Depth map quality metric for three-dimensional video

Donghyun Kim[a], Dongbo Min[a], Juhyun Oh[ab], Seonggyu Jeon[b], Kwanghoon Sohn[a]
[a]Dept. of Electrical and Electronics Eng., Yonsei University, Seoul, Korea
[b]Korean Broadcasting System, Seoul, Korea

ABSTRACT

In this paper, we propose a depth map quality metric for three-dimensional videos which include stereoscopic videos and autostereoscopic videos. Recently, a number of researches have been done to figure out the relationship of perceptual quality and video impairment caused by various compression methods. However, we consider non-compression issues which are induced during acquisition and displaying. For instance, using multiple cameras structure may cause impairment such as misalignment. We demonstrate that the depth map can be a useful tool to find out the implied impairments. The proposed quality metrics using depth map are depth range, vertical misalignment, temporal consistency. The depth map is acquired by solving corresponding problems from stereoscopic video, widely known as disparity estimation. After disparity estimation, the proposed metrics are calculated and integrated into one value which indicates estimated visual fatigue based on the results of subjective assessment. We measure the correlation between objective quality metrics and subjective quality results to validate our metrics.

Keywords: Three dimensional video, quality assessment, depth map

1. INTRODUCTION

A number of researches have been developed recently to measure the perceptual quality of image and video. We are expecting more extensive researches on 3D video quality assessment with the increasing demand on 3D video. To standardize the method of subjective quality assessment, ITU has recommendations for monoscopic television and stereoscopic television pictures [1][2]. Although subjective assessment is the most reliable method to gather the quality information from video, it is expensive and time consuming. Many objective quality assessment metrics are proposed to measure the perceptual quality which can substitute the results of subjective assessment. Recently, many researches have been done to figure out the relationship of perceptual 3D video quality and impairment posed by compression. For instance, quality of various compression method including JPEG and H.264 are investigated [3], and the effect of asymmetric coding of stereoscopic videos is studied [4].

Nowadays, novel approaches are focused which utilize depth information instead of directly dealing with stereoscopic video or multi viewpoint video. PHILIPS developed an autostereoscopic display device which has 2D plus depth input format. In addition, Ad Hoc Group on Free Viewpoint Television in MPEG concentrated on Free Viewpoint Video (FVV) and 3DTV systems, including representation, generation, processing, coding and rendering of MVD format data. [5]. Moreover, several quality assessment issues in depth map coding for 2D plus depth format are investigated. They give comparative studies on different techniques for depth-image compression and measure the quality of rendered image [5][6][7].

Three dimensional television (3DTV), which could provide sense of presence, however it has severe problem of causing visual fatigue by various reasons besides the compression impairment. For instance, video acquisition using multiple cameras may yield impairment such as misalignment, color imbalance and asynchronization. In addition, various 3D display devices and viewing conditions should be considered for 3D video quality assessment. There was a study to explore the stereoscopic human visual characteristics by subjective evaluation with manipulating camera parameters as camera separation, convergence and focal length to measure perceived quality and naturalness [8]. 3D consortium established safety guideline for 3D videos to create comfortable 3D contents and prevent visual fatigue by recommending the disparity threshold for comfortable viewing [9]. Furthermore, parallax adjustment algorithm was proposed to enhancement of viewer comfort by studying the relationship between visual comfort and parallax on stereoscopic HDTV [10].

khsohn@yonsei.ac.kr; phone 82 2 2123-2879; diml.yonsei.ac.kr

Stereoscopic Displays and Applications XX, edited by Andrew J. Woods, Nicolas S. Holliman,
John O. Merritt, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 7237, 723719
© 2009 SPIE-IS&T · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.806898

In this paper, we propose a depth map quality metric for three-dimensional videos which include stereoscopic videos and autostereoscopic videos. We demonstrate that the depth map can be a useful tool to find out the implied impairments. Depth map utilized through this paper, is acquired by solving corresponding problems from stereoscopic video. The proposed quality metrics using depth map are depth range, vertical misalignment, temporal consistency. Depth (disparity) range is the most important metric, because they directly affect to perceived depth in 3D display with several acquisition parameters (baseline, focal length shooting distance, resolution) and viewing condition parameters (viewing distance, resolution, display size). We use Scale-Invariant Feature Transform (SIFT) to extract initial feature disparity value, then we perform dense disparity estimation using initial disparity as a search range. Excessive horizontal and vertical disparity is a well-known problem that causes visual fatigue. Temporal consistency is also considered, which measures whether depth map has a stable value or fluctuating value through consecutive frames. For the applications which utilize dense depth map or 2D plus depth video format, temporal consistency is important quality factor. These proposed metrics are calculated and integrated into one value which indicates estimated visual fatigue based on the results of subjective assessment. Finally, we measure the correlation between objective quality metrics and subjective quality results to validate our metrics.

