Student Conference Proceedings
Vol 1, No 1, 2025, Article ID 1937, 4 Pages

Student Conference Lübeck

Proceedings Article

Evaluation of Full-Reference Image Quality Assessment Metrics for Artifact Sensitivity in Lung CT Images for Radiotherapy

Cassandra Krause[a,*] · Goran Stanic[b] · Kristina Giske[b] · Mattias Heinrich[c]

[a] Student of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[b] Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[c] Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[*] Corresponding author, email: cassandra.krause@student.uni-luebeck.de; mattias.heinrich@uni-luebeck.de

Received 03 February 2025; Accepted 06 June 2025; Published online 15 July 2025

Abstract

In adaptive radiotherapy, the quality of images acquired during treatment has to be determined. For this purpose, the sensitivity of full-reference image quality assessment metrics to CT artifacts in lung images is investigated. From an examplary patient image, distortion-free and distorted images were created by simulating artifacts in the image and sinogram space. Two experiments were conducted to analyze the metric sensitivity to realistically distorted images and the behavior for different distortion levels. The results show that pixel-level intensity difference (PID) metrics are sensitive to global transformations, such as rotation, shift, and motion. Gradient-based metrics like gradient magnitude similarity deviation (GMSD) react to blurring, truncation and aliasing artifacts, whereas a histogram-based metric like Jensen–Shannon divergence (JSD) can be used for noise, breathing, metal and other artifacts. A combination of a PID metric, GMSD and JSD provides a comprehensive quality assessment in the context of longitudinal image monitoring along radiotherapy treatment.

I. Introduction

In radiotherapy, the treatment plan is calculated on a planning CT (pCT), which is acquired prior to treatment. During treatment, a series of images is acquired to visualize anatomical changes that could have an impact on the required dose distribution, triggering plan adaptation. This is a typical phenomenon for lung cancer patients, as changes in position, size and breathing patterns are likely. Further, artifacts can influence the quality of images, which consequently affects the quality of the adapted plan [1]. In general, the gold standard for judging the quality of CT images and CT-pCT fusions is visual assessment by specialized radiooncologists. However, this is very observer-dependent and the impact on dosimetric changes is not clear. Having reliable automated algorithms that can determine the quality changes of longitudinal images compared to pCT more efficiently would be beneficial. This gives rise to a research area focusing on full-reference image quality assessment (FR-IQA) metrics, which measure the similarity between two images [2].

Wang et al. and Ohashi et al. [2]-[3] analyzed various FR-IQA metrics for CT images and compared them to subjective radiologists' assessments. However, their analysis was limited to two types of distortion - blurring and noise - which do not cover the sensitivity of the given metrics on various relevant artifacts of CT images.

The objective of this work is to investigate the sensitivity of various metrics on different CT-specific artifacts and find a combination of metrics that are able to give an informative impression of the image quality. To achieve

this, a three-component metric evaluation system was implemented. Subsequently, the results are presented and discussed, followed by a brief summary and an outlook on future research.

II. Materials and Methods

In the following, the implemented system for the generation of distorted images, the comparison of these images, and the visualization are explained in detail.

II.I. Metric Evaluation System

The system implemented to evaluate metrics, called the Metric Evaluation System (MES), consists of three components: the Artifact Simulator, the Image Comparison Module, and the Metric Visualizer. A distortion-free CT image is loaded into the Artifact Simulator, which generates a series of specified distorted images that contain typical CT imaging artifacts. When a pair of original and distorted images is given to the Image Comparison Module, different FR-IQA metrics are calculated between them. The resulting values are passed into the Metric Visualizer, which plots the metric values for each transformation.

II.I.1. Artifact Simulator

The goal of the Artifact Simulator is to generate CT-specific artifacts on a given CT image. Artifacts in CT images are mainly additional structures within the reconstructed data that are not present in the object during image acquisition [4].

The Artifact Simulator takes two inputs: the original image and an artifact ID. Each artifact has additional parameters that specify the transformation, such as kernel sizes or angles. The input image has to be normalized to the range [0, 1] by applying Min/Max Scaling. The resulting transformed image is also clipped to this range. Since there is a need for paired data in order to calculate the metrics, artifacts are added to the original image. The following radiotherapy-typical artifacts are implemented: *rotate*, *blur*, *shift*, *noise*, *ring*, *cupping*, *truncation*, *breathing*, *aliasing*, *metal*, *motion*, *shading* and *streaking*. Rotate, blur, shift and cupping are applied in the image space, whereas the other more complex ones are created by manipulating the sinogram. The sinogram is generated using the *skimage radon transform* method. For the backpropagation the *skimage iradon sart* method is applied with reconstruction angles in the range [0,360], whereas for breathing artifacts, a reconstruction angle in the range [0,180] is used [4]-[5]. Figure 1 shows resulting distorted images for an exemplary set of artifacts.

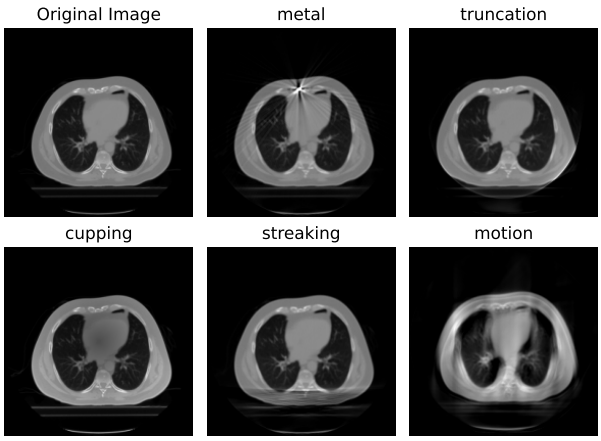

Figure 1: Transversal slice through the thorax of a representative lung cancer patient CT: original CT scan and the distorted images for metal, truncation, cupping, streaking, and motion artifacts.

II.I.2. Image Comparison Module

The Image Comparison Module calculates metric values between the original and its distorted version. As input, it takes an image pair and a metric ID. The investigated metrics can be divided into different groups according to how they are calculated: *pixel-level intensity difference (PID) metrics*, *structural similarity metrics*, *information-theoretic metrics* and *correlation-based metrics* [6].

PID metrics determine the similarity between two images by comparing the corresponding pixel intensities. For this class, mean absolute error (MAE), mean squared error (MSE), normalized root mean squared error (NRMSE), and peak signal-to-noise ratio (PSNR) are investigated.

Structural similarity metrics focus on perception, analyzing features like luminance and contrast. This group includes structural similarity index (SSIM), multiscale SSIM (MS-SSIM), and feature similarity index (FSIM).

Information-theoretic metrics analyze shared information and feature fidelity. Thereby, normalized mutual information (NMI), Jensen-Shannon divergence (JS) and visual information fidelity (VIF) are used.

The correlation-based metrics evaluate similarity by assessing the linear relationship or gradient consistency. Here, the normalized cross correlation (NCC) and the gradient magnitude similarity deviation (GMSD) are investigated.

II.I.3. Metric Visualizer

The Metric Visualizer plots the different resulting metric values per metric. For comparability, the same image pair is used for all metrics. Additionally, the best-achievable-case value is visualized for each metric, which is calculated between the original image and its identity.

10.18416/SCP.2025.1937

II.II. Experiments

A 2D CT image of the chest is used to analyze the metrics. Two experiments were conducted to determine the sensitivity of the metrics to the artifacts. In experiment 1 a plot for one distortion is created by using one realistic transformation for each artifact. The parameters are empirically selected based on the visual perception and expert opinion. The investigated parameters for both experiments are shown in Table 1. For metal, shading, and streaking artifacts, there are no additional parameters.

Table 1: Overview of the applied parameters in the generation of the artifacts. Experiment 1 (Exp. 1) shows the values for the analysis of one realistic version of the images. Experiment 2 (Exp. 2) shows the ranges for the analysis of multiple levels of distortion. The range is given as follows: start, stop, step size.

Artifact	Parameter	Exp. 1	Exp. 2
rotate	angle	5	1 to 19 in 2°
blur	kernel size	5	3 to 21 in 2 px
shift	shift in x/y	5	1 to 19 in 2 px
noise	noise level in %	0.05	0.01 to 0.1 in 0.01 %
ring	number of rings	25	10 to 28 in 2 rings
cupping	strength	0.7	0.1 to 1 in 0.1 steps
truncation	threshold from bottom	25	10 to 28 in 2 rows
breathing	number of streaks	300	50 to 410 in 40 streaks
aliasing	repeat factor	3	1 to 10 in 1 columns
motion	threshold for shift	100	1 to 19 in 2 rows

Experiment 2 investigates how the metrics react to different levels of distortion. To this end, the metrics were calculated on images with varying levels of distortion, which were adjusted via the artifact parameters. Metal, shading, and streaking artifacts are excluded, as they are not adjustable.

III. Results and Discussion

Figure 2 shows the results for experiment 1. The resulting metric values for four metrics are presented, demonstrating the sensitivity of each metric group to different distorted images. For all PID metrics, the bar plot shows that the artifacts with the highest dissimilarity are the rotation, shift, cupping, and motion artifacts, which are exemplarily presented by the MSE and the PSNR - artifacts affect the entire image or a significant portion of it. For PID metrics even small changes across a large number of pixels can result in a substantially higher error compared to localized artifacts.

Concerning the information-theoretic metrics, the JSD has the highest dissimilarity for cupping, breathing, and motion artifacts. It also delivers higher dissimilarities for aliasing, metal, noise, and shading artifacts compared to the other artifacts. The GMSD shows the highest sensitivity for the rotation, shift and motion artifact.

Analyzing one realistically deformed image could be misleading, since the metric magnitude is not the only characteristic of the sensitivity. It is also important to

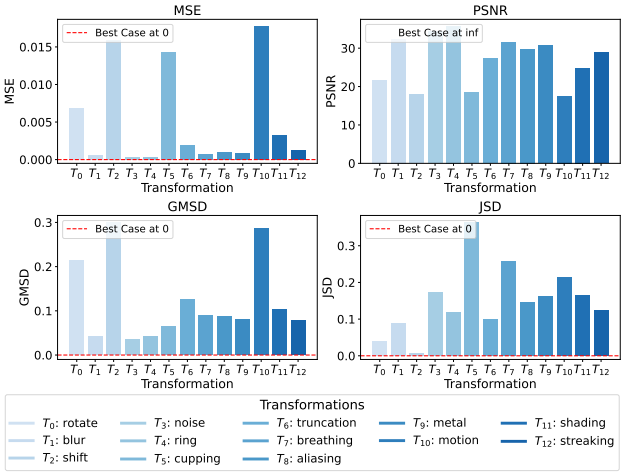

Figure 2: Results of the first experiment of the comparison of one realistic version of the artifacts. Here, the plots are presented for MSE, PSNR, SSIM, and JSD for all artifacts.

analyze the magnitude change of metrics when the distortion is enhanced. The changes for different distortion levels, investigated in experiment 2, are visualized in Figure 3. Here, the PID metrics, which are represented by the MSE and PSNR, also show that shift, cupping, motion and rotation are most sensitive to such metrics, and the metric values also increase with the level of distortion. JSD is most sensitive for cupping but motion, truncation, blur, and aliasing also lead to an increase in the metric value for a higher level of distortion. The GMSD is sensitive to shift, motion, rotation, aliasing, truncation, as well as blurring.

These results highlight that there is no one metric that is sensitive to all CT-specific artifacts. Therefore,

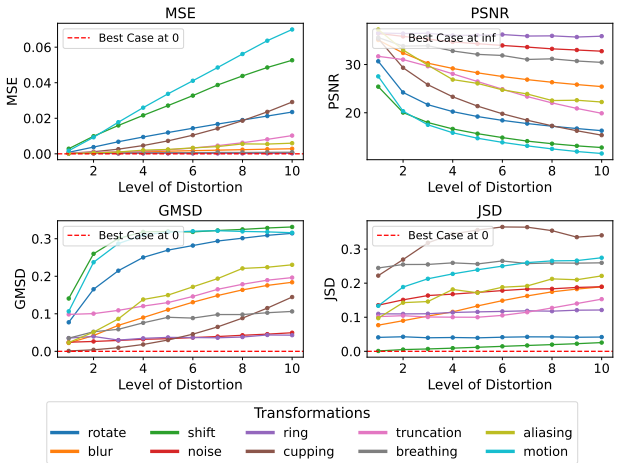

Figure 3: Results of the second experiment comparing multiple levels of distortions of the artifacts. The plots are presented for MSE, PSNR, SSIM, and JSD for all artifacts implemented to allow changes in distortion strength.

10.18416/SCP.2025.1937

it is useful to use a mixture of metrics combining the sensitivity to different artifacts.

A combined analysis of both experiments was conducted to rank the metrics and artifacts into categories of good, medium, and bad. This ranking was based on the intensity of their magnitudes relative to other metrics in experiment 1 and on whether their values increased or decreased with the level of distortion in experiment 2.

After ranking the metrics for both experiments, the best combination is chosen, identifying sensitive metrics for each artifact as shown in Table 2. It can be seen that integrating one PID metric, GMSD, and JSD is essential to ensure a comprehensive assessment of all artifact types. Including results from all three metrics, it becomes possible to identify which artifacts are present and how they may impact the image. It should be mentioned that so far this study was applied only to a single image. Analyzing a broader image set would enhance the metric stability and avoid the location-specific nature of the artifacts.

Table 2: Overview of the metrics which are most sensitive for the artifacts in the investigated experiments.

Artifacts	Sensitive Metrics
rotate, shift, motion	MAE, MSE, NRMSE, PSNR, GMSD
blur	FSIM, VIF, GMSD
noise	PSNR, SSIM, FSIM, VIF, JSD
breathing, metal, streaking	JSD
cupping	MAE, MSE, NRMSE, PSNR, JSD
truncation	GMSD
aliasing	PSNR, SSIM, FSIM, VIF, GMSD, JSD
shading	MAE, MSE, JSD

In radiotherapy, the automated metric monitoring could standardize the image assessment by reducing the reliance on subjective visual evaluations. Furthermore, automated systems can detect subtle variations in image quality that might be missed by human observers. This capability enables the early identification of issues that could potentially compromise dosimetric accuracy. By integrating these methods into clinical practices, treatment planning can be optimized while maintaining high-quality care and minimizing the risk of errors.

IV. Conclusion

The goal of this study is to investigate the sensitivity of several FR-IQA metrics regarding the image quality degradation in lung CT images. Various artifacts were simulated in the image or sinogram space to achieve paired data of distortion-free and distorted images. First, experiment 1 is conducted to determine the metric values for one realistic version of the artifacts. Secondly, the change of the metric values was analyzed by applying different levels of distortion. It was shown that the PID metrics are sensitive to global transformations or those which manipulate bigger regions, e.g., rotation, shift, cupping or motion. The GMSD could be used for blur, truncation and aliasing artifacts. The JSD is practicable for noise, breathing, metal, streaking, cupping, aliasing or shading artifacts. It was shown that there is no single solution for all artifacts. Further analysis indicates that a combination of one PID metric, the GMSD and the JSD could give a more complete impression of the quality of an image.

Since this study was only applied to one image, there should be further research into applying it to different images to achieve an averaged result. Moreover, to analyze how realistic the created images are, the judgment of an experienced radiologist should be included. It could also be beneficial to focus on the artifacts that are most relevant for radiotherapy. This approach would allow the development of more specifically tailored metrics that focus on detecting and evaluating the most relevant artifacts. If these adapted metrics manage to only react to a specific artifact, it would open up possibilities to build an artifact detector system for clinical use.

Acknowledgments

The work has been carried out at the department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ) and supervised by the Institute of Medical Informatics, Universität zu Lübeck.

Author's statement

Conflict of interest: Authors state no conflict of interest. DeepL and ChatGPT were used for the linguistic fine-tuning of this manuscript.

References

[1] A. Cole, C. Veiga, U. Johnson, D. D'Souza, N. Lalli, and J.McClelland. Toward adaptive radiotherapy for lung patients: Feasibility study on deforming planning ct to cbct to assess the impact of anatomical changes on dosimetry. *Physics in Medicine and Biology*, 63, 17 2018, doi:10.1088/1361-6560/aada96.

[2] G. Wang, C. Sun, Y. Liu, and H. Yang. Optimization of ct image quality assessment metric based on genetic algorithm. *2022 IEEE 10th International Conference on Computer Science and Network Technology, ICCSNT 2022*, pp. 14–18, 2022, doi:10.1109/ICCSNT56096.2022.9972914.

[3] K. Ohashi, Y. Nagatani, M. Yoshigoe, K. Iwai, K. Tsuchiya, A. Hino, Y. Kida, A. Yamazaki, and T. Ishida. Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *Journal of Digital Imaging*, 36:2623–2634, 6 2023, doi:10.1007/S10278-023-00875-0/FIGURES/8.

[4] J. F. Barrett and N. Keat. Artifacts in ct: Recognition and avoidance. *Radiographics*, 24, 6 2004, doi:10.1148/RG.246045065.

[5] F. E. Boas and D. Fleischmann. Ct artifacts: Causes and reduction techniques. *Imaging in Medicine*, 4:229–240, 2 2012, doi:10.2217/iim.12.13.

[6] A. Bandi, P. Adapa, and Y. Kuchi, The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges, 2023. doi:10.3390/fi15080260.