LOW-DOSE COMPUTED TOMOGRAPHY IMAGE DENOISING BASED ON JOINT WAVELET AND SPARSE REPRESENTATION

by

Samira Ghadrdan

MASc., K. N. Toosi University of Technology, Tehran - Iran, 2010

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2014

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

LOW-DOSE COMPUTED TOMOGRAPHY IMAGE DENOISING BASED ON JOINT WAVELET AND SPARSE REPRESENTATION

Master of Applied Science

2014

Samira Ghadrdan

Electrical and Computer Engineering

Ryerson University

Abstract

One of the most challenging issues in low dose computed tomography (CT) imaging is image denoising and signal enhancement. Sparse representational methods have shown initial promise for these applications. In this thesis we present a wavelet based sparse representation denoising technique utilizing dictionary learning and clustering. By using wavelets we extract the most suitable features in the images to obtain accurate dictionary atoms for the denoising algorithm. To achieve improved results we also lower the number of clusters which reduces computational complexity. In addition, a single image noise level estimation is developed to update the cluster centers in higher PSNRs. A new image enhancement technique is developed for low-dose CT images to improve the quality of image for diagnostic purpose and reduce the blurring artifacts. The accuracy along with the computational efficiency of the proposed algorithm are then compared with recent approaches and clearly demonstrate the improvement of the proposed algorithm proposed in this thesis.

Acknowledgements

Table of Contents

List of Figures

List of Tables

List of Acronyms

AEC: Automatic Exposure Control

BM3D: Block-Matching and 3D filtering

CSR: Clustering based Sparse Representation

CT: Computed Tomography

CTDI: Computed Tomography Dose Index

DCT: Discrete Cosine Transform

DLP: Dose Length Product

kNN: K-Nearest Neighbour

LASSO: Least Absolute Shrinkage and Selection Operator

LSSC: Learned Simultaneous Sparse Coding)

mSv: milli-Sieverts

NLM: Non Local Means

PCA: Principle Component Analysis

PDE: Partial Differential Equation

PSNR: Peak Signal to Noise Ratio

SCoTLASS: Simplified Component Technique-LASSO

SD: Standard Deviation

SSIM: Structural Similarity Index

SVD: Singular Value Decomposition

ROI: Region Of Interest

Chapter 1. Introduction

Computed Tomography (CT) scan is one of the most used image modalities for diagnosis due to the preservation of important details and structures inside the body. CT scanners reconstruct a cross-sectional image by measuring x-ray attenuation properties of the body from multiple directions. The radiation dose involved with a typical CT scan is 1-14 milli-Sieverts (mSv) depending on the exam, which is not negligible since it is almost equal to the annual dose, 1–10 mSv, received from different natural sources of radiation, such as radon and cosmic radiation [3]. With the increasing number of people undergoing CT scans, the public health may be of significant risk due to radiation exposure by CT scan.

One of the major recent goals in CT research and image processing is reducing the harmful patient radiation dose and then surpassing the noise on the low- dose CT images by post processing. Due to a complex relation between image noise and scan parameters and spatial position [4], finding a distribution of noise in the final CT image is difficult. The noise is usually unknown and the noise

variance is a variable parameter. Different algorithms have been proposed to reduce the CT noise. The first category removes the noise in the projection data before image reconstruction while in the second category, algorithms reduce noise during the CT reconstruction phase. These algorithms perform denoising through optimization of objective functions [5], [6]. The most common methodology is noise reduction algorithms of the reconstructed CT images. A critical aspect is to preserve edges and small important structures for diagnosis while denoising. The conventional edge-preserving methods in frequency domain are wavelet-based methods and in spatial domain are partial differential equation (PDE) based methods [7], [8].

Recently Sparse Representation has been used as a dominant tool for image noise removal and preservation of important information and edges. This method is a non-local model that reconstructs the signal based on a set of basic vectors called dictionary atoms. A suitable dictionary can be selected by utilizing either analytical or adaptive dictionary techniques. Adaptive dictionaries are constructed based on training of different patches of the noisy image. In contrast, analytical dictionaries are fixed with regards to the nature of the image using stationary basis functions like cosine functions or wavelet functions. To find an optimal solution an iterative process is used to modify a chosen small subset of the training set. The well-known adaptive dictionary called K-SVD method [9] proposed by Elad and Aharon is the state of the art in this field. On the other hand, Non-Local models such as Non Local Means (NLM) [10] making use of the repetitive structures in an image and by exploring the similarity between patches have led to a successful denoising algorithm, Block-Matching and 3D filtering(BM3D) [11] has shown to be superior among all non-local models. Combining these two complementary models, a clustering based sparse representation (CSR) algorithm has been proposed [1]. CSR algorithm unifies both models and formulates a double header l_1-optimization problem. Key advantage of this proposed method includes both sparsity and clustering (location related constraint) thus generating a sparser solution and better denoising results. More details about these methods are provided in chapter 2.

In this thesis, a new edge preserving approach is proposed to combine conventional methods and sparse representation to denoise low dose CT images more efficiently. Using wavelets, we extract the features that are most suitable for denoising and edges preservation. The ultimate goal is to preserve as much high frequencies existing in the image as possible which is used to learn the dictionary later. Wavelet denoising is usually done by thresholding the wavelet coefficients. By introducing a balance between the measure of the energy loss after the denoising process and the percentage of the relative sparsity (the number of resulting zero coefficients in the denoised image), wavelet denoising preserves the detailed information and at the same time reduces unwanted noise. Thus, the proposed method chooses coefficients that serve in the best interest for edge preservation while keeping the frequencies that the algorithm is working based on them and the algorithm also reduces unwanted noise that makes the dictionary learning step difficult and less accurate. Adaptive strategy is used to learn the dictionary in this approach. The dictionary atoms are learned from k-means clustering and Principal Component Analysis (PCA) of each cluster. In this process we can find the accurate clusters and construct the dictionary with fewer atoms and as a result the computational complexity is reduced.

A single image noise level estimation is developed in the algorithm to update the cluster centers more accurately. Utilizing a patch based noise level estimation; the weak patches from a single noisy image are selected. The selection is based on the gradients of the patches and their statistics and then PCA is used to do the noise estimation from the selected patches. It is a fast and accurate method to estimate the variance of the noise. This method performs effectively for removing additive Gaussian noise from images, and has also been adapted to the non-Gaussian noise in CT images.

Preforming the denoising algorithm on the low-dose medical images result in some blurring and loss of details since the texture is so similar beyond the edges. Hence, the proposed method is followed by an image enhancement method based on fusion to highlight the detailed textures and edges. The fusion based enhancement method combines the second level (low-low) wavelet coefficients of the noisy and

the denoised image. The selected coefficients in noisy image contain the important information about the texture not influenced by the noise. Therefore the resulting coefficients can take advantage of the highlighted detailed information and edges and an inverse wavelet transform conclude the algorithm with a high performance low-dose image denoising.

Summarizing the contributions in this thesis:

- A joint wavelet sparse representation denoising algorithm on the low-dose CT images is presented.

- A preprocessing step is proposed to learn a better dictionary with fewer number of clusters which brings higher computational efficiency.

- A new noise level estimation introduced to the algorithm to have an accurate cluster center updates at each iteration.

- A Fusion-based Image enhancement is proposed to recover the blurred edges and important structures.

This thesis is organized as follows. In Chapter 2, an introduction to computed tomography imaging, the dose measurement and reduction is reviewed. A literature review on different low-dose CT image denoising is presented in this chapter as well. In Chapter 3, the sparse representation problem is described and some of its existing solutions are introduced, in detail. Also, dictionary learning and sparse principal component is studied. Chapter 4, provides analysis on clustering and clustering-based sparse representation, our algorithmic methodology and contribution in this thesis in details. Chapter 5, contains the main results of the proposed methods on both synthetic and medical CT images along with a simulation-based study of its performance and evaluation. Chapter 6 provides concluding remarks and discussion of the new proposed methods and the possible future works.

Chapter 2. Computed Tomography Imaging

2.1. Computed Tomography

One of the most powerful techniques for creating 2-D and 3-D cross-sectional images is Computed Tomography (CT). CT images are obtained from the back projected data using flat x-ray images. Due to the characteristics of the different internal structures of the body, the acquired back projection data contains multiple information such as dimensions, shape, internal defects, and density in the resulting CT images. Figure 2.1 shows a schematic of a generic CT system with a single row of detectors. Typical CT machine have multiple rows of detectors operating side by side.
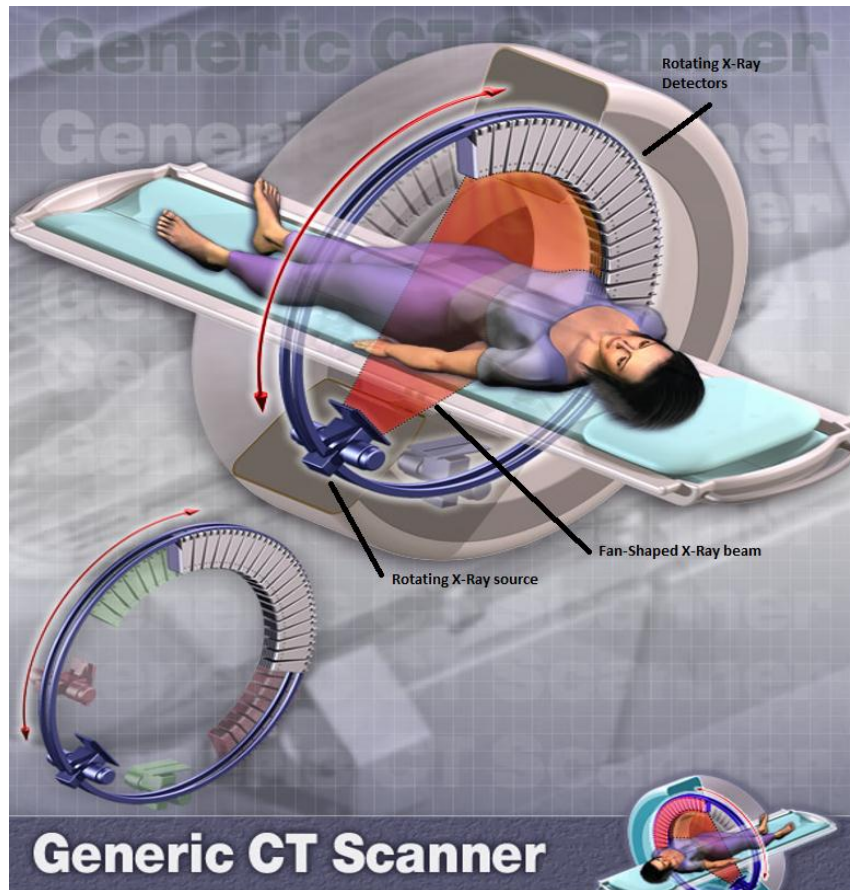
Figure 2.1. Generic CT scanner [12]

The patient is placed on a motorized stage that is between a radiation source and an imaging system. Imaging system contains rotating x-ray detectors which are connected to a computer so that x-ray images collected can be correlated to the position of the patient. As the x-ray source which creates a fan-shaped beam and detectors are rotating, the table is continuously moving to produce a spiral or helical scan. Multiple rows of detectors in the imaging system help to have as many slices as 320 to reduce the overall time of scanning and create a 2-D image like a film radiograph. Special computer software is designed to produce cross-sectional images of the patient's body as if it was being sliced. Figure 2.2 depicts the side view of scanning principals as the patient transports in the scanning system.
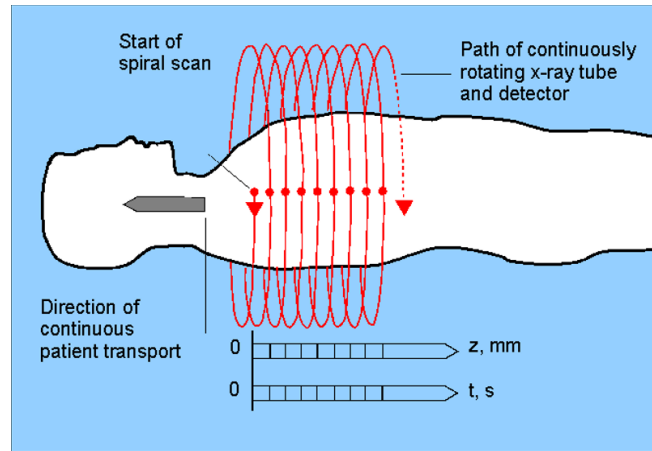
Figure 1.2. Scanning Principle [12]

High Quality CT scanners permit images less than 1 mm thick and precise reformatted images. Due to these precious capabilities, utility of CT has been increased greatly and it is replacing other radiographic examinations more often. The growth rate of utilizing CT scans in United States confirms the above statement. Approximately 13 million CT scans were performed in the United States in 1990 [13]. In 2000, the number of CT scans more than tripled to approximately 46 million [13] . The estimated number of CT scans for 2006 is 62 million [13]. Using the high quality CT imaging, patients can benefit from a quicker and more accurate diagnosis and precise internal body structures information for planning therapeutic procedures. However, in spite of the extensive contributions of CT to healthcare, some attention should be considered to the health risk associated with the ionizing radiation received during a CT exam [14].