2. QUALITY METRIC USING DEPTH MAP

In general, compressed video is hard to recover its original quality because the high frequency component of video was discarded to reduce bit rate. However, when considering non-compression related impairment of 3D video, for instance extreme parallax by excessive camera baseline, can be modified by scaling their parallax to appropriate for specific viewing conditions. Fig. 1 shows the process of quality enhancement of 3D video by using 3D quality factor database. 3D quality factor database is collected from subjective assessment of various low quality video generated by various compressor & camera structure. When building the quality database, we should include the test video from various video acquisition methods, display devices and viewing conditions. Also, variance between the assessor's quality results should be collected as well as the average of assessment results because safety of 3D videos is the most important factor. After building the 3D quality factor database, every 3D video should examine objective quality before ill-quality videos are exposed to viewers and can be modified in quality enhancement stage.

We regard depth map as a useful tool to find out the implied impairments. Therefore, we propose a depth map quality metric for three-dimensional videos which include stereoscopic videos and autostereoscopic videos. Depth map of stereoscopic videos represents the information of horizontal and vertical disparity distribution which affect to visual fatigue. In addition, the quality of autostereoscopic video is directly related to depth map because color view and depth are exploited in multiview rendering process. In the remainder of this chapter, we will demonstrate the depth map acquisition method and explain the detail of the proposed quality metric using depth map which are depth range, vertical misalignment and temporal consistency.

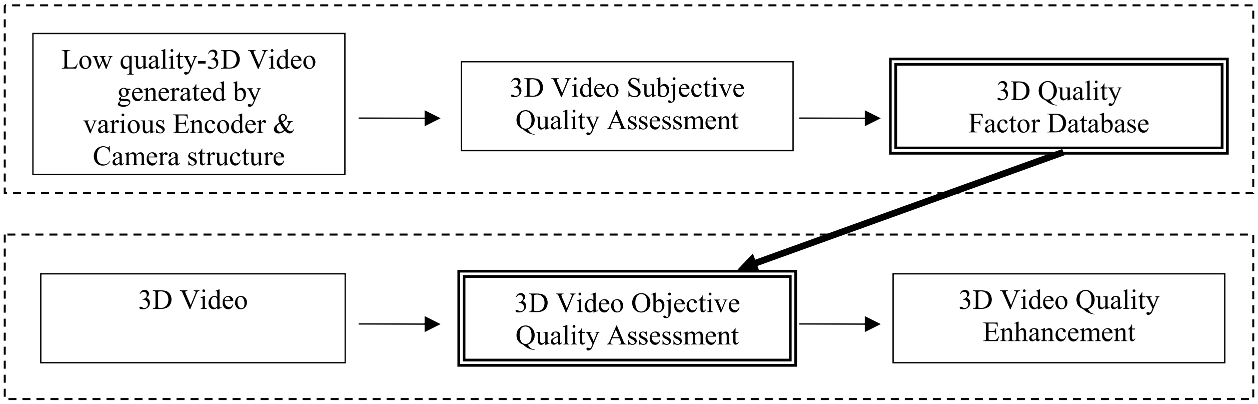

Low quality-3D Video generated by various Encoder & Camera structure
3D Video Subjective Quality Assessment
3D Quality Factor Database
3D Video
3D Video Objective Quality Assessment
3D Video Quality Enhancement

Fig. 1. 3D video quality assessment and 3D video quality enhancement

2.1 Depth map acquisition

There are several methods to obtain depth map. One of the methods is using depth camera such as range sensor, structured light, TOF (Time-Of-Flight) range camera, LIDAR (Light Detection and Ranging) and so on. Structured light projects a known pattern of pixels on to objects, while TOF range camera emits modulated light and measures the reflectance of the emitted light at objects and LIDAR measures properties of scattered light to find range of objects. Although each sensor has its own strength in specific environment, they have restrictions of the sensing range, color of objects and lighting conditions.

