame that are coded by temporal prediction. Sub-macroblocks coloured white (note: only the top left hand corner of each sub-block has been coloured white) denote areas coded by binocular prediction. Note that the person closest to the camera is predominantly coded via temporal prediction. This is because the closer an object is to the camera, more binocular disparity it would show in the stereo scene. Thus better predictions with smaller motion vectors could be found via temporal prediction as compared to binocular prediction. On the other hand, in background areas which are far from the camera, binocular displacements are minimum. Thus binocular predictions are much better than temporal predictions, especially if moving far away objects are present in the background or illumination differences occur between the time intervals of the two frames. This fact is proven by figure 9, as it is clearly seen that binocular prediction has been used in background regions and texture-less foreground regions which behaves similar to background regions.

Figure 11 illustrates a bar-chart showing the relative compression of left/right sequences of ‘crowd’ stereo sequence at different levels of quantization, assuming an equivalent level of quality is maintained between left and right sequences. It shows that approximately 50% relative compression is achieved for right sequence as compared to the left at high compression levels.



(a) Frame 15 of Left (b) Frame 15 of Right

(c) Frame 13 of Left (d) Frame 13 of Right

Figure 8. Stereo Video Sequence 'Booksale'

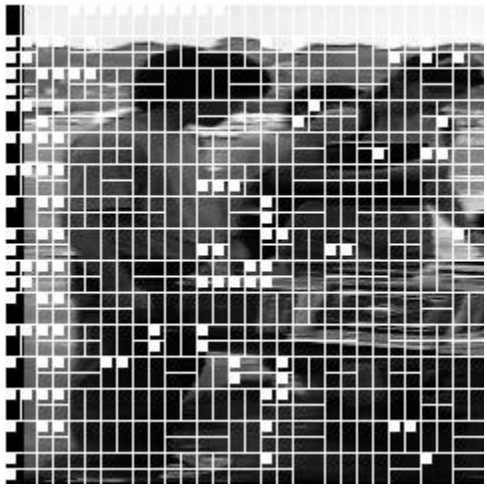


Figure 9. Macroblock Partitioning in Prediction

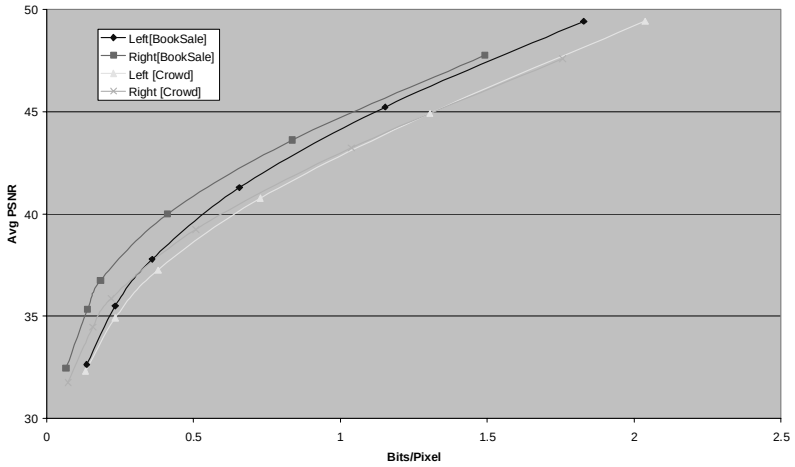


Figure 10. Rate-Distortion Performance

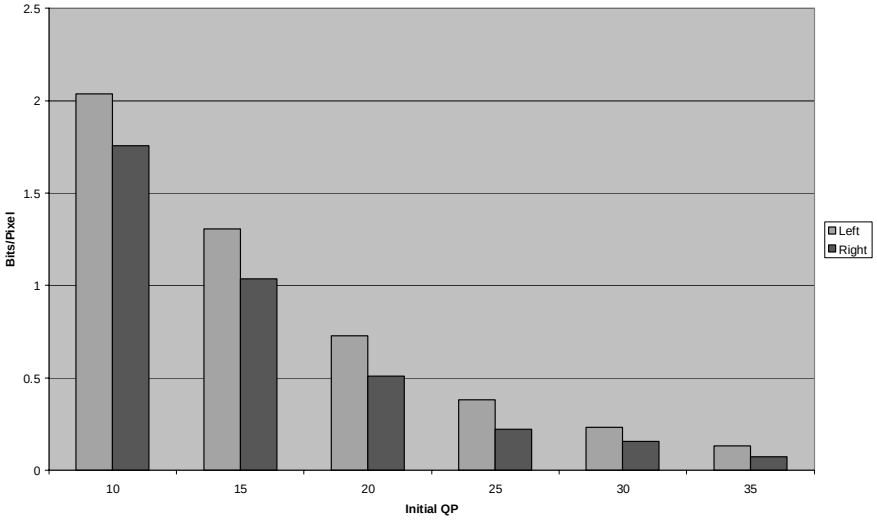


Figure 11. Relative Compression at Equivalent Quality

Note that the standard practice in coding stereo images/video is to code the reference frame/s at a slightly higher quality as compared to predicted frame/s. In our experiments above we have maintained the quality of the two bit streams at similar levels, by using a single quantization process, related to the single encoding approach used. However our design is capable of following the standard practice. The results thus obtained would show more significant improvements in relative compression levels between the left and right sequences, as compared to those illustrated in figure 11.

5. CONCLUSION