Considering the growth of the population undergoing CT scans and their exposure to the radiation, the public health risk may be significant. One researcher claimed that a population as big as 0.4% of the United States patients diagnosed by cancer, may be attributable to the radiation from CT studies based on CT usage data from 1991–1996 [13]. It was determined that 1.5-2% of cancers may eventually be resulted from the ionization radiation in CT scanning, when organ specific cancer risk was adjusted for

current levels of CT usage [13]. Based on the mentioned study and similar ones the CT community was forced to review the prescribed amount of radiation for CT scans, especially for pediatric patients. Eventually this progress resulted in an effort to minimize CT doses and optimize image quality.

2.2. Fundamental concepts in radiation dose

The primary step is to distinguish between the "radiation exposure" and "radiation dose (absorbed radiation dose)". The former relates to the quantity of ionization events in air produced by x-ray photons, where the latter describes the amount of radiation energy deposited in the patient's body as a result of exposure. Radiation exposure is usually a measured quantity, whereas the absorbed dose is typically calculated based on the exposure and estimates of energy absorption per body mass unit (eg, kilograms of body weight).

2.2.1. Scanner output

One of the first metrics used to describe the radiation dose from the CT scanner is Computed Tomography Dose Index (CTDI). When CTDI was defined in the early days of CT, it was measured with thermoluminescent dosimeters. This measurement is rarely performed since it is time consuming and labor intensive and also requires many exposures for each beam width, phantom size, tube potential setting (kV), and position in the field of view. The resultant parameter was referred to as the absorbed dose, and the SI unit of measurement is the gray (Gy). CTDI can be measured by calculating the integral under the radiation dose profile in the z-axis in Figure 2.2 of a single scan. The maximum of the radiation dose profile is called the "peak dose." Figure 2.3 shows a sample of radiation dose profile in a single CT scan. Radiation dose should drop sharply at the edges of a scan in an ideal case. However in reality, the radiation dose profile decreases gradually at the edges of a scan with tails at both side. The reason lie behind the x-ray beam divergence and internal radiation scatter by the body tissues.
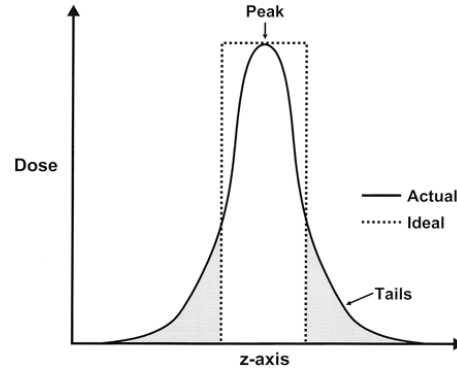
Figure 2.3. Distribution of radiation dose in a single CT scan [15]

There are couple of variants of CTDI based on specific description on the steps of the measurement and calculation process. These include the $CTDI_{100}$, the weighted CTDI ($CTDI_w$) and Volume CTDI ($CTDI_{vol}$). The most relevant one is $CTDI_{vol}$ which based on a directly and easily measured quantity provides a single parameter to describes the radiation delivered to the scan volume for a standardized (CTDI) phantom [16]. The other concept that is useful in representation of the energy delivered by a given scan protocol is Dose-Length Product (DLP). It is an indicator of the integrated radiation dose of an entire CT examination. DLP is defined as following:

$$DLP = \text{CTDI}_{\text{vol}} \times Scan\ length \qquad (2.1)$$

2.2.2. Effective Dose

The Effective dose, E, stands for the non-uniform radiation absorption of partial body exposure relative to the whole body radiation dose. It is not a measurement of dose, but a concept that allows the comparison of the risk from an exposure to radiation between different CT examination protocols. The SI unit of measure is expressed in milli-Sieverts (mSv). The generic calculation method for effective dose is based on DLP and a set of coefficients k (unit $mSv.mGy^{-1}.cm^{-1}$), where the values of k are dependent only on the region of the body being scanned. Therefore E can be calculated as:

9

$$E = k \times DLP \tag{2.2}$$

Comparing the patient effective dose (from 1-10 mSv in U.S.) and background radiation dose from natural sources (equivalent to averages 3 mSv per year in U.S.), patients are better to consider the risk associated with medical doses.

2.2.3. General principles of ALARA

At the purpose of radiation protection, a guideline has been defined as following:

1. The exam must be medically indicated to prevent unnecessary scans (exposures).

2. The exam dosage must be As Low As Reasonably Achievable (ALARA) based on the diagnostic task.

3. Although the limited dosage level is recommended by consensus organizations, for medically-necessary exams limits are not applied [17].

The dose management guideline is designed to use the right dose for a CT examination considering the specific patient attenuation and the diagnostic task. Moreover undertaking a CT exam must be appropriate for the individual patient. Both clinicians and radiologists share the responsibility. As a result the first thing to do is to choose the most suitable imaging modality for the required medical diagnostic task, and to optimize the technical aspects of the test, in such way that we are applying the dosage as low as possible and still obtaining the desired level of image quality.

2.3. Dose reduction strategies

There are different factors used to change the radiation dose amount: the tube current (amperage), slice scan time, and tube peak kilovoltage. Since the tube current and slice scan time have the same effect on radiation dose, they usually are taken together as mAs in radiography. Thus decreasing the

mAs (by reducing tube current or slice scan time) decreases the dose proportionally: 150 mAs deliver half the dose of 300 mAs. Decreasing peak kilovoltage also decreases radiation dose, because less energy is carried by the beam. Another factor is slice thickness, slice spacing, and helical pitch that affect dose as well.

2.3.1. Fixed tube current (technique charts)

Due of the normalized nature of CT data, CT images are not like traditional radiographic imaging in the sense of being too dark or too light (over-exposed) and the image always appears properly exposed. As a result, there is no pressure on CT users decrease the tube-current-time product (mAs) for small patients in the technical aspect; however, it is the CT operator responsibility to take patient size into account when selecting the parameters that affect radiation dose, such as mAs. To obtain the proper mAs, the operator is provided with a chart on the mAs as a function of size. By changing the tube potential and exposure time we can reach the appropriate exposure to the patient. Usually the tube potential (kV) and rotation time is fixed for a specific application. To minimize motion blurring, rotation time should decrease and the lowest kV consistent with the patient size should be selected to maximize image contrast. Hence the first parameter to adapt to patient size is the tube current.

An extensive research has been done to find a precise relationship between the mA and patient size. The results all confirm that the reduction should not be based on patient weight rather; it should be based on the overall attenuation, or thickness, of the anatomy of interest. The mAs reduction from an adult to a newborn when we are doing a head CT imaging should be around a factor of 2 to 2.5 and in case of the CT imaging of the body, a reduction in mAs of a factor of 4 to 5 is typical from adult to infants, while for obese patients, an increase of a factor of 2 is appropriate [14].

2.3.2. Tube current (mA) modulation

Patient radiation absorption is a variable factor based on the both projection angle and anatomic region which is not considered when using a fixed tube current. Collecting data from different body parts is based on their attenuation. Therefore data can be acquired with substantially less radiation from the parts with less attenuation without negatively affecting the final image noise. Another fact which is important to take into account is dose reduction for projections of the part with limited interest. The modulation of tube current can be done angularly about the patient, along the long axis of the patient or incorporate both in order to adapt to attenuation differences within the patient.

2.3.2.1. Angular (x,y) mA modulation

To address the x-ray attenuation around the patient Angular (x,y) mA modulation is utilized by changing the mA value as the x-ray source rotates around the patient. The CT operator selects an initial mA value and then the mA is modulated upward or downward from the initial value with a period of one gantry rotation. Angular mA modulation optimizes mA selection for each angle to provide the least radiation dose for the required level of image quality. As the x-ray tube rotates, the mA can be varied according to the attenuation information from the CT radiograph, or in near real-time according to the measured attenuation from the 180° previous projection.

2.3.2.2. Longitudinal (z) mA modulation

Changing the mA value along the z axis is Longitudinal (z) mA modulation which addresses the varying attenuation of the patient among anatomic regions. Figure 2.4 shows a sample of longitudinal modulation. In this example the prescribed tube current curve is determined by using attenuation data from the CT projection radiograph and a manufacturer-specific algorithm. Therefore the only parameter that the operator defines for the system is the quality of the resulted image and the determination of the tube current is done by the attenuation information of the CT radiograph.

2.3.2.3. Angular and Longitudinal (x,y,z) mA modulation

Combining the previous two method, Angular and longitudinal (x,y,z) mA modulation vary the mA both during rotation and along the z axis of the patient. The task of the CT operator is still selecting the desired level of image quality. This method advantages adjusting the x-ray dose based on all 3-dimention attenuation and therefore is the most comprehensive approach to CT dose reduction.
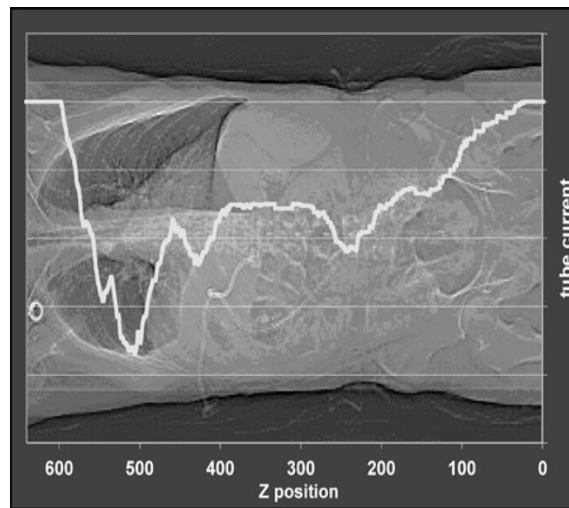
Figure 2.4. Relative tube current superimposed on a CT projection radiograph. [14]

2.3.3. Automatic exposure control (AEC)

One of the advanced techniques in dose reduction is Automatic Exposure Control (AEC), which aims to automatically modulate the tube current. AEC algorithms have demonstrated dose reductions of about 20–40% when image quality is appropriately specified. When using an AEC, the modulation can occur in near-real time by using a feedback mechanism and thus may be fully preprogrammed, or incorporate pre-programming and a feedback loop. AEC not only modulates the tube current but also determines and delivers the right dose for any patient in order to have a proper image for diagnosis.

