Design and Evaluation of a 3D Video System Based on H.264 View Coding

Hari Kalva, Lakis Christodoulou, Liam M. Mayron, Oge Marques, and Borko Furht

Dept. of Computer Science and Engineering
Florida Atlantic University
Boca Raton, FL 33431
Phone: 561-297-2885

ABSTRACT

Recent advances in video compression and 3D displays have necessitated a further understanding and development of 3D video coding algorithms. The emergence of low cost autostereoscopic displays is expected to drive the growth of 3DTV services. This paper discusses key issues that affect the quality of 3D video experience on autostereoscopic displays. The characteristics of the human visual system can be exploited to compress individual stereo views at different qualities without affecting the perceptual quality of the 3D video. The H.264/AVC video coding algorithm was used to compress each view. We examine the bounds of asymmetric stereo view compression and its relationship to eye-dominance based on a user study. This paper also presents the design and development of a modular video player with stereoscopic and multi-view capabilities including a discussion of useful tools for accelerating the development and enhancing flexibility. The experimental results indicate that eye-dominance influences 3D perception and as a result will impact the coding efficiency of 3D video.

Categories and Subject Descriptors

I.4.2 [Image Processing and Computer Vision]: Compression (Coding) – Approximate methods

General Terms

Algorithms, Performance, Human Factors, Experimentation

Keywords

H.264, 3DTV, Eye dominance, asymmetric view coding

1. INTRODUCTION

The recent interest in 3D and multi-viewpoint (MV) TV can be attributed, in part, to the success of the MPEG-4 AVC/H.264 video coding standard. The coding gains made possible by H.264 can be applied to provide enhanced services such as multi-viewpoint TV and 3D television. Another reason for the increasing interest in 3D TV is the recent advances in the display technologies that have lowered the cost of stereoscopic projectors and 3D displays. While these technological advances have renewed interest in 3D/multi-view coding, the successful deployment of 3D services still faces key challenges. The current state of the technology and the maturity of the marketplace indicated that this is the right time to overcome barriers to 3D and MV TV services.

The digital video revolution launched by the MPEG-1 and MPEG-2 video coding standards also resulted in an active 3D and multi-view video coding research [1, 2]. The MPEG-2 multi-view profile is a form of temporal scalability that encodes left view of the stereo pair as a base layer and the right view is coded as a temporal enhancement. Existing studies on the quality of 3D video are based on MPEG-2 view coding and not applicable to H.264 based coding that is expected to be used in 3D TV services [3]. The studies also did not use autostereoscopic displays which are expected to be the dominant display types for 3D TV [4]. MPEG-2 based coding is inefficient compared to H.264 based view coding. Furthermore, the coding artifacts in MPEG-2 and H.264 are different and are likely to have different effects on the 3D perception. The quality of a 3D video experience is influenced by the type of displays used. A good summary of the perceptual quality requirements and evaluations for 3D video is presented in [4]. Our current focus is on developing efficient coding and representation algorithms for 3D and multi-view video. We are using H.264 as the basis for view coding and autostereoscopic displays for rendering the 3D video.

One of the reasons for the lack of success of 3D TV so far is the ease-of-use of the 3D TV and the viewing comfort. Most of the displays today use standard TV with anaglyph video and a pair of glasses to generate 3D perception. Watching such TV is straining to the eye. Even the current generation autostereoscopic displays have limited viewing angle and are not suitable for viewing for longer periods. The application where 3D video has had reasonable success are the applications where viewing comfort is secondary to the objective; applications such as security, medicine, design automation, and, scientific visualization.

This paper is organized as follows. Section 2 gives an overview of the 3D/multiview video system we are developing including a short overview of stereo perception in the human visual system. The player architecture and tools used are discussed in section 3. Section 4 presents the experimental methodology and the results are discussed in section 5. Finally, conclusions are presented in section 6.

