PREDICTING 3D QUALITY BASED ON CONTENT ANALYSIS

Philippe Hanhart and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

Development of objective quality metrics that can reliably predict perceived quality of 3D video sequences is challenging. Various 3D objective metrics have been proposed, but PSNR is still widely used. Several studies have shown that PSNR is strongly content dependent, but the exact relationship between PSNR values and perceived quality has not been established yet. In this paper, we propose a model to predict the relationship between PSNR values and perceived quality of stereoscopic video sequences based on content analysis. The model was trained and evaluated on a dataset of stereoscopic video sequences with associated ground truth MOS. Results showed that the proposed model achieved high correlation with perceived quality and was quite robust across contents when the training set contained various contents.

***Index Terms*—** 3D, objective quality, subjective quality, content analysis, quality prediction

1. INTRODUCTION

With the rapid growth of 3D video technologies, the design of objective quality assessment methods, i.e., metrics, that can reliably predict the quality of 3D content as perceived by the end user, is of crucial importance. Subjective tests are time consuming, expensive, and not always feasible. Therefore, objective measurements are needed, especially to assess advances in the design of coding technologies. Despite the efforts of the scientific community in recent years, 3D video quality assessment is still an open challenge. There are no metrics that are widely recognized as reliable predictors of human 3D quality perception. PSNR is commonly accepted and used by video coding experts to evaluate the performance of coding algorithms, even though its correlation with human perception of visual quality is known to be limited.

PSNR values below 25 dB and over 40 dB are often considered as bad and excellent quality, respectively. However, the exact relationship between PSNR values and perceived quality has not been established yet. This relationship should consider non-linearities and saturation effect of the human visual system (HVS). As it was shown that PSNR is strongly content dependent [1], this relationship should also be determined for each content separately.

Korhonen and You [2] have found a strong correlation between the parameters of an exponential function, which was used to map PSNR values to mean opinion scores (MOS), and the spatial and temporal activity of a set of six 2D video sequences. Based on this finding, they have used a linear regression to estimate the parameters of the mapping function based on the spatial and temporal activity of the six contents.

Liao *et al.* [3] have shown how the Quality of Experience (QoE) of a set of 2D video sequences was correlated with objective quality metrics, video content characteristics, and device features. From these results, a linear mapping between multi-scale structural similarity (MS-SSIM) and QoE was proposed. The authors assumed that the parameters of the linear mapping can be accurately estimated from the amount of spatial details, motion level, display resolution, and device type, but this assumption was not investigated.

In this paper, we investigate the prediction of perceived quality of stereoscopic video sequences based on PSNR and content analysis. We propose a model based on a logistic function to map the PSNR values to perceived quality, which should better represent the saturation effect of the HVS when compared to linear or exponential mapping. The parameters of the mapping function were predicted using 2D and 3D content features, which were extracted from the original sequences. Each parameter of the logistic function was predicted from two content features. To select the most relevant features for each parameter, the dataset was split into training and testing sets and the model was trained on the training set. To evaluate how well the proposed model predicts perceived quality, the trained model was applied to the testing set.

A dataset of eight stereoscopic contents with associated ground truth MOS was used [4, 5]. The dataset is composed of six natural contents and two computer-generated contents. These contents are commonly used by MPEG, VCEQ, and other researchers to evaluate the performance of 3D video compression algorithms. The subjective results have been collected during the evaluations of the MPEG Call for Proposals on 3D Video Coding Technology [4]. Only the results for the 3-view configuration, fixed stereo pair, of the two best AVC proposals and two best HEVC proposals were used as ground truth. The PSNR was computed as the average PSNR of the left and right views of the displayed stereo pair.

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1) and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

The remainder of the paper is organized as follows. Section 2 describes the proposed model. In Section 3, the evaluation of the proposed model is reported and analyzed. Finally, Section 4 concludes the paper.

2. PROPOSED MODEL

This section describes the feature extraction and feature selection processes used to predict the parameters of the mapping function of the proposed model.

2.1. Feature extraction

Both 2D and 3D features were extracted from the original video sequences. For the 2D features, the well-known spatial perceptual information (SI) and temporal perceptual information (TI) [6] are often used to characterize the amount of spatial detail of a picture and temporal changes of a video sequence, respectively. These two features were used by Korhonen and You [2] to map PSNR values to perceived quality in the case of 2D video sequences (see Section 1). In this paper, the temporal perceptual information and a modified version of the spatial perceptual information, referred to as $\tilde{SI}$, were used. $\tilde{SI}$ was computed using a Sobel kernel multiplied by $\frac{1}{8}$. The 2D features were computed on the luminance component of each content.

