Zerotree Entropy Based Coding of Stereo Video Sequences

S. Thanapirom[1], W.A.C. Fernando[1], E.A. Edirisinghe[2]
[1] Telecommunications Program, Asian Institute of Technology, P.O. BOX4, Klong Luang, Pathumthani 12120, Thailand
Tel. +66-2-524-5746, Fax. +66-2-524-5730
Email: fernando@ait.ac.th
[2] Department of Computer Science, Loughborough University, Leicestershire, LE11 3TU, UK
Tel. +44-1509-228234, Fax. +44-1509-211586
Email: E.A.Edirisinghe@lboro.ac.uk

Abstract: Over the past 30 years, many efficient 2D video coding techniques have been presented and developed from many research centers for commercialization. However, direct application of these monocular compression schemes is not optimal for stereo video coding. In this paper, we present a new technique for coding stereo video sequences based on Discrete Wavelet Transform (DWT). The proposed technique exploits Zerotree Entropy Coding (ZTE) that makes use of the wavelet block concept to achieve low bit rate stereo video coding. The one of two image streams, called main stream, is independently coded by modified MPEG-4 encoder and the other stream, called auxiliary stream, is coded by predicting from its corresponding image, its previous image or its follow image.

1. Introduction

Recently, digital video coding technologies have proceeded and targeted for a wide range of emerging applications such as mobile communication, video on demand, digital TV/HDTV broadcasting, video conferencing, DVD, surveillance, telemedicine, distance learning applications and multimedia image/video data base services. Many video coding standards have been established for different applications such as H.261, H.263 for video-conferencing applications, MPEG-1, MPEG-2 for multimedia applications and MPEG-4. These video coding standards have many similarities, differences and optional modes. However, these 2D visual communication technologies have matured so fast and been available for real-time visual communication, they might not be sufficient for the increasing new demands for realism or more natural representations of the scene. Various 3D [1] technologies have been investigated to support the demand for realistic environments. The recent advances in autostereoscopic display technology have enabled users to experience stereoscopic vision without the aid of special eyeglasses or helmet-mounted display kits that often result in noticeable user discomfort. Currently these cutting-edge developments are driving stereo imaging into further heights by widening its scope to cover a more diverse application area that includes CAD/CAM, remote surveillance, navigation, medical imaging, 3D visual communications, telemedicine, telerobotics, HDTV, entertainment and virtual reality.

A stereo vision is a simple way to provide 3D perception. To perceive stereo vision, we acquire two pictures of the same scene from two horizontally separated positions and then presenting the left frame to the left eye and the right frame to the right eye. The human brain can process the difference between these two images to yield 3D perception. Thus, every 3D image can be represented by two 2D image frames. These frames are said to form a stereo image pair. If a stereo pair, left and right images, is necessary to be stored and transmitted, twice as many bits will be required to represent stereo pair compared to a single image without exploiting the inherent redundancy.

Therefore, limited channel bandwidth is, as for conventional video, the main bottleneck for making these systems possible since stereo video requires to transmit or to store enormous amount of data. As a result, an efficient compression algorithm [2] will be essential to reduce the bandwidth requirements while maintaining the perceptual visual quality at the decoder.

Since the two images are projections of the same scene from two nearby points of view, they are bound to have a lot of redundancy between them. By properly exploiting this redundancy, the two image streams might be compressed and transmitted through a single monocular channel's bandwidth without excessive degradation of the perceived stereoscopic image quality. In the proposed algorithm, we make use of the high correlation between two image streams and ZTE to compress stereo video stream. One of two image streams, called main stream, is first encoded as a reference sequence using MPEG-4 [3,4] based on DWT and ZTE [5,6]. Then, the disparity information of another image sequence, called auxiliary stream, is estimated using mean difference of corresponding image blocks. Using the disparity information, auxiliary stream is encoded by choosing the best estimation from its corresponding main frame, its previous frame or its follow frame. The difference between the present frame and the predictive frame of auxiliary frame together with disparity are encoded using some parts of coder in main stream to receive further compression.

Rest of the paper is organized as follows. In section 2, ZTE is described in detail including with a brief discussion of characteristics of ZTE coding and the differences between ZTE and Embedded Zerotree Wavelet (EZW). A proposed method for coding stereo sequences using ZTE is given in section 3. The last section presents the conclusion and future work.

