Depth Based Perceptual Quality Assessment for Synthesised Camera Viewpoints

Erhan Ekmekcioglu, Stewart Worrall, Demuni De Silva,
Anil Fernando, and Ahmet M. Kondoz

I-Lab Multimedia Communications Research Group
University of Surrey
Guildford GU2 7XH, Surrey, United Kingdom

Abstract. This paper considers the visual quality assessment for view synthesis in the context of 3D video delivery chain. It is targeted to perceptually quantify the reconstruction quality of synthesised camera viewpoints. It is needed for developing better QoE models related to 3D-TV, as well as for a better representation of the effect of depth maps on views synthesis quality. In this paper, existing 2D video quality assessment methods, like PSNR and SSIM, are extended to assess the perceived quality of synthesised viewpoints based on the depth range. The performance of the extended assessment techniques is measured by correlating multiple sample video assessment scores to that of the Video Quality Metric (VQM) scores, which are a robust reflector of real subjective opinions.

Keywords: 3DTV, Free-viewpoint Video, Video Quality Assessment, Depth Map, Multi-view Video.

1 Introduction

The quality assessment of video has been a challenging research task. The subjective assessment is very time consuming, costly and cannot definitely be conducted in real time. PSNR cannot accurately model the perceptual quality, since the human perception of image/video distortions and human visual system properties are not taken into account [1]. Video Quality Metric (VQM) [2] measures the perceptual attributes of various video impairments and combines them into a single metric. Despite its complexity, VQM has high correlation with subjective video quality assessment.

For stereoscopic video, a number of quality assessment methodologies are available, although there is not yet a universal standard. PSNR does not give information about depth perception. The reason is that the attributes associated with 2D video cannot be directly used in measuring the real naturalness, depth perception and the overall image quality of 3D video. In [3], it is found that the output from the VQM metric can be mapped so that it correlates strongly with both the overall viewer perception of image quality and depth perception.

F. Alvarez and C. Costa (Eds.): UCMEDIA 2010, LNICST 60, pp. 76–83, 2012.

The quality assessment of arbitrary camera viewpoint synthesis via a pair of cameras is also an open research issue. Especially, the recent advancements in 3D-TV delivery systems necessitate the correct analysis of synthesised viewpoints that drive next generation multi-view displays. One of them is the 3D HD multimedia delivery chain envisaged in the recent EU ICT FP7 project called DIOMEDES that aims at content aware 3D delivery. In the same context, by incorporating the effect of depth maps on the perceived visual quality, better depth extraction and compression algorithms can be generated. In [4] authors have analysed the effect of depth map compression on the geometric distortions generated in the 3D space and accordingly on synthesised camera viewpoints. In [5], authors have implemented a mode decision algorithm for depth map compression in 3D-TV systems based on the effect on view synthesis without saliency adaptation. The work presented in this paper proposes a depth range and depth consistency based adaptation for the view synthesis quality assessment that will improve both the multi-view depth estimation and compression within DIOMEDES.

The 3D Video (3DV) Ad Hoc group formed under MPEG utilises a modified version of PSNR to evaluate the quality of view synthesis, which is called PSPNR (Peak Signal to Perceptual Noise Ratio) [6]. The authors in [6] introduce a Just Noticeable Difference (JND) model into pixel categorization, which is based on some human visual system traits. However, this technique doesn’t take into consideration any kind of adaptation to the scene depth.

Based on the knowledge that low depth areas are more vulnerable to rendering distortions than high depth areas, this paper proposes an unequal weighting based quality evaluation approach to be applied on most commonly used 2D video quality assessment tools: PSNR and SSIM. The weighting is based on the scene depth information. The proposed framework also takes into consideration the scene motion activity in time and adapts a factor in the final score, that is related to temporal consistency of non-moving background objects during view synthesis.

Section 2 presents the proposed framework. Section 3 gives the details of the experiments and discusses the results. Section 4 gives the subjective evaluation results to justify the results in section 3. Finally, section 4 concludes the paper.

