SSIM - robustness of the image quality metric

S. Schauman[1], K. Setsompop[1]
[1]Department of Radiology, Stanford University, Stanford, CA, US

Target audience: Researchers that use SSIM as a method to assess their own and others' image reconstruction algorithms.

Purpose: Structural similarity index measure, SSIM, is a commonly reported metric to assess image quality, however, it can be sensitive to image scaling among other parameters unrelated to visual quality. This investigation explores the robustness of the metric.

Methods: SSIM is a local metric between 0 and 1 where a higher number indicates better correspondence. It is defined between two patches of images X, and Y as: $SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$, where $\mu_{x/y}$ is the mean intensity in the patch of image X and Y respectively. $\sigma_{x/y}$ is the standard deviation of values of each image, and σ_{xy} the covariance between the two images. $c_{1/2}$ are constants that are determined based on the dynamic range, L, of the images such that $c_1=(0.01L)^2$ and $c_2=(0.03L)^2$. However, default implementations of the metric vary. Two commonly used implementations are shown in Table 1. The effect of varying definitions of L and windowing was compared between the implementations.

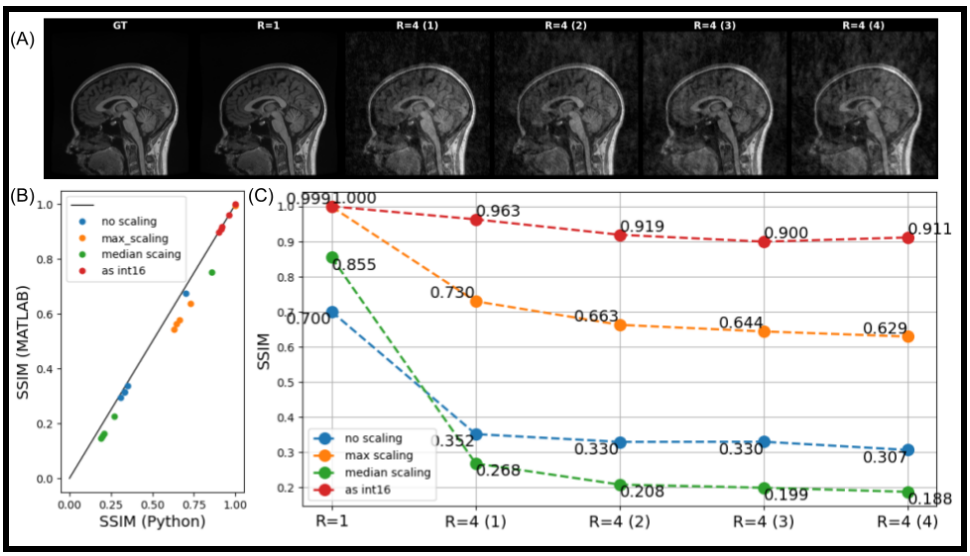

Fig. 1 - (A) Images used in SSIM comparisons. (B) SSIM measurement differences between MATLAB and scikit.image (Python). (C) SSIM measurements for the different images with different input scaling or data type.

Package name	Dynamic range determination	Patch selection
MATLAB	intX or uintX = 2^X - 1 single or double = 1	Gaussian window SD=1.5
Scikit.image (Python)	intX or uintX = 2^X - 1 float = 2	7x7 window

We compared the effect of small data differences and resulting normalization. A simple structural brain image was transformed to k-space data using an FFT. Complex Gaussian noise was then added and the data was undersampled by a factor R=4 using four different Poisson disk sampling patterns. Reconstruction was then performed by a simple zero-filled iFFT (Fig. 1A).
The resulting images were compared with the original image using the default SSIM implementations in MATLAB and scikit.image. By default MATLAB had the data stored as *double* and Python had it stored as *float*. The GT values were in the range [0.0, 5710.0]. The reconstructions had a lower dynamic range (e.g. the first R=4 reconstruction had a range [1.2, 4413.9]. For each assessment the input images were either divided by their maximum or median value, or simply converted from float/double to int16. Code for experiments are shared on https://github.com/SophieSchau/ssim_as_metric.

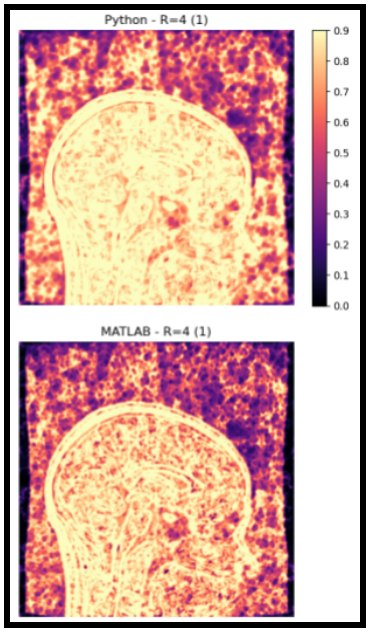

Fig. 2 - SSIM map for scikit.image and MATLAB implementations (R=4 (1) scaled by the image maximum). Windowing affects mean result. Background has large effect on mean SSIM

Results: MATLAB and scikit.image had a clear bias between them when the input data was scaled. Better correspondence was observed for no scaling and using the same data type - int16 (Fig. 1B). Scaling the input data affected the SSIM values drastically. For the same undersampled reconstruction, results varied between 0.18 and 0.9 (almost the full range of SSIM) depending on how the data was normalized or what data type was used. The ordering of the different undersampled reconstructions were also not constant among different SSIM estimation methods (Fig. 1C). SSIM map differences between MATLAB and scikit.image were also evident because of different windowing methods used in the implementations (Fig. 2).

Discussion: The results show that comparing SSIM using the same input data type (int16) has the largest correspondence between scikit.image and MATLAB and is most in line with recommendations for setting L correctly [1]. However, this implementation also had a smaller dynamic range. In these experiments, and many uses of SSIM, the effect of background was neglected, but can affect the results as seen in Fig. 2. It is also not clear what windowing approach is the best for medical imaging where images of varying resolution are studied.

Conclusion: Data type, scaling, and dynamic range should be carefully controlled when reporting SSIM. For benchmarking purposes these are important considerations.

References: [1] Zhou, W., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*. Vol. 13, Issue 4, April 2004, pp. 600–612.