Efficient Measurement of Stereoscopic 3D Video Content Issues

Stefan Winkler

Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore
and
Cheetah Technologies, Pittsburgh, PA, USA

ABSTRACT

The paper presents metrics to estimate a number of spatial and temporal parameters relevant for stereoscopic 3D video content. Based mainly on view differences and disparity, the aim of these metrics is to check for common issues with 3D content that might make viewers uncomfortable. The algorithms are designed for high computational efficiency to permit real-time video content analysis, and are shown to be robust to common video impairments.

Keywords: Depth perception, disparity, 3DTV, viewing comfort

1. INTRODUCTION

Measuring the quality of stereoscopic 3D video is gaining importance, with an increasing amount of content being produced and consumed in 3D. As the technology becomes more widely adopted and mature, quality issues rise to the forefront of concerns.

Quality issues for images and traditional 2D video have been studied quite extensively,[1] and commercial quality assurance (QA) tools are already being deployed to monitor video quality in real time. Most of these tools are designed to pick out common spatial and temporal distortions of the video resulting from compression and transmission.

Stereoscopy adds another layer of complexity on top of the common 2D impairments from video compression, network impairments, etc.[2] Furthermore, stereoscopic content may have potential physiological effects: if 3D is not produced, processed and presented correctly, it can make viewers dizzy or nauseous. This underlines that 3D viewing comes with more severe concerns than 2D. One of the primary practical goals must be to minimize or prevent possible discomfort caused by 3D content.

Many current 3D quality metrics choose a rather simplistic approach of extending 2D quality measurement to 3D by combining quality measurements done separately on left and right views; these are mainly targeted at the evaluation of asymmetric stereo coding. Only recently, more general methods for 3D quality assessment taking into account additional parameters have been proposed.[3,4] However, little consideration has been given to 3D video content issues and computational efficiency so far.

This paper presents metrics for a number of parameters relevant for stereoscopic 3D video content, based mainly on view differences and disparity. Note that video distortions such as those introduced by compression are not the main focus of these metrics; instead the aim is to enable checks of 3D content for issues that might make viewers uncomfortable. They would need to be combined with other 2D quality metrics in order to estimate the overall quality of a stereoscopic 3D presentation. The important feature of these metrics is high computational efficiency to permit real-time video content analysis.

E-mail: stefan.winkler@adsc.com.sg. Web: adsc.illinois.edu and www.cheetahtech.com.

2. DISPARITY ESTIMATION

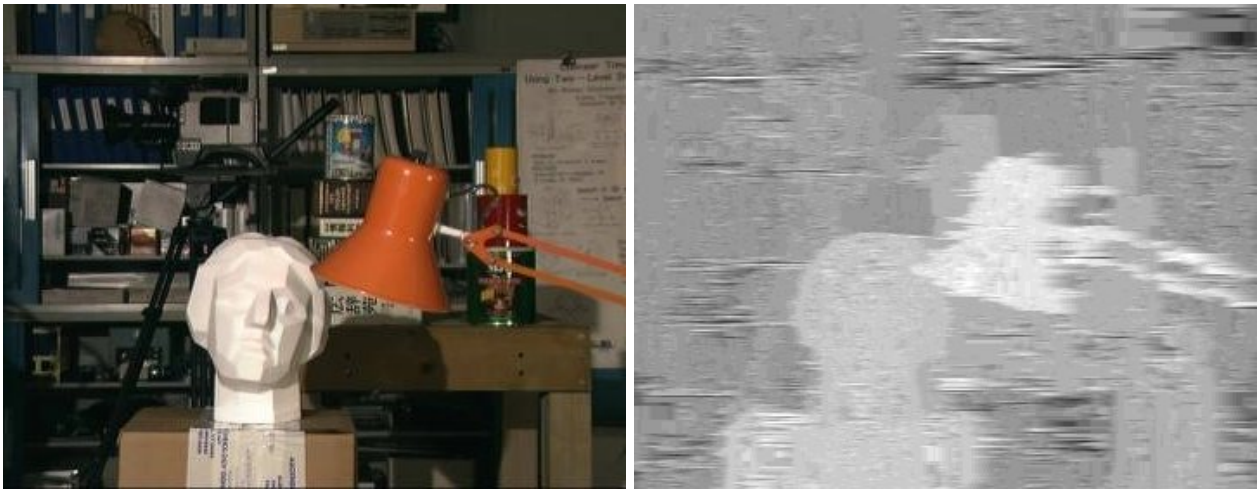
Disparity estimation is a prerequisite for computing depth metrics when the 3D content is represented in separate left and right views. The approach used here attempts to match individual scan lines of a frame to compute disparity at the pixel level. The specific method is based on work by Takaya[5] and relies on "dynamic time warping" (DTW), a well-known technique to find the optimal alignment between two given sequences.[6]

