EE669 Homework #6

Yifan Wang
wang608@usc.edu

November 24, 2019

1 SSIM for Image Quality Assessment

(1)

Advantages

1. MSE is a simple function and can be easily computed without expensive cost on memory or time. Besides, there is no parameters needed to be fine tuned when computing it, which means the evaluation metric is the same under any conditions for any data type.
2. It uses L_p norm as distance metric which has some characteristics like none negativity, identity, symmetry and triangular inequality which make it a nice prosperity in R^n space.
3. The physics meaning is clear. L_2 norm distance shows the energy of error signal which can be valid after many transformations like linear or Fourier transformation. This advantage make widely use in signal processing field where it can be used on many transformation results and make these result comparable.
4. It is a nice metric when doing optimization, since it would always show the difference with desire output. Guided by it, it is easy to find the one that minimize the error. Both gradient and Hessian matrix is relative easy to compute. L_2 norm is super nice when the condition matrix is bad. It trends to prevent the optimization result from being stacking in some local minima.
5. MSE is additive for independent source of distortions.
6. It was the first metric introduced to evaluate the performance, so that it is used by many other algorithms as a benchmark.

Disadvantages

1. MSE would not consider human's perception difference for various distortion. It would regard all distortion in same weight. However, human eyes would be more sensitive to noise and blur while less sensitive to mean luminance shift. If treat them as the same weight would conflict to human perceptions and give some bad score of MSE even though human cannot find the difference.
2. MSE is super sensitive to rotation or pixel shifting. Since it needs matched pixel pairs to compute. However, some pixel shift or rotation would not affect human's visual experience since it remains the raw image.
3. MSE cannot represent signal fidelity which is independent to input signal and error signal.
4. It would not include the direction information but only the norm. While direction information is super important when we evaluate a signal.

(2)

Formular:

SSIM is based on the theory that human eye would emphasis more on structure information. So that it would compare the structure information between raw and target image. Using formula:

$$SSIM(x,y) = l(x,y)^{\alpha} c(x,y)^{\beta} s(x,y)^{\gamma}$$

where

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$

$$c(x,y) = \frac{2\delta_x\delta_y + c_2}{\delta_x^2 + \delta_y^2 + c_2}$$

$$c(x,y) = \frac{2\delta_{xy} + c_3}{\delta_x\delta_y + c_3}$$

$$\delta_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

where x, y are input image pairs, μ_x is the mean of local image (receptive field) x, δ_x is variance of local image x, δ_{xy} is the covariance between x and y. c_1, c_2, c_3 are small positive number which helps to prevent denominator from being zero.

$l(x,y)$ represents the luminance difference, $c(x,y)$ considers the difference of contrast and $s(x,y)$ consider the structure information. α, β and γ are positive numbers. Which is use to control the importance of each components in the equation (structure would account for more importance). Typically, it is calculated in a 11×11 gaussian window on luma image only.

Characteristic:

SSIM is symmetric and range from -1 to 1. 1 means input image is the same as reference image. It is computed by moving a window in a local region which result a SSIM map. Then the SSIM score of entire image is computed by averaging the SSIM values across the image. (some adaptive space-variant weighting SSIM can be used as well.) It can perform well across many image distortion types which may sensitive to human eyes. While luminance shifting or contrast stretching which would not degrade image structure or human visualization experience will have high SSIM. These advantages are bought by the design which focus on structure difference while incorporating with other important perceptual phenomena.

Structural information is about the edges and other parts which have strong inter-correlations especially in the neighbourhood. These dependency would give important information about structure of object which is an important part when human judge image qualities. Besides, it would regard distortion in dark region as less important than bright region like human eye.

Drawback:

However, SSIM would be sensitive to translation, scaling or rotations where the structure would be shifted and cannot be found in a local window.

