DELIVERING ENHANCED 3D VIDEO

Contributors

Yury Gitman
Lomonosov Moscow State University

Can Bal
University of California, San Diego

Mikhail Erofeev
Lomonosov Moscow State University

Ankit Jain
University of California, San Diego

Sergey Matyunin
Lomonosov Moscow State University

Kyoung-Rok Lee
University of California, San Diego

Alexander Voronov
Lomonosov Moscow State University

Jason Juang
University of California, San Diego

Dmitriy Vatolin
Lomonosov Moscow State University

Truong Nguyen
University of California, San Diego

Autostereoscopic displays are expected to gain higher popularity in comparison with devices that require a viewer to wear special glasses to see 3D. Nonetheless, the industry might not be ready to deliver high quality 3D to viewer. In this article we examine each stage of the 3D content lifecycle from its creation to display in the user's home. We present various algorithms for solving existing problems in each of these steps.

To avoid the creation of low quality 3D, we propose a set of methods for automatic quality assessment. To enable easy multiview content creation, a disparity estimation method is proposed. We discuss two methods of efficient 3D compression to save bandwidth in wireless channels. We propose a system for 3D display quality assessment to ensure that 3D video quality is not affected by the display device. Finally, we describe a system for carrying out subjective comparisons that will assist in further improvement of the above-mentioned methods.

Introduction

The challenge of video transport in future wireless networks includes the increasing prevalence of new data types like HD and 3D. While HD (high-definition) video increases data size by expanding resolution, 3D (stereoscopic) video increases it through the use of dual video streams, one intended for each eye of the viewer, or even more than ten video streams for multiview autostereoscopic displays, enabling the watching of 3D from several points of view without special glasses. Raw 3D video data is thus at least twice the size of raw 2D video, although much work has been done to develop compression standards (for example, MVC extensions to H.264/AVC) that improve data size in practice.

A viewer's perceptual system relies on several cues to extract information about the relative distance between objects within 3D space. These can be divided into two groups: monocular cues and binocular cues. While 2D video systems use only monocular cues (such as motion parallax, perspective, and occlusions), 3D video systems also utilize binocular cues to give the viewer an even stronger sensation of scene depth. For example, *binocular parallax* refers to the difference between left and right images, something the human visual system automatically translates into perceived depth. *Eye convergence* refers to the manner in which human eyes point inward at one another as an object gets closer to the viewer.

While conventional approaches to 3D video formatting like Side-By-Side and Top-and-Bottom incorporate separate left and right images into a single video frame or temporally interlace left and right images sequentially within the data stream (frame sequential), newer approaches leverage the notion of a *depth map*.

Depth maps provide information on the relative distances of surfaces within a scene from a viewpoint and, when combined with 2D or multiview video data, can be used to reconstruct a second view frame at the decoder using depth-image-based rendering (DIBR) techniques. Depth maps may be used to adjust the disparity between views on different displays and to generate multiple views in autostereoscopic displays. They may also be associated with compression strategies for transmitting 3D video in capacity-constrained wireless networks.

We believe the challenge of transporting 3D video over wireless networks of the future requires an end-to-end approach that considers all phases of the transport pipeline: content creation, delivery, and display. Content creation determines the encoding and formatting of data, matching data to user display requirements and creating opportunities for data compression. Delivery considerations include the role of data compression and robustness against errors that occur due to wireless channel effects. Display considerations focus on device characteristics, as well as complexity and performance of rendering and playback. How to define and measure quality in 3D video playback for a particular device and 3D encoding scheme is a key challenge.

"We believe the challenge of transporting 3D video over wireless networks of the future requires an end-to-end approach that considers all phases of the transport pipeline…"

This article reviews our research contributions to each part of this end-to-end pipeline within the VAWN research program. The next section, "Content Creation: Disparity Estimation and Processing, and Quality Metrics," discusses our work on content creation, including disparity estimation, which is a building block for depth map construction, the handling of common errors in depth map construction, and our development of VQMT3D, a tool for 3D video quality assessment. In the section "Delivery," we review our contributions to 3D data delivery, including multiview video compression based on depth map propagation, multiview video plus depth coding with depth-based prediction mode, and mixed resolution stereoscopic coding. In the section "Display," we discuss our work on 3D displays including autostereoscopic displays, tools for subjective testing, and automatic device testing. In the section "Conclusion," we close with recommendations to the industry community.

Content Creation: Disparity Estimation and Processing, and Quality Metrics

There are two common approaches to creating 3D content (besides rendering): capturing with a stereo camera system or converting from a 2D video. Capturing with a stereo camera rig seems to be the most natural way to create stereo video, but typically, captured video suffers from various mismatches in camera settings and inaccurate spatial alignment. This implies the need for 2D-to-3D conversion instead of capturing for some scenes, even for feature films. Converting from a 2D video to 3D has its own difficulties.

The process of 2D-to-3D conversion first requires the content creator to determine the distance of each pixel from the camera for each one of the frames of the entire 2D video. These distance values associated with each pixel

are then stored as a separate 2D video called the *depth map*. While conversion can potentially allow the production of 3D content without any impairments, the quality of the converted video strongly depends on the quality of the depth maps. The labor-intensive nature of depth map creation typically leads to low-quality conversion. In this section, we will also discuss methods for visual quality estimation for both captured and converted content.

It is also possible to capture 3D content with more than two cameras. Such a multiview video capture process is even more challenging because the complexity of camera alignment increases with the number of views. This is the main reason for limited availability of multiview video content which requires autostereoscopic displays. Thus multiview content creators commonly prefer an intermediate approach; that is, capturing stereo video with subsequent conversion. In the case of stereo-to-multiview conversion, the depth map associated with each of the views can be estimated without manual labor. For this, the pixels of the left-view image are matched with pixels of the right-view image in order to estimate a shift map, referred to as the *disparity map*. Given the disparity map, the depth map can easily be calculated, which enables synthesis of additional views.

In this section, we present methods for 3D content creation, beginning with our disparity estimation algorithm.

Disparity Estimation

Disparity estimation is an integral problem associated with the delivery of 3D content (2D + depth, multiview + depth format), and it plays a direct role in both compression efficiency and the need for wireless network capacity to deliver 3D content to mobile devices. In a two-camera imaging system, disparity is defined as the vector difference between the object point in each image relative to the focal point. It is this disparity that allows for depth estimation of objects in the scene via triangulation of the object point in each image. Figure 1 shows

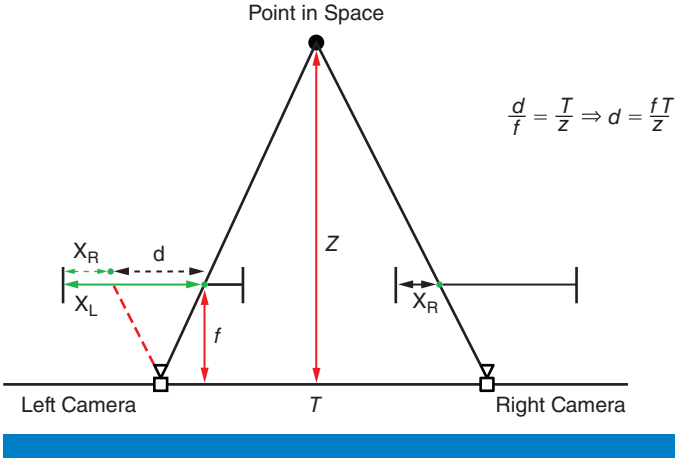


Figure 1: Relating depth to disparity
(Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

the inverse relationship between depth z and disparity d, as identified next to the figure.

The proposed disparity estimation algorithm consists of five main components: similarity measure, support weight, disparity computation, occlusion filling, and *total variation* (TV) refinement. The block diagram of the overall system is shown in Figure 2.

For similarity measure, we propose a three-mode census transform with a noise buffer to be more tolerant of image noise in flat areas and a cross-square census to increase the reliability of census measure. We suggest the effective combination of three cost measures formulated as

$$C_0(q, q_d) = 3 - \exp\left(-\frac{\Delta H_{qqd}}{\gamma_H}\right) - \exp\left(-\frac{\Delta I_{qqd}}{\gamma_I}\right) - \exp\left(-\frac{\Delta G_{qqd}}{\gamma_G}\right) \quad (1)$$

where ΔH_{qqd}, ΔI_{qqd}, and ΔG_{qqd} stand for census, color, and gradient respectively.

For support weight, the adaptive support weight is based on the strength of grouping by similarity and proximity. We suggest the following conditional adaptive support weight:

$$w(c, q) = \begin{cases} g_{\gamma_s}(\Delta S_{cq}) & \text{if } \Delta S_{cq} < E \\ g_{\gamma_s}(\Delta S_{cq})\, g_{\gamma_p}(\Delta p_{cq}) & \text{otherwise} \end{cases} \quad (2)$$

where r_s and r_p are empirical similarity parameters, ΔS_{cq} is the RGB color difference between the center pixel and the neighboring pixel, Δp_{cq} is the spatial distance between pixel c and pixel q, and E is a color difference threshold determining the similar color between two pixels.

For disparity computation, once the support weights are calculated, the aggregated cost is computed by aggregating the raw similarity measures, scaled by the support weights in the window. The aggregated matching cost between pixel c and pixel c_d is given in the weighted mean as

$$A(c, c_d) = \frac{\Sigma q \in W_c, q_d \in W_{cd}\, w(c, q) w(c_d, q_d) c_0(q, q_d)}{\Sigma q \in W_c, q_d \in W_{cd}\, w(c, q) w(c_d, q_d)} \quad (3)$$

where W_c and W_{cd} represent the left and right support windows, respectively, and the function $w(c_d, q_d)$ is the support weight of pixel q_d in the right window. After the aggregated matching costs have been computed within the disparity range, the disparity map is obtained by determining the disparity d_p of each pixel p through the Winner-Takes-All (WTA) algorithm.

For occlusion filling, first a left-right consistency check is performed to detect unreliable pixels. The unreliable pixels are defined as the ones that have nonmatching disparities on the left and right images.

Then, the reliable pixels within a cross-based neighborhood vote for the candidate disparity value at (x, y). The pixels with missing disparity values are then filled with the majority votes of the reliable pixels in the voting region.

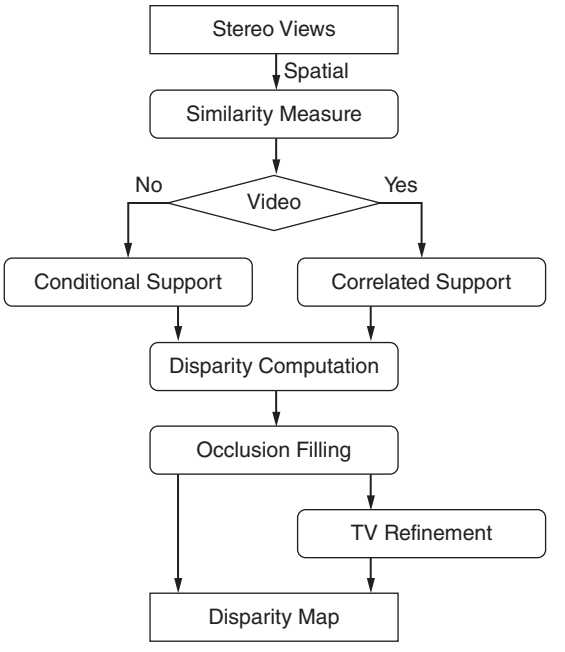


Figure 2: Block diagram of the proposed disparity estimation approach
(Source: Zucheul Lee, PhD thesis: "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

The final step in the algorithm is the TV refinement. The block diagram illustrated in Figure 3 captures how this refinement process works at a high level. TV refinement involves solving the following minimization problem:

$$\underset{\mathbf{f}}{\text{minimum}}\ \mu\|\mathbf{f}-\mathbf{g}\|_1+\|\mathbf{Df}\|_2 \tag{4}$$

where $g = vec(g(x, y, t))$ and $f = vec(f(x, y, t))$ are the initial disparity and the optimization variables, respectively. The operator **D** is the spatiotemporal gradient operator that returns the horizontal, vertical, and temporal forward finite difference of **f**.

