Stereo Video Coding Based on Interpolated Motion and Disparity Estimation

J. N. Ellinas, M. S. Sangriotis
University of Athens, Department of Informatics, Ilissia, 157 84 Athens, Greece
jellin@teipir.gr, sagri@di.uoa.gr

Abstract

In this paper, a new optimised method of coding stereoscopic image sequences is presented and compared with already known methods. Two basic methods of coding a stereoscopic image sequence are the compatible and joint. The first one uses MPEG for coding the left channel and takes advantage of the spatial disparity redundancy between the two sequences for coding the right channel. The second one employs MPEG for coding the left channel but takes advantage of both temporal redundancy among the right channel frames and spatial redundancy between the corresponding frames of the two channels. The proposed method, which is called IMDE, estimates the P and B type of frames of the right channel by an interpolative scheme that takes in to account both the temporal and disparity characteristics. Investigating the effectiveness of the joint motion and disparity vectors estimation as well as the choice of the weighting factors that participate in the proposed interpolative scheme optimises the whole framework.

1. Introduction

The stereoscopic vision is based on the projection of an object on two horizontally slightly displaced image planes and has an extensive range of applications as 3-D television, 3-D image sequences applications, robot vision, virtual machines, medical surgery etc. In that sense, two pictures of the same object taken from two nearby points form a stereo pair and contain all the depth information that are necessary so that if these pictures are seen by the human eyes to give the stereoscopic feeling. These demanding areas impose the need for the development of more efficient compression techniques of a stereo image pair or a stereo image sequence. Typically the transmission or the storage of a stereo image sequence requires twice the bandwidth of a monocular video system. The objective on a bandwidth-limited transmission system is to develop an efficient coding scheme that will exploit the redundancies of the two video streams, that is, the intra-channel and inter-channel correlation or similarities. A typical compression scenario is the more effective motion compensation of individual video channels and the reduction of the disparity between the left and right frames. Basically, there are two methods of handling the motion and disparity fields in a stereoscopic video.

The first method, based on intensity, estimates the motion field between succeeding frames of a video channel or disparity correspondences between corresponding frames of a stereo image pair by handling the intensities of the individual frames [1]. Several stereo compression algorithms have been developed that use block matching or alternative implementations, as hierarchical disparity estimation [2], multiresolution block matching [3], block matching with geometric transform [4], etc.

The second method, based on object handling, firstly defines or extracts the features of the participating objects in the processed scene and then estimates the temporal or disparity fields between the frames of the stereoscopic video [5]. Apart from the above-mentioned methods, several others have been proposed that try to improve the performance or to combine the characteristics of them as the segmentation based coding [6], stereo image projection [7], post-compensation residual coding [8], overlapped block disparity compensation [9], a hybrid scheme between block and object based technique [10], etc.

In this paper a new stereo image sequence compression scheme is proposed, which is called IMDE (Interpolated Motion and Disparity Estimation) and belongs to intensity methods, that uses the MPEG standard for coding the left or main image stream and an interpolative scheme, which is applied on the corresponding motion and disparity frames, for predicting and coding the P and B type of frames of the right or auxiliary image stream. The estimated residual frames of both channels are coded by using a DWT (Discrete Wavelet Transform) followed by a morphological encoder, which represents the wavelet coefficients by using morphology. Furthermore, the performance of the proposed interpolative scheme is investigated with respect to the joint estimation of motion and disparity vectors as well as the optimized choice of the weighting factors of the predicted frames that participate in this scheme.

This paper is organized as follows. Section II provides a brief description of disparity, the two basic stereoscopic compression techniques used for performance comparison and the principle of operation of the proposed coder. In section III the analysis of the proposed interpolative scheme is explained and experimental results are presented and commented in Section IV. Finally, conclusions are summarized in Section V.

