EE550 - Image and Video Processing - Lab 5 Report

Baris Sevilmis
Lab Date: 27/11/2019
Due Date: 12/12/2019

1 Overview

Fifth lab assignment of EE550 focuses on Visual Quality Assessment on various Point Cloud Objects and their projected images. Assessment of quality is a complex task, since definition of quality could not be simplified. Rather than a specific definition, quality is generally defined by the task itself. Generalized definitions are relatively superficial, even in case they are correct, but they lack content. Therefore, there are different quality measurement techniques used to measure the quality. Quality assessment will done mainly on two tracks, namely subjective and objective techniques. Corresponding methods and results will be discussed in the following sections with help of various figures and tables. As a last important remark before discussing about the quality, content to be discussed are demonstrated in Figure 1.

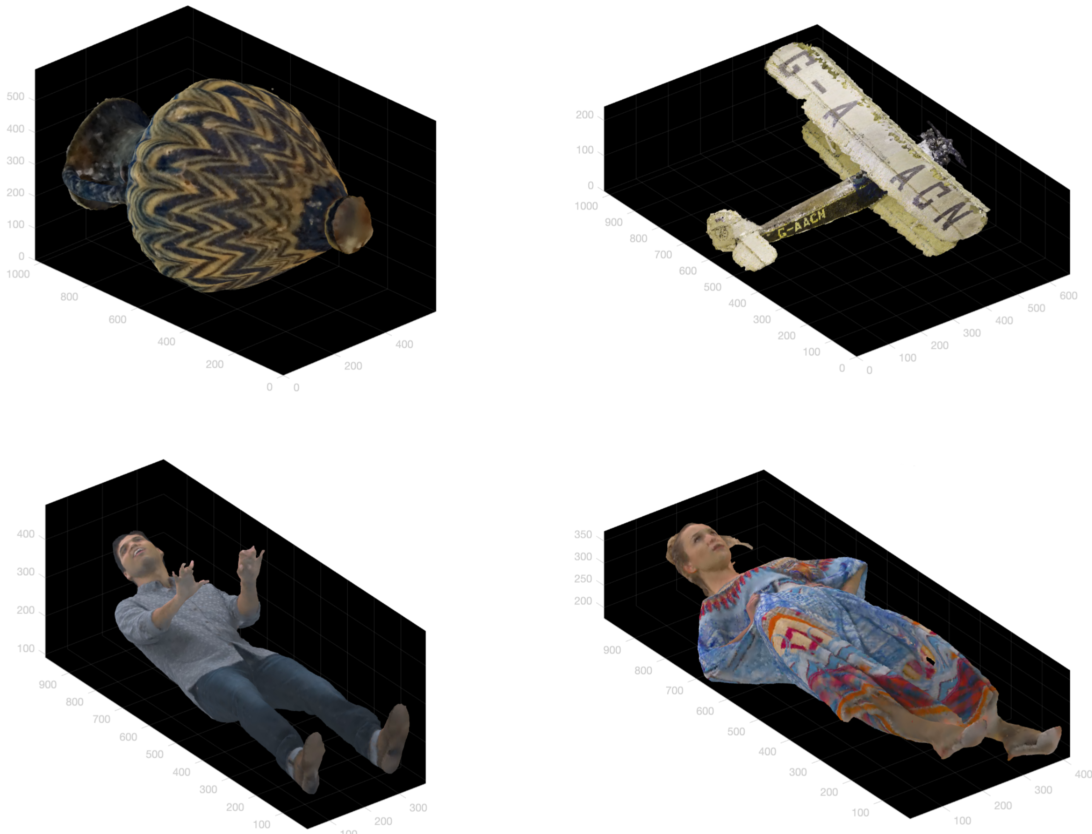

Figure 1: Contents

2 Subjective Quality Assessment

Subjective quality assessment is a method based on human subjects rating the quality or the level of impairment of the the multimedia material, or to submit their preference between two different versions of a multimedia content. To assess quality of content or level of impairment, normally a categorical scale is used, which in our case is $\in [1, 7]$. 7 can be considered as *Excellent* in case of quality assessment or *Imperceptible* in case of level of impairment. On the other hand, 1 corresponds to either *Bad* or *Very Annoying* in consideration of the same cases as above. These grading scale are used during evaluations of a limited group of test subjects. These participants are normally sampled as representatives of end-users. These experiments should have to be done under very specific constraints and very specific environments for the results to be robust, reliable and scalable.

2.1 Outlier Detection

Outlier detection is a corner stone for subjective quality assessment, as human based assessment errors exist no matter the testing environment or testing constraints. Therefore, outlier detection and removal is used as it finds subject scores deviating from the overall scores.
Outlier detection technique is not the same method as it was in the video quality assessment, instead Pearson correlation coefficient is used. Pearson coefficient is defined as following in 1:

$$PLCC(x,y) = \frac{n * \sum_{i=1}^{N} x_i * yi - \sum_{i=1}^{N} x_i * \sum_{i=1}^{N} y_i}{\sqrt{n * \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} * \sqrt{n * \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}} \tag{1}$$

$$r_i(x,y) = \{PLCC(x,y) | i \in [1,2]\} \tag{2}$$

To be more specific, coefficient indicates a correlation level between raw scores of one subject and MOS over all subjects. MOS is explained in a detailed manner in the next subsection. To threshold outliers, two different $PLCC$ scores are calculated. These scores are called r_1 & r_2 as in formula 2. r_1 stands for the MOS of all subjects and the individual score of one subject for the stimulus i, and n is total number of stimuli, whereas r_2 indicate the MOS of all subjects and the MOS of one subject for all variations (i.e., stimuli) of content i, and n is total number of contents. After both r_1 and r_2 are computed, threshold operation happens, in which if both of the values are below a certain distinct threshold, subject is removed as an outlier. In our case, if $r_1 < 0.75$ & $r_2 < 0.8$ corresponding subject had to be removed, and luckily no outliers were found within the subjective scores.

2.2 Mean Opinion Score(MOS) & Confidence Interval(CI)

Mean Opinion Score, namely MOS, is a methodology to combine user based grades for each stimuli. This method provides a reliable subjective score in terms of subject scores. MOS is computed as following:

$$MOS_j = \{\frac{\sum_{i=1}^{N} m_{ij}}{N} | (i,j) \in (U,V)\} \tag{3}$$

where N is the number of subjects and m_{ij} is the score of subject i for the stimulus j. In addition to MOS, the confidence interval, namely CI, is calculated. This provides information upon the relationship between the estimated mean values based on a sample of the population and true mean values of the entire population. Due to a small number of subjects, t distribution is used to compute CI in the following way:

$$CI_j = \{t(1-\alpha/2, N-1) * \frac{\sigma_j}{\sqrt{N}} | j \in [0, N_s]\} \tag{4}$$

where $t(1-\alpha/2, N-1)$ corresponds to a two tailed Student's t-distribution. with $N-1$ degrees of freedom and a desired significance level of α(equal to 1-degree of confidence). N is the number of subjects, and σ_j is the standard deviation of the scores assigned to stimulus j.

In our case, α is picked as 0.05(degree of confidence of 95%), where $100x(1-\alpha)\%$ of the intervals will contain the exact true value. MOS results are plotted with their corresponding CIs for fixed bitrate values. This is done for each of the 4 content separately. In addition, we have for each content all 4 different codecs, which are plotted on the same plot with other codecs with respect to the same content. Resulting MOS and corresponding CI values for each codec and content are depicted in Figure 2.

