ANALYSIS OF SSIM PERFORMANCE FOR DIGITAL CINEMA APPLICATIONS

Fitri N. Rahayu, Ulrich Reiter, Marlon T.M. Nielsen, Touradj Ebrahimi, Peter Svensson, Andrew Perkis

Centre for Quantifiable Quality of Service in Communication Systems (Q2S)[1]
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

ABSTRACT

Visual quality is one of the most important issues in Digital Cinema applications, and the most practical method to measure visual quality is using objective metrics. Several objective metrics that take into account perceived visual quality have been developed in the past few years; one of them is based on the structural similarity paradigm. To analyze the performance of these objective metrics for digital cinema applications, a subjective visual quality assessment in digital cinema environments using contents from the Digital Cinema Initiative (DCI) Standard Evaluation Material has been conducted in a previous study. Since in practice digital cinema utilizes only high quality imagery, in this paper we analyze further a subset of high-quality stimuli from the previous study.

Index Terms— Digital Cinema, Subjective Quality Assessment, Objective Metrics, Perceived Quality

1. INTRODUCTION

Visual quality is always an important issue in any multimedia application, including Digital Cinema applications. The most practical and applicable method of measuring visual quality is using objective metrics. Peak-signal-to-noise-ratio (PSNR) is the most popular metric nowadays. However, PSNR is a traditional pixel based metric that is acknowledged to possess little correlation with the human visual system (HVS). As an alternative, some objective metrics that take the human visual system into account have been proposed.

Objective metrics that are going to be widely used in any application are the metrics that produce the best performance; these metrics can estimate satisfactorily the perceived visual quality of users in their intended applications. The ultimate measure of perceived visual quality for digital cinema applications is to conduct subjective visual assessments using human subjects in a controlled realistic environment. The collected data from such experiments is the ground truth, and the objective metric that has the best performance is the metric that has the best correlation with the ground truth.

One of the types of objective metrics that takes into account the Human Visual System (HVS) is the group of metrics developed based on a structural similarity paradigm, which was introduced by Wang and Bovik [1][2]. In their study, structural similarity measurement (SSIM) — both single scale and multi scale — has shown a good correlation with perceived quality, outperforming PSNR. For that reason, SSIM has a potential for measuring perceived visual quality in digital cinema applications, which motivated us to study the suitability of SSIM — both single scale and multi scale — to measure the perceived quality of images taken from Digital Cinema Initiative (DCI) Standard Evaluation Material (StEM) [3].

In a previous study, a subjective visual assessment was carried out in a DCI-specified digital cinema in Trondheim, Norway [3]. The stimuli used in the subjective assessment were six 2K images taken from StEM [4]. Because only the luminance component of images was taken into account in that study, the luminance component was extracted from each image resulting in six gray scale 2K images. The subjective assessment was performed by examining a range of JPEG2000 compression errors introduced by varying bit rates. 8 different coding conditions were applied to create 8 processed images from each source image. These selected conditions covered the whole range of quality levels (from "bad" to "excellent"), with subjects being able to detect the change in quality from each quality level to the next.

To validate the results of single-scale SSIM and multi-scale SSIM, we investigated the correlation between the measurement data from these metrics and the subjective data collected from our subjective visual assessment in Digital Cinema. Our study showed that in the case of digital cinema environment and content, SSIM does not exhibit the same type of performance that has been reported in the literature, when compared to PSNR metric [3].

However, in practice, digital cinema applications only make use of high quality imagery and utilize JPEG2000 coding conditions that do not produce bad quality level responses from the users. Therefore, it might be necessary to analyze separately the performance of these objective

metrics for a subset of the data from the stimuli that produced a quality level of fair and higher. We include a quality level of fair because we think that a satisfactory objective metric must be able to report a threshold level of perceived quality in which the quality level below the threshold is unacceptable for application users. In this paper, we report the results of a performance study of the single-scale SSIM (SS-SSIM) and the multi-scale SSIM (MS-SSIM) metrics in measuring the perceived quality of 2K digital cinema content by means of such selected subjective and objective data.