The other method to obtain depth map is stereo matching algorithm from stereoscopic video. Depth map, which is utilized through this paper, is acquired by solving corresponding problems from stereoscopic video. We use Scale-Invariant Feature Transform (SIFT) to extract initial feature disparity value, then we perform dense disparity estimation using initial disparity as a search range. Feature based disparity estimation is required before the dense disparity estimation because conventional disparity estimation algorithm assume stereoscopic video was acquired with parallel camera structure which implies the video have only negative disparity values. However, most of the stereoscopic video has both positive and negative disparity values which form the actual 3D objects into front direction and backward direction of 3D monitor.

Fig. 2 shows the proposed quality metrics using depth map through the format conversion for various 3D display devices. We have 3 quality metrics using depth map which are vertical misalignment, depth range and temporal consistency. We will demonstrate the details of each metrics later. The input is stereoscopic video which can be converted into disparity shifted stereoscopic video, one color view plus one depth view, and multi viewpoint video. At first, high frequency components of images are selected as feature points, then they are tracked across the left and the right images. Then, dense disparity estimation is performed with belief propagation algorithm. Stereo matching algorithm using belief propagation was proposed by Sun et al that formulate the stereo matching problem as a Markov network and solved it using Baysian belief propagation [11]. Although it has a weak point of sensitivity about noise or similar pixel values, it is known as excellent global method among various disparity estimation algorithms. Finally, 1 view plus 1 depth format is obtained by combining one view of stereoscopic video and corresponding dense depth map. Multiview video is obtained by employing intermediate view rendering algorithm by using stereoscopic videos and estimated depth map video.

Fig. 3 shows image shift for disparity estimation and disparity adjustment for comfort viewing. We use image shift stage in twofold in our algorithm. The first is to use image shift as an initial process for dense disparity estimation. Images which contain both positive and negative disparity values are shifted to have only negative disparity values. The second is to adjust disparity values only using image shift method. Stereoscopic video with excessive disparity values are modified by shifting the images inner or outer direction.

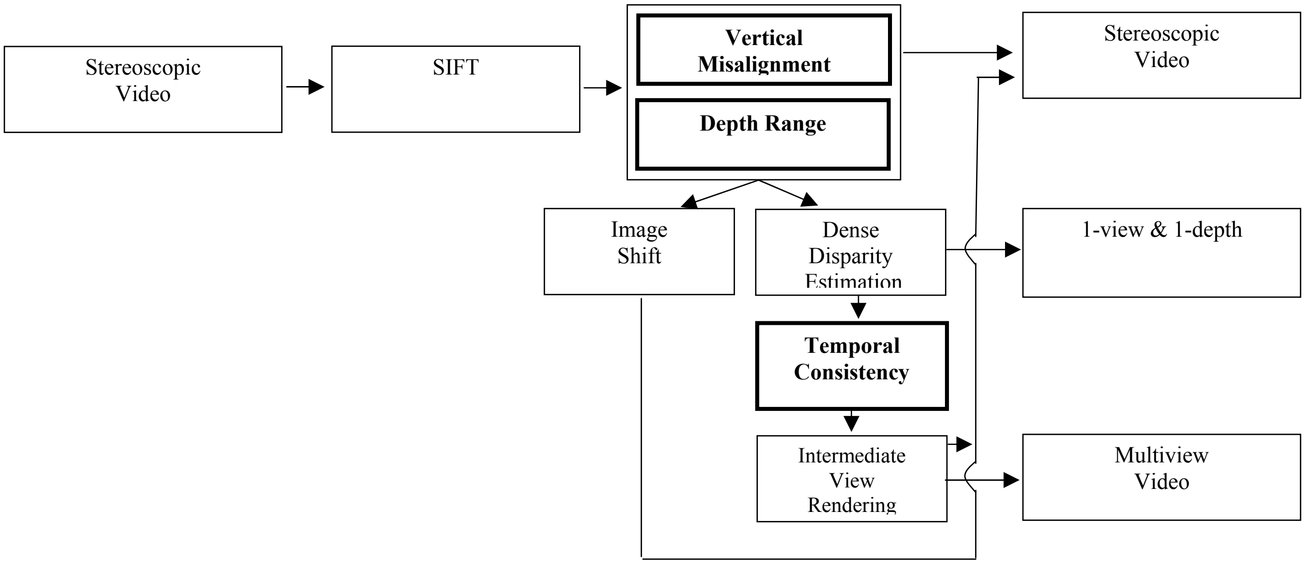


Fig. 2. Proposed quality metric using depth map and format conversion for various 3D displays