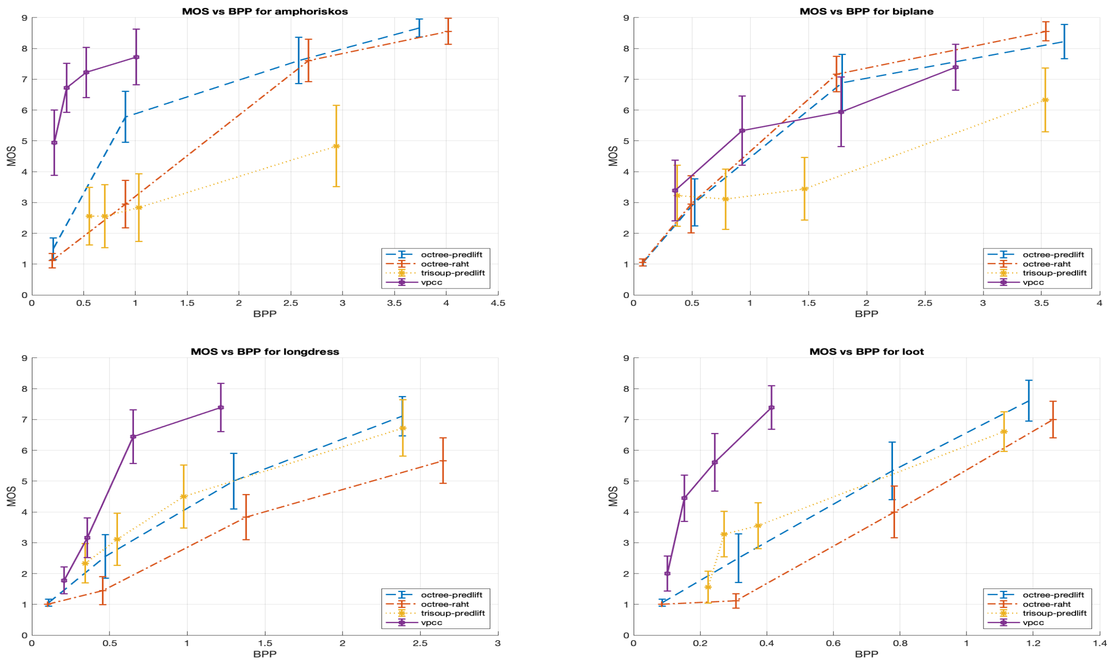

Figure 2: MOS vs Bitrates

MOS and CI results for each content differ from each other in most of the cases, though they follow similar patterns. It would be most reasonable to conclude on performance of VPCC as in each content MOS values have high values. Another interesting fact would be the high performance of VPCC given the low bitrate level. Although *octree-predlift*, second best codec, is performing decent as well, its most optimal level is achieved at higher bitrates in comparison with VPCC. Other codecs are not practicing as well as the mentioned codecs, although depending on content they do show promising results as well, such as in biplane content.

3 Objective Quality Assessment

Subjective quality assessment is a difficult and time-consuming process, both for the test subjects and test makers. Additionally, they can not be applied in case of real-time in-service quality evaluation. Therefore, objective quality assessment methods are produced, which do not heavily rely on human assessment but more focused on quantitative features. There are many different objective metrics available, and some of the are used for the sake of experimentation. These methods are explained in detail in further subsections.

3.1 Point based objective quality metrics

Point based objective quality metrics are mainly based on geometric and color oriented assessments. These methods are briefly discussed in further subsections, and then their results will be shown specific for each content. Mean Squared Error(MSE) and Hausdorff(HAU) metrics are used for the total error computation in case of geometry metrics. For the sake of convenience, our distorted model will be denoted as A and reference model will be denoted as B. As a last detail, all of the point based metrics are done with respect to symmetry, meaning that A & B are both in role of distorted and reference model. Depending on their error metrics, either minimum or the maximum of these values are provided as the resulting score.

$$d_{MSE}^{A,B} = \frac{1}{N_A} \sum_{k=1}^{N_A} (e(a_k, b_i))^2 \tag{5}$$

$$d_{HAU}^{A,B} = \max_{\forall k \in A} (e(a_k, b_i)) \tag{6}$$

3.1.1 Point-To-Point Metric

Point-to-point metrics depend on geometric distances of the corresponding coordinates between the reference and the model under evaluation. Given A is distorted content and B is the reference model, closest neighbors from model B to distorted model A are chosen for each point in A. Then, their euclidean distances are taken into consideration, which produces error:

$$e(a_k, b_i) = \{\sqrt{(a_k - bi)^2} | i \in B, k \in A\} \tag{7}$$

3.1.2 Point-To-Plane Metric

Point-to-plane metrics depend on the projected error across the normal vector of the corresponding reference point. To be more precise, for each point a_k of the distorted model A, its nearest neighbor b_i from the reference model B is chosen. In this case, error is calculated as the following:

$$e(a_k, b_i) = \{(a_k - bi) * \overrightarrow{n_{b_i}} | i \in B, k \in A\} \tag{8}$$

3.1.3 Plane-To-Plane Metric

Plane-to-plane metrics depend on the angular similarity of tangent planes that correspond to associated points between the reference and the model under evaluation. In particular, for each point a_k of the distorted model A, its nearest neighbor b_i from the reference model B is determined. Angle between the normal vectors of a_k and b_i are computed, and minimum of these angles are picked. This is denoted by the angular similarity, which has a value in range of [0, 1]. At the end, following formula is used to compute plane-to-plane error:

$$e(a_k, b_i) = \{1 - \frac{2\theta}{\pi} | \min_{\measuredangle \in (A,B)} \theta\} \tag{9}$$

3.1.4 Color Metric

Color-only metrics use a similar approach as the previous Point based metrics, in which for each point a_k of the distorted model A, its nearest neighbor b_i from the reference model B is determined. The most widely used metric is the PSNR, and there are various PSNR based color only metrics. Computations can be either in RGB or YUV channels of the requested model. For the sake of convenience, following formula will be used to compute errors(in this case, Y channel):

$$e(a_k^Y, b_i^Y) = |a_k^Y - b_i^Y| \tag{10}$$

These error values can be used again to produce formulas 5 and 6. In order to create the actual metric scores, there are various methods. First methodology used to create $PSNR_{RGB}$ score is as below, which utilizes RGB channel errors overall:

$$PSNR_{RGB} = 10 * log_{10} \frac{255^2}{MSE} \tag{11}$$

Second method is as the following where color channels are utilized distinctly:

$$PSNR_{RGB} = (PSNR_R + PSNR_G + PSNR_B)/3 \tag{12}$$

On the other hand, third method utilizes mainly the luminance plane of the model, as well as the U and V channels, by the following formula:

$$PSNR_{YUV} = (6 * PSNR_Y + PSNR_U + PSNR_V)/8 \tag{13}$$

3.1.5 Point based metric results

Figure 3, 4, 5 & 6 are demonstrating Point based objective quality scores for each content. Figures consist of objective metrics in the given order from left to right: Point-to-Point MSE, Point-To-Plane MSE, Plane-To-Plane MSE, Point-To-Point HAU, Point-To-Plane HAU, PSNR-RGB(11), PSNR-RGB(12), PSNR-YUV(13). Each Figure, 3 to 6, depict results of these objective metrics on each content separately. Obviously, HAU(6) based objective metrics are not really reliable. On the other hand, MSE scores and Color-based PSNR scores are as expected and much more reliable. In terms of codecs, outstandingly VPCC ensures best scores, which has high objective scores(Color-based metrics) for relatively lower bitrate values, and provides really small MSE errors again for relatively lower bitrate values. Worst codec seems to be octree-raht, especially for Color-based metrics. Predlift based codecs are in the middle in terms of their performances. For lower bitrate values, trisoup-predlift codec performs better, but for higher bitrate values octree-predlift codec has a better performance. Octree based codecs outperform other codecs given higher bitrate values in any case.