This paper is organized as follows. Section 2 introduces the structural similarity measurement paradigm. Section 3 describes the subjective quality assessment in the DCI-specified digital cinema environment and its result. Section 4 discusses the test result, and section 5 concludes the paper.

2. STRUCTURAL SIMILARITY MEASUREMENT

Natural image signals are highly structured; their pixels demonstrate strong dependencies, which contain important information about the structure of the objects in the visual scene [1]. It is assumed that the measurement of structural information changes provides a good estimation of the perceived image distortion because the human visual system is highly adapted to extract structural information [1]. Suppose $\mathbf{x}$ and $\mathbf{y}$ were two image signals; if one of the signals had perfect quality, then the similarity measure could be utilized to measure quantitatively the quality of the second signal.

Structural information in an image is defined as attributes that represent the structure of objects in the scene. The quality assessment uses local luminance and contrast because luminance and contrast can vary across a scene. The similarity measurement system is based on three comparisons: luminance, contrast, and structure. The luminance comparison function $l(x,y)$ is estimated as a comparison of the mean intensity of two discrete signals, μ_x and μ_y, which is defined by

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1},\qquad(1)$$

in which the constant C_1= 6.50 [1] is included to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. The contrast comparison function $c(\mathbf{x},\mathbf{y})$ takes a similar form, based on the standard deviation of the two signals, σ_x and σ_y,

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}.\qquad(2)$$

Here, again, the constant C_2= 58.52 [1] is included to avoid instability when $\sigma_x^2 + \sigma_y^2$ is very close to zero. The third comparison — the structure comparison function $s(\mathbf{x},\mathbf{y})$ — is defined as follows:

$$s(\mathbf{x},\mathbf{y}) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}.\qquad(3)$$

To avoid instability when $\sigma_x\sigma_y$ is very close to zero, a constant C_3= 29.26 [1] is incorporated. The general form of the SS-SSIM index between signals $\mathbf{x}$ and $\mathbf{y}$ becomes:

$$SS-SSIM(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^\alpha \cdot [c(\mathbf{x},\mathbf{y})]^\beta \cdot [s(\mathbf{x},\mathbf{y})]^\gamma,\qquad(4)$$

with $\alpha > 0$, $\beta > 0$, $\gamma > 0$ being parameters used to adjust the relative importance of the three components. In the single scale approach, the parameter values are set to $\alpha = \beta = \gamma = 1$.

The perceptibility of image details depends on the sampling density of the image signal and the distance of the image plane from the observer. When these factors vary, the subjective evaluation of a given image varies, too. For this reason, a multi scale variant of structural similarity has been developed to incorporate image details at different resolutions [2]. Taking the reference and the distorted image signals as input, the method iteratively applies a low-pass filter and down-samples the filtered image by a factor of 2. For example, at the j-th scale, the reference and the distorted image signals are low-pass filtered and down-sampled 2 j-1 times.

The overall computation is acquired by combining the measurement at different scales using

$$MS-SSIM(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M}[c_j(\mathbf{x},\mathbf{y})]^{\beta_j}[s(\mathbf{x},\mathbf{y})]^{\gamma_j},\qquad(5)$$

in which the original image is indexed as scale 1. This definition also comprises the single scale measurement as the special case $M = 1$.

The exponents α_M, β_j, and γ_j are used to adjust the relative importance of the three components, and M is set to 5. Based on a subjective parameterization test [2], the resulting parameters are $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$, respectively.

We also conducted a subjective parameterization test in order to obtain parameters that take into account the digital cinema viewing conditions. We used the same methodology as in [2] except for the test environment setup, which is a DCI-specified commercial digital cinema in Trondheim, Norway. Based on our test [3], the obtained parameters are $\beta_1 = \gamma_1 = 0.1587$, $\beta_2 = \gamma_2 = 0.2329$, $\beta_3 = \gamma_3 = 0.2298$, $\beta_4 = \gamma_4 = 0.2008$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1778$, respectively.

3. SUBJECTIVE QUALITY ASSESSMENT IN DIGITAL CINEMA ENVIRONMENT

3.1. Laboratory Set Up

The assessment has been conducted at a commercial digital cinema in Trondheim, Norway. The DCI-specified cinema setup is considered to provide ideal viewing conditions.

24

Figure 1 shows a view of the auditorium. Table 1 gives the specifications of the auditorium.

Fig. 1. Ullman auditorium of Nova Kinosenter.

Table 1. Ullman auditorium specifications.

DISPLAY	
Screen (H x W)	5 x 12 m
Projection Distance	19 m
Image Format	WS 1:1.66
	WS 1:1.85
	CS 1:2.35
HALL	
Number of Seats	440
Width	18.3 m
Floor area	348 m^2
Built Year	1994

The digital cinema projector used is a Sony CineAlta SRX-R220 4K projector, one of the most advanced projectors in digital cinema installations around the world (for more detail on this projector see [5]). The notable guideline relevant for this kind of environment is recommendation ITU-R BT.1686 [6], which provides recommendations on how to perform on-screen measurements of the main projection parameters of large screen digital imagery applications, based on presentation of programs in a theatrical environment. However, the projector's installation, calibration, and maintenance has been performed by Trondheim Kino. Therefore, it did not seem necessary to perform any additional measurements of contrast, screen illumination intensity and uniformity, or any other measurement.

3.2. Test Methods and Conditions

Recommendation ITU-R BT.500-11 [7] provides a thorough guideline for the test methods and the test conditions of subjective visual quality assessments. There are several stimulus viewing sequence methods described in [7], which can be classified into two categories: single stimulus (the subjects are presented with a sequence of test images and are asked to judge the quality of each test image) and double stimulus (the subjects are presented with the reference image and the test image before they are asked to judge the quality of the test image). Differentiating between levels of high quality images requires a test method that possesses a higher discriminative characteristic. Based on the outcome of our pilot test, the most suitable test method is the Simultaneous Double Stimulus test method, in which the subjects are presented with the reference image on the left hand side of the screen and the distorted test image displayed on the right hand side, as illustrated in Figure 2. Subjects judge and vote the quality of the right hand side of the image using the quality scale illustrated in Figure 3.

Fig. 2. Display format of Simultaneous Double Stimulus.

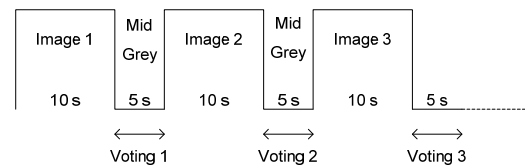
Fig. 3. Ten point quality scale.

Fig. 4. Presentation structure of the test.

A viewing distance of 2H was selected for the experiment [8]. The test was conducted as a single session, and prior to the main session, a training session was conducted. The length of the main session was 15 minutes. The presentation structure of the test is illustrated in Figure 4.

25

3.3. Subjects

A proper evaluation of visual quality requires human subjects with good visual acuity and high concentration, e.g. young persons such as university students. 29 subjects (10 female, 19 male) participated in the evaluation tests performed in this work. Their age ranged from 21 to 32 years old. All subjects reported that they had normal or corrected to normal vision.

3.3. Subjective Data Analysis

In this section, we will show and analyze the subset of Mean Opinion Score (MOS) results from the subjective quality assessment. Here, we consider only data from the stimuli that produced a MOS higher than 4.5. Before processing the resulting data, a post-experiment subject screening was conducted to exclude outliers. The scores of each subject for the reference images were examined; as a result, one subject was excluded because this subject showed randomness due to scoring low for the quality of reference images. In addition, we also used the method described by VQEG [9] to detect outliers.

We averaged the votes for all 28 subjects to obtain the Mean Opinion Score. We disregarded the votes below "fair". Figure 5 illustrates the MOS result with its 95% Confidence Intervals (CI). Our previous study using the same subjective data showed that the behavior of a codec is generally content dependent [3]. However, if we only consider test images that produce MOS higher than 4.5, there is a tendency that at higher compression rates the differences between contents (in terms of MOS score) become minor. We also observed that the images processed in the 0.2 bpp coding condition already produced MOS higher than 7, which means that subjects were satisfied with the image quality regardless the content. When the artifacts of the distorted images are easily noticed due to very low rate values, then the differences of perceived quality among contents become more noticeable, too.
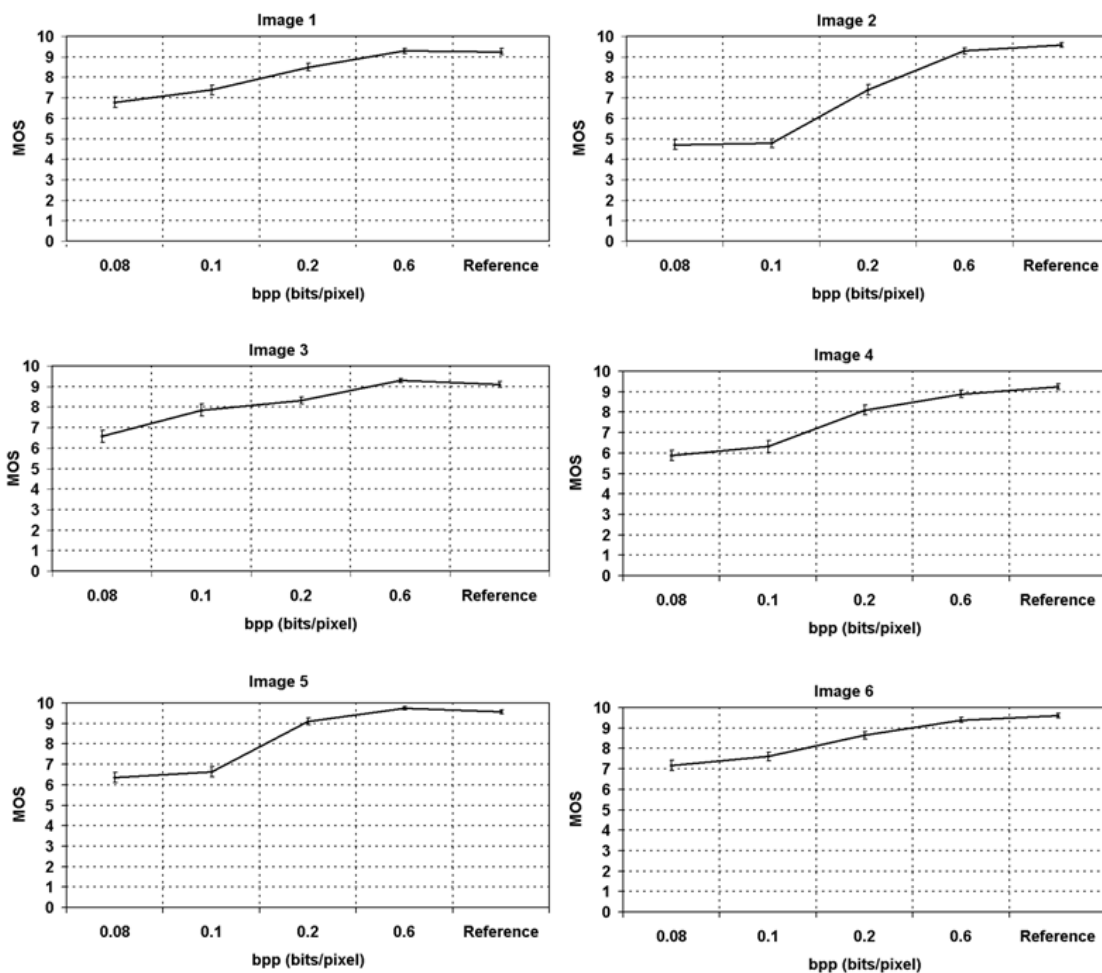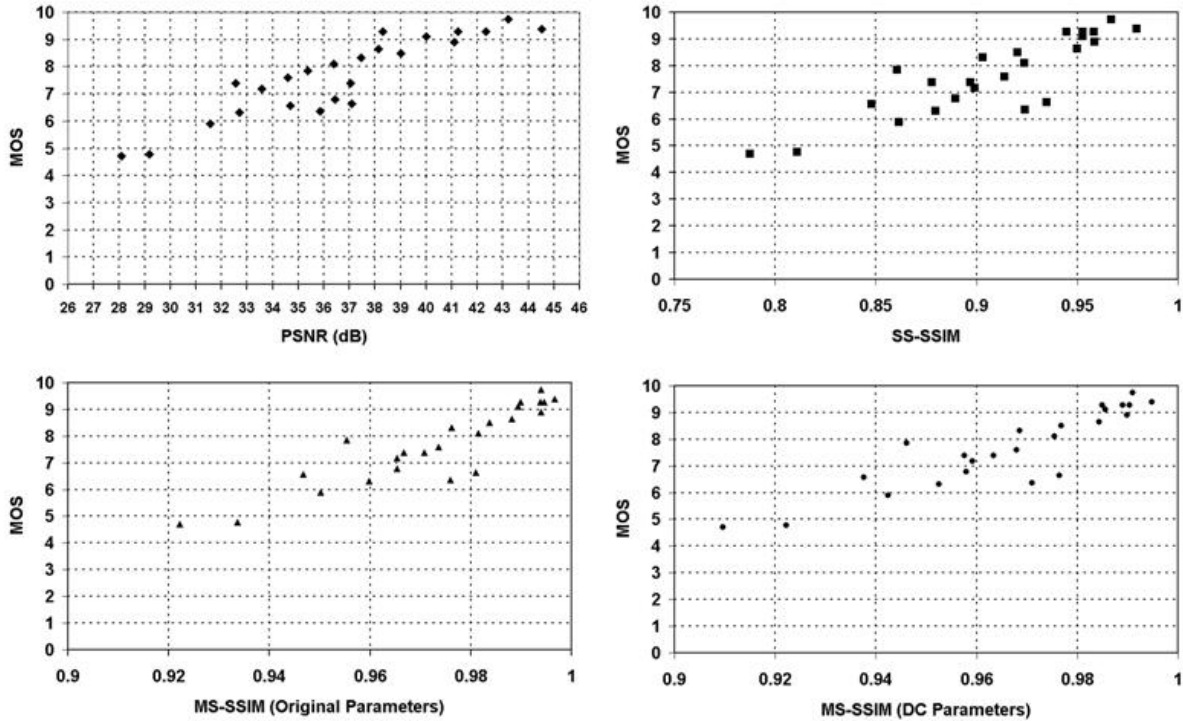

Fig. 5. MOS score of each image vs. bit rate

26

Fig. 6. Scatter plots of MOS vs. metrics result.

4. RESULT

We studied the performance of objective models in [3]. They are PSNR, SS-SSIM, MS-SSIM (using original parameters) and MS-SSIM (using DC parameters). In this section, we evaluate further the performance of these objective models in the higher image-quality range by analyzing a subset of the subjective data. The evaluation is done by statistically comparing the objective measurement data from each model with the subjective data. The scatter plots of MOS versus objective metrics are illustrated in Figure 6. The scatter plots representation provides a practical and intuitive representation of correlation between subjective and objective data. A good correlation among subjective and objective data can be observed from the distribution of the points of the graph along the diagonal that divides the graph in two identical parts. Figure 6 shows that there is no noticeable difference in distribution of the points among scatter plots from each model. It indicates that the performances of all models are similar.

The linear Pearson's correlation coefficient for each metric according to the corresponding MOS score is computed. The respective correlation coefficients are reported in Table 2. Figure 7 shows the Pearson's correlation and their associated 95% confidence intervals for each metric.

To check the significance of the difference between the correlation coefficients, the statistical significance test is conducted. No significant difference between coefficients is used as H_0 hypothesis. The test uses the Fisher-z transformation. The normally distributed statistics Z_N is determined for each comparison and compared against the 95% t-Student value for the two-tail test — t(0.05)=1.96. The calculated Z_N for each correlation coefficient comparison is shown in Table 3. If Z_N is higher than 1.96, there is a statistically significant difference with 0.05 significance level between correlation coefficients. All calculated Z_N values based on MOS are lower than 1.96, which means statistically, there are no significant differences between correlation coefficients of all models.

Table 2. Correlation coefficients.

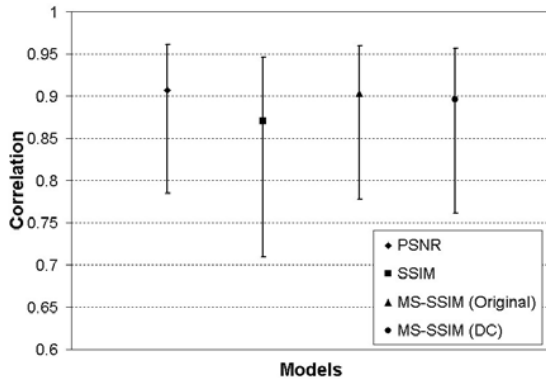| Objective Model	Pearson
PSNR	0.907
SS-SSIM	0.871
MS-SSIM (original parameter)	0.903
MS-SSIM (DC parameters)	0.896

27

Fig. 7. Pearson's correlation coefficient.

Table 3. Significance of the difference between correlation coefficients.

Models Comparison	Z_N
PSNR vs. SS-SSIM	0.56
PSNR vs. MS-SSIM	0.06
PSNR vs. MS-SSIM (DC parameters)	0.19
SS-SSIM vs. MS-SSIM	0.5
SS-SSIM vs. MS-SSIM (DC parameters)	0.36
MS-SSIM vs. MS-SSIM (DC parameters)	0.13

It is important to note that evaluating metrics performance of quality image measure is not a trivial task. In the case of digital cinema applications, the distortion types (JPEG2000 compression) and the image types are limited. Therefore there is a possibility that limited scope of test have effect on approximately statistically equivalent result.

5. CONCLUSION

Based on the calculated correlation coefficient values, the PSNR has the highest correlation with subjective data, followed by MS-SSIM and then SS-SSIM. However, there are no significant differences between correlation coefficients of objective metrics investigated in this paper. Hence, based on this result, there is no objective model that comes out as best performer from a statistical point of view, if we take into account only higher quality data.

In the case of high quality digital cinema content, these results show that structural similarity based metrics do not exhibit the same type of performance that has been reported for lower quality / lower resolution images in the literature, when compared to PSNR metric.

11. REFERENCES

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale Structural Similarity for Image Quality Assessment," *Proc. of 37th IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, 9-12 November 2003.

[3] F.N. Rahayu, U. Reiter, T. Ebrahimi, A. Perkis, and P. Svensson, "SS-SSIM and MS-SSIM for Digital Cinema Applications," *Human Vision and Electronic Imaging XIV Proceeding*, vol. 7240, San Jose, 19-22 January 2009.

[4] DCI Digital Cinema Initiatives, http://www.dcimovies.com.

[5] SONY, "4K Digital Cinema Projectors SRX-R220/SRX-R210 Media Blok LMT-100 Screen Management System LSM-100," *Technical Specification*, 2007.

[6] Recommendation ITU-R BT.1686, "Methods of measurement of image presentation parameters for LSDI programme presentation in a theatrical environment," ITU-R, Geneva, Switzerland, 2004.

[7] Recommendation ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, Geneva, Switzerland, 1974-2002.

[8] Baroncini, V., "Report of the activities at Q2S Centre of Excellence at the Norwegian University of Science and Technology in Trondheim," Internal Report, Q2S-NTNU, 2008.

[9] VQEG, "Multimedia Group Test Plan Version 1.21," http://www.vqeg.org, 2008.

28