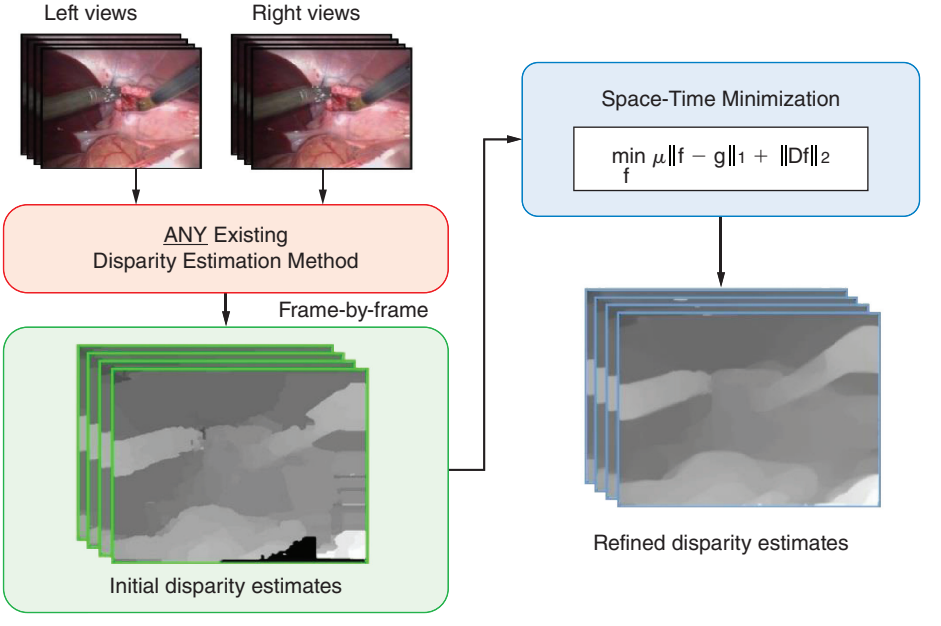


Figure 3: Space time minimization overview
(Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

An augmented Lagrangian method solves the above minimization problem by the following steps (at the *K*th iteration):

$$\mathbf{f}_{k=1}=\ \underset{\mathbf{f}}{\text{argmin}}\ \frac{\rho_o}{2}\ \|\mathbf{r}_k-\mathbf{f}+\mathbf{g}\|^2+\rho_r\|\mathbf{u}_k-\mathbf{Df}\|^2+\mathbf{z}_k^T\mathbf{f}+\mathbf{y}_k^T\mathbf{Df},$$

$$\mathbf{v}_{k+1}=\mathbf{Df}_{k+1}+\frac{1}{\rho_r}\mathbf{y}_k,$$

$$\mathbf{u}_{k+1}=\max\left\{\mathbf{v}_{k+1}-\frac{1}{\rho_r},0\right\}\cdot\frac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|2},$$

$$\mathbf{r}_{k+1}=\max\left\{\left|\mathbf{f}_{k+1}-\mathbf{g}+\frac{1}{\rho_o}\mathbf{z}_k\right|-\frac{\mu}{\rho_o},0\right\}\cdot sign\left(\mathbf{f}_{k+1}-\mathbf{g}+\frac{1}{\rho_o}\mathbf{z}_k\right),$$

$$\mathbf{y}_{k+1}=\mathbf{y}_k-\rho_r(\mathbf{u}_{k+1}-\mathbf{Df}_{k+1})$$

$$\mathbf{z}_{k+1}=\mathbf{z}_k-\rho_o(\mathbf{r}_{k+1}-\mathbf{f}_{k+1}+\mathbf{g})$$

In the iterative method described above, the first problem (known as the f-subproblem) can be solved using Fast-Fourier Transform for fast computation.

The performance evaluation makes us of the Middlebury datasets with ground truth disparity maps provided by the Middlebury online benchmark.[1][2] Table 1 summarizes the quantitative results taken from the Middlebury database methods. Our method achieves excellent results, ranking 13th out of about 130 methods and it is the best performing local method.

Methods	Rank	Avg Err(%)	Tsukuba	Venus	Teddy	Cones
Proposed	13	5.12	2.10	0.12	5.46	2.12
PatchMatch	15	4.59	2.09	0.21	2.99	2.47
CostFilter	20	5.55	1.51	0.20	6.16	2.71
InfoPermeable	21	5.51	1.06	0.32	5.60	2.65
GeoSup	28	5.80	1.45	0.14	6.88	2.94
AdaptDisCalib	37	6.10	1.19	0.23	7.80	3.62
SegmentSupport	53	6.44	1.25	0.25	8.43	3.77
AdaptWeight	67	6.67	1.38	0.71	7.88	3.97

Table 1: Local method performance evaluation on the Middlebury datasets. (Source: Zucheul Lee, PhD thesis, "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

To assess the performance of the proposed method quantitatively on stereo videos, we use five synthetic stereo videos with ground truth disparity from the University of Cambridge.[3] We compare three methods without occlusion filling to compare their performance. The LASW method ranks 67th and the Cost-filter, which is one of the best performing local methods, ranks 20th on the Middlebury benchmark. Table 2 shows the average percentage of bad pixels over all frames and illustrates that the proposed method has the best performance.

Video	# of frames	LASW	CostFilter	Proposed method
Tunnel	99	1.435%	2.157%	0.997%
Book	40	5.933%	4.919%	3.601%
Temple	99	10.15%	10.70%	10.36%
Street	99	9.978%	7.305%	7.246%
Tanks	99	5.714%	4.826%	4.811%

Table 2: Performance comparison of methods on five stereo videos (Source: Zucheul Lee, PhD thesis, "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

"…the depth map is critical for efficiently displaying and transmitting 3D content…"

Depth Upsampling and Processing

As mentioned earlier, the depth map is critical for efficiently displaying and transmitting 3D content, especially for multiview displays where multiple views can be generated at the display using the depth map rather than transmitting each view separately. The more accurate the depth map, the more efficiently the content can be compressed. This reduces the capacity needed when sending over a wireless network and allows a higher quality rendering to be achieved. A depth map is typically estimated using stereo vision systems and the disparity estimation procedure; however recent advances in capture technology also allow capturing the depth maps directly using real-time depth sensors. Regardless of whether the depth maps are estimated or directly captured, depth maps contains errors. These errors can be roughly categorized into two broad categories:

- Errors in transition areas: Inadequate calibration, occlusion areas, or motion artifacts often lead to wrong depth values at object boundaries when aligned with color images.
- Random noise on geometrically flat or smooth surfaces: Properties of the object surface, lighting conditions, or systematic errors may generate noise on the surface.

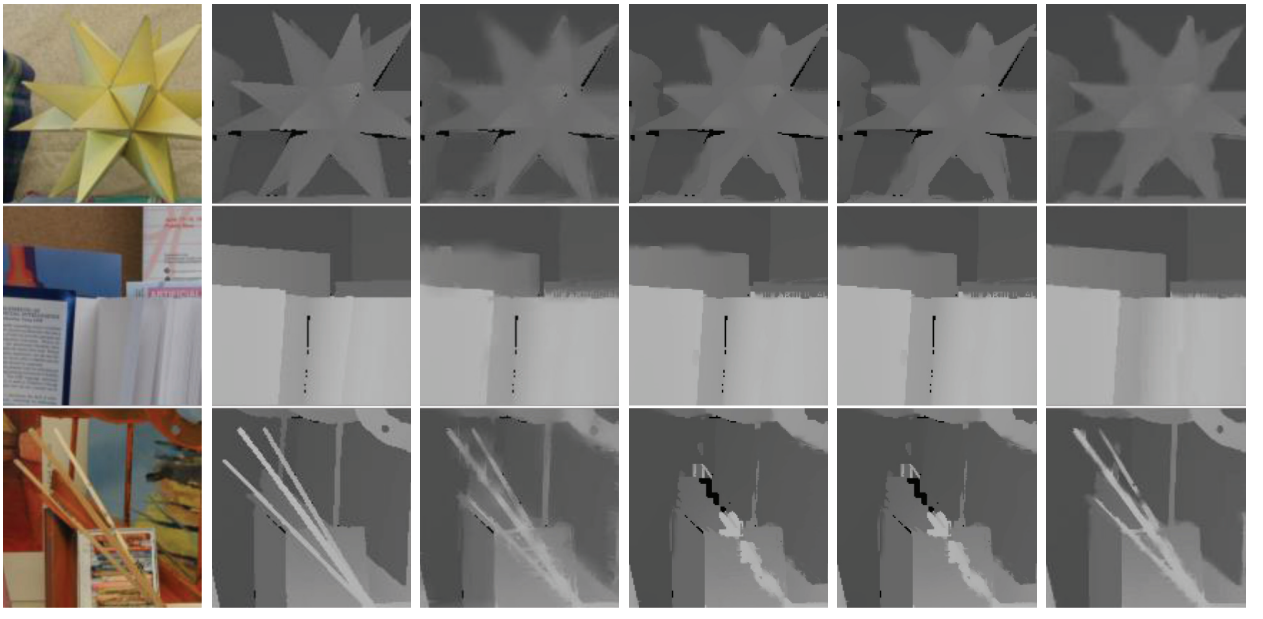
In our work, we investigate a method that can fix both of these errors. Our method takes a color image I and a corresponding lower resolution depth map D as inputs. The process consists of upsampling, sample selection, sample refinement, and robust multilateral filtering of the depth map. Before the refinement step, we begin by measuring the depth reliability and finding the unreliable regions. In this sample selection stage, for every pixel in the unreliable region, we collect depth samples from reliable regions and select the best sample that yields the highest fidelity. Then these samples are refined by sharing their information with their neighbors' selected samples. Finally, a robust multilateral filter is applied to reduce noise in smooth areas, while preserving sharpness along the edges.

The proposed method has been implemented with GPU programming and tested on a computer with an Intel® Core™ i7 2.93-GHz CPU and an NVIDIA GeForce GTX 460* graphics card. Our implementation can produce an output of 26 fps on average for a 640×480 input video.

For a quantitative comparison with the state-of-the-art methods presented by Garcia et al.[4], we utilize the *Moebius*, *Books*, and *Art* scenes from the Middlebury dataset.[5] In this dataset, the disparity maps are of the same resolution as the color images. Hence, we generated the input disparity maps by downsampling the ground truth by a factor of 3x, 5x, and 9x.

Garcia et al.[4] use the structural similarity (SSIM) measure to compare the evaluated methods; however it is not appropriate for evaluation in this context. SSIM cannot yield meaningful results for the regions with unknown depth values and Middlebury's ground truth disparity maps contain such regions. Instead, for a fair comparison, we calculate the average percentage of bad pixels

with an error threshold of 1 for all known regions. In this measure, pixels with a disparity error greater than the threshold are regarded as bad pixels. This is the same scoring scheme employed in the Middlebury evaluation. Figure 4 and Table 3 show that our method performs better than all the competing methods.



(a) Color image (b) Ground truth (c) JBU (d) PWAS (e) AMF (f) Proposed

Figure 4: Visual comparison on the Middlebury dataset. The upsampling methods include: (c) JBU[6], (d) PWAS[7], (e) AMF[4], (f) proposed method
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

Dataset		JBU[6]	PWAS[7]	AMF[4]	Proposed
	3x	7.43	4.68	4.5	**3.62**
Moebius	5x	12.22	7.49	7.37	**4.87**
	9x	21.02	12.86	12.75	**9.02**
	3x	5.4	3.59	3.48	**2.38**
Books	5x	9.11	6.39	6.28	**3.58**
	9x	15.85	12.39	12.24	**7.11**
	3x	15.15	7.05	6.79	**5.07**
Art	5x	23.46	10.35	9.86	**6.91**
	9x	38.41	16.87	16.87	**11.7**

Table 3: Quantitative comparisons (average percentage of bad pixels)
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

In addition, we further evaluate our refinement method by applying the algorithm to all the disparity maps generated by the methods submitted to the Middlebury stereo evaluation. Figure 5 shows the improvement in terms of the percentage of bad pixels. Note that the proposed method improves the

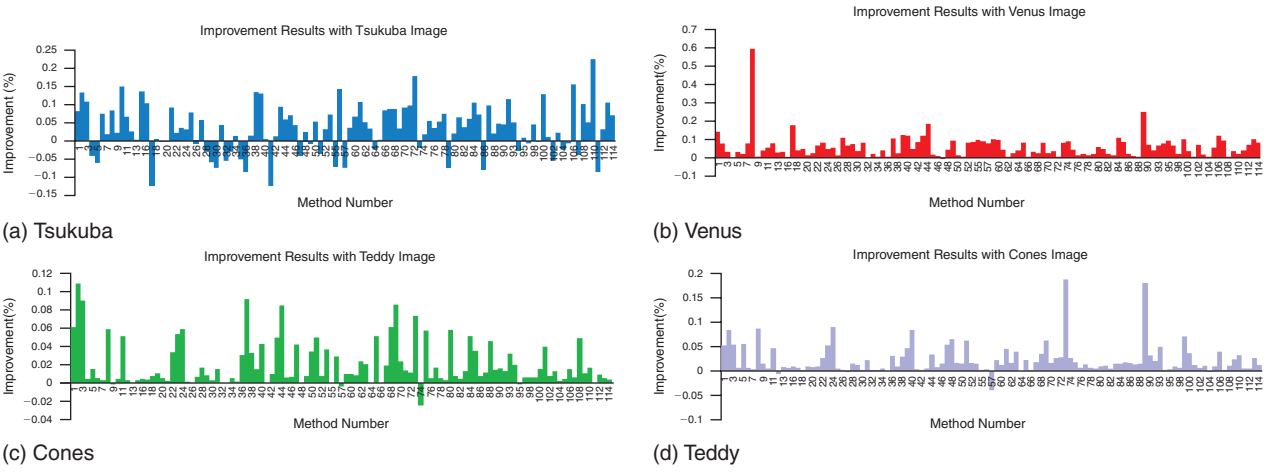


Figure 5: Percentage improvement in terms of number of bad pixels after applying the proposed algorithm to all the 109 methods on the Middlebury stereo evaluation[2]
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

disparity estimates for the majority of the methods. One limitation of the proposed algorithm is that its performance drops when the input images are small and complex, or when the initial disparity estimates contain significant errors.

VQMT3D: Video Quality Measuring Tool for 3D

The quality of experience for any video viewing is important, especially for 3D and multiview content. As such, it's important to understand all the different aspects of 3D that affect the end user quality of experience. As such when it comes to evaluating the rate vs. distortion/quality tradeoff for delivering the content, the communications rate, pre- or postprocessing, and end user quality can all be jointly optimized. Therefore, significant effort in this project went towards understanding 3D video quality impairments and towards creating tools that can help identify these issues. A cinematographer of a 2D film should consider many factors simultaneously, such as scene composition, color camera settings, light, focus position, depth of field, and amount of zoom. A cinematographer of a stereoscopic film additionally pays attention to depth budget distribution and various cameras cross-settings, such as geometry, color, blurriness, and time synchronization.

We have checked the quality of 30 well-known films (besides converted films) and found more than 1000 frames with inter-view mismatches or excessive parallax.[8] In summary, we have concluded that manual quality control leaves a lot of artifacts even in released versions, resulting in less consistent and thus less redundant video-content, which is harder to compress. Therefore process automatization is a necessary and important requirement in stereoscopic video creation.

We started a project to automatically detect all common problems (Video Quality Measurement Tool 3D). Within the project, we developed a distributed system that produces per-frame charts of each metric and automatically extracts problem frames. Finally, the quality-estimating report is generated. We have already published four reports and collected feedback. More than 20 stereographers provided valuable feedback on found artifacts, and their comments were included in our reports.