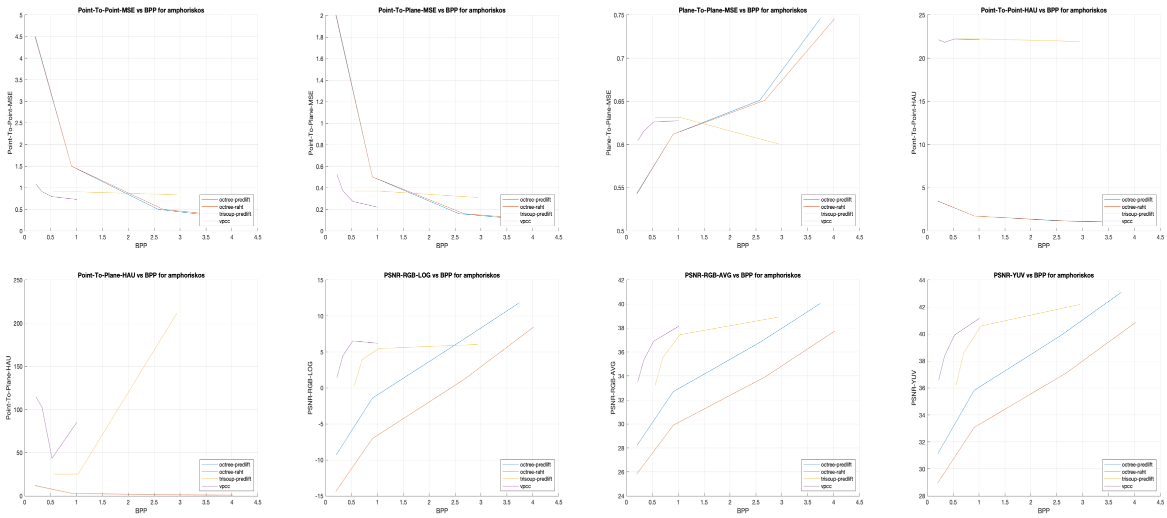

Figure 3: Point based metrics for amphoriskos

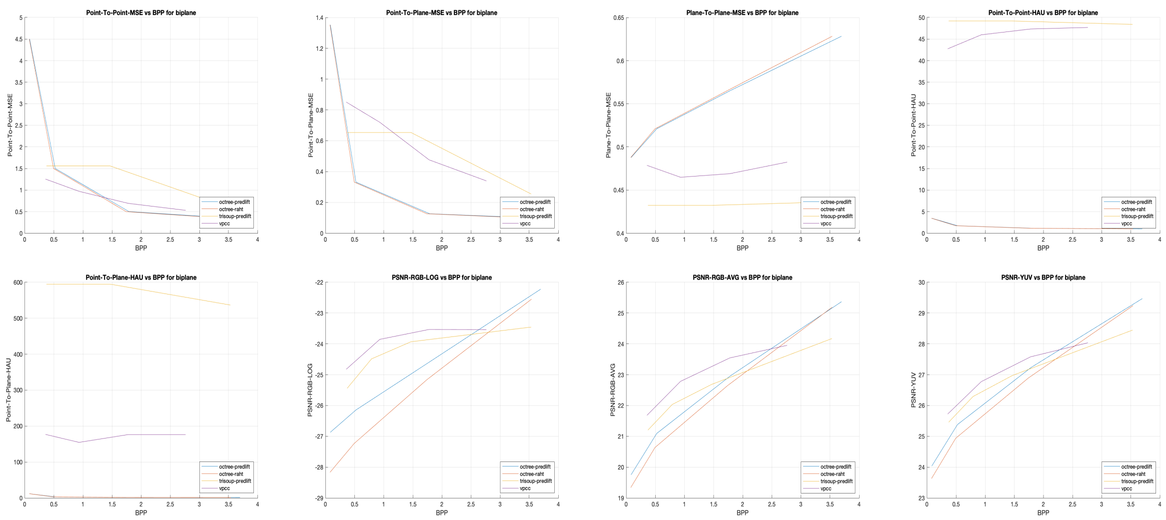

Figure 4: Point based metrics for biplane

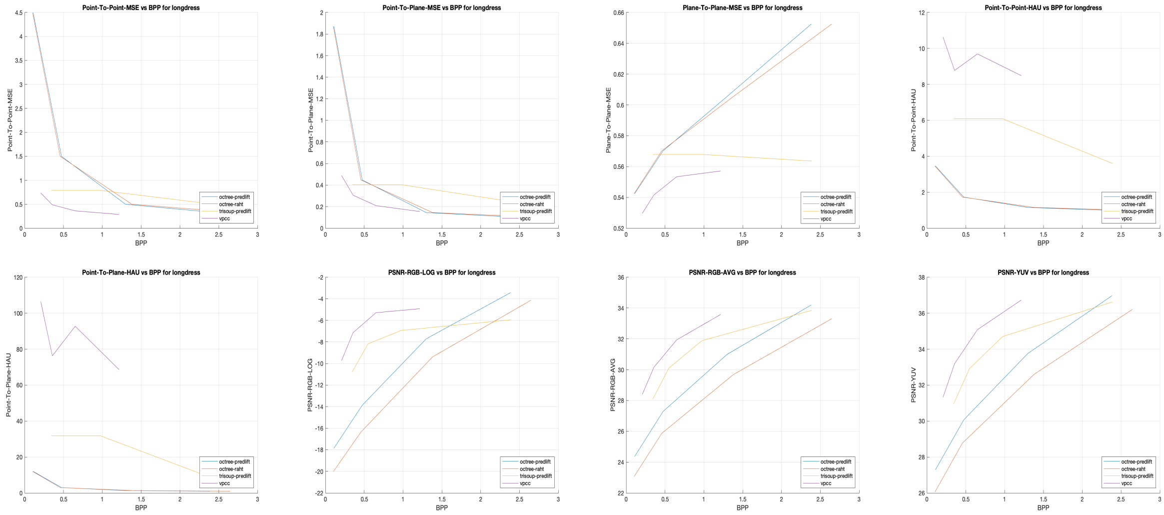

Figure 5: Point based metrics for longdress

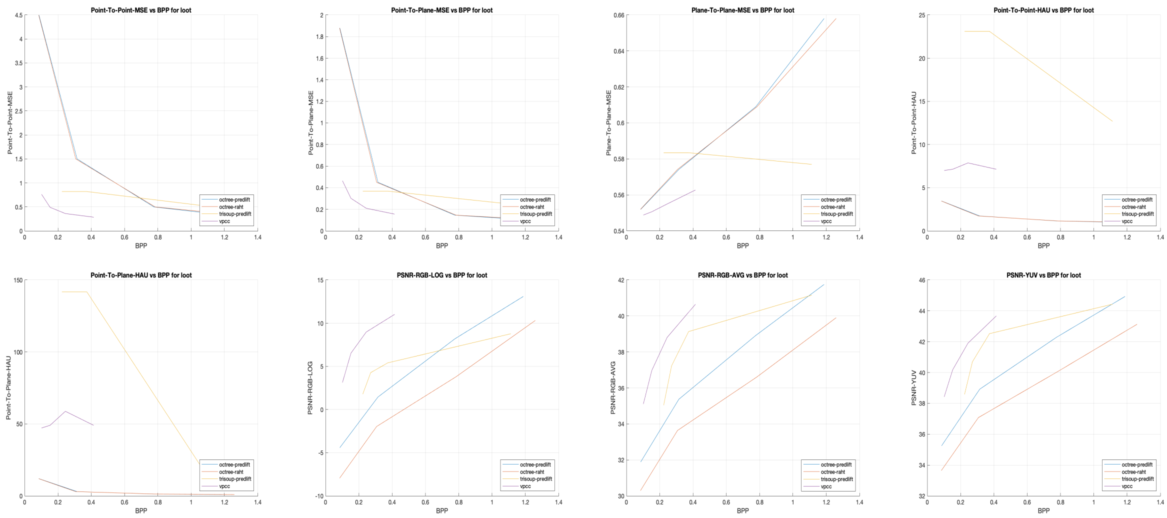

Figure 6: Point based metrics for loot

3.2 Full Reference(FR) Metric

FR metrics is one of the main objective quality assessment methods. In our case, PSNR, SSIM, and MS-SSIM are used as FR metrics.
Peak Signal-To-Noise Ratio, namely PSNR, is a single frame based metric defined as in 14 & 15:

$$PSNR = \{10 * log_{10} \frac{(2^B - 1^2)}{MSE} | B := [1, N](bitdepth)\} \tag{14}$$

$$MSE = \{\frac{1}{MN} \sum_{y=1}^{M} \sum_{x=1}^{N} (Ref(x,y) - Proc(x,y))^2 | (x,y) \in U, V)\} \tag{15}$$

where U,V are the image dimensions. On the other hand, Ref is the reference image and $Proc$ is the processed image. PSNR is a full reference (FR) metric, as both reference and input image are inputs to the method itself.

Structural Similarity Index Metric, namely SSIM, is based on a hypothesis built upon Human Visual Systems, namely HVS, ability to extract structural information from the scenery. Structural information is considered as the features that stand out as the structure of objects in the scene. It is independent of average luminance and contrast. Moreover, distortion in processed image can be estimated by the structural information change between the *Ref* and *Proc*. The similarity measure is conventionally calculated on luminance channel of the image, namely Y. It compares original and distorted images in terms of the luminance, the contrast and the structure. SSIM map is obtained by multiplying the values of three comparison functions, which are the luminance comparison