2. Overview

2.1. Disparity in stereoscopic vision

The problem of finding the points of a stereo pair that correspond to the same 3-D object point is called correspondence. The correspondence problem is simplified in a one-dimensional problem if the used cameras are coplanar. The distance between two points of the stereo pair images that correspond to the same scene point is called disparity. The estimation of this distance (disparity vector or DV) is very important in stereo image compression, because one image (target) can be predicted from the other (reference) along with the disparity information. Then, the difference from the original one (disparity compensated difference or DCD) is evaluated, so that the redundant information not to be encoded and transmitted. The Disparity Compensated Difference (DCD) for the target image is defined as:

$$DCD(x, dv) = I_R(x) - I_L(x + dv) \quad (1)$$

where I_R, I_L are the intensity values for the Right and Left images respectively and dv is the disparity vector, over a window searching area, which is defined as:

$$dv(x) = \arg\min_{dv} \left| \; DCD(x, dv) \; \right| \quad (2)$$

In a stereoscopic video stream the redundancies that can be exploited in order to increase compression are the temporal redundancies, which regard motion in both streams, and spatial redundancies, which regard disparity between corresponding frames of the two streams. Hence, the difference frames for motion, DFD (Displaced Frame Difference) and disparity, DCD (Disparity Compensated Difference), are encoded along with the motion (MV) and disparity vectors (DV).

2.2. Stereoscopic image sequence compression

Stereoscopic image sequence compression can be obtained by one of the three methods as in Figure 1.

The *simulcast* method is analogous to the motion-JPEG of a monoscopic video stream and inherits the advantages and disadvantages of the independent coding. The *compatible* method utilizes MPEG standard for the main stream and takes advantage of the spatial correlation between the corresponding frames of the main and auxiliary video streams in order to increase the overall compression factor. The *joint* method employs MPEG standard for the main stream together with the motion and disparity fields of the auxiliary channel frames. The proposed method is an optimised version of *joint* that provides an interpolative estimation of the auxiliary channel P and B frames and uses a morphological encoder based on DWT.

Figure 2 illustrates the *compatible* stereo video coding method in more detail. The characteristics of this method, if the used Group Of Pictures is IBBPBBPBB, are:

- The main stream is coded as in MPEG standard.
- The auxiliary stream frames are coded by estimating the disparity predicted frames from the corresponding ones of the main stream.

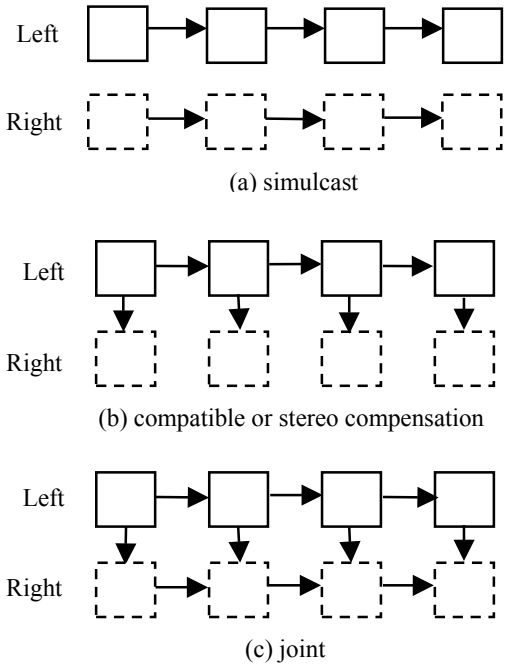


Figure 1. Typical methods of stereoscopic video coding

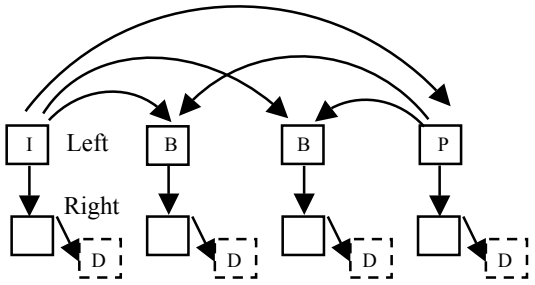


Figure 2. *Compatible* method of stereo coding

Figure 3 illustrates the proposed stereo video coding method, with the following characteristics:

- The main stream is coded as in MPEG standard.
- The I-frame of the auxiliary stream is coded by using the disparity predicted frame.
- The P-frame of the auxiliary stream is coded by using the predicted frame, which results by interpolating the motion predicted frame from the same channel and the disparity predicted frame from the other channel.
- The B-frame of the auxiliary stream is coded by using the predicted frame, which results by interpolating the forward and backward motion predicted frames from the same channel and the disparity predicted frame from the other channel.