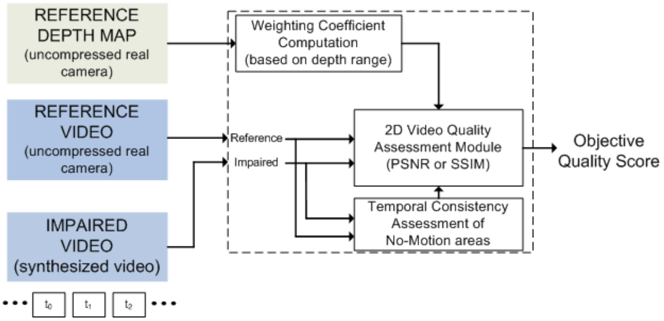

REFERENCE DEPTH MAP (uncompressed real camera)
REFERENCE VIDEO (uncompressed real camera)
IMPAIRED VIDEO (synthesized video)
Weighting Coefficient Computation (based on depth range)
Reference
Impaired
2D Video Quality Assessment Module (PSNR or SSIM)
Temporal Consistency Assessment of No-Motion areas
Objective Quality Score
t_0
t_1
t_2

Fig. 1. Block diagram illustration of the proposed view synthesis quality assessment framework

2 Proposed Virtual Viewpoint Quality Assessment Framework

The proposed framework includes depth range and scene motion activity adaptive partitioning of error map. Accordingly, the final score of view synthesis quality per video frame is computed as the weighted average of the corresponding error map, where the weighting is performed via the mentioned adaptation. Figure 1 shows the illustrative block diagram of the overall process. The reference video signal is a captured video and the reference uncompressed depth map represents the original scene depth information from the corresponding viewpoint (where the synthesis takes place). The impaired video signal is the synthesised camera viewpoint. According to the weighting coefficient calculation process during quality assessment, some video frame regions, which are far away from the capturing camera, are given less importance when incorporating them in the overall visual quality score and some frame sections are given zero importance which makes them get excluded during error signal computation.

The next two sub-sections describe the weighting coefficient computation based on scene depth information, temporal consistency assessment in no-motion video frame sections.

2.1 Weighting Computation Based on Depth Range

In measuring the objective quality of a video frame, all pixels or blocks are assigned the same weight in contributing to the final frame quality. In a synthesised video frame, it is more likely to observe rendering related distortions in the vicinity of front objects. Assuming that human visual perception is more affected by the distortions happening on the front part of the scene it is planned to give more weight to such frame regions in calculating the 'perceived' frame distortion.

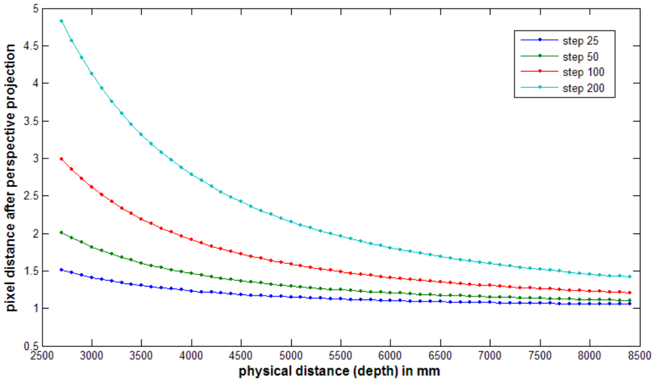


Fig. 2. Pixel distance between the neighbouring pixels after perspective projection, where the step sizes in mm represent the depth difference between the neighbouring pixels

The weighting coefficients assigned to individual video frame pixels vary between 0 and 1. In the proposed approach, the weighting coefficient function is selected as a piecewise linear function, where the weighting coefficient becomes 0 when the local scene depth is above a certain upper threshold (Z_f) and becomes 1 when the local scene depth is under a certain lower threshold (Z_n). In the depth zone, which lies between the corresponding lower and upper thresholds, the weighting coefficients change linearly, inversely proportional to the local scene depth range. Equation 1 depicts this.

$$WC(x,y)=\begin{cases} 0, & Z(x,y)>Z_f \\ \dfrac{(Z(x,y)-Z_f)}{(Z_n-Z_f)}, & Z_n \le Z(x,y) \le Z_f \\ 1, & Z(x,y)<Z_n \end{cases} \tag{1}$$

WC(x,y) stands for the weighting coefficient at the corresponding pixel position *(x,y)* in the video frame. Similarly, *Z(x,y)* refers to the physical distance (depth) of the corresponding pixel position *(x,y)*. Z_n and Z_f are the lower and the upper threshold depth values, which represent the sensitivity levels of the textural connectivity of corresponding video frame regions to break-downs during view synthesis. Considering two neighbour pixels in the source viewpoint, *(x,y)* and *(x-1,y),* a certain amount of depth difference, which is not expected to break the connectivity between them after warping to another viewpoint may cause so, if the depth levels of them are sufficiently low. Considering the pinhole camera model and using the computation of coordinate transforms, the horizontal pixel distance Δd between these two neighbour pixels after warping to the coordinates of the target viewpoint (*C'*) is