2. Zerotree Entropy Coding

ZTE [5] is a new efficient method for coding wavelet transform coefficients of motion-compensated video residuals or of video frames. It was proposed by S.A. Martucci and I. Sodagar in 1996. It employs the idea that, in the hierarchical decomposition of the wavelet transform, every coefficient at a given scale, excluding the highest frequency subband, can be associated to a set of coefficients of same orientation at the next finer scale. The coefficient at the coarse scale is called the *parent*, and the four coefficients representing the same spatial location and of similar orientation at the next finer scale are called *children*. According to this relationship, we can build a data structure called a *wavelet tree*. Figure 1 shows a wavelet tree root at a coefficient in the highest frequency subband, HH_3. At the next lower frequency subband, the parent-child relationship is defined such that each parent has three children, one in each subband at the same scale and spatial location but of different orientation.

In ZTE coding, the coefficients of each wavelet tree are rearranged to form a wavelet block as shown in Figure 2. Each wavelet block contains all coefficients in the same wavelet tree in the frame. The wavelet blocks are located at the same corresponding spatial location as their wavelet tree. To use ZTE, a symbol is assigned to each node in a wavelet tree describing the wavelet coefficients corresponding to that node.

The quantization of wavelet coefficients is not mentioned in ZTE but can be taken either before employing ZTE coding or during the ZTE process. Any quantization scheme can be used. By combining quantization with the construction and coding of zerotrees, only a single pass through the data is required and quantization and bit allocation can be done adaptively.

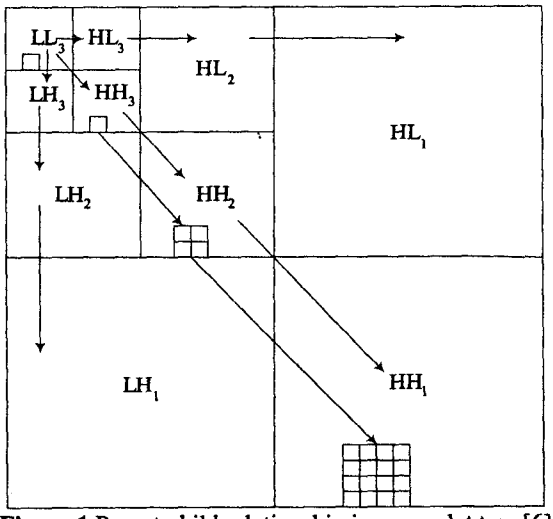

Figure 1 Parent-child relationship in a wavelet tree [6]

The wavelet trees are coded by scanning each tree from the root in the lowest frequency band through the children, and assigning one of three symbols to each node encountered [5]: *zerotree root*, *valued zerotree root* or *value*. A zerotree root denotes a coefficient that is the root of a zerotree. A zerotree exists at any wavelet tree node where the coefficient is zero and all the node's children are themselves zerotrees. A valued zerotree root is a node where the coefficient has nonzero amplitude and all four children are zerotree roots. A value symbol identifies a coefficient with amplitude either zero or nonzero, but also with some nonzero descendant. The scanning process of each tree can stop at zerotree root or valued zerotree root symbols. The list of nonzero quantized coefficients that correspond one-to-one with the valued zerotree root symbols are encoded using an alphabet that does not include zero. The remaining coefficients, which correspond one-to-one to the value symbols, are encoded using an alphabet that includes zero. For any node reached in a scan without node's children, zerotree root and valued zerotree cannot apply. Therefore, bits are saved by not encoding any symbol for this node and encoding the coefficients along with those corresponding to the value symbol.

2.1 Characteristics of ZTE coding

The ZTE algorithm does not generate an embedded bitstream as EZW [7,8] does, but it provides edibility and other advantages over EZW coding, including additional improvement in coding efficiency.

If quantization is performed during the tree construction, it is possible to specify the dynamic global quantizer step size for each wavelet block or the dynamic local quantizer step size for each group of coefficients in a wavelet block. These quantizers are adjusted according to coefficient representing in frequency band or/and in a wavelet block thus provide content based-coding. Further, the bit usages are available to the quantizer for adaptation purposes.

