TEMPORAL AND SPATIAL SCALING FOR STEREOSCOPIC VIDEO COMPRESSION

Anil Aksay[1], *Cagdas Bilen*[1], *Engin Kurutepe*[2], *Tanır Ozcelebi*[2], *Gozde Bozdagi Akar*[1],
M. Reha Civanlar[2], *A. Murat Tekalp*[2]
1 Electrical and Electronics Engineering Department, Middle East Technical University
Ankara, Turkey
2 College of Engineering, Koç University, Istanbul, Turkey
{cbilen, anil, bozdagi}@eee.metu.edu.tr, {ekurutepe, tozcelebi, rcivanlar, mtekalp}@ku.edu.tr

ABSTRACT

In stereoscopic video, it is well-known that compression efficiency can be improved, without sacrificing PSNR, by predicting one view from the other. Moreover, additional gain can be achieved by subsampling one of the views, since the Human Visual System can perceive high frequency information from the other view. In this work, we propose subsampling of one of the views by scaling its temporal rate and/or spatial size at regular intervals using a real-time stereoscopic H.264/AVC codec, and assess the subjective quality of the resulting videos using DSCQS test methodology. We show that stereoscopic videos can be coded at a rate about 1.2 times that of monoscopic videos with little visual quality degradation.

1. INTRODUCTION

The stereoscopic video structure consists of two video sequences captured by closely located (similar to the distance between two eyes) cameras. The close location of cameras in these applications results in a high redundancy between the sequences from each camera. The efficiency of coding these sequences as simulcast (coding each sequence as monoscopic video) can be improved by the use of a different codec structure that removes these redundancies between sequences from closely located cameras.

A stereoscopic video codec based on H.264/AVC is introduced and discussed in [1]. In this codec structure the left frames only refer to the other left frames whereas the right frames refer to all the previous frames for compatibility with monoscopic H.264/AVC standard. More research goes on for multi-view coding with a similar kind of idea.

We have implemented a H.264/AVC based multi-view codec [2] which removes those redundancies in between two monoscopic videos. However the performance of this codec is video dependent and coding gain is sometimes below %20.

There are two different theories about the effects of unequal bit allocation between left and right video sequences, namely *fusion theory* and *suppression theory* [3][4][5]. In fusion theory, it is believed that the stereo distribution must be equally made for the best human perception. On the other hand, in suppression theory, it is believed that the highest quality image in the stereo-pair determines the overall perception performance. Therefore, according to this theory, we can compress the target image as much as possible to save bits for the reference image, so that the overall distortion is the lowest. If we assume that the overall distortion measure of a stereo-pair will be a weighted average of the individual images, we can define weighting coefficients between right and left image distortion values to take different amount of contributions from each picture into account.

In monoscopic video coding, chrominance values are usually subsampled, since Human Visual System (HVS) is less sensitive for chrominance values. Similar to this behavior and theories for stereo perception, it is reported in [6] that, HVS can use the high frequency information in one of the videos if the other video is low pass filtered. The authors proposed using spatial subsampling in one of the videos to reduce bandwidth requirements without any visual quality degradation. Authors also tried temporal scaling, but visual quality results show that spatial scaling gives more promising results.

In this work, we have implemented temporal and spatial downsampling in our H.264/AVC based multi-view codec. We have experimented with several parameters and conducted a subjective quality test with the coded sequences. In the following sections, we will explain the structure of codec and the modifications, quality test setup and the results.

2. CODING

2.1. MMRG Multiview Codec [2]

MMRG Multiview Codec is a H.264/AVC based codec which is designed for coding multiview sequences, i.e. sequences with camera number greater than 1 [2]. By setting the number of cameras to two, stereoscopic sequences can be coded using this codec. In stereoscopic coding with standard compatible mode, any standard H.264/AVC decoder can decode the sequence as a monoscopic sequence since left channel is coded independent of right channels. In order to decode right channel as well, special decoder is necessary. The structure of the encoder and decoder is shown in Figure 1 and Figure 2 respectively. In order to improve the coding gain without any significant perceptual quality loss, we added two modes called spatial and temporal scaling.