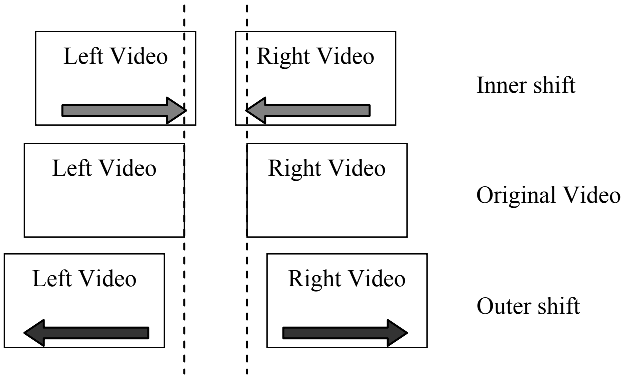


Fig. 3. Image shift for disparity estimation and disparity adjustment

2.2 Quality metric using depth map

Depth (disparity) range is the most important metric, because they directly affect to perceived depth in 3D display with several acquisition parameters (baseline, focal length shooting distance, resolution) and viewing condition parameters (viewing distance, resolution, display size). Many researches have been conducted to figure out disparity fusional limit of human stereo visual system. 3D Consortium [9] reported fusional limit to ± 2 degrees and recommended to maintain disparity less then ± 60 minutes. Also, they recommended that it is desirable to minimize the time using large disparity close to a fusional limit. However, Yano et al. [12] recommended less fusional limit to ± 0.2D (Diopters) which can be converted to ± 0.82degrees. This inconsistency of the results occurred due to the individual differences in age, viewing time and eye distances. Moreover, visual fatigue can occur even stereoscopic fusion is possible. Therefore, overall visual fatigue is collected from assessors in several discrete scales rather than measuring only fusional limit. Equation (1) shows how to measure disparity range of one stereoscopic image frame. It represents the difference the upper average and lower average of ordered horizontal disparity of each feature. The maximum and minimum feature values are excluded to decrease the sensitivity of error occurred from disparity estimation.

$$Disparity\ range = \sum_{f=fn*f_1}^{fn*f_2}(D_h(f) - D_h(fn-f)) \tag{1}$$

where $D_h(f)$ represents ordered horizontal disparity, fn , f_1 , f_2 represent the number of features, upper and lower range in percentages, respectively.

Vertical misalignment is a well-known problem that causes visual fatigue. It is calculated from SIFT results which are previously used in depth range metric. Vertical misalignment can be caused by convergence stereoscopic camera structure or caused by slightly tilted stereo camera pair. Panum's fusional area defines the limit of horizontal disparity and vertical disparity in retinal domain [13]. It describes area, within which different points projected on to the left and right retinas, producing binocular fusion and sensation of depth. Hence, vertical misalignment should be limited within Panum's fusional area as horizontal disparity. Otherwise, excessive disparity could cause double vision or severe visual fatigue. Equation (2) shows how to measure vertical disparity of one image frame. It adopts same structure of equation (1) and represents the difference the upper average and lower average of ordered vertical disparity of each feature.

$$Vertical\ disparity = \sum_{f=fn*f_1}^{fn*f_2}(D_v(f) - D_v(fn-f)) \tag{2}$$

where $D_v(f)$ represents ordered vertical disparity, fn , f_1 , f_2 represent the number of features, upper and lower range in percentages, respectively.

If vertical misalignment is detected in stereoscopic videos, one can simply translate one of the stereoscopic videos and crop them into same resolution. Otherwise, image rectification can be applied to a pair of stereoscopic videos. Image rectification transforms each stereoscopic video so that the pairs of conjugate epipolar lines become parallel to one of the video axes.

