Empirical Study of 3D Video Source Coding For Autostereoscopic Displays

Roger Cheng
University of Illinois
Dept. of Electrical and Computer Engineering
Urbana, IL 61801
rcheng2@uiuc.edu

Klara Nahrstedt
University of Illinois
Dept. of Computer Science
Urbana, IL 61801
klara@cs.uiuc.edu

ABSTRACT

The recent commercial availability of autostereoscopic displays has led to a rise in interest in 3D video. The need to store and transmit 3D video has created some interesting challenges. 3D video contains both color and depth information, and should be treated differently from 2D video for optimal results. This paper explores ideas for efficient 3D video source coding, tests these ideas in a human subject test with 27 subjects, and analyzes and discusses the results. It is concluded that for the specific display and codec used, 3D video file sizes can be reduced to about one-quarter of their original sizes without significant degradation in quality.

Categories and Subject Descriptors: H.5.2 User Interfaces: Evaluation/methodology, I.4.2 Compression(Coding): Approximate methods

General Terms: Experimentation, Human Factors

Keywords: ACM proceedings, 3D video, autostereoscopic display, source coding, video quality, applied psychology

1. INTRODUCTION

An autostereoscopic display provides 3D images and depth perception without requiring the viewer to wear special glasses or headgear [3]. They provide an interesting experience when fed with 3D videos. 2D videos can be converted to 3D format, but for the best quality 3D videos should be captured or rendered natively. Unlike 2D videos, 3D video files require storing both color and depth information. Much research has been done on 2D video coding and compression, and that domain is well understood. However, 3D video coding and compression is not as well understood; currently, the information in the color and the depth domains are frequently stored with the same resolution and compressed with the same parameters that a 2D video would use. Since the human visual system perceives 3D video differently than it does 2D video, this is likely not the best implementation. In order to determine how best to deal with the additional dimension of depth and in order to optimize the source coding of 3D videos, we need to examine how sensitive the human visual system is to certain factors of 3D videos.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

2. 3D VIDEO SOURCE CODING

Some previous work has been done on the source coding of 3D videos. Both Stelmach and Tam [5] and Christodoulou et al. [1] looked at the effect that coding the left eye and right eye views (of a video) at different bitrates had on subjective image quality. The former party concluded that disparate quality coding results in an overall quality around the average of the quality of the two views, while the latter party concluded that an overall "good experience" was attained when the lower quality image was unacceptable by itself. In a separate paper Tam and Zhang [6] said that the same depth frame information can be repeated across separate color frames without visible artifacts.

We performed a survey of a number of 3D videos, and we saw that most depth frames are comprised of simple shapes and edges. Furthermore, by looking at the DFT (discrete Fourier transform) of the depth frames, we see that the vast majority of frequency domain content is concentrated at the low frequency bands. The simple nature of the depth frames, compared to their color frame counterparts, is one major reason it is hypothesized that the depth information can be stored effectively at a lower resolution.

3. TEST METHODOLOGY

As previously mentioned, the goal of this study is to determine how sensitive the human visual system is to certain factors of 3D videos, and to what extent these factors can be modified without exhibiting a significant degradation in quality. This section describes the methodology used in designing and conducting the experiment.

3.1 System Description

The Philips 42-3D6W01 autostereoscopic display was used for the research described in this paper. It features a 42 inch LCD panel (measured diagonally) with a 2D resolution of 1920x1080, coupled with a fixed-lenticular screen that provides the capability for 3D effects. The optimal distance from which to view the screen is 3 m [2] (subjects were tested at this distance).

Philips terms its 3D frame format "2D plus depth" [2]. Figure 1 shows an example 3D frame created by Philips. In short, the color information is fed to the left hand side of the screen, the depth information to the right hand side, and the display uses this information to create the 3D image that the

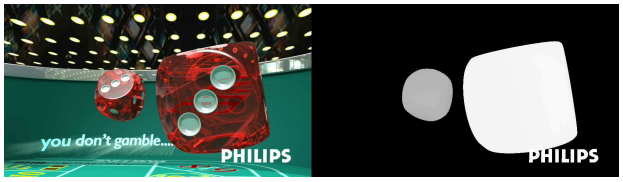

Figure 1: A 3D video frame

viewer sees. Note that "depth" in this case does not refer to absolute depth or Z-value, but rather disparity, which is inversely proportional to absolute depth. Three bytes per 3D pixel are used for color, and one byte per 3D pixel is used for depth. Three bytes are actually allocated to each depth pixel, and so two bytes per 3D pixel are wasted. This is a fairly inefficient implementation, however, this study and paper will not focus on improving this aspect, and we will assume that only four bytes per 3D pixel are being used.