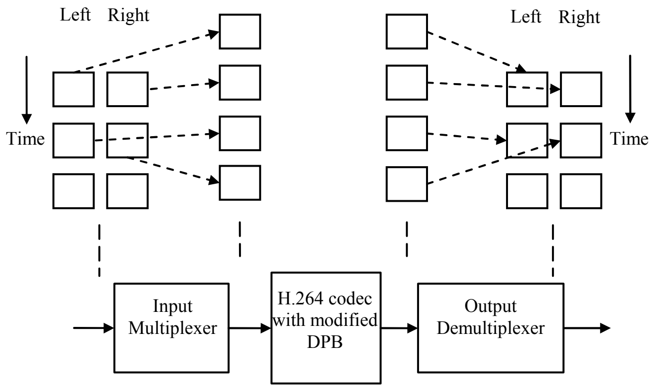

Figure 1: Stereoscopic Encoder

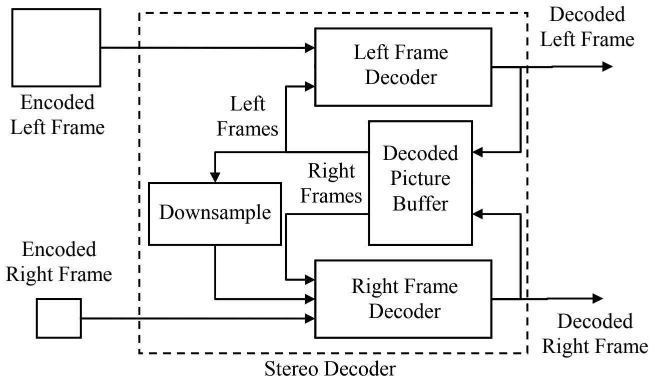

Figure 2: Stereoscopic Decoder

2.2. Spatial Scaling

The spatial scaling mode corresponds to downsampling the right video by a predefined scale prior to encoding in order to improve the coding gain. The implementation of downsampling the image consists of both decimation and low-pass filtering in order to prevent the aliasing. For spatial scaling following filters are used:

13-tap downsampling filter:
{0,2,0,-4,-3,5,19,26,19,5,-3,-4,0,2,0}/64
11-tap upsampling filter:
{1, 0,-5, 0, 20, 32, 20, 0,-5, 0, 1}/64

Filters are applied to all Y, U, and V channels and in both horizontal and vertical directions. The picture boundaries are padded by repeating the edge samples. These filters are currently used in Scalable Video Coding extension of H.264/AVC [7] and explained in [8]. In order to keep filtering process simpler in both encoder and decoder, we have implemented downscaling by factors of 2 (dyadic sampling) in both dimensions. Although the spatial scaling is applied to the right channel only, left frames are also temporarily scaled just for disparity estimation required for right frame coding.

2.3. Temporal Scaling

Temporal scaling mode corresponds to the decimation of right video in time, i.e. frame dropping in the right sequence. The implementation of temporal downsampling is done by sending all the macro blocks of dropped frame as skipped mode of the H.264/AVC standard. In our codec notation, temporal scaling of n denotes encoding 1 frame out of n frames and dropping the remaining n-1 frames.

3. TEST METHOD

We have adapted DSCQS Test method where non-experts and inexperienced assessors are used. The two videos are evaluated by the assessor on a continuous scale ranging from 0 to 100 with help of two sliders.

3.1. Test Methodology

Multiple assessors are shown two conditions, A and B (two stereoscopic images), consecutively one of which is always the source and the other is the tested condition applied on the source. The identity of the images, whether it is the source or the test condition, should be known by the experimenter but not by the assessors. The next pair of conditions is shown after the assessors establish an opinion.

Analysis Method: For the analysis of the test results, each evaluation is graded between 0-100 and the difference between the scores of source image and the test condition is calculated to find the score of that test condition on that image by the assessor. After all these scores are calculated, the values are normalized to fit in 0-100. And as a final step, to find the scores of each algorithm (test condition) the average of all the scores over the assessors and images are taken. Scores of the algorithms can be compared with their closeness to the number to which zero score is mapped during the normalization process.