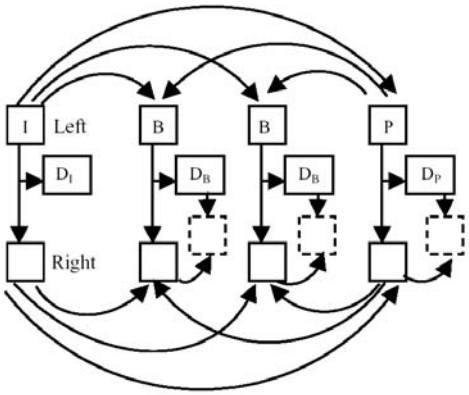


Figure 3. IMDE method of stereo coding

The above-mentioned procedure presents beneficial characteristics in stereo video coding, especially after vector and weighting factors optimization. Both motion and disparity compensation is applied on reconstructed frames by using the full search BMA, with blocks of size 16X16 pixels and searching area of 6 pixels.

2.3. Coding based on the morphological representation of DWT coefficients

The conventional wavelet image coders decompose a "still" image into multiresolution bands [11], providing better compression quality than the so far existed DCT transform. An alternative adaptive wavelet packet scheme can enhance the benefits of this transform [12]. The types of wavelet decomposition suffer from the fact that they include all the coefficients, which are spread within the subbands, in the transmitted sequence, even those that are zero or nearly zero and their absence would have little effect on the reconstructed image quality. The statistical properties of the wavelet coefficients led to the development of some very efficient algorithms as, the embedded zero tree wavelet coder (EZW) [13], the coder based on set partitioning in hierarchical trees (SPIHT) [14], the coder based on the morphological representation of wavelet data (MRWD) [15], and the embedded block coding with optimized truncation of the embedded bit streams (EBCOT) [16].

The MRWD algorithm, which is used in the present work, exploits the intra-band clustering and inter-band directional spatial dependency of the wavelet coefficients, Figure 4, in order to predict them in a hierarchical manner starting from the coarsest scale by using a structuring element 3X3 for the morphological dilation operation. A dead-zone uniform step size quantizer quantizes all the subbands and the coefficients of the coarsest detail subbands constitute binary images that contain the significant and insignificant generated than a predefined threshold are called significant.

The intra-band dependency of wavelet coefficients or the tendency to form clusters, suggests that applying a morphological dilation operator may capture the significant neighbors. The finer scale significant

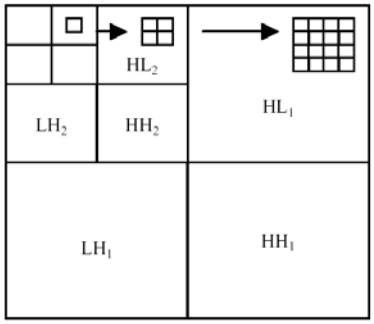


Figure 4. Inter-band spatial dependency in wavelet decomposition

coefficients, children subbands, may be predicted from the maps of significance of the coarser scale, parent subbands, by applying the same morphological operator to an enlarged neighborhood because the children subbands have double size than their parents. Each of these two partitions can be further partitioned into two groups. The significant partition is divided into two groups that contain significant coefficients that are truly significant and insignificant coefficients that were predicted to be significant. The insignificant partition is divided into two groups that contain insignificant coefficients that are truly insignificant and significant coefficients that were predicted to be insignificant. This partitioning reduces the overall entropy but there is an overhead by the needed side information.