3.2 Test Materials

Twelve 3D videos in Philips' format were downloaded from www.philips.com/3dsolutions and www.wowvx.com. A 35 to 40 second long clip was extracted from each of these videos, and these clips formed the bases of the test videos. Since some videos had audio tracks and others did not, all existing audio tracks were edited out so that sound was not a factor.

The videos were divided up into 4 groups of 3 videos each, and each group was tested on different factor(s). For each test, each video was modified to 2 separate levels (explained below). Each video was first converted to full uncompressed .avi format before the modifications were done. The modified .avi files were then encoded with the Microsoft WMV Series 9 codec with the standard "video only" profile (as recommended by Philips). Since the encoding process introduces additional distortion to the modified videos, the original videos were also re-encoded with the same process to ensure fairness in this regard.

3.3 Test Subjects and Schedule

Volunteers were gathered to serve as test subjects. Since it is helpful to have some background in image and video coding and processing in order to evaluate the quality of the videos, students studying ECE and CS were targeted (although all interested people were invited). The subject pool consisted of 27 people, of which there were 19 males and 8 females, 19 ECE/CS students and 8 "others", ranging from 21 to 35 years of age.

For each of the 4 groups of videos, each subject was shown 3 chosen video clips, comprising one video of each different content and one of video each different quality. This process prevents subjects from directly comparing two different quality variants of the same video content. Between subjects, the content and variant association of the chosen clips was shuffled according to a "Latin Squares" test schedule, in order to minimize the influence of the order of content and variant.

3.4 Test Metrics and Phase Approach

Since it is difficult to attain an objective measurement of quality, subjective criteria were used. Both quantitative and qualitative test metrics were used. The subjects were asked to fill out a questionnaire in which they evaluate the videos that they saw. After watching each video, the subjects were asked the following three questions:

- Rate the quality of 3D effect in this video on a scale from 1 (low quality, unimpressive) to 10 (high quality, very impressive)
- Rate the overall quality of this video on a scale from 1 (low quality, unwatchable) to 10 (high quality, very clear, detailed, and smooth)
- Briefly describe the above qualities, including specifically what was good or bad, whether there was a consistent or fluctuating quality, and pointing out anything interesting or peculiar

At the end of each group of three videos, one additional question was asked:

- For the video(s) that you thought had poorer quality, what in particular could be improved?

The experiment was divided into 3 phases. Each phase tested 9 subjects, which resulted in 3 tests per unique video. The motivation for this approach is to help select suitable factors and parameters to test. For example, if it is determined in phase I that altering the frame rate yields a very significant drop in scores, then for phase II frame rate should either be dropped from testing or should be modified more conservatively.

4. RESULTS

4.1 Phase I

These factors were tested, in the following order:

- 3D awareness: Firstly, the 3D effect was slowly curtailed as the video played along, until finally there was none. This tests whether "change blindness" [4] occurs, which states that the human visual system is particularly bad at recognizing certain changes in videos. Secondly, all depth values were fixed at 128 (which is the middle of the range of 0 to 255). Thus the subjects would see an image projected onto the same depth plane that conveyed no useful or interesting depth information.
- Frame rate: The original videos were encoded at 30 fps, which is high enough to ensure smooth 2D videos. Since depth perception provides an additional "distraction" to the viewer, it was hypothesized that a lower frame rate is sufficient for 3D videos to look smooth. The altered videos were encoded at 20 and 15 fps.
- Depth interpolation: Instead of storing a unique depth frame for each color frame, only one out of every n depth frames was kept, and the missing depth frames were computed via linear interpolation of the stored depth frames. The altered videos stored every third and fifth frame.

- Depth resolution: Each depth frame was decimated in both the horizontal and vertical directions, so that only every *nth* pixel would be stored (in each direction). Then the lower resolution depth frame was interpolated upwards to the original resolution, for playback purposes. In order to avoid aliasing, the frames had to be filtered with a low-pass filter with a cutoff frequency of π/n prior to the downsampling operation and again after the upsampling operation. The altered videos were decimated by 8 and 16.

Figure 2 summarizes the 3D and overall quality scores for this phase. Note that for each group, the left line represents the original video, the middle line represents the modified, middle quality video, and the right line represents the modified, lowest quality video; for example, for the frame rate group, the left line represents 30 fps, the middle line 20 fps, and the right line 15 fps. For each line, the mark is the mean score, and the top and bottom of the bar represent one standard deviation above and below the mean, respectively.

The 3D awareness test resulted in very low scores for the modified videos, suggesting that the subjects could easily pick out the modified videos. For the frame rate test, the 3D scores of the modified videos was slightly higher, and their overall scores were slightly lower. For the depth frame interpolation test, using every third frame resulted in a surprisingly good score while using every fifth frame resulted in a much worse score. Finally, for depth frame resolution, the 3D scores took a significant hit with the modified videos, while the overall scores only decreased slightly.