The most important feature of this coder is that in the ZTE technique quantization is done explicitly and therefore can be optimized and dynamically adapted to scene content. A second feature is that the scanning and encoding of wavelet coefficients is done in an order that exploits the close connection between the coefficients and what they represent in the frame, thereby allowing allocation of bits on an object-by-object basis. Thirdly, in ZTE coding, the use of zerotrees has been enhanced by defining a new set of symbols designed specifically for very low-bit rate coding of video. In place of popular EZW algorithm, ZTE is claimed to provide greater flexibility, adaptability and improved coding efficiency.

2.2 Differences between ZTE and EZW

Like EZW, ZTE exploits the similarities among the wavelet coefficients at the same spatial location in all scale and organizes wavelet coefficients into wavelet trees and then uses zerotrees to reduce the number of bits required to represent those trees. Nevertheless, ZTE differs from EZW in four major ways [6]: 1) Quantization is explicit instead of implicit and can be performed distinct from the zerotree growing process or can be incorporated into the process, thereby making it possible to adjust the quantization according to where the transform coefficient lies and what it represents in the frame. 2) Coefficient scanning, tree growing, and coding are done in one pass instead of bit-plane-by-bit-plane. 3) Coefficient scanning is changed from subband-by-

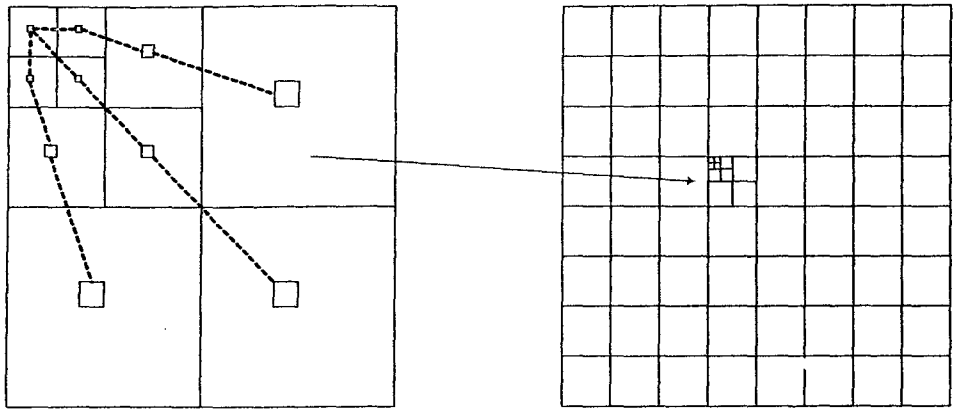

Figure 2 Rearranging a wavelet into wavelet block [6]

subband to a depth-first traversal of each tree. 4) The alphabet of symbols for classifying the tree nodes is changed to one that performs significantly better for very low bit-rate encoding of video.

3. ZTE for stereo video coding

To reduce numbers of bit rate represent stereo video sequence while maintaining the perceptual visual quality at the decoder, we utilize Zerotree Entropy Coding (ZTE) that makes the use of wavelet block concept to achieve low bit rate stereo video coding according to the result in [6] for 2D video, a zerotree video coder performs better on I-frames and equal or better on P-frames when compare to EZW. The additional feature of ZTE in stereo video coding is that quantization is done explicitly and therefore can be optimized and dynamically adapted to scene content. A second feature is that the scanning and encoding of wavelet coefficients is done in an order that exploits the close connection between the coefficients and what they represent in the frame, thereby allowing allocation of bits on an object-by-object basis. We take advantages of binocular redundancy between stereo images (high correlation between stereo images) as well as the temporal redundancy between consecutive images. The resulting disparity and motion maps [9] can be used as efficient representations of 3D visual sequence data. The basic strategy will be as follows. First, the video frames in two image sequences are defined in terms of layers of *video object planes* (VOP). Each VOP is a video frame of a specific object of interest to be coded. We apply motion and disparity based segmentation to extract the objects in the scene. The coder encodes VOP of main stream by using a modified MPEG-4 video encoder which implements DWT, Overlapping Block Motion Compensation (OBMC) to reduce the artificial block discontinuities and ZTE. We denote the I, P, B pictures of the main stream using I_M , P_M , and B_M. The corresponding pictures in the auxiliary stream are represented by I_A , P_A , and B_A .Then we apply auxiliary stream and VOP of main stream into disparity estimation and compensation. After that the coder chooses the best estimation VOP of auxiliary stream from its corresponding object plane, its previous object plane frame or/and its follow object plane frame [10]. The decision is depend on which types of VOP picture in main stream are coded 1) a VOP- I_A frame is coded by using disparity compensation prediction which respects to the corresponding VOP- I_M frame. The total bit count is greatly reduced compare to independent coding. 2) a VOP- P_A frame is coded by selecting the best estimation between the previous reference VOP frame and the corresponding VOP-P_M frame. 3) similarly, a VOP- B_A frame is coded by selecting the best estimation among the previous reference VOP frame, the follow reference VOP frame and the corresponding VOP-B_M frame. The prediction process is shown in Figure 3. The residuals of each VOP frame in auxiliary stream is then transformed by DWT, quantized, coded by ZTE and at last coded by arithmetic coding. If the best prediction frame is achieved by disparity estimation, the disparity is coded by Huffman coding and transmitted together with coded VOP frame.