Temporal consistency measures whether depth map has a stable value or fluctuating value through continuous frames for each pixel. The video captured from stereoscopic camera do not have a depth map fluctuation problem. However, applications which utilize dense depth map or 2D plus depth video format, temporal consistency should be considered. Because most of disparity estimation algorithms are conducted frame by frame, the result may include fluctuating depth values in non-motion area. One of the methods to increase temporal consistency is to consider both neighbor frames and previously estimated neighbor disparity map in the process of disparity estimation. Because it increases the complexity of disparity estimation, we proposed conditional depth filtering as a post processing. Equation (3) shows simple conditional filtering method, and equation (4) represents to measure the depth fluctuation condition map. It is consisted by two parts, which are color transition and depth fluctuation. Basic assumption to find the region of depth fluctuation is to search the area with less color variance and larger depth variance.

$$Out = (1 - condition) \times original + conditon \times filtered \tag{3}$$

$$condition(i,j) = color\ transition \times depth\ fluctuation \tag{4}$$

$$= e^{-\frac{1}{\sigma_1^2}\sum_{j\in f}\sum_{i\in R^2}(c(i,j)-c(i,j+1))^2} \times (1 - e^{-\frac{1}{\sigma_2^2}\sum_{j\in f}(d(i,j)-d(i,j+1))^2})$$

Where, f and R^2 represent the number of neighbor frames and kernel for $c(i,j)$ and $d(i,j)$ which indicate i-th pixel in j-th frame of color and depth image. σ_1, σ_2 represent the weighting factors for sum of squared difference of color and depth image, respectively.

We notice that depth error tends to appear one of the frames due to independent depth estimation process between neighboring frames. That is the reason that color transition is calculated across the frames and kernel of current frame, and depth fluctuation is calculated only across the neighboring frames.

3. EXPERIMENTAL RESULTS

3.1 Subjective assessment to measure visual fatigue

The participants were non-experts but who were familiar with 3D display device and they were screened for color vision, visual acuity and stereo acuity. Our test platforms are 24 inch polarized stereoscopic display device which offers resolution of 960 × 1280 in stereoscopic mode and 20 inch autostereoscopic display device which offers 9 viewpoint and 1600 ×1200 resolution.

20 assessors participated in subjective assessment with ages from 25 to 35. We decided to select 3D display familiar assessors who can exactly describe their depth perception and visual fatigue. They watched randomly-ordered stereoscopic video clips for ten seconds twice at the distance of 3H and assigned a grade from 1 to 10 about visual fatigue of 3D video. We followed the recommendations for subjective assessment of stereoscopic television pictures from ITU [2]. Since there is no reference video to be compared with, the assessment method could be classified into single-stimulus scaling,

We conducted subjective assessment with computer graphic video shown in Fig. 4. By using computer graphic video, we could easily control the baseline of cameras and alignment of cameras. Objects are placed on 5 by 5 meters space as shown in right side of Fig. 4. The robot which has a height of 1.5 meter, was filmed with stereoscopic camera from the distance of 2 meter. Four camera baselines which were 4cm, 8 cm, 12 cm, 16 cm, and three vertical misalignments which were tilted 2.5 cm, 5.0 cm, 7.5 cm, were used through the test. We used smaller alternations in vertical misalignment than camera baseline because human visual systems are more sensitive to vertical misalignment. Fig. 5 shows the results of subjective assessment of visual fatigue levels while varying baseline and vertical misalignment. The result shows linear decrease with the increase of camera baseline rather than the increase of vertical misalignment. We noticed that tolerance to vertical misalignment was increased, when camera baseline became shorter.

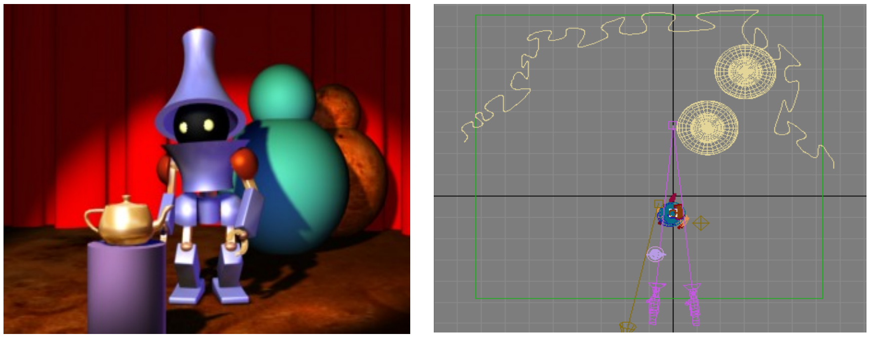


Fig. 4. Camera baseline controlled computer graphic (Robot)