The sequences are warped in a nonlinear fashion to match each other. In the context of disparity estimation here, it is applied as follows: Based on the assumption that disparities should be primarily horizontal in a stereo image, the method processes the views scan-line by scan-line. DTW is used to compute the spatially varying shift between the scan-lines from the left and right views/images.* The result is an estimate of the disparity at each pixel, or in other words a disparity map. Downsampling (or alternatively median filtering) may be applied to the frames before processing in order to reduce noise as well as computation time.

The method has the following benefits:

- Performance: disparity estimation is the most computationally demanding task in stereo processing; with proper down-sampling, even high-definition (HD) content can be processed within a few milliseconds per frame on a standard PC, which is essential for real-time content monitoring.
- Resolution: potentially pixel-level precision for disparity (although this has to be traded off with performance and noise).
- Flexibility: the trade-off between resolution and performance can easily be fine-tuned as necessary.
- Robustness: disparity estimates are largely correct, without major outliers.

Figure 1 shows the disparity map computed using this algorithm for a sample image. While the resulting disparity maps are quite noisy, this is not a big concern for the content metrics proposed below, since we are mainly interested in the overall disparity range and distribution rather than the exact values at every pixel. As will be shown below, the method is sufficiently accurate for the measurements of interest and also robust to various image distortions.



(a) Original image (b) Disparity map

Figure 1. Tsukuba head-and-lamp scene.[7]

* A list of DTW implementations can be found at http://en.wikipedia.org/wiki/Dynamic_time_warping.

3. 3D CONTENT METRICS AND VALIDATION

We define several parameters and metrics to detect and quantify common spatial and temporal issues with 3D content that are listed in Table 1. An experimental validation of each metric is also presented.

Table 1. 3D content metrics.

View Mismatch	ΔLR
Disparity Range	$[D_{\min}; D_{\max}]$
Divergence	
Disparity Change	ΔD

3.1 View Mismatch

Unwanted mismatches between corresponding left and right views may arise at various stages of the production and distribution chain, if any of the following are not matched:[8,9] camera optics and sensors; white balance; shutter speed; aperture; gamma; geometry (camera angle and position; picture skew or cropping). Most of these mismatches can be corrected through careful calibration or during post-production.

Compression can also lead to view differences in terms of artifact severity (blockiness, blur) and time-varying quality (e.g. different coding parameters). Similarly, network impairments and error propagation may affect the views to different extents, especially when they are contained in separate streams. Views that are out-of-sync even by only 20-30 ms (e.g. in field-sequential displays) can cause depth errors.[10,11]

If any of these view differences become too severe, the HVS may be unable to fuse the two images into a consistent 3D percept and instead alternate between the two views. This is also known as binocular rivalry.

We use histogram correlation to express the magnitude of the mismatch between the two views. Histogram correlation is commonly used in video scene change detection,[12] and we found it to respond well to undue variations in color distribution; however, it would not be able to detect those subtle view differences which are amenable to post-production editing.

First, the luminance histograms of left and right views are computed, respectively (we use 256 bins). The linear (Pearson) correlation coefficient ρ_{LR} between these two histograms is used to quantify how well the views match.