Some films are captured in 2D and then converted to S3D at the postprocessing stage to avoid various mismatches during S3D capturing. However this approach produces its own specific artifacts, with no reliable classification. Currently, we propose algorithms to detect the cardboard effect and edge-sharpness mismatch artifacts. In future work, we plan to investigate other artifacts.

We believe that our work will motivate the development of stereoscopic video quality standards. In relation to VAWN program goals, that means that network traffic will be decreased since low-quality stereoscopic video contains inter-view mismatches requiring extra bits from encoder.

"We believe that our work will motivate the development of stereoscopic video quality standards."

Quality Issues in Stereo Video Capturing

Usually, the binocular impairments belong to one of the following types:

- geometry mismatch
- color mismatch
- sharpness mismatch
- time asynchrony

We designed a set of methods to automatically estimate each of them excluding time asynchrony. Another problem with stereoscopic videos is excessive parallax that causes accomodation-vergence conflict. We have a method to detect potentially problematic frames with respect to excessive parallax.

Excessive parallax. Many authors[9][10][11][12][13][14] consider the accommodation-vergence problem caused by excessive parallax to be the primary cause of visual discomfort and fatigue associated with viewing 3D. The problem, in essence, is that under natural viewing conditions, vergence and accommodation stimuli are equal to each other, whereas in stereo films, they can be significantly different due to excessive horizontal parallax. Normally, even in natural viewing, there exists a certain degree of tolerable mismatch between accommodation and vergence, since neither mechanism is ideal and both have limited accuracy and sensitivity. A large depth of focus makes accurate accommodation useless because retinal image quality is constant when changing accommodation by (0.2–0.3) D, an exact tuning to the object distance appears to be excessive, and accommodation mechanisms perform only minimal adjustment. From a functional point of view, high accuracy is not very important. On the contrary, a viewer must see clearly both the object of interest and its surroundings. Some quantitative data on this topic is provided by Daum.[15]

We designed an algorithm to detect excessive-parallax scenes.[16] It can be briefly summarized by the following steps:

1. Rough estimation of the inter-view disparity map
2. Disparity map filtering
3. Creation of a distribution graph that shows the number of pixels according to the disparity in the timeline. Beforehand, we ask the user to input the display system characteristics where the maximum artifact-free parallax level can be obtained.
4. Detection of the problem scenes.

An example of parallax monitoring graph is shown in Figure 6. We then mark frames that have a parallax exceeding a certain value.

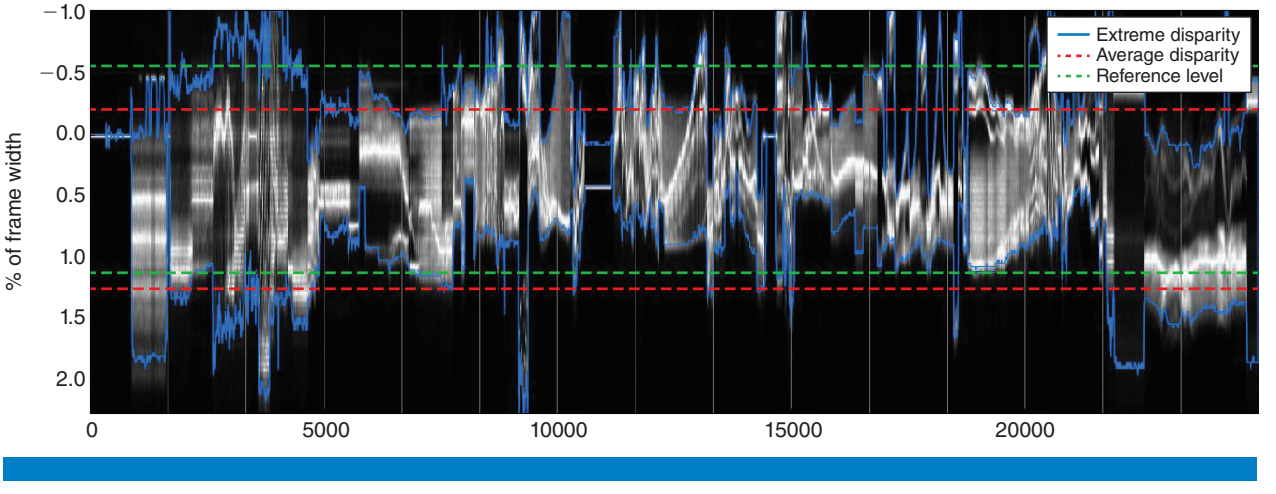


Figure 6: Example of a parallax control chart for the movie *Pina*. The parallax clearly differs from scene to scene and a large parallax is achieved only in a small portion of scenes.
(Source: Lomonosov Moscow State University, 2013)

Geometry distortions. When stereo images are displayed on a screen frontoparallel to the viewer so that horizontal lines on the screen are parallel to the line joining the eyes, all the conjugated points corresponding to virtual objects in the spatial scene should have only horizontal displacement, not vertical. Indeed, in the real world, any point in space, along with the two eyes of the viewer, defines the epipolar plane that intersects the screen in a horizontal line. Therefore, the rays directed from the eyes to a single object should intersect the screen at points located on horizontal lines; that is, at points displaced horizontally on the screen. Thus, to simulate a real scene, a correct stereo pair should have no vertical disparities, and on the screen, the left and right images must be presented without vertical disparities.

At the same time, it is necessary to take into account that in natural vision, vertical disparities are usually present. For example, in cases where the distances from the left and right eyes to the object in view are different, the two retinal images will be of different sizes in both horizontal and vertical dimensions. Psychophysical experiments reveal that the vertical-disparity gradients

significantly affect perception of 3D form, depth, and size.[17][18][19] Regarding quantitative data, Stevenson and Schor in 1997[20] found that matching stereo images is not restricted to epipolar lines and that people are able to estimate depth accurately even in cases of vertical disparities up to $45'$ (retinal angular minute). Currently we estimate vertical parallax and tilt.[16][21]

Color mismatch. It is well known that under natural viewing conditions involving a real scene, the binocular visual mechanisms use not only geometric disparities (cues based on the relative positions of objects in depth) but also differences in luminance, contrast, and color between the left and right images, as well as asynchrony in their temporal changes.[22] In particular, local differences in luminance are characteristic of images containing lustrous surfaces and transparent objects. In this way, color and luminance differences can cause false perception. Our system of binocular vision, however, is sufficiently robust to handle color differences between views. Recently, new data has surfaced indicating that depth perception can survive significant interocular differences in luminance levels up to 60 percent.[23][24]

"Our system of binocular vision, however, is sufficiently robust to handle color differences between views."

Our metric of color differences is described elsewhere.[16][21][25]

Sharpness mismatch. The effect of blur can significantly affect binocular perception. Specialists in binocular vision have long been aware that when image blur is asymmetric, the quality of the binocular percept is determined mainly by the sharper of the two images. In particular, this conclusion was mentioned by Stelmach et al.[26] This conclusion is true only for a certain range of stimulus parameters, however. A more recent study[27] has shown that the perceived quality of an asymmetrically degraded image pair is roughly the average of both perceived qualities when one of the two views is degraded by very strong JPEG compression. Kooi and Toet[28] also note that fusing a good image with a blurred image requires a few more seconds than fusing the original image pair. Unfortunately, insufficient data is available (concerning various questions that need to be clarified) to perform a comprehensive analysis of the issue.

A detailed description of our sharpness-mismatch detection algorithm is presented elsewhere.[21][25] The metric is designed to detect differences in high frequencies caused by focus mismatches and also by inaccurate postprocessing, differences in motion blur, and asymmetric compression.

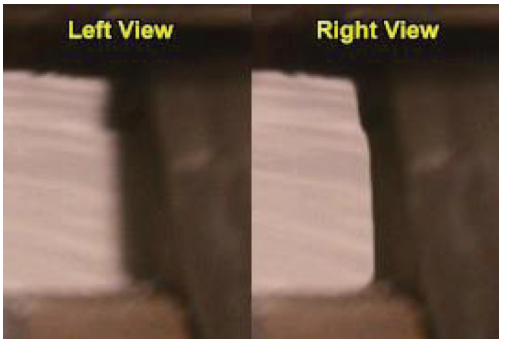
Quality Issues in Stereo Video Conversion

Converted videos are potentially better for VAWN goals since there exists a ground-truth sequence of depth maps, thus they can be easily represented in 2D+Z or MVD formats (we will discuss the advantages of these formats for compression in the section "Delivery"). However, the process of conversion requires significant manual work (including depth drawing), thus the price/quality ratio is low, resulting in audience distrust. Our research on the quality of captured stereoscopic video motivates us to study the quality of 2D-to-3D conversion. We believe that our work will improve quality of stereo creation by 2D-to-3D conversion.

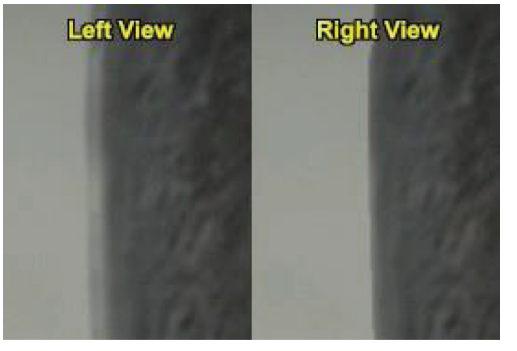
"Our research on the quality of captured stereoscopic video motivates us to study the quality of 2D-to-3D conversion."

Edge-sharpness mismatch. The term edge-sharpness mismatch (ESM) describes defective stereo pairs with specific asymmetric impairments. It refers to any inconsistencies in object edges between the stereoscopic views (edge-sharpness variation, edge doubling, and so on). Under the viewing conditions in a real environment, such situations rarely occur. In the case of 2D-3D conversion, however, the likelihood of ESM occurrence can be rather high. During the 2D-3D conversion workflow, ESM can be caused (besides an inaccurate depth map) by:

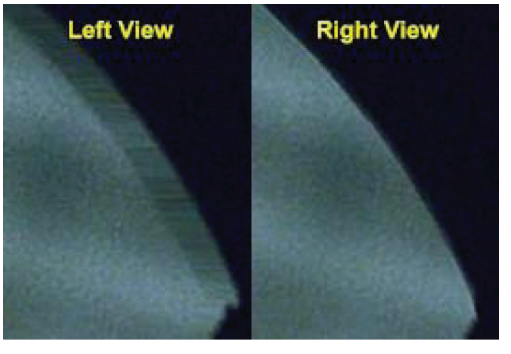
- Use of a "rubber sheet" occlusion-filling technique, defined as warping the pixels surrounding the occlusion regions to avoid the explicit occlusion-filling step (Figure 7a)
- Lack of proper alpha-channel treatment (Figure 7b)
- Simple occlusion-filling techniques where background or foreground pixels are stretched across the entire occluded region (Figure 7c)



(a) Result of "rubber sheet" occlusion filling



(b) Result of the absence of proper alpha channel



(c) Result of a very simple occlusion-filling technique based on stretching foreground and background pixels

Figure 7: Examples of edge-sharpness mismatch. Each example is a magnified fragment of a stereoscopic picture
(Source: Lomonosov Moscow State University, 2014)

Our proposed approach[29] is based on edge detection and matching:

1. Disparity map construction
2. Estimation of edge map for each view
3. Matching edge pixels using disparity map

4. Per-pixel edge sharpness mismatch estimation
5. Rejecting the results when the background changes significantly

Cardboard effect. Cardboard effect is a term referring to an unnatural flattening of objects in perceived visual images (the objects look like pieces of cardboard placed parallel to the screen). Long before video professionals gained widespread experience with stereoscopic movies, stereo photographers observed the cardboard effect. These photographers analyzed its causes and tried to formulate shooting conditions that minimize it.[30] In the case of 2D-3D conversion, the cardboard effect refers to mismatch between the perceived frontal size of the object (the visual angle occupied by the object in the visual field) and its perceived depth (thickness).

The proposed algorithm[29] for flat foreground objects detection consists of the following steps:

1. Disparity map construction
2. Mean-shift segmentation of the obtained map
3. Calculation of the median disparity value for each segment
4. Calculation of the variance around this value

Delivery

The main problem underlying the VAWN program is the rising amount of video content being transferred over wireless networks while network bandwidth remains the same. The increasing popularity of 3D video makes the problem even worse because of the additional data required for 3D. Hence the delivery of 3D video with the least amount of bitrate increase is desired. Although some communication networks cannot be extended for various reasons (such as cost and physical limitations), a 3D-specific video codec can help minimize this additional bitrate. However, most existing 3D video compression methods yield a significant bitrate increase over their 2D counterpart to achieve the same perceptual quality.

The current standard for 3D video compression is the multiview coding (MVC) extension of H.264/AVC. Bitrates generated by MVC are linearly proportional to the number of encoded views[31], and thus not scalable for use with autostereoscopic displays. Moreover, with multiview video representation, the number and location of the views are restricted to the captured data. In this work we pay attention to promising encoding techniques based on depth map storage that address the limitations of MVC. Specifically, two approaches to compact depth map storage are presented based on existing MVD and 2D+Z formats (Figure 8).