3.2. Display System

Subjective evaluation of the encoded stereo videos was conducted at Koç University using a pair of Sharp MB-70X projectors. Light from one of the projectors is polarized in clockwise direction and light from the other projector in counter-clockwise direction using circular polarization filters. Although circular polarization is more expensive, it is preferable over linear polarization in stereo projection applications due to its rotation invariant polarization properties. When the light from the projectors is linearly polarized there are four distinct polarization planes at work: two from the projectors and two from the viewer's glasses. In such a setup, the projection planes of corresponding projectors and eye glasses must be parallel to each other and orthogonal to the other projector and eye glass to ensure only the light from the desired projector is let through unhindered, and the light from the other projector is completely blocked out. In such a setup, the subjects should be prevented from tilting their heads to preserve the orientations of the filters. Otherwise, light from one projector will start leaking to the opposite eye, spoiling the illusion of 3D. In the extreme case when a subject's head tilts by 45^{o} the polarization effect will disappear completely and the subject will only see two garbled images on top of each other. Circular polarization on the other hand is invariant to the orientation of the filters. Therefore, in our setup, the subject's glasses will optimally filter out the light from the opposite projector regardless of the position of the subject's head.

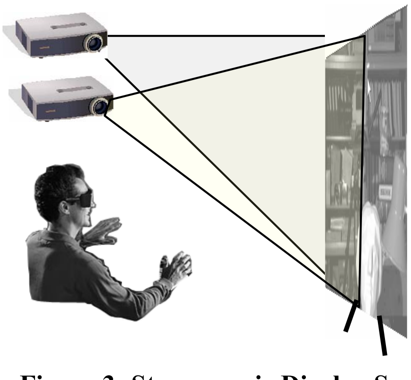

Figure 3: Stereoscopic Display System

In the experiment setup front projection is used to save space. Both projectors are aligned to project onto a "Silver Screen" covered with a neutral grey reflective dielectric material to preserve the polarization of light during reflection. The subjects wear glasses which have matching filters with the projectors to ensure that light from one projector is only seen with one eye. This enables us to feed left and right images to left and right eye of the subject to create the illusion of 3D.

The projectors were driven by a single high-end PC with two display outputs using the extended desktop feature of Windows XP. This setup results in a virtual desktop of 2048x768 pixels, each projector displaying only one half of the extended desktop at 1024x768 pixels native resolution. But since the projectors are aligned to project on the same area on the silver screen this wide desktop appears to be a normal 1024x768 desktop and a pixel, p_1, with coordinates (x, y) projects on the same physical spot on the screen as another pixel, p_2, with coordinates (x + 1024, y). Using this relation, left and right videos can be easily shown on the left and right halves of the extended desktop, such that they exactly overlap with each other on the Silver Screen.

The Silver Screen is 2500mm by 1900mm in size and the assessors sit approximately 3 meters away from the screen. The test videos are shown at their native resolution. We do not perform any upsampling to fit the videos to screen resolution since that could result in additional artifacts, which might bias and distort the assessors' subjective evaluation. In accordance with DSCQS guidelines, the empty area of the screen is set to 50%-gray during playback of test sequences.

4. EXPERIMENTS

In this experiment, we investigated effects of spatial and temporal scaling in stereoscopic videos. In order to meet time requirements of assessment test, we use only 4 video sets with 8 algorithms.

Assessors: 21 assessors (13 female, 8 male with average age 24) with ages ranging from 19 to 36, volunteered to participate in the experiment. The participants were non-experts in the area of picture quality and were screened for color vision, stereo depth perception and visual acuity.

Each assessor is well informed on the test process and test materials (possible quality defects) before the test and they are assisted during the whole test procedure. DSQCS test method with the second variant mentioned above is used as the test methodology. At each step two video sequences, original left and right videos and processed left and right videos are used. We will call those 4 videos an evaluation pair. In the experiments, original videos are also repeated as a processed video in order to test the performance of the test.

At the beginning of the test, 5 random evaluation pairs are shown to the assessors and these 5 evaluation pairs are not evaluated since they provide stabilization of the perception of assessors. The test material is shown in a random order for each assessor. The randomization is done both among evaluation pairs and among the set of video sequences in the pair.