3. The proposed interpolative scheme

3.1. Interpolative procedure in stereo video

In a monoscopic video stream, the transmitted residual B-frame is the difference of the initial frame and the predicted one, which is estimated by interpolating a previous and a next reference frame. The achieved compression by the interpolative prediction is larger, because the energy of the residual frame becomes smaller. The predicted B-frames of the auxiliary channel are estimated by interpolating the motion compensation predicted frames from the reconstructed I or P reference frames and the disparity compensation predicted frame from the corresponding reconstructed B-frame of the main channel. Hence, the transmitted difference has been exempted, in a great deal, from the temporal and spatial correlating characteristics of the frame. As Figure 5 shows, the interpolation must be applied to every macroblock of the frame and this is described as follows:

$$\hat{b}(v_f, v_b, v_d) = w_f \tilde{b}_f(v_f) + w_b \tilde{b}_b(v_b) + w_d \tilde{b}_d(v_d) \qquad (3)$$

where b, v, w stand for macroblock, vector, weighting factor and indexes f, b, d denote forward, backward, disparity respectively. The (~) sign indicates that the used macroblocks are the reconstructed ones.

At this stage, the two motion weighting factors together are considered to play an equal role with

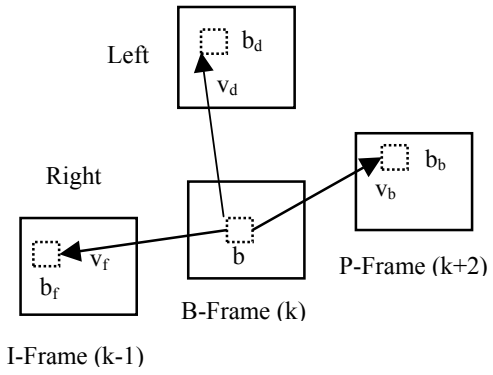


Figure 5. Interpolative scheme for stereoscopic video coding

disparity weighting factor and for this reason they are set to: w_f=w_b=0.25 and w_d=0.5. The aim is to find:

$$(v_f^{opt}, v_b^{opt}, v_d^{opt}) = \arg\min_{S}(b - \hat{b}(v_f, v_b, v_d)) \qquad (4)$$

where S is the block searching area. Usually a good approximation of the solution of equation (4), which reduces the number of the performed matches, comes from the best independent matching of the frames that participate in the interpolation, that is:

$$v_b^{opt} = \arg\min_{S}(b - \tilde{b}_b(v_b))$$

$$v_f^{opt} = \arg\min_{S}(b - \tilde{b}_f(v_f)) \qquad (5)$$

$$v_d^{opt} = \arg\min_{S}(b - \tilde{b}_d(v_d))$$

However this solution is sometimes sub optimal. The method that provides a better solution is the joint full search that has been developed for monoscopic video [17]. The extension of the joint search method to stereoscopic video, which is for first time applied and tested, can be described as follows:

- The best independent vectors and the minimum difference according to (5) are estimated.
- The vectors v_f, v_b are kept constant and vector v_d is updated for a lower minimum of the difference.

$$v_d^{opt} = \arg\min_{S}((b - w_f\tilde{b}_f(v_f) - w_b\tilde{b}_b(v_b)) - w_d\tilde{b}_d(v_d)) \qquad (6)$$

- The vectors v_f, v_d are kept constant and vector v_b is updated for a lower minimum of the difference.

$$v_b^{opt} = \arg\min_{S}((b - w_f\tilde{b}_f(v_f) - w_d\tilde{b}_d(v_d)) - w_b\tilde{b}_b(v_b)) \qquad (7)$$

- The vectors v_b, v_d are kept constant and vector v_f is updated for a lower minimum of the difference.

$$v_f^{opt} = \arg\min_{S}((b - w_b\tilde{b}_b(v_b) - w_d\tilde{b}_d(v_d)) - w_f\tilde{b}_f(v_f)) \qquad (8)$$

- The above procedure is iterated until the minimum value no longer becomes lower.

3.2. Estimation of weighting factors in the interpolation

The weighting factors for the frames that participate in an interpolative procedure are usually w_f=w_b=0.5, in a monoscopic video. However, it is possible their values to be inversely proportional to the time distance from the reference frames [18]. In this work, the motion weighting factors, which are forward and backward, are kept in a constant relation, 7:3 for the first and 3:7 for the second B-frame and their relation with disparity weighting factor is investigated. The sum of the three factors must be unity.

