Video Clarity
Tools for Video Analysis

1566 La Pradera Dr
Campbell, CA 95008
www.videoclarity.com
408-379-6952

White Paper – Advancing To Multi-Scale SSIM

Video Clarity, Inc.

Advancing To Multi-Scale SSIM

Since it was first created at the University of Texas in 2002, the Structural SIMilarity (SSIM) image assessment algorithm has become a valuable tool for still image and video processing analysis. SSIM provided a big advance over MSE (Mean Square Error) and PSNR (Peak Signal to Noise Ratio) techniques because it much more closely aligned with the results that would have been obtained with subjective testing. Today, an even better version of SSIM has been developed, which is called Multi-Scale SSIM, or MS-SSIM. For objective image analysis, this new method represents as significant an advancement over SSIM as the advancement that SSIM provided over PSNR.

Subjective and Objective Testing

Subjective testing is the gold standard of video analysis, because it uses human volunteers to compare video test sequences that have been processed, such as video that has gone through compression or has experienced distortion or loss during transmission. Because it uses actual human volunteers, subjective testing is considered to be a very accurate model of the human visual system.

Unfortunately, subjective testing is very expensive and difficult to do correctly. Since each person can have a different perception of image quality, multiple volunteers must be used to ensure that the test results are statistically significant. A considerable amount of time can be consumed in gathering the required number of test subjects and actually performing the test. For the test results to be valid, the viewing conditions must be stringently controlled to eliminate possible biases that could be introduced by incorrect viewing angle and distance, improper lighting, or other factors that could interfere with the reliability or repeatability of the tests.

Objective testing uses mathematical techniques to analyze images in order to simulate the results of subjective testing. Algorithms have been developed to analyze images (or image sequences) that can be done standalone or by comparison with a known reference image (or sequence). Objective testing is highly repeatable, in that the algorithms will always produce the same result when presented with the same set of input images. As compared to subjective testing, objective algorithms can be run extremely fast and without human intervention. With the right hardware, depending on the complexity of the algorithm being used, objective results can be obtained in real time while a video is playing. These factors combine to reduce the costs of objective testing to a small fraction of the costs of subjective testing, making it the best choice for repetitive and in-service testing.

When image quality measurement results are reported, it is common to use either the Mean Opinion Score (MOS) scale or the Difference (or Differential) Mean Opinion Score (DMOS) scale. The MOS scale is based on the way that viewers grade images, ranging from “Unacceptable” (unwatchable) to “Excellent” (perfect). DMOS compares two images that have been evaluated using the MOS scales. When one of the two images in the comparison is a high-quality original, and the other has been degraded in some manner, the amount of difference between the MOS scores of the two images gives a DMOS score that indicates how severely the image has been degraded.

Structural SIMilarity (SSIM)

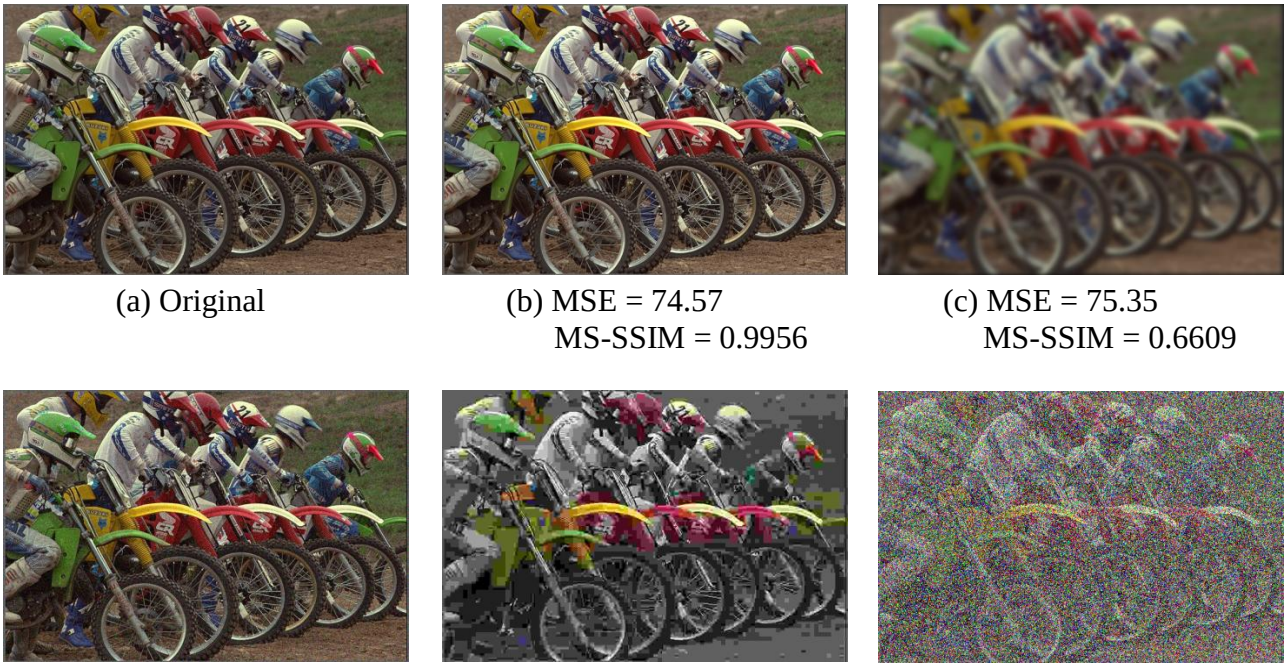
SSIM was developed to more accurately model the way that humans perceive image quality than the mathematical models used previously. MSE and PSNR weight every change in pixel values equally, whether or not the change would be noticeable to a human observer. This equal weighting could, for example, cause a high difference score to be attributed to a pair of images where the brightness or contrast was changed, even though such changes are not particularly important to human observers. SSIM would be more likely to score the image pair similarly, because the structure of the two images would closely match.

SSIM compares three major aspects of a pair of images to derive a measurement of how different they would appear to a human observer:

- Change in Luminance, which compares the brightness of the two images. The human visual system is not particularly sensitive to the absolute level of brightness of an image, but is sensitive to differences in brightness between two images.
- Change in Contrast, which looks for differences in the range between the brightest and darkest extent of the two images. As in the case of luminance, the human visual system is not particularly sensitive to the absolute amount of contrast difference in an image.
- Correlation, which compares the fundamental structure (represented by the local luminance patterns) of the two images to determine if they are similar or different, as they would appear to human observers. The parameter that is actually measured to determine correlation is called the "covariance" of the two images. If both images are closely matched, then the covariance will be high; in areas where the two images are different the covariance will decrease. (Note that this step is performed after the average luminance of the images has been equalized and the overall contrast of the two images has been normalized.)

The SSIM process begins with the two images that are to be compared being scaled to the same size and resolution, to allow for a pixel-by-pixel comparison. Then, a fixed size window is selected within the images where the mathematical comparisons take place. Measurements for each of the three aspects described above are then combined into an overall score that represents the quality level of the image. After each comparison, the window is moved to another portion of the image and then repeated. Scores for each window location are accumulated and averaged to yield an image difference value (expressed as a DMOS) for the overall image.

Figure 1 illustrates the advantages of SSIM over MSE image evaluation. The image (from the University of Texas LIVE image database) was distorted in five different ways to achieve similar MSE scores, but with radically different MS-SSIM. In figure 1, (b) has increased contrast, (c) has been blurred, (d) has Gaussian noise added, (e) shows JPEG compression, and (f) has salt and pepper noise added. In each case, the MS-SSIM score shows a significantly better ability to track and match with perceived image quality than MSE. (Note: it may be easier to view the distortions by zooming in on the images.)



(a) Original

(b) MSE = 74.57
MS-SSIM = 0.9956

(c) MSE = 75.35
MS-SSIM = 0.6609

(d) MSE = 74.57
MS-SSIM = 0.9592

(e) MSE = 77.90
MS-SSIM = 0.6609

(f) MSE = 75.64
MS-SSIM = 0.4145

Figure 1

Multi-Scale SSIM (MS-SSIM)

Using SSIM as a basis, MS-SSIM extends the technique by making multiple SSIM image evaluations at different image scales. This is accomplished by repeatedly performing the image analysis in multiple iterations, with each successive image pair that is downsampled by a factor of two from the previous iteration. One concept that drove the development of MS-SSIM was the observation that the performance of SSIM was dependent on the scale of the images, including changes in image viewing distance. With MS-SSIM, the scale of the images becomes less important, to the point where images that have been upconverted or downconverted (say from High Definition to Standard Definition) can still be objectively compared.

The MS-SSIM process begins with the same procedure as SSIM, in that the two images that are being compared are scaled to the same size and resolution, normalized for luminance and contrast, and then analyzed using a sliding window. The images are compared using the same criteria as SSIM: change in luminance, change in contrast, and correlation (or structure). Scores for each comparison are accumulated, and then the images are downsampled, and the comparisons are repeated. This process continues for a fixed number of iterations; for most implementations, between three and five iterations are used.

Because there are comparison results from each of the various image scales, the results must then be combined into an overall score for the image. Since human perception of the impact of distortions varies with image scale, different weightings are given to the results at each image scale. These weights were derived from experiments using multiple image sets and human observers that were told to identify images that had the same amount of perceived distortion at each scale. One interesting result of this analysis was that the luminance comparison only needs to be done at the smallest scale (most highly downsampled) whereas the contrast and structure comparisons are accumulated at each image scale.

To justify the extra workload that MS-SSIM places on an analysis system, the results must be better. This is indeed the case, as has been shown in several studies which are detailed in the following section. It is interesting to note that the improvement of MS-SSIM relative to SSIM is similar to the improvement that SSIM achieved relative to PSNR. With the processing power that is available today in measurement equipment, the incremental cost of using MS-SSIM is low relative to the value of the increased accuracy that can be achieved by using the algorithm.

MS-SSIM vs. SSIM by the Numbers

Several large scale tests have been done to assess the ability of both SSIM and MS-SSIM to provide objective scores that most closely approximate the subjective scores for a collection of image pairs. The closer the match, the better the objective algorithm is for simulating how human viewers would evaluate the same images.

This testing begins with a set of high-quality images. These images are processed to add various types of distortion, and a pool of test images is created. Volunteers then perform subjective testing of the images in the pool to create one set of results. A second set of results is created by using an objective test to also score all of the images in the pool, both pristine and degraded. These two sets of results are then compared. If these two sets of results are mathematically correlated, it means that the objective measurement method is a good predictor of how a human viewer would rate the image. Higher levels of correlation or lower residuals (the amount of variance that remains after the correlations are removed) indicate better performance by the objective model.

This testing process is repeated for each of the objective tests being evaluated. (In reality, the subjective testing is done once and this same set of results is used for each comparison). So, for example, if the goal was to compare the performance of the SSIM algorithm to the MS-SSIM, then the evaluation process would be done twice: first the SSIM results would be compared to the subjective testing results, and then the MS-SSIM results would be compared to the same subjective results. The results of these two comparisons could then be examined to determine which algorithm more closely correlated with the subjective (human) results. This was exactly the process used in three major sets of results that looked at various forms of objective testing.

The first set of results was published in a paper entitled “Multi-Scale Structural Similarity for Image Quality Assessment” by Z. Wang, E. Simoncelli and A. Bovik; an invited paper for the 37th IEEE Asilomar Conference on Signals, Systems and Computers, held in Pacific Grove, CA, on November 9-12, 2003. In this paper, MS-SSIM was compared to the PSNR and the SSIM algorithms by comparing each of them to the results that were obtained from subjective testing of 344 images with MOS data obtained by averaging 13 to 25 subjective scores that were given by human observers.

Model	Non-linear regression correlation coefficient	Spearman rank-order correlation coefficient	Mean absolute error	Root mean squared error	Outlier ratio (%)
PSNR	0.905	0.901	6.53	8.45	15.7
SSIM (M=2)	0.963	0.959	4.21	5.38	2.62
MS-SSIM	0.969	0.966	3.86	4.91	1.16

Five results are given for each of the algorithms being evaluated in the preceding table. The first two results show the correlation of the algorithm with the subjective results, with a higher value being a better score. The remaining three results show the amount of the variation in the subjective data that was not explained (the “error”) by the objective algorithm. One each of the five criteria, MS-SSIM was the best performer.

Note that in the preceding table that the SSIM result is annotated with an “M=2.” In this particular test, the SSIM algorithm was run on the original images as well as on images that had been scaled down from the original. Five sets of SSIM results were calculated, with M indicating the downscaling counter for the image, where a value of M=1 indicates that no downscaling was used, and a value of M=5 indicating that the image had been downscaled four times. The result table shows the results for M=2 because this was the best result obtained out of the five different scales used for SSIM.

The second set of results, which is available on-line at http://www.ponomarenko.info/tid2008.htm, compares eighteen different objective testing methodologies. This test used possibly the largest set of results for subjective image evaluation ever compiled, with total of 838 human observers performing 512856 evaluations of relative visual quality in image pairs. A total of 25 reference and 1700 distorted images were used in the test.

Each of the eighteen objective testing methods was evaluated in comparison to this large database of subjective results. Two different types of correlation analysis were used to do this comparison: Spearman and Kendall. Objective results that more highly correlated with the subjective results (i.e. more closely modeled human observations) were ranked higher in the following tables. Out of the eighteen objective models that were evaluated, MS-SSIM ranked highest for both correlation tests.

Rank	Measure	Spearman correlation	Rank	Measure	Kendall correlation
1	MS-SSIM	0.853	1	MS-SSIM	0.654
2	SSIM	0.808	2	SSIM	0.605
3	VIF	0.75	3	VIF	0.586
4	VSNR	0.705	4	VSNR	0.534
5	VIFP	0.655	5	VIFP	0.495
6	NQM	0.624	6	PSNR-HVS	0.476
7	UQI	0.6	7	NQM	0.461
8	PSNR-HVS	0.594	8	PSNR-HVS-M	0.449
9	XYZ	0.577	9	UQI	0.435
10	IFC	0.569	10	XYZ	0.434
11	PSNR-HVS-M	0.559	11	IFC	0.426
12	PSNRY	0.553	12	PSNRY	0.402
13	PSNR	0.525	13	WSNR	0.393
14	MSE	0.525	14	LINLAB	0.381
15	SNR	0.523	15	SNR	0.374
16	WSNR	0.488	16	DCTUNE	0.372
17	LINLAB	0.487	17	PSNR	0.369
18	DCTUNE	0.476	18	MSE	0.369

The third set of results comes from a paper entitled "Study of Subjective and Objective Quality Assessment of Video" by K. Seshadrinathan et al., which was published in June 2010 in *IEEE Transactions on Signal Processing*. In this paper, four sets of degraded video sequences were used; one set each that simulated the degradations caused by wireless network data loss, IP network packet loss, H.264 compression, and MPEG-2 compression. A summary combining all the data from the four image sets was also produced.

The results in this paper show that MS-SSIM performed better than both PSNR and SSIM, as summarized in the following four tables. The first two tables show two different measures of correlation, with higher scores meaning higher correlation. A higher correlation indicates that the objective algorithm more closely matches the subjective tests done with human observers, and hence indicates the algorithm's greater ability to mimic human scoring. The third and fourth tables show the variance of the residuals, with lower scores being better. Overall, the data in these four tables shows a substantial, significant improvement in the ability for MS-SSIM to objectively score images more closely to human observers.

Spearman Rank Order Correlation Coefficient

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.4334	0.3206	0.4296	0.3588	0.3684
SSIM	0.5233	0.455	0.6514	0.5545	0.5257
MS-SSIM	0.7285	0.6534	0.7051	0.6617	0.7361

Linear Correlation Coefficient

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.4675	0.4108	0.4385	0.3856	0.4035
SSIM	0.5401	0.5119	0.6656	0.5491	0.5444
MS-SSIM	0.717	0.7219	0.6919	0.6604	0.7441

Variance of the residuals between individual subjective scores and VQA algorithm prediction

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	189.77	171.83	193.18	179.04	201.07
SSIM	180.59	164.33	166.02	165.83	184.99
MS-SSIM	156.77	140.78	159.37	152.21	153.97

Variance of the residuals between VQA algorithm predictions and DMOS values

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	86.87	75.66	97.84	81.78	101.55
SSIM	77.46	67.91	69.98	68.24	85.36
MS-SSIM	53.07	43.58	63.15	54.3	54.15

To understand why MS-SSIM is better than SSIM, it's helpful to consider what is being modeled – the human visual system (HVS). The HVS is particularly adept at recognizing objects in their natural environment at multiple scales, that's why we can easily recognize faces of people we know whether they are five feet or fifty feet away from us. Distortions that are introduced by modern compression and video transport systems are also often multi-scale, so they also can be better evaluated using MS-SSIM. Even more important, SSIM breaks down if the video resolution is changed i.e. scaled down or up, or if the viewing distance to the images is changed. The performance of MS-SSIM does not have these same limitations.[1]

[1] Regarding scaled images/videos, or images/videos viewed at different distances, there has not yet been a large scale human study done, it is based on observations that MS-SSIM does not break down by many authors.

Conclusion

While there is no disputing the ultimate superiority of subjective testing for evaluating images and signals, such testing can be hard to do correctly, is very expensive, and is quite time consuming. Accordingly, if a suitable objective testing algorithm can be developed to closely match the results that would be obtained with subjective testing, then the objective algorithm can be used to simplify and automate video image comparison. An added benefit of objective testing is repeatability, because an algorithm will always return the same results for a given image pair. This enables many forms of testing that are difficult or expensive to do with subjective testing, such as fine-tuning the configuration of a video compression encoder and decoder pair. Repeatability also allows test equipment located at different points within a distribution network to create results that can be compared and analyzed to compare the quality of various technologies and delivery network providers.

As the results presented in this paper clearly indicate, MS-SSIM represents a significant improvement over SSIM, making it one of the most accurate objective image measurement techniques in use today. Based on the research summarized in this paper, any user who requires critical image quality measurements should chose the more representative and accurate MS-SSIM for testing in place of the good, but superseded SSIM. MS-SSIM on the DMOS scale is featured in Video Clarity's ClearView product line, which offers a comprehensive set of tools for measuring audio and video quality.