4. Conclusion and future work

The proposed compression method makes use of ZTE and high correlation in a stereo pair to achieve low bit rate stereo video. In the proposed scheme, the main stream is coded by DWT based MPEG-4 encoder with ZTE. To extract the object, we use the motion and disparity based segmentation. Since the redundant information between left and right images is quite large, an auxiliary stream can be encoded using the disparity based compensation prediction. However, sometimes the motion based compensation provides the better prediction than disparity compensation. The best prediction among these two types of compensation is chosen such that the high compression can be achieved. Future work is required to examine the proposed stereo coder with actual stereo video sequences.

References

[1] Woon-Tack Woo, http://vr.kjist.ac.kr/~3D/.

[2] M.G. Perkins, "Data Compression of Stereopairs", IEEE Trans. on communications, Vol. 40, no. 4, pp. 684-696, 1992.

[3] M. Ghanbari, "Video coding: an introduction to standard codecs", The institution of Electrical Engineers, London, United Kingdom, 1999.

[4] T. Sikora, "The MPEG-4 Video Standard Verification Model", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 7, no. 1, Febeuary, 1997.

[5] S.A. Martucci and I. Sodagar, "Zerotree entropy coding of wavelet coefficients for very low bit rate video", Proc. 1996 IEEE Int. Conf. Image Processing, Lausanne, Switzerland, September, 1996.

[6] S.A. Martucci, I. Sodagar, T. Chiang and Y. Zhang, "A Zerotree Wavelet Video Coder", IEEE Trans. on circuits and systems for video technology, Vol. 7, no.1, pp.109-118, February, 1997.

[7] J.M. Shapiro, "An embedded hierarchical image coder using zerotrees of wavelet coefficients", IEEE Data Compression Conference, Snowbird, UT, pp. 214-223, 1993.

[8] J.M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients", IEEE Trans. on Signal Processing, Vol. 41, pp.3445-3462, 1993.

[9] P.D. Gunatilake, M.W. Siegel, A.G. Jordan, "Compression of Stereo Video Streams", Department of Electrical & Computer Engineering and School of Computer Science, Carnegie Mellon University.

[10] M. Siegel, S. Sethuraman, J.S. McVeigh and A. Jordan, "Compression and Interpolation of 3D-Stereoscopic and Multi-View Video", The Robotics Institute, Carnegie Mellon University, Pittsburgh PA, 1997.