Major manufacturers of CT imaging systems now have an Automatic exposure control systems on heir scanners. The principal of AEC is the same for all but it is implemented differently on each of them in

terms of defining the output image quality. Table 2.1 gives a summary of the different implementations from four major CT vendors:

Table 2.1. Summary of the four most common automatic exposure control strategies [18]

Manufacturer	AEC trade name	Image quality reference	Goal
General Electric	Auto mA, Smart mA	Noise index	Maintain a constant noise level (defined in noise index), using tube currents within prescribed minimum and maximum values
Toshiba	SureExposure	Standard deviation (high quality, standard, low noise)	Maintain a constant noise level (defined in standard deviation valuesfor each protocol), using tube currents within preset minimum and maximum values
Siemens	CARE Dose4D	Quality reference mAs	Maintain the same image quality (varying noise target for different attenuation level) with reference to a target effective mAs level for a standard-sized patient
Philips	DoseRight	Reference image	Keep the same image quality as in the reference image, regardless of attenuation level

2.4. CT Noise

By using the graphic cursor to display the pixels of the CT image of a uniform phantom such as phantom containing all water, it is shown that the pixel CT numbers are vibrating around 0, rather than being uniform (zero is what is assumed for water). It means that by reading the pixel CT numbers, values of 0, +1, +2, −1, and etc. can be seen. The reason lies behind the fact that we use a limited number of photons to form the image. In radiography, image noise is related to the numbers of x-ray photons contributing to each small area of the image. In CT, image noise is resulted from the number of x-rays contributing to each detector measurement. Number of x-ray detected and therefore CT noise are bases on different factors as follows [14]:

- *X-ray tube amperage:* By changing the mA value, the beam intensity changes proportionally and as a result the number of detected x-rays. So to the purpose of dose reduction if we halve the mA value, the beam intensity halves as well and therefore the number of x-rays detected by each measurement. Thus we have a noisier image but lower radiation dosage.

- *Scan (rotation) time:* Changing the scan time means reducing or increasing the duration of each measurement and therefore respectively decreasing or increasing the number of detected x-rays in the proportional manner. Since both the tube amperage and scan time act in the same way in contributing to the noise, usually they are taken into account together as mA × s, or mAs.

- *Slice thickness:* Another factor is based on changing the thickness of slices thus the beam width entering each detector and as a result the number of detected x-rays approximately in the proportional manner. For instance if we change the slice thickness from 10mm to 5mm, we have approximately halved the x-rays entering each detector.

- *Peak kilovoltage:* For the dose reduction decreasing the peak voltage is another important factor. By decreasing the peak kilovoltage, we decrease the number of detected x-rays. Therefore, decreasing the kilovoltage increases image noise and slightly increases the subject contrast.

Another factor affecting the noise is using a reconstruction filter without changing the number of detected x-rays.two types of filter is usually used: Smooth filters which blur the noise, reducing its visual impact in contrast to sharp filters which enhance the noise. The preferences in the images of soft tissue is to blur the interfering noise, so the smooth filters are appropriate and in images with structures such as edges and bone, blur is interfering comparing to noise so the sharp filters is used. Figure 2.5 shows examples of CT images with noise contributed from reducing amperage with two different filters.
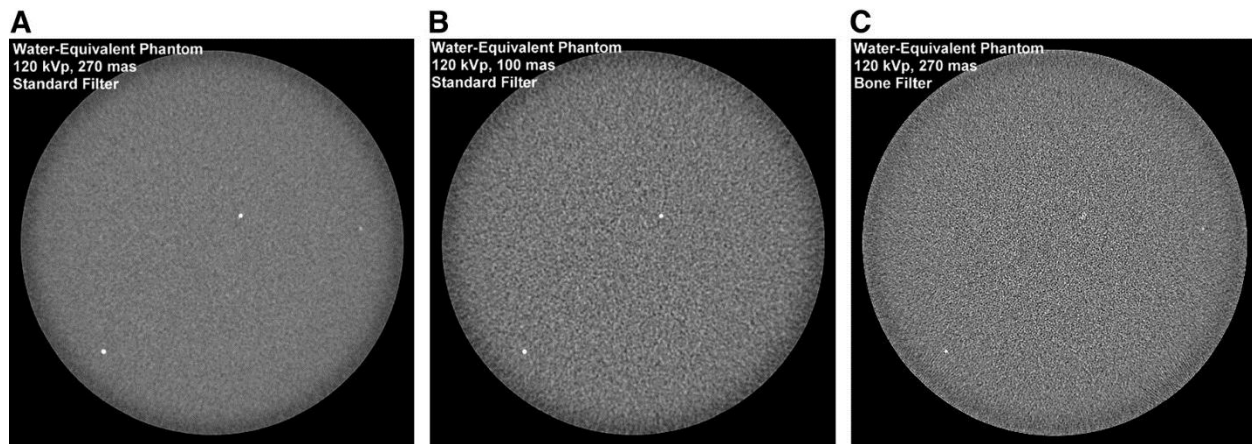
Figure 2.5. (A and B) Comparison of noise from scans using 270 mAs (typical clinical value) and 100 mAs. (C) Appearance of image noise is strongly affected by reconstruction filter; sharp filter such as bone also sharpens (enhances) appearance of noise. [15]

The measurement of CT noise can be done by measuring the fluctuation which is made by using the region of interest (ROI) selection of a uniform phantom. There is an ROI function available on most of the CT scanners which helps user to select the region that they want to find the noise on and then average and standard deviation (SD) of the CT numbers of the pixels in the selected region will be calculated. The SD indicates the magnitude of random fluctuations in the CT number and thus is related to noise: The larger the SD, the higher the image noise [15].

2.5. Introduction to CT noise reduction methods

A significant area of research has been dedicated to the noise reduction of the reconstructed CT images to reach a better quality images for clinical diagnosis with lower radiation dose. As explained earlier the most common category of medical image denoising is based on reducing noise on the reconstructed image while preserving edges and small important details.

The most common standard image denoising method in spatial domain is Partial Differential Equation (PDE) method. Perona and Malik first proposed the PDE based denoising method in the form of diffusion. [7] They developed a diffusion factor as a function of brightness gradient which was

16

controlling the diffusion; meaning that it weakens or stops at the edges. The main disadvantage of this method is the unacceptable amount of noise remained on the edges because of the diffusion reduction on the edges and also high computational complexity of the algorithm. To prevent smoothing across edges, a new anisotropic diffusion proposed in [19] to solve the first problem by substituting a diffusion tensor instead of diffusion coefficient.

In frequency domain wavelet-based denoising methods are mostly used. In these types of methods the Gaussian noise is reduced in a wavelet domain using an orthogonal basis. The first step is to obtain the wavelet coefficient and then wavelet denoising is usually done by thresholding the wavelet coefficients. In these methods certain frequencies corresponding to the noise which are less than a pre-set threshold value will be shrunk to zero. Using the fact that noise coefficients usually have smaller magnitude comparing to the structures coefficients, we can surpass the noise by shrinking the coefficient smaller than a well-chosen threshold value. But choosing an optimal threshold value is still a major concern. Different methods have been proposed bringing this concern to attention, which has led to adaptive thresholding methods, for example in [20] the Bayesian framework is used to derive the threshold. The proposed threshold is adaptive to each subband because it depends on data-driven estimates of the parameters.

One of the other useful denoising methods is Total-Variation regularization method. The TV method was proposed by Rudin, Osher and Fatemi [21] to surpass noise and any blurring in an image, while preserving edges. This technique is commonly used in image processing and aims to minimize an energy function of a TV form. The energy function deals with the image gradient, so to effectively remove noise and streaking artifacts, high spatial gradient parts in the reconstructed CT images should be removed by minimizing the energy function. TV approach tends to penalize the image gradient uniformly without taking into account the image structure which results in an over-smoothed reconstructed image with smeared edges of fine structures. Based on the importance of edge information in medical images, the

17

performance of TV method is not acceptable in some cases and a lot of approach has been made to solve this problem. For instance, a compressive sensing edge enhancive reconstruction algorithm has been proposed in [22] for an MRI image reconstruction problem.

Recently, sparse representation based algorithms have shown promising results in the field of image processing specially image denoising. There are couple of algorithms proposed to use sparse representation and fixed dictionaries applying on CT images. Since these dictionaries contain frequency information they do not add more value comparing to transform domain denoising methods. Sparse representation and K-SVD dictionary learning is the well-known method in this field and is used to remove the additive white-Gaussian noise from the image. As explained before the dictionary is trained in contrast to fixed analytical dictionaries, therefore it enables a better description of the image in the dictionary contents. The algorithm output is a weighted average of the reconstructed image patches. Some efforts have been taken to apply the mentioned method on CT images in [9] [23] [24]. The main disadvantage of these methods is over-smoothed edges in the reconstructed image which highlights the problem of diagnosing in medical images.

Previously described denoising methods are explained in more details in the following chapter. In these methods noise is assumed to be an additive noise.

2.5.1. Wavelets in Image Denoising

Discrete wavelet transform (DWT) are used to concentrate the energy of a signal in a small set of coefficients. The wavelet transform of the noisy image results in a small number of coefficients with high Signal Noise Ratio (SNR) and a large number of coefficients with low SNR due to the mentioned characteristics. Therefore, a denoised image can be reconstructed using inverse wavelet transform after deleting or lowering the effect of coefficients with low SNR.

Fourier Transform or any orthogonal signal representation can be used to do the same denoising progress, but Fourier domain procedures results in over-smoothed edges since the filtering is done as a spatially global operation.

Any functions can be represented as a superposition of wavelet functions which helps decomposing the function into different scale levels where each level is further decomposed with a resolution adapted to that level. As shown in figure 2.6.a , DWT divides image into four sub bands and sub sampled. The sub bands are the results of different filters in horizontal and vertical axis. The sub bands are labeled as LH1, HL1 and HH1 and LL1. To move further to the next coarse level, only LL1 will be divided to four sub bands, since LL1 corresponds to coarse level coefficients. The two-level decomposition is depicted in figure 2.6.b.

This results in two-level wavelet decomposition as shown in Fig. 1(b).

LL1	HL1
LH1	HH1

a)

LL2	HL2	HL1
LH2	HH2	
LH1		HH1

b)

Figure 2.6. Image decomposition.