4.2 Phase II

These factors were tested, in the following order:

- Depth resolution: Since decimating by 8 and 16 resulted in significantly lower scores, more conservative values of 2 and 4 were chosen.
- Depth quantization: Instead of using all integers from 0 to 255 for the depth values, a coarser granularity would be used. It is hypothesized that the human eye cannot distinguish between similar, but different depth values. When quantizing by a factor of n, only multiplies of n are used. The modified videos were quantized by factors of 4 and 8.
- Frame rate: Since using 20 and 15 fps did not yield a significant drop in scores, the envelope was pushed further by using 12 and 10 fps.
- Depth range: Instead of using the full range of 0 to 255, all depth values would be scaled back to occupy a smaller range. Scale factors of 1/2 and 1/4 were chosen (resulting in max values of 127 and 63, respectively)

The results are summarized in Figure 3. Similarly to the 3D awareness test in phase I, the depth range test resulted in very low scores for the modified videos. Also similarly to phase I's results, the frame rate test resulted in no significant deviation in scores between the three quality variants. For the depth quantization test, a small improvement in scores was seen when quantizing by 4 while a small drop was seen with a factor of 8. Perhaps the most interesting result was seen in the depth resolution test, as a sizable increase in scores was seen when decimating by 4.

4.3 Phase III

Unlike in phases I and II, several factors were modified and tested in each group of videos. The three factors (and respective parameters) used were depth resolution (decimating by 2 and 4), depth quantization (by 2 and 4), and frame rate (15 and 10 fps). These factors and parameters were selected based on the fact that in phases I and II they resulted in nearly constant scores. In three of the groups of videos, two factors were chosen, while in the fourth group, all three factors were chosen. For consistency reasons, the "better" parameters were paired together, and likewise for the "worse" parameters; i.e., for the frame rate / depth resolution test, one video was decimated by 2 and encoded at 15 fps, and the other was decimated by 4 and encoded at 10 fps.

The results are summarized in Figure 4. Perhaps not surprisingly, the largest drops in scores were exhibited when all three factors were modified simultaneously. Frame rate / resolution resulted in small drops in both scores. Quantization / resolution resulted in a slightly higher score for the first modification level, and a slightly lower one for the second modification level. Frame rate / quantization exhibited constant overall scores, and actually resulted in slightly higher 3D scores.

4.4 Summary of qualitative results

The most frequent comment given was that the videos were too blurry; this happened with both the original and the modified videos. This is in part due to the relatively low resolution of the color information in the videos (960x540) coupled with the large screen size. By design, the resolution is required to be 960x540 for this particular display. This is also likely in part due to the way the 3D effect is generated. The fixed-lenticular screen on top of the LCD panel contains cylindrically shaped lenses that redirect light from the LCD panel; the lenses certainly add some kind of distortion.

The subjects had trouble identifying and explaining the true nature of the problems they were seeing. For the 3D awareness test, 5 of the 9 subjects said something to the effect of "the video did not look 3D," but nobody explained that the 3D effect was being curtailed or that the image was projected onto the same depth plane. Of the 4 people who did not make this comment, 3 of them did score the altered videos lower, perhaps indicating that they realized that something was odd but could not explain what it was. For the frame rate test, only 4 of the 27 subjects mentioned that the motion in the video was jerky or that the frame rate should be increased for better quality. Since the average scores across different frame rates were nearly the same in this test, it is believed that the scores were largely a function of the video's content rather than its frame rate.

4.5 Discussion of results

In the cases where the modified videos received slightly higher scores than the originals, it is most likely due to the inherent randomness in empirical studies. No modified videos received high enough scores such that it could be concluded that the modification actually made the video look better. On the other hand, on the 3D awareness, depth range, and depth interpolation tests, the modified videos received low enough scores such that it is very likely the case that those modifications made the videos look worse. These conclusions were drawn using the one-sided Student's t-test