Test Material: As the test material, four different stereoscopic video pairs are used: Balloons (720x480, 25 fps, 10 seconds), Botanical (960x540, 15 fps, 5 seconds), Flowerpot (720x480, 25 fps, 10 seconds), Train (720x576, 25 fps, 10 seconds). 8 different algorithms are applied on these videos as shown in Table 1. Sample frames are shown in Figure 4 and Figure 5. As a result a total of 41 evaluation pairs, including first 5 stabilizing pairs, are shown to the assessors and it is assured that each test does not take more than 30 minutes.

ORIG	Original
SIMUL	Simulcast coding
S1T1	Stereo coding, no spatial, no temporal scaling
S1T2	Stereo coding, no spatial, temporal scaling 2 for right frames
S1T2L	Stereo coding, no spatial, temporal scaling 2 for left and right frames
S1T3	Stereo coding, no spatial, temporal scaling 3 for right frames
S2T1	Stereo coding, spatial scaling 2, no temporal scaling
S4T1	Stereo coding, spatial scaling 4, no temporal scaling
S4T3	Stereo coding, spatial scaling 4, temporal scaling 3 for right frames

Table 1 Algorithms applied to test videos

	BALN	FLOW	BOTA	TRAIN	Av.
SIMUL	2.000	2.000	2.000	2.000	2.000
S1T1	1.901	1.927	1.452	1.881	1.790
S1T2	1.606	1.692	1.289	1.601	1.547
S1T2L	1.324	1.450	0.923	1.336	1.258
S1T3	1.489	1.586	1.228	1.492	1.449
S2T1	1.242	1.267	1.065	1.252	1.207
S4T1	1.091	1.095	1.012	1.085	1.071
S4T3	1.053	1.069	1.006	1.049	1.044

Table 2 Normalized Bitrates of Algorithms

5. RESULTS

All the test videos are encoded with the modes explained in Table 1. Intra period of 25 and Quantization Parameter (QP) of 28 are used while encoding. Total bitrate for simulcast coding is interpreted as twice the data required compared to single view coding and the bitrates of all other algorithms are normalized accordingly and can be found in Table 2.

By only spatial subsampling of right video with 2 in both dimensions we have approximately match 1.2 times the single view bitrate.

After all the assessors finish the test, the scores are evaluated and normalized. Average MOS scores and confidence intervals for each algorithm is shown in Table 3 and Table 4. Due to the normalization, 0 (best quality) is mapped to 38, and the success of the algorithms can be measured by closeness of their mean to 38. As it can be seen, the mean of the original video is also not exactly 38, which is due to the misjudgment of the assessors and it is expected.

Simulcast (SIMUL) coding and stereo coding without scaling (S1T1) have similar or better performances over original video. Since QP is low, reconstructed video quality is visually lossless (with average PSNR of 36 dB) and misjudgment is expected for these algorithms as well. Also DCT based coded images are reported [9] to be preferred by assessors comparing to original.

Algorithm	MOS	Normalized Bitrate	SNR Overall	SNR Left	SNR Right
ORIG	43.60	N/A	INF	INF	INF
SIMUL	41.93	2	36.09	36.05	36.13
S1T1	42.11	1.79	35.96	36.05	35.88
S2T1	46.68	1.207	31.05	36.05	28.8
S1T2	48.68	1.547	32.35	36.05	31.11
S1T2L	48.71	1.258	31.15	31.19	31.11
S4T1	53.02	1.071	27.82	36.05	25.15
S1T3	55.57	1.449	30.58	36.05	28.92
S4T3	59.26	1.044	26.13	36.05	23.37

Table 3 MOS Scores of algorithms

We can see from the results that scaling with 3 or 4 in both spatial and temporal domain is not acceptable. According to the bitrate and MOS score, only spatial scaling looks like the optimum solution. Spatial scaling by 4 corresponds to 16:1 reduction in image size; therefore its performance is not acceptable. Spatial scaling with non-dyadic factors and better filters for upsampling might keep the visual quality at desired levels with bitrate similar to single view coding bitrates.