function, contrast comparison function and the structural comparison function. Each of them are computed by calculating correlation between corresponding channels in *Ref* & *Proc*. Final SSIM value is obtained by taking the mean SSIM value by computing the mean over the SSIM map.

The Multi-Scale Structural Similarity Index Metric(MS-SSIM) is an extension of the SSIM to additionally consider the fact that the perceivableness of each frame/image impairments vary depending on the sampling density of the frame/image signal. Operations are same as in the SSIM, however the correlation scores are computed at various spatial scales.

Figure 7, 8, 9 & 10 demonstrate results of requested FR metrics on each content separately. VPCC codec has the best objective score - bitrate tradeoff as in Point-based objective metrics. For low bitrates, objective scores result in higher values in comparison with other codec. For some specific contents such as longdress and loot, VPCC achieves in the highest values among all codecs given same bitrate values. Codec trisoup-predlift seems to be relatively inconsistent in terms of performance and additionally, bitrate does not affect its performance as the other codecs. Overall, its performance can be considered as non-reliable. Octree based codecs on the other hand ensure consistent behaviour among all of the contents. Their performance increase with higher bitrates is noticeable, as well as the cases, in which they overperform VPCC. However, their performance-bitrate tradeoff is not as well as VPCC, as for low bitrates octree based codec perform relatively poor.

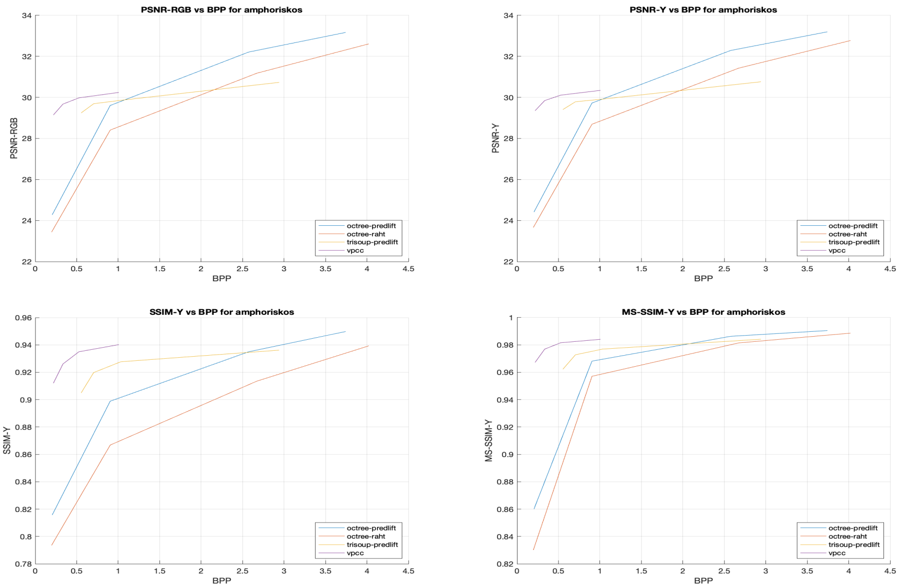

Figure 7: FR metrics for amphoriskos(projected)

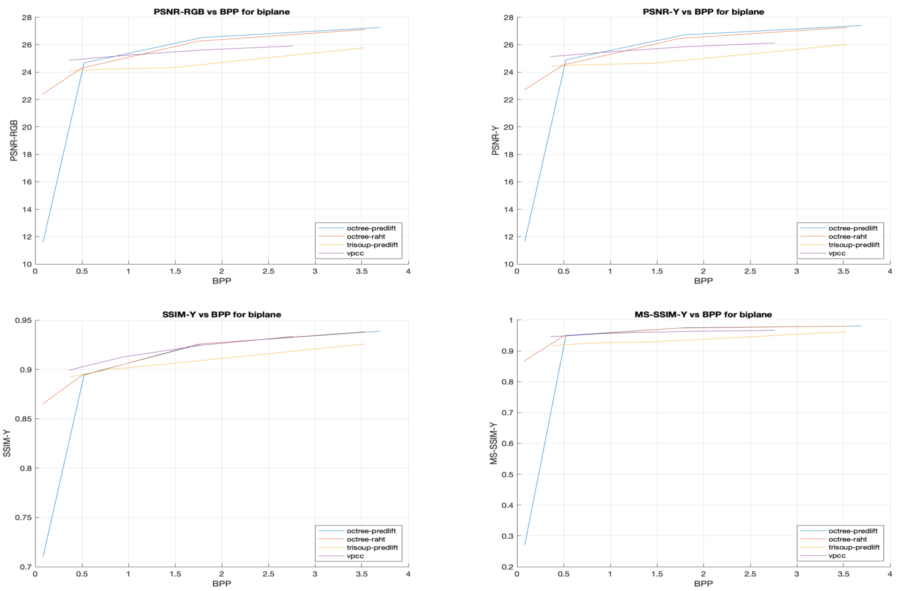

Figure 8: FR metrics for biplane(projected)