To demonstrate the effectiveness of this quality factor, Figure 2 contrasts the histogram correlation ρ_{LR} of matching stereo pairs with those for unmatched views for a short image sequence. A clear distinction can be made.

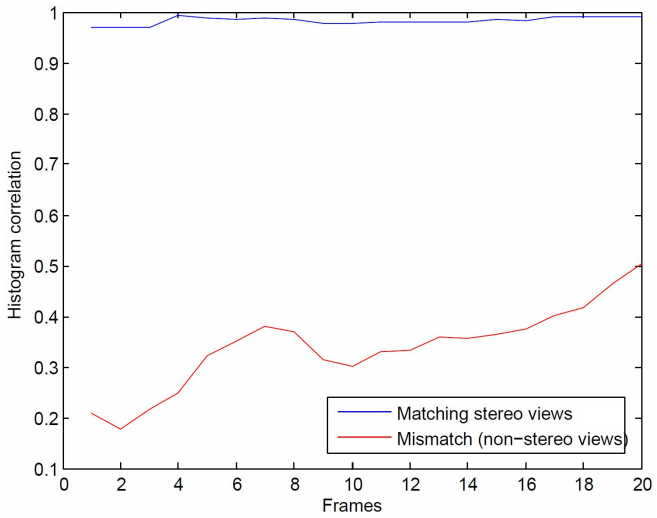


Figure 2. Histogram correlation ρ_{LR} of matching stereo pairs (blue) vs. unmatched views (red).

The final measure of the mismatch between the two views is defined as:

$$\Delta LR = 1 - \rho_{LR} \tag{1}$$

and can be expressed as a percentage.

3.2 Disparity Range

The conflict between accommodation and vergence is one of the main reasons for discomfort.[13,14] There is a certain range of both vergence and accommodation where the images from the two eyes can be comfortably fused into a single percept by the human visual system.[15] This range is also known as Panum's fusion area. Points outside of it result in double vision. The size of Panum's fusion area determines the range of depth in a scene (also called "depth bracket") that can be comfortably presented to a viewer. It depends on the angular disparity as well as the spatial and temporal properties of the content.[16] Ideally, stereoscopic presentations should be displayed such that they fall entirely within this range.[17,18]

There are many approaches to characterize the size of the comfortable viewing zone; its dependence on display size, viewing distance, and the amount of ambient light further complicates matters. As a rough guideline, the minimum/maximum disparity should be less than 2-3% of screen width, but this needs to be adjusted in accordance with the given viewing conditions. Tam et al.[19] provide an in-depth analysis of the various factors and limits.

Disparity Range measures the range of pixel disparities between the left and right view, using the method for disparity estimation described in Section 2. It is expressed as the range of disparities $[D_{\min}; D_{\max}]$ of a majority of pixels in a given frame; we use 90% here, but this can be adjusted to trade off robustness to noise with sensitivity.

For a detailed evaluation of the *Disparity Range* estimate, we use the New Tsukuba Stereo Dataset;[†] it contains ground truth disparity maps for 1800 frames from a simulated camera fly-through of a computer-generated office environment, featuring a wide variety of content and lighting conditions.[7] The original size of these frames is 640×480; they were downsampled 5 times to 128×120 for disparity estimation as described above (this is a typical size at which real-time processing is possible on an average CPU).

The results for the "fluorescent" set are shown in Figure 3 (because the stereo camera setup is perfectly parallel, all disparities are positive in this dataset, which also means that most minimum disparities are very small). Error statistics for different percentile levels are given in Table 2. The data show that the disparity range metric is generally accurate, with errors in an acceptable range.

Table 2. Disparity error statistics (estimate vs. ground truth in pixels).