(3)

By watching at these five images, I think $lena\_distor2.raw$ would have higher PSNR and SSIM, since it looks almost the same as the raw image. While $lena\_distor5.raw$ would have worst PSNR and SSIM, it was blurred and all the edge information was destroyed. While $lena\_distor1.raw$ looks nice, I don't think it would have high PSNR, because its contrast has been modified which make it not suit for computing PSNR, on the other hand, its SSIM would be nice, since all the structure information are retained.

	lena_distor1.raw	lena_distor2.raw	lena_distor3.raw	lena_distor4.raw	lena_distor5.raw
MSE	288.752	289.000	289.057	288.996	288.172
PSNR	23.5255	23.5218	23.521	23.5219	23.5343
SSIM	0.9418	0.9803	0.7390	0.7369	0.6959

Table 1: Evaluation results of image set 1

It is surprising that PSNR for all these five images are the same which is beyond my exception. This can be a case that shows PSNR is not a nice metric when evaluating an image. SSIM would be a better one, since SSIM score for these five images highly matched with my prediction. This shows that SSIM is more close to human visual systems.

(4)

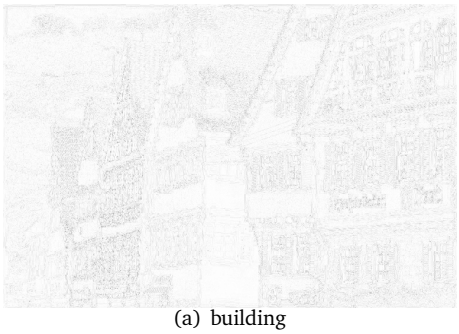
	lena_distor6.raw	lena_distor7.raw	lena_distor8.raw	lena_distor9.raw	lena_distor10.raw
MSE	133.088	193.936	520.45	550.044	1847.88
PSNR	26.8894	25.2542	20.967	20.7268	15.4641
SSIM	0.9204	0.8540	0.7430	0.7424	0.3605

Table 2: Evaluation results of image set 2

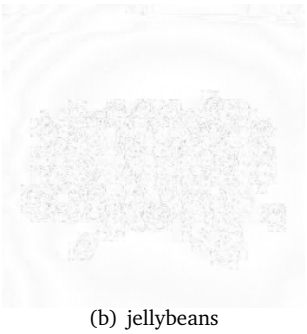
These five images contains shift right, shift up, rotate counter-clockwise, rotate clockwise and crop. It would be obvious that $lena\_distor10.raw$ would have the lowest PSNR and SSIM, since this image is cropped which make all pixel shift a lot resulting low PSNR. Besides, due to the cropping and resize, since structure information is neglected, SSIM score would not be too high either. While for the rotation, it uses black pixels to fill the blank part which make me to believe that it would have lower PSNR than shifting which uses gray pixel to fill the blank part. Considering SSIM, I think rotation would have almost same result as shifting since SSIM is sensitive to them.

The result score conforms my prediction. While shifting different direction would make so large difference is beyond my exception. It shows that even though SSIM is more close to human perception than PSNR, there remains some difference human eyes are not sensitive to such small shifting. Sensitive to spatial shifting is a drawback of SSIM.

(5)

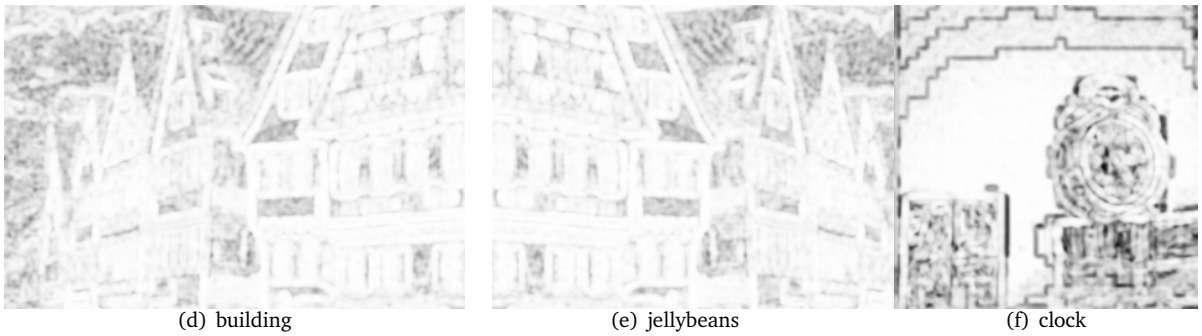


(a) building



(b) jellybeans

(c) clock



(d) building (e) jellybeans (f) clock

Figure 0: Error map, first row: absolute error map, second row: SSIM index map

	buildings_jpg.raw	clock_jpg.raw	jellybeans_jpg.raw
MSE	207.455	108.66	50.2112
PSNR	24.9616	27.7701	31.1228
SSIM	0.9283	0.8631	0.9108

Table 3: Evaluation results of image set 3

It can be found SSIM would focus on the difference in edges and other structure information. Due to the blocking effect of JPEG, SSIM is not such high. From error map in $Figure1$, it can be found that error mainly errors in place where blocking effects happen and place where includes many fine details. Beacuse for an image with so many blocking effect, its Q factor when doing JPEG compression is quite low. In that case many details are discarded (high frequency part), and higher error.

