SciMedCentral
JSM BIOMEDICAL IMAGING DATA PAPERS

Research Article

Big Data from CT Scanning

Qingsong Yang[1], Mannudeep K. Kalra[2#*], Atul Padole[2], Jia Li[3], Elizabeth Hilliard[4], Rongjie Lai[3] and Ge Wang[1#*]

[1]Department of Biomedical Engineering, Rensselaer Polytechnic Institute, USA
[2]Divisions of Thoracic and Cardiac Imaging, Department of Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[3]Department of Mathematics, Rensselaer Polytechnic Institute, USA
[4]Department of Physics and Psychology, Rensselaer Polytechnic Institute, USA
[#]Both authors contributed equally

***Corresponding authors**
Ge Wang, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, USA; Tel: 151-827-637-26; Email: ge-wang@ieee.org
Mannudeep K. Kalra, Divisions of Thoracic and Cardiac Imaging, Department of Imaging, Massachusetts General Hospital, Harvard Medical School, USA. Tel: 161-764-309-53; Email: MKALRA@mgh.harvard.edu

Submitted: 06 March 2015
Accepted: 18 March 2015
Published: 20 March 2015

Keywords
- Computed tomography
- Big data
- Low dose
- Image denoising

Abstract

Over 100-million of x-ray CT scans are performed worldwide each year. In most cases, a scan projection or sonogram data are discarded after images are read. This represents a huge waste of big data, and an opportunity to develop new methods for better image reconstruction and high dose efficiency. Here we present an initial attempt to archive, utilize and share big data from CT scanning. In this project, CT scans of several cadavers are used as examples. The data were collected at Massachusetts General Hospital at multiple different radiation dose levels for different x-ray spectra, and with representative reconstruction techniques. Hence, this database is more informative than others of this kind as prior knowledge to improve image reconstruction and image analysis, and reduce radiation dose.

ABBREVIATIONS

NI: Noise Index; FBP: Filtered Back-Projection; ASiR: Adaptive Statistical Iterative Reconstruction; MBIR: Model-Based Iterative Reconstruction; ROI: Region Of Interest; BM3D: Block-Matching and 3D Filtering; NLM: Non-Local Means Filtering; MSE: Mean Squared Errors

INTRODUCTION

X-ray CT has been an important imaging tool since its invention in the 1970s. Over 100 million of x-ray CT scans are conducted annually around the world. With increasing use of CT, there has been a public concern over the involved x-ray radiation dose and potential risks. Many techniques were then developed to reduce the radiation dose without compromising the diagnostic performance [1,2]. A direct way to reduce the radiation dose is to lower the x-ray tube current and voltage. This will decrease the number or energy of received x-ray photons and increase image noise, which can be handled with more advanced methods such as iterative algorithms. Currently, the penalty terms for iterative image reconstruction are rather generic, such as the sparsity and low-rank requirements. Hence, we are motivated to extract more specific prior knowledge from existing CT scans of the same and other patients and bring the image reconstruction strategy to the next level.

Several relevant techniques were studied by data scientists in the past decades. For instance, data mining methods extract hidden patterns from large data sets; and machine learning methods make predictions based on rules learned from training data. Up to now, these two approaches have achieved tremendous successes in the field of artificial intelligence. Especially, a sparse representation using a trained dictionary has been proved to be an efficient way for image denoising and restoration [3,4]. Deep neuron network learning is another technique that has been adapted to image denoising [5–7]. A recent study on applying dictionary learning to medical image reconstruction was reported in [8].

The maturing machine learning and data mining techniques allow us to utilize existing CT data and images efficiently and effectively. However, there has never been such a dataset available for this purpose. The Visible Human Project is a high-profile data project [9]. Despite its success in human anatomy visualization applications, this project is limited to only two male and female cadavers that were scanned using the conventional CT protocol. As a significant extension, large CT image data of several cadavers were obtained for advanced feature extraction to help image reconstruction. In the following section, we will discuss how these CT data were acquired and reconstructed and how to access the data for new applications. As an example, image denoising was performed to demonstrate the workflow. Finally, we discuss further topics and conclude the paper.

DATA ACQUISITION AND IMAGE RECONSTRUCTION

Similar to the Visible Human Project, the CT images were collected from cadavers. This method has two advantages. First, x-ray dose is not a problem, permitting repeated CT scans with high and low tube currents and offering both gold standards and clinical emulations. Second, the cadaver is stationary, so no motion artifacts exist for perfect image registration.

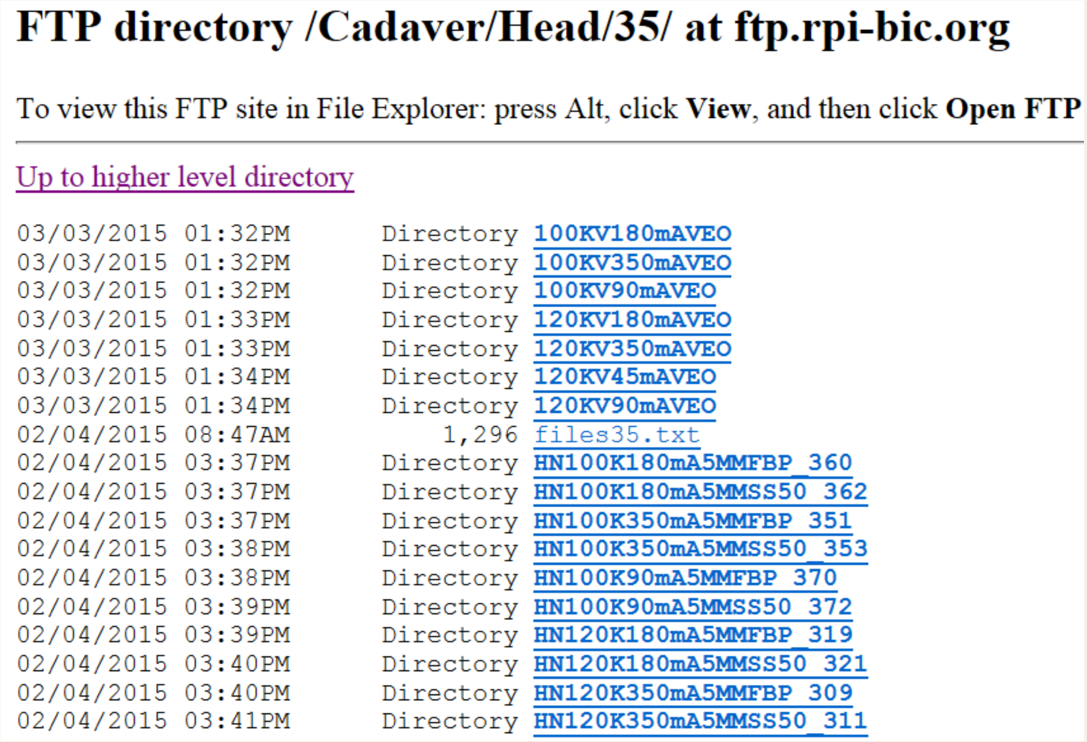
A GE Discovery 750 HD was used for cadaver scanning and image reconstruction. As one of the most popular CT scanners, it was designed for high definition imaging and up to 50% dose reduction [10]. All the cadavers were scanned under 140kVp, 120kVp, 100kVp and 80kVp x-ray spectra. GE uses a noise index (NI) to define image quality. The noise index is approximately equal to standard deviation of CT number in the central region of the image of a uniform phantom [2,11]. In the database, four noise indices of 10, 20, 30 and 40 were acquired for each x-ray spectrum. For complete de-identification, we scanned these cadavers were scanned without including any identifiers (such as name, medical record number, age, date of birth, race, ethnicity, gender, address, scanning physician, and referring physician).

The scanner provides three options for image reconstruction. One is filtered back-projection (FBP), which is the most commonly used technique on the current commercial CT scanners because of its fast speed and high performance in most cases [12]. The second is adaptive statistical iterative reconstruction (ASiR), which improves image quality statistically compared to FBP in the case of noisy data [13,14]. The third reconstruction technique is Veo, which is the world's first commercial model-based iterative reconstruction (MBIR) product. MBIR takes all the data acquisition processes into account during image reconstruction. Experimental results have shown that among these representative reconstruction algorithms, MBIR provides images of the best quality with the lowest dose [15-17]. More technical details on ASiR and MBIR can be found in [18,19] and references cited therein.

CT IMAGES

Except some of these dataset available on this journal webpage, we set up an FTP server and uploaded some of these data. To access the data, users can contact us with the intended use of the data and their contact information. Once this information is received, we initiate regulatory approval process (generally less than 4 weeks) prior to providing free access on the following website http://www.rpi-bic.org/resources/x-ray-ct-image-database/ (Figure 1) shows a snapshot of the web page and (Figure 2) shows the FTP site. Since the CT images cover three parts of the human body (head, chest and abdomen), we organized these images into

Figure 1 A snapshot of data web page.



FTP directory /Cadaver/Head/35/ at ftp.rpi-bic.org

To view this FTP site in File Explorer: press Alt, click **View**, and then click **Open FTP**

Up to higher level directory

```
03/03/2015 01:32PM      Directory 100KV180mAVEO
03/03/2015 01:32PM      Directory 100KV350mAVEO
03/03/2015 01:32PM      Directory 100KV90mAVEO
03/03/2015 01:33PM      Directory 120KV180mAVEO
03/03/2015 01:33PM      Directory 120KV350mAVEO
03/03/2015 01:34PM      Directory 120KV45mAVEO
03/03/2015 01:34PM      Directory 120KV90mAVEO
02/04/2015 08:47AM          1,296 files35.txt
02/04/2015 03:37PM      Directory HN100K180mA5MMFBP_360
02/04/2015 03:37PM      Directory HN100K180mA5MMSS50_362
02/04/2015 03:37PM      Directory HN100K350mA5MMFBP_351
02/04/2015 03:38PM      Directory HN100K350mA5MMSS50_353
02/04/2015 03:38PM      Directory HN100K90mA5MMFBP_370
02/04/2015 03:39PM      Directory HN100K90mA5MMSS50_372
02/04/2015 03:39PM      Directory HN120K180mA5MMFBP_319
02/04/2015 03:40PM      Directory HN120K180mA5MMSS50_321
02/04/2015 03:40PM      Directory HN120K350mA5MMFBP_309
02/04/2015 03:41PM      Directory HN120K350mA5MMSS50_311
```

Figure 2 A snapshot of data FTP site.

the corresponding categories or directories. In each category, a cadaver ID is used for identification. As mentioned before, there are four spectra, four noise indices, and three reconstruction techniques. Image slices associated with the same scanning prototype and reconstruction method are in a single folder that indicates all of this information. In (Figure 3), we provide a tree view of how we organize these data.

The CT images were reconstructed slice by slice and stored in the DCM and IMA formats. Any DICOM viewer, e.g., Clear Canvas and Image J, can be used for visualization. MATLAB has functions *dicominfo* and *dicomread* to load data. In the DCM format, HU values are stored in 16-bit signed integers with an offset of 1024 between the true and stored values. That is, the stored value minus 1024 is the truth. On the other hand, the IMA format uses 16-bit unsigned integers to store the HU values also with the 1024 offset, as in the DCM format. Outside the field of view (FOV), the stored HU values were set to -2000. To show how to read and show the data, an MATLAB example is as follows:

```
display_window = [-160,240];
info = dicominfo ('1.3.46.670589.33.1.10204558493204245575.226
38314293567849424.DICOM');
img0 = dicomread(info); img0 = img0 - 1024;
figure; imshow(img0,window);
info = dicominfo('VA0035B.CT._.0381.0050.2012.02.21.11.51.35.31
2500.80726356.IMA');
img1 = dicomread(info); img1(img1==-2000) = 0; img1 = img1 -
1024;
figure; imshow(img1,display_window);
```

In (Figure 4), we display several CT images at the same position. It is observed from these images that the larger the noise index is, the worse the image quality is. From the same dataset, Veo provides the least noisy reconstruction, followed by ASiR. FBP produces the noisiest images, especially in the case of low-dose data. In (Table 1), the means and standard deviations in regions of interest (ROI) marked by red circles in (Figure 1) are listed.

IMAGE DENOISING

A research utility of these CT images is to test the post-processing techniques for CT imaging. In our study, we compared four image denoising techniques: the total variation minimization (ROF) [20], the non-local means filtering (NLM) [21], the block-matching and 3D filtering (BM3D) [22] and a wavelet frame based method (WFM) [23,24]. The experimental results are in (Figure 2b). Since the HU value ranges from -1024 to maximum 3200 (the upper bound may be different for different images), we used a window transform to scale the image values to 0~1 for convenience. Then, we ran the noise suppression algorithms to obtain the results in (Figure 5). We computed the mean square errors (MSE) to quantify the denoising effect, where the highest quality image (Figure 4c) served as the reference. The representative results are in (Table 2).

DISCUSSION AND CONCLUSION

While the current database is still limited, we are interested in expanding it to cover the CT scans extensively in the future. Given the outstanding environment and facilities at MGH, dual-energy and spectral CT datasets can also be collected. At least, CT images can be utilized after IRB-required modifications. This can be a huge resource for the research community, and eventually

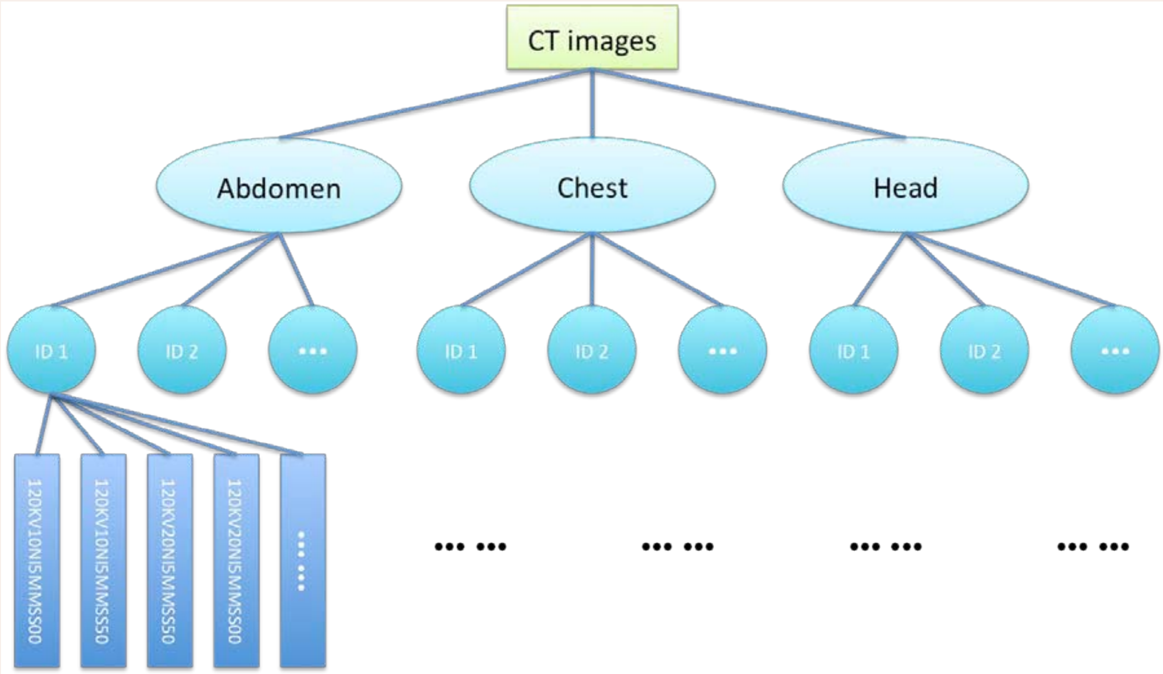


Figure 3 Tree view of the database of CT images.

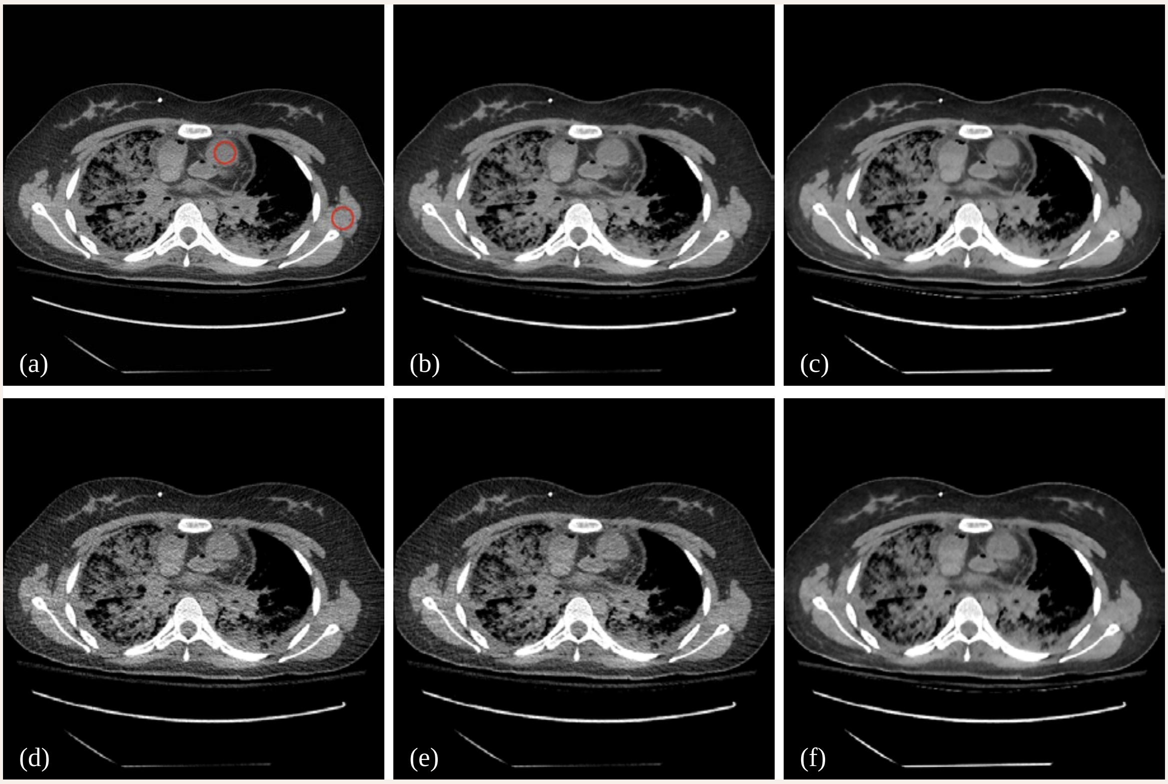


Figure 3 Tree view of the database of CT images.

be impactful on the development of commercial reconstruction software.

In conclusion, we have reported a unique dataset of CT images. One of its key characteristics is the quantity and variety already significantly larger than that from the famous Visible Human Project. Our database contains images at four noise levels under four x-ray spectra and reconstructed using three algorithms respectively. We have described how to access to this dataset through an FTP server and how to read the images. We

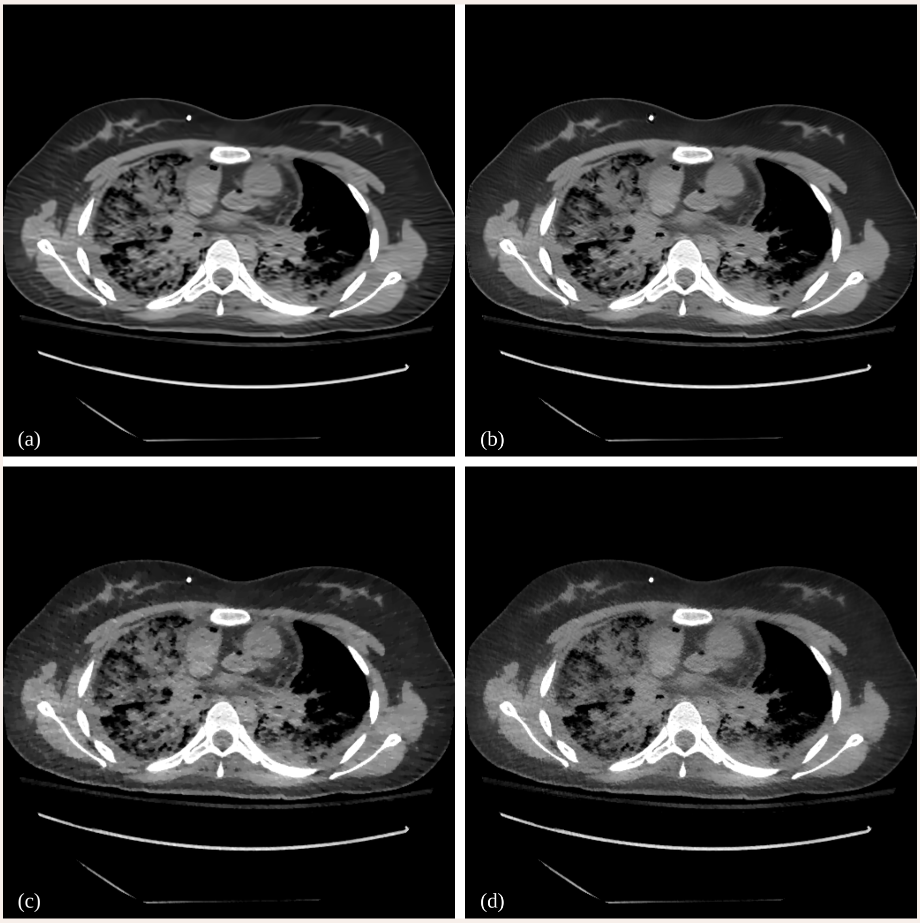


Figure 3 Tree view of the database of CT images.

Table 1: Means and standard deviations of the CT number in two ROIs in (Figure 4).

		NI=10			NI=40		
		FBP	ASiR	Veo	FBP	ASiR	Veo
ROI1	Mean	38.670	38.915	37.739	38.242	36.508	33.331
	Std.	19.247	14.587	10.287	37.619	27.690	11.525
ROI2	Mean	59.721	59.345	63.600	56.909	57.533	62.051
	Std.	18.531	15.033	11.935	32.026	24.059	11.578

Abbreviations: ROI: Region Of Interest; FBP: Filtered Back-Projection; Asir: Adaptive Statistical Iterative Reconstruction; NI: Noise Index

Table 2: Mean square errors of the denoised images using BM3D, NLM, ROF and WFM respectively.

	Noisy data	BM3D	NLM	ROF	WFM
MSE	0.0614	1.404e-4	1.428e-4	1.589e-4	1.417e-04

Abbreviations: MSE: Mean Square Error; BM3D: Block-Matching and 3D Filtering; ROF: Rudin-Osher-Fatemi Image Denoising; WFM: Wavelet Frame Based Method.

have demonstrated a demo application of this dataset, which is an image denoising example. As a follow-up work, we will focus on dictionary learning from big data from CT scanning to help image reconstruction at low-dose level.

ACKNOWLEDGEMENTS

This work was partially supported by NIH under Grant No. U01 EB017140 and Grant No. R01 EB016977.

REFERENCES

1. Yu L, Liu X, Leng S, Kofler JM, Ramirez-Giraldo JC, Qu M, et al. Radiation dose reduction in computed tomography: techniques and future perspective. Imaging Med. 2009; 1: 65-84.
2. McCollough CH, Primak AN, Braun N, Kofler J, Yu L, Christner J. Strategies for reducing radiation dose in CT. Radiol Clin North Am. 2009; 47: 27-40.
3. Elad M, Aharon M. Image denoising via sparse and redundant

representations over learned dictionaries. IEEE Trans Image Process. 2006; 15: 3736-3745.

4. Mairal J, Elad M, Sapiro G. Sparse representation for color image restoration. IEEE Trans Image Process. 2008; 17: 53-69.

5. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. JMLR. Org. 2010; 11: 3371–3408.

6. Burger HC, Schuler CJ, Harmeling S. Image denoising: Can plain Neural Networks compete with BM3D? Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE. 2012; 2392–2399.

7. Jain V, Seung S. Natural image denoising with convolutional networks. Advances in Neural Information Processing Systems. 2009; 769–776.

8. Xu Q, Yu H, Mou X, Zhang L, Hsieh J, Wang G. Low-dose X-ray CT reconstruction via dictionary learning. IEEE Trans Med Imaging. 2012; 31: 1682-1697.

9. Ackerman MJ. The visible human project. Proc IEEE. IEEE; 1998; 86: 504–511.

10. http://www3.gehealthcare.com/en/education/product_education_clinical/tip_applications/computed_tomography_hq_class/discovery_ct750_hd

11. McCollough CH, Bruesewitz MR, Kofler JM Jr. CT dose reduction and dose management tools: overview of available options. Radiographics. 2006; 26: 503-512.

12. Herman G. Fundamentals of computerized tomography: image reconstruction from projections 2009.

13. Silva A, Lawder H, Hara A. Innovations in CT dose reduction strategy: application of the adaptive statistical iterative reconstruction algorithm. 2010.

14. Prakash P, Kalra M. Reducing abdominal CT radiation dose with adaptive statistical iterative reconstruction technique. 2015.

15. Katsura M, Matsuda I, Akahane M, Sato J, Akai H, Yasaka K, et al. Model-based iterative reconstruction technique for radiation dose reduction in chest CT: comparison with the adaptive statistical iterative reconstruction technique. Eur Radiol. 2012; 22: 1613–1623.

16. Deák Z, Grimm JM, Treitl M, Geyer LL, Linsenmaier U, Körner M, et al. Filtered back projection, adaptive statistical iterative reconstruction, and a model-based iterative reconstruction in abdominal CT: an experimental clinical study. Radiology. 2013; 266: 197–206.

17. Pickhardt PJ, Lubner MG, Kim DH, Tang J, Ruma JA, del Rio AM, et al. Abdominal CT with model-based iterative reconstruction (MBIR): initial results of a prospective trial comparing ultralow-dose with standard-dose imaging. AJR Am J Roentgenol. 2012; 199: 1266-1274.

18. Hsieh J, Nett B, Yu Z, Sauer K, Thibault JB, Bouman CA. Recent advances in CT image reconstruction. Curr Radiol Rep. Springer. 2013; 1: 39–51.

19. Beister M, Kolditz D, Kalender WA. Iterative reconstruction methods in X-ray CT. Phys Med. 2012; 28: 94-108.

20. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Phys D Nonlinear Phenom. Elsevier; 1992; 60: 259–268.

21. Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. Computer Vision and Pattern Recognition, 2005 CVPR 2005 IEEE Computer Society Conference on. IEEE; 2005; 60–65.

22. Dabov K, Foi A, Katkovnik V, Egiazarian K. BM3D image denoising with shape-adaptive principal component analysis. SPARS'09-Signal Processing with Adaptive Sparse Structured Representations. 2009.

23. Cai JF, Dong B, Osher S, Shen Z. Image restoration: Total variation, wavelet frames, and beyond. J Am Math Soc. 2012; 25: 1033–1089.

24. Dong B, Li J, Shen Z. X-ray CT image reconstruction via wavelet frame based regularization and Radon domain inpainting. J Sci Comput. Springer. 2013; 54: 333–349.

Cite this article

Yang Q, Kalra MK, Padole A, Li J, Hilliard E, et al. (2015) Big Data from CT Scanning. JSM Biomed Imaging Data Pap 2(1): 1003.