"In this work we pay attention to promising encoding techniques based on depth map storage…"

Multiview + depth format (MVD) assumes depth map storage and transmission for one view or several views. The presence of a depth map significantly decreases the complexity of intermediate view generation. This format is also compatible with display systems with two views. The size of additional data is even larger here than for pure multiview formats. However,

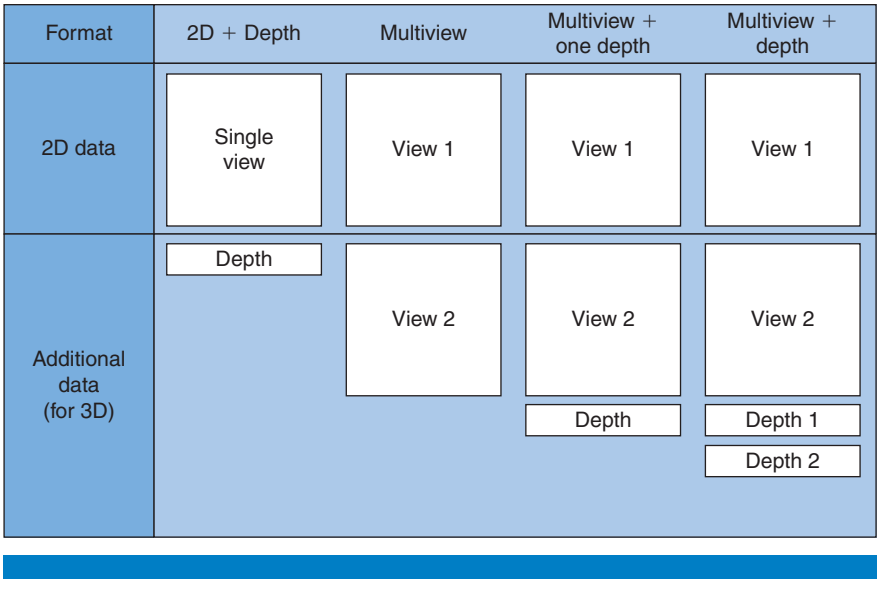


Figure 8: Comparison of 3D video formats in terms of additional data required for extending 2D video to 3D
(Source: Lomonosov Moscow State University, 2013)

the additional depth information enables new coding tools that exploit the correlation between the views of a 3D video better than existing tools. With depth-based 3D video, it is possible to synthesize virtual views and use them as a means of prediction within the codec. Currently, the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) is working on the standardization of a depth-based 3D video representation and its coding methods.

The 2D+depth format is supported by the 2D-to-3D video conversion industry, Dolby 3D format[32], and several TV set manufacturers. This format provides a 3D video experience with minimal additional data. The depth map can be compressed very effectively so almost all available bandwidth is used for 2D streaming. Consequently, this format is the most promising one in terms of compatibility with 2D devices and minimal additional data. This section presents three main options for reducing the overall bitrate needed to deliver 3D content: (1) 2D+depth compression, (2) MVC+depth compression, and (3) mixed resolution coding.

Multiview Video Compression Based on Depth Map Propagation

Let's consider the additional data that should be added to a 2D video stream to provide a 3D video experience. To minimize this data, we use the most effective 3D video format: 2D+depth.

The naïve solution for 2D+Z compression is using conventional 2D video codecs for both 2D and depth maps. This approach doesn't take advantage of the strong correlation between the 2D view and depth map. The depth map also has a structure different from the video structure, and since it has

no texture, conventional video compression approaches are not efficient. There are several methods for utilizing correlation in 2D video and depth maps. For example, Choi et al.[33] use a 2D image to increase the frame rate and resolution of the depth map obtained using a depth sensor. Modified cross-bilateral filtering is used to increase spatial resolution. Frame rate is doubled using temporal interpolation based on motion vectors estimated from 2D video. Depth map restoration from sparse key frames requires a more complex interpolation procedure because of the larger difference between distant frames.

De Silva et al.[34] perform joint compression of video and depth maps using motion vectors estimated from the source video. The input scene is segmented and extracted background is transmitted independently. Depth maps are considered only for foreground objects. Additionally, motion vectors are estimated and transmitted with the compressed video stream.

Our proposed approach is based on sparse representation of depth maps. The general pipeline is shown in Figure 9, where only downsampled key frames for a depth map are compressed. In the simplest case, the encoder for a depth map only selects every *n*th frame, downsamples it, and compresses using any conventional image or video codec, yielding a very low bitrate for the depth map.

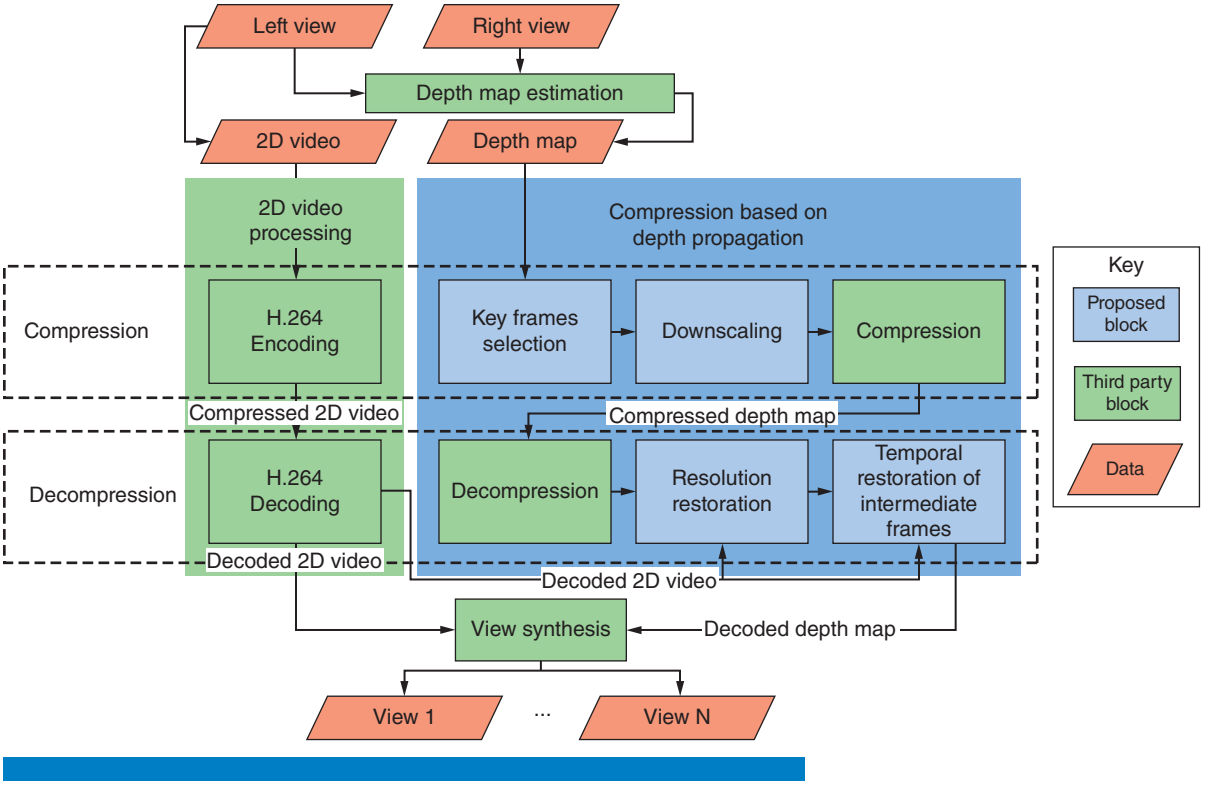


Figure 9: Depth propagation-based compression scheme for 3D video (Source: Lomonosov Moscow State University, 2013)

The decoder decompresses the 2D video stream and then uses it for depth map decompression. First the decoder restores spatial resolution of the depth map for key frames using appropriate decoded high-resolution 2D frames. This can be done using numerous depth map upscaling methods.[4][6][35] The method is used by YUVsoft Depth Upscale[35], where rough edges in a low-resolution depth map are enhanced using a high-resolution 2D color frame.

Then the decoder restores missing depth maps for the rest of the frames. We use block-based motion estimation to get information about object motion from the 2D video. The depth map is assumed to be correlated to 2D video motion flow and the depth map can be interpolated using key frames and motion information. Appropriate nonlinear edge-aware postprocessing is used to conceal motion compensation artifacts. Such a propagation procedure is held for each video interval between key frames. Depth is propagated from the first and the last depth key frame of the interval. Then the two depth maps are merged according to the degree of confidence in motion information, based on Simonyan et al.[36]

The final decoding step depends on the target device, where the required views are generated using a dedicated algorithm.

Quantitative Results

The performance of depth-propagation–based compression is compared with depth map compression using x264 codec. 2D video compression is the same for both cases, so the bitrate of 2D video is not considered. We consider only the bitrate of additional data (that is, depth map bitrate).

"Due to the complexity of stereoscopic perception there is no generally accepted method for quality measurement."

Due to the complexity of stereoscopic perception there is no generally accepted method for quality measurement. Research in this direction is still in progress.[37][38] The most common approach is using a conventional 2D quality metric (such as PSNR). Direct measurements of the depth map difference before and after compression are not relevant because the depth map influences synthesized view quality in a nontrivial way. Therefore 2D quality metrics are applied to synthesized views, for example by Lee et al.[39] Both methods were used for quality evaluation of the propagation-based approach.

In our experiments, we used constant intervals between key frames. We tested intervals of 10, 20, 40, and 100 frames. Depth key frames were downscaled using factors 1, 2, and 4. For key frame compression we used the JPEG 2000 image codec. Full-resolution key frames are restored using YUVsoft Depth Upscale.[35] The full depth map is restored from key frames using YUVsoft Depth Propagation.[40] We took the best settings of the proposed propagation scheme in terms of quality and bitrate.

The proposed technique demonstrates very promising results (Figures 10 and 11). It outperforms depth map coding using x264 up to 15 times in bitrate while preserving the same quality. The highest gain is achieved on the lowest bitrates.

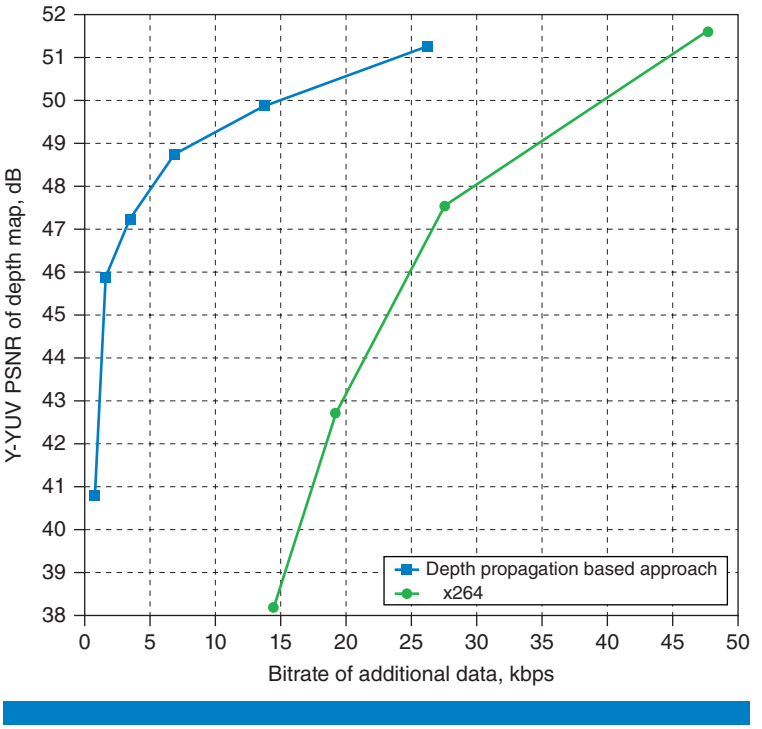


Figure 10: Comparison of depth map compression using depth propagation based approach and x264. Only the bitrate of additional data is considered.
(Source: Lomonosov Moscow State University, 2013)

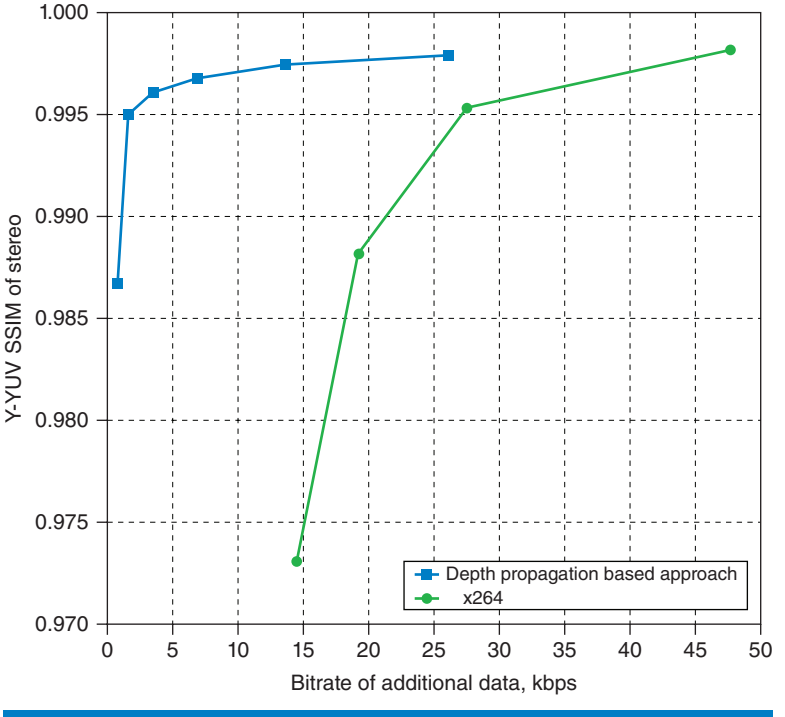


Figure 11: Comparison of reconstructed stereo for depth propagation based compression and x264. Only the bitrate of additional data is considered.
(Source: Lomonosov Moscow State University, 2013)

While the depth-propagation–based approach demonstrates good potential compression, there are a number of ways it can be improved. One of the improvements is adaptive key frames selection. Static scenes require a lower bitrate, so the algorithm can use sparser key frames. Dynamic scenes must be encoded using dense key frames to achieve acceptable quality. Starting from a sparse arrangement, we add key frames to maximize the target metric. An example of per-frame SSIM of stereo for a part of the *Basketball* sequence is presented in Figure 12. Adaptive key frames selection demonstrates even more gain in comparison with the basic approach with equidistant key frames, shown in Figure 13.