2 VMAF for Video Quality Assessment

(1)

Video Multimethod Assessment Fusion (VMAF) is an objective video quality metric. It can predict subjective video quality based on input and reference video sequences. It can be used to evaluate different video codecs, encoders etc.

Basic:

VMAF combines several basic quality metrics by using a machine learning algorithm. In this case, a SVM is used to weight different metric scores to produce a more accurate score which is close to opinion scores obtained through a subjective experiment

The basic metrics used in VMAF includes Visual Information Fidelity (VIF), Detail Loss Metric (DLM), motion etc. VIF and DLM are both image quality measure.

1. VIF is based on the premise that image quality is complementary to information fidelity loss. In VMAF, they used 4 loss of fidelity score from 4 different scale as elementary metric.

2. DLM is based on the rationale of separately measuring the loss of details. It affects the content visibility and the redundant impairment which distracts viewer attention.

Motion. Which is a measure of temporal difference between two adjacent frames by calculating the average absolute pixel difference for the luminance component.

SVM is used to combine these basic metrics and produce a score which is more close to human perception. Train label comes from the objective score given by many people.

New:

VMAF has got several improvement based on the framework introduced in 2016. Since it uses a SVM inside to combine each elementary metrics, final accuracy would by huge influence by the training data. So that they were continuing improve the dataset and the opinion score which helped to improve the VMAF score. Secondly, due to the difference of user environment (Smart phone, hdtv, computer etc), they are developing VMAF score based on user environment. It is obvious that because of the limited size of the screen of smart phone, people would not capture some fine detail loss. While for playing videos on a HD or 4K TV, which has higher resolution and larger screen, people would need much more details to achieve the same visual experience as that when using smart phone.

(2)

$Figure2$ shows a screenshot of 99^{th} frame from these videos. It can be found $FoxBird\_20\_288\_375.yuv$ and $CrowdRun\_03\_288\_375.yuv$ having more blurring and blocking which is extremely obvious in edge part so that they would have lowest evaluation score. While $FoxBird\_95\_1080\_5800.yuv$ and $CrowdRun\_90\_1080\_15000.yuv$ are the best. These two videos looks almost the same as the original one (I cannot tell any difference).

(a) ori (b) 20_288_375 (c) 55_480_750 (d) 95_1080_5800

(e) ori (f) 03_288_375 (g) 50_1080_4300 (h) 90_1080_15000

Figure 1: 99^{th} frame

```
./ffmpeg \
  -pix_fmt yuv420p -s 1920x1080 -i input.yuv \
  -pix_fmt yuv420p -s 1920x1080 -i ori.yuv \
  -lavfi libvmaf='log_path=./vmaf_score.txt:psnr=1:ssim=1' \
  -f null -
```

FoxBird	_20_288_375.yuv	_55_480_750.yuv	_95_1080_5800.yuv
PSNR	31.256	34.560	45.578
SSIM	0.9448	0.9752	0.9987
VMAF	26.000	60.164	98.299
CrowdRun	_03_288_375.yuv	_50_1080_4300.yuv	_90_1080_15000.yuv
PSNR	23.003	28.384	34.207
SSIM	0.7701	0.9585	0.9926
VMAF	10.461	66.251	96.541

Table 4: Evaluation results using VMAF

Considering SSIM, it's surprising that $FoxBird\_20\_288\_375.yuv$ would have such a high SSIM score. Considering the fact the largest value of SSIM score is 1, I think this score cannot represent the huge gap from my

impression when watching these two video (One is super good while other is terrible from my perception). By simply observing the SSIM score would give me a feeling that both videos' quality is nice. On the other hand VMAF score shows this huge difference of my visual experience. This proves that VMAF score is more close to human's visual system.