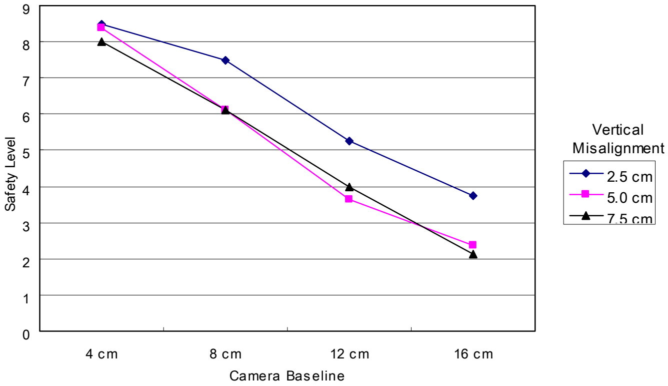


Fig. 5. Results of subjective assessment of visual fatigue while varying baseline and vertical misalignment

3.2 Results of depth map quality metric

In order to evaluate the proposed algorithm, several videos were used. We used stereoscopic image sequences called ‘Dance’, ‘Concert’, ‘Moly’, ‘Temple’, ‘Family’, ‘Building’ and ‘Breakdancers’ as shown in Fig. 6. Videos were captured by Stereopia, KBS, 3D Korea and Microsoft, and they were resized to the resolution of 3D monitor.

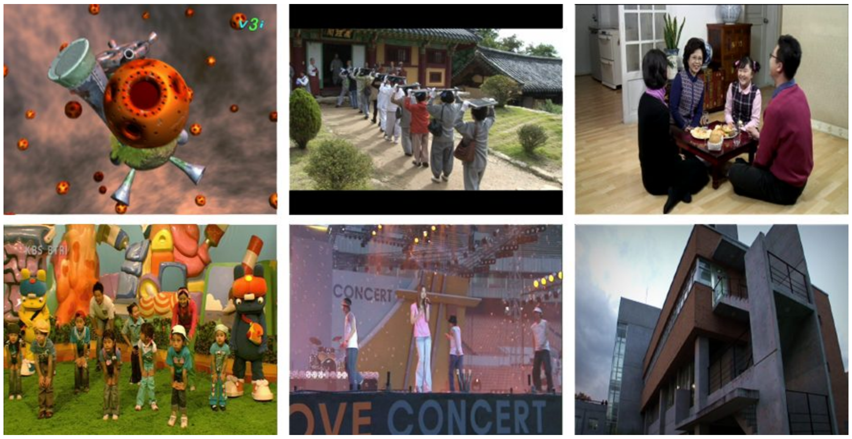
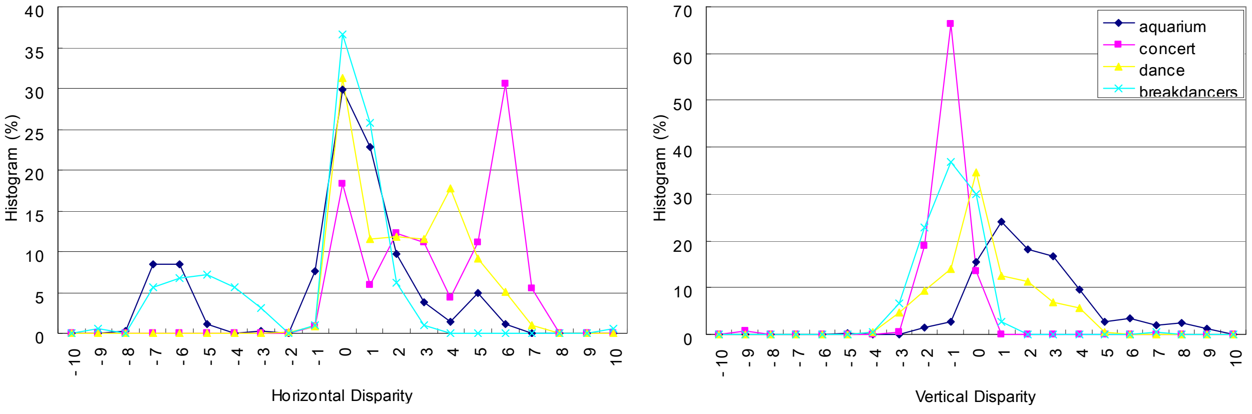


Fig. 6. Test video sets (Moly, Temple, Family, Dance, Concert, Buliding).