The normalized magnitude or energy of a macroblock b_{ij} of the residual B-frame is:

$$E_{ij} = \frac{1}{16\times16}\sum_{k=1}^{16}\sum_{l=1}^{16}\{b_{ij}(k,l) - \hat{b}_{ij}(k,l;v_f,v_b,v_d)\}^2 \qquad (9)$$

where, $\hat{b}_{ij}$ is the predicted block from (3) after the previously described vector optimization procedure and the definition of the weighting factors. The total normalized energy of the residual frame is:

$$E_{tot} = \sum_{i=1}^{M/16}\sum_{j=1}^{N/16} E_{ij} \qquad (10)$$

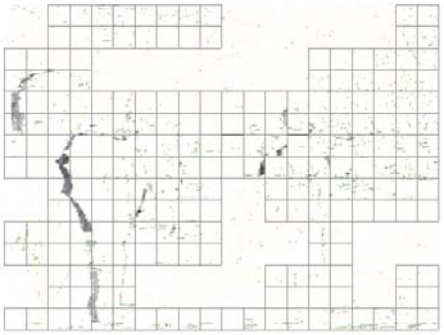
where i=1…M/16 and j=1…N/16, for an image of size MXN and blocks of 16X16 pixels.

The energy of the residual B-frame depends on the proper selection of the weighting factors between motion and disparity. The reduction of this energy provides, for a given bit-rate, higher PSNR for the reproduced frame as the resulting distortion is decreased. Basically the weighting factors should be estimated for every 16X16 macroblock and their choice must minimize the magnitude of each macroblock and consequently the energy of the entire residual frame. Instead of estimating the weighting factors for every macroblock, which is time consuming and bit-rate expensive, the following method is proposed:

- Initially the value of 0.5 is assigned to disparity weighting factor and the values of the motion weighting factors are evaluated according to the previous definition.
- The resulting residual B-frame is processed by the quadtree methodology. The frame is divided into four equal blocks if and only if the difference between the maximum and the minimum intensity of the block is greater than a threshold. This procedure is repeated until a certain block size or this difference becomes lower than the predefined threshold, which is set to 10% in this work. In this way the residual is partitioned in two areas, one with low energy and another, which concentrates most of the energy. Therefore the weighting factors of the high-energy region will affect mostly the total energy of the residual frame. This suggests that the optimisation of the weighting factors will be the same effective if it is applied on the high-energy homogeneous region

created. The weighting factors for the high-energy region will be estimated by minimizing the energy of this region. The other set of weighting factors, for the low energy region, may hold their initial values i.e 50% for motion and 50% for disparity. Figure 6(a) shows the application of quadtree on a residual B-frame for a threshold value of 10% and smallest block size of 16X16 pixels. The weighting factors for motion and disparity have a relation 50-50% in Figure 6(a), whereas this relation becomes 80-20% in Figure 6(b) by applying weighting factors optimization.

It is obvious from the two figures that the energy in the second case is much lower. The mean variance is approximately half of the first case. This reduction of energy means that the compression factor for the specific residual will be increased.



(a)

(b)

Figure 6. B-frame residuals for "crowd" sequence motion-disparity: (a) 50-50% (b) 80-20%

4. Experimental results

The proposed coder of a stereoscopic image sequence, called IMDE, that estimates the P and B-frames of the auxiliary channel by interpolating the motion and disparity predicted frames of the right and left channel respectively, has been applied on two stereoscopic sequences "crowd" and "booksale" [19]. The size of each image is 320X240 pixels, the type of sequence is IBBPBBPBB and the macroblock size is 16X16 pixels.