$$\Delta d = 1 + a \cdot \frac{Z(x-1,y)-Z(x,y)}{Z(x-1,y)\cdot Z(x,y)} \tag{2}$$

where a is a constant derived from the rotational, translational and affine parameters of the source and target viewpoints. The graph shown in Figure 2 depicts the computed Δd value according to the pixel's depth values. The physical distance axis corresponds to *Z(x,y)*. Several depth differences between neighbour pixels are considered (from 25 mm to 200 mm apart). Z_f is defined as the depth value *min(Z(x,y), Z(x-1,y))* such that $\Delta d = 1$. Pixels with a depth value greater than Z_f are not vulnerable to separation and visual distortion after warping, for a pre-determined step depth difference. Similarly, Z_n is defined as the depth value *min(Z(x,y), Z(x-1,y))* such that $\Delta d = 2$. Pixels with a lower depth than Z_n, are separated by more than 1 pixels after perspective projection and therefore, the corresponding frame regions are subject to visual distortions. Hence, such frame sections are given most significance by assigning them the highest coefficient, when computing the overall perceptual frame distortion.

2.2 Temporal Consistency Assessment in Motionless Frame Regions

The proposed virtual viewpoint quality assessment framework also takes into account the temporal consistency among successive synthesised frames. This functionality is achieved by considering the motion activity in both the synthesised video and the original video. Basically, motionless frame regions might have suspiciously high motion activity in the synthesised video, causing a flickering effect.

First, the frame absolute difference D_o is computed between the current time colour texture frame and the previous time colour texture frame of the original camera (O), such as

$$D_o = |O(t) - O(t-1)| \tag{3}$$

Second, the frame absolute difference D_s is also computed between the current time colour texture frame and the previous time colour texture frame of the synthesised camera (S), such as

$$D_S = |S(t) - S(t-1)| \tag{4}$$

At pixel locations *(x,y)*, if $D_o(x,y) < m$, where m represents a motion activity threshold, and $D_S(x,y) > m$, then the corresponding pixel locations are flagged as temporally suspicious pixels with a risk of flickering. Hence, a coefficient of '1' is given to such pixels, where a '0' coefficient is given to all remaining pixels, during the frame temporal difference calculation. The frame temporal difference metric is selected either as PSNR or SSIM, where the reference video frame is the previous time instant frame of the synthesised video source and the impaired video frame is the current time instant frame of the synthesised video source. The calculated total temporal frame synthesis error corresponds to the total amount of flickering activity which is not present in the original camera source.

Once the two sets of weighting coefficients are computed, they are multiplied with the calculated error frame. Error frame indicates the calculated MSE map if PSNR is utilised or the calculated structural similarity index map if SSIM is utilised. The corresponding final objective score *Final* is computed as

$$Final(n) = c_1 \cdot Final_s(n) + c_2 Final_t(n) \tag{5}$$

where n is the frame index, and subscripts s and t correspond to the final scores considering the depth range adaptation and the temporal consistency check, respectively. c_1 and c_2 are the coefficients of individual metrics that contribute to the overall perceived distortion. In this case, this relationship is assumed to be additive and the contributions from each is considered to be equal, i.e. $c_1 = c_2 = 0.5$.

3 Experimental Results

To evaluate the performance of the proposed framework, a number of camera viewpoints are synthesised using colour texture videos and depth maps, with various kinds of artefacts added to them. Two different multi-viewpoint video sequences (*Akko&Kayo* and *Newspapers*) are used in the experiments. For each test video sequence, two target camera viewpoints (camera #1 from cameras #0 and #2, camera #2 from cameras #1 and #3) are used for view synthesis.

To distort colour texture video sequences, two different quantisation step sizes are used. For distorting the depth map sequences, four different operations are utilised, one deployed each time. One source of distortion is the quantisation distortion. The second source of distortion is added via passing the depth map videos through a low pass filter (by down-sampling and up-sampling with a ratio of ½ and 2, respectively). In this way, the high frequency components in the depth map videos are eliminated. The third source of artificial distortion is added by shifting the object borders. In the experiments, the object borders in the depth map videos are shifted to left and right separately, by 5 pixels. The final source of distortion is introduced to the depth map videos by adding artificial local spot errors in certain regions to create temporal inconsistency in the synthesised videos. All combinations of distorted colour texture and depth map video sequences are used for synthesising the test videos. In total, 64 different synthesised videos are used for quality assessment experiments.