(a) (b)

Fig. 7. Results of depth map quality metric for test videos

Fig. 7 shows the histograms of horizontal disparity and vertical disparity calculated by equation (1), (2) on ‘Aquarium’, ‘Dance’, ‘Concert’ and ‘Breakdancers’ videos. In Fig. 7 (a), less histogram distribution which exceeds the fusional limit discussed in previous chapter indicates less visual fatigue, while concentrated histogram to center region implies less visual fatigue in Fig. 7 (b).

Multiple metrics combining scheme is required, because we applied multiple quality metrics. Since horizontal disparity and vertical disparity showed independency in the assessment results shown in Fig. 5, we used additive first order linear regression of two metrics (horizontal disparity and vertical disparity). At this time, temporal consistency metric was excluded because of the difficulty in acquiring three items-varying test videos. Linear regression coefficients were calculated through the results of computer graphic ‘Robot’ sequence using linear regression which maximizes the correlation between objective and subjective quality. Then, we measure the correlation between objective quality metrics and subjective quality results to validate our metrics using test video sets shown in Fig. 6. Fig. 8 shows the graph with subjective quality in x-axis and proposed metrics in y-axis. Each point in Fig. 8 indicates one of the test videos. The correlation was 71% between the proposed quality metrics and subjective quality assessment results. We concluded that the proposed metric predicted the overall quality of test videos, but rather lower correlation result was due to the absence of reference video to compare with.

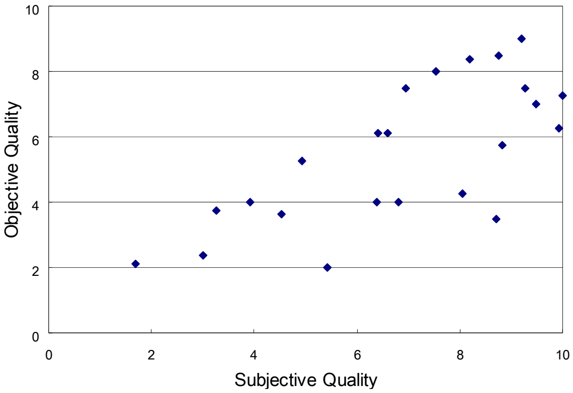


Fig. 8. Subjective and objective quality of test videos

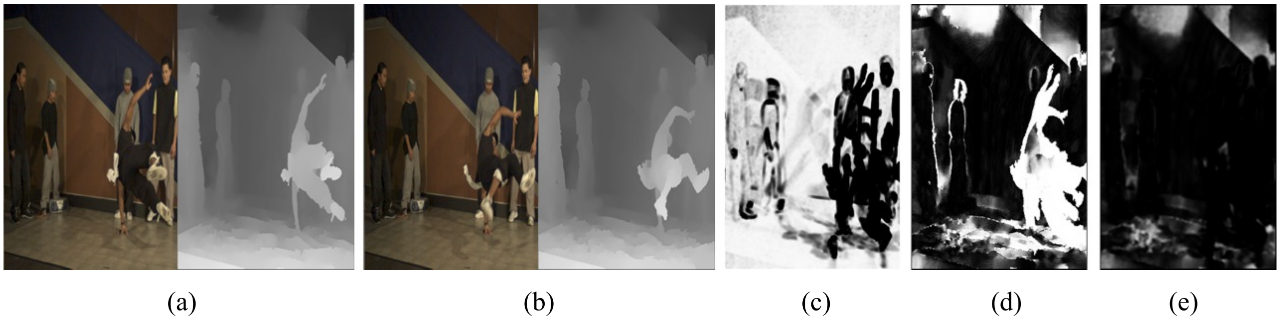


Fig. 9. Results of temporal consistency metric (fluctuation detection) (a) frame #17 (b) frame #18 (c) motion variance map (d) depth variance map (e) depth map fluctuated area