In this paper we have proposed an extension to a standard H.264 video CODEC, which will be capable of efficient stereoscopic video coding. We have used a single encoder/decoder stereoscopic video CODEC design capable of producing a fully compliant H.264 bit stream enabling monocular video decoding using a standard H.264 decoder and stereoscopic video decoding using the proposed, H.264 stereoscopic video decoder. We have shown that the special design of the joint motion-disparity estimation algorithm enables exploitation of *worldline* correlation that may be present in stereoscopic video sequences. A comprehensive analysis of the design aspects has been provided to explain the functionality of the proposed CODEC. An analysis based of popular test stereoscopic video sequences has been provided to evaluate the rate-distortion performance of the proposed CODEC. Currently we are carrying out further research into methods of speeding up the encoding/decoding operations of the proposed stereoscopic video CODEC.

6. REFERENCE

[1] "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC," in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050, 2003.

[2] Wiegand, T.; Sullivan, G.J.; Bjntegaard, G.; Luthra, A. "Overview of the H.264/AVC video coding standard" IEEE transactions on circuits and systems for video technology, vol. 13, no. 7, July 2003

[3] S.Sethuraman,"Stereoscopic image sequence compression using multi-resolution and quadtree decomposition based disparity and motion-adaptive segmentation," Ph.D. dissertation, Dept. of Electrical and Computer Engineering, Carnegic Mellon Univ., Pittsburgh, PA, 1996.

[4] Thanapirom, S., Fernando, W.A.O.P. and Edirisinghe, E.A., "Zerotree Entropy Based Coding of Stereo Video Sequences" , Proceedings of the 17th International Technical Conference on Circuits/Systems, Computers and Communications , IEEE, ITC-CSCC , Taiwan, 2002, pp. 908-911

[5] ISO/IEC 13 818-2, AMD 3, "MPEG-2 multiview profile," ISO/IEC JTC1/SC29/WG11, document no. N1366.

[6] Ohm J.R., Muller K., "Incomplete 3-D multiview representation of video objects" IEEE transactions on circuits and systems for video technology, vol. 9, no. 2, march 1999.

[7] J.-R. Ohm : "Stereo/Multiview Encoding Using the MPEG Family of Standards," invited paper, Proc. SPIE Vol. 3639, pp. 242-253, Stereoscopic Displays and Virtual Reality Systems VI, Jan. 1999.

[8] Woontack Woo "Rate-Distortion Based Dependent Coding For Stereo Images and Video: Disparity Estimation and Dependent Bit Allocation" PhD thesis, Faculty of the Graduate School University of Southern California, Dec 1998.

[9] N. Grammalidis, M.G Strintzis, "Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multi- view Image Sequences" IEEE Transaction on Circuits and Systems for Video Technology Vol 8. No 3. June 1998

[10]Luo Yan, Zhang Zhaoyang, and An Ping "Stereo Video Coding Based on Frame Estimation and Interpolation" IEEE TRANSACTIONS ON BROADCASTING, VOL. 49, NO. 1, MARCH 2003