Table 1. Performance of *compatible* and IMDE methods

Sequence	Right Image Sequence					
	I		B		P	
	PSNR	bpp	PSNR	bpp	PSNR	bpp
Crowd-1	32.03	0.37	31.88	0.22	31.40	0.39
Crowd-2	32.03	0.37	32.66	0.22	32.35	0.27
Book-1	32.63	0.31	32.24	0.27	31.70	0.48
Book-2	32.63	0.31	33.20	0.25	32.72	0.29

(1) *compatible* (2) IMDE

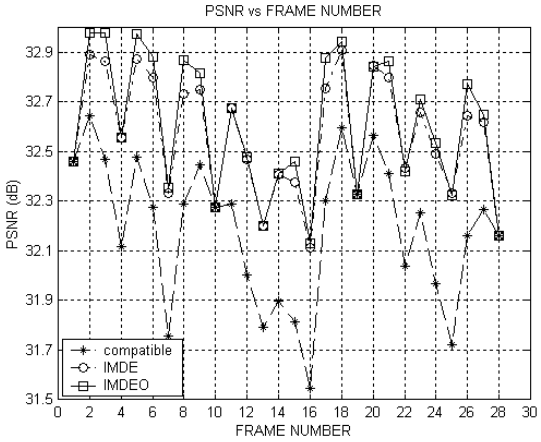


Figure 7. PSNR vs Frame Number

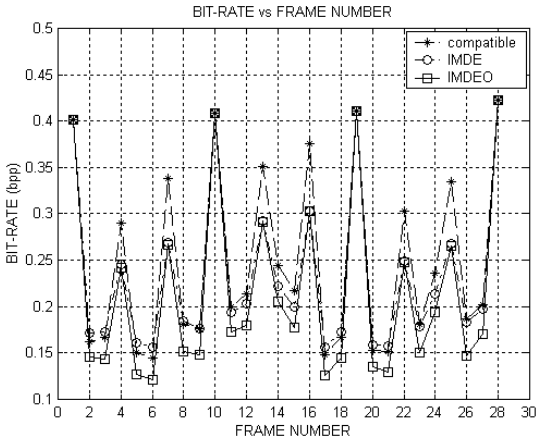


Figure 8. Bit-Rate vs Frame Number

The quality measure of the reproduced images is estimated by PSNR. The total bit-rate is the entropy of the DWT subband coefficients, after their morphological representation and partitioning by the morphological encoder and the vectors that are used for a prediction. The vectors are DPCM encoded, since their transmission must be lossless.

Table 1 provides the average performance of the two tested stereo video sequences for *compatible* and IMDE compression methods. The values indicated for B and P

frames of the auxiliary sequence have not been optimized either for motion and disparity vectors or for weighting factors. It is apparent, for both image sequences, that the proposed method provides better overall performance. The above-proposed vector and weighting factors optimization may further enhance this performance.

Figure 7 and Figure 8 show the total PSNR and bit-rate for the "crowd" stereo sequence respectively, for the three tested methods. The weighting factors for IMDE method were set: w_f=0.35, w_b=0.15, w_d=0.5 for B-frames and w_f=0.5, w_d=0.5 for P-frames. IMDEO (IMDE Optimum) denotes the enhanced performance of IMDE method after vector and weighting factors optimization for P and B-frames. For example the optimization of the vectors and weighting factors, for the first B-frame of the auxiliary image sequence, provides: w_f=0.56, w_b=0.24, w_d=0.2. This estimation results in better total PSNR and bit-rate for the specific frame. Applying the optimization to the whole image sequence it can be seen that, for the 7th P-frame there is an increment of 0.6 dB in PSNR, whereas for the 11th B-frame, this increment is 0.4 dB for about the same bit budget. It is obvious that the IMDEO method outperforms the *compatible* method for P or B-frames.

5. Conclusions

A stereoscopic image sequence can be coded in an effective way if the redundant information that exists between the frames of the same channel (temporal redundancy) and the corresponding frames of the two channels (disparity redundancy) is taken into account. Among the typical methods of coding, the most attractive are the *compatible* and *joint*. The *compatible*, compresses the auxiliary channel by taking into account the spatial redundant information between the two channels. The *joint* compresses further the auxiliary P and B-frames by taking into account both the temporal and the spatial redundancy. The proposed scheme is a version of the *joint* method that estimates P or B frames of the auxiliary sequence by interpolating the motion and the disparity compensation predicted frames. The motion and disparity vectors are optimized, by applying an exhaustive search, so that an optimal MAD for every macroblock of the processed frame to be obtained. Furthermore, minimizing the energy of the residual frame may optimize the weighting factors. The minimization process is applied on the high-energy region that may be formed by a quadtree procedure on the residual frame. The experimental results show that there is an improvement in the overall performance with the proposed stereo coding scheme, over the typical *compatible* method of coding.