2. OVERVIEW OF MULTIVIEW VIDEO SYSTEM

We are currently developing a 3D/multi-view video coding system with an initial focus on security and surveillance. The goal of this project is to develop technologies and tool for efficient compression, communication, and playback of multi-view and 3D video.

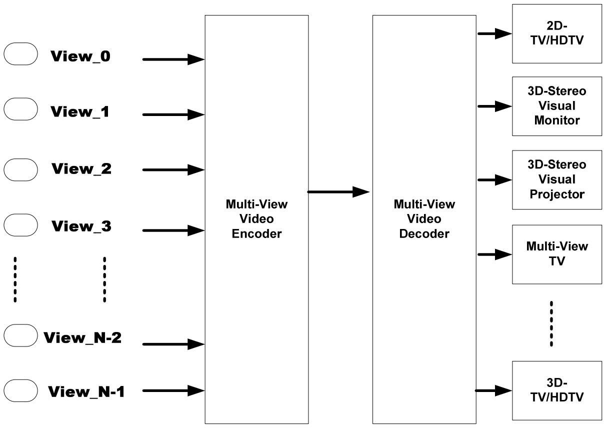

Figure 1. 3D/Multiview video system

Figure 1 shows the general architecture of a multi-view video system. The multiple views are encoded at the sender by exploiting the large amount of redundancies among the views. We use H.264 as the core compression engine with inter-view prediction to increase compression efficiency [5]. The coded views are communicated to the receiver where the decoded views are rendered on an appropriate display. The 3D displays use a pair of coded views to display 3D video with depth perception.

2.1 Brief Overview of Binocular Vision

The human visual system receives two separate projections of a scene; one from each eye. The eyes are separated by an average horizontal distance of 6.3 cm [7]. The stereoscopic image is an image synthesized by the monocular left-eye-view and the monocular right-eye-view causing relative viewing projections described with high correlation, but with different image information. The left and right eye views are combined resulting in a single 3D percept. The combined visual perception of the scene is also known as binocular fusion. Binocular suppression is property where portions of the view in one eye are suppressed by the corresponding view of the other eye. The possibilities of dominance and suppression mechanisms during the binocular fusion exist, but their impact is not yet well understood [7]. Experiments have shown that when the left and right eye views are combined the higher quality view is able to mask coding artifacts in the lower quality view [3,8].

The process of binocular fusion in the human visual system results in the comparison and combination of the left and right eye views to generate a single 3D percept. The left and right eye views have to be presented to the users using 3D display means to give the sensation of 3D and depth perception. The left and right eye views can be encoded and sent to the receiver and the stereo views can be generated at the receiver. The properties of binocular fusion make possible encoding of left and right eye views at different bitrates. This asymmetric view coding has been exploited to improve compression efficiency [3,8]. The H.264 video coding used in our system is much more efficient than MPEG-2 and also has support for de-blocking that improves the perceptual quality of video. The effects of these improved compression algorithms and autostereoscopic displays on the 3D video quality cannot be understood from the past MPEG-2 based studies.

The two main approaches to delivering 3D video are 1) stereo coding where the left and right views are encoded and 2) depth image based rendering (DIBR) where a single view and an associated depth map are transmitted to the receiver [9]. DIBR systems synthesize the left and right views at the receiver based on the single view and the depth information. These two approaches have their advantages and disadvantages. However, from a production and compatibility point of view the stereo coding methods are more suitable. Furthermore, the free viewpoint TV (FTV) based on multi-view video coding (MVC) is gaining momentum and this makes DBIR approaches unnecessary as the MVC is sufficient to generate the left and right views necessary for the 3DTV.

3. STEREOSCOPIC & MULTI-VIEW VIDEO PLAYER

While the study stereoscopic visual stimuli is not new, it is a field that has seen renewed interest due to advances in capturing videos, mediums for broadcasting, autostereoscopic displays, and other viewing techniques. This section presents the architecture of a modular video player with stereoscopic and multi-view capabilities.

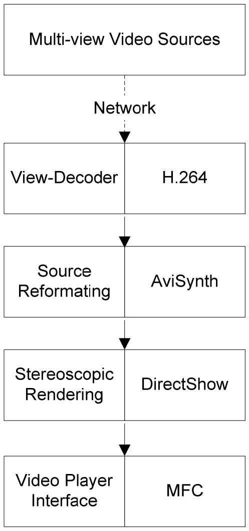

Figure 2. 3D/MV player architecture

3.1 Architecture

The player was implemented and tested on the Microsoft Windows XP platform. The Microsoft DirectShow framework was used for the capture and transform functions. MFC was used to implement the interface. An open source project, AviSynth [10], was used for some preprocessing tasks. The player takes a pair of views as input and renders them in a format suitable for the target display (anaglyph, Sharp 3D display, side-by-side, etc.).

The inputs can be from video decoded from the network or from local video sources (e.g., files, cameras). Figure 2 shows this general architecture.

DirectShow is a component of DirectX. DirectShow offers a modular architecture that allows runtime reuse of modules (known as DirectShow filters). The framework allows reusing existing filters for video capture, decoding, and rendering. Filters are connected via compatible terminals, known as pins. A collection of connected filters is referred to as a graph. A minimal graph consists of a source filter to decode media, a transform filter to perform a meaningful operation on the media, and a render filter to display the result on screen or write it to disk. Because our player deals with known and widely available video codecs we are not concerned with source and render filters. Additionally, the use of AviSynth abstracts an even wider variety of file formats that could not normally be played back (for example, raw YUV files) by presenting them as uncompressed AVI data to the player. Instead, the transform filter is where the majority of the processing takes place. In our project the transform filter changes depending on the choice of output format (monoscopic or a specific stereoscopic format).

There are many choices for the implementation of an interface. One option is simply to write a series of DirectShow filters that can be used with a variety of preexisting media players. The existing players lack support for multi-view and 3D sources and player thus needed a new interface. Windows MFC provides as much control over the interface as needed in a Windows environment and is well-documented.

We chose to include AviSynth in our project for several reasons. It is an open source project that has been in use for several years. As a result we trust the validity of its functionality, such as color-space conversions, and can verify the implementation for ourselves. Using AviSynth resulted in considerable time savings, enabling us to focus our work on our primary goal of rendering stereoscopic video.

3.2 Stereoscopic Video Playback

One of the challenges of displaying stereoscopic video is the wide variety of video formats. Stereoscopic video is typically available as independent left and right sequences or as a single video formatted with the left and right views side-by-side or top-to-bottom.

In the implemented solution we use the versatile AviSynth scripting language to help format stereo video data consistently for the stereo player. AviSynth is a frame server. It performs a variety of transformations on video files on-the-fly without creating other files. To the player application the AviSynth script appears as an uncompressed AVI file. In practice we found AviSynth to provide a useful layer of abstraction between the source data and the player, greatly reducing the complexity of the player.

The user must be able to specify the format of the source video data. For example, if we desire to playback left and right video data encoded in two separate files the AviSynth script needed would ensure that the videos are of equal length and resolution and then place them side-by-side with the left source to the left. This is the format that is expected by the video player. Similar transformations can be made for other formats. If the source is a single video in the side-by-side format no changes are needed. The AviSynth script needed to format the video for playback can be generated with the assistance of a GUI and does not need to be written by the user. The specification of a video format and the generation of the corresponding AviSynth file are performed only once.

3.3 Multi-view Playback

Our architecture supports the playback of monoscopic and stereoscopic multi-view video. We describe the location of cameras (or viewpoint of video sources) available for the user to select. Certain combinations of cameras (viewpoints) are indicated as valid pairs for stereo viewing. The user can then select this pair for stereoscopic viewing.

4. EXPERIMENTAL METHODOLOGY

The goal of this work is to understand the impact of the compression advances in H.264 video and the display advances in the autostereoscopic displays on the quality of the 3D video experiences. We are currently conducting a large user study to evaluate the impact of asymmetrically coded 3D views on the quality of the 3D video rendered on the Sharp autostereoscopic display. The goal of this study is to understand the bounds of asymmetric coding, relationship between the eye-dominance and 3D quality of asymmetrically coded video, and to understand the effects of the H.264 coding features that improve perceptual video quality. The results are reported based on the evaluations from 14 users that have evaluated the subjective quality so far.

The sequences used for these experiments are the Akko & Kayo and the Ballroom sequences created for 3D/mulitview coding work currently underway in the MPEG committee [11]. A pair of views from these sequences was chosen to render stereo video. The video sources are 10 seconds long, 640x480 resolution, 30 FPS, and available in YUV 4:2:0 format. The Akko & Kayo sequence is made specifically for this research and has a number of carefully selected objects that help evaluation of 3D sequences well. The Ballroom sequences capture ballroom dancing and show dancers at multiple levels of depth.

The test sequences were created to test 3D video at different levels of quality. The quality was varied by encoding the left and right eye views at different qualities. Two test cases were created for each video sequence: 1) right eye view at a high quality with left eye view quality varying and 2) left eye view kept constant at a high quality and the right eye view quality varying. The high constant quality views were encoded at a PSNR of 42.5 dB, considered broadcast quality, and the quality of the other view is varied from 42.5 dB to 28 dB. The discussion presented here uses PSNR for quality and deliberately avoids using bitrate as there is no standard way of encoding 3D video yet and the same quality can be achieved at different bitrates depending on the coding and prediction modes used.

Subjects were recruited to participate in this research and evaluate the 3D viewing experiences. This is an ongoing study and the results reported are for 16 subjects evaluating the test sequences. The participants evaluated the overall quality of video (without looking for specific artifacts) on the standard subjective evaluation scale from 1 to 5 (1-bad, 2-poor, 3-fair, 4-good, 5-excellent). Most of the participants have had 3D movie experience in the past but this evaluation was the first experience with autostereoscopic displays. Before beginning the evaluations, the participants were shown four high quality 3D video sequences including the two test sequences without any compression.

We used the Sharp LL-151-3D autostereoscopic display to render the stereoscopic videos. The display is 15-inches, XGA resolution (1024 by 768 pixels). This display which uses lenticular imaging techniques and renders depth very accurately gives a true 3D experience. The perception of depth is achieved by a parallax barrier that diverts different patterns of light to the left and right eye. It should be noted that our player architecture accommodates a variety of formats for 3D playback and can be extended to include others.

4.1 Quality Evaluation Tests

The users evaluated test sequences at a variety of qualities. The 10 second test sequences were presented in a random order on the 15-inch Sharp autostereoscopic 3D displays with a 5 second gray level image in between the test sequences. Figure 3 shows the presentation order used in the experiments. Each participant evaluated a total of 34 ten second 3D clips. The experiments used two different sequences encoded at varying qualities. To evaluate the impact of asymmetric coding, the test sequences were encoded such that quality of one view of the stereo pair is kept constant at a high quality while the quality of the other stereo view is varied from high to low quality. We used video coded at 42.5 dB as a high quality point and the lowest quality video was coded at 28 dB. The tests were evaluated with 16 participants with eight left-eye dominant and eight right-eye dominant. The equal number of left and right eye dominant participants is a coincidence and was not by design. The dominant eye test was conducted using the commonly used hole-in-the-card test. The data collected included handedness and eyedness.

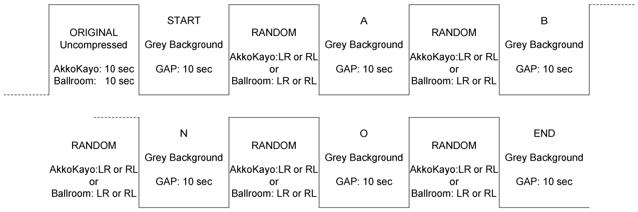

Figure 3. Timing of subjective 3D Image Quality of each random constructed video set.

5. RESULTS AND DISCUSSION

The quality of the 3D video experienced primarily depends on the coding artifacts present in the individual views and the type of 3D display. The influence of the different types of artifacts present in the individual views is not well understood. The quality of a single 2D view alone is not an indication of the 3D quality. Developing objective quality metrics for 3D quality is thus very difficult and subjective evaluation is the primary means of evaluating 3D video quality.

5.1 3D Video Quality and Eye Dominance

While it has been known that human have a preference of one eye over the other, the significance of this preference is not well understood. Humans are mostly right handed (90%) and about 70% are right eyed, 20% left eyed, and 10% exhibit no eye preference [12]. The larger number (50%) of left-eye dominant participants in the 3D evaluation can perhaps be explained by the fact that all the participants are from the college of engineering. A recent study suggested that the eye dominance just indicates individual sighting preferences and has no function in binocular vision [13]. A more recent study, however, found that eye dominance improves the performance of visual search tasks by perhaps aiding visual perception in binocular vision [14]. Our results also suggest a role for eye dominance in binocular vision.

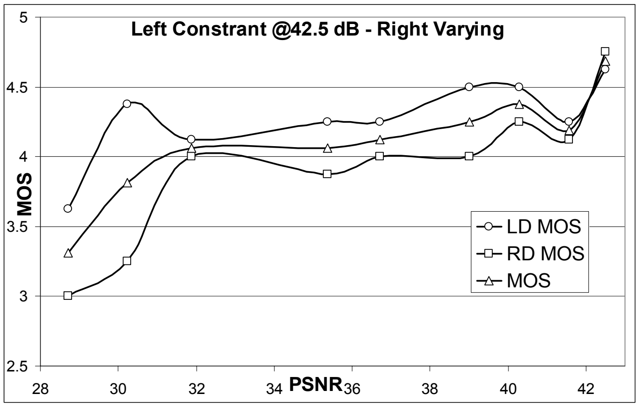

Figure 4. Mean opinion scores for asymmetric view coding with left eye view at a higher quality

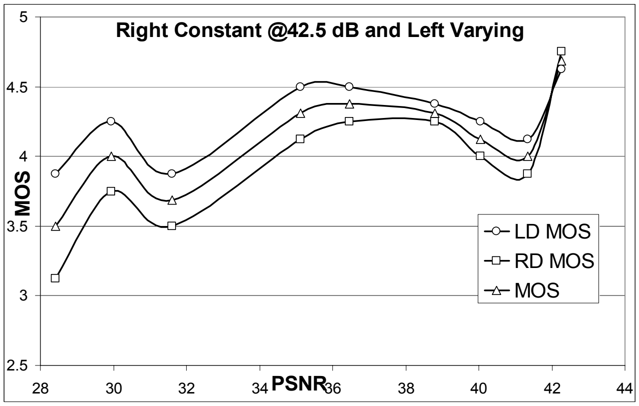

Figure 5. Mean opinion scores for asymmetric view coding with right eye view at a higher quality

Mean opinion scores were computed for the test sequences based on subjective evaluations. Figures 4 and 5 show the mean opinion scores (MOS) for the Akko and Kayo sequence with right eye view kept constant at 42.5 dB and the left eye view coded at lower qualities. A second set of sequences were also evaluated with left eye view encoded at 42.5 dB and right eye view quality varied from 42.5 dB to 28 dB. The figures show the MOS for all the users, the right-eye dominant users, and left-eye dominant users. The figures show that eye dominance does impact 3D perception. Right eye dominant users seem to be more sensitive to the asymmetric video quality. As the quality of the right (left) view increases, the difference between the left-eye and right-eye dominant users decreases.

The MOS is about one point higher for left-eye dominant users when one the views is encoded at a lower quality. The increased sensitivity of right-eye dominant users puts constraints on the lower bound of view quality in asymmetric view coding. Further study is necessary to understand why the right eye dominant users might be more sensitive to asymmetric video coding. The role of eye dominance has significant implications on the asymmetric view encoding of stereo views. The stereo views have to be encoded at a sufficiently high quality so that the right-eye dominant population does not experience poor 3D quality.

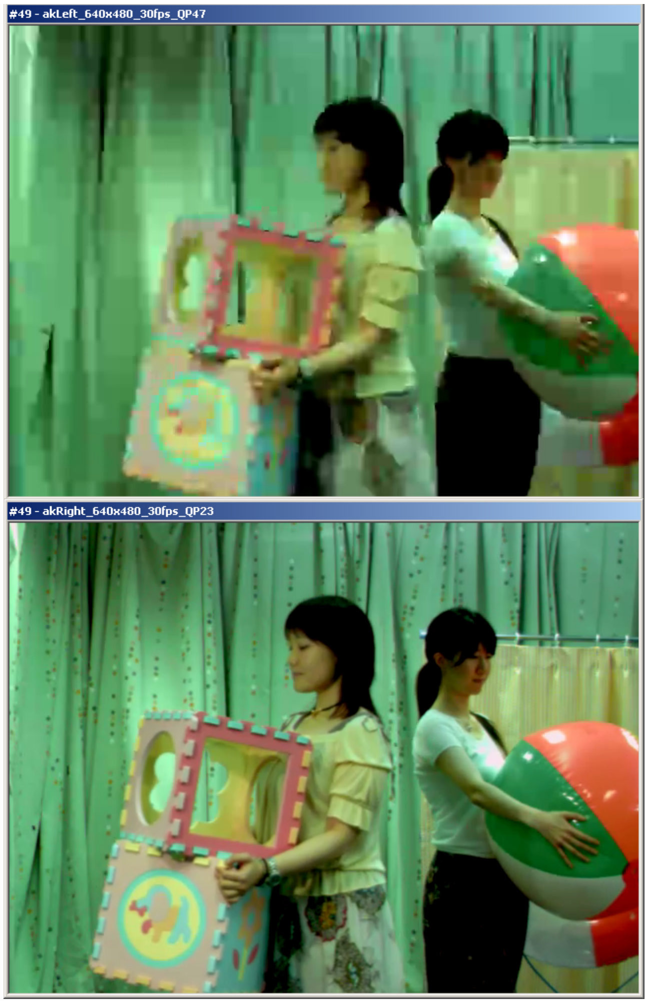
#49 - akLeft_640x480_30fps_QP47
#49 - akRight_640x480_30fps_QP23

Figure 6. Snapshots of the left view coded at a low quality (above) and right view coded at a high quality (below)

3D compression with H.264 view coding performs very well under asymmetric view coding. Figure 6 shows the quality of the left and right eye views that resulted in a MOS close to 4. The binocular mixture in the human visual system suppresses this poor quality and gives the users a reasonably good 3D experience. The low quality left eye view in this case was encoded at a very low quality and is completely unacceptable by itself. As shown in the figure, the low quality left-eye view lost significant picture details due to quantization. The pattern on the background is lost and the facial features are completely blurred. However, when combined with a high quality right eye view, the 3D/depth perception is well preserved. The resulting 3D view has blocking artifacts on the background but contains all the background and foreground details that are lost in the left-eye view.

Binocular vision is not the only source of depth perception. The monocular views contain depth cues which are combined with the disparity information to give the depth perception. The asymmetric view coding principle can be further exploited by coding the low quality view such that the visual cues that contribute to depth perception are coded with a higher quality compared with the regions without any depth cues. Similarly, flat regions in a picture (regions without depth) can be compressed more than the regions with objects present. The presence of an edge is one simple metric that can be used to drive such adaptive compression in asymmetric view coding. The blocks with edges can be coded with higher quality compared to the edge-free blocks in the picture. The impact of these adaptive coding techniques on the eye dominance also needs to be studied.

6. CONCLUSION

This paper presents a 3D video system with asymmetrical view coding. The characteristics of the human visual system are exploited to encode stereo views with asymmetric quality without affecting the quality of the 3D experience. Architecture of a 3D/multiview video player was presented. The player is based on the DirectShow and AviSynth frameworks and renders 3D video on autostereoscopic displays. The paper reports the results of experiments designed to understand the bounds of asymmetric view coding using H.264 video compression and autostereoscopic displays. Experiments were conducted to evaluate the 3D video quality with one eye view kept constant at a high quality and the other eye view encoded with decreasing quality. The results of these asymmetric view coding experiments suggest the influence of eye-dominance on the perceived video quality. The role of eye dominance will have significant implications on the asymmetric view encoding and as a result on the coding efficiency of 3D video.

7. ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research (ONR) under the Center for Coastline Security Technology grant N00014-05-C-0031.

8. REFERENCES

[1] B.L. Tseng and D. Anastassiou, “Multi-Viewpoint Video Coding with MPEG-2 Compatibility,” *IEEE Trans. Circuits and Systems for Video Tech.*, Vol. 6, No. 4, Aug. 1996, pp. 414-419.

[2] A. Puri, R. V. Kollarits and B. G. Haskell, “Stereoscopic Video Compression Using Temporal Scalability,” *Proc. SPIE Visual Communications and Image Processing'95*, Taiwan, May 1995.

[3] Lew B. Stelmach, W. James Tam, “Stereoscopic image coding: Effect of disparate image-quality in left- and right-eye views”, *Signal Processing: Image Communication*, Vol. 14, pp.111-117, 1998.

[4] N.A. Dodgson, “Autostereoscopic 3D Displays,” *Computer*, Volume 38, Issue 8, Aug. 2005 pp. 31 – 36.

[5] L.M.J. Meesters, W.A.Jsselsteijn, and P.J.H. Seuntiens, "A Survey for Perceptual Evaluations and Requirements of Three-Dimensional TV," *IEEE Trans. Circuits Syst.Video Technol.*, Vol. 14, No. 3, Publisher, Location, pp. 381-391, March 2004.

[6] H. Kalva and B. Furht, "Hypercube based inter-view prediction for multi-view video coding," *Proceedings of the 2nd Workshop on Immersive Communication and Broadcast Systems (ICOB)*, October 2005.

[7] O. Schreer, P. Kauff, and T. Sikora, edts., "3D Video Communications" *Wiley* 2005.

[8] Daniel V. Meegan, Lew B. Stelmach, and W. James Tam, "Unequal Weighting of Monocular Inputs in Binocular Combination: Implications for the Compression of Stereoscopic Imagery", *Journal of Experimental Psychology: Applied*, Vol. 7(2) 143-153, Jun 2001.

[9] C. Fehn, "A 3D-TV approach using depth-image-based rendering (DIBR)", *Proc. of VIIP 03*, Benalmadena, Spain, Sept. 2003.

[10] AviSynyth, http://www.avisynth.org

[11] ISO/IEC JTC1/SC29/WG11, "Survey of Algorithms used for Multi-view Video Coding (MVC)," *MPEG Document MPEG2005/N6909*, January 2005.

[12] D.C. Bourassa, I.C. McManus, and M.P. Bryden, "Handedness and eye-dominance: A meta-analysis of their relationship," *Laterality*, Vol 1, No. 1, 1996, pp. 5–34.

[13] A.P. Mapp, H. Ono, and R. Barbeito, "What does the dominant eye dominate? A brief and somewhat contentious review," *Perception & Psychophysics*, Vol. 65, No. 2, 2003, pp. 310– 317.

[14] E. Shneor, and S. Hochstein, "Eye dominance effects in feature search," *Journal of Vision*, 5(8), 699a, http://journalofvision.org/5/8/699/, doi:10.1167/5.8.699.