Figure 2.7 shows the basic wavelet based image denoising flowchart. This method is based on the fact that noise usually stays on the fine-grained structure in the image and since DWT is a scale based decomposition, therefore applying DWT recovers the coefficients at the finer scales which include noise. Deleting these coefficients is the process of filtering on the basis of scale. Selecting a threshold which corresponds to noise level and setting the coefficients lower than the threshold value to zero results in

discarding the noise. The edge relating coefficients on the other hand, are usually above the threshold. The denoised image is the reconstructed image using the IDWT.

```
Noisy Image
    │
    ▼
   DWT
    │
    ▼                    ┌──────────────┐
Thresholding  ◄────────► │ • Calculate  │
    │                    │   threshold  │
    │                    └──────────────┘
    │                    ┌──────────────┐
    │           ◄────────► │ • Apply      │
    │                    │   threshold  │
    ▼                    └──────────────┘
  IDWT
    │
    ▼
Denoised
Image
```
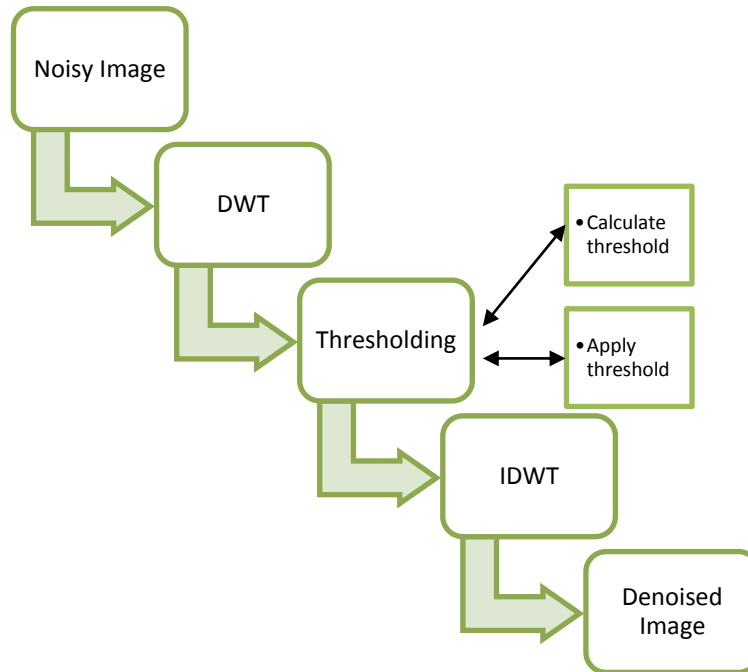
Figure 2.7. wavelet based Image Denoising flowchart.

Wavelet transform based image denoising is directly affected by wavelet thresholding which is a signal estimation technique. It surpasses noise by discarding coefficients that are insignificant comparing to a threshold value depending completely on a selection of a thresholding parameter and based on that the efficacy of denoising can be determined.

Threshold selection is an important question when denoising is going to be done. Selecting a threshold to a small value may yield a result close to the input, but the result may still be noisy. On the other hand, a threshold with a large value produces a signal with a large number of zero coefficients which means a smooth signal. Having a high degree of smoothness, however, destroys details and in image processing may cause blur and artifacts.

There are different thresholding methods such as: Hard thresholding, Soft thresholding, Semi-soft Thresholding and Quantile thresholding. Soft thresholding method is mostly used, since soft thresholding results in better denoising performance than other denoising methods [16].

2.5.2. PDE Techniques in Image Denoising

Recently modern PDE techniques in image processing have been introduced, although some traditional image denoising methods can be interpreted from PDE algorithms. One of the traditional examples is the classical Gaussian filter for image denoising which is defined by convolving the noisy image u_0 with the Gaussian kernel, $G(x,t) = \frac{1}{2\pi t} e^{\left(-\frac{x^2}{2t}\right)}$.

$$u = G * u_0 = \int u_0(y) G(x-y,t) dy. \tag{2.3}$$

This denoised image u is actually the solution $u(x,t)$ of the following diffusion PDE,

$$u_t(x,t) = D\Delta u(x,t), \quad u(x,0) = u_0(x) \tag{2.4}$$

where Δ is the Laplace operator , and $D = \frac{1}{2}$ is diffusive coefficient .

These techniques have shown success in the past two decades. The reason behind that is the ability to handle geometrical features unlike the traditional statistical or Fourier/wavelet based approaches. Usually two different strategies are used to design PDE techniques based on the application in image processing which it will be used.

1. Construct PDE-based evolution processes and incorporate geometry in the equations.

2. Develop the image processing activities in variational framework and compute the minimizers by deriving the corresponding Euler-Lagrange equations.

In both of them the image processing target can be achieved solving PDE's.

2.5.2.1. Anisotropic Diffusion for Denoising

The purpose of image denoising is to surpass the unwanted disturbances in images. Usually noise is highly localized and oscillatory, like white noise and pepper-and-salt noise, which makes it really difficult to separate noise from local and discontinuous edges. Diffusion would be a good option since its nature is an anti-oscillation procedure. As mentioned earlier, the classical Gaussian filter for denoising is equivalent to the linear isotropic diffusion. The main drawback of isotropic diffusion is that it smoothen the sharp edges, corners and other geometrical features existing in the noisy image itself. This is due to the uniqueness of the denoising produce in all orientations without recognizing the presence of spatially coherent discontinuities - edges. Besides that increasing the diffusive coefficient D will increase the speed of smoothing out.

To overcome this drawback, Perona-Malik [7] proposed using anisotropic diffusion instead,

$$u_t = \nabla.\left(D(x, u, \nabla u)\nabla u\right) \tag{2.5}$$

The diffusive coefficient D, is data dependent and should be able to recognize the edges, in order to do that the PDE stops diffusion across the discontinuities. Thus, D is designed to satisfy the following requirements,

$$D = \begin{cases} large, & when \ |\nabla u| \ is \ small \ on \ intra-regions, \\ small, & when \ |\nabla u| \ is \ large \ near \ edges. \end{cases} \tag{2.6}$$

By choosing the above criteria for choosing D, the algorithm only smoothens out the oscillations away from edges but not across them. For example Perona and Malik decided to select D as:

$$D = g(|\nabla u|^2)$$

<div align="right">(2.7)</div>

where g is a smooth positive concave function satisfying $g(+\infty) = 0$. For example, g can be assumed as below where σ and $b > 0$ are constants.

$$g(|\nabla u|^2) = e^{\frac{-|\nabla u|^2}{2\sigma^2}}, \quad \text{or } g(|\nabla u|^2) = \frac{1}{1+b|\nabla u|^2}$$

Practically selection of the anisotropic diffusion is problematic in the beginning of the process when the noisy image have a highly oscillatory noise, because $|\nabla u|$ is large almost everywhere, so D is small everywhere. Therefore using the diffusion is that much of help in this case. To overcome this difficulty, the use of a mollified image in g has been proposed in [25], with the form:

$$u_t = \nabla.(g(|\nabla(G_\sigma * u)|^2)\nabla u) \ , \ u(x,0) = u_0(x)$$

<div align="right">(2.8)</div>

where G_σ is again the Gaussian kernel .

2.5.3. Total Variation Image Denoising

Total Variation (TV) in the mathematical view is a quantity used to measure the oscillations in functions. In the other denoising can be defined as reducing the uncorrelated local oscillation in an image. Therefore it can be concluded that the oscillatory noise greatly increases the TV norm and the denoising process can be reducing the total variation of images. The well-known TV algorithm proposed in [21]

$$\min_u \int |\nabla u| dx \ \ subject \ to \ \ \parallel u - u_0 \parallel_2 \leq \sigma$$

<div align="right">(2.9)</div>

where σ is the noise variance. The purpose of this objective function is to reduce oscillations in the reconstruction, and the defined constraint term is a fitting requirement. The non-constraint minimization problem can be formulated as

$$\min_u \int |\nabla u| dx + \frac{\lambda}{2} \parallel u - u_0 \parallel_2 ,$$

(2.10)

λ is a lagrangian multiplier and $\lambda \geq 0$ as a balance factor between oscillations and fidelity. Decreasing the λ will result in a fewer details in the denoised images.

One of the advantageous characteristics of TV algorithm is preserving the sharp edges while doing the reconstruction. This confirms that TV model is reducing the noise while keeping the edges. Another attraction of TV denoising is its geometrical properties. An equivalent coarea formula can be used for functions with finite TV semi norms.

$$\int |\nabla u| dx = \int_{-\infty}^{+\infty} \int_{\{u=\gamma\}} ds d\gamma ,$$

(2.11)

Here the term $\int_{\{u=\gamma\}}$ is the length of the level set $\{u = \gamma\}$. The geometric connection of TV minimization is more visible if we analyze the optimization (2.10) by calculus of variation. The Euler-Lagrange equation of the minimizer is the following:

$$-\nabla \left(\frac{\nabla u}{|\nabla u|} \right) + \lambda (u - u_0) = 0.$$

(2.12)

To represent the curvature of the image, first term of the equation is used; which makes the method more geometric friendly. For noisy pixels, the jumps are isolated and their curvature is large. They will be wiped out much quicker than the edges that are coherent jumps with relatively smaller curvature.

One of the most common solutions to (2.10) is the gradient descent method, which introduces an evolution PDE,

$$u_t = \nabla\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(u - u_0). \tag{2.13}$$

This solution is an anisotropic diffusion requirement. So if the algorithm faces a sharp edge, *D* will be zero and no diffusion is performed across the edge.

2.5.4. CT Image Denoising Using Sparse Representations and Dictionary Learning

All of the methods explained in the previous sections suffer from high computational complexity and acquisition of an accurate and explicit modeling of the system and noise. To overcome the mentioned issues, another approach is used which is adaptive processing tools that can automatically adjust and implicitly learn the data statistics. The advantage is performing the learning process offline and as a result lower computational complexity and improved performance due to trained system is expected. An adaptive learning and sparse representations of medical images is explained in this section applying on low dose scans. The algorithm is based on imposing the sparsity in a patch-wise manner to the CT scans using the dictionary. In the first step the image will be transformed into a linear combination of dictionary atoms. The famous dictionary learning K-SVD algorithm has been proposed in [2]. A denoising method based on dictionary learning is presented in [9]. In the rest of this section we will present a CT noise reduction method using sparse model.

2.5.4.1. Learning Adaptive and Sparse Representations of Medical images

In [26], Stagliano et.al. proposed a method using adaptive dictionary learning and sparse representation of medical images. K-SVD dictionary learning algorithm is used here adapted to CT denoising. This

procedure is consisted of two stages: the training stage and the denoising stage. In the training stage, the K-SVD algorithms employed to optimize

$$\lim_{D,x_i} \sum_i \| y_i - Dx_i \|_2^2 \ \ subject \ to \ \ \| x_i \|_0 < T \tag{2.14}$$

Where D is the dictionary and x_i is the sparse representation of the noisy image patches y_i, such that less than T non-zero coefficients are used to reconstruct each patch. $\| \|_0$ denotes the number of non-zero coefficients in a matrix. The optimization procedure starts with assuming a fixed dictionary and x_i is optimized then using the obtained sparse matrix; the dictionary is going to be optimized till convergence. Here is a summary of the two main steps:

- *Sparse coding*: Assuming a fixed dictionary, sparse representation of each patch is calculated. This is done using the Orthogonal Matching Pursuit [9].

- *Dictionary update*: for each dictionary column d_i, minimize the representation error [9].

After training the dictionary, denoising framework starts. The noisy image can be represented as:

$$Y = Z + n \tag{2.15}$$

Assuming a noise component n adding to an original image Z, The cost function for denoising can be written as:

$$\arg\min_z \| Z - Y \|_2^2 + \lambda \| . \|_p \tag{2.16}$$

λ is the Lagrange multiplier. In a sparse framework it is then formulated to solve for,

$$\mathrm{X}, \mathrm{Z} = \arg\min_{X,Z} \| Z - Y \|_2^2 + \lambda \| DX - Z \|_2^2 + \sum_i \mu_i \| x_i \|_0 \tag{2.17}$$

Solving a problem for each patch, the problem will change to:

$$x_i, Z = \arg\min_{x_i, Z} \parallel Z - Y \parallel_2^2 + \lambda \parallel Dx_i - R_i Z \parallel_2^2 + \sum_i \mu_i \parallel x_i \parallel_0 \qquad (2.18)$$

The R is the window operator which selects the i-th patch from the image. Changing the Lagrange multiplier changes the similarity of the denoised output image with the noisy image. The last term is implicitly accounted for in the sparse coding stage and can be removed from the equation. Hence denoising solution becomes:

$$Z = \frac{(\lambda Y + \sum_i R_i^T Dx_i)}{(\lambda I + \sum_i R_i^T R_i)} \qquad (2.19)$$

In other words, the solution is computed as a weighted average of the patches. The OMP algorithm is used to approximate the solution to the following:

$$\arg\min \parallel x_i \parallel_0 \; subject\; to \; \parallel Dx_i - R_i Z \parallel_2^2 < (C\sigma^2) \qquad (2.20)$$

Here, the error tolerance is defined as the noise variance of the image, multiplied by a gain factor C. The benefit of using adaptive dictionary and sparse representation using a set of noisy images as training samples and its flexibility to capture the distribution of data is investigated.