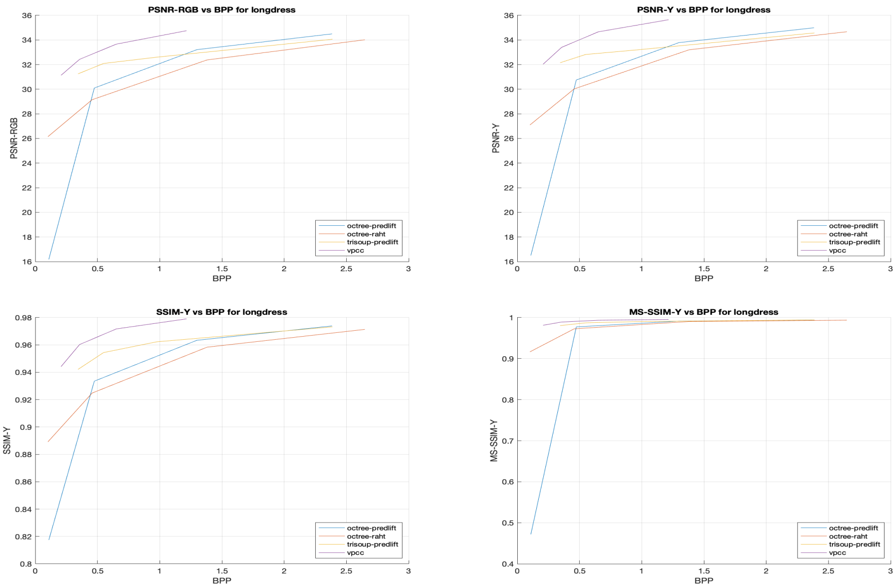

Figure 9: FR metrics for longdress(projected)

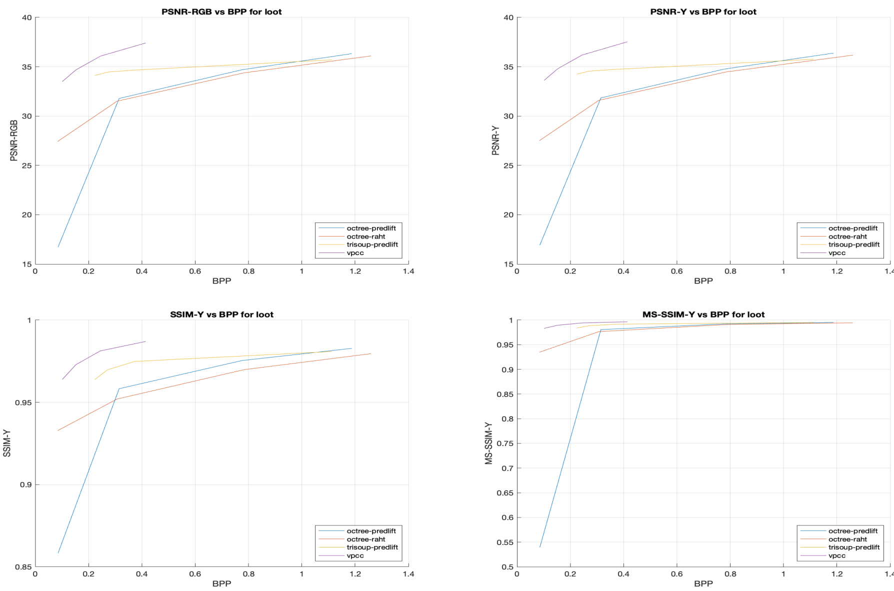

Figure 10: FR metrics for loot(projected)

4 Pearson & Spearman Correlation and RMSE

Pearson, Spearman Coefficients and RMSE are performance metrics for comparison of objective and subjective scores. These resulting correlation and mean squared error score metrics are used to benchmark the performance of objective metrics. There are two attributes used to compare the prediction performance of the different metrics:

Accuracy is the ability of a metrics to predict subjective ratings with the minimum average error. Root Mean Square Error, namely RMSE, is used to compute accuracy as the following:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{n}} \tag{16}$$

where Pearson correlation is defined as:
Pearson correlation results between $[-1, 1]$ indicating strong positive correlation in case of 1 and strong negative correlation in case of -1.

Monotonicity is the second attribute, where Spearman Rank Correlation Coefficient, RC, is used. It describes the relationship between MOS and predicted values.

$$RC = \{1 - \frac{6 * \sum_{i=1}^{N}(R(x_i) - R(y_i))^2}{n * (n^2 - 1)} | x_i, y_i = 1, ..., n\} \tag{17}$$

4.1 No Polynomial Fitting

This subsection concentrates on the raw, no linear or cubic fitting, objective scores versus the subjective MOS scores. Figure 11 demonstrates all of the MOS versus Objective metric scores. Every plot corresponds to one of the different objective metric comparison. On the other hand, comparison of each content for a specific case are plotted on the same graph. Table 1 represents PLCC, SROCC & RMSE scores for each objective metric for all contents. Negative correlative results in terms of MSE, except plane to plane which is a specific case, and positive correlative results for Color-based and FR-based metrics indicate our objective score calculation worked as intended. Obviosly, HAU metrics result poorly as it was depicted in previous Point-based metric score plots and as it was depicted in Figure 10. HAU RMSE scores are another indicator for HAU to be not relied as a metric instead of MSE. On the other hand, best correlative results are obtained by using Point-based MSE metrics, whereas FR-based metrics ensure relatively scalable results. Color-based metrics are promising reliable and fair results in comparison with HAU, however not as well as MSE-based and FR-based metrics. Overall issue seems to be RMSE values tend to be high, which is solved by doing various polynomial fitting. Table 2, 3, and 4 demonstrate the content specific PLCC, SROCC & RMSE values in the respective order. Above mentioned results can be seen in a much more detailed manner. Lastly in terms of projection based metrics, different weighting methods were used, but only a specific one was picked. There exist 6 different projection references and there may be different weighting strategies. Unfortunately, it would take unreasonable amount of time to try all of them to find the best set of parameters. Therefore, weighting technique used consists of non-gray pixel amount within the reference images. Weight of a certain reference is determined by computing the non-gray pixel amount$(! = 192)$ for that reference and multiplying the specific score by that value. At the end, all of these values are combined and divided by the total amount of non-gray pixels in terms of normalization. As a consequence, resulting FR scores prove to be much more reliable. Lastly, as an improvement to these objective metrics, there is one main improvement that would be very beneficial. Instead of MSE, various other error calculation methods should be tried such as Hamiltonian distance. MSE is almost used by each of these metrics, and

HAU is the only different alternative in case of this lab, which is not reliable or robust. As mentioned, different alternatives of MSE should be evaluated as well.

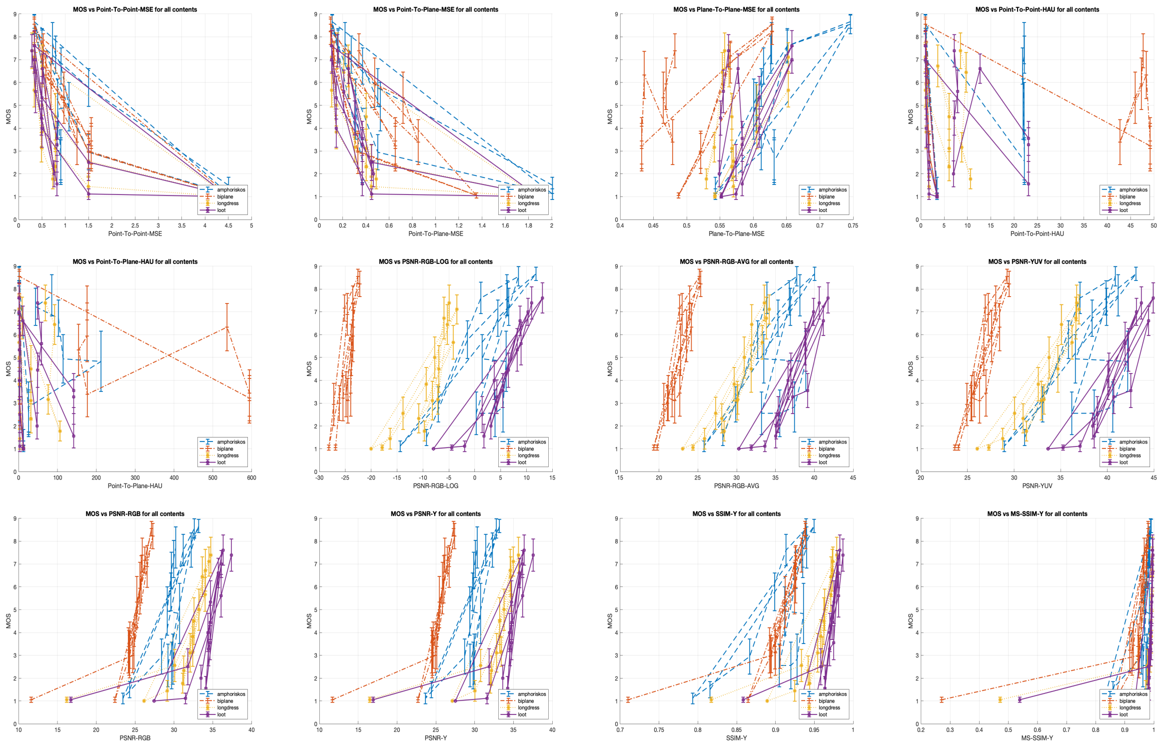

Figure 11: No Fitting: MOS vs Objective Scores

Table 1: No Fitting: Overall

Overall	PLCC	SROCC	RMSE
MSE-Po2Po	-0.6804	-0.7794	4.7061
MSE-Po2Pl	-0.6668	-0.7790	4.8585
MSE-Pl2Pl	0.4326	0.4524	4.5822
HAU-Po2Po	0.0059	-0.2029	17.2914
HAU-Po2Pl	-0.0425	-0.1981	158.6558
Color-PSNR-OA	0.2743	0.3533	16.7103
Color-PSNR-AVG	0.3522	0.3717	27.1506
Color-PSNR-YUV	0.3696	0.3849	30.3933
FR-PSNR-RGB	0.4241	0.4053	25.4700
FR-PSNR-Y	0.4119	0.3936	25.7778
FR-SSIM	0.4781	0.4455	4.2841
FR-MSSSIM	0.4285	0.5761	4.2565

Table 2: No Fitting: PEARSON

PEARSON	amphoriskos	biplane	longdress	loot
MSE-Po2Po	-0.6962	-0.8424	-0.6658	-0.6635
MSE-Po2Pl	-0.6985	-0.8257	-0.6633	-0.6478
MSE-Pl2Pl	0.7262	0.5737	0.4955	0.5486
HAU-Po2Po	-0.1621	-0.0433	0.0100	-0.1531
HAU-Po2Pl	0.0324	-0.1746	0.0560	-0.1972
Color-PSNR-OA	0.7474	0.8209	0.8671	0.9292
Color-PSNR-AVG	0.7017	0.9297	0.9338	0.9272
Color-PSNR-YUV	0.7007	0.9280	0.9376	0.9161
FR-PSNR-RGB	0.8067	0.6880	0.7087	0.6610
FR-PSNR-Y	0.8096	0.6764	0.6915	0.6610
FR-SSIM	0.7390	0.7106	0.7403	0.6583
FR-MSSSIM	0.6950	0.5373	0.4346	0.4138

Table 3: No Fitting: SPEARMAN

SPEARMAN	amphoriskos	biplane	longdress	loot
MSE-Po2Po	-0.8119	-0.9297	-0.8654	-0.9231
MSE-Po2Pl	-0.8089	-0.7372	-0.8595	-0.8994
MSE-Pl2Pl	0.6459	0.3523	0.3654	0.4675
HAU-Po2Po	-0.5078	-0.3020	-0.0784	-0.1880
HAU-Po2Pl	-0.3289	-0.3437	-0.1524	-0.2857
Color-PSNR-OA	0.7717	0.8271	0.9412	0.9794
Color-PSNR-AVG	0.6642	0.9257	0.9647	0.9441
Color-PSNR-YUV	0.6642	0.9345	0.9706	0.9412
FR-PSNR-RGB	0.8247	0.9433	0.9824	0.9706
FR-PSNR-Y	0.8439	0.9492	0.9676	0.9618
FR-SSIM	0.7717	0.9639	0.9824	0.9647
FR-MSSSIM	0.8409	0.9051	0.9765	0.9676

Table 4: No Fitting: RMSE

RMSE	amphoriskos	biplane	longdress	loot
MSE-Po2Po	5.2621	4.8933	4.2329	4.3632
MSE-Po2Pl	5.5392	5.0725	4.2930	4.4247
MSE-Pl2Pl	5.2088	4.9079	3.9888	4.1057
HAU-Po2Po	12.8113	30.4404	3.9579	9.4633
HAU-Po2Pl	67.7108	300.0540	43.4664	64.6545
Color-PSNR-OA	6.1080	29.5986	13.8019	3.6138
Color-PSNR-AVG	29.5594	17.6526	25.8960	33.0551
Color-PSNR-YUV	32.6377	21.8404	28.7965	36.3807
FR-PSNR-RGB	24.4433	19.7349	27.3512	29.3234
FR-PSNR-Y	24.5819	19.9715	28.0994	29.4155
FR-SSIM	4.9704	4.5620	3.6742	3.7946
FR-MSSSIM	4.9249	4.5306	3.6564	3.7845

4.2 Linear Fitting

Linear fitting is the first polynomial fitting used on the raw objective scores. Figure 12 demonstrates the results of MOS values versus linearly fitted Objective scores. As in the previous lab assignment, linear fitting causes resulting PLSS, SROCC & RMSE scores to be positive as well as MOS ve Objective result metrics. As before, HAU outcomes prove to unreliable. Negative correlation is represented as positive correlation in case of linear fitting, and for this reason MSE based Point to Point and Point to Plane have the same change in terms of increase instead of decrease. In fact, Table 5, 6, 7 & 8 are establish an evidence for these plots, in which PLCC, SROCC & RMSE scores are given. As before, Table 5 have the overall results for all the metrics, whereas Table 6 has the PLCC score, Table 7 has SROCC scores, and Table 8 has the RMSE scores. Table 6, 7 & 8 should be used to check resulting values in a more detailed manner since the resulting values are content specific. In terms of comparison, Table 5 would be more than enough. In the first instance main difference appears to be decrease in RMSE scores and the negative to positive conversion in case of MSE based Point to Point and Point to Plane scores. Decrease in RMSE values, especially the reliable and scalable metrics as have been mentioned in the previous sections, denote linear fitting works in favor metrics. Best set of metrics to consider would be again MSE-based(geometric) and FR metrics because of the correlative and error results and Color-based metrics can be considered as well. HAU should not be considered as a reliable metric. In terms of weighting, same method as in no fitting was used, as it ensured the best FR metric scores. In terms of improvement, it would be same as in no fitting case. In other words, different error metrics do need to be evaluated as well instead of MSE.

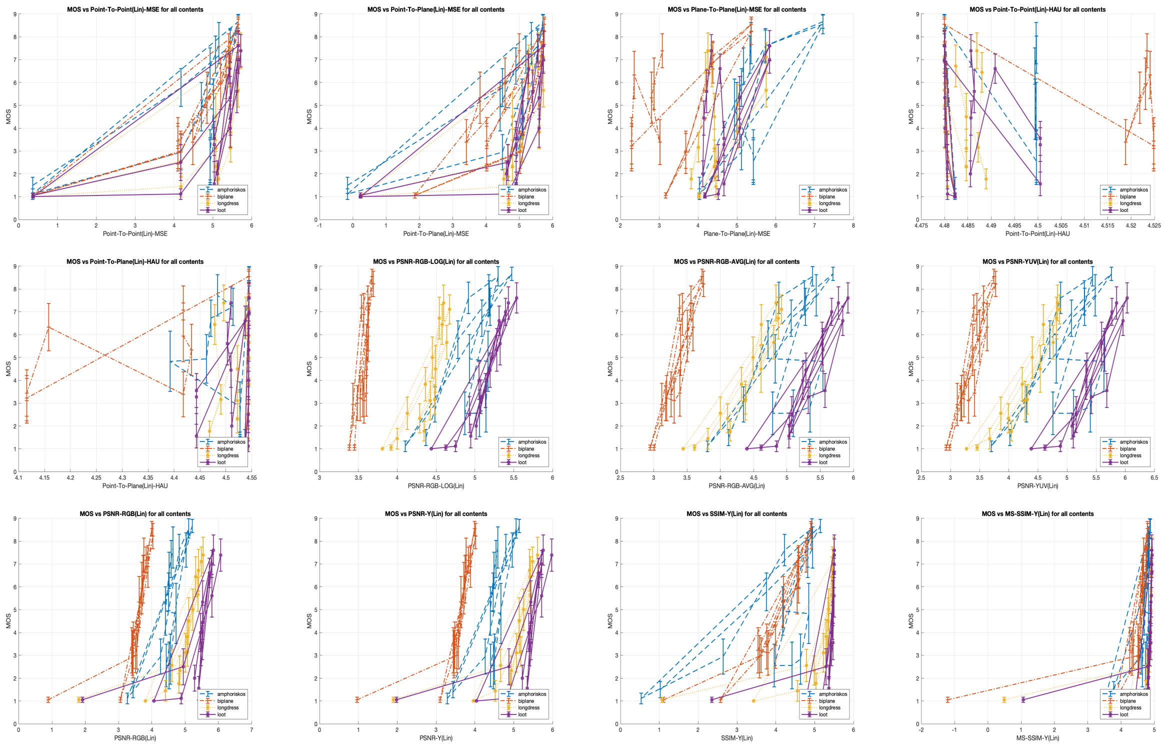

Figure 12: Linear Fitting: MOS vs Objective Scores

Table 5: Linear Fitting Overall

Overall	PLCC	SROCC	RMSE
MSE-Po2Po	0.6804	0.7794	1.7617
MSE-Po2Pl	0.6668	0.7790	1.7915
MSE-Pl2Pl	0.4326	0.4524	2.1672
HAU-Po2Po	0.0059	-0.2029	2.4037
HAU-Po2Pl	0.0425	0.1981	2.4016
Color-PSNR-OA	0.2743	0.3533	2.3116
Color-PSNR-AVG	0.3522	0.3717	2.2498
Color-PSNR-YUV	0.3696	0.3849	2.2335
FR-PSNR-RGB	0.4241	0.4053	2.1768
FR-PSNR-Y	0.4119	0.3936	2.1904
FR-SSIM	0.4967	0.4452	2.0863
FR-MSSSIM	0.4285	0.5761	2.1718

Table 6: Linear Fitting: PEARSON

PEARSON	amphoriskos	biplane	longdress	loot
MSE-Po2Po	0.6962	0.8424	0.6658	0.6635
MSE-Po2Pl	0.6985	0.8257	0.6633	0.6478
MSE-Pl2Pl	0.7262	0.5737	0.4955	0.5486
HAU-Po2Po	-0.1621	-0.0433	0.0100	-0.1531
HAU-Po2Pl	-0.0324	0.1746	-0.0560	0.1972
Color-PSNR-OA	0.7474	0.8209	0.8671	0.9292
Color-PSNR-AVG	0.7017	0.9297	0.9338	0.9272
Color-PSNR-YUV	0.7007	0.9280	0.9376	0.9161
FR-PSNR-RGB	0.8067	0.6880	0.7087	0.6610
FR-PSNR-Y	0.8096	0.6764	0.6915	0.6610
FR-SSIM	0.7287	0.8765	0.6274	0.4821
FR-MSSSIM	0.6950	0.5373	0.4346	0.4138

Table 7: Linear Fitting: SPEARMAN

SPEARMAN	amphoriskos	biplane	longdress	loot
MSE-Po2Po	0.8119	0.9297	0.8654	0.9231
MSE-Po2Pl	0.8089	0.7372	0.8595	0.8994
MSE-Pl2Pl	0.6459	0.3523	0.3654	0.4675
HAU-Po2Po	-0.5078	-0.3020	-0.0784	-0.1880
HAU-Po2Pl	0.3289	0.3437	0.1524	0.2857
Color-PSNR-OA	0.7717	0.8271	0.9412	0.9794
Color-PSNR-AVG	0.6642	0.9257	0.9647	0.9441
Color-PSNR-YUV	0.6642	0.9345	0.9706	0.9412
FR-PSNR-RGB	0.8247	0.9433	0.9824	0.9706
FR-PSNR-Y	0.8439	0.9492	0.9676	0.9618
FR-SSIM	0.7717	0.9639	0.9824	0.9647
FR-MSSSIM	0.8409	0.9051	0.9765	0.9676

Table 8: Linear Fitting: RMSE

RMSE	amphoriskos	biplane	longdress	loot
MSE-Po2Po	1.9857	1.4372	1.7564	1.8222
MSE-Po2Pl	1.9745	1.5886	1.7363	1.8440
MSE-Pl2Pl	2.0270	2.4433	2.0206	2.1508
HAU-Po2Po	2.6390	2.3976	2.2176	2.3410
HAU-Po2Pl	2.6325	2.3873	2.2293	2.3391
Color-PSNR-OA	2.3025	2.6263	1.9776	2.2943
Color-PSNR-AVG	2.2240	2.6097	1.7926	2.2968
Color-PSNR-YUV	2.1936	2.5972	1.7347	2.3211
FR-PSNR-RGB	2.2643	2.3850	1.8569	2.1659
FR-PSNR-Y	2.2887	2.3881	1.9154	2.1406
FR-SSIM	2.1908	1.8291	1.9240	2.3586
FR-MSSSIM	2.3625	2.1076	2.0335	2.1700

4.3 Cubic Fitting

Cubic fitting is the last fitting method, where polynomial fitting is of degree 3 instead of 1 as in linear fitting case. Figure 13 demonstrates the results of MOS values versus cubic fitted Objective scores. Table 9 depicts the Objective metrics PLCC, SROCC & RMSE scores over all of the contents. Moreover, Table 10, 11 & 12 shows PLCC, SROCC & RMSE scores for each specific content distinctly, therefore provides a better overview of these results. In case of cubic fitting, PLCC, SROCC & RMSE results got better for each metric. However, HAU metrics still lack reliable results. On the other hand, MSE based metrics, especially Point to Point and Point to Plane, are now highly positive correlated(change from negative to positive holds also in cubic fitting case). FR metric and Color-based metric results also indicate higher correlation. RMSE scores are way less than the RMSE scores in No Fitting case, as well as the linear fitting case. Most precise weighting method is as before in no fitting and linear fitting, to be more specific non-gray pixel amount is the weighting feature. Lastly, as an improvement, different error metrics can be used as well instead of MSE and HAU. In fact, a better error metric may affect highly correlated results to an even higher correlation level, in which most reliable Objective metric can be selected without a hesitation. In addition, weighting method can changed, as it seems PLCC, SROCC & RMSE outcomes for FR metrics are not affected of fitting as much as Point and Color based metrics. Therefore, a different weighting technique might affect FR-based Objective metrics in a positive manner, for the reason of using FR metrics in addition to Point based metrics as an additional precise metric set.

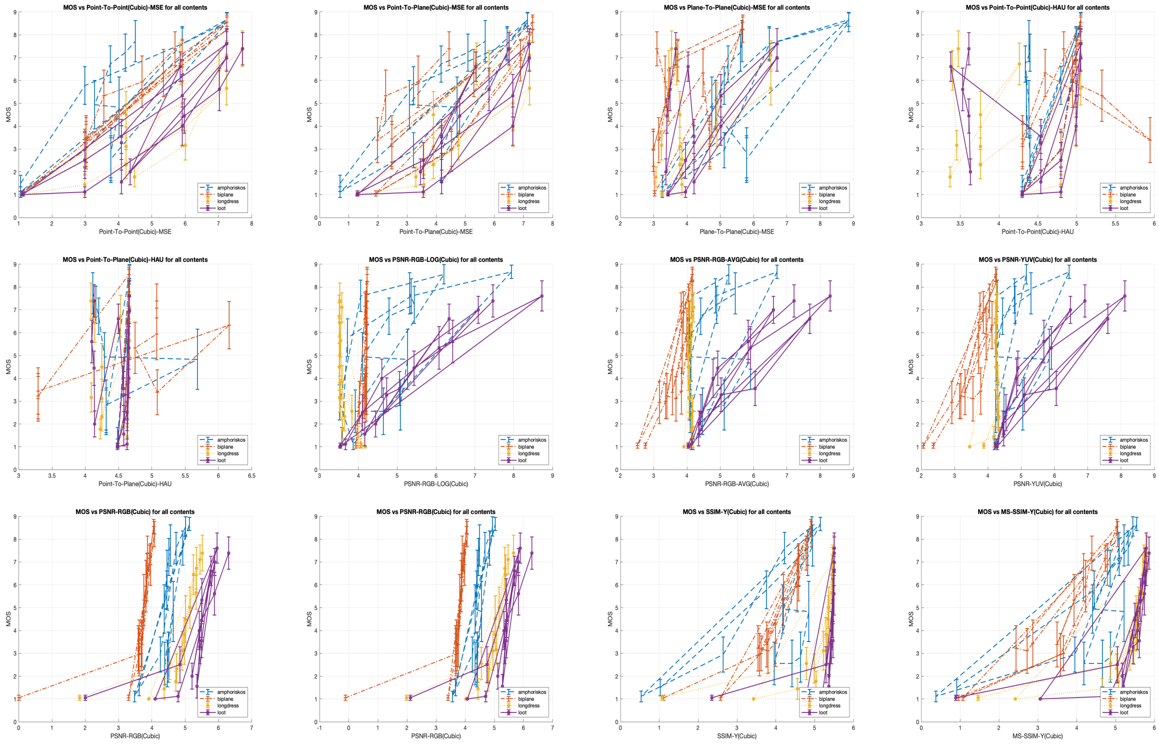

Figure 13: Cubic Fitting: MOS vs Objective Scores

Table 9: Cubic Fitting: Overall

Overall	PLCC	SROCC	RMSE
MSE-Po2Po	0.8141	0.7879	1.3958
MSE-Po2Pl	0.7925	0.7790	1.4659
MSE-Pl2Pl	0.5582	0.5262	1.9944
HAU-Po2Po	0.2301	0.4057	2.3393
HAU-Po2Pl	0.1797	0.3044	2.3646
Color-PSNR-OA	0.4722	0.4303	2.1189
Color-PSNR-AVG	0.4406	0.3485	2.1579
Color-PSNR-YUV	0.4366	0.3030	2.1625
FR-PSNR-RGB	0.4288	0.4053	2.1715
FR-PSNR-Y	0.4191	0.3936	2.1825
FR-SSIM	0.4967	0.4452	2.0863
FR-MSSSIM	0.5943	0.5752	1.9331

Table 10: Cubic Fitting: PEARSON

PEARSON	amphoriskos	biplane	longdress	loot
MSE-Po2Po	0.8359	0.9840	0.8778	0.9118
MSE-Po2Pl	0.8493	0.8115	0.8448	0.8755
MSE-Pl2Pl	0.7280	0.4931	0.4799	0.5576
HAU-Po2Po	0.6196	0.4379	0.0415	-0.0968
HAU-Po2Pl	0.0477	0.4016	-0.0399	-0.1048
Color-PSNR-OA	0.6672	0.7300	-0.7212	0.9620
Color-PSNR-AVG	0.5199	0.8875	0.1950	0.9270
Color-PSNR-YUV	0.4984	0.8965	0.4327	0.9100
FR-PSNR-RGB	0.8175	0.5805	0.7188	0.7125
FR-PSNR-Y	0.8253	0.5486	0.7265	0.7300
FR-SSIM	0.7287	0.8765	0.6274	0.4821
FR-MSSSIM	0.7424	0.9189	0.6966	0.6337

Table 11: Cubic Fitting: SPEARMAN

SPEARMAN	amphoriskos	biplane	longdress	loot
MSE-Po2Po	0.8119	0.9800	0.8654	0.9231
MSE-Po2Pl	0.8089	0.7372	0.8595	0.8994
MSE-Pl2Pl	0.6459	0.6070	0.3654	0.4675
HAU-Po2Po	0.6359	0.6573	0.0784	0.0666
HAU-Po2Pl	0.2104	0.5689	0.0607	0.1643
Color-PSNR-OA	0.7187	0.8300	-0.7794	0.9794
Color-PSNR-AVG	0.4978	0.9257	0.0941	0.9412
Color-PSNR-YUV	0.5523	0.9345	0.0647	0.9382
FR-PSNR-RGB	0.8247	0.9433	0.9824	0.9706
FR-PSNR-Y	0.8439	0.9492	0.9676	0.9618
FR-SSIM	0.7717	0.9639	0.9824	0.9647
FR-MSSSIM	0.8409	0.9051	0.9765	0.9676

Table 12: Cubic Fitting: RMSE

RMSE	amphoriskos	biplane	longdress	loot
MSE-Po2Po	1.8203	0.9377	1.4083	1.2714
MSE-Po2Pl	1.5863	1.5469	1.3709	1.3440
MSE-Pl2Pl	1.8428	2.2754	1.8925	1.9381
HAU-Po2Po	2.4496	2.2188	2.2344	2.4439
HAU-Po2Pl	2.6385	2.2042	2.2218	2.3683
Color-PSNR-OA	2.0170	2.4066	2.3303	1.6333
Color-PSNR-AVG	2.2893	2.2896	2.1439	1.8831
Color-PSNR-YUV	2.3328	2.2478	2.0815	1.9695
FR-PSNR-RGB	2.3278	2.3672	1.8324	2.1172
FR-PSNR-Y	2.3712	2.3735	1.8652	2.0781
FR-SSIM	2.1908	1.8291	1.9240	2.3586
FR-MSSSIM	1.9461	1.8248	1.8827	2.0703