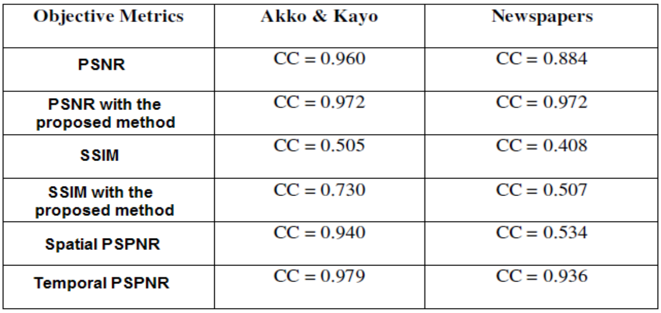
The performances of six different objective quality metrics are computed, taking the VQM scores as the reference, i.e. as mean subjective opinions. It is chosen as such, because VQM [2] has a high correlation with the subjective assessment scores despite its computational complexity.

Table 1 shows the correlation coefficient results of the mentioned objective quality assessment methods.

When the proposed framework is applied on the conventional PSNR metric, an increase by 0.012 and 0.088 in *CC* is obtained for *Akko&Kayo* and *Newspapers* sequences, respectively. In the case of SSIM metric, these numbers are 0.225 and 0.099, respectively. The improvements in the correlation coefficients are significant.

According to the results in Table 1, Spatial PSPNR and Temporal PSPNR metrics that are extended from the PSNR metric, show a better performance than the conventional PSNR metric in terms of assessing the perceptual video quality.

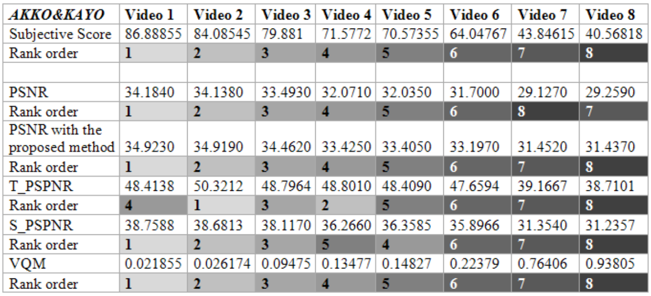
Table 1. Correlation coefficient scores of the evaluated objective metrics



Objective Metrics	Akko & Kayo	Newspapers
PSNR	CC = 0.960	CC = 0.884
PSNR with the proposed method	CC = 0.972	CC = 0.972
SSIM	CC = 0.505	CC = 0.408
SSIM with the proposed method	CC = 0.730	CC = 0.507
Spatial PSPNR	CC = 0.940	CC = 0.534
Temporal PSPNR	CC = 0.979	CC = 0.936

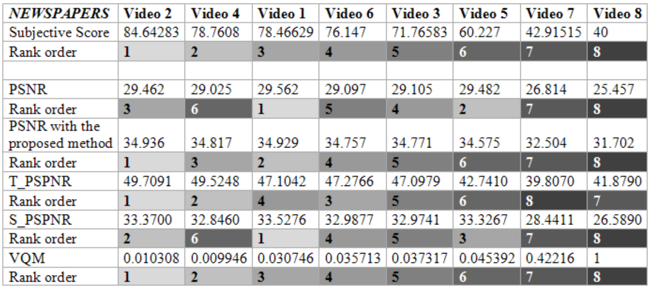
However, this improvement is not very significant. Excluding the Temporal PSPNR and comparing PSNR and Spatial PSPNR to each other, it is observed that the improvement in *CC* is in the range of 0.018 to 0.052.

Table 2. Subjective rankings for Akko & Kayo



AKKO&KAYO	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8
Subjective Score	86.88855	84.08545	79.881	71.5772	70.57355	64.04767	43.84615	40.56818
Rank order	1	2	3	4	5	6	7	8
PSNR	34.1840	34.1380	33.4930	32.0710	32.0350	31.7000	29.1270	29.2590
Rank order	1	2	3	4	5	6	8	7
PSNR with the proposed method	34.9230	34.9190	34.4620	33.4250	33.4050	33.1970	31.4520	31.4370
Rank order	1	2	3	4	5	6	7	8
T_PSPNR	48.4138	50.3212	48.7964	48.8010	48.4090	47.6594	39.1667	38.7101
Rank order	4	1	3	2	5	6	7	8
S_PSPNR	38.7588	38.6813	38.1170	36.2660	36.3585	35.8966	31.3540	31.2357
Rank order	1	2	3	5	4	6	7	8
VQM	0.021855	0.026174	0.09475	0.13477	0.14827	0.22379	0.76406	0.93805
Rank order	1	2	3	4	5	6	7	8