Chapter 3. Sparse Representation and Dictionary Learning

This chapter describes the sparse representations of signals and how to find them based on an overcomplete dictionary that contains a set of elementary signals. If the overcomplete dictionary is designed based on the noise-free part of image, sparse representations can be used for image denoising. This dictionary can be obtained using K-SVD algorithm, which is described in this chapter. The dictionary learned using K-SVD algorithm is adapted to the noisy image as a training sample and it is used to denoise the image, on which it was trained. This is due to the fact that white Gaussian noise is not learned by that dictionary.

3.1. Sparse Representation

Sparse representation is reconstructing a signal based on a set of elementary vectors to represent the signals compactly by only few of these vectors. Sparse representation can be obtained from different signal and image processing methods for the applications of compression, feature extraction, image inpainting and denoising.

Given an overcomplete dictionary, $D \in \mathbb{R}^{n \times K}$, where $k \geq n$, and a signal $y \in \mathbb{R}^n$, the sparse representation problem can be formulated by:

$$\min_{a} \| a \|_0 \ subject \ to \ Da = y \tag{3.1}$$

where $\|.\|_0$ is the l_0- norm that shows the nonzero elements of a vector: $\|.\|_0 = \sum_{j=0}^{K} |a_j|^0$

For a known error tolerance ϵ, the problem can be formulated with l_2 - norm:

$$\min_{a} \| a \|_0 \ subject \ to \ \| Da - y \|_2 < \epsilon \tag{3.2}$$

Or for a present number of non-zero coefficients, T, in the sparse representation by

$$\min_{a} \| Da - y \|_2 \ subject \ to \ \| a \|_0 < T \tag{3.3}$$

The sparsest solution, results to a most compact representation of signals. If ϵ is selected such as it is larger than the coherence of noise to all dictionary atoms, sparse representation can be used to remove noise from signals. Different methods have been proposed to solve the mentioned problem; the most used category is pursuit methods which will be discussed in the section 3.2.

Uniqueness of sparse representations is related to the properties of the dictionary. The dictionary limits the sparsity of signals that are to be sparsely represented by its atoms. This sparsity guarantees also stability in recovering the solution of (3.2). In this thesis, all dictionary atoms are linearly independent and they are assumed to be normalized with respect to l_2-norm.

Linear independency of dictionary atoms can be measured using the mutual coherence $\mu(D)$ of a dictionary D. This measure is defined as the maximum absolute scalar product between each pair of normalized dictionary atoms:

$$\mu(D) = \max_{i \neq j} \frac{d_i^T d_j}{d_i^T d_i d_j^T d_j} \tag{3.4}$$

Mutual coherence for orthogonal dictionaries is equal to zero while $\mu(D)$ for overcomplete dictionaries is larger than zero. If the mutual coherence is small, dictionaries are called incoherent and deviate only slightly from orthonormal and therefore they are desirable for sparse representation purposes. The worst case is when there exist parallel atoms in the dictionary that results in $\mu(D) = 1$, which causes confusion in finding sparse representations. Using Gram matrix $G = D^T D$ of a dictionary, the mutual coherence can be computed as the off-diagonal entry of the Gram matrix with maximal absolute value:

$$\mu(D) = \max_{1 \leq i,j \leq K, i \neq j} |G_{i,j}| \tag{3.5}$$

According to [27], the following relation holds for full rank dictionaries:

$$\mu(D) > \sqrt{\frac{K - n}{n(K - 1)}} \tag{3.6}$$

Therefore with $K \gg n$, the complexity of $O\left(\frac{1}{\sqrt{n}}\right)$ can be expected for the mutual coherence.

Spark of the dictionary σ defined as the smallest number of atoms that form a linearly dependent set; is another measure for dictionary properties. [28] The relationship between spark and mutual coherence can be formulated as:

$$\sigma(D) = 1 + \frac{1}{\mu(D)} \qquad\qquad (3.7)$$

The uniqueness of the solution to (3.1) as a linear representation is guaranteed if $\| a \|_0 < \frac{\sigma(D)}{2}$ meaning that the sparsest representation has $\frac{\sigma(D)}{2}$ non-zero elements.

As explained earlier solving the sparse representation problem is an optimization problem to find the minimal number of atoms to reconstruct the signal based on K atoms in the dictionary .Assuming the sparse representation consisting of m atoms, the minimum computational complexity of this technique is $O(Km)$ [28].

Solving the sparse representation problem with minimal approximation error is an NP-hard problem. This is implied by possible NP-completeness of minimizing the approximation error. A reason behind that is that linear combination of atoms for sparse representation of signal cannot be progressively refined. This is due to the fact that slightly different signals can be represented with a completely different set of atoms. Another reason is that sparse representations do not have any optimal substructure property.

Different optimization methods were introduced to aim to solve this problem such as greedy pursuits, convex and non-convex optimization, based on stochastic modeling and exhaustive search. The most simple and efficient technique is greedy algorithms. These algorithms gradually increase the selected support of coefficient vector to reduce the reconstruction error. These methods are preferred especially when the size of the problem is large. Here we have large size dictionaries and so greedy algorithms are used and are explained briefly in the following section.

3.2. Pursuit Methods

Pursuit methods are one of the most common used greedy algorithms. These methods calculate the inner product of the atoms with the residuals and then select the one with the largest correlation.

Under the following condition $\| a \|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(D)}\right)$, the pursuit methods will find sufficiently sparse solutions [29].

In the case of presence of the noise, greedy algorithms result in a locally stable sparse representation [28] under two conditions: The dictionary is mutually incoherent and it results to sufficiently sparse representations for the ideal noiseless signal. Following the mentioned conditions, greedy algorithms such as MP and OMP recover the ideal sparse signal representation, with an error proportionally to the noise level which can be used for signal denoising.

As explained earlier, MP recursively finds the highly correlated signal components with few dictionary atoms and then replaces them with a sparse linear combination of these atoms. As long as the highly correlated atoms are not chosen, in the first iterations, the representation error decays quickly. Later on by selection of similar atoms, the residual decays slowly and they would behave like realizations of white noise. Therefore, the noise free part of the signal can be recovered, if the original signal is well represented by few dictionary atoms and if ϵ fits the present noise.

Table 3.1. The pseudo algorithm of MP algorithm

Objective: Approximation of the solution to the optimization problem, $\min_a \| a \|_0$ subject to $\boldsymbol{y} = \boldsymbol{Da}$

Given: Dictionary $\boldsymbol{D}$, the input signal $\boldsymbol{y}$ (training signal), and the error tolerance ϵ

Initialization: $k = 0$, Initial approximation $\boldsymbol{a^0} = \boldsymbol{0}$, Residual $\boldsymbol{r^0} = \boldsymbol{y} - \boldsymbol{Da^0}$ and $\boldsymbol{S^0} = \{\}$

Main Iteration: $k = k + 1$

- **Sweep Stage:** Find the error $e(j) = \min_a \| \boldsymbol{d}_j z_j - \boldsymbol{r}^{k-1} \|$ for all j using optimal solution $z_j = \boldsymbol{d}_j^T \boldsymbol{r}^{k-1} / \| \boldsymbol{d}_j \|_2^2$

- **Update Stage:** Update S, Find a minimizer, j_0 of $e(j): \forall 1 \leq j \leq m$ and $e(j_0) \leq e(j)$ then update $\boldsymbol{S}^k = \boldsymbol{S}^{k-1} \cup \{j_0\}$
- **Update the Solution:** Set $a^k = a^{k-1}$ and update $a^k(j_0) = a^k(j_o) + z_j$

- **Stopping Criterion:** $\| \boldsymbol{r} \|_2 \leq \epsilon$, otherwise apply another iteration

Output: $\boldsymbol{a} = \boldsymbol{a}^k$

3.2.1. Orthogonal Matching Pursuit

The matching pursuit is able to optimally decompose the signal in the case that atoms are pairwise orthogonal. However, there is no guarantee that it converges into an efficient response if the basis waveforms are not orthogonal. The Orthogonal Matching Pursuit (OMP), proposed by [30], is an alternative solution based on the matching pursuit which provides a fast convergence with non-orthogonal dictionaries. Adding an additional orthogonalization step, MP algorithm is extended to OMP algorithm, in which the original signal is orthogonally projected to the span of previously selected atoms. Therefore the resulting residuals after applying this step is orthogonal to all previously selected atoms. This ensures the linear independency of the recently chosen atoms for sparse representation with previous ones.

Table 3.2. The pseudo algorithm of OMP algorithm

Objective: Approximation of the solution to the optimization problem, $\min_a \| a \|_0$ subject to $y = Da$

Given: Dictionary D, the input signal y (training signal), and the error tolerance ϵ, and the scalar $0 < t < 1$

Initialization: $k = 0$, Initial approximation $a^0 = 0$, Residual $r^0 = y - Da^0$ and $S^0 = \{\}$

Main Iteration: $k = k + 1$

- **Sweep Stage:** Find the error $e(j) = \min_a \| d_j z_j - r^{k-1} \|$ for all j using optimal solution $z_j = d_j^T r^{k-1} / \| d_j \|_2^2$

- **Update Stage:** Update **S**, Find a minimizer, j_0 of $e(j) : \forall 1 \leq j \leq m$ and $e(j_0) \leq e(j)$ then update $S^k = S^{k-1} \cup \{j_0\}$
- **Update the Solution: calculate** a^k, the solution to $\| y - Da \|_2^2$ with respect to S^k

- **Stopping Criterion:** $\| r \|_2 \leq \epsilon$, otherwise apply another iteration

Output: $a = a^k$

The error value in the sweep stage is the following:

$$e(j) = \min_{a} \parallel \boldsymbol{d}_j z_j - \boldsymbol{r}^{k-1} \parallel_2^2 = \parallel \boldsymbol{r}^{k-1} \parallel_2^2 - \frac{(d_j^T \boldsymbol{r}^{k-1})^2}{\parallel \boldsymbol{d}_j \parallel_2^2} \tag{3.8}$$

Similar to MP algorithm, the largest inners product between the residuals and dictionary atoms is when the error is minimum. $\parallel \boldsymbol{y} - \boldsymbol{Da} \parallel_2^2$ is updated with respect to $\boldsymbol{a}$ in the update stage using the support $\boldsymbol{S}^k$. Therefore the problem is reduced to $\parallel \boldsymbol{y} - \boldsymbol{D}_{\boldsymbol{S}^k} \boldsymbol{a}_{\boldsymbol{S}^k} \parallel_2^2$, using the dictionary $D_{\boldsymbol{S}^k}$ with size $n \times \boldsymbol{S}^k$ containing the corresponding atoms and the non-zero coefficients $\boldsymbol{a}_{\boldsymbol{S}^k}$.