Acknowledgements

This work was supported in part by the Research Committee of the University of Athens under the project Kapodistrias.

6.References

[1] M. G. Perkins, "Data compression of stereopairs", *IEEE Trans. Commun.*, vol. 40, pp. 684-696, Apr. 1992.

[2] S. Sethuraman, A. G. Jordan, M. W. Siegel, "Multiresolution based hierarchical disparity estimation for stereo image pair compression", *Proc. of the Symposium on Application of subbands and wavelets*, March 1994.

[3] D. Tzovaras, M. G. Strintzis, H. Sahinoglou, "Evaluation of multiresolution block matching techniques for motion and disparity estimation", *Signal Proc: Image Commun.*, vol. 6, pp. 59-67, 1994.

[4] M. Ghanbari, et al., "Motion compensation for very low bit-rate video, *Signal Proc: Image Commun.*, vol. 7, pp. 567-580, 1995.

[5] D. Tzovaras, N. Grammalidis, M. G. Strintzis, "Object-based coding of stereo image sequences using 3D motion/disparity compensation*", IEEE Trans. CSVT*, vol. 7, pp. 312-327, Apr. 1997.

[6] S. Sethuraman, M. W. Siegel, A. G. Jordan, "Segmentation Based Coding of stereoscopic Image Sequences", *Proc. of the SPIE*, San Jose, 1996.

[7] H. Aydinglou, H. Hayes, "Stereo image coding: A projection approach*", IEEE Trans. IP*, vol. 7, pp. 506-516, Apr. 1998.

[8] M. S. Moellenhoff, M. W. Maier, "Transform coding of stereo image residuals", *IEEE Trans. IP*, vol. 7, pp. 804-812, June 1998.

[9] O. Woo, A. Ortega, "Overlapped block disparity compensation with adaptive windows for stereo image coding", *IEEE Trans. CSVT*, vol. 10, pp. 194-200, Mar. 2000.

[10] J. Jiang, E. A. Edirisinghe, "A Hybrid Scheme for Low Bit-Rate Coding of Stereo Images", *IEEE Trans. IP*, vol. 11, no. 2, pp. 123-134, Feb. 2002.

[11] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform", *IEEE Trans. IP*, vol. 1, no. 2, pp. 205-220, Apr. 1992.

[12] K. Ramchandran, M. Vetterli, "Best Wavelet Packet Bases in a Rate-Distortion Sense", *IEEE Trans. IP*, vol. 2, no. 2, pp. 160-175, Apr. 1993.

[13] J. M. Shapiro, "Embedded Image Coding Using Zero trees of Wavelet Coefficients", *IEEE Trans. SP*, vol. 41, no. 12, pp. 3445-3462, Dec. 1993.

[14] A. Said, W. A. Pearlman, "A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees", *IEEE Trans. CSVT*, vol. 6, no. 3, pp. 243-250, June 1996.

[15] S. D. Servetto, K. Ramchandran, M. T. Orchard, "Image Coding based on a Morphological Representation of Wavelet Data", *IEEE Trans. IP*, vol. 8, no. 9, pp. 1161-1174, Sept. 1999.

[16] D. Taubman, "High Performance Scalable Image Compression with EBCOT", *IEEE Trans. IP*, vol. 9, no. 7, pp. 1158-1170, July 2000.

[17] Siu-Wai Wu, A. Gresho, "Joint Estimation of Forward and Backward Motion Vectors for Interpolative Prediction of Video", *IEEE Trans. IP*, vol. 3, no. 5, pp. 684-687, Sept. 1994.

[18] A.Puri et al., "Video coding with motion compensated interpolation for CD-ROM applications", *Image Commun.,* vol. 2, no. 2, pp. 127-144, Aug. 1990.

[19] AVDS sequences from Carnegie Mellon University, Pittsburgh, PA, USA.