EE550 - Image and Video Processing - Lab 4 Report

Baris Sevilmis
Lab Date: 13/11/2019
Due Date: 28/11/2019

1 Overview

Third lab assignment of EE550 focuses on Visual Quality Assessment on various videos. 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 lack content. Therefore, there are different quality measurement techniques used to measure the videos by usually comparing it to its reference video. These reconstructed videos, namely quality measured videos, are obtained by different algorithms, which are to be mentioned in further sections. However, our main task is to compare these algorithms rather than explaining them, and compute their relative scores and efficiencies. Further sections & subsections provide more information on subjective and objective quality assessment techniques.

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 is $\in [1, 5]$. 5 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 Spatial & Temporal Perceptual Information(SI & TI)

Spatial and Temporal Perceptual Information are essential parameters for Subjective Quality assessment, namely SI and TI values. They are used to determine the possible amount compression and the suffered level of impairment during the scene transmission over a fixed-rate digital transmission service channel.
SI is constructed upon Sobel filtering, but not as in Edge detection. There is no thresholding operation. Each video frame (Luminance plane) F_n is filtered with Sobel filter. Therefore, video frames are first converted from RGB to YUV, and Y channel, namely 2D Luminance plane, is used for filtering. Then, standard deviation of each of these filtered frames are calculated, and this results in a time series of spatial information. Maximum of these time series is picked as final SI value.

$$SI = \{max_n(\sigma(std(F_n)))|F_n \in [0, N]\} \quad (1)$$

where N is the total number of video frames.
On the other hand, TI is constructed upon the motion difference property, $M_n(i, j)$, which is the difference between the pixel values, again 2D Luminance planes are used, between consecutive frames of the video. Location of these pixels in image spaces are same. $M_n(i, j)$ is defined as following:

$$M_n(i, j) = \{F_n(i, j) - Fn - 1(i, j)|(i, j) \in (U, V)\&n \in [0, N]\} \quad (2)$$

where U,V are the frame dimensions, and N is the total amount of frames.
As in SI, standard deviation of each $M_n(i, j)$ are computed over time, and their maximum value is chosen as TI. TI is defined as following:

$$TI = \{max_n(std(M_n(i, j)))|(i, j) \in (U, V)\&n \in [0, N]\} \quad (3)$$

These methods are used to evaluate possible compression rates and level of impairment for each of our 5 contents. SI and TI values are demonstrated on the bar graph in Figure 1. Results make sense, as TI values in high in highly motion, camera itself or video content, videos have higher values as well, f.e Campfire, but lower in videos such as TreeShades. On the other hand, SI values are higher in videos with more features in terms of edge and corner points such as Runners TreeShades but have lower values videos containing less features such as Campfire.

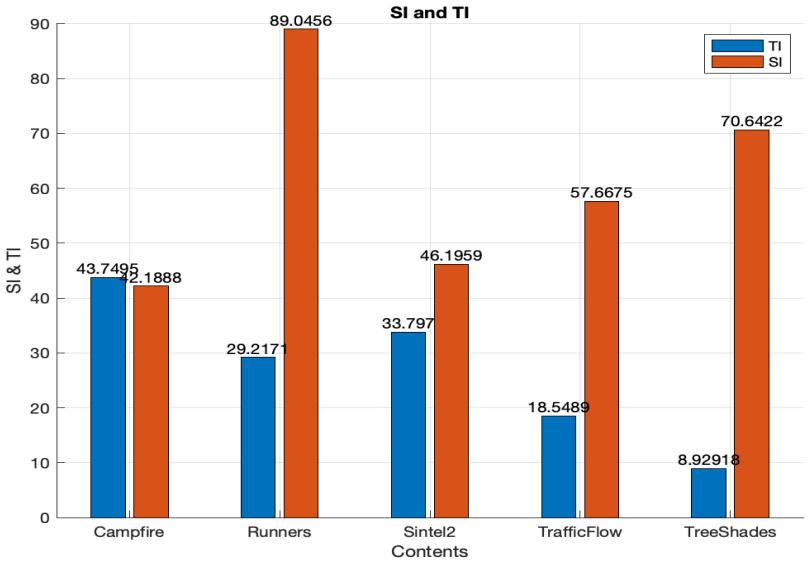

Figure 1: SI & TI values for Reference Contents

2.2 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.
To be more specific, if subjects have kurtosis coefficient of a subject is in range of $[2, 4]$, then distribution is considered as a Gaussian distribution. In order to reject a subject & remove its corresponding scores, a confidence interval is determined. If scores of a specific subject are distributed normally, for each score larger than 2σ from the mean of the scores of a test sequence i, P_i,a counter, is incremented by 1. For each score less than 2σ from the mean, Q_i, another counter, is incremented by 1. If the distibution is non-gaussian, then $\sqrt{20}\sigma$ is used as the limit instead of 2σ, and procedure is the same. In conclusion, considering total number of stimuli N_s, scores of a subject are removed if the following formulas 4 & 5 are satisfied:

$$\{\frac{\sum_{i=1}^{N_s}(P_i + Q_i)}{N_s} > 0.05 | i \in [0, N_s]\} \tag{4}$$

$$\{\frac{\sum_{i=1}^{N_s}(P_i - Q_i)}{\sum_{i=1}^{N_s}(P_i + Q_i)} > 0.05 | i \in [0, N_s]\} \tag{5}$$

In our case, there are 23 subjects that are tested, and at the end, only 19 of them remain. In other words, there are 4 outliers.

2.3 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:

$$MOS_j = \{\frac{\sum_{i=1}^{N} m_{ij}}{N} | (i, j) \in (U, V)\} \tag{6}$$

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:

$$CI_j = \{t(1 - \alpha/2, N - 1) * \frac{\sigma_j}{\sqrt{N}} | j \in [0, N_s]\} \tag{7}$$

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 5 content seperately. In addition, we have for each content all 4 different codecs, which are plotted on the same

plot with other codecs with respct to the same content. Resulting MOS and corresponding CI values for each codec and content are depicted in Figure 2.

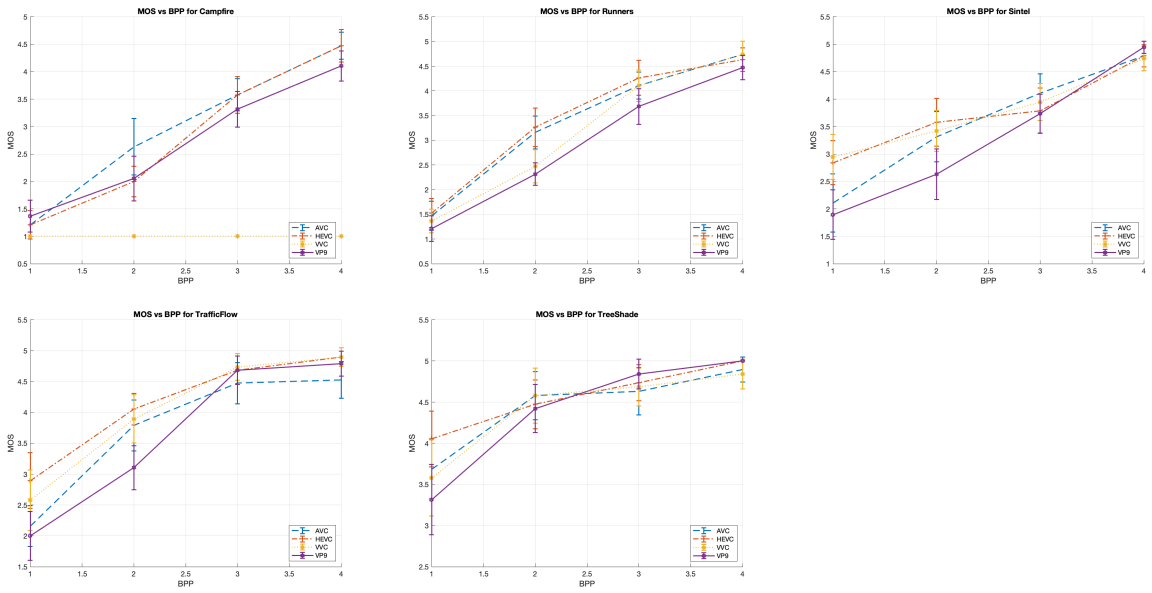

Figure 2: MOS vs Bitrates

MOS and CI results for each content differ from each other in small manners. From the overall perspective, there isn't a best codec as some codecs perform better than others in some of the contents. For the sake of generalization, it would be reasonable to conclude that codec HVEC is among the best, and VP9 is vice versa. In most of the contents, HVEC provides more either best or the second best result, and its efficiency is preserved at higher/lower bitrates as well. On the other hand, excluding high bitrates of TreeShades, VP9 does not produce as reliable and robust MOS and CI as the other codecs. AVC ensures nearly as successful MOS and CI as HVEC for each distinct content as well.

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 Full Reference(FR) metrics-PSNR,SSIM,MS-SSIM

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 8 & 9:

$$PSNR = \{10 * log_{10}\frac{(2^B - 1^2)}{MSE} | B := [1, N](bitdepth)\} \tag{8}$$

$$MSE = \{\frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}(Ref(x,y) - Proc(x,y))^2 | (x,y) \in U, V)\} \tag{9}$$

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 3, 4, 5, 6, & 7 demonstrate the plots of mentioned FR metrics against different bitrate values. Both Y & RGB channels are used for the sake of experimentation. All of these Figures represent FR metric results on certain content, and for a specific content all the mentioned FR metrics both on Y & RGB frames are depicted. As expected, FR metrics have higher values for higher bitrate values. Reasoning shines out such that at higher bitrate values *Proc* are much more similar to its corresponding *Ref*. For the sake of comparison, starting with best codec performance would be more simple, although it is hard to determine considering all the contents and FR metrics both for Luminance plane(Y) and RGB based images. For contents, Runners, Trafficflow and TreeShades, VVC performs really well. All these mentioned contents have low TI values, therefore it could be inferred that VVC performs well on videos with low level of motion. However, as TI value increases VVC performs very poor as its core decreases. VP9 has relatively low FR scores in each case compared to HVEC and AVC. However, it performs better than VVC given high TI content. On the other hand, HVEC and AVC perform really well for each content, suggesting they have low correlation with video motion itself. They result in reliable FR scores for each content such that they are not only resulting in high scores but similar score increase for specific bitrate values in each content.

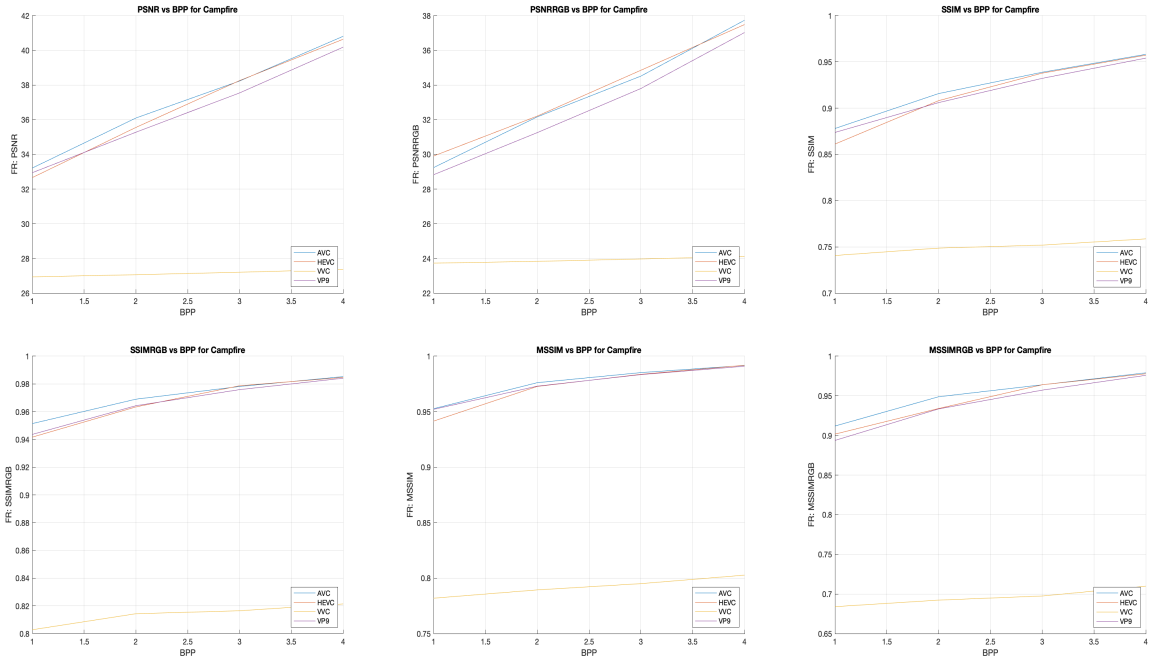

Figure 3: FR metrics for CampFire

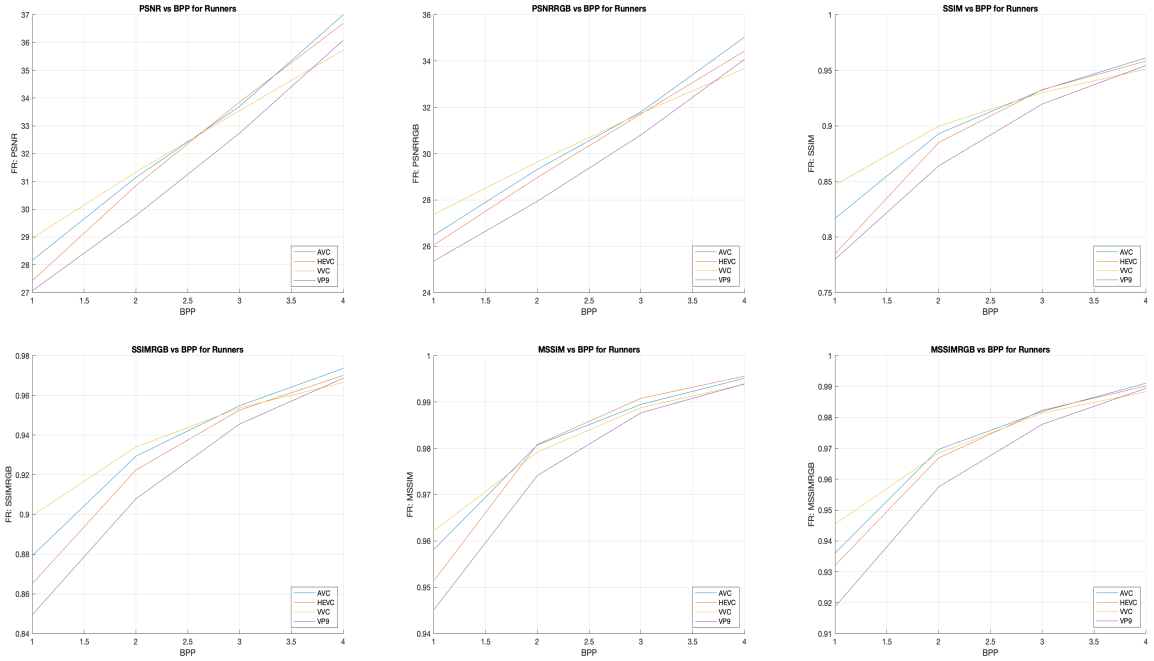

Figure 4: FR metrics for Runners

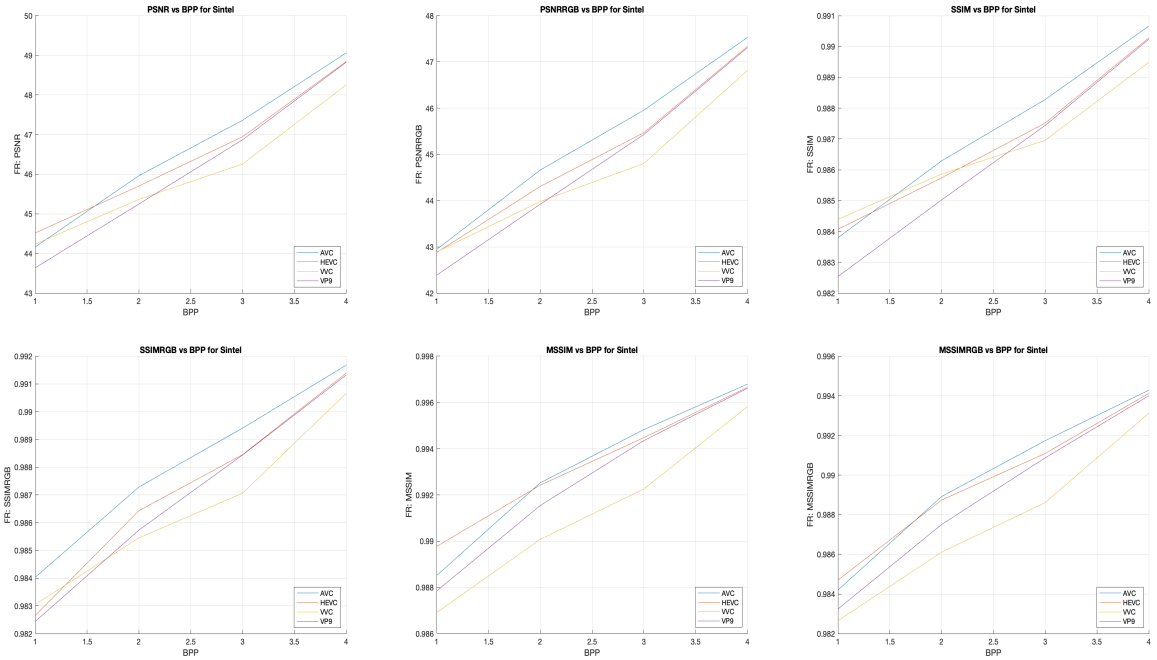

Figure 5: FR metrics for Sintel

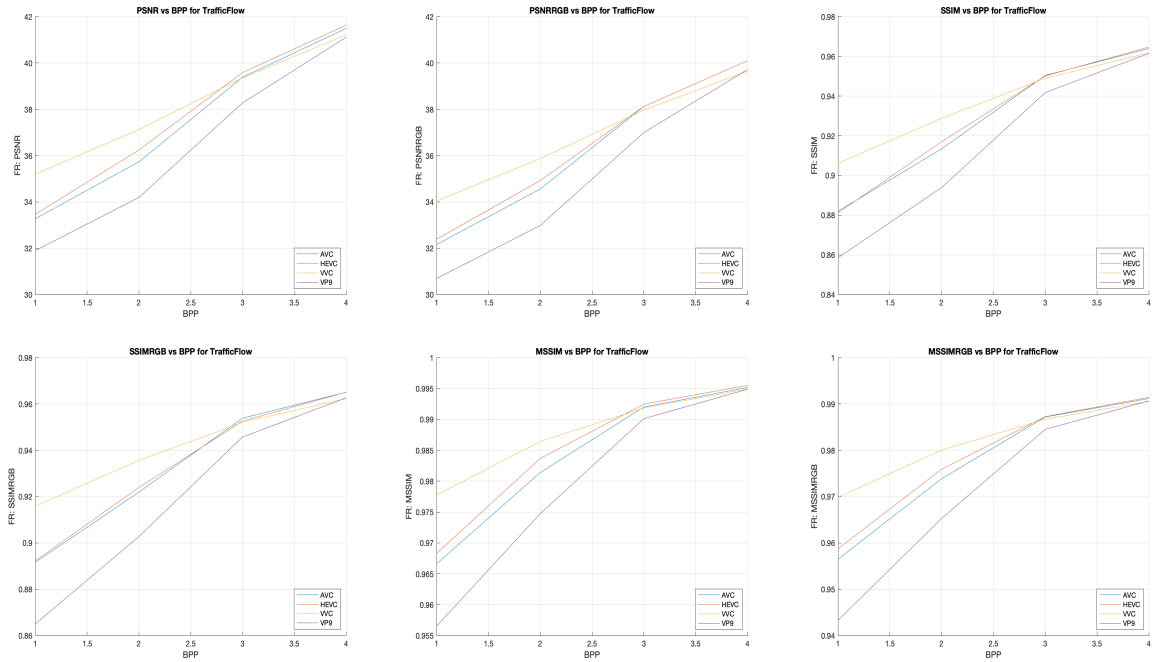

Figure 6: FR metrics for TrafficFlow

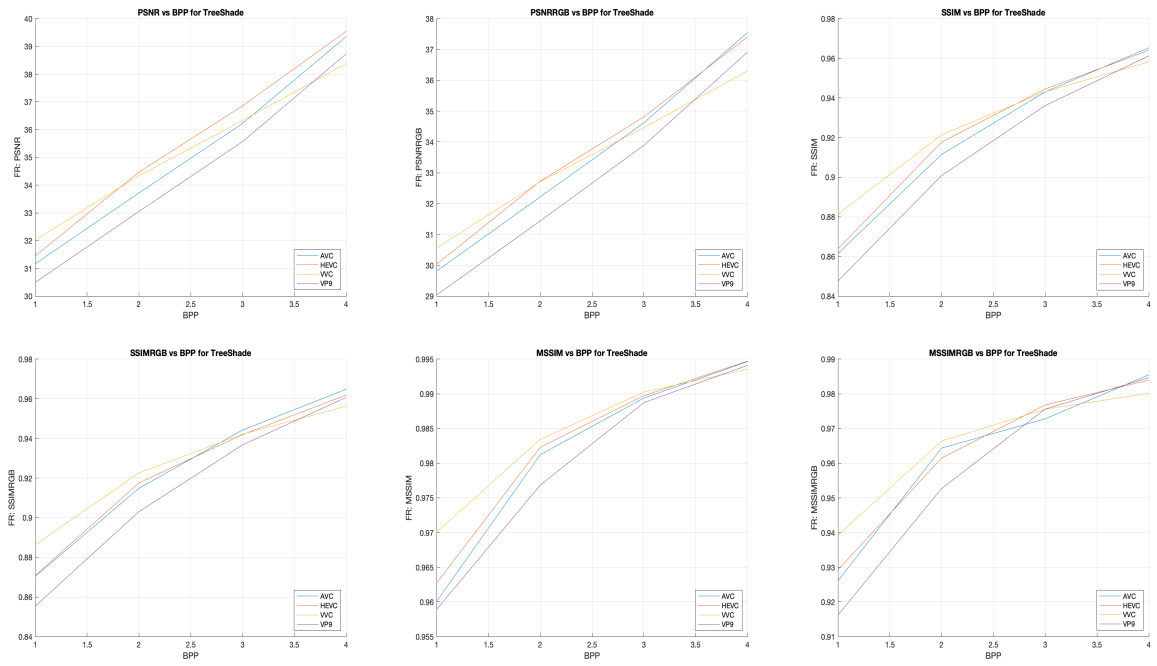

Figure 7: FR metrics for TreeShades

3.2 No Reference(NR) metrics-Brisque,Niqe,Piqe

No reference metrics, namely NR, have no reference images in comparison to FR metrics. There are various methods to compute NR scores.

BRISQUE is the one the main NR techniques. It stands for blind/referenceless image spatial quality evaluator. It does not compute any distortion specific features, but uses scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions.

NIQE is another NR technique and it stands as Naturalness Image Quality Evaluator. It is option-unaware and distortion-unaware, meaning that there is no training on databases involving human judgements or specific distortions. Unlike BRISQUE, NIQE uses only Natural Scene Statistic(NSS) from a corpus of natural images. It is not tied to any specific distortion type, but proves nearly as reliable scores as BRISQUE with much lower complexity. Smaller scores indicate better quality.

PIQE is also another NR metric, and stands as Perception-Based Image Quality Evaluator. It is option un-aware and unsupervised, therefore needs no training. It estiamtes block-wise distortion and measures the local variance of perceptibly distorted blocks to compute the quality score. Input image is seperated into B_k blocks of size nxn. Criterion to label a block is the following:

$$B_k = \begin{cases} U_k, & \text{if } v_k < TU \\ SA_k, & \text{if } v_k \geq TU \end{cases}$$

where v_k is the variance in a given block. TU is set to 0.1 by empirical observations.

A distorted block is assigned a score given the type of distortion, to be more specific: additive white noise criterion and/or noticeable distortion criterion. Less are the NR metric scores, better are the results. For higher bitrates, all of the NR metric scores should be decreasing. As expected, they are decreasing. Resulting Figures 8 to 12 demonstrate BRISQUE, NIQE and PIQE score results for specific contents at specific bitrate values. All of them decrease for higher bitrates for expected. In terms of codec analysis, it is hard to specify the best codec as there is no perfect codec having lowest score for each NR metric in any of the contents. HVEC and VP9 shows promising scores for each of the NR metrics, as their plots are similar in terms of bitrates. HVEC and VP9 have pretty similar scores, and are performing well at each of the contents. AVC has a similar score structure, however does not perform as well as HVEC and VP9. Lastly, VVC results in the worst scores overall, excluding some exceptional cases such as PIQE in Sintel2. To conclude, codecs HVEC and VP9 prove the most reliable scores overall in terms of BRISQUE, NIQE and PIQE considering every content. VVC has the worst performance, and AVC is in between, such that it performs better than VVC, but worse than HVEC and VP9.

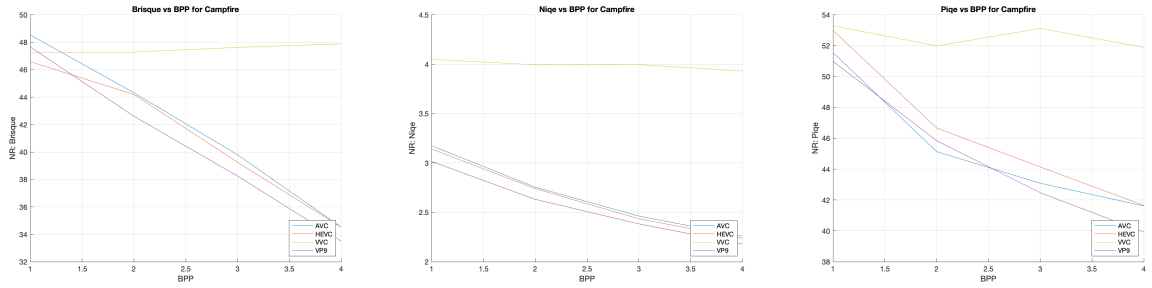

Figure 8: NR metrics for CampFire

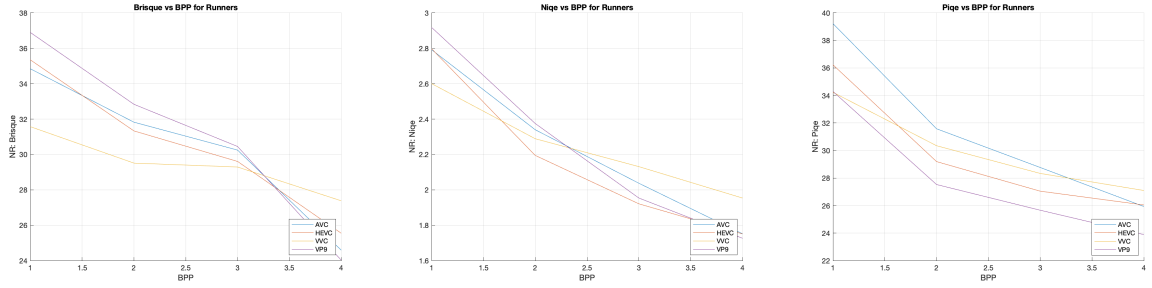

Figure 9: NR metrics for Runners

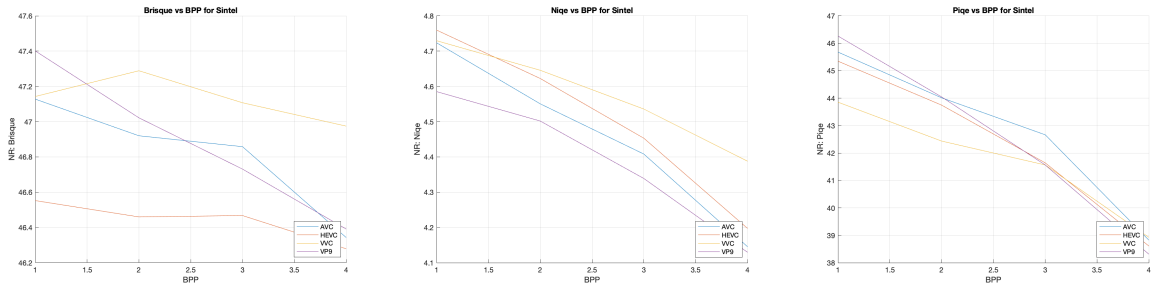

Figure 10: NR metrics for Sintel

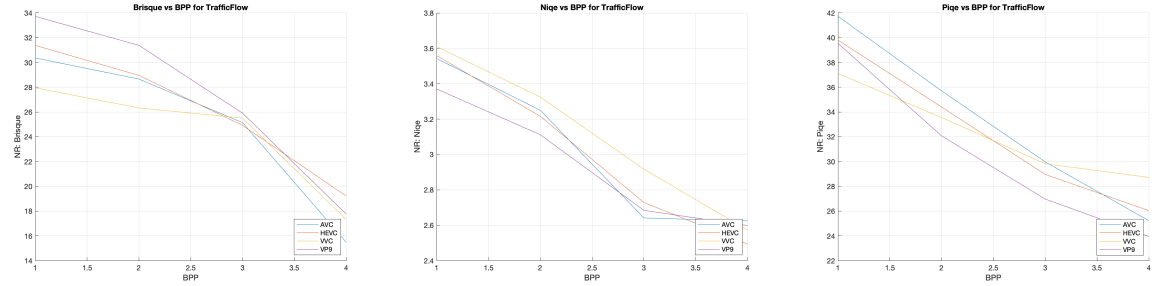

Figure 11: NR metrics for TrafficFlow

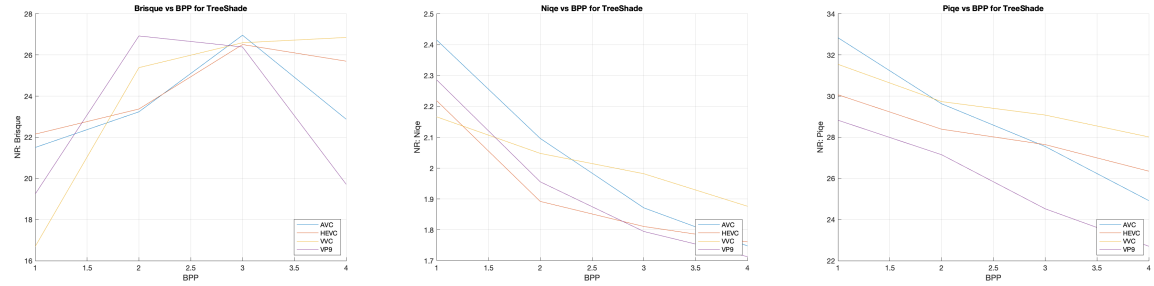

Figure 12: NR metrics for TreeShades

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:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{n}} \tag{10}$$

where Pearson correlation is defined as:

$$CC = \frac{n * \sum_{i=1}^{N} x_i * yi - \sum_{i=1}^{N} x_i * \sum_{i=1}^{N} y_i}{\sqrt{n * \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} * \sqrt{n * \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}} \tag{11}$$

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.

$$RC = \{1 - \frac{6 * \sum_{i=1}^{N} (R(x_i) - R(y_i))^2}{n * (n^2 - 1)} | x_i, y_i = 1, ..., n\} \tag{12}$$

4.1 No Polynomial Fitting

Figure 13 depicts results of each codec for each content in a single figure in terms of subjective versus objective scores. Same colors indicate same content. In this case, resulting pure FR and NR scores are plotted against MOS values. Our resulting opinions in previous sections, are proved once again in Figure 13 in terms of FR and NR scores.

Table 1 demonstrates overall Pearson and Spearman coefficients computed upon pure data as well as the RMSE score. Pearson and Spearman results have high values in terms of FR metrics and vice versa for NR metrics. As mentioned in the section overview, strong positive correlation results in values close to 1 and strong negative correlation results in values close to -1. Therefore, these existing values make sense. Only existing issue is that for some metrics RMSE values are really high such as PSNR, BRISQUE and PIQE. Table 2, 3 & 4 contain Peason, Spearman and RMSE scores for each content-FR/NR metric combination. These tables contain more detailed information in comparison to Table 1. Resulting score explanation is same as the explanation for Table 1.

Table 1: No fitting: OVERALL

OVERALL		PEARSON	SPEARMAN	RMSE
PSNR	Y	0.9969	0.9747	33.6294
	RGB	0.9871	0.9747	30.1489
SSIM	Y	0.9855	0.9747	2.1216
	RGB	0.9720	0.9747	2.0898
MS-SSIM	Y	0.9622	0.9747	2.0848
	RGB	0.9844	0.9747	2.1008
BRISQUE	Y	-0.9722	-0.9747	40.6245
NIQE	Y	-0.9987	-0.9747	1.6584
PIQE	Y	-0.9708	-0.9747	44.6427

Table 2: No fitting: PEARSON

PEARSON		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.9236	0.9651	0.9587	0.9312	0.8912
	RGB	0.9346	0.9616	0.9512	0.9307	0.8922
SSIM	Y	0.8463	0.9537	0.9720	0.9496	0.9391
	RGB	0.7428	0.9521	0.9285	0.9470	0.9376
MS-SSIM	Y	0.7305	0.9583	0.9047	0.9582	0.9419
	RGB	0.7657	0.9599	0.9182	0.9549	0.9596
BRISQUE	Y	-0.9709	-0.8650	-0.6475	-0.7806	0.6522
NIQE	Y	-0.8805	-0.9565	-0.8111	-0.8802	-0.8913
PIQE	Y	-0.9596	-0.8590	-0.9539	-0.8982	-0.7150

Table 3: No Fitting: SPEARMAN

SPEARMAN		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.9830	0.9632	0.9610	0.9175	0.9536
	RGB	0.9726	0.9543	0.9492	0.9057	0.9462
SSIM	Y	0.9830	0.9632	0.9610	0.9043	0.9536
	RGB	0.9859	0.9543	0.9227	0.9057	0.9418
MS-SSIM	Y	0.9859	0.9705	0.9051	0.9175	0.9551
	RGB	0.9785	0.9661	0.9051	0.9131	0.9757
BRISQUE	Y	-0.8835	-0.8881	-0.6873	-0.8498	0.3744
NIQE	Y	-0.9637	-0.9381	-0.8138	-0.8984	-0.9108
PIQE	Y	-0.9548	-0.8189	-0.9551	-0.8837	-0.7664

Table 4: No Fitting: RMSE

RMSE		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	32.2101	28.9763	42.7387	33.6424	30.7480
	RGB	28.7717	27.1162	41.3243	32.3366	29.0374
SSIM	Y	1.9439	2.6240	2.7686	3.1079	3.5684
	RGB	1.9132	2.6020	2.7681	3.1037	3.5670
MS-SSIM	Y	1.9088	2.5685	2.7630	3.0597	3.5104
	RGB	1.9279	2.5759	2.7661	3.0654	3.5289
BRISQUE	Y	40.8761	27.5231	43.2341	22.5910	19.5035
NIQE	Y	2.0011	1.9168	1.3999	1.5945	2.5750
PIQE	Y	45.2952	27.0192	38.8948	28.9033	23.7831

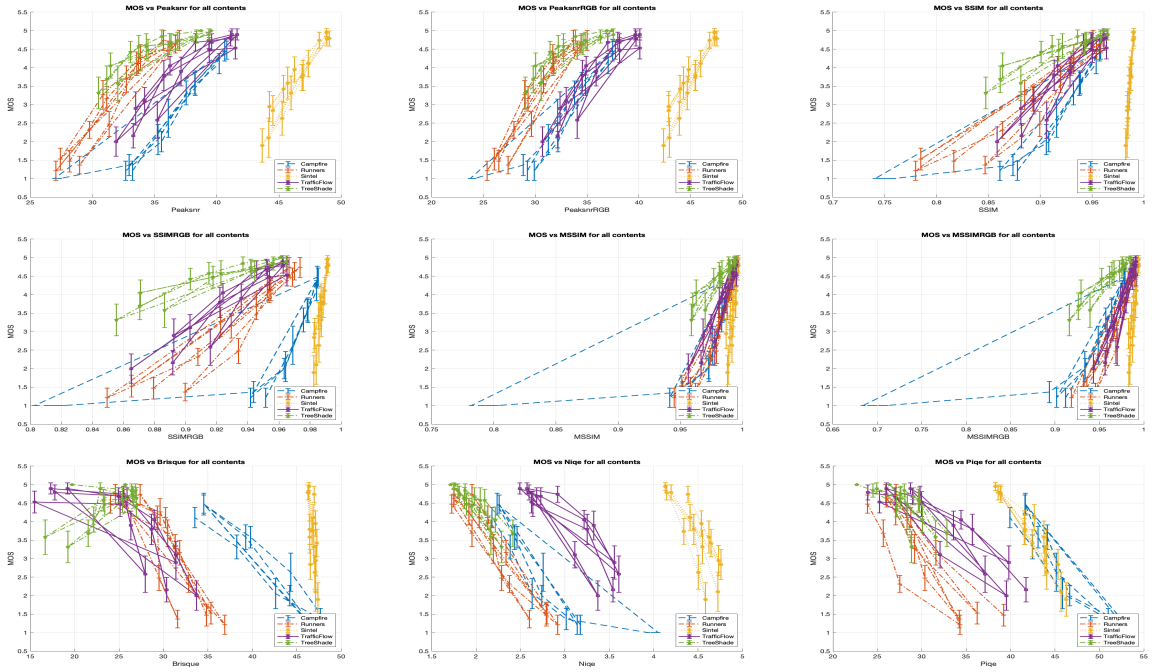

Figure 13: MOS vs Objective Ratings for Overall- No Fit

4.2 Linear Fitting

First polynomial fitting that is used is Linear Fitting. Linear fitting is used on raw scores that are mentioned in the previous subsection. Figure 13 depicts results all contents together with each codec in MOS and objective score comparison as in Figure 12. Resulting Figure 13 is really similar to Figure 12. Increase of FR metric scores for each content-codec combination with increased bitrates are certain. However, there is increase of NR metrics in same set conditions as well, instead of decrease of scores. This is the effect of polynomial fitting, more specifically in this case linear fitting. However when compared to Figure 12, although the decrease changed to increase, change itself is same as before. Table 5, 6, 7 & 8 are ensuring effect of polynomial fitting on MOS versus Objective Score comparison and corresponding plots as well. As it can be seen, Pearson and Spearman correlation coefficents are same as in previous sections table set(1, 2, 3 & 4). However, NR metrics changed their signs, but their values remain the same. Lastly, main advantage of linear fitting can be observed by looking at RMSE results, as all of the RMSE results dropped significantly to lower values after linear fitting. Table 5 provides Spearman, Pearson and RMSE scores for all the contents at the same time, nevertheless Table 6, 7 & 8 ensure these scores specific to each subject. There are some exceptional cases such as Spearman correlation coefficient to be strongly negative in case of SSIM-Sintel combination, but overall results are promising in terms of both FR and NR metrics.

Table 5: Linear Fitting: OVERALL

OVERALL		PEARSON	SPEARMAN	RMSE
PSNR	Y	0.9969	0.9747	1.2454
	RGB	0.9871	0.9747	1.1448
SSIM	Y	0.9855	0.9747	1.0759
	RGB	0.9720	0.9747	1.8055
MS-SSIM	Y	0.9622	0.9747	1.2522
	RGB	0.9844	0.9747	0.4532
BRISQUE	Y	0.9722	0.9747	1.0030
NIQE	Y	0.9987	0.9747	1.4744
PIQE	Y	0.9708	0.9747	0.8763

Table 6: Linear Fitting: PEARSON

PEARSON		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.9236	0.9651	0.9587	0.9312	0.8912
	RGB	0.9346	0.9616	0.9512	0.9307	0.8922
SSIM	Y	0.8463	0.9537	0.9720	0.9496	0.9391
	RGB	0.6067	0.8726	-0.9312	0.9211	0.9476
MS-SSIM	Y	0.7305	0.9583	0.9047	0.9582	0.9419
	RGB	0.9725	0.9709	0.9192	0.9578	0.9565
BRISQUE	Y	0.9709	0.8650	0.6475	0.7806	-0.6522
NIQE	Y	0.8805	0.9565	0.8111	0.8802	0.8913
PIQE	Y	0.9596	0.8590	0.9539	0.8982	0.7150

Table 7: Linear Fitting: SPEARMAN

SPEARMAN		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.9830	0.9632	0.9610	0.9175	0.9536
	RGB	0.9726	0.9543	0.9492	0.9057	0.9462
SSIM	Y	0.9830	0.9632	0.9610	0.9043	0.9536
	RGB	0.5538	0.9514	-0.9227	0.9057	0.9418
MS-SSIM	Y	0.9859	0.9705	0.9051	0.9175	0.9551
	RGB	0.9785	0.9661	0.9051	0.9131	0.9757
BRISQUE	Y	0.8835	0.8881	0.6873	0.8498	-0.3744
NIQE	Y	0.9637	0.9381	0.8138	0.8984	0.9108
PIQE	Y	0.9548	0.8189	0.9551	0.8837	0.7664

Table 8: Linear Fitting: RMSE

RMSE		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	1.1912	0.9734	1.1738	0.7578	1.1935
	RGB	1.0767	0.9892	1.2234	0.7401	1.1703
SSIM	Y	0.8535	0.5435	1.2300	0.6278	0.9780
	RGB	1.3878	0.8426	0.9792	0.6653	1.0469
MS-SSIM	Y	1.0865	1.0656	0.9081	0.7993	0.8457
	RGB	0.3494	0.6728	1.0649	0.5184	1.1144
BRISQUE	Y	1.1023	1.1889	1.3267	0.7587	0.7144
NIQE	Y	1.5123	1.2658	1.2172	0.9182	0.6250
PIQE	Y	0.8506	1.3091	1.0146	0.5447	0.3955

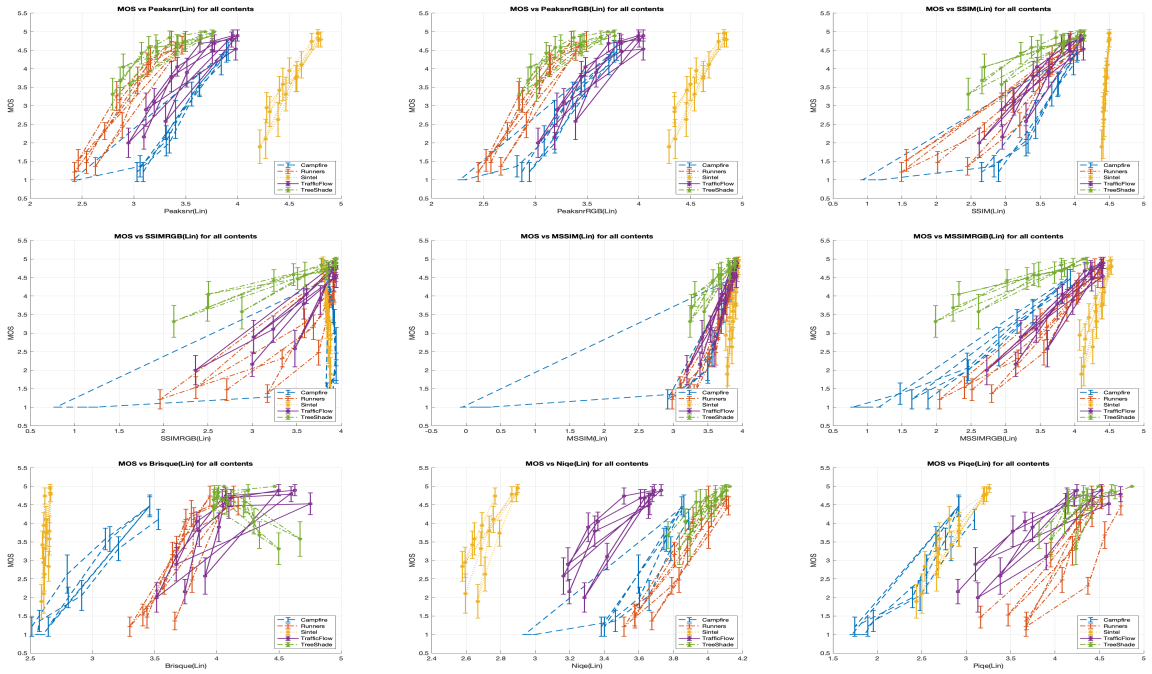

Figure 14: MOS vs Objective Ratings for Overall- Linear Fit

4.3 Cubic Fitting

Cubic fitting is another polynomial fitting done on the raw values. Instead of fitting the raw results into a linear polynomial form, cubic fitting converts these results to a polynomial of third degree. Figure 14 depicts results for cubic fitting for each metric separately. Each content-codec combination are plotted into the same graph per FR/NR metric. Resulting plots are similar overall to plots in Figure 13 for linear fitting. In order to comprehend the this Figure in full detail, Table 9, 10, 11 & 12 should be investigated. As in linear fitting RMSE values are much lower compared to raw comparison scores, however differ metrics such as PSNR(Y) and MS-SSIM(Y) with linear fitting in terms of performance. After checking each content performance separately in Table 10, 11, & 12, it would be reasonable to suggest that cubic fitting is an overfit in some cases(F.e Spearman-Sintel). However, RMSE scores prove that using cubic fitting results in better values in comparison with raw(no fitting) performance comparison. Comparison of cubic

and linear fitting depends on content a lot as well, but for the sake of explanation and our experimentation, it could be indicated that linear fitting performs better than cubic fitting overall given the Pearson, Spearman coefficients as well as the RMSE scores, although there are some exceptional cases as mentioned.

Table 9: Cubic Fitting: OVERALL

OVERALL		PEARSON	SPEARMAN	RMSE
PSNR	Y	0.9143	0.8721	1.5639
	RGB	0.9846	0.9747	1.2418
SSIM	Y	0.9407	0.9747	1.0584
	RGB	0.1409	0.2052	1.8055
MS-SSIM	Y	0.9839	0.9747	0.3682
	RGB	0.9927	0.9747	0.4532
BRISQUE	Y	0.2633	0.1539	1.2628
NIQE	Y	0.9897	0.9747	1.0408
PIQE	Y	0.9226	0.9747	0.6157

Table 10: Cubic Fitting: PEARSON

PEARSON		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.7086	0.9586	0.8204	0.8805	0.9522
	RGB	0.7990	0.9661	-0.9517	0.9274	0.9499
SSIM	Y	0.8800	0.9497	-0.9679	0.9339	0.9408
	RGB	0.6067	0.8726	-0.9312	0.9211	0.9476
MS-SSIM	Y	0.9677	0.9788	0.9076	0.9630	0.9373
	RGB	0.9725	0.9709	0.9192	0.9578	0.9565
BRISQUE	Y	0.3942	0.8724	-0.6487	0.8839	-0.5563
NIQE	Y	0.9705	0.9450	-0.8067	0.8114	0.8847
PIQE	Y	0.9101	0.8201	0.9554	0.7646	0.6401

Table 11: Cubic Fitting: SPEARMAN

SPEARMAN		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	0.9028	0.9632	0.8521	0.7320	0.9447
	RGB	0.9726	0.9543	-0.9492	0.8675	0.9462
SSIM	Y	0.9785	0.9485	-0.9610	0.8262	0.9403
	RGB	0.5538	0.9514	-0.9227	0.9057	0.9418
MS-SSIM	Y	0.9859	0.9705	0.9051	0.9175	0.9551
	RGB	0.9785	0.9661	0.9051	0.9131	0.9757
BRISQUE	Y	0.1218	0.8881	-0.6873	0.8940	-0.3375
NIQE	Y	0.8790	0.9381	-0.8138	0.9043	0.9108
PIQE	Y	0.9548	0.8189	0.9551	0.8837	0.7664

Table 12: Cubic Fitting: RMSE

RMSE		CampFire	Runners	Sintel	TrafficFlow	TreeShades
PSNR	Y	1.2362	0.5088	0.9358	0.7834	0.8214
	RGB	1.0017	0.5451	1.0598	0.7641	0.7691
SSIM	Y	0.9258	0.4133	1.0281	0.4734	0.8797
	RGB	1.3878	0.8426	0.9792	0.6653	1.0469
MS-SSIM	Y	0.3891	0.4089	0.9875	0.3623	0.9835
	RGB	0.3494	0.6728	1.0649	0.5184	1.1144
BRISQUE	Y	1.3075	1.0071	1.2875	0.6685	0.6560
NIQE	Y	1.2065	0.9747	1.1529	1.2496	0.2529
PIQE	Y	0.5880	1.2708	0.8417	0.6883	0.4755

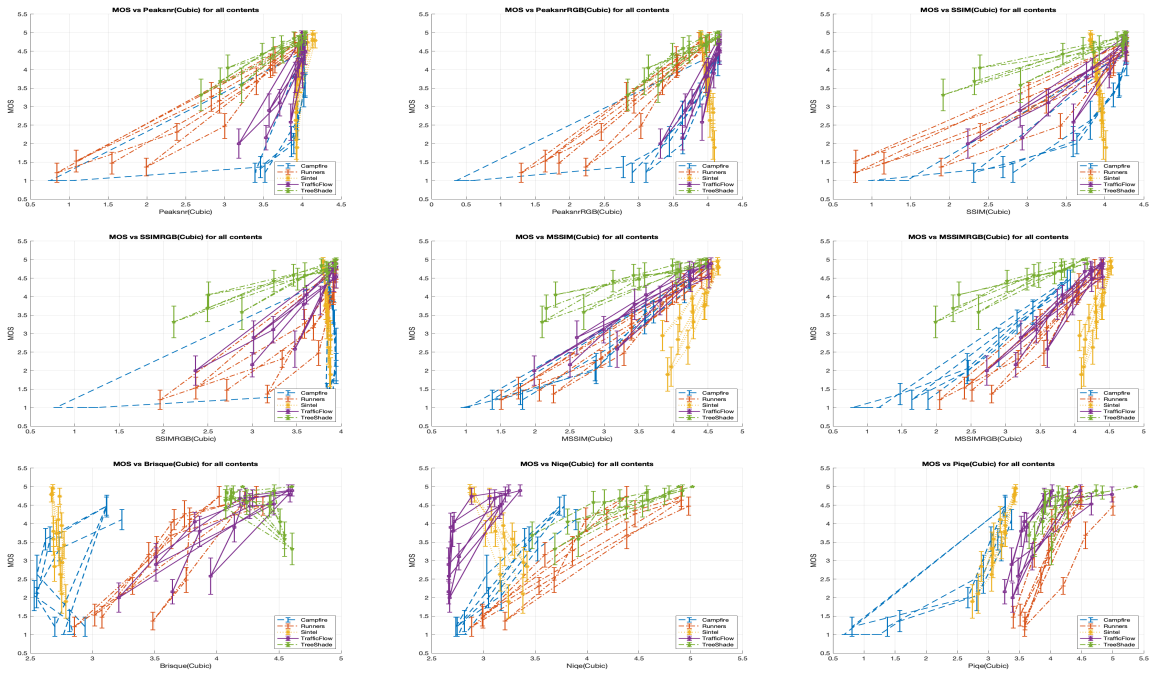

Figure 15: MOS vs Objective Ratings for Overall- Cubic Fit