3.3. K-SVD

All of the explained sparse representation methods work for known and fixed dictionaries. Quality of the sparse representation is affected by the selection of a dictionary in terms of the level of sparsity, the error of representation, the desired characteristic extractions and etc. As for analytical dictionaries, each dictionary highlights a special property of the input signal. For example the DCT dictionary extracts localized frequency domain information, whereas the wavelet conducts a multi-resolution decomposition. All these dictionaries are independent of input signal contents regardless of what characteristics they are focused on. However, to reach the maximum sparsity, it is of great interest to create a dictionary well adapted to the input signal. To obtain the best approach to this matter, different algorithms have been proposed. Dictionary learning is used to create a dictionary that best represents a given signals. In order to do this, the signals are contaminated in a matrix and the matrix will be sparse factorized to be represented into the dictionary matrix and a sparse coefficient matrix.

3.3.1. K-means and K-SVD

Generalizing the K-means method, the K-SVD method is proposed by [2] to address the dictionary learning problem. A dictionary of codewords $C = [c_1, c_2, \ldots, c_L]$ based on a training algorithm is calculated in K-means method. Using the minimum distance $\parallel y - c_i \parallel_2^2 \geq \parallel y - c_k \parallel_2^2, \forall i \neq k$, each

input signal is represented to its closest code word. To summarize k-means method, one can say that it will be done by the several iterations of the following two steps:

- Performing a nearest-neighbor criterion on the training signals to the most similar of the K codewords entries.

- Updating each codeword with the mean value of the signals in that set, which was assigned to this entry in the previous step.

K-means is the extreme case of the sparse representation problem with only one nonzero coefficient and the coefficient is 1. The K-SVD method is a generalized K-means method in which the representation of each signal can be done by more nonzero coefficients with arbitrary values between 0 to 1.

As mentioned earlier, given a set of N training signals in a matrix $Y \in \mathbb{R}^{n \times N}$, the K-SVD method adapts the dictionary D to best represent these signals. The purpose is to divide the matrix Y into the dictionary D containing K atoms, and matrix A consisting sparse representation coefficients for each of the N signals Y:

$$Y = DA \tag{3.9}$$

For this case, the minimization problem is,

$$\min_{D,A}\{\parallel Y - DA \parallel_F^2\} \; subject \; to \quad \parallel a_i \parallel_0 \leq L \,, \forall i \tag{3.10}$$

Similarly here at each iteration, two steps are performed. The first step considers that the dictionary, D, is fixed and using any possible method such as OMP, determines the sparse vector,A.This step is called a sparse coding stage. In the second step, the obtained sparse vector is fixed and the dictionary is updated to minimize the optimization problem (3.10).This step is called the dictionary update stage. To perform the second step, the dictionary columns $\tilde{d}_i$, are updated individually in which each column is calculated

to minimize the mean square error of the input signal reconstruction using all other dictionary columns (atoms). The problem of updating only one column can be addressed using the singular value decomposition (SVD). For updating only the column, d_k, the following optimization problem should be solved,

$$\| Y - DA \|_F^2 = \| Y - \sum_{j=1}^{L} d_j a_j^T \|_F^2 = \| \left(Y - \sum_{j \neq k} d_j a_j^T \right) - d_k a_k^T \| = \| E_k - d_k a_k^T \|_F^2 \tag{3.11}$$

The column, d_k, should be updated to reduce the reconstruction error, E_k, obtained by all other atoms. Using the SVD, new values for both d_k and a_k^T are obtained. However, a_k^T may have many nonzero coefficients which is against the sparsity constraint. The key to solve this problem is to consider only nonzero coefficients and corresponding input signals. Then, the obtained updated result corresponds to nonzero coefficients of existing sparse vector and it maintains the sparsity condition. For each atom update, a set of indexes of atoms which are involved to represent the updating atom is defined as,

$$\omega_k = \{i \mid 1 \leq i \leq L, A_k^T(i) \neq 0\} \tag{3.12}$$

Now, a matrix $\Omega_k \in \mathbb{R}^{N \times lenght(\omega_k)}$ is formed with ones on $[col: \omega_k(i) \quad row: i]$ positions and zeros elsewhere. Now, the transform $a_k^R = a_k^T \Omega_k$ only keeps nonzero coefficients. The same thing is applied on the input signals, $Y_k^R = Y \Omega_k$, which shrinks the input signals to relative input signals which are represented with the k-th atom. Similarly, it is applied on the error, $E_k^R = E_k \Omega_k$, and the minimization problem is modified,

$$\tilde{d}_k = \arg \min_{d_k} \| E_k^R - d_k a_k^R \| \tag{3.13}$$

Now, we can decompose E_k^R using the SVD into $E_k^R = U\Delta V^T$. $\tilde{d}_k$ is the first column of U and a_k^R is the

first column of V multiplied by $\Delta(1,1)$. In Table , the pseudo-algorithm of the K-SVD method is presented

which shows the simplicity of this method.

Table 3.3. The pseudo-algorithm of the K-SVD method.

Objective: Train dictionary D to sparsely represent the training data $\{y_i\}_{i=1}^M$

Given: the input signal y (training signal), and the error tolerance ϵ

Initialization: $k = 0$, initialize the dictionary $D \in \mathbb{R}^{n \times k}$ by random entries or k randomly chosen examples

Main Iteration: $k = k + 1$

- **Sparse Coding:** Use any pursuit algorithm to solve $\arg\min_A \parallel DA - Y \parallel$ subject to $\parallel A_0 \parallel \leq T_0$

- **Dictionary Update:** For each column in D, d_k ,

    a) Find ω_k, according to (3.12).

    b) Compute the representation error E_k^R

    c) Apply SVD decomposition and find $\tilde{d}_k$ and a_k^R

- **Update the Solution: calculate** a^k, the solution to $\parallel Da - Y \parallel_2^2$ with respect to S^k

- **Stopping Criterion:** $\parallel d_k a_k - Y \parallel_2^2 \leq \epsilon$, otherwise apply another iteration

Output: D

If the sparse coding stage is performed perfectly, the convergence of the K-SVD algorithm is guaranteed.

In each sparse coding step, the total representation error $\parallel DA - Y \parallel_F^2$ is decreased. In addition, the

mean squared error (MSE) is reduced in the dictionary update stage.

The sparsity constraint is not violated. Thus, convergence of the K-SVD algorithm to a local minimum is

guaranteed, if the pursuit algorithm robustly finds a solution.

3.3.2. Denoising with the K-SVD Algorithm

M.Elad has proposed the image denoising problem based on sparse representation problem and the K-SVD algorithm in [9] to recover the ideal underlying image x, that has been corrupted with additive, zero-mean, white, and Gaussian noise n with standard deviation σ

$$y = x + n \tag{3.14}$$

The K-SVD algorithm has been used here to learn the dictionary from the noisy image y. The learnt dictionary represents the structure of the underlying noise-free image x. The main point here is the stationary Gaussian white noise is not learned by that dictionary.

By decomposing the image into small patches of size $\sqrt{n} \times \sqrt{n}$, we can be free of the high dimension and of course the curse of dimensionality and the small patches can be interpreted to signals stored in vectors $y \in \mathbb{R}^n$. The denoising of the patches can be done by computing their noiseless sparse approximation for a given dictionary, and then reconstructing them from the sparse representation.

After finding the sparse representation $\hat{a}$ using the below relationship, the denoised patch is obtained from $\hat{x} = D\hat{a}$:

$$\hat{a} = \min_{a} \| a \|_0 \ subject \ to \ \| Da - y \|_2^2 \leq L \tag{3.15}$$

L is defined using the value of ϵ and σ of the patch. Changing the constraint to a penalty, the optimization problem can be formulated as:

$$\hat{a} = arg \min_{a} \| Da - y \|_2^2 + \mu \| a \|_0 \tag{3.16}$$

For a certain value of μ, both problems are equivalent. In the first step the K-SVD denoising algorithm trains the dictionary on patches of the noisy image Y, and in the next step based on the resulting dictionary reconstructs the denoised image X.

The optimization problem for denoising can be formulated as follows:

$$\hat{X} = arg\min_{X,D,A} \left\{ \lambda \parallel Y - X \parallel_2^2 + \sum_{ij} \mu_{ij} \parallel a_{ij} \parallel_0 + \sum_{ij} \parallel Da_{ij} - R_{ij}X \parallel_2^2 \right\} \qquad (3.17)$$

R_{ij} is a window operator which extracts the patch (i,j) from the image. The trend of the patch selection is generally overlapping in order to prevent blocking artifacts at the edges of patches. As the first term stand, λ is a Lagrange multiplier which controls how close the denoised output image $\hat{X}$ will be to the noisy image. The parameter μ_{ij} determines the sparsity of the patch (i,j).

To solve the above problem, first, the initial dictionary is selected; it can be an analytical for example DCT or adaptive dictionary based on training on noisy patches from the image. The denoised output image X is initialized with the input noisy image, $X = Y$. The next step is applying the K-SVD algorithm for several iterations with the following steps:

- Sparse Coding Stage: The sparse representation vectors a_{ij} of each patch $R_{ij}X$ is computed. Any pursuit algorithm can be used to approximate the solution of the sparse representation problem:

$$\forall_{ij} \min \parallel a_{ij} \parallel subject\ to\ \parallel Da_{ij} - R_{ij}X \parallel_2^2 \leq (C\sigma^2) \qquad (3.18)$$

The error tolerance is defined as the noise variance of the image, multiplied by a gain factor C.

- Dictionary update stage: The update will be performed as described in previous section, on the patches of the noisy image.

When the training is over, the output denoised image X is computed by (3.19). The resulting image have a different noise level so a recalculation for σ is needed.

$$\hat{X} = arg\ \min_{x}\left\{\lambda \parallel Y - X \parallel_2^2 + + \sum_{ij} \parallel Da_{ij} - R_{ij}X \parallel_2^2\right\} \qquad (3.19)$$

The closed-form solution of this simple quadratic term is:

$$\hat{X} = \left(\lambda I + \sum_{ij} R_{ij}^T R_{ij}\right)^{-1}\left(\lambda Y + \sum_{ij} R_{ij}^T D\hat{a}_{ij}\right) \qquad (3.20)$$

By averaging the denoised patches, the solution can be reached, which are obtained from the coefficients a_{ij} and the dictionary D.

3.4. Principal Component Analysis

The Principle Component Analysis (PCA) is an unsupervised dimension-reduction technique to reduce the dimensionality of a data set with a large number of interrelated variables. PCA seeks the linear combination of the original variables such that the derived maximal variables capture maximal variance. The PCA method transforms the space represented by the current variables into a new set of variables which are uncorrelated. Only the most uncorrelated variables are kept to reduce the dimensionality while maintaining the data set variation [31]. The PCA can be computed via the Singular Value Decomposition (SVD) of the Data Matrix, or using eigenvalue decomposition of the data covariance matrix [32].

In detail, let the data X be a $n \times p$ matrix, where n and p are the number of observations and the number of variables, respectively. Suppose the data has a zero empirical mean value (mean of columns are zero). The singular value decomposition of X is $X = UDV^T$ in which U is a $n \times n$ matrix of eigenvectors of XX^T, D is a $n \times p$ rectangular diagonal matrix and V is a $p \times p$ matrix of eigenvectors of $X^T X$. The principal components are columns Z_i of $Z = UD$ with the variance of D_{ii}^2/n. V is the corresponding loading vector of principal components. Usually the first $q\ \{q \ll \min(n, p)\}$ principal

components are chosen to represent the data, thus a great dimensionality reduction is achieved. The success of PCA is due to the following two important properties:

1. Principal components sequentially capture the maximum variability among the columns of X, thus quarantining minimal information loss.