Mittal *et al.* [7] have proposed that 3D images have certain statistical properties that can be captured using simple statistical measures of the disparity distribution. They used statistical features from disparity and disparity gradient maps to predict the Quality of Experience of 3D images and video sequences. Thus, the following 3D features were computed on the disparity map D of each content, according to [7]:

1. mean disparity $\mu = \mathrm{E}[D]$,
2. median disparity $med = median(D)$,
3. disparity standard deviation $\sigma = \sqrt{\mathrm{E}[(D-\mu)^2]}$,
4. kurtosis of disparity $\kappa = \frac{\mathrm{E}[(D-\mu)^4]}{(\mathrm{E}[(D-\mu)^2])^2}$,
5. skewness of disparity $skew = \frac{\mathrm{E}[(D-\mu)^3]}{(\mathrm{E}[(D-\mu)^2])^{(3/2)}}$,
6. mean differential disparity $\mu_d = \mathrm{E}[\delta D]$,
7. differential disparity standard deviation $\sigma_d = \sqrt{\mathrm{E}[(\delta D-\mu_d)^2]}$,
8. kurtosis of differential disparity $\kappa_d = \frac{\mathrm{E}[(\delta D-\mu_d)^4]}{(\mathrm{E}[(\delta D-\mu_d)^2])^2}$,
9. skewness of differential disparity $skew_d = \frac{\mathrm{E}[(\delta D-\mu_d)^3]}{(\mathrm{E}[(\delta D-\mu_d)^2])^{(3/2)}}$

where the differential disparity (δD) was computed using a Laplacian operator on the disparity map. The 3D features were computed on a frame-by-frame basis and then averaged across frames.

Therefore, a total of eleven features, two 2D features and nine 3D features, were extracted for each content.

2.2. Mapping function

To consider non-linearities and saturation effect of the human visual system, a logistic function was used to predict perceived quality from PSNR values:

$$MOS_p(PSNR) = a + \frac{b-a}{1+\exp\left[-c\left(PSNR-d\right)\right]} \quad (1)$$

where the parameters c and d are related to the slope and translation of the logistic function, respectively, and can be controlled independently. The parameters a and b were determined as follows. The subjective scores range R is typically divided into five parts of equal lengths, which are associated with distinct quality levels. By varying the bit rate, the quality of the video sequence varies from the lowest quality level to the highest quality level. Therefore, we assumed that the horizontal asymptotes of the logistic function are associated with the lowest and highest quality levels for the lowest and highest bit rates, respectively:

$$\begin{aligned} \lim_{PSNR \to 0} MOS_p(PSNR) = a = R_{10\%} \\ \lim_{PSNR \to +\infty} MOS_p(PSNR) = b = R_{90\%} \end{aligned} \quad (2)$$

To determine the optimal values c_o and d_o for each content of the dataset, a fitting using Equation 1, partially constrained by Equation 2, was performed between the PSNR values and ground truth MOS, for each content separately.

2.3. Feature selection

The total number of extracted features (see Section 2.1) is higher than the number of contents in the dataset. Therefore, the number of features used to predict the parameters c and d of the mapping function in Equation 1 needs to be restricted. To avoid the risk of over-fitting, only two features out of eleven were used to predict each parameter of the logistic function:

$$c = \alpha f_1 + \beta f_2 + \gamma \quad (3)$$

$$d = \delta f_3 + \epsilon f_4 + \zeta \quad (4)$$

where f_1, f_2, f_3, and f_4 are content features, and α, β, γ, δ, ϵ, and ζ are coefficients.

To determine which extracted features should be used to predict the parameters of the mapping function, the proposed model was trained on a subset of contents of the dataset. For each combination of two features, a least square regression was performed to determine the coefficients of Equation 3. The pair of features which obtained the best correlation with the optimal parameters c_o of the contents in the training set was chosen to predict the parameter c of the contents in the testing set. Similarly, for each combination of two features, a least square regression was performed to determine the coefficients of Equation 4. The pair of features which obtained the best correlation with the optimal parameters d_o of the contents in the training set was chosen to predict the parameter d of the contents in the testing set.