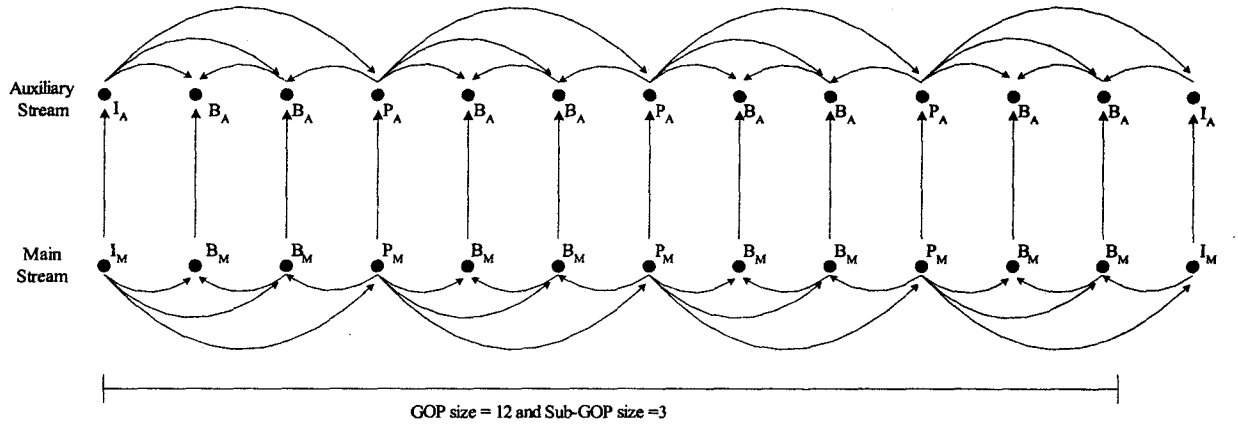

Figure 3 Disparity and motion compensated prediction in stereo video sequences

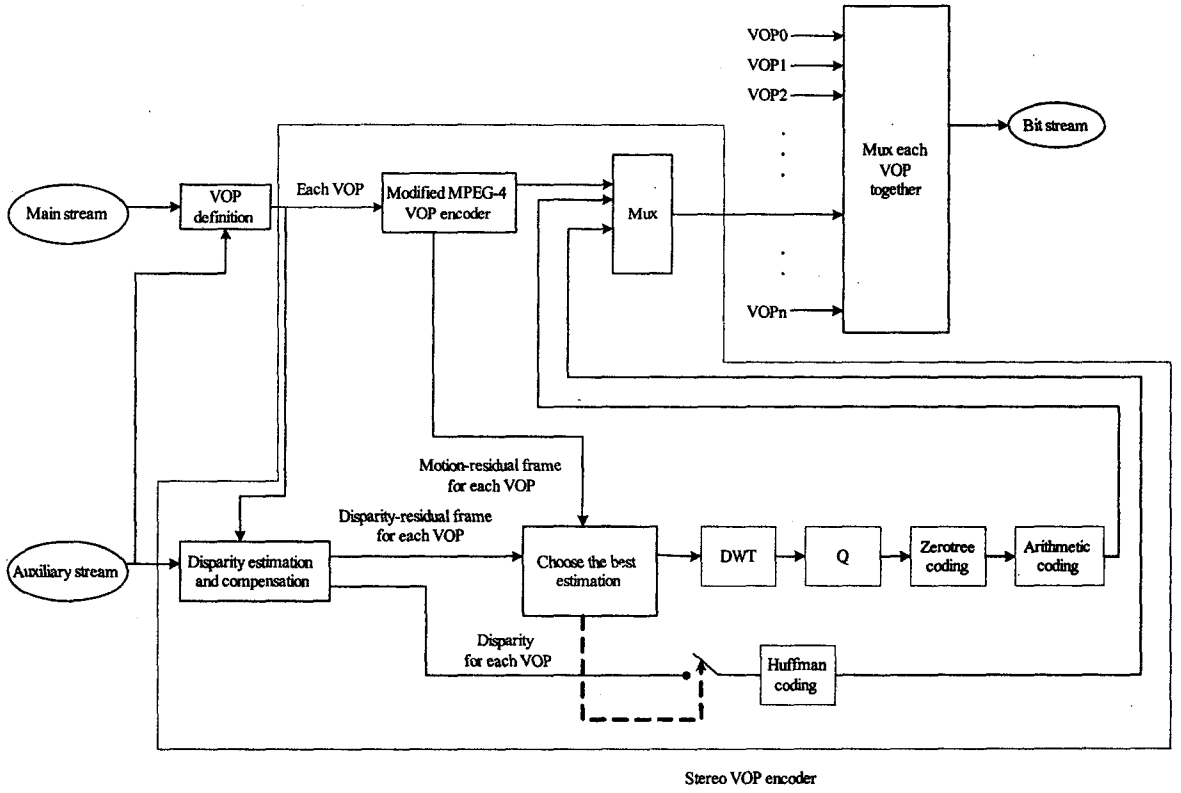

Figure 4 The proposed stereo video encoder