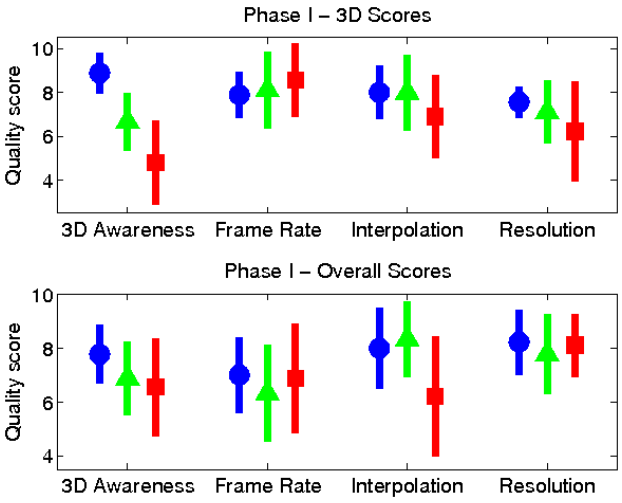

Figure 2: Phase I

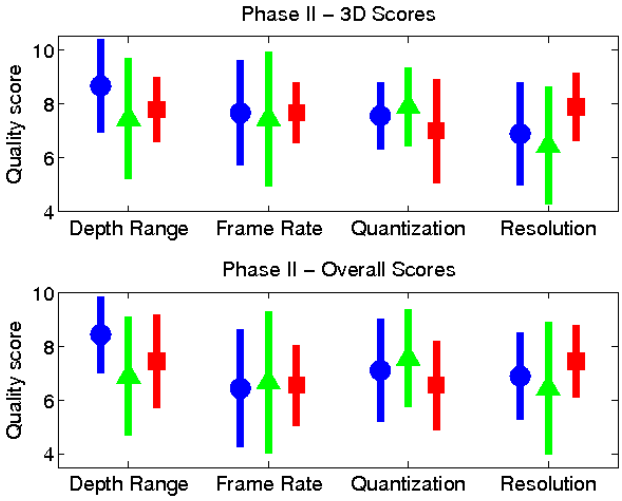

Figure 3: Phase II

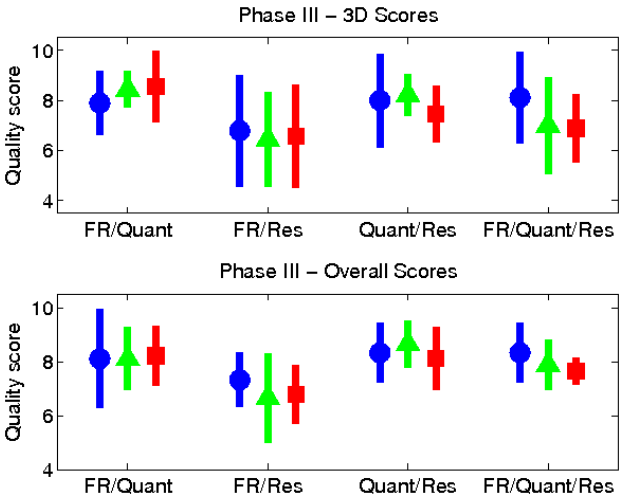

Figure 4: Phase III

with an alpha threshold of 0.05. Note that the t-test assumes normality of data, which was verified with the D'Agostino-Pearson test; all sets of data passed the test, although admittedly it is difficult for a data set of nine samples to fail this test.

One limitation of this study is the previously mentioned fact that we are using subjective criteria in our evaluations. Each person has a different perception of what "good" and "bad" looking videos are, and each person returned a set of scores with different means and variances. However, it turns out that when each subject's set of scores was normalized to have zero mean and unit variance, the results did not change; for example, the frame rate test still showed nearly equal means while the 3D awareness test still showed large deviations in means. Bias in terms of video content was also present, meaning that some videos received higher scores because they contained more interesting material. However, just as in the subject normalization case, when the scores were normalized according to their content, the results did not change.

5. CONCLUSIONS AND FUTURE WORK

It is fairly safe to say that Philips' current way of storing 3D video files is inefficient, and improvements can be made that lower file sizes without lowering quality. Based on the collected test results and factoring in ease or difficulty of implementation, reducing the frame rate is the best improvement, followed by depth frame decimation. If the frame rate is reduced to 10 fps and the depth resolution decimated by a factor of 4, this would reduce file sizes to a factor of 49/192 (just over one-quarter) of their original size. Admittedly, it is difficult to explain why the average viewer cannot see a difference in video quality after these operations; this matter is outside the scope of this paper and could make for an interesting study.

The conclusion drawn is based on the specific hardware and codec used, and further testing would be required to be able to generalize the conclusion. Also, there are several new directions that can be explored through further testing. For example, bitrate, an important 2D video compression parameter, was not tested in this study. If a method was developed that allows the color information and depth information to be coded at different bitrates, it would be interesting to see what an optimal allocation ratio would be between these two fields.

6. REFERENCES

[1] *3D TV Using MPEG-2 and H.264 View Coding and Autostereoscopic Displays*, 2006.

[2] *Philips 3D Solutions: 3D Interface Specifications*, 2006.

[3] N. Dodgson. Autostereoscopic 3d displays. *Computer*, 2005.

[4] D. J. Simons and M. S. Ambinder. Change blindness: Theory and consequences. *Current Directions in Psychological Science*, 14:44–48, 2005.

[5] L. B. Stelmach and W. J. Tam. Stereoscopic image coding: Effect of disparate image-quality in left and right eye views. *Signal Processing: Image Communication*, 14:111–117, 1998.

[6] W. J. Tam and L. Zhang. 3d-tv content generation: 2d-to-3d conversion. *IEEE ICME*, 2006.