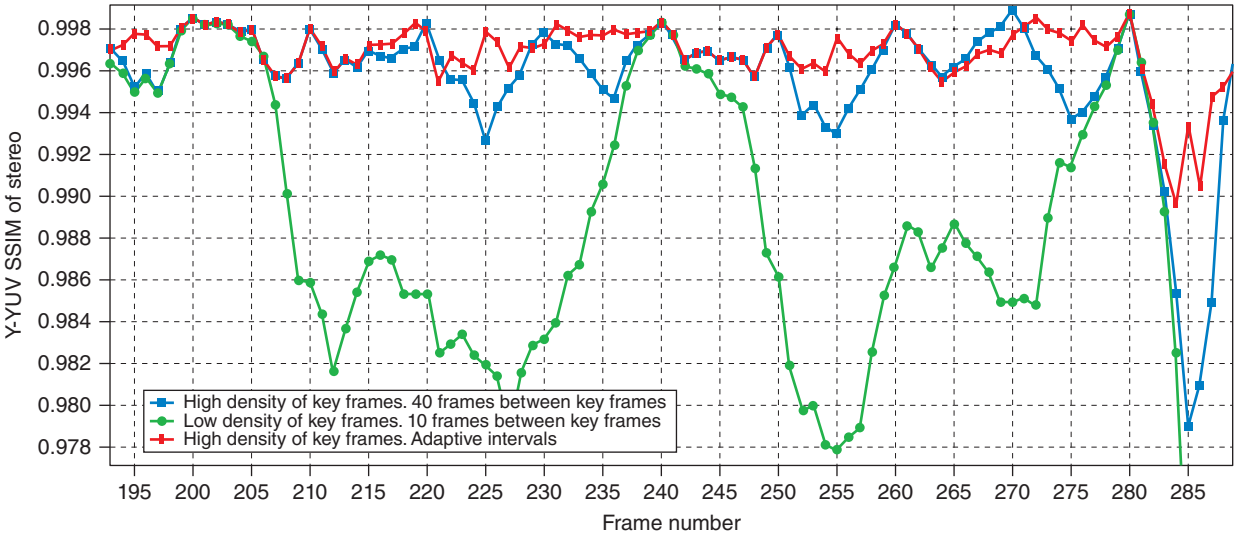


Figure 12: Per-frame SSIM scores of proposed depth-propagation-based compression with different key strategies for choosing frames
(Source: Lomonosov Moscow State University, 2013)

Advantages and Disadvantages

The proposed technique demonstrates good compression performance, where a quality 3D experience is delivered with minimal additional bandwidth. Compression efficiency is achieved at the expense of high computation cost of decompression. Depth map propagation on the decoder requires a rather large buffer of decoded 2D frames. The number of frames depends on the compression rate, which can reach tens of frames. Consequently, a small delay at the decoder is inevitable. The depth-propagation–based approach inherits all the advantages and disadvantages of the 2D+Z video format. This format and MVD support an arbitrary number of displayed views and arbitrary parallax tuning in a reasonable range. This capability is important due to the variety of existing 3D display sizes and formats. Each of them

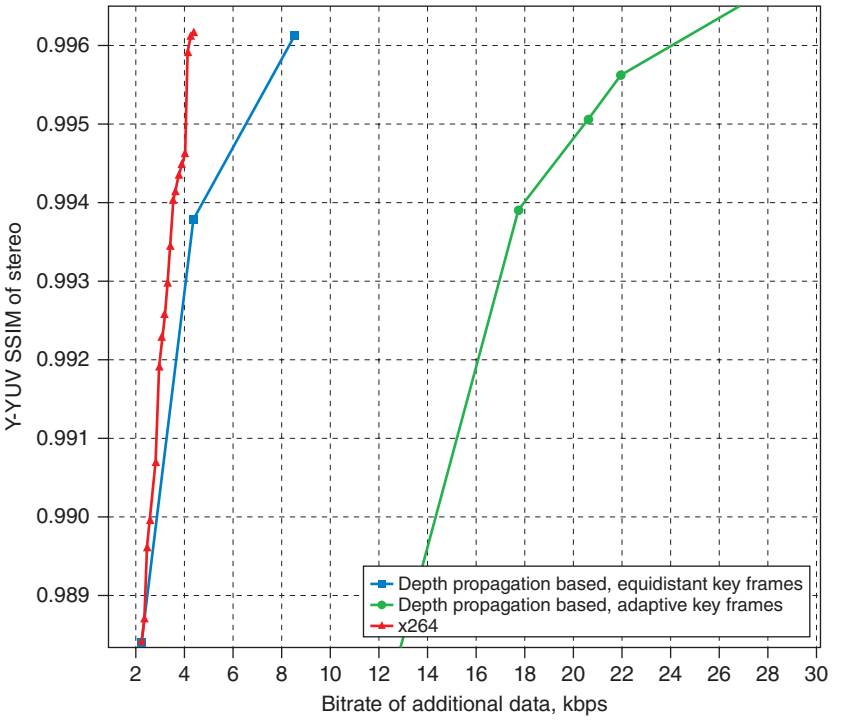


Figure 13: Comparison of reconstructed stereo for depth-propagation–based compression and x264 on a sequence from the *Pirates of the Caribbean* trailer. Only the bitrate of additional data is considered (Source: Lomonosov Moscow State University, 2013)

requires appropriate stereo parameters. The 2D+Z format is also supported as a native format in a variety of displays. The 2D+Z format is impossible to use for correct processing of semitransparent objects. However, this drawback can be concealed by additional data for semitransparent region representation. The quality of the final image strongly depends on the quality of stereo occlusion processing. The inpainting algorithm for these areas is critical, although the inpainting area is typically small. On small screens, parallax must be high, but the overall size of the display is low. On the large screens, parallax must be small, so the area to fill is not very big. This allows one to successfully apply inpainting algorithms.

Multiview Video Plus Depth Coding with Depth-based Prediction Mode

An alternative to the 2D+depth compression technique discussed in the previous section is to utilize depth information as a complement to existing multiview coding (MVC). This section describes how this can be done and quantifies the resulting improvements in quality and bitrate needed to deliver the content.

Depth-based Prediction Mode (DBPM)

DBPM allows the use of a synthesized reference picture for prediction without any high level syntax changes to the MVC and requires only simple macroblock-level syntax changes to the standard. It can be used concurrently with existing prediction modes of MVC without introducing a significant overhead due to changes in syntax.[41][42]

In our codec, the base (first) view video and its associated depth map are encoded with H.264/AVC individually. Then while encoding a frame of an additional view, a virtual view is rendered from the base view data at the corresponding time instance. The virtual view is rendered simply by the well-known Depth Image Based Rendering (DIBR)[43] algorithm by projecting the base view pixels onto the current viewpoint, without any hole filling at the disocclusion regions. Once the proposed mode is signaled to the decoder, the decoder refers to this reference virtual view and copies the collocated macroblock for prediction. If there is additional residual information in the bitstream it also adds the residual information. An illustration of DBPM is provided in red in Figure 14a along with the DIBR operation, which is depicted in blue. A block diagram of the proposed codec is provided in Figure 14b.

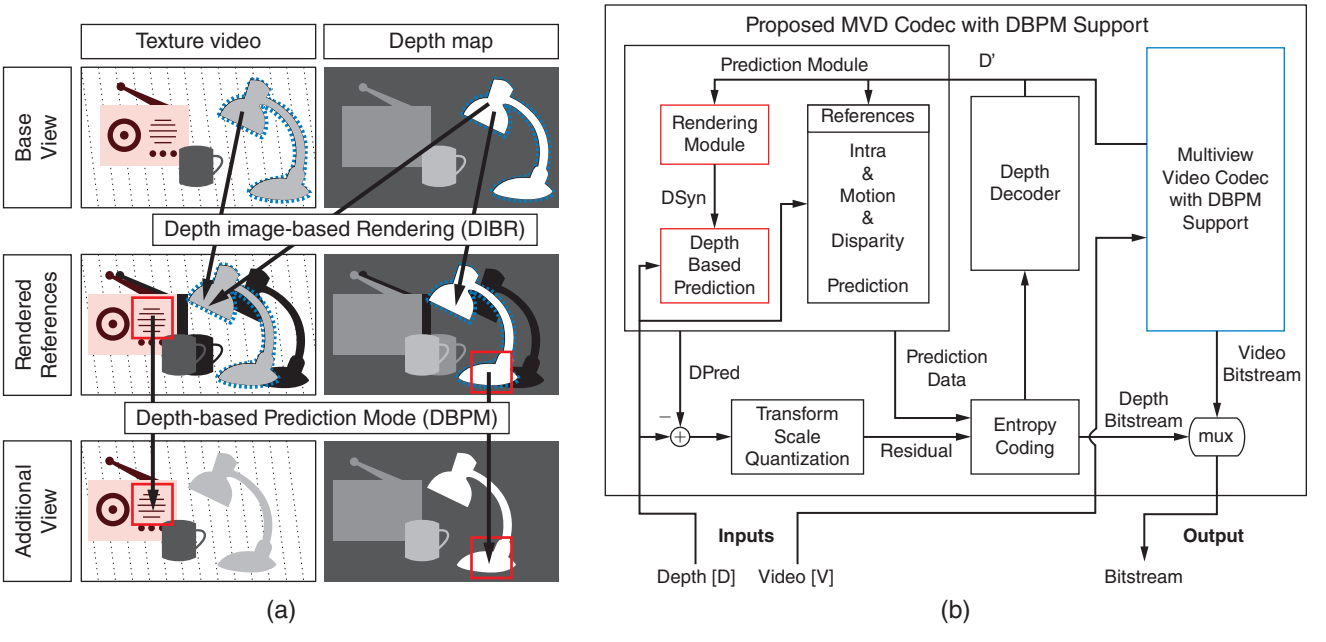


Figure 14: (a) Illustration of the depth-based prediction mode (b) Block diagram of the proposed MVD codec with DBPM support

(Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

Rate-Distortion Analysis

We compare the coding performance of the proposed codec with MVC using rate-distortion (RD) curves and Bjontegaard Delta Rate (BD-Rate).[44] RD curves measure the coding performance of a codec in terms of a quality metric (such as PSNR) and the corresponding bitrate levels. BD rate values measure the percent bitrate savings between two RD curves, where negative values signify gain. We analyze the rate-distortion performance of DBPM in two different contexts. First, we provide results for the proposed MVD codec. The proposed codec encodes the texture videos and the depth maps, both with DBPM support. In comparison, we use MVC to encode texture and depth channels disjointly. Second, to isolate the contribution of the DBPM support for depth maps, we analyze its prediction performance in the context of depth map coding.

The BD rate results for coding MVD data using the proposed codec versus MVC are reported in Table 4. These results show that the proposed codec can achieve up to 9.2 percent bitrate savings with DBPM support, and as expected, the gains vary depending on the depth map quality. For example, *GTFly* and *UndoDancer* are among the sequences with the largest gain since they are computer-generated sequences with ground truth depth maps. In comparison, the depth maps of the *Newspaper1* sequence are noisy and consist of both temporal inconsistencies and spatial errors. This leads to inefficient coding of the depth maps and geometric distortions in the rendered references for DBPM. Thus, *Newspaper1* is among the sequences that benefited the least from DBPM support.

Depth QP	PoznanHall2	PoznanStreet	UndoDancer	GTFly	Kendo	Balloons	Newspaper1
26	–5.26	–3.81	–7.06	–9.19	–2.03	–2.57	–2.55
31	–5.35	–4.04	–5.52	–9.03	–2.53	–3.16	–3.17
36	–4.95	–3.75	–3.93	–8.51	–2.84	–3.33	–3.52
41	–4.49	–3.09	–2.94	–7.50	–2.88	–3.53	–3.64

Table 4: BD rate (%) for coding MVD data, 3 views—measured against MVC (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 6, pp. 995–1005, Jun. 2014)

BD rate allows measurement of bitrate savings in a concise manner, yet it fails to associate these savings with the absolute number of bits saved. Hence, in addition to the BD rate results, we also provide the RD curves for the *GTFly*

sequence, which yields the most gain for the proposed codec. Figure 15 shows that the proposed codec can deliver the same quality of MVD data with up to 900 kbps less bitrate than MVC.

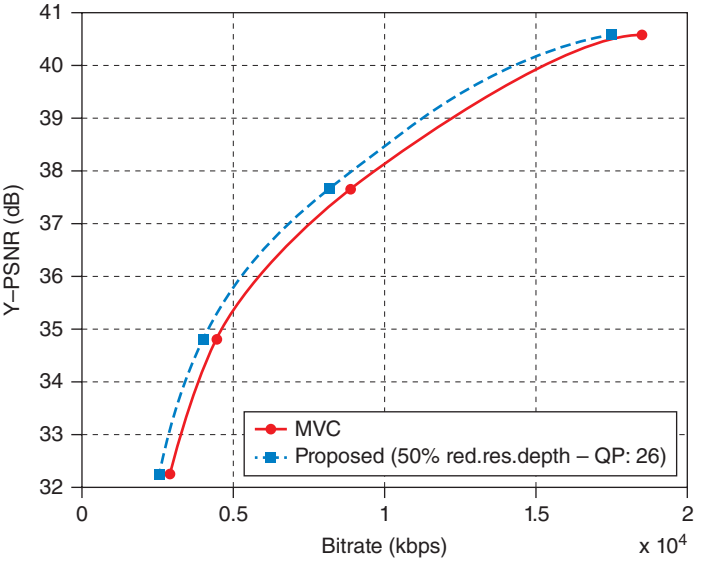


Figure 15: RD curves for *GTFly*, 3 views, coding MVD data (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

We also provide BD rate results for depth map coding in Table 5. Since depth maps consist of piece-wise smooth regions, DBPM faces stronger competition against existing MVC prediction modes. Looking at the results in Table 5, DBPM proves to be successful with up to 9.9 percent bitrate savings when the depth maps are accurate. On the other hand, for depth maps with limited accuracy, the encoder chooses DBPM infrequently and the DBPM-enabled codec starts to yield slightly worse performance than MVC. These bitrate losses are limited to around or less than 1 percent, and they are due to the syntax overhead introduced by DBPM.