According to the video characteristics (slow motion video), temporal scaling in either right channel or both channels might give good results as well. By analyzing the characteristics of the video in each GOP (each chunk of video sequence that can be decoded without use of other parts of the sequence), appropriate scaling can be applied to decrease bitrate without visual quality degradation.

	BALN	FLOW	BOTA	TRAIN
ORIG	43.0 +- 0.6	44.1 +- 0.8	42.0 +- 0.4	45.2 +- 0.8
SIMUL	42.4 +- 0.9	40.0 +- 0.8	43.7 +- 1.0	41.7 +- 1.3
S1T1	41.5 +- 1.1	39.7 +- 0.9	45.5 +- 0.9	41.8 +- 1.5
S2T1	46.9 +- 1.0	47.5 +- 1.2	46.0 +- 1.0	46.3 +- 1.1
S1T2	45.3 +- 1.4	45.8 +- 1.3	57.0 +- 1.8	46.6 +- 1.1
S1T2L	47.2 +- 1.5	51.7 +- 1.8	52.0 +- 1.1	43.9 +- 1.4
S4T1	51.5 +- 1.6	52.9 +- 1.7	54.0 +- 1.7	53.7 +- 1.4
S1T3	56.5 +- 1.7	54.7 +- 1.7	60.0 +- 1.7	51.0 +- 1.4
S4T3	53.3 +- 1.6	61.2 +- 1.5	67.5 +- 1.7	55.0 +- 1.6

Table 4 MOS Scores and confidence intervals for each video sequence

Figure 4 Sample frames of Train sequence

Figure 5 Sample frames of Botanical sequence

6. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed several sampling methods for decreasing the bitrate of stereoscopic video sequences. We have implemented the algorithms on a H.264/AVC based stereoscopic codec and assess the subjective quality of the methods with DSCQS test.

We have shown that stereoscopic videos can be coded at a rate about 1.2 times the monoscopic video with little visual quality degradation. We believe that this ratio can be lowered approximately close to 1 by appropriate filtering for each GOP with optimized filtering parameters.

In the future we are planning to investigate adaptive scaling of right channel according to the video characteristics.

7. ACKNOWLEDGEMENTS

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

The authors would like to thank Gokhan Demirel for his help while conducting subjective tests.

MMRG H.264/AVC Multiview Extension Codec can be downloaded from METU MMRG (Middle East Technical University MultiMedia Research Group) Website: http://mmrg.eee.metu.edu.tr/multiview

8. REFERENCES

[1] B. Balasubramaniyam, E. Edirisinghe, H. Bez, "An Ex-tended H.264 CODEC for Stereoscopic Video Coding", Proceedings of SPIE 2004.

[2] C. Bilen, A.Aksay, G. Bozdagi Akar, "A Multi-View Video Codec Based on H.264", ICIP 2006 (accepted).

[3] B. Julesz, "Foundations of Cyclopeon Perception," The University of Chicago Press, 1971.

[4] I. Dinstein, M.G. Kim, A. Henik, and J. Tzelgov, "Compression of stereo images using subsampling transform coding," Optical Engineering, vol. 30, no. 9, pp. 1359-1364, Sept. 1991.

[5] W. Woo, A. Ortega, "Optimal Blockwise Dependent Quantization for Stereo Image Coding," IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, pp. 861-867, September, 1999.

[6] Lew B. Stelmach, Wa James Tam, Dan Meegan, André Vincent: "Stereo image quality: effects of mixed spatio-temporal resolution", IEEE Trans. Circuits Syst. Video Techn. 10(2): 188-193 (2000)

[7] Julien Reichel, Heiko Schwarz and Mathias Wien, "Scalable Video Coding – Working Draft 3", JVT-P201, Poznan, PL, 24-29 July, 2005.

[8] C. A. Segall, "Study of Upsampling/Down-Sampling for Spatial Scalability", JVT-Q083, Nice, FR, PL, 14-21 October, 2005.

[9] A. Schertz, "Source coding of stereoscopic television pictures," in Proc. IEE Inter. Conf. Image Processing and its Applications, Maastricht, The Netherlands, Apr. 7–9, 1992, pp. 462–464