Percentile	RMSE	μ	σ	ratio	ρ
5%	14.14	–9.73	10.26	N/A	N/A
50%	4.39	0.31	4.38	1.2%	89%
95%	12.08	–0.75	12.06	2.1%	90%

Since the New Tsukuba Stereo Dataset contains no noise or other types of image distortions, we further evaluate the robustness of the Disparity Range estimate using a sequence created from the IRCCyN/IVC 3D Images Database.[‡] It contains six different stereoscopic images, each of which is present in original undistorted form and in 15 distorted versions.[20] Distortions include three different types of processing (JPEG and JPEG2000 compression as well as blurring), which were applied symmetrically to the stereo pairs. The images are concatenated into a sequence of 96 frames.

Figure 4(a) shows the *Disparity Range* measurements on this image sequence. It demonstrates that the *Disparity Range* metric is largely unaffected by image distortions such as compression and blur, some of which reach rather severe levels in this database.

[†] http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php

[‡] http://www.irccyn.ec-nantes.fr/spip.php?article876

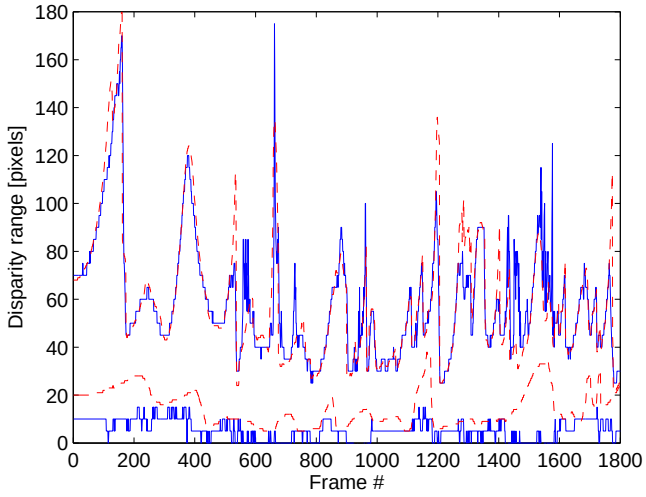
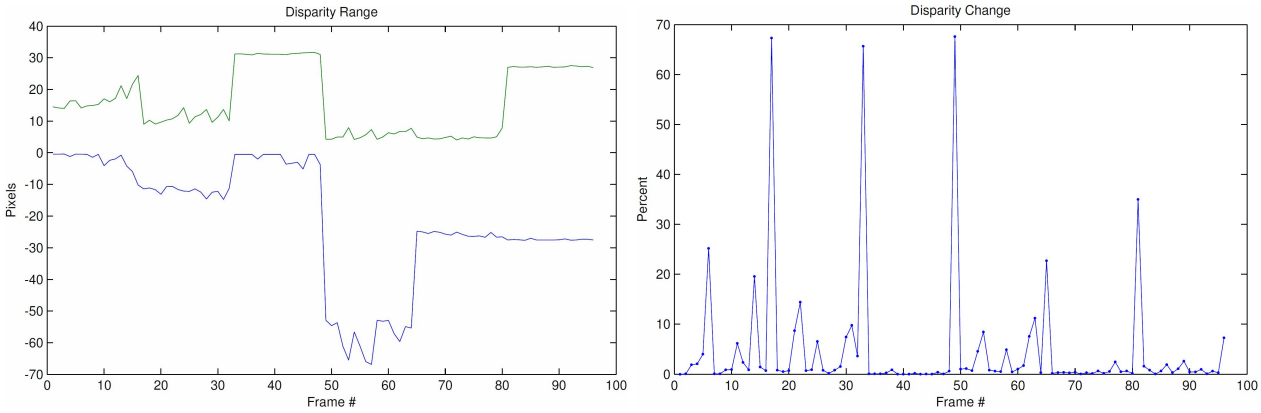


Figure 3. Disparity range for New Tsukuba Stereo Dataset sequence (dashed/red: ground truth, solid/blue: estimate).



(a) Disparity range (bottom and top lines indicate $D_{\min}$ and $D_{\max}$, respectively)

(b) Disparity change ΔD

Figure 4. Disparity range (a) and change (b) for the IVC 3D image sequence (the picture changes every 16 frames).