Table 3. Subjective rankings for Newspapers



NEWSPAPERS	Video 2	Video 4	Video 1	Video 6	Video 3	Video 5	Video 7	Video 8
Subjective Score	84.64283	78.7608	78.46629	76.147	71.76583	60.227	42.91515	40
Rank order	1	2	3	4	5	6	7	8
PSNR	29.462	29.025	29.562	29.097	29.105	29.482	26.814	25.457
Rank order	3	6	1	5	4	2	7	8
PSNR with the proposed method	34.936	34.817	34.929	34.757	34.771	34.575	32.504	31.702
Rank order	1	3	2	4	5	6	7	8
T_PSPNR	49.7091	49.5248	47.1042	47.2766	47.0979	42.7410	39.8070	41.8790
Rank order	1	2	4	3	5	6	8	7
S_PSPNR	33.3700	32.8460	33.5276	32.9877	32.9741	33.3267	28.4411	26.5890
Rank order	2	6	1	4	5	3	7	8
VQM	0.010308	0.009946	0.030746	0.035713	0.037317	0.045392	0.42216	1
Rank order	1	2	3	4	5	6	7	8

4 Subjective Test Results

To further justify the performance of the proposed framework applied on the conventional video quality evaluation methods, subjective tests are conducted using a subset of the synthesised test videos. 15 non-expert subjects are asked to assess the visual quality of different synthesised videos (from 1 to100) in random order, taking the original camera view as reference. For each test, the standard deviation of the subject's opinion is calculated and accordingly, the outliers are eliminated. Table 2 and Table 3 show the average subjective scores as well as the scores of PSNR, PSNR with the proposed framework, *S_PSPNR* and *T_PSPNR* for *Akko&Kayo* and *Newspapers* sequences, respectively. The quality ranking order of each synthesised video for the corresponding objective assessment metrics as well as subjective scores, are provided on each table. The results verify the assumption that VQM can be used as a robust reflector of the subjective scores. According to the results, the quality ranking according to the metric on which the proposed framework is applied, has a better correlation with the quality ranking according to the subject opinion than the quality ranking according to *S_PSPNR*.

5 Conclusion

This paper addressed the issue of evaluating the perceptual quality of the synthesised videos to be considered within the 3D-TV delivery systems. A new full-reference perceptual quality assessment framework is designed and tested using some state-of-the-art video quality assessment metrics, like PSNR and SSIM. The proposed framework assigns more importance to regions of a synthesised frame that are relatively more open to synthesis related distortions. This is achieved by exploiting the depth map at the target camera position. Another feature of the proposed framework is that temporal consistency is tracked within the synthesised videos. Accordingly, better performance is achieved in reflecting real subject opinions, with respect to PSNR, SSIM and PSPNR. The presented quality assessment idea can be incorporated with multi-view depth estimation and coding systems for better 3D scene reconstruction quality in high quality entertainment multimedia delivery.

Acknowledgement. This work has been supported by the DIOMEDES (*Distribution of Multi-View Entertainment Using Content Aware Delivery Systems*) project, funded under the European Commission ICT 7th Framework Programme (www.diomedes-project.eu).

References

[1] Girod, B.: What's wrong with Mean-Squared Error. In: Watson, A.B. (ed.) Digital Images and Human Vision, ch. 15, pp. 207–220. The MIT Press (1993)

[2] Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Transactions on Broadcasting 50(3), 312–322 (2004)

[3] Hewage, C.T.E.R., Worrall, S., Dogan, S., Kondoz, A.M.: Quality Evaluation of Colour plus Depth Map Based Stereoscopic Video. IEEE Journal of Selected Topics in Signal Processing 3(2), 304–318 (2009)

[4] Merkle, P., Morvan, Y., Smolic, A., Farin, D., Mueller, K., de With, P.H.N., Wiegand, T.: The Effects of Multiview Depth Video Compression on Multiview Rendering. Signal Processing: Image Communication 24, 73–88 (2009)

[5] De Silva, D., Fernando, W.A.C., Kodikara Arachchi, H.: A New Mode Selection Technique for Coding Depth Maps of 3D Video. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), Dallas (March 2010)

[6] Zhao, Y., et al.: Perceptual measurement for evaluating quality of view synthesis. In: MPEG Doc. M16407, Maui, USA (April 2009)