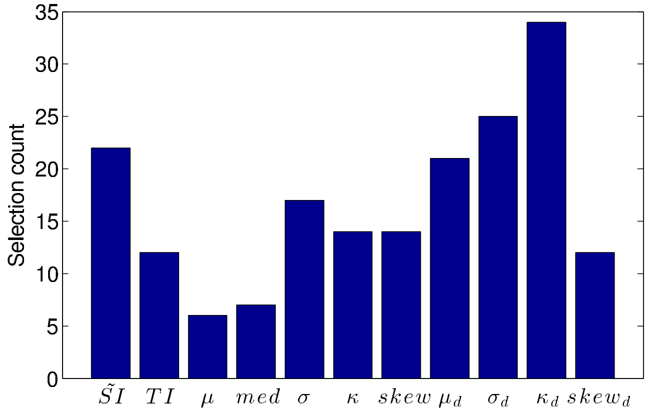

Fig. 1: Histogram of features selected across all train-test trials to predict the parameter c.

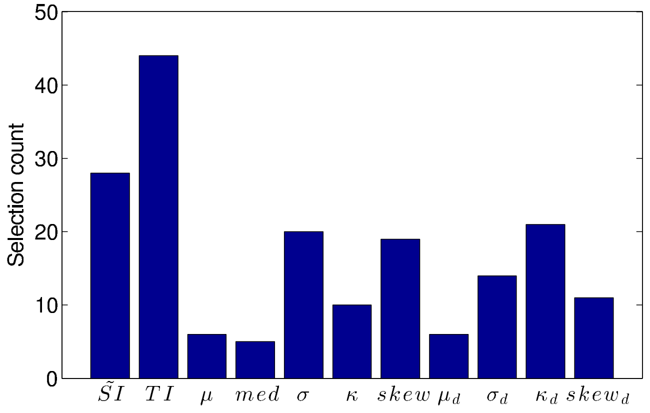

Fig. 2: Histogram of features selected across all train-test trials to predict the parameter d.

3. PERFORMANCE EVALUATION

To evaluate the performance of the proposed model, a dataset of 3D video sequences with associated ground truth subjective scores, containing a total of $n = 8$ contents, was used. The contents were divided into a training set and a testing set. The size of the training set was varied between five and seven contents to evaluate the influence of the training set size. For a fixed training set of size k, all possible $\binom{n}{k}$ combinations to split the contents into training and testing sets were realized to evaluate the robustness of the proposed model across contents. For each train-test trial, the model was trained on the training set according to Section 2.3 and the performance of the trained model was evaluated on the testing set.

3.1. Selected features

Figure 1 and Figure 2 show the histograms of features selected across $\binom{8}{7}+\binom{8}{6}+\binom{8}{5} = 92$ train-test trials to predict the parameters c and d, respectively. To predict the parameter c, no feature, except κ_d, was selected in more than a third of the train-test trials. Features extracted from the differential disparity map were more often selected than features extracted from the disparity map. This result is intuitive since the differential disparity map is related to occluded areas. Whereas the temporal activity was used to model the slope of the exponential function in [2], the TI feature was selected only 12 times out of 92 train-test trials. Regarding the prediction of the parameter d, the TI and $\tilde{SI}$ features were selected in almost half and a third of the train-test trials, respectively. However, the translation of the exponential function in [2] was modeled using the spatial activity. This difference might come from the fact that the training contents only covered a limited range of spatial activity and no general trend could be drawn.

3.2. Performance indexes

The following properties of the prediction of perceived quality were considered: accuracy, monotonicity, and consistency.

The Pearson linear correlation coefficient (PCC), root-mean-square error (RMSE), and mean absolute error (MAE) were computed between predicted quality, MOS_p, and ground truth scores, MOS, to estimate the accuracy. The root-mean-square error is defined as follows:

$$RMSE = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(MOS_i - MOS_{pi})^2}$$

where M is the total number of points.

To estimate monotonicity and consistency, the Spearman rank order correlation coefficient (SROCC) and the outlier ratio (OR) were computed between MOS_p and MOS, respectively. The outlier ratio is defined as follows:

$$OR = \frac{\text{total number of outliers}}{M}$$

$$\text{outlier if: } |MOS_i - MOS_{pi}| > CI_i$$

where M is the total number of points and CI_i is the 95% confidence interval corresponding to MOS_i.

3.3. Anchors

To compare the performance of the proposed model to useful reference points, a fitting using Equation 1, unconstrained, was performed between the PSNR values and ground truth MOS to determine all four parameters (a, b, c, and d). In this case, no prediction was performed and all eight contents were used. The fitting was applied in two different ways:

A. on all contents at once,

B. on each content separately.

Table 1: Performance indexes of the anchors.

Anchor	PCC	SROCC	RMSE	MAE	OR
A	0.3926	0.3973	1.4592	1.1057	0.7344
B	0.9462	0.9015	0.3723	0.2879	0.2109

In the latter case, the performance indexes were computed separately on each content and then averaged across contents. Anchor A does not consider content characteristics since all contents are mixed. Therefore, the proposed model must show better performance than anchor A to be considered as valid. However, anchor B does consider all contents characteristics as the fitting is applied on each content separately. Thus, this anchor should provide upper bounds on PCC and SROCC as well as lower bounds on RMSE, MAE, and OR for comparison with the proposed model. Table 1 reports the performance indexes of the two anchors.

3.4. Results

Table 2 reports the mean value and standard deviation of the performance indexes across $\binom{n}{k}$ train-test trials of the proposed model for different training set sizes. For each train-test trial, the best features selected on the training set (with a frequency shown in Figure 1 and Figure 2) were used to predict the parameters of the mapping function for the testing set. Whereas the PCC and SROCC were quite high over the different training set sizes, the RMSE and MAE increased significantly for $k < 7$. Since the mapping function was applied on each content separately in the proposed model, the PCC and SROCC values were quite high when compared to anchor A. Nevertheless, if the mapping function had a wrong slope or translation, namely if there was an error in the prediction of c or d, the RMSE, MAE, and OR values increased significantly compared to anchor B. For $k = 7$, the standard deviation of the PCC and SROCC was quite low, which indicates that the proposed model was quite robust across contents when the training set contained various contents. However, in some cases for $k < 7$, the predicted quality scores had a negative correlation with the ground truth MOS, which explains the high standard deviation for PCC and SROCC. This indicates that the training set should contain different contents covering a wide range of spatiotemporal characteristics. In general, predicted quality always achieved a high correlation with perceived quality when compared to anchor A, which does not consider content characteristics in the fitting process. This result indicates that content analysis can improve the accuracy of the mapping of PSNR values to perceived quality.

Table 2: Performance indexes of the proposed model.

(a) Mean value

	PCC	SROCC	RMSE	MAE	OR
$k = 7$	0.9341	0.9015	1.2181	1.0104	0.6250
$k = 6$	0.8743	0.8711	2.0893	1.7975	0.7143
$k = 5$	0.7815	0.7863	2.2065	1.9106	0.7437

(b) Standard deviation

	PCC	SROCC	RMSE	MAE	OR
$k = 7$	0.0595	0.0888	1.0914	0.9299	0.3204
$k = 6$	0.2552	0.2486	1.7016	1.6262	0.2395
$k = 5$	0.4559	0.4500	1.7753	1.7014	0.2351

4. CONCLUSION

In this paper, we proposed a model to predict perceived quality of stereoscopic video sequences based on content analysis. A logistic function was used to map the PSNR values to perceived quality. The parameters of the mapping function were predicted using 2D and 3D content features. Results showed that the proposed model achieved high correlation with perceived quality and was quite robust across contents when the training set contained various contents. This finding indicates that perceived quality can be predicted from PSNR values based on content analysis and that subjective tests might not be always required.

To extend our work, different metrics will be considered instead of PSNR in future investigations. The correlation between the parameters of the logistic function and additional features will be investigated as well. A dataset containing more contents will be used to further evaluate the performance of the proposed model.

5. REFERENCES

[1] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[2] J. Korhonen and J. You, "Improving Objective Video Quality Assessment with Content Analysis," in *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, Jan. 2010.

[3] Y. Liao, A. Younkin, J. Foerster, and P. Corriveau, "Achieving High QoE Across the Compute Continuum: How Compression, Content, and Devices Interact," in *7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, Jan. 2013.

[4] ISO/IEC JTC1/SC29/WG11, "Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology," Doc. N12347, Geneva, Switzerland, Nov. 2011.

[5] P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Two Synthesized Views Using Objective Metrics," in *7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, Jan. 2013.

[6] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Apr. 2008.

[7] A. Mittal, A.K. Moorthy, J. Ghosh, and A.C. Bovik, "Algorithmic assessment of 3D quality of experience for images and videos," in *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop*, Jan. 2011, pp. 338–343.