2. Principal components are uncorrelated to provide a better analysis on the data with converting the data into separable modes.

Due to the useful characteristics of the PCA, different applications have been derived based on this method such as handwritten character recognition [33], human face recognition [34], gene expression data analysis [35] and etc. The basic PCA defines the principal components as a linear combination of all the original variables which have non-zero values. Figure 3.1 depicts a linear combination of PCA components to represent as image with $q_1 = 0.078$, $q_2 = 0.062$, $q_3 = -0.182$ and $q_4 = 0.179$.

Figure 3.1. Reconstruction of the image from four basis vectors which can be displayed as images. $q_1 = 0.078$, $q_2 = 0.062$, $q_3 = -0.182$ and $q_4 = 0.179$. [36]

This problem is known as the loading problem and results in a difficult interpretation of the results [37]. This problem is addressed in a very simple way which is defining a threshold to set loadings below the threshold equal to zero, and is called *Simple Thresholding*. Based on a linear regression method, Lasso, proposed by [38], Sparse PCA is proposed by [37] to address this loading problem. We first introduce the Lasso method and then continue to describe the sparse PCA approach.

3.4.1. Multiple Linear Regression, Lasso and Elastic Net

It is desirable not only to achieve the dimensionality reduction but also to reduce the number of explicitly used variables. An ad hoc way to achieve this is to artificially set the loadings with absolute values smaller than a threshold to zero. Multiple linear regression is a method which models the relationship between an measured variable (response vector), $y = (y_1, \ldots, y_n)^T$, and a set of explanatory variables (predictors), $X_j = (x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$. Linear regression methods are a penalized least squares method, imposing a constraint on the l_1 norm of the regression coefficients. The linear regression is used to find the relative strength between Y and X_j.

$$y = X\beta + \varepsilon \tag{3.21}$$

where ε models an additive noise and β is a p-dimensional regression coefficients.

To maintain the sparseness of the result, the Least Absolute Shrinkage and Selection Operator (Lasso) method combines the regression model with an additional constraint on the regression coefficients.

$$\beta_{lasso} = \arg \min_{\beta} \| Y - \sum_{j=1}^{p} X_j \beta_j \|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.22}$$

In order to mix the linear regression model with an l_1 optimization problem on the regression coefficients, the Lagrange coefficient λ is introduced in this equation.

The most relevant limitation of this work is that the number of variables selected by the lasso method is limited by the number of observations. To overcome this drawback, Elastic Net proposed by [39] which generalizes the lasso as following.

$$\beta_{en} = (1 + \lambda_2) \left\{ \arg\min_{\beta} \| Y - \sum_{j=1}^{p} X_j \beta_j \|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.23}$$

The elastic net penalty is a convex combination of the lasso penalty and a ridge penalty. It is shown that the elastic net can potentially include all variables in the fitted model, so the elastic net solves the limitation of the lasso method in terms of the number of selected variables [39].

3.4.2. Sparse Principal Component Analysis

Simplified Component Technique-LASSO (SCoTLASS), has been proposed by [40] which directly imposes l_1 constraint on the PCA. Since SCoTLASS method doesn't provide a convex optimization problem, it is suffering from high computational complexity which makes it an impractical method. The SCoTLASS optimization problem is,

$$a_k = \arg\min_{a_k} \left\{ a_k^T (X^T X) a_k \right\} \text{ subject to } a_k^T a_k = 1 \text{ and } \sum_{j=1}^{P} |a_{k,j}| \le t \tag{3.21}$$

The Sparse PCA method was first proposed by [37]. The Sparse PCA combines the Lasso, multiple linear regression method, and the concept of SCoTLASS in order to sparsify the loading coefficients while holding the maximum data variance. The sparse PCA is performed in two steps.

- First step: PCA is performed using the SVD method.

- Second step: A suitable sparse approximation based on the Lasso method is fulfilled using the following equation:

$$\hat{\beta} = arg \min_{\beta} \| Z_i - X\beta \|^2 + \lambda \| \beta \|^2 + \lambda_1 \| \beta \|_1 \qquad (3.22)$$

where $Z_i = U_i D_{ii}$ is the i-th principal component and $\| \beta \|_1 = \sum_{j=1}^{P} |\beta_j|$ are the l_1-norm. The i-th approximated loading $\hat{V}_i = \frac{\beta}{\|\beta\|}$ which is a sparse approximated model of V_i. In fact, the term $X\hat{V}_i$ approximates Z_i. The larger value of λ leads to the production of more zero coefficients in β. In this approach the principle components should be determined individually and then sparsify the loadings in the second step. The reader is referred to [37] in which a numerical solution is provided for this problem.

Chapter 4. Methodology

Sparse representation and an adaptive dictionary can be utilized to remove the noise effectively. Inspiring from the repeated patterns especially at the edges of medical images, structural clustering with sparse representation is combined. In this case the denoising algorithm can benefit from grouping the similar patches and present higher sparsity in denoising. The basic idea behind the CSR model is to treat the local and nonlocal sparsity constraints (associated with dictionary learning and structural clustering respectively) as peers and incorporate them into a unified variational framework.

4.1. Clustering-based Sparse Representation Model

Assuming an image I represented with the set of sparse coefficients $a = \{a_i\}$, each patch I_i at the spatial location i, extracted from the image using a rectangular windowing operator R_i can be shown as;

$$I_i = R_i I \tag{4.1}$$

The solution to the above problem can be easily obtained using Least-Square solution, which is an averaging method on an overlapped patches.

$$I = \left(\sum_i R_i^T R_i \right)^{-1} \left(\sum_i R_i^T I_i \right) \tag{4.2}$$

It is of interest when overlapping is allowed since it helps preventing artifacts in the edges and therefore, such patch-based representation is highly redundant and the recovery of I from $\{I_i\}$ becomes an over-determined system.

Assuming dictionary Φ has been used for sparse representation, the relationship between each patch to its sparse coefficients $\{a_i\}$ is by the following equation:

$$I_i = \Phi a_i \tag{4.3}$$

Thus the image can be obtained using equation (4.3),

$$I = Da = \left(\sum_i R_i^T R_i \right)^{-1} \left(\sum_i R_i^T \Phi a_i \right) \tag{4.4}$$

Assuming $Y = I + N$ is the noisy image and λ is the standard Lagrangian multiplier, the solution to the sparse representation problem can be formulated using the following variational problem,

$$a = \arg\min_a \frac{1}{2} \parallel Y - Da \parallel_2^2 + \lambda \parallel a \parallel_1 \tag{4.5}$$

As explained in the previous chapter a lot of efforts has been done to solve the above convex optimization problem, from designing an efficient dictionary to reducing the complexity of it while having a robust solution.

There is just one drawback to all mentioned solutions which is that the sparse coefficients a are NOT randomly distributed shown in Figure 4.1. Exploiting a *location related* constraint can help to achieve higher sparsity since the location uncertainty is often related to the nonlocal self-similarity of image signals. Applying such non-linear constraint (location-related constraint) is possible by using clustering. There are many different clustering algorithms reported in the literature such as kmeans, K-Nearest Neighbour (kNN), spectral and ect. Data clustering and sparse representation are two tools acting on different levels, middle level and low level respectively, therefore it is usually difficult to develop a connection between these two.
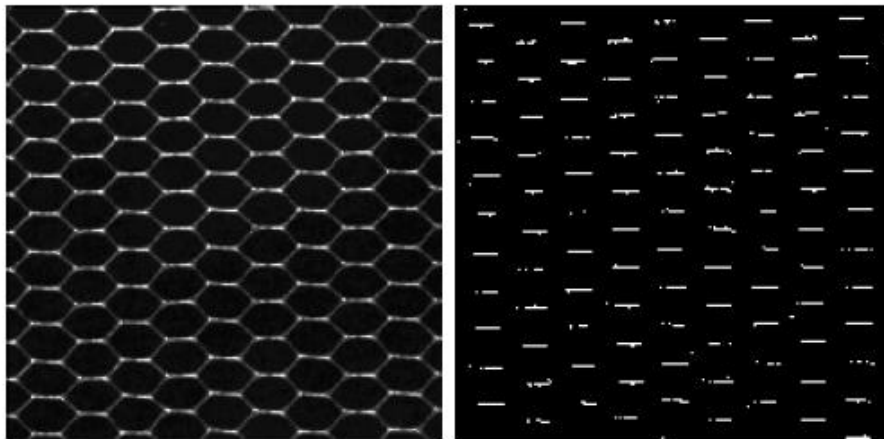
Figure 4.1. Limitation of K-SVD: a) an image of regular texture; b) spatial distribution of sparse coefficients corresponding to the 6th basis vector (note that their locations are not random) [1]

In order to show that how nonlocal self-similarity can result in higher sparsity, the following cost function is reviewed.

$$(a, \mu) = \arg\min_{a,\mu} \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2 + \lambda_1 \parallel a \parallel_1 + \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} \parallel \Phi a_i - \mu_k \parallel_2^2 \qquad (4.6)$$

where μ_k is the centroid of the kth cluster C_k of coefficients a. In other words, the weighting coefficients a are re-encoded with respect to μ_k in the new clustering-based regularization. As a consequence of exploiting nonlocal self-similarity, further compression and sparser representation can be obtained. BM3D and Learned Simultaneous Sparse Coding (LSSC) are proposed based on a similar observation about clustering and sparsity with different approach. [41] [42]

Rewriting the equation 4.6, a suitable formulation can be reached, in which all centroid vectors are represented with respect to the same dictionary Φ as I_i: $\mu_k = \Phi\beta_k$

$$(a, \beta) = \arg\min_{a,\mu} \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2 + \lambda_1 \parallel a \parallel_1 + \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} \parallel \Phi a_i - \Phi\beta_k \parallel_2^2 \qquad (4.7)$$

Taking an advantage of the unitary property of Φ, $\parallel \Phi a_i - \Phi\beta_k \parallel_2^2 = \parallel \alpha_i - \beta_k \parallel_2^2$. Thus, the following joint optimization problem can be obtained.

$$(a, \beta) = \arg\min_{a,\mu} \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2 + \lambda_1 \parallel a \parallel_1 + \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} \parallel a_i - \beta_k \parallel_2^2 \qquad (4.8)$$

The l_2-norm is replaced by l_1-norm in this regularization based on the successful result of compressed sensing [43].

$$(a, \beta) = \arg\min_{a,\mu} \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2 + \lambda_1 \parallel a \parallel_1 + \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} \parallel a_i - \beta_k \parallel_1 \qquad (4.9)$$

CSR model is offering higher sparsity level by combining dictionary learning and structural clustering. In this method β_k's are learned through structural clustering to encode a_i's at a higher level. It is shown in Figure 4.2 that CSR appears to be sparser by re-encoding a into $\gamma = a - \beta$ and exploiting the nonlocal similarity. The comparison of the learned dictionaries of K-SVD and CSR is depicted in Figure 4.3.
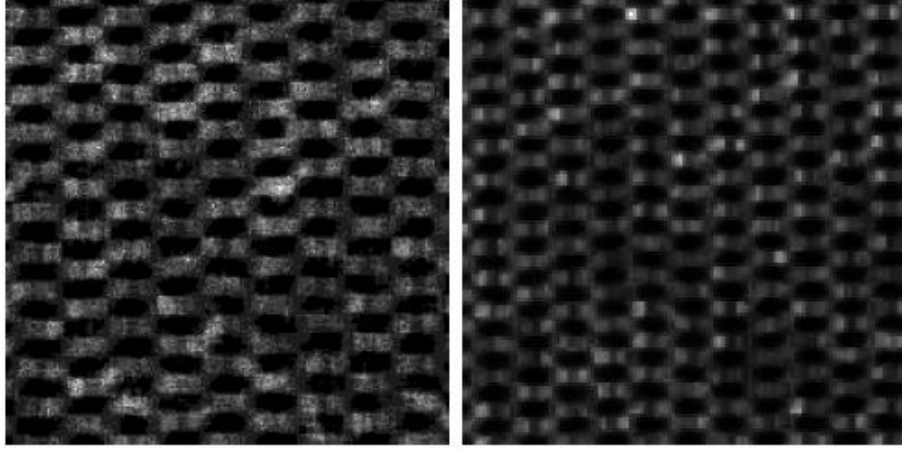
Figure 4.2. Comparison of sparsity distribution between K-SVD and CSR: a) spatial distribution of a plotted on a block-level (B = 8); b) spatial distribution of $\gamma = a - \beta$ plotted on a block-level (B = 7); note that how the introduction of β (cluster centroids) makes the CSR sparser. [1]