3.3 Divergence

A disparity greater than the inter-ocular distance would force the eyes to diverge and place the object beyond infinity, which is impossible in nature and should be avoided. Therefore, the maximum positive disparity on screen should not exceed the interocular distance of the viewer; in other words, positive (divergent) disparity should not be more than 5-6 cm.

Naturally, in the image domain, this is screen- and resolution-dependent (note that it does *not* depend on viewing distance). For high-definition (HD) video displayed on a 42" screen for example, this corresponds to roughly 5% of the width of the video frame, which is about 100 pixels.

3.4 Disparity Change

Temporal depth discontinuities occur when the depth or depth distribution of a scene changes. Rapid depth variations can result in viewer discomfort, because the human visual system (HVS) is unable to follow the changes and to reconstruct depth properly.[17,21] This is a common problem at transitions (e.g. scene cuts). Rapid depth variations can be even more detrimental to viewing comfort than a large depth bracket.[18] In general, depth changes should happen slower and less frequently than in 2D.

Disparity Change measures the temporal change of disparity distributions between two consecutive frames. As mentioned earlier in Section 3.1, histogram correlation is commonly used for detecting scene changes in video. Therefore, we adapt it here again for use with the disparity maps.

First, a 256-bin histogram of the disparity map is computed for every video frame (cf. Section 2). The linear (Pearson) correlation coefficient ρ_D between the disparity histograms of the current frame and the previous frame is used to quantify the change in disparity. The final measure of the disparity change between two frames is defined as:

$$\Delta D = 1 - \rho_D \tag{2}$$

and can be expressed as a percentage.

To evaluate the *Disparity Change* estimate ΔD, we again use the sequence of 96 images created from the IVC 3D Images Database. Figure 4(b) shows the measurements of disparity change. As one would expect, the changes in disparity are greatest at scene cuts in this particular sequence (occurring at multiples of 16 frames).

A test on a longer and more realistic compressed video sequence (about 2 minutes of a soccer match) shows the reliability of the metric (Figure 5), with scene cuts being clearly identified and rated according to their severity in terms of disparity change.

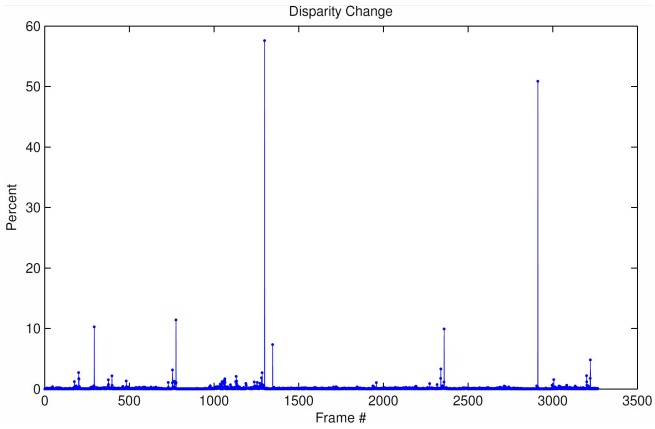


Figure 5. Disparity change ΔD for the soccer video.

4. CONCLUSIONS

We presented metrics that can detect and quantify some common spatial and temporal issues with 3D content, namely view mismatch, divergence, disparity range, and disparity change, all of which may introduce discomfort in viewers if they become too large. The metrics are robust to compression and various other types of image distortions.

Computational efficiency was a prime consideration in the design of the measurement algorithms; as a result, they are suitable for real-time 3D video monitoring applications. Even for HD content, all metrics can be computed within a few milliseconds per frame. The bulk of the computation is due to disparity estimation from the stereo images, which is performed using an efficient method based on dynamic time warping.

The disparity estimation is designed for 3D content stored as separate left and right views; however, the metrics can also be applied to 2D+depth representations, where the disparity information is readily available.

While we have demonstrated the validity of the metrics in terms of estimating various 3D content parameters, their perceptually acceptable ranges still need to be verified. The 3D content metrics could be easily integrated and combined with other quality metrics (e.g. 2D quality assessment of left and right views) in order to measure viewing comfort or 3D quality of experience (QoE).

ACKNOWLEDGMENTS

The author is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

REFERENCES

[1] Winkler, S. and Mohandas, P., "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Transactions on Broadcasting* **54**, 660–668 (Sept. 2008).

[2] Winkler, S. and Min, D., "Stereo/multiview picture quality: Overview and recent advances," *Signal Processing: Image Communication* **28**, 1358–1373 (Nov. 2013).

[3] Lambooij, M., IJsselsteijn, W., Bouwhuis, D. G., and Heynderickx, I., "Evaluation of stereoscopic images: Beyond 2D quality," *IEEE Transactions on Broadcasting* **57**, 432–444 (June 2011).

[4] Hewage, C. T. E. R. and Martini, M. G., "Quality of experience of 3D video streaming," *IEEE Communications Magazine* **51**, 101–107 (May 2013).

[5] Takaya, K., "Dense stereo disparity maps – real-time video implementation by the sparse feature sampling," in [*Proc. Conference on Machine Vision Applications*], (June 13–15, 2011).

[6] Müller, M., "Dynamic time warping," in [*Information Retrieval for Music and Motion*], ch. 4, Springer (2007).

[7] Martull, S., Peris, M., and Fukui, K., "Realistic CG stereo image dataset with ground truth disparity maps," in [*Proc. ICPR TrakMark Workshop*], (Nov. 2012).

[8] Woods, A. J., Docherty, T., and Koch, R., "Image distortions in stereoscopic video systems," in [*Proc. SPIE Stereoscopic Displays and Applications*], **1915** (Feb. 1993).

[9] Goldmann, L., De Simone, F., and Ebrahimi, T., "Impact of acquisition distortion on the quality of stereoscopic images," in [*Proc. International Workshop on Video Processing and Quality Metrics (VPQM)*], (Jan. 13-15 2010).

[10] Burr, D. C. and Ross, J., "How does binocular delay give information about depth?," *Vision Research* **19**, 523–532 (1980).

[11] Goldmann, L., Lee, J.-S., and Ebrahimi, T., "Temporal synchronization in stereoscopic video: Influence on quality of experience and automatic asynchrony detection," in [*Proc. International Conference on Image Processing (ICIP)*], (Sept. 26-29 2010).

[12] Pindoria, M., "Scene segmentation using multiple metrics," Tech. Rep. WHP 210, BBC (March 2012).

[13] Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S., "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision* **8**, 33 1–30 (March 2008).

[14] Ukai, K. and Howarth, P., "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays* **29**, 106–116 (Feb. 2008).

[15] Semmlow, J. L. and Heerema, D., "The role of accommodative convergence at the limits of fusional vergence," *Investigative Ophthalmology & Visual Science* **18**, 970–976 (Sept. 1979).

[16] Schor, C. and Tyler, C., "Spatio-temporal properties of Panum's fusional area," *Vision Research* **21**(5), 683–692 (1981).

[17] Yano, S., Ide, S., Mitsuhashi, T., and Thwaites, H., "A study of visual fatigue and visual comfort for 3D HDTV/HDTV images," *Displays* **23**, 191–201 (April 2002).

[18] Yano, S., Emoto, M., and Mitsuhashi, T., "Two factors in visual fatigue caused by stereoscopic HDTV images," *Displays* **25**, 141–150 (Sept. 2004).

[19] Tam, W. J., Speranza, F., Yano, S., Shimono, K., and Ono, H., "Stereoscopic 3D-TV: Visual comfort," *IEEE Transactions on Broadcasting* **57**, 335–346 (June 2011).

[20] Benoit, A., Le Callet, P., Campisi, P., and Cousseau, R., "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing* (2008).

[21] Nojiri, Y., Yamanoue, H., Ide, S., Yano, S., and Okano, F., "Parallax distribution and visual comfort on stereoscopic HDTV," in [*Proc. International Broadcasting Convention (IBC)*], (2006).