Authors' accepted manuscript of the article published in
2015 IEEE International Conference on Image Processing (ICIP)
http://dx.doi.org/10.1109/ICIP.2015.7351345

LIMITATIONS OF THE SSIM QUALITY METRIC IN THE CONTEXT OF DIAGNOSTIC IMAGING

Jean-François Pambrun and Rita Noumeir

École de technologie supérieure (ÉTS)
Department of Electrical Engineering
1100 Notre-Dame Ouest, Montréal, Canada

ABSTRACT

Lossy image compression is increasingly used in medical applications, but great care must be taken to ensure that no diagnostically relevant features are altered. Guidelines based on compression ratios are often use to mitigate this issue, but are criticized due to the considerable compressibility variations between images. Objective image quality assessment metrics should be used instead, but the most common, mean squared error, is known to be poorly correlated with our perception of quality. Structural similarity (SSIM) is probably currently the most popular alternative, but it is also increasingly criticized. Using computed tomography simulations, this paper shows some of the limitations of SSIM when used with medical images : uniform pooling, distortion underestimation near hard edges, instabilities in regions of low variance and insensitivity in regions high intensities. Furthermore, this paper demonstrates the effect of these limitations when SSIM is used to bound compression in a block coder such as JPEG 2000.

Index Terms— SSIM, diagnostic imaging, medical imaging, image compression, image quality assessment

1. INTRODUCTION

Efficient image compression is required in the medical domain to handle the increasing amount of data generated by diagnostic imaging devices. Lossless compression techniques can help reduce storage requirements and increase effective transfer rates by reducing file sizes by up to two thirds, but lossy compression is needed to achieve better performance. However, in doing so, great care must be taken to ensure that no diagnostically relevant features are altered.

To address this issue, many two-alternative forced choice (2AFC) studies were conducted with trained radiologist in order to find safe and optimal compression ratios for different modalities. These studies are now the basis of local and national guidelines. However, it has been suggested[1] that image compressibility varies widely with image content, even within modalities, and that the use of accurate quality metrics is required to establish viable guidelines. The current go-to objective quality metric, mean squared error (MSE), is widely known to be poorly correlated with human perception of image quality and many alternatives have since been proposed. Structural similarity[2] (SSIM) is one of these new metrics that are getting the most attention, but its limitations are not well understood. This paper will shed lights on some of these limitations from the perspective of diagnostic imaging. Specifically, we will explore the issues related to uniform pooling, distortion underestimation near hard edges and regions of insensitivity or instabilities using computed tomography phantom simulations. Furthermore, we will illustrate the challenges of using SSIM to replace MSE to bound compression in block coder such as JPEG 2000.

2. PREVIOUS WORK

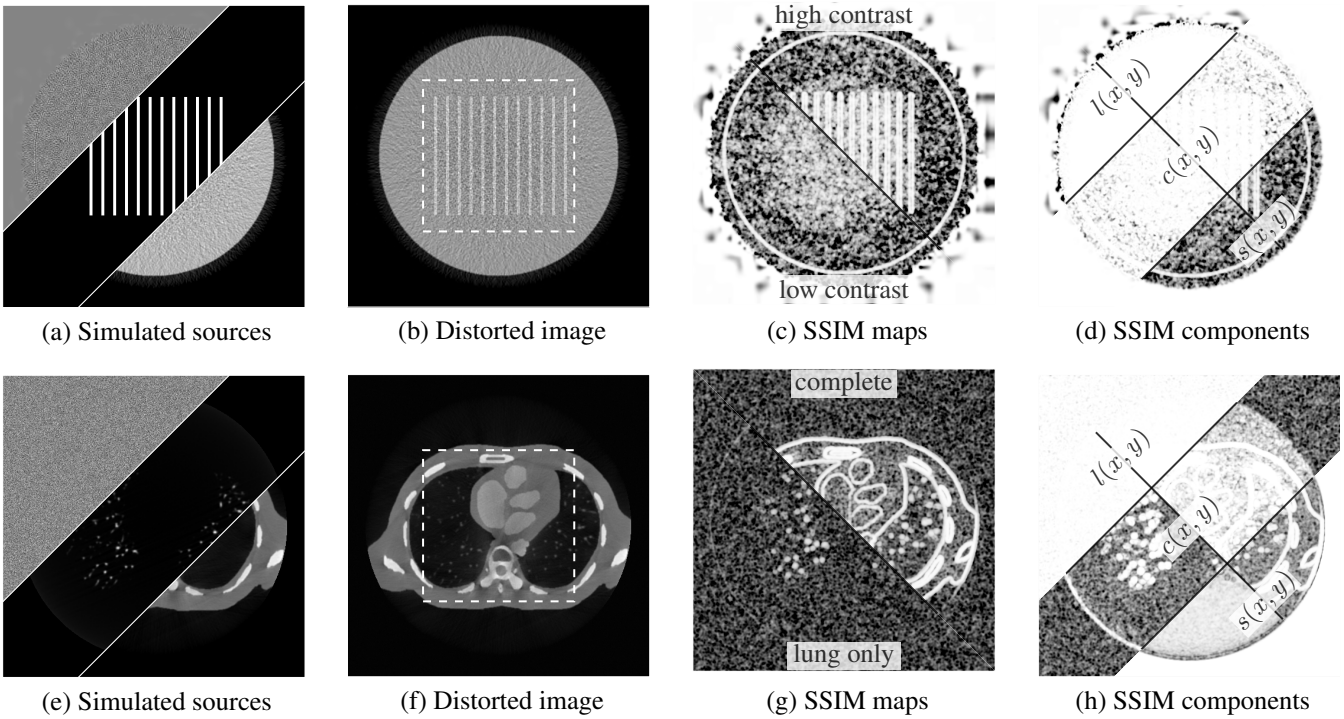
Since its publication, SSIM has grown increasingly popular in the image and video compression fields. It has already been used to replace MSE in rate-distortion optimisation algorithms[3, 4] in order to optimize or bound compression. However, the effectiveness of SSIM was mostly demonstrated with natural image databases subjected to heavy distortions such as the LIVE image quality assessment database[5]. These databases are important tools used to evaluate the performance of image quality assessment (IQA) metrics in the context of streamed videos or pictures displayed on web pages, but their results may not translate well to text images, graphics or diagnostic images. Some studies involving diagnostic images and trained radiologist have shown SSIM to be either on par[6, 7] with or better[8] than MSE, but they have not explored its limitations.

SSIM is computed from three distinct terms : luminance (mean), contrast (variance) and structure (correlation) which are respectively defined as follow :

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$



(a) Simulated sources (b) Distorted image (c) SSIM maps (d) SSIM components

(e) Simulated sources (f) Distorted image (g) SSIM maps (h) SSIM components

Fig. 1. Source images and their associated SSIM maps. a) Noise (top-left), grating signal (center) and background water cylinder (bottom-right) used to compose image b). c) SSIM map using low (bottom-left) and high (top-right) contrast grating signals. d) individual SSIM components of c). e) Noise (top-left), lung (center) and thoracic (bottom-right) phantom simulations used to compose image f). g) SSIM maps of the lung only(bottom-left) and complete thoracic simulations(top-right). h) individual SSIM components of g).

The statistics, μ, σ and σ_{xy}, are all computed in a sliding Gaussian weighted window usually about 11 pixels wide. The three terms are then multiplied together at each pixel location to produce a SSIM map which is subsequently uniformly pooled to obtain a single SSIM index. An index of one indicates a perfect reconstruction while zero is the lower bound.

An empirical and formal analysis[9] showed evidence of a close relationship between SSIM and MSE suggesting that the performance of the SSIM may be much closer to the MSE that we might expect. Furthermore, a thorough theoretical investigation[10] of SSIM showed that 1) SSIM and MSE are composed of the same parameters combined in different ways, 2) luminance and variance terms of SSIM are dice coefficients and depends on the absolute values of the input parameters (μ and σ) and 3) it is unstable when these input parameters are close to zero. SSIM's dependence on input parameters may be especially problematic with diagnostic images because of the increased dynamic range and the presence of large regions of low variance or average. Figure 2 shows the effect of this dependency when distortions of 5 in average and variance, $\mu_y = \mu_x + 5$ and $\sigma_y = \sigma_x + 5$, are plotted against base parameters, μ_x and σ_x, ranging from 0 to 30. Even with recommended regularisation parameters C_1 and C_2, which were introduced for this purpose, the steep slopes means that SSIM is still very unstable when the variance or luminance of the reference image are low. Furthermore, these terms become insensitive when the base variance or mean are very high. These conditions are fairly uncommon in natural images, but are ubiquitous in medial imaging.

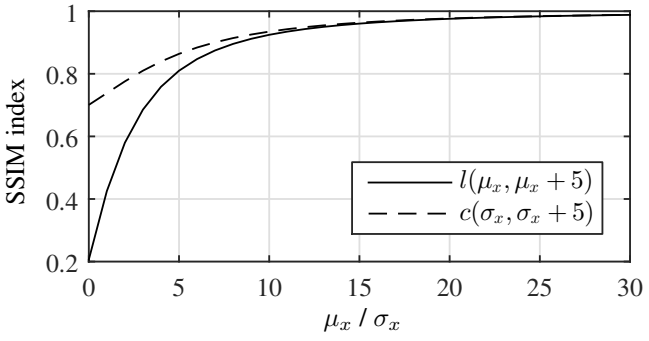


Fig. 2. $l(\mu_x, \mu_y)$ and $c(\sigma_x, \sigma_y)$ terms with $\mu_y = \mu_x + 5$ and $\sigma_y = \sigma_x + 5$ plotted against low values of mean and variance with recommended stabilising parameters C_1 and C_2.

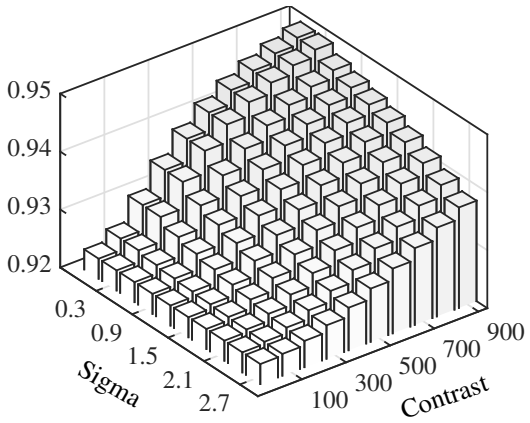
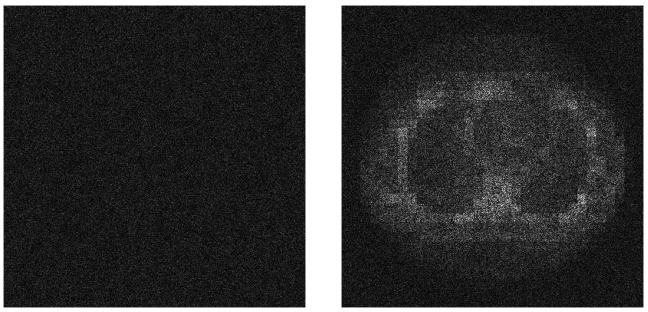


Fig. 3. SSIM index of the center region of Fig. 1b with constant noise and a grating pattern of varying contrast and hardness.



(a) Constant MSE (b) Constant SSIM

Fig. 4. Random Gaussian noise independently generated on blocks of 16×16 pixels to obtain a) constant MSE and b) constant SSIM.

3. METHODOLOGY

The following experiments were performed to illustrate the aforementioned limitations of SSIM in the medical domain. In both cases, the SSIM index were computed on the full range images and was configured for a dynamic range of 2500. Several stabilisation parameters were tested, but default values were used as they provided the best compromise between sensitivity and instability. SSIM indexes were computed using uniform pooling, however distortion (DW-SSIM) and information (IW-SSIM)[11] wighted techniques showed similar results. Both phantoms are simulated using a GPU-accelerated Monte Carlo x-ray transport simulation (MC-GPU)[12] to generate a realistic computed tomography image.

The first phantom(Fig. 1b) is composed of a water cylinder surrounded by air with an added grating pattern of different hardness, simulated by a Gaussian filter with a sigma ranging from 0.3 to 2.7, and a contrast between 0 and 1000. Such high contrast is common in computed tomography image which are represented on the Hounsfield scale where air is defined as -1000 Hounsfield Unit (HU), water as 0 HU and bones are represented by value above 1000 HU. The same noise image (Fig. 1a, top-left), obtained by extracting the difference image between the reference phantom and a 24:1 JPEG 2000 compressed version, was used throughout the experiment. SSIM indexes were computed for the entire image and for a region of interest (ROI) shown in Figure 1b with dashed lines. The SSIM index of the entire image is higher because the region outside of the acquisition field of view (FOV) is mostly unaffected by compression. The $l(x, y)$ term has almost no influence in the ROI because μ_x is in the insensitive zone of the dice coefficient. On the other hand, outside of the FOV where there is little variations, $l(x, y)$ and $c(x, y)$ are in their unstable zones leaving almost only the stabilising terms. Figure 1c and d shows SSIM maps with different contrast and, even though the noise is strictly identical, they are significantly different in the ROI because hard edges propagate high SSIM values across the sliding window. Figure 3 show the SSIM indexes of the ROI with different hardness and contrast.

The second phantom was used to simulate the type of images used by radiologists to search for lung nodules. The background is a thoracic phantom complete with bone, muscles and soft tissues (Fig. 1e, bottom-right) and randomly generated additive Gaussian noise with a variance of 20 (Fig. 1e, top-left) is used. An offset is added to both reference and distorted images to ensure there are no negative values. In this case, the diagnostically relevant information is exclusively located in the low contrast structures of the lungs shown in Figure. 1e (center). The lung and background are simulated separately using MC-GPU and fused together in the sinogram domain thus keeping the diagnostically relevant information and the background apart. Fusing the background to the lung simulation increases the SSIM index from 0.9395 to 0.9628 in the center ROI (shown in Fig. 1f) even though the noise remains constant and the diagnostic value in unchanged. SSIM maps in Fig. 1 g) and h) shows the same behavior around edge as in the previous example. Again, $l(x, y)$ does not contribute in the ROI and $c(x, y)$ is in the unstable zone outside of the FOV and for most of the lung only image.

Finally, because it has been suggested that SSIM is better correlated with our perception of quality than MSE, it is tempting use it to replace MSE in the post compression rate-distortion optimisation algorithms of image coders. Unfortunately, the limitations exposed earlier have an almost prohibitive effect on this approach. Figure 4 compares the noise needed to a) produce a constant PSNR of 30 and b) a constant SSIM of 0.99 on the thoracic phantom in non-overlapping windows of 16×16 pixels. This results is similar to what would have been obtained with a JPEG 2000 coder bounded by SSIM with code-blocks of 8×8 coefficients. SSIM produces significantly more artefacts around edges in accordance with the visual masking model traditionally used with natural

images. In practice, however, the level of distortion introduced around edges is not visually equivalent.

4. DISCUSSION

Several observations can be derived from these results. In both our experiments, the luminance terms did not contribute to the SSIM index in the ROI because the average pixel values placed it in the insensitive zone of the dice coefficient. Conversely, the area outside of the phantoms are in the unstable zone of both luminance and variance terms because their mean values are very close to zero. Because the image statistics are heterogeneous, it was not possible to adjust the stabilising parameters to obtain good performance across all regions. These conditions, common to diagnostic images, combined with uniform pooling results in unreliable SSIM measurements.

Furthermore, hard edges saturates the $c(x, y)$ and $s(x, y)$ terms for the width of the sliding window resulting in overestimation of the SSIM index. This may be considered in line with the principles of visual masking that occurs when a strong signal overshadows another, but, in this case, the effect appears too severe. Assumptions based on natural images and typical imaging applications may be inappropriate for the medical domain. For instance, a faint signal essential to diagnostic close to a high contrast structure (ex. a lung nodule near soft tissues) could be considered visually masked and of no value in another context. This issue could be mitigated by slightly modifying the SSIM algorithm to use edge preserving filtering techniques instead of the Gaussian weighted average window to compute the required statistics.

5. CONCLUSION

In this paper, we have shown through analytical and empirical evidence that : 1) because dice coefficients are dependent on the base value, the luminance term can be either unstable or insensitive in regions of low or high average intensities, 2) the variance term can be unstable in regions of low variance, 3) the variance and structure terms underestimate distortions near hard edges and finally 4) uniform pooling should be avoided when image statistics are not fairly homogeneous. These characteristics are common with diagnostic imaging and, consequently, great care must be taken when using SSIM in the medical domain. Moreover, SSIM, in its current form, in not well suited to replace MSE in rate-distortion allocation algorithms of image coders. Future work should include developing alternative pooling techniques and exploring the use of edge preserving filtering techniques in SSIM.

6. REFERENCES

[1] J-F. Pambrun and R. Noumeir, "Compressibility variations of JPEG2000 compressed computed tomography," in *Proc IEEE Eng Med Biol Soc*, 2013, pp. 3375–3378.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process*, vol. 13, no. 4, pp. 600–612, 2004.

[3] S. Wang, S. Ma, and W. Gao, "SSIM based perceptual distortion rate optimization coding," *Image Processing*, vol. 7744, no. 1, pp. 774407–8, 2010.

[4] T. Richter and KJ. Kim, "A MS-SSIM optimal JPEG 2000 encoder," in *Data Compression Conference*. 2009, pp. 401–410, Ieee.

[5] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik, "LIVE image quality assessment database release 2," `http://live.ece.utexas.edu/research/quality.`, 2009.

[6] KJ. Kim, B. Kim, R. Mantiuk, T. Richter, H. Lee, H-S. Kang, J. Seo, and KH. Lee, "A comparison of three image fidelity metrics of different computational principles for JPEG2000 compressed abdomen CT images," *IEEE Trans. Med. Imag*, vol. 29, no. 8, pp. 1496–1503, 2010.

[7] V. Georgiev, A. N Karahaliou, S. Skiadopoulos, N. Arikidis, A. D Kazantzi, G. S Panayiotakis, and L. Costaridou, "Quantitative visually lossless compression ratio determination of JPEG2000 in digitized mammograms," *Journal of digital imaging*, vol. 26, no. 3, pp. 427–439, 2013.

[8] I. Kowalik-Urbaniak, D. Brunet, J. Wang, D. Koff, N. Smolarski-Koff, E. Vrscay, B. Wallace, and Z. Wang, "The quest for 'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images," in *SPIE Medical Imaging*, 2014, pp. 903717–16.

[9] R. Dosselmann and XD. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 81–91, 2011.

[10] G. Palubinskas, "Mystery behind similarity measures MSE and SSIM," *Proc. of ICIP, IEEE*, 2014.

[11] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.

[12] A. Badal and A. Badano, "Accelerating monte carlo simulations of photon transport in a voxelized geometry using a massively parallel graphics processing unit," *Medical physics*, vol. 36, no. 11, pp. 4878–4880, 2009.