4.1.1. Iterative Reweighted and Regularized l_1-Minimization

The solution to the above *double-header* l_1-optimization problem can be achieved through an iterative algorithm alternatively updating a and β. Using the proposed surrogate functions in [44], an iterative shrinkage operator is used to update a for fixed β.

$$a_j^{(i+1)} = \begin{cases} S_{\tau_1, \tau_2}\left(u_j^{(i)}\right) & \beta_j \geq 0 \\ -S_{\tau_1, \tau_2}\left(-u_j^{(i)}\right) & \beta_j < 0 \end{cases} \tag{4.10}$$

where

$$u^{(i)} = \frac{1}{c} D^T (I - D a^{(i)}) + a^{(i)} \qquad (4.11)$$

superscript (i) represent the iteration number and subscript j is used to demonstrate the j-th entry in a vector. $\tau_1 = \frac{\lambda_1}{c}, \tau_2 = \frac{\lambda_2}{c}$ where c is a constant guaranteeing the convexity of surrogate function.

In CSR algorithm computational complexity is reduced due first, implementing the ideas from [45] on variational image restoration and reweighted l_1-optimization [46] to adaptively adjust the two regularization parameters τ_1, τ_2. Based on the conclusion in [45] the regularization parameter should be inversely proportional to signal to-noise-ratio (SNR); the same is proposed in [43]. Thus, the updating strategy is chose as the following:

$$\tau_1 = c_1 \frac{\sigma_\omega^2}{\sigma_a}, \tau_2 = c_2 \frac{\sigma_\omega^2}{\sigma_\gamma} \qquad (4.12)$$

where σ_ω^2 is noise variance, $\gamma = a - \beta$ and c_1, c_2 are two predefined constants (we usually set $c_1 < c_2$)

Second, estimating the recovered image using the work [47]

$$X^{(i+1)} = \tilde{S}\left((1 - \delta)X^{(i)} + \delta Y\right) \qquad (4.13)$$

where $\tilde{S} = D \circ S \circ R$ denotes the projection onto the regularization constraint set and

$$(1 - \delta)X^{(i)} + \delta Y = X^{(i)} + \delta(Y - X^{(i)}) \qquad (4.14)$$

is the operator implementing the idea of iterative regularization. Where δ is a small positive number controlling the amount of noise going back to the iteration.
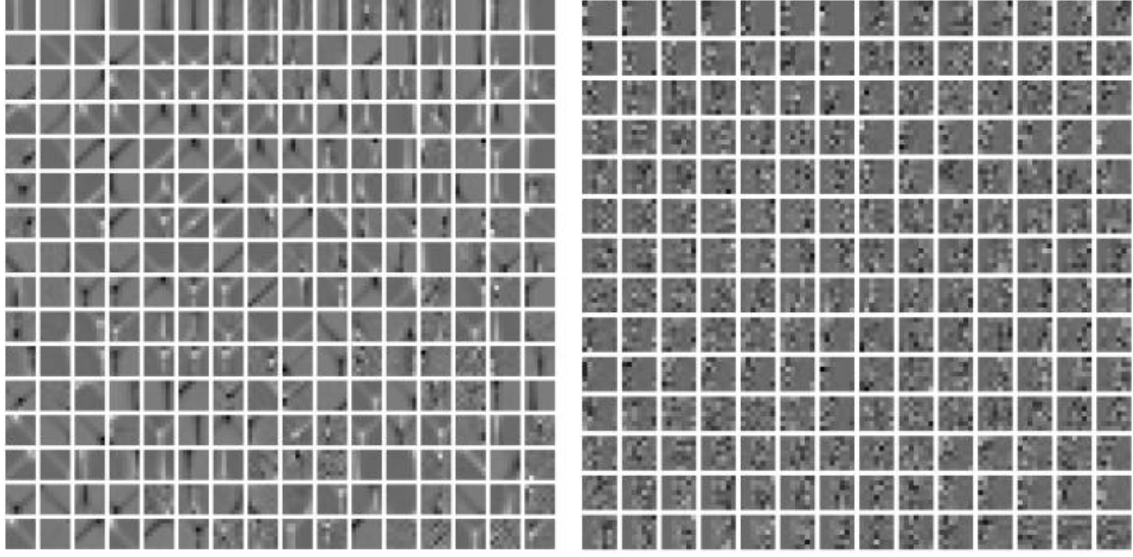
Figure 4.3. Comparison of learned dictionaries from the test image *D*34 between a) K-SVD (*K* = 256) and b) CSR (only four out of 64 sets of dictionaries is displayed). [1]

4.1.2. Bayesian Interpretation of CSR Denoising

A Bayesian interpretation of the CSR denoising algorithm is investigated in this section. The relationship between sparse representation and Bayesian denoising is well reviewed before. Based on equivalence between variational and Bayesian image restoration, the extension of the connection between dictionary-based to clustering-based framework is described here.

CSR is developed based on the idea that the centroid of *K* clusters (β_k) are as the peer hidden variables to sparse coefficients a. This concept can be used to justify the importance of resolving the location-related uncertainty associated with images. In order to do this, the following maximum a posterior (MAP) estimation problem is formulated.

$$(a, \beta) = \arg \max_{a, \beta} \log P(a, \beta | \mathbf{Y}) \tag{4.15}$$

The above equation can be rewritten using the Bayesian formula using likelihood and prior distributions:

51

$$(a, \beta) = \arg\max_{a,\beta} \log P(\boldsymbol{Y}|a, \beta) + P(a, \beta) \tag{4.16}$$

The first term can be formulated to (4.17) using the degradation model $\boldsymbol{Y} = \boldsymbol{I} + \boldsymbol{W}$ and the second term using statistical modeling to (4.18).

$$P(\boldsymbol{Y}|a, \beta) = \frac{1}{\sqrt{2\pi\sigma_\omega^2}} \exp(-\frac{1}{2\sigma_\omega^2} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2) \tag{4.17}$$

$$P(a, \beta) = P(\beta|a)P(a) = P(\gamma|a)P(a) \tag{4.18}$$

where $\gamma = a - \beta$ shows the deviation from each cluster. At this point, the clustering-based differential prediction can be defined as another level of sparse coding strategy so γ is approximately independent from a. Choosing to model both a and γ by i.i.d. Laplacian distribution, (4.18) can be rewritten as:

$$P(a, \beta) = \prod_i \frac{1}{\sqrt{2\sigma_a}} \exp(-\frac{\parallel a_i \parallel_1}{\sigma_a}) \times \prod_k \prod_i \frac{1}{\sqrt{2\sigma_\gamma}} \exp(-\frac{\parallel a_i - \beta_k \parallel_1}{\sigma_\gamma}) \tag{4.19}$$

And finally to solution to a, β can be reached at:

$$a, \beta = \arg\min_{a,\beta} \parallel \boldsymbol{Y} - \boldsymbol{D}a \parallel_2^2 + \frac{2\sqrt{2}\sigma_\omega^2}{\sigma_a} \sum_i \parallel a_i \parallel_1 + \frac{2\sqrt{2}\sigma_\omega^2}{\sigma_\gamma} \sum_k \sum_i \parallel a_i - \beta_k \parallel_1 \tag{4.20}$$

which is equivalent equation to (4.6) by choosing $\lambda_1 = \frac{2\sqrt{2}\sigma_\omega^2}{\sigma_a}, \lambda_2 = \frac{2\sqrt{2}\sigma_\omega^2}{\sigma_\gamma}$.

4.2. Wavelet Preprocessing

In low-dose CT images, the structural details and information is still available, but the noise often blurs the edges or confuses some of the detailed characteristics. To improve the performance of the sparse representation we want to use a preprocessing method to magnify important structures and details of images in order to learn a more accurate adaptive dictionary.

One of the most used conventional methods to denoise the medical images is wavelet denoising since it preserves edges and important image structures. Wavelet denoising is usually done by thresholding the wavelet coefficients, thus, it is really important how to choose the coefficients to serve the best for edge preserving while keeping the frequencies that the algorithm is working based on them. The global threshold can be calculated by

$$\lambda = \sigma \sqrt{2logN} \tag{4.21}$$

where σ is the noise vaiance and N the size of image. The global threshold is not really of interest since the result is an over-smoothed image. To overcome the smoothing problem and finding the optimal threshold, the *square root balance-sparsity norm* approach is used. First, a thresholds array t, which contains uniformly distributed values between 0 and 1 is defined. The t array is used to define two curves: the percentage of L2-norm recovery (the measure of the energy loss after the denoising process using the value in t) and the percentage of the relative sparsity (the number of resulting zero coefficients in the denoised image). The intersection of two curves is t_{opt} and *the square root balance-sparsity norm* threshold is defined using equation (4.24),

$$\lambda = \sqrt{t_{opt}} \tag{4.22}$$

By introducing this balance to the system the wavelet denoising preserves the detailed information and at the same time reduces the unwanted noise which makes the clustering step difficult and less accurate. Fourier spectrum of the noisy image (Man) and clustering step (dictionary leaning) input in CSR and our proposed method, joint Wavelet and CSR (WCSR), are compared in Figure 4.5. The figure illustrates the preservation of the high frequencies which are the selected features for clustering and dictionary learning corresponding to the images in Figure 4.4.
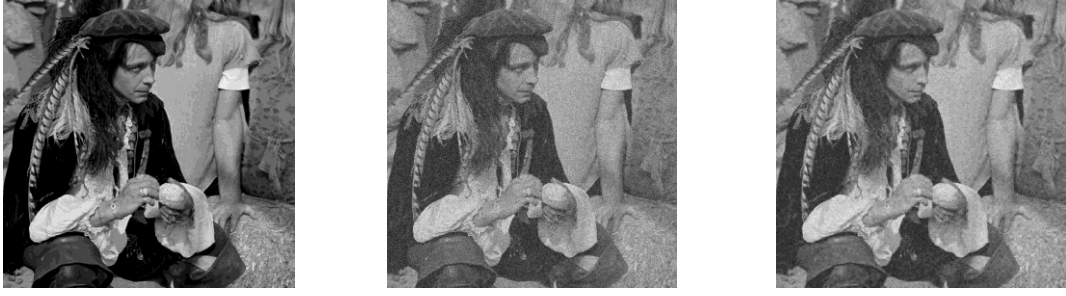
<center>(a)   (b)   (c)</center>

Figure 4.4.   a) noisy image (Man), b) clustering based sparse representation (CSR) and c) proposed joint wavelet and clustering sparse representation (WCSR)

<center>(a)   (b)   (c)</center>

Figure 4.5.   Fourier spectrum of the a) noisy image (Man), b) clustering based sparse representation (CSