Basics of X-ray CT reconstruction
Principles and applications of iterative reconstruction

Takumi Ohta*

Abstract

This article describes the principles and applications of iterative reconstruction in X-ray computed tomography. We use several real examples to show how iterative reconstruction can produce higher-quality reconstructed images than the conventional reconstruction method.

1. Introduction

X-ray CT (computed tomography) is a method to investigate the internal structure of a sample using X-rays. In principle, CT uses a computer to reconstruct and visualize three-dimensional internal structures of a sample from a collection of its two-dimensional projections. The reconstructed images can then be subjected to various types of analysis. Regardless of the method and purpose of the analysis, it is important to obtain high-quality images to ensure the best possible results.

In CT, images are recovered from a set of projections through a process called tomographic reconstruction. This process is an essential step in CT image processing, and it affects the quality of the reconstructed data in a major way. In certain cases, conventional reconstruction methods are too unreliable and produce images unfit for further analysis. Under these circumstances, iterative reconstruction (IR) methods may perform much better.

This article describes the principles and applications of IR methods. Section 2 describes the procedure of obtaining projection images and notes some important precautions for CT measurements. Principles of conventional reconstruction algorithms are shown in section 3. The principles and features of the IR method are described in section 4. Finally, section 5 showcases situations where IR works well while the conventional reconstruction method struggles.

2. Method of CT Measurement

The main components of a CT apparatus are an X-ray source, a detector, and a sample stage. Depending on the apparatus, the X-ray source and the detector may rotate around a fixed sample stage, or the sample stage may rotate while the X-ray source and the detector are fixed. Since both setups use the same reconstruction methods, we assume the former configuration in the following explanation.

X-rays generated in the X-ray source pass through the sample, where they are attenuated, and they are then detected by the detector (Fig. 1). In a CT measurement, this is repeated for a range of angles by rotating the X-ray source and the detector. The X-ray shadow of the sample, which is also called a projection image, contains partial information about the distribution f of the linear attenuation coefficient of the sample. Projection data can be rearranged and viewed as a sinogram, where one axis represents the projection angle and the others represent the positions along a horizontal row and a vertical column of the projection images. The attenuation of X-rays by a material is described by the Beer–Lambert law (Equation 1):

$$I = I_0 \exp(-\int f dl), \tag{1}$$

where I is the detected intensity of the X-ray beam, I_0 is the beam intensity without attenuation, and l is the X-ray path. The Beer–Lambert law describes the exponential attenuation of the beam in a material based on its linear attenuation coefficient. The intensity I and I_0 are functions of sinogram coordinate $s=(u, v, \theta)$, where θ is the projection angle, and u and v are the coordinates of the detector. The distribution of the linear attenuation coefficient is a function of the sample coordinate $r=(x, y, z)$. Note that the coordinates may be omitted in the following text for the sake of brevity.

Here we must note that when there is a difference between measurement conditions and the conditions

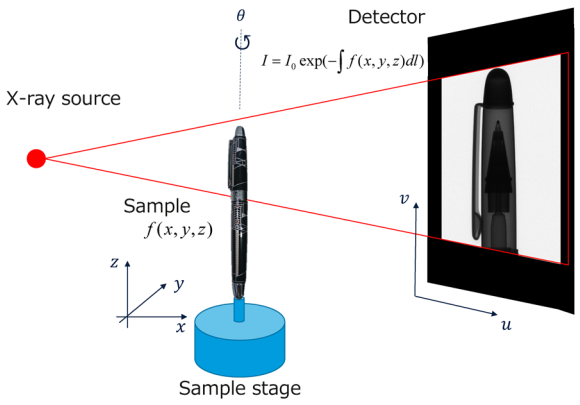

Fig. 1. Acquisition of projection images.

* XRD Application & Software Development, X-ray Instrument Division, Rigaku Corporation.

assumed during reconstruction, the reconstructed image will contain false structures—artifacts—in addition to legitimate structures of the sample. For example, if the geometry of the CT device is misaligned and different from the geometry assumed by the reconstruction algorithm, the reconstructed image may be blurred. Although these differences can be corrected as a pre-processing step before reconstruction, it is often faster and more reliable to ensure there is no misalignment before conducting a CT measurement.

3. Principle of Conventional Reconstruction

Tomographic reconstruction determines the three-dimensional distribution f of the linear attenuation coefficient from a set of two-dimensional projection images measured at a range of angles (Fig. 2). To see this, let us apply a logarithmic transform to Equation 1 as follows:

$$g := \log\left(\frac{I_0}{I}\right) = \int f\, dl. \tag{2}$$

The result g represents the absorption of X-rays, and Equation 2 linearizes the exponential character of the intensity values in raw projection images. Since the right side is a function of f, we can write it more concisely as

$$g = Af. \qquad \text{why can we make this linear?} \tag{3}$$

The operator A transforms the function f into the projection images g, and is called the Radon transform[1]. From this point of view, reconstruction is simply the operator B that transforms the projection images g to reconstructed image f (inverse Radon transform):

$$f = Bg. \tag{4}$$

In the case of the parallel beam method, the operator B has an exact analytic form, whose implementation is called the FBP (Filtered Back-Projection). A more detailed description will be left to the literature [2], but the basic functionality of FBP is as follows: projection images are convolved with (in real space) or multiplied by (in frequency space) a filter such as the Ram-Lak filter and then projected backward (the direction of projection is reversed). The Ram-Lak and similar filters emphasize high-frequency components (details) in the projection data and enable us to obtain sharp reconstructed images.

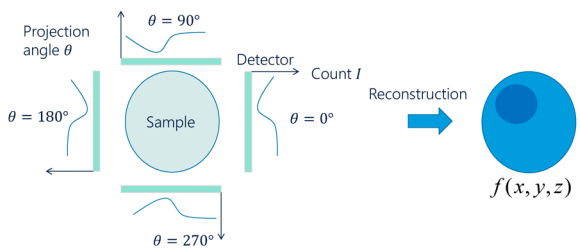

Fig. 2. CT measurement and reconstruction.

The extension of FBP to a three-dimensional cone-beam geometry is called the FDK (Feldkamp, Davis, Kress) method[3]. Although it is not an exact reconstruction algorithm, it is widely used because of its speed and ease of implementation. FBP-type methods were used for a long time because of their analytical exactness and low cost of calculation. However, they are highly susceptible to noise and non-standard or imperfect scan conditions, which can easily lead to the formation of severe artifacts in the reconstructed images.

For example, if FBP reconstruction is performed with an insufficient number of projections, radial streak artifacts will appear around high-contrast structures. This, along with the sensitivity of the method to noise, makes FBP unsuitable for fast measurements[4]. Furthermore, when the sample extends past the measurement field in the direction perpendicular to the rotation axis, low-frequency noise appears in the reconstructed image and a bright ring, also called a truncation artifact, appears around the edge portion of the reconstructed image. These artifacts reduce the accuracy of the analysis and need to be removed. Additionally, FBP-type methods cannot be used for measurements with unconventional scan geometries[5].

In some cases, depending on the sample, it may not be possible to adjust the measurement conditions to suit the conventional reconstruction method. For example, when the sample is susceptible to deformation or when it is necessary to reduce the radiation dose, measurement time must be shortened. Further, when the measurement field of view must be reduced to increase resolution, it is inevitable that the sample protrudes from the measurement field of view. In such cases, reconstruction using the IR method can be applied to better deal with the imperfect measurement conditions.

4. Principle of IR

The IR method iteratively updates a reconstruction image until the projections of the image match the measured projections[6]. The merit of this method is that it produces high-quality images even in cases where analytical reconstruction fails. Additionally, we can incorporate prior information about the sample into the reconstruction process, further increasing the quality of the final image. However, the main disadvantage of IR is that it takes a long time to calculate and requires a high-performance computer. For a long time, this has prevented IR methods from being widely adopted in practice.

The amount of time and hardware resources needed by IR methods has been greatly reduced in recent years due to developments in computer technology and the algorithms utilized in these methods. In this article, we explain the principles of the IR method using the proximal gradient method, and we introduce techniques for obtaining high-quality reconstruction images at high speed.

We reformulate the above discussion to better understand the operations performed in the IR method.

Equation 3 describes the relationship between voxels of the reconstruction image and pixels of the projection images, and can be rewritten as follows:

$$g(s)=\sum_r A(s, r)f(r). \quad (5)$$

The matrix A is called the system matrix or coefficient matrix, whose matrix element $A(s, r)$ represents the ratio of voxel at r projected onto position s in the sinogram. By solving this matrix equation, we can obtain the reconstruction image.

In IR, we do not directly solve Equation 5. Instead, we find a solution by minimizing the evaluation function J, which consists of the sum of squares of differences between the left and right sides of the equation.

$$J(f)=\frac{1}{2}\|Af-g\|^2 \quad (6)$$

In this article, the evaluation function is minimized using the proximal gradient method by the gradient descent method and regularization. First, the gradient descent method is explained.

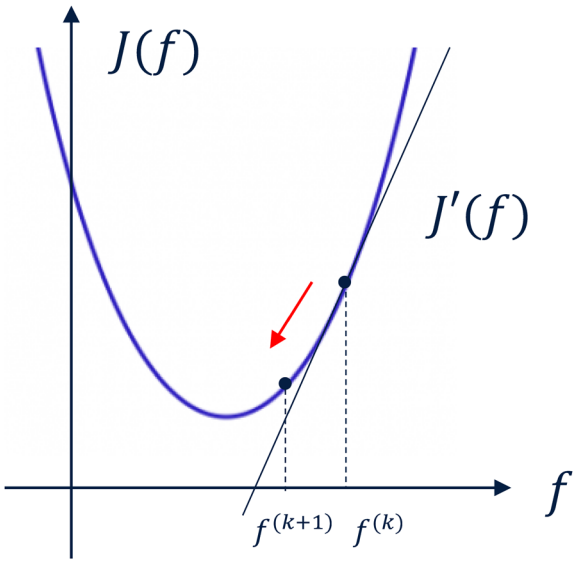

Fig. 3. One-dimensional illustration of optimization using the gradient descent method.

In the gradient descent method, we calculate the gradient (differential) of the evaluation function with respect to the reconstruction image and update the image in the direction of the gradient's steepest descent (Fig. 3).

$$f^{(k+1)}=f^{(k)}-\alpha\nabla J(f^{(k)})=f^{(k)}-\alpha A^T(Af^{(k)}-g) \quad (7)$$

The parameter α represents the degree to which the image is changed in the gradient direction. By repeating this update, we converge to a reconstructed image that minimizes the evaluation function.

Equation 7 can be decomposed into the following four steps:

1) Project the reconstructed image from the k-th iteration step using A.
2) Calculate the difference image between these projections and the measured projection images.
3) Back-project the difference image with A^T (the transpose of A).
4) Add the back-projected difference image (multiplied with α) to $f^{(k)}$ to get the updated reconstructed image $f^{(k+1)}$.

A^T acts as the back-projection operator because it transforms sinogram values into voxel values. That is, in IR, we can obtain the reconstructed image by repeating the projection and back-projection operations. The initial reconstructed image for IR methods is often simply a homogeneous image of zeroes, but it can also be, for example, the output of a conventional reconstruction method.

In the IR method, the quality of the reconstructed images can be improved by including prior information about the images in the evaluation function. A common example of prior information is the sparseness of the image. Sparseness means that there are many pixels with a value of 0, which means there is little essential information contained within them. This occurs when the sample occupies only a small part of the measurement field of view, while the rest is occupied by air. L1 regularization is one method to enforce sparseness at the time of reconstruction by adding the L1 norm of the pixel values to the evaluation function. When including regularization, the IR method can be expressed as:

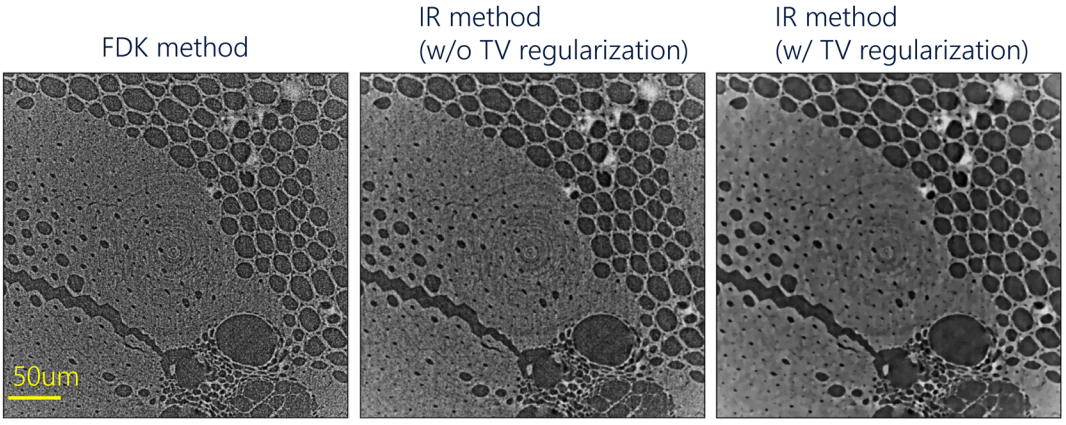

Fig. 4. Comparison of images reconstructed using different reconstruction methods.

$$f^{(k+1)} = \mathrm{prox}_{\alpha R}\left(f^{(k)} - \alpha\nabla J(f^{(k)})\right)$$
$$= \mathrm{argmin}_f\, R(f) + \frac{1}{2\alpha}\left\|f - f^{(k)} + \alpha\nabla J(f^{(k)})\right\|^2, \quad (8)$$

where the $R(f)$ is the regularization term and *prox* is the proximal operator. In the case of L1 norm mentioned above, this proximal operator is equivalent to a soft-thresholding operator[7].

It is not always necessary to use the sparseness of the original image itself, and the sparseness of a conversion of this image may be used. For example, when a sample contains many homogeneous regions, its differential image will have many zero parts. Therefore, it is possible to perform a transform that differentiates the reconstruction image and reconstructs it by adding its L1 norm to the evaluation function. This is called the TV (Total Variation) regularization[8], and it allows us to obtain a reconstruction image with reduced noise while preserving sharp edges. However, note that increasing the degree of regularization (increasing the parameter indicating the strength of the regularization term) can lead to loss of details of the sample.

Let us compare reconstructed images obtained by the FDK method and the IR method. Figure 4 shows a bamboo skewer sample measured with a Rigaku nano3DX, reconstructed using FDK (left), the IR method (center), and the TV regularized IR method (right). We can see that the reconstruction image of the FDK method is the noisiest of the three, while the noise is reduced by using the IR method. The IR method with TV regularization has the least noise, and the voids and tissues can be clearly distinguished.

In IR methods, as the number of iterations is increased, the reconstruction image approaches convergence and becomes clearer, while the calculation time increases proportionally. Convergence may require a large number of iterations in some cases, so methods have been devised to accelerate this process. Here, we introduce the OS (Ordered Subset) method[9] and the Nesterov acceleration method[7, 10].

In the OS method, projection images are divided into several subsets, and projection or back-projection is performed for each subset separately and in order. By doing so, the number of image updates per iteration increases and the image converges faster. However, increasing the number of subsets too much will lead to increased computational overhead and the possibility of artifacts occurring due to each image update using only a small number of projections.

Next, in the Nesterov acceleration method, image updates use not only the image from the previous step, but also the image from the second-to-last step. This is a method that can be used not only in IR using the gradient descent method but also in conjunction with the proximal operator. Using these acceleration methods on the examples shown in this article, we can obtain reconstructed images in 10 iterations with the initial image as a uniform image.

5. Applications of IR Method

This section showcases several applications of the IR method. We will pay special attention to how the image quality is improved by the IR method when the measurement is performed under conditions that cause artifacts in conventional reconstruction. Below, we use three types of samples for comparison: a bamboo skewer, bread, and an electronic component.

First, we compare the degree of image quality deterioration when the number of projection images is reduced for the bamboo skewer sample. For comparison, we use the same measurement data and restrict the number of projections used at the time of reconstruction.

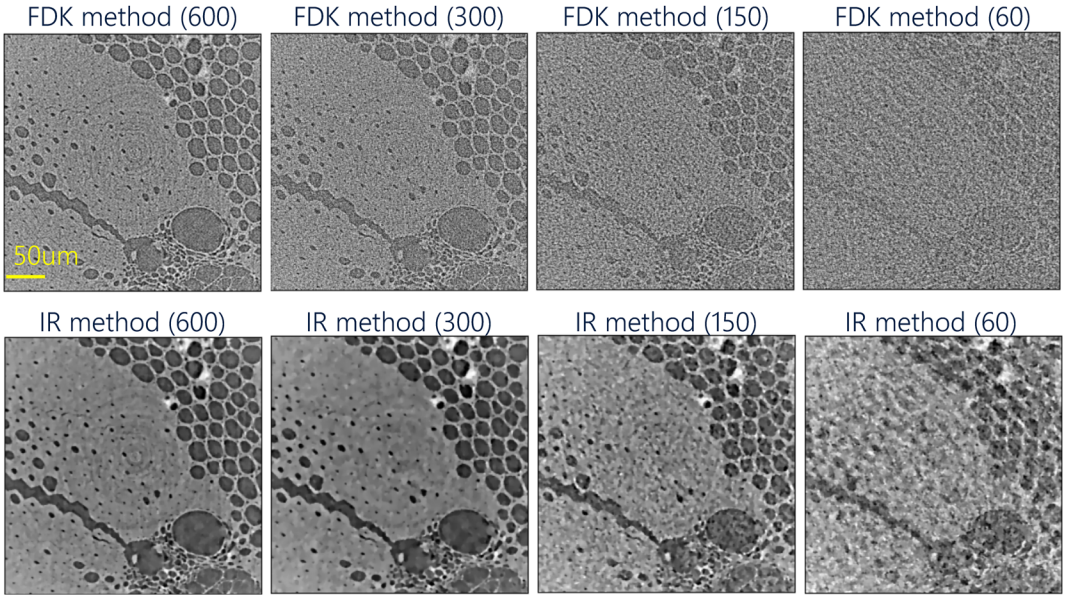

Fig. 5. Comparison of images reconstructed using different numbers of projections. The numbers in the brackets represent the number of projection images used during reconstruction.

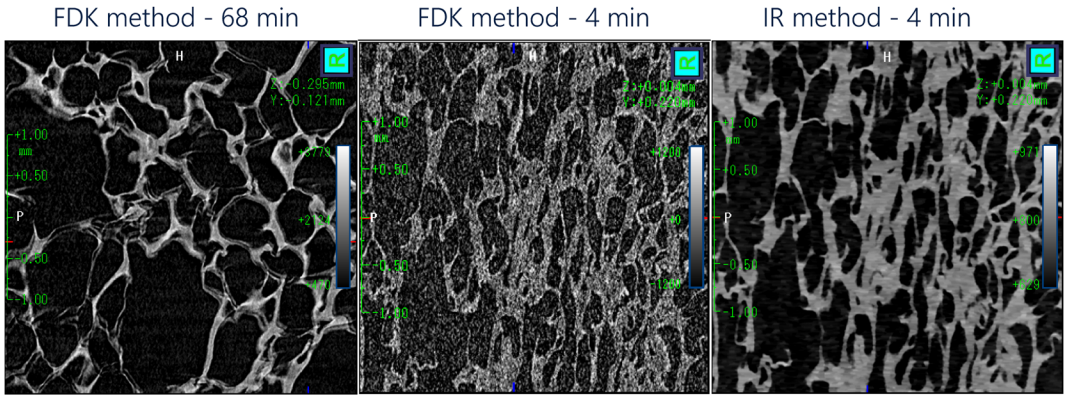

Fig. 6. Comparison of images reconstructed from data with different measurement times.

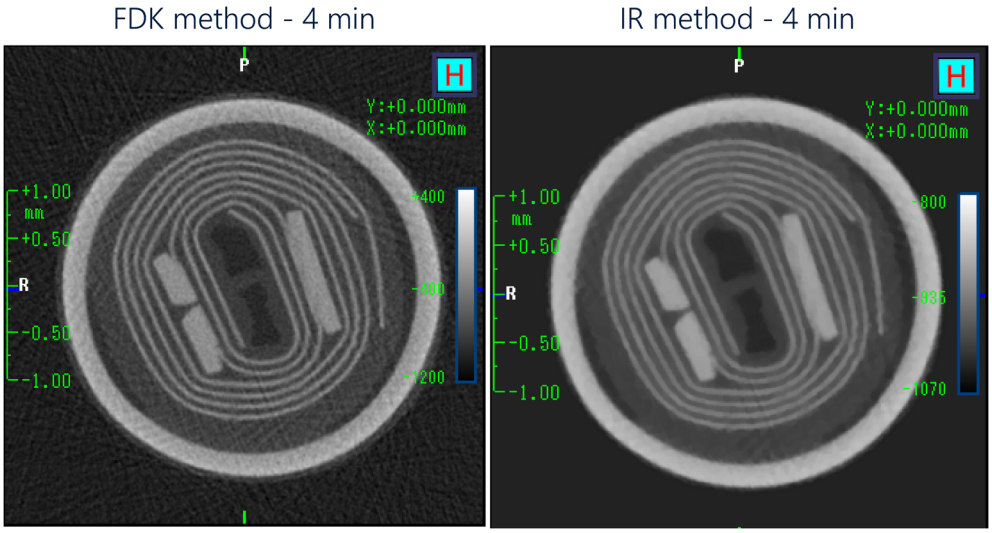

Fig. 7. Comparison of the reconstructed images of a capacitor sample with high-contrast structures.

The upper row of Fig. 5 shows images reconstructed by the FDK method, while the lower row contains the results of the IR method. We can see that noise increases as the number of projections used for reconstruction decreases. Comparing reconstructed images for the same numbers of projection images, we can see that the IR method produces images with less noise than the FDK method.

Next, we compare image quality when the measurement time is shortened for a fast measurement of the bread sample. The measurement was performed using a Rigaku CT Lab HX. The measurement time needs to be kept short because the sample tends to deform over time. This is shown in the left part of Fig. 6, where an image reconstructed from a long (68-minute) measurement contains blurred edges and artifacts due to factors such as evaporation of water in the sample. Shortening the measurement time to reduce this effect in turn increases the noise contained in the measured data. In such a case, reconstruction using the IR method is effective. The center and right parts of Fig. 6 show images of a fast measurement (4 minutes) reconstructed by the FDK method and the IR method, respectively. It can be seen that the IR method produced images with much less noise than the FDK method. This experiment demonstrates that the IR method can be used to drastically shorten measurement times to avoid sample motion without sacrificing image quality. Shorter measurement times can also be beneficial for a number of other reasons, such as decreasing the radiation dose absorbed by the sample. For instance, in the medical field, short-time or low-power measurements are carried out in order to reduce the radiation dose absorbed by the human or animal patient.

Lastly, we compare the image quality in reconstructions of the capacitor sample, which contains high-contrast components. The measurement was performed with Rigaku CT Lab HX again. Figure 7 shows images reconstructed by the FDK method (left), with streak-like artifacts near high-contrast parts of the sample, and the IR method (right), which significantly reduces this type of streak artifacts.

6. Conclusion

In this article, we have familiarized ourselves with the principles and applications of the iterative reconstruction (IR) method. Through multiple applied examples, we have seen that this method has drastically increased image quality compared to the conventional

reconstruction method in various scenarios. For a sample whose shape tends to change over time, long-duration measurements are not feasible, and fast measurements become necessary. Fast or low-power measurements may also be used to reduce radiation dose. However, noise increases drastically in such measurements, complicating any diagnosis performed using the reconstructed images. In such cases, the IR method can prove to be very effective by allowing us to decrease the measurement time or radiation dose without a significant loss in image quality.

This time, we have explored only some of the more basic applications of IR, and we have seen how this method can provide high-quality images suitable for further processing, increasing the accuracy of various analyses. Additionally, various advanced correction methods, such as motion corrections and beam hardening corrections, can be performed within the framework of the IR method. We will focus on such advanced uses of the IR method in a future article.

References

(1) A. G. Ramm and A. I. Katsevich: *The Radon Transform and Local Tomography*, CRC Press, Boca Raton–New York–London–Tokyo, (1996).
(2) A. C. Kak and M. Slaney: *Principles of Computerized Tomographic Imaging*, IEEE, (1988).
(3) L. Feldkamp, L. C. Davis, and J. W. Kress: *J. Opt. Soc. Am. A-opt. Image Sci. Vis.* **1**(1984), (6), 612–619.
(4) S. UK, F. Morin, V. Bousson, R. Nizard, G. Bernard, and C. Chappard: *J. Surg. Res.*, **5**(2022), 115–133.
(5) B. De Samber, J. Renders, T. Elberfeld, Y. Maris, J. Sanctorum, N. Six, Z. Liang, J. De Beenhouwer, and J. Sijbers: *Opt. Exp.*, **29**(2021), 3438–3457.
(6) R. Gordon, R. Bender, and G. Herman: *J. Theor. Biol.*, **29**(1970), 471–481.
(7) A. Beck and M. Teboulle: *SIAM J. Imag. Sci.*, **2**(1), (2009), 183–202.
(8) L. Rudin, S. Osher, and E. Fatemi: *Phys. D*, **60**(1992), 259–268.
(9) H. M. Hudson and R. S. Larkin: *IEEE Trans. Med. Imag.*, **13**(1994), (4), 601–609.
(10) Y. Nesterov: *Soviet Mathematics Doklady*, **27**(1983), 372–367.