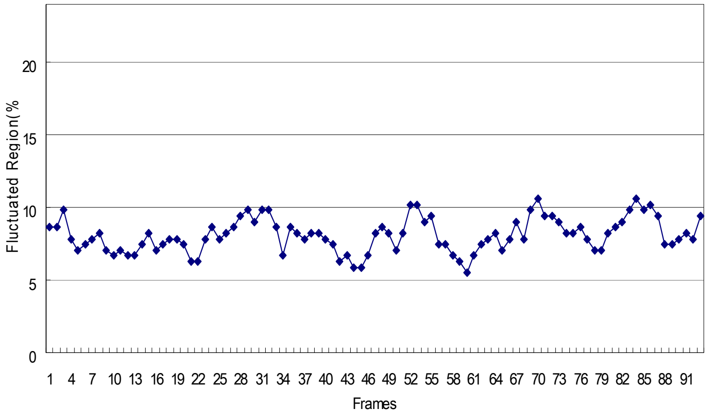


Fig. 10. Results of temporal consistency of 'breakdancers' sequence

Fig. 9 shows the process of detecting temporal consistency between 17th frame and 18th frame. Fluctuation in depth map is observed in the upper wall in Fig. 9 (a) and (b) even they are consecutive frames. Fig. 9 (c) is motion variance map which detects less color variance area, while Fig. 9 (d) is depth variance map which detects larger depth variance area. Fig. 9 (e) is estimated fluctuated area in depth map by equation (4). Fig. 10 shows the results of temporal consistency metric. Conditional median filtered result estimated by equation (3) was displayed on autostreoscopic display, and showed increased stability in fluctuating area while maintaining details in remaining area.

4. CONCLUSION

In this paper, we propose a depth map quality metric for three-dimensional videos. The proposed quality metrics using depth map are depth range, vertical misalignment and temporal consistency. To obtain depth map, feature based disparity vectors is estimated by using Scale invariant feature transform. After disparity estimation, proposed metrics are calculated and integrated into one value which indicated visual fatigue. Coefficients used in integration are calculated by regression which maximizes the correlation between objective and subjective quality. Then, we measure the correlation between objective quality metrics and subjective quality results to validate our metrics. The correlation value is 71% between the proposed quality metrics and subjective quality assessment results. For future work, we will figure out more accurate depth fusional limit and consider another depth fusion factors as contrast, size and focus of object, to reduce visual fatigue of 3D videos.

ACKNOWLEDGEMENT

This work was partially supported by the IT R&D program of MKE/IITA [2007-F036-02 Objective quality assessment system based on human visual system] and partially supported by Korean Broadcasting System.

REFERENCES

[1] "Subjective Assessment of Stereoscopic Television Pictures," ITU, Recommendation BT.1438 (2000).

[2] "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU, Recommendation BT.500-10, (2000).

[3] P. Gorley, N. Holliman, "Stereoscopic Image Quality Metrics and Compression," Stereoscopic Displays and Virtual Reality Systems XIX, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol.6803 (2008)

[4] Piter J. H. Seuntiens, Lydia M. J. Meesters, Wijnand A. IJsselsteijn, "Perceived Quality of Compressed Stereoscopic Images: Effects of Symmetric and Asymmetric JPEG Coding and Camera Separation," ACM Transactions on Applied Perception, Vol.3, No.2, 95–109 (2006).

[5] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P.H.N. de With, and T. Wiegand, "The Effect of Depth Compression on Multiview Rendering Quality," 3DTV Conference, (2008).

[6] Y. Morvan, D. Farin, P. H.N. de With, "Depth-Image Compression based on an R-D Optimized Quadtree Decomposition for the Transmission of Multiview Images", IEEE International Conference on Image Processing, (2007).

[7] P. Merkle, A. Smolic, K. Müller, T. Wiegand, "Multi-view Video Plus Depth Representation and Coding", IEEE International Conference on Image Processing, (2007).

[8] W.A.Ijsselsteijn, H.de Ridder, J.Vliegen, "Subjective Evaluation of Stereoscopic Images: Effects of Camera Parameters and Display Duration," IEEE Transaction on CSVT, Vol.10, No.2 (2000).

[9] "3DC Safty Guidelines for Popularization of Human-friendly 3D," 3D Consortium (2006).

[10] Y. Nojiri, H.Yamanoue, S.Ide, S.Yano, F.Okano, "Parallax Distribution and Visual Comfort on Stereoscopic HDTV," IEIC Technical Report, Vol.102, No.224 (2002).

[11] D. Sharstein, R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," Proc. of IEEE Workshop on Stereo and Multi-Baseline Vision, 131-140 (2001).

[12] S. Yano, M. Emoto, T. Mitsuhashi, "Two Factors in Visual Fatigue Caused by Stereoscopic HDTV Images," Displays, Vol. 25, Issue 4, 141-150 (2004).

[13] D.Qin, M.Takamatsu, Y.Nakashima, "Measurement for the Panum's Fusional Area in Retinal Fovea Using a Three-Dimention Display Device," Journal of Light & Visual Environment, Vol. 28, Issue 3, 126-131 (2004)