PoznanHall2	PoznanStreet	UndoDancer	GTFly	Kendo	Balloons	Newspaper1
1.03	–1.69	–2.98	–7.56	0.08	0.32	0.75

Table 5: BD rate (%) for coding depth maps, 3 views—measured against MVC (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

Mixed Resolution Stereoscopic Coding

A final technique to reduce the required over-the-air bitrate is to blur one eye's view compared to the other, allowing for a lower overall bitrate. Mixed resolution stereoscopic coding (MRSC) relies on the perceptual phenomenon of binocular suppression, where if one eye's view of the world is blurry while the other eye's view is sharp, then the fused 3D percept of the scene will appear almost as sharp as the high resolution view and will be faithfully represented in depth.[45] Mixed resolution coding implements this idea by transmitting a stereo pair comprised of one full resolution image and one lower resolution image.

"Mixed resolution stereoscopic coding (MRSC) relies on the perceptual phenomenon of binocular suppression…"

One concern for MRSC is that one eye continually receives a low resolution or blurry input. In the following subsection, we investigate the perceptual response for two methods of MRSC.[46] The first method, single-eye blur, is to blur all frames of the video corresponding to one of the eyes. The second method, alternating-eye blur, is to blur alternate frames of each view, such that there is one blurry and one sharp frame at each time instance, and the view that is blurred alternates with each frame.

There are applications, such as high quality viewing or decoder-side processing, for which a full-resolution stereo pair is beneficial. A super-resolution algorithm for single-eye blur MRSC videos is presented by Jain and Nguyen.[47]

Quality Experiment

In order to compare the perceived quality of the two processing methods, we asked subjects which of the pair of videos, each processed according to one of the two methods, they preferred. We used four high quality stereoscopic video clips, four blur levels corresponding to a diameter of 2, 4, 8, or 16 pixels (1.1 arcmin to 9.0 arcmin) of a disk kernel, and three frame rates: 30 Hz, 60 Hz, and 120 Hz. Each of the 48 unique test conditions was repeated four times for a total of 192 trials, which were tested in random order. Stimuli were presented on a pair of CRT displays viewed through a mirror stereoscope at a distance of 6.4 feet (6° horizontal per eye). Twenty-three subjects participated.

Figure 16 shows the proportion of trials where subjects preferred single-eye blur over alternating blur, as a function of blur diameter, for the three different frame rates. We did not see any consistent difference in preferences for blur type between the four different source videos, so we have combined data across that factor.

For a refresh rate of 30 Hz it is clear that single-eye-blur is preferred. For 60 Hz and 120 Hz, there is no real evidence of a preference below blur diameters of 8 pixels.

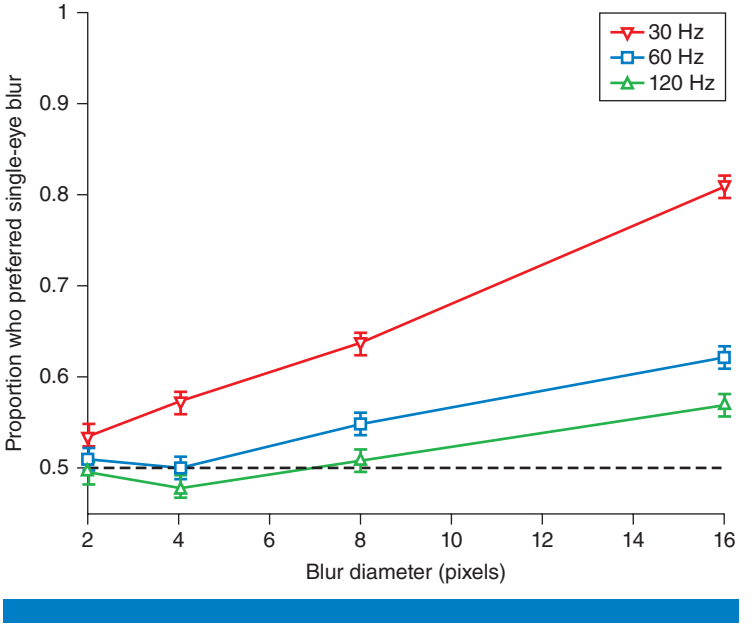


Figure 16: Proportion of trials in which single-eye blur was preferred over alternating-eye blur
(Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

Fatigue Experiment

In this experiment, we compiled video clips into a single 5-minute source video and looped it twice to make a 10-minute exposure. We showed subjects this video processed according to each of the two processing methods and had them continuously rate their visual comfort level using a slider with scores ranging from 1 (very uncomfortable) to 5 (very comfortable) using the system presented by Jain et al.[48] Twenty-two subjects participated.

The mean scores across subjects over the duration of the test videos are shown in Figure 17. For both test methods, the mean scores generally range from 3 to 4 (fair to good comfort). The alternating blur method is rated higher, with an overall mean of 3.86 compared to 3.74 for the single-eye blur.

The dashed lines in Figure 17 indicate the ends of each of the four clips, and the dotted lines indicate a scene change. In addition to the general preference for alternating blur over time, there is some dependence on the type of content as certain clips are less straining to watch than others. For instance, the "Looney Tunes" clip from about 2:05–4:25 and again from 7:05–9:25 reflects the largest difference in scores between the two methods, with the alternating blur being preferred. The animation contains high contrast, flat textures, and sharp edges. All of these features produce

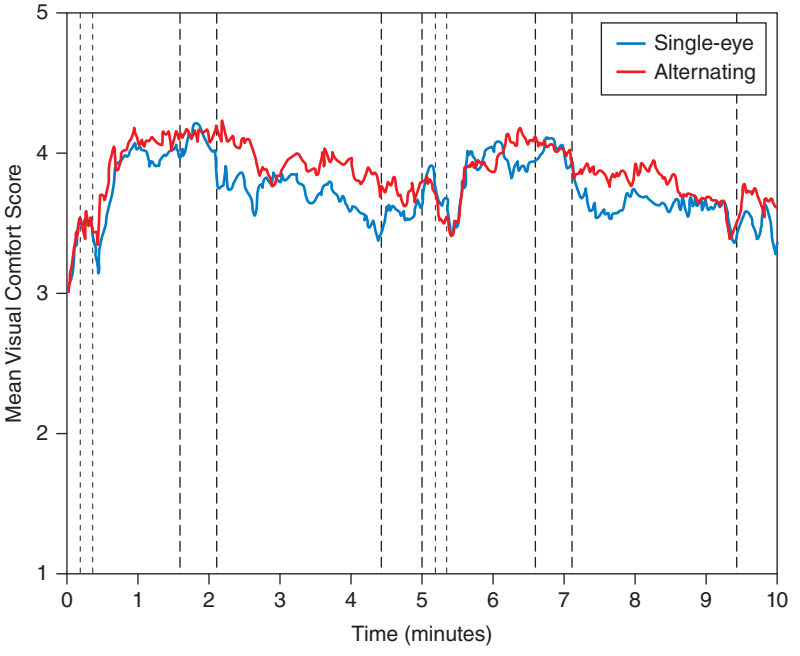


Figure 17: Mean scores across subjects over duration of video for each processing method. Dashed lines separate each video clip in the sequence
(Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

artifacts under asymmetric blur, which are quite salient in the single-eye blur case.

The mean scores are relatively high for both methods considering the level of blur applied to the sequence. A 20-pixel diameter was chosen for the disk filter, which corresponds to a downsampling ratio of 8.33 in each dimension or about 69.44 overall.

Display

Usually, the quality of compression is measured by a rate-distortion ratio, thus there exist two ways to decrease traffic over networks:

- to decrease the bitrate for the same distortion level,
- to decrease the level of distortions for the same bitrate.

This section is mainly about the second approach, but viewing the distortion and quality from the end user perspective. For 3D viewing, quantifying quality is still an unknown and difficult task. Without a way to quantify quality for different bit rates and distortions, it is difficult if not impossible to optimize the end-to-end delivery system, which was an objective of the VAWN research. The final viewing quality is determined by the quality of the video itself and

the display equipment. Thus, it is important to study the quality of the whole stereoscopic video life cycle and the problems of automatic stereoscopic-display adjustment and choice of the proper content. The first part of this section will cover the unique processing requirements needed to render content for autostereoscopic displays, which relies heavily on an accurate depth map as discussed earlier. Then a tool will be presented that helps to capture subjective quality scores efficiently, which can hopefully be used more broadly with the industry in order to better quantify 3D video quality for different content, bitrate, and distortions. Finally, another tool is presented to help automatically compare the quality of the end display, since displays have a significant impact on the end user quality. It is hoped that these tools will help move the industry towards a deeper understanding of the end user perception of 3D video quality.

Autostereoscopic Display

Various autostereoscopic display technologies have surfaced in recent years. In general, they work by projecting images of a scene into the space in front of them to create two or more spatially separated perspectives. Then, based upon where an individual stands in reference to the display, each eye will perceive a different viewing angle, which leads to a disparity between what each eye sees. This disparity is translated into depth perception by the human visual system.

Although we typically use an 8-view display for a richer 3D effect, for simplicity Figure 18 illustrates a display with just two views using a sheet of

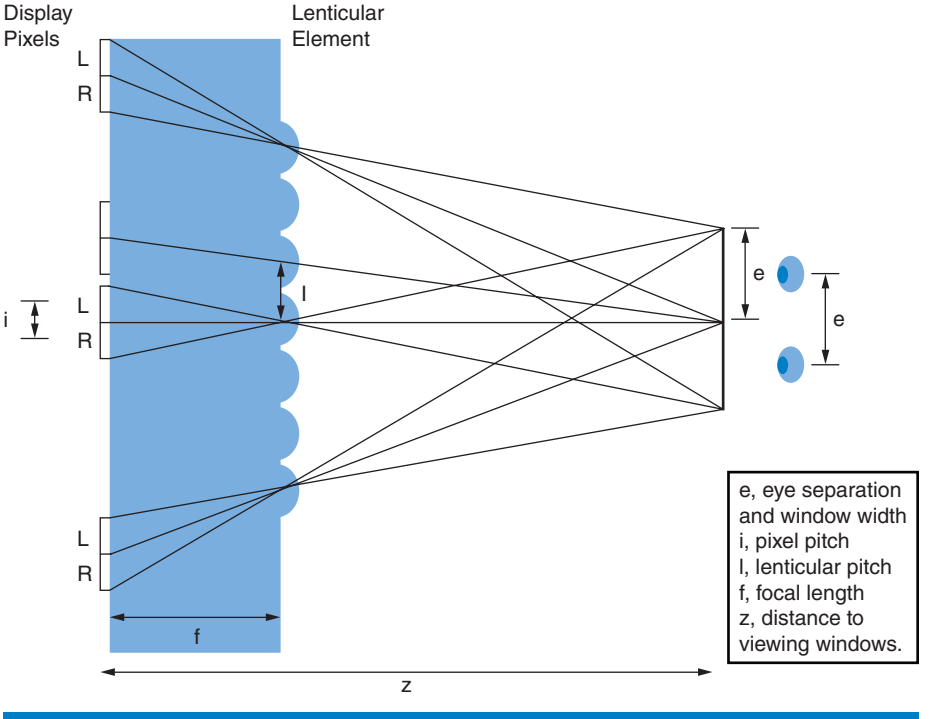


Figure 18: A top-down view of a 2-view lenticular display (Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

lenses called a lenticular sheet. The curved lenses angle the light emitted from a traditional LCD in such a way that the image incident on the left eye is slightly different from that of the right eye, creating a 3D sensation. As can be readily seen from Figure 18, with just a 2-view display, only one viewer can perceive 3D from a fixed location. Using an 8-view display allows for eight viewing zones with eight different perspectives, also known as sweet spots. This effectively allows all onlookers to see 3D from multiple vantage points. A typical problem with autostereoscopic displays is that moving between these sweet spots can produce an experience that resembles double vision when, for example, the eyes are seeing half of two separate views.

While autostereoscopic displays offer users the ability to see 3D without having to wear any specialized glasses, they require multiple viewpoints of the scene in order to display content. Typically, they require five, eight, or nine stereoscopically aligned views in order to properly display a 3D effect. In our case, this means that we would need to construct a camera system with at least eight cameras. In addition, there are strict requirements that the cameras must be as nearly identical to each other as possible, and extremely and fixed in orientation for an accurate 3D representation. Therefore, it is impractical to capture data with such camera systems. Instead, we limit ourselves to capturing the data with just two cameras. With that, it becomes a matter of providing a robust stereo-to-multiview conversion solution to take in the camera input and visualize it on an autostereoscopic display. Figure 19 shows the result of our work rendering the remaining six stereoscopically aligned views given a pair of images as input.

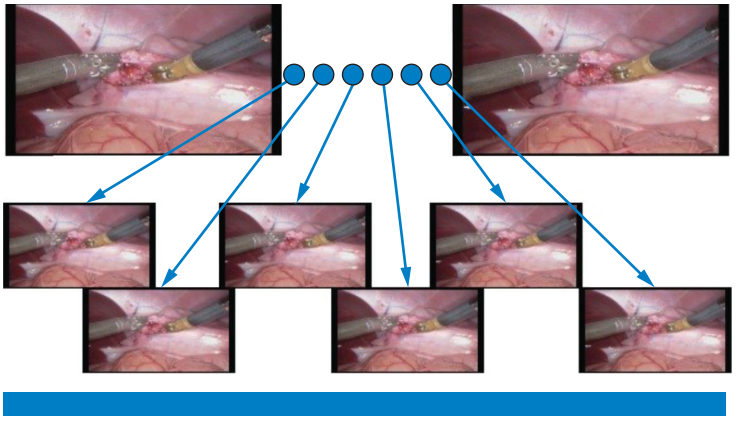


Figure 19: Rendering of six interpolated views from a stereo pair (Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

Tally: A Subjective Testing Tool

Many labs around the world conduct research that relies heavily on perceptual experiments with video. Commonly, data is collected by asking subjects to write their responses to stimuli using pen and paper, and then manually entering this data into computerized spreadsheets. Not only is this process extremely slow,

it is also prone to error. Some researchers have written custom software to automate this process, but it is not generally applicable, is not made widely and freely available, does not work for both 2D and 3D content, and does not permit testing multiple subjects simultaneously.

We developed Tally[48], a subjective testing tool, as a web-based system to solve these problems. Additionally, Tally's web-based design allows data to be accessible from anywhere, allows many people to use the same system with their individual history and data securely saved and privately accessible, allows experiments to be repeated with identical parameters and methods, and allows remote collaboration between labs through a sharing feature.

Our system consists of three major pieces: the desktop application, the web front end, and the server back end. The basic workflow of a subjective experiment is depicted in Figure 20. Prior to the experiment, the researcher uses the web front end to create a test and run it. Then, subjects log into the web front end (website) and select the appropriate test to begin. Once they are ready, the server tells the desktop application which video to play. The desktop application receives the command and plays the video to the display device.

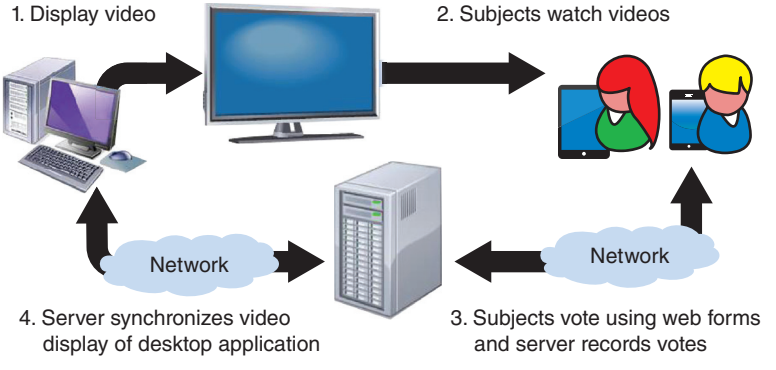


Figure 20: Workflow of the subjective testing tool (Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

Note that only the file name is sent across the network; the actual videos are stored locally on the machine connected to the display. Subjects then vote on the video using any web-enabled device such as a smartphone, tablet, laptop, or desktop, and their scores are transmitted to the server and recorded. Once the video is done playing, the server tells the desktop which video to play next, and the process repeats until all test videos have been shown. After the test, the server automatically aggregates the data and makes it available for download in several different formats.

Tally is free, open source, cross-platform, and very customizable. Any web-enabled device can be used to vote, most any video player can be used to play the videos, and any display device can be used to show the videos. We natively

support most of the standard test methods of the ITU[49], but also allow for custom test methods to be added. Tally, along with full documentation and installation instructions, is available for download at the project website (http://github.com/canbal/Tally).

Automatic Device Testing

Finally, the display device itself has a significant impact on the end user perceived quality of the 3D video. Therefore, it's important to understand how to evaluate the quality of the display and how different displays compare. Eventually, this information could be used to optimize the compression and bitrates needed to deliver a good quality of experience that takes into account the specific characteristics and quality of the display. The problem of viewing-device fair comparison existed long before the market of 3D viewing devices started growing and some solutions were proposed.[50] Nowadays the problem is urgent again. The 3D viewing devices are much more diverse than 2D ones and the space of their characteristics has more dimensions. Figure 21 shows a partial classification of existing device types.

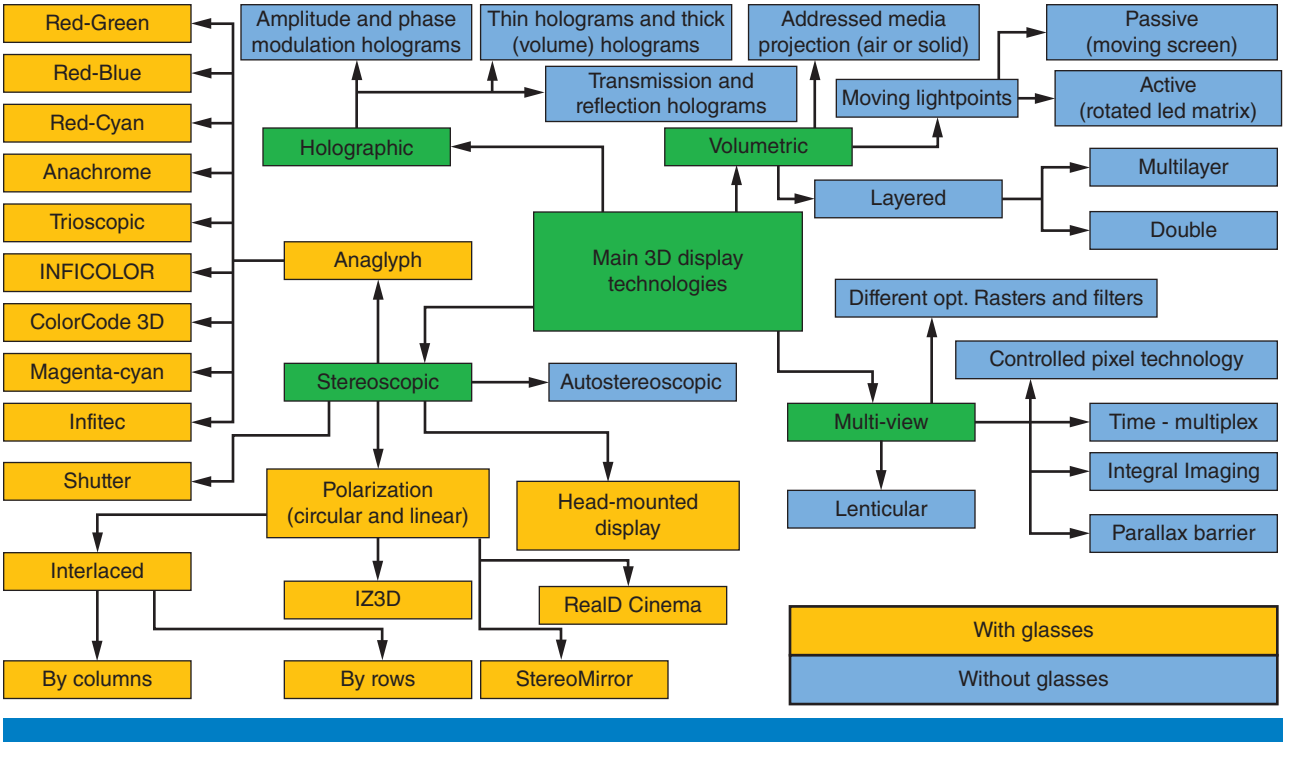


Figure 21: Partial classification of existing stereoscopic devices (Source: Lomonosov Moscow State University, 2012)

Absence of a fair comparison methodology leads to unfair competition between manufacturers and undermines user confidence in the whole market. Creation of easy-to-use software that performs a complete estimation of 3D viewing device characteristics and a database with a detailed description of each device is needed.

Content creators are also interested in understanding end-user display devices because they could provide device-specific content and require device-specific settings for that content.

Proposed Pipeline

Our proposed pipeline is based on the well-known concept of special *patterns* that enable estimation of each individual device characteristic. However, understanding *patterns* is not an easy-to-use approach, especially when the required result is a quantitative one. Consequently, we propose a system requiring from the user a series of device shots and locating automatically the relative position of the shot using special QR codes placed behind the screen, robustly estimating device characteristics at each spatial position in front of the device (normally, manufacturers declare the best value). The result of the measurement is continuous maps of each characteristic spatial distribution interpolated from points of shots. We briefly illustrate how the system works in the Figure 22. Some common problems for specific technologies are:

- Crosstalk
- Geometric distortions (stereoscopic two-projector systems)
- Time asynchrony (stereoscopic two-projector systems)
- Brightness decrease

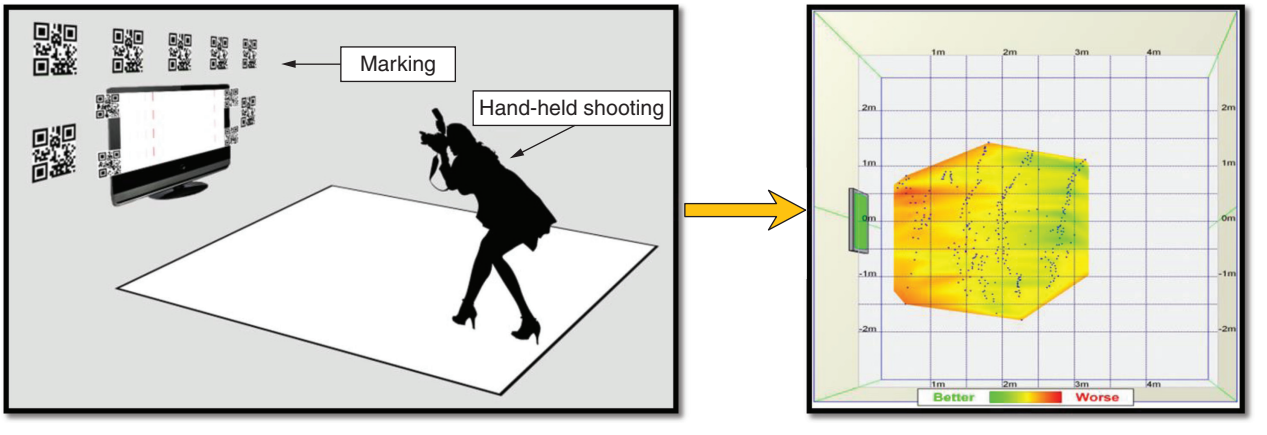


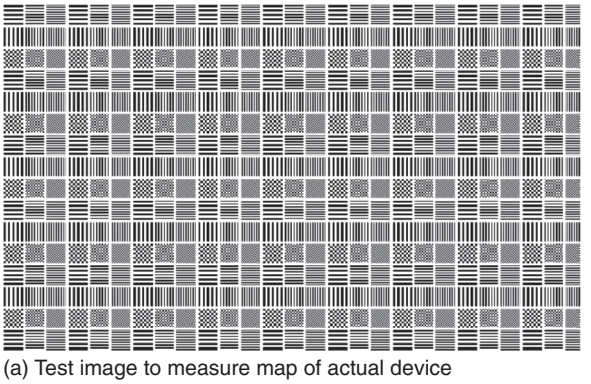
Figure 22: Brief illustration of our proposed pipeline. A user performs a series of shots from different positions in front of the device. The system determines each shot position relative to the device using QR codes placed behind the device and estimates a set of available characteristics. Finally, sparsely estimated characteristics are interpolated into a continuous map
(Source: Lomonosov Moscow State University, 2013)

Moreover, some integral characteristics are important to improve viewing quality:

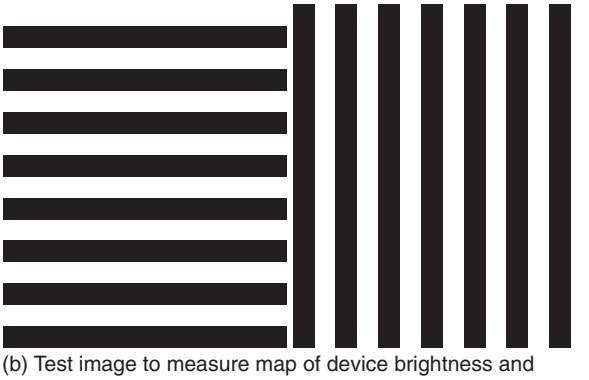
- Optimal observing distance
- Width of view zones (autostereoscopic displays)
- Actual resolution
- Actual number of views (autostereoscopic displays)

Estimated Characteristics

Brightness and crosstalk. Viewing quality of a 3D device can not be expressed by one single value. Actually, viewing quality significantly changes with respect to the observer's position. Mainly, it is determined by changes in brightness and amount of crosstalk (view mixes). To simplify the testing process, we use the pattern from Figure 23b, enabling us to measure brightness and crosstalk maps for each view simultaneously.



(a) Test image to measure map of actual device resolutions



(b) Test image to measure map of device brightness and crosstalk

Figure 23: Examples of test images to measure characteristics of a 3D viewing device (Source: Lomonosov Moscow State University, 2012)

Resolution. Most of the S3D viewing devices assume decreases in horizontal resolution up to the number-of-view-zone times. The exact number depends on the amount of crosstalk. Normally, the resolution reduction is spatially uniform, thus we measure only one number (not a map). We show the test image we used to estimate actual resolution in Figure 23a.

Conclusion

Although most of the wireless optimization techniques studied as part of the VAWN research program were focused on the delivery of 2D content, it's important to also understand how these same concepts could be applied to 3D eventually. As such, this research was intended to increase the fundamental understanding of the various factors that impact 3D content delivery, from content creation to compression to end user quality prediction. The studies mentioned in the previous sections lead us to several conclusions, which we would like to communicate to the industry community:

- 3D content creation requires more quality control than 2D creation does. The VQMT3D project revealed that most 3D films contain numerous impairments that can potentially cause eyestrain and headaches. Recent waining interest in 3D can be explained by it's unacceptable quality.

To avoid further decrease in interest, the industry needs to introduce strict quality standards to the 3D creation stage. Additionally, high quality 3D can be better compressed than video with a large amount of stereoscopic impairments.

- Autostereoscopic multiview devices are expected to gain popularity over displays that require users to wear glasses. Capturing content for the autostereoscopic displays with camera arrays is currently impractical. The content for such devices should be generated using depth maps, which can either be estimated from data captured by stereo cameras or captured by real-time depth sensors.
- 3D representations employing depth maps look promising due to their scalability (that is, support for various autostereoscopic displays) and low amount of additional data in comparison with 2D streams. Unfortunately, as autostereoscopic multiview devices gain popularity, due to its additional bitrate requirements the MVC standard will not suffice to deliver 3D videos to multiview displays over current modern wireless networks. Hence it is important to invest effort into developing 3D codecs specific to depth-based 3D video representations.
- It is important to study the quality of the entire end-to-end 3D system and not focus only on the rate distortion ratio of the codec. No one will watch a video with significant artifacts introduced at the creation stage despite the absence of distortions introduced during delivery and display stages. Any quality gain at the delivery stage can easily be negated at the display stage due to a low quality display. Therefore a quality standard for 3D displays should be created. Also an update to the conventional ITU recommended methodologies for subjective experiments is required. The software tools developed for measuring 2D video quality can no longer be used for 3D. Our open source software, Tally, is a good candidate solution to this problem both for industrial and scientific communities.

In this article, we presented several issues arising in the 3D-video life cycle (content creation, delivery, processing, and display) and proposed several concepts to mitigate these issues. A considerable amount of work still remains to be done. Results reported on depth-map–based compression schemes must be validated by conducting extensive subjective tests.

We would like to highlight the work that should be done to better understand how various factors influence 3D visual quality. In the content creation section, we have proposed several methods to detect artifacts that, according to medical experts' opinions, can potentially cause eyestrain. Measurement of the correlation between proposed metrics values and actual eyestrain is still an open problem. Actually, determining the value of the viewer's eyestrain is a challenge on its own, although some promising work exists in this field.[51] We hope that the research studies presented in this article will help to bring high-quality 3D video to every home while avoiding wireless network overload.

References

[1] Scharstein, D. and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2003, pp. I–195–I–202.

[2] Scharstein, D. and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002.

[3] Richardt, C., D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *European Conference on Computer Vision (ECCV)*, vol. 6313, 2010, pp. 510–523.

[4] Garcia, F., D. Aouada, B. Mirbach, T. Solignac, and B. Ottersten, "A new multi-lateral filter for real-time depth enhancement," in *Advanced Video and Signal-Based Surveillance*, 2011, pp. 42–47.

[5] Hirschmuller, H. and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[6] Kopf, J., M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, 2007.

[7] Garcia, F., B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *International Conference on Image Processing (ICIP)*, 2010, pp. 2805–2808.

[8] "Video Quality Measuring Tool 3D Project," http://www.compression.ru/video/vqmt3d.

[9] Rushton, S. K. and P. M. Riddell, "Developing visual systems and exposure to virtual reality and stereo displays: Some concerns and speculations about the demands on accommodation and vergence," *Applied Ergonomics*, vol. 30, no. 1, pp. 69–78, 1999.

[10] Hoffman, D. M., A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, pp. 1–30, 2008.

[11] Ukai, K. and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, no. 2, pp. 106–116, 2008.

[12] Howarth, P. A., "Potential hazards of viewing 3-D stereoscopic television, cinema and computer games: A review," *Ophthalmic and Physiological Optics*, vol. 31, no. 2, pp. 111–122, 2011.

[13] Tam, W. J., F. Speranza, S. Yano, K. Shimono, and H. Ono, "Stereoscopic 3D-TV: Visual comfort," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 335–346, 2011.

[14] Banks, M. S., J. C. A. Read, R. S. Allison, and J. W. Simmon, "Stereoscopy and the human visual system," *Motion Imaging Journal*, vol. 121, no. 4, pp. 24–43, 2012.

[15] Daum, R. M., "Clinical management of binocular vision: Heterophoric, accommodative and eye movement disorders," in *Optometry & Vision Science*, vol. 71, 1994, p. 414.

[16] Voronov, A., A. Borisov, and D. Vatolin, "System for automatic detection of distorted scenes in stereo video," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2012, pp. 138–143.

[17] Ogle, K. N., "Induced size effect: I. a new phenomenon in binocular space perception associated with the relative sizes of the images of the two eyes," *Archives of Ophthalmology*, vol. 20, no. 4, pp. 604–623, 1938.

[18] Gåding, J., J. Porrill, J. E. W. Mayhew, and F. J. P., "Stereopsis, vertical disparity and relief transformations," *Vision Research*, vol. 35, pp. 703–722, 1994.

[19] Allison, R. S., B. J. Rogers, and M. F. Bradshaw, "Geometric and induced effects in binocular stereopsis and motion parallax," *Vision Research*, vol. 43, no. 17, pp. 1879–1893, 2003.

[20] Stevenson, S. B. and C. M. Schor, "Human stereo matching is not restricted to epipolar lines," *Vision Research*, vol. 37, no. 19, pp. 2717–2723, 1997.

[21] Voronov, A., D. Vatolin, D. Sumin, V. Napadovsky, and A. Borisov, "Methodology for stereoscopic motion-picture quality assessment," in *Stereoscopic Displays and Applications*, vol. 8648, 2013.

[22] Anderson, B. L., "A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions," in *Perception*, vol. 26, 1997, pp. 419–454.

[23] Boydstun, A., J. Rogers, L. Tripp, and R. Patterson, "Stereoscopic depth perception survives significant interocular luminance differences," in *Journal of the Society for Information Display*, vol. 17, 2012, pp. 467–471.

[24] Patterson, R., A. Boydstun, J. Rogers, and L. Tripp, "Stereoscopic depth perception and interocular luminance differences," in *Digest of Technical Papers*, vol. 40, 2009, pp. 815–818.

[25] Voronov, A., D. Vatolin, D. Sumin, V. Napadovskiy, and A. Borisov, "Towards automatic stereo-video quality assessment and detection of color and sharpness mismatch," in *International Conference on 3D Imaging (IC3D)*, 2012, pp. 1–6.

[26] Stelmach, L., W. Tam, D. Meegan, A. Vincent, and P. Corriveau, "Human perception of mismatched stereoscopic 3D inputs," in *International Conference on Image Processing (ICIP)*, vol. 1, 2000, pp. 5–8.

[27] Seuntiens, P., L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," in *ACM Transactions on Applied Perception*, vol. 3, 2006, pp. 95–109.

[28] Kooi, F. and A. Toet, "Visual comfort of binocular and 3D displays," in *Displays*, vol. 25, 2004, pp. 99–108.

[29] Aleksei, B., B. Aleksandr, D. Vatolin, and M. Erofeev, "Automatic detection of artifacts in converted s3d video," in *Stereoscopic Displays and Applications*, 2014.

[30] Ivanov, B. T. and L. A. L., "Stereoscopic photography," *The Art Magazine*, 1959.

[31] Müller, K., P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proceedings of IEEE*, vol. 99, no. 4, pp. 643–656, 2011.

[32] "Dolby gets support of the foundry, Cameron—Pace group for glasses-free 3D," http://goo.gl/BgyOjL.

[33] Choi, J., D. Min, and K. Sohn, "2d-plus-depth based resolution and frame-rate up-conversion technique for depth video," *IEEE Transactions on Consumer Electronics*, vol. 56, pp. 2489–2497, 2010.

[34] De Silva, D. V. S. X., W. A. C. Fernando, and S. L. P. Yasakethu, "Object based coding of the depth maps for 3d video coding," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1699–1706, 2009.

[35] "YUVsoft depth upscale," http://goo.gl/mrqEFI.

[36] Simonyan, K., S. Grishin, and D. Vatolin, "Confidence measure for block-based motion vector field," in *Graphicon*, 2008, pp. 110–113.

[37] Hewage, C. T. E. R. and M. G. Martini, "Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3d video," in *ICIP,* IEEE, 2010, pp. 4017–4020.

[38] Kim, W.-S., A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, pp. 721–724.

[39] Lee, C., J.-I. Jung, and Y. .-S. Ho, "Inter-view depth pre-processing for 3d video coding," in *ISO/IEC JTC1/SC29/WG11, m22669*, 2011, pp. 1–7.

[40] "YUVsoft depth propagation," http://goo.gl/IDdcFe.

[41] Bal, C. and T. Q. Nguyen, "Depth-based prediction mode for 3D video coding," in *International Conference on Image Processing (ICIP)*, 2013.

[42] Bal, C. and T. Q. Nguyen, "Multiview video plus depth coding with depth-based prediction mode," *Circuits and Systems for Video Technology*, 2013.

[43] Fehn, C., "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93–104, 2004.

[44] Bjontegaard, G., "Calculation of average PSNR differences between RD-curves," *VCEG-M33*, Mar. 2001.

[45] Julesz, B., *Foundations of Cyclopean Perception* (Univ. Chicago Press, 1971).

[46] Jain, A. K., A. E. Robinson, and T. Q. Nguyen, "Comparing perceived quality and fatigue for two methods of mixed resolution stereoscopic coding," *Circuits and Systems for Video Technology*, 2013.

[47] Jain, A. K. and T. Q. Nguyen, "Video super-resolution for mixed resolution stereo," in *International Conference on Image Processing (ICIP)*, 2013.

[48] Jain, A. K., C. Bal, and T. Q. Nguyen, "Tally: A web-based subjective testing tool," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 128–129.

[49] ITU-R, "Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Tech. Rep., 2012.

[50] "The Lagom LCD monitor test pages," http://www.lagom.nl/lcd-test/.

[51] Chen, W., "Multidimensional characterization of quality of experience of stereoscopic 3d tv," PhD dissertation, Université de Nantes, 2012.

Author Biographies

Yury Gitman (ygitman@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2014. His main research interests are still image inpainting, visual attention modeling, edge aware filtering, and video matting. His recent work include semiautomatic visual attention models (http://compression.ru/video/savam) and the objective video matting benchmark (http://videomatting.com).

Can Bal (Cbal@ucsd.edu) is currently a PhD candidate in the Electrical and Computer Engineering Department at the University of California, San Diego. He received his BS and MS in Electrical and Electronics Engineering from Bilkent University, Turkey in 2007 and 2009 respectively. His research interests are in the field of 3D video compression and include depth-based 3D video coding, virtual view synthesis algorithms, and analysis of perceptual quality for 3D video.

Mikhail Erofeev (merofeev@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2013. Currently he is a PhD student in MSU's Graphics and Media Lab. His research interests are video and image matting, machine learning, 3D video generation and visual saliency modelling. Mikhail is one of the major contributors to the video matting methods benchmark (http://videomatting.com).

Ankit K. Jain (ankitkj@ucsd.edu) received a BS from Stanford University, Stanford, California, in 2005, and MS and PhD from the University of California, San Diego, in 2010 and 2014, respectively, all in Electrical Engineering. From 2005 to 2008, he was with the Embedded Digital Systems Group, MIT Lincoln Laboratory, Lexington, Massachusetts. He is currently a member of the technical staff at Pelican Imaging Corporation, Santa Clara, California. His research interests include image processing, 3-D video processing, computer vision, and human binocular vision.

Sergey Matyunin (smatyunin@graphics.cs.msu.ru) is a PhD student in the Graphics and Media Lab at Lomonosov Moscow State University. He received his specialist degree at MSU in 2011. The topics of his research are digital image processing, 3D video compression, and 2D-to-3D conversion.

Kyoung-Rok Lee (krl006@ucsd.edu) received a B.Eng. degree in Computer Engineering from the Kyungpook National University, Daegu, Korea, in 2008, an MS in Computer Science from the University of California, San Diego, in 2010, and a PhD in Electrical Engineering from the University of California, San Diego in 2014. His research interests are in video processing and computer vision including depth refinement, Simultaneous Localization And Mapping (SLAM), and 3D reconstruction.

Alexander Voronov (avoronov@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2011. As a member of the MSU Graphics and Media lab in 2008–2013, he participated in research on S3D-video generation and post-processing.

During 2012–2013 he was the head of the VQMT3D research project, which is dedicated to S3D video objective quality estimation and artifact analysis (http://compression.ru/video/vqmt3d/). Since 2013 Alexander has worked at Intel as a software engineer in the area of video compression.

Jason Juang (jajuang@ucsd.edu) is a PhD student in the Video Processing Lab in the Department of Electrical Computer Engineering, University of California, San Diego, led by Prof. Truong Nguyen. He received his MS at University of California, San Diego, and his BS at National Taiwan University. He is specializing in the field of computer vision and graphics, specifically in stereo vision. He did an internship at Qualcomm R&D and has worked previously at Tetravue.

Dmitriy Vatolin (dmitriy@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 1996 and got his PhD in 1999. The theme of his PhD thesis was "Optimization methods of fractal image compression." Dmitriy has been teaching a computer graphics course at MSU since 1997. He wrote the book *Algorithms of Image Compression* in 1999 and coauthored *Methods of Data Compression* in 2003. Dr. Vatolin supervised collaborative research projects with Intel and Samsung. Three PhD students have graduated under his supervision. He created one of the largest websites devoted to data compression (http://compression.ru). He teaches courses on methods of 3D and 2D video and image processing and compression.

Truong Q. Nguyen (tqn001@eng.ucsd.edu) [F'05] is currently a professor in the ECE Department of the University of California, San Diego. His current research interests are 3D video processing and communications and their efficient implementation. He is the coauthor (with Prof. Gilbert Strang) of a popular textbook, *Wavelets & Filter Banks,* Wellesley-Cambridge Press, 1997. He has over 400 publications. Prof. Nguyen received the IEEE Transaction in Signal Processing Paper Award (1992). He received the NSF CAREER Award in 1995. He served as associate editor for *IEEE Transaction on Signal Processing*, *Signal Processing Letters*, *IEEE Transaction on Circuits & Systems*, and *IEEE Transaction on Image Processing*.