Multi-view Video Compression for 3D Displays

Matthias Zwicker
University of California
San Diego

Sehoon Yea
Mitsubishi Electric
Research Laboratories

Anthony Vetro
Mitsubishi Electric
Research Laboratories

Clifton Forlines
Mitsubishi Electric
Research Laboratories

Wojciech Matusik
Adobe Systems, Inc.

Hanspeter Pfister
Harvard University

Abstract—**Multi-view 3D displays are preferable to other stereoscopic display technologies because they provide autostereoscopic viewing from any viewpoint without special glasses. However, they require a large number of pixels to achieve high image quality. Therefore, data compression is a major issue for this approach. In this paper, we describe a framework for efficient compression of multi-view video streams for multi-view 3D displays. We enhance conventional multi-view compression pipelines with an additional pre-filtering step that bandlimits the multi-view signal to the display bandwidth. We show that this pre-filtering step leads to increased image quality compared to state-of-the-art multi-view coding at equal bitrate. We present results of an extensive user study that corroborate the benefits of our approach. Our work suggests that any multi-view compression scheme will benefit from our pre-filtering technique.**

I. Introduction

Multi-view 3D displays offer viewing of high-resolution stereoscopic images from arbitrary positions without glasses. These displays consist of view-dependent pixels that reveal a different color to the observer based on the viewing angle. View-dependent pixels can be implemented using conventional high-resolution displays and parallax-barriers, lenticular sheets, or holographic screens. Today, commercial availability ranges from multi-view desktop monitors [1] to large-scale displays based on multi-projector systems [2], [3].

Multi-view 3D displays feature several advantages over competing autostereoscopic display technologies, such as stereoprojection systems using shuttered or polarized glasses. Most importantly, automultiscopic displays do not require users to wear any special glasses, which leads to a more natural and unrestricted viewing experience. They also do not require head tracking to provide motion parallax; instead, they provide accurate perspective views from arbitrary points inside a viewing frustum *simultaneously*. They are truly multi-user capable, since none of the display parameters needs to be adjusted to an individual user. For these reasons, we believe that multi-view 3D displays will become the device of choice for a large number of applications such as scientific visualization or remote collaboration. They have the potential to replace conventional 2D displays in the mass markets of digital entertainment [4].

However, the amount of data that needs to be processed, rendered, and transmitted to such displays is an order of magnitude larger than for systems based on stereo-image pairs. Therefore, data compression is of paramount importance for such systems. We describe a framework for efficient compression of multi-view video streams that complements current techniques. Our approach reduces the required data rate to a minimum by taking into account the multi-dimensional display bandwidth. A more detailed description of this technique has been presented earlier [5].

The limited bandwidth of multi-view 3D displays corresponds to a *shallow depth of field.* This means that only those scene elements that are within a certain distance from the display plane can be shown sharply. Scene elements that appear at larger distances become increasingly blurry. We improve standard multi-view compression by adding a pre-filtering step that bandlimits the input signal to the display bandwidth. Pre-filtering has two desirable effects: First, it removes high frequencies that would appear as aliasing, and second, it reduces the signal bandwidth.

We evaluate our approach using an extensive user study that corroborates the benefits of the pre-filtering step. We show that, at equal signal bitrate, our approach leads to higher perceived image quality compared to state-of-the-art multi-view coding without pre-filtering. Our work suggests that any compression scheme for multi-view 3D displays will benefit from our pre-filtering technique.

II. Previous Work

We distinguish three approaches to characterize display bandwidth of multi-view 3D displays. The first one, proposed by St. Hilaire [6], builds on wave optics. A second approach, as described by Halle [7], is based on simple geometric considerations. The third approach [8] is based on a ray space representation of multiview 3D displays. It casts the analysis of display bandwidth as a multidimensional sampling problem in three- or four-dimensional ray space. This approach is related to the concept of light fields [9], which has been studied extensively in the computer graphics community. The frequency analysis of light fields, also known as *plenoptic sampling theory*, has been studied by Chai et al. [10] and Isaksen et al. [11]. An analysis of the display bandwidth using plenoptic sampling theory [10] reveals important properties, such as the shallow depth of field of practical displays.

978-1-4244-2110-7/08/$25.00 ©2007 IEEE

Moller et al. [12] describe a method to prevent interperspective aliasing that is based on St Hilaire's [6] display bandwidth analysis. Unfortunately, this approach requires the knowledge of per pixel scene depth. In addition, it leads to a spatially varying 2D filter. Zwicker et al. [8] derive a low-pass filter directly from the ray-space sampling grid of the multiview 3D display. This approach prevents aliasing within each view as well as inter-perspective aliasing. It does not require the knowledge of scene depth and it is implemented as a linear convolution rather than relying on spatially varying filtering. Therefore, we base our pre-filtering technique on this approach.

Multiview 3D displays require, at least, an order of magnitude more samples than conventional 2D displays to achieve comparable image quality because of the higher dimensionality of the input signals. Therefore, data compression plays a crucial role in making these displays practical. Compression of multi-view video data is a highly active area of research, and standardization efforts for multi-view video compression are well under way in the MPEG-4 community. Various extensions of the H.264/MPEG-4 AVC video compression standard to the multi-view setting have been proposed recently [13].

However, none of the previous multi-view video or light field compression techniques take the three- or four-dimensional bandwidth of multi-view 3D displays into account. This means that parts of the frequency content of the encoded signal will appear as interperspective aliasing when rendered on a 3D display. This can reduce image quality and lead to inefficient compression. Our multi-view compression scheme includes a low-pass filtering stage to ensure that the encoded signal does not exceed the bandwidth of a target 3D display. This approach has two advantages over previous techniques. First, it avoids interperspective aliasing artifacts, and second, our approach increases compression efficiency.

III. Compression Pipeline

Our compression pipeline for multi-view 3D displays consists of two main steps. In the first step, described in Section III-A, we perform a display pre-filtering operation. This step removes frequency content from the input signal that is beyond the Nyquist limit of the display. Because these frequencies would appear as aliasing on the multi-view display, the pre-filtering step does not reduce image quality. However, it increases the compression efficiency by zeroing out parts of the spectrum of the input signal. In the second step of our pipeline, we run the pre-filtered signal through a state-of-the-art multi-view compression algorithm, which we summarize in Section III-B.

A. Display Pre-filtering

Multi-view 3D displays seek to reproduce the full *light field* [9], [14] of an input scene. The underlying idea of the display pre-filtering step is to parameterize the light rays emitted by the display by their intersection with two parallel planes. The intersection coordinates of each ray correspond to a point in *ray space*, and the set of all rays forms a higher-dimensional, quadrilateral sampling grid in ray space. This sampling grid determines the Nyquist limit of the display, or the display bandwidth. The ideal display pre-filter can now be characterized as a rectangular box in the frequency domain of ray space [8].

Chai et al. [10] introduced a frequency analysis of light fields using the two plane parameterization of ray space. They showed that the spectrum of light fields has a typical bow tie shape as depicted in Figure 1. The horizontal and the vertical axis represent spatial (θ) and angular frequencies (ψ). The minimum and maximum slopes here correspond to the minimum and maximum depth in the scene captured by the light field. Chai et al. also showed that light fields acquired by camera arrays are often undersampled in the angular domain and suffer from aliasing. This is illustrated in Figure 1a, where spectral replicas of the sampled light field overlap. If we were to apply the display pre-filter directly to this data as shown in 1b, the band-limited signal would still suffer from aliasing artifacts, which were already present in the *input* signal.

To avoid this situation, we first oversample the signal in the angular direction such that it is *free of aliasing within the display bandwidth.* This means that we interpolate more views at a smaller spacing in the angular domain than the display actually provides. We effectively prevent aliasing if none of the bow tie spectra except the central one overlaps with the display prefilter. . This is shown in Figure 1c. We then band-limit the oversampled signal by convolving it with the display pre-filter as illustrated in Figure 1d. We implement this step as a convolution with a Gaussian filter in the spatial domain [8]. Of course, other filter kernels could be used alternatively. After pre-filtering we subsample at the original display resolution.

B. Multi-view Compression

One solution for compressing multiview videos is to encode each view independently using a state-of-the-art video codec such as H.264/AVC [15]. The main advantage of this approach is that current standards and existing hardware could be used. To achieve further gains in coding efficiency, extensions to the H.264/AVC standard are now being developed to exploit not only the redundancy in pictures over time, but also the redundancy between pictures in different camera views.

It has been shown that coding multiview video with inter-view prediction does give significantly better results compared to independent coding of each view [13]. Improvements of more than 2 dB have been reported for the same bit-rate, and subjective testing has indicated that the same quality could be achieved with approximately half the bit-rate for a number of test sequences. A more comprehensive review of recent developments in multi-view coding can be found in [16]. All compression experiments that follow utilize inter-view prediction using an algorithm based on Merkle et al.'s approach [13].

Since our pre-filtering approach suppresses high frequency of the input signal to avoid anti-aliasing, the multiview signal

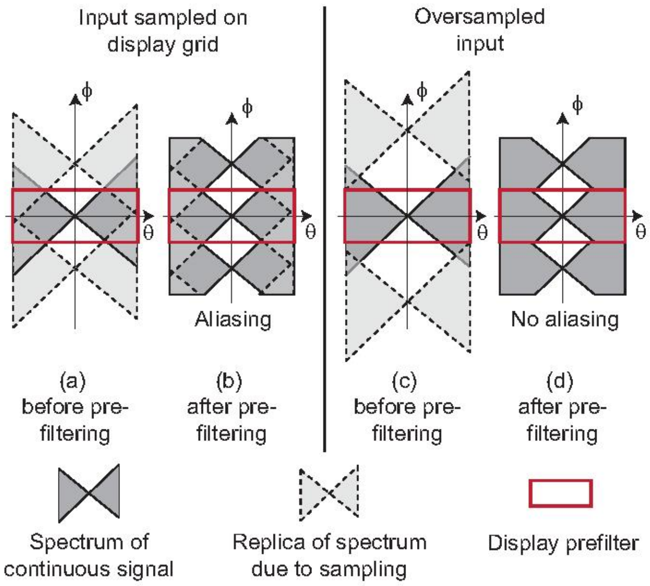

Fig. 1. Display pre-filtering without oversampling leads to aliasing artifacts, as shown on the left. Oversampling the input avoids these problems, as shown on the right. Note that the display prefilter is a unit square. The visualization is stretched horizontally to emphasize the difference in resolution between spatial sampling (the resolution of the multiview display) and angular sampling (the number of views).

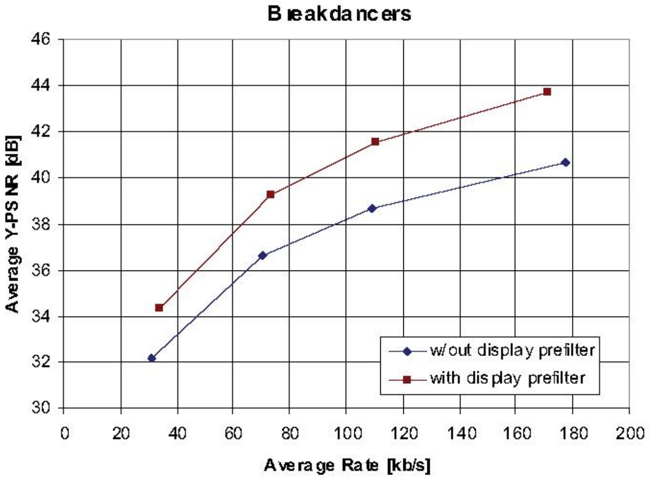

Fig. 2. Comparison of RD curves for breakdancer sequence with and without pre-filtering.

becomes even easier to compress. To demonstrate the reduction in data rate that is possible, we plot the rate-distortion curves comparing the quality of the compression of multiview videos with and without pre-filtering at different bitrates in Figure 2. We performed the measurements using the *breakdancers* data set. These plots show that the rate could be reduced by approximately half in the medium to higher rate ranges. It is important to note that this should not be viewed as a gain in coding efficiency since the references used for each curve are indeed different. The purpose of these plots are just to demonstrate the degree of rate savings that are achieved when the multiview signal has been pre-filtered with the primary purpose of removing anti-aliasing artifacts.

We compare the result of compression of pre-filtered views and original views in Figure 3. The images are from the *Waterfall* test sequence, which was also included in our user study (Section IV). We show results of compression without pre-filtering at the top, and with pre-filtering at the bottom. We reduced the bitrate of both sequences to 110 kbps per second. The images in Figure 3 are simulated display views. The foreground character shows stronger blocking artifacts in the version without pre-filtering, at the top, than with pre-filtering, at the bottom. In addition, pre-filtering removes ghosting artifacts, which appear without pre-filtering in the background in the top image.

IV. User Study

We conducted a preferential study designed to shed light on the effect of our pre-filtering approach on user preference when viewing compressed 3D videos. Twelve subjects participated in our study, six males and six females between the ages of 23 and 45 years old. These individuals were recruited from an on-line community bulletin board and from the administrative and speech departments at our organization. Participants from outside our organization were paid $10 compensation for their time.

A. Method and Procedure

Subjects were first shown an example video that shipped with our display in order to demonstrate the capabilities of the device. We used a 23” display by Newsight [1] that provides eight views with a resolution of 640×384 pixels each. For most of our participants, this was their first experience viewing a 3D display. All of our participants were able to perceive depth in the image, and all of them had normal or corrected normal vision.

Participants were shown a series of video pairs with a short segment of blank grey video inserted between them. Each video in the pair contained the same content, compressed with and without pre-filtering as described in Section III. Participants were allowed to view the pair of videos as many times as they wanted to in order to answer the question, “Which video do you prefer overall?” Five different video clips were used, ranging in length between six and ten seconds. These included a variety of different content - a video of a ballerina, a video of several break-dancers, a synthetic scene of a model dragon, a man standing in front of a waterfall, and a man standing in front of a pedestrian walkway. All video sequences have eight views with a resolution of 640×384 pixels. Each video pair was compressed at three different bitrates for a total of 15 pairs. We manually adjusted the quality parameter of the compression algorithm to achieve similar bitrates with and without pre-filtering. We report the bitrates for the test sequences in Table I. Although they do not match exactly, we verified empirically that the remaining differences are too small to influence the perceptual study. In

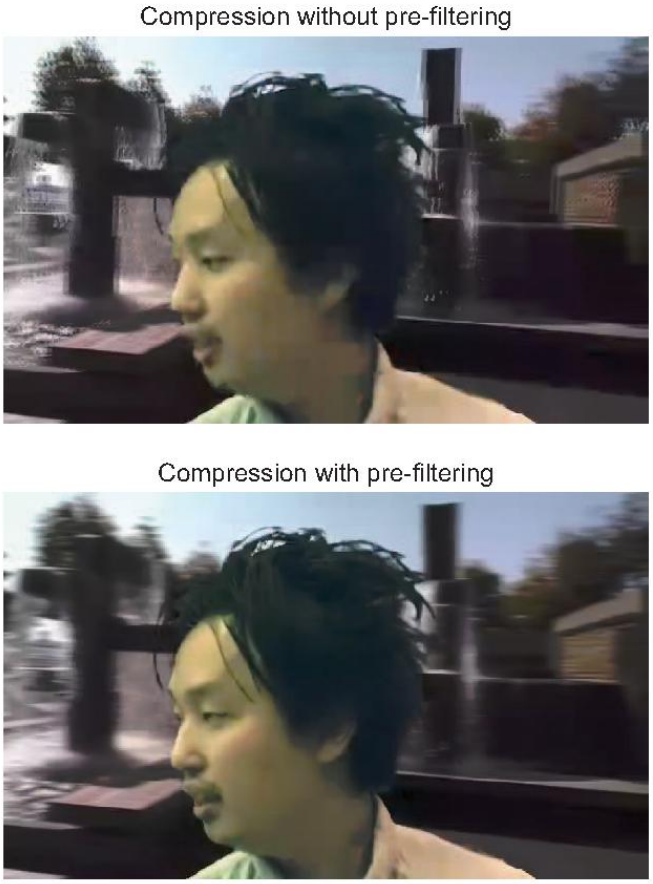
Compression without pre-filtering
Compression with pre-filtering

Fig. 3. Comparison of compressed frames of a video sequence with and without pre-filtering. The images show simulated views of a multi-view 3D display. The version without pre-filtering at the top shows stronger blocking artifacts than the version with pre-filtering at the bottom. In addition, pre-filtering avoids ghosting artifacts, which appear without pre-filtering in the background in the top image.

Scene	Low	Medium	High
Walkway	52.8/51.2	82.6/83.2	138.4/136.4
Breakdancers	37.8/40.6	46.3/48.9	70.2/73.0
Waterfall	58.0/61.2	91.3/95.1	132.1/128.7
Dragon	59.4/54.4	121.9/124.9	179.4/181.8
Ballet	31.9/32.4	60.3/59.4	122.7/115.7

TABLE I
LOW, MEDIUM, AND HIGH BITRATES IN KBPS FOR THE FIVE TEST SEQUENCES WITH/WITHOUT PRE-FILTERING. WE MANUALLY ADJUSTED THE PARAMETERS OF THE COMPRESSION ALGORITHM TO OBTAIN SIMILAR BITRATES WITH AND WITHOUT PRE-FILTERING.

summary, our design was:

- 2 compression techniques (with/without pre-filtering),
- 5 video clips (ballerina, break-dancers, dragon, waterfall, and walkway),
- 3 bitrates (low, medium, and high),
- 12 participants,
- resulting in 180 trials in total.

We had two experimental hypotheses:

- As a group, participants would prefer video clips compressed using the pre-filtering technique over clips compressed without pre-filtering.
- The preference for videos rendered using the pre-filtering technique over those without pre-filtering would be inversely correlated with the bitrate of the encoded video.

B. Results and Discussion

As predicted by hypothesis one, our participants preferred the pre-filtering technique in a majority of the experimental trials (60.7% vs. 39.3% of trials for anti-aliased and bilinear respectively), with nine of our twelve participants preferring the pre-filtered technique overall. While this difference is not statistically significant, the lack of significance is likely due to the strong preference for one technique or the other on the part of three of our participants and the resulting large standard deviations for the mean preference scores. Figure 4 shows the mean preference scores for both compression techniques.

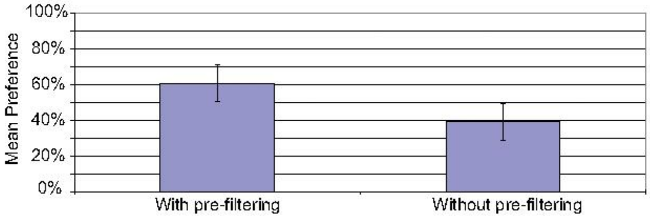

Fig. 4. The percent of trials in which participants preferred each of the rendering techniques. Bars indicate standard error.

In accord with hypothesis two, there appears to be an interaction between compression technique and bitrate, as shown in Figure 5. As seen in the figure, the lower bitrates resulted in a higher preference for the pre-filtered technique, while the highest bitrate was the only bitrate in which our participants preferred the technique without pre-filtering overall.

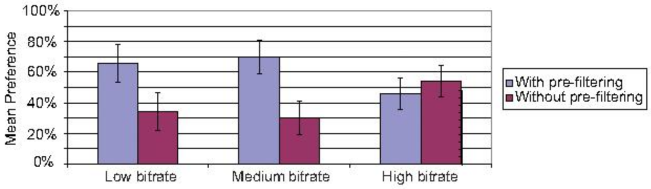

Fig. 5. The percent of trials in which participants preferred each rendering technique for each of the three bitrates.

This interaction may be due to the participants' multiple viewing of the videos during the experiment. Each scene used in our study had an object or person that was the focus of the scene. During the first viewing of the scene, our participants tended to focus on this main object; however, after viewing the videos several times, they began to inspect the background, the foreground, or other secondary objects in the scene. When the bitrate was high, and compression artifacts were few, the compression without pre-filtering produces clearer images in these regions of the scene farthest from the focal point. While this clarity results in object ghosting, which is a quality that the majority of our participants identified as distracting, participants were able to make out more detail in their subsequent viewings of the video. Therefore, we hypothesize that the preference for pre-filtered rendering would grow for 3D video viewed only a single time.

There also appears to be an interaction between the video clip shown and the rendering technique preferred by our participants, indicating that content is an important consideration when choosing compression techniques. Figure 6 shows the mean preference for both compression techniques for each of the five video clips used in the study. The scene with the ballerina is the only scene for which our participants preferred the compression without pre-filtering. This scene included not only a dancing ballerina, which appears in focus on the display, but also a dance partner that is closer to the viewer and slightly out of focus. Several participants mentioned that they could not see as much detail of the partner when viewing the pre-filtered version of the video.

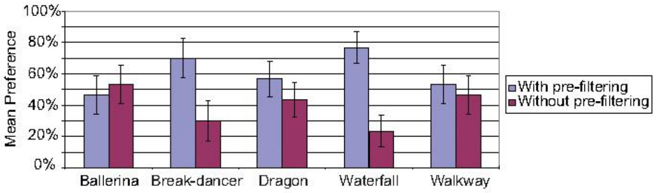

Fig. 6. The percent of trials in which participants preferred each rendering technique for each of the five different video clips.

As shown in Figure 6, the waterfall clip resulted in the largest preference difference between rendering techniques. This clip contains a man's face in the foreground, and a complex moving background. Compression without pre-filtering wastes many bits on the complex background, such that the face in the foreground exhibits significantly more compression artifacts compared to the pre-filtering version. Because humans are very sensitive to the qualities of faces, these artifacts may have driven up our participants' preference for the pre-filtered version of this video. Without pre-filtering, the motion of the waterfall in the background also interacts with ghosting artifacts to produce visual noise.

V. Conclusions

We described a framework for multi-view video compression for 3D displays that takes into account the multi-dimensional display bandwidth. We apply a pre-filtering step that band-limits the input to the display bandwidth before compression. Since the display bandwidth imposes a shallow depth of field, this removes high frequencies from the input signal, making it easier to compress. In addition, if pre-filtering is omitted, high frequencies that appear out of focus on the display lead to ghosting artifacts. The pre-filtering step also avoids these *inter-perspective aliasing* artifacts. Therefore, pre-filtering is beneficial for compression for two reasons: it reduces compression artifacts, and it avoids aliasing.

We evaluated our technique with a preferential user study. We prepared pairs of multi-view video sequences of the same scene, compressed at the same bitrate, with and without our pre-filtering approach. We asked our subjects to indicate their preference for each pair of sequences. We found that pre-filtering is an important parameter to optimize the visual quality of compressed 3D videos.

The perceptual evaluation of compression techniques will play an important role in making multi-view 3D displays practical. However, much more work needs to be done in this area. Testing should be performed on sequences much longer than ours to take into account eye strain. The amount of pre-filtering could be adjusted to better explore the trade-off between ghosting and blurriness. In terms of compression, scalable techniques need to be developed that can produce the right amount of depth of field for different displays with different numbers of views.

Acknowledgments

We are grateful to Microsoft Research for the *breakdancers* data set and to Aseem Agarwala for the waterfall scene.

References

[1] Newsight, http://www.newsight.com/, 2006, cited January, 2006.

[2] T. Agocs, T. Balogh, T. Forgcs, F. Bettio, E. Gobbetti, and G. Zanetti, "A large scale interactive holographic display," in *Proc. IEEE VR 2006 Workshop on Emerging Display Technologies*, 2006.

[3] Holografika, "Holovizio displays," http://www.holografika.com/, 2006, cited September, 2006.

[4] L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, no. 3, pp. 381–391, 2004.

[5] M. Zwicker, S. Yea, A. Vetro, C. Forlines, W. Matusik, and H. Pfister, "Display pre-filtering for multi-view video compression," in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, 2007, pp. 1046–1053.

[6] P. S. Hilaire, "Modulation transfer function and optimum sampling of holographics stereograms," *Applied Optics*, vol. 33, no. 5, February 1994.

[7] M. Halle, "Holographic stereograms as discrete imaging systems," in *Practical Holography VIII*, ser. SPIE Proceedings, vol. 2176, 1994, pp. 73–84.

[8] M. Zwicker, W. Matusik, F. Durand, and H. Pfister, "Antialiasing for automultiscopic 3d displays," in *Eurographics Symposium on Rendering*, 2006.

[9] M. Levoy and P. Hanrahan, "Light field rendering," in *Computer Graphics*, ser. SIGGRAPH 96 Proceedings, New Orleans, LS, Aug. 1996, pp. 31–42.

[10] J. X. Chai, S. C. Chan, H. Y. Shum, and X. Tong, "Plenoptic sampling," in *Computer Graphics*, ser. SIGGRAPH 2000 Proceedings, Los Angeles, CA, July 2000, pp. 307–318.

[11] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Computer Graphics*, ser. SIGGRAPH 2000 Proc., Jul. 2000.

[12] C. N. Moller and A. Travis, "Correcting interperspective aliasing in autostereoscopic displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 2, pp. 228–236, March/April 2005.

[13] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient compression of multiview video exploiting inter-view dependencies based on H.264/AVC," in *Proc. IEEE Int'l Conf. Multimedia & Expo*, 2006.

[14] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen, "The lumigraph," in *Computer Graphics*.

[15] ITU-T Rec. & ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services," 2005.

[16] ISO/IEC JTC1/SC29/WG11, "Survey of algorithms used for multi-view video coding (MVC), MPEG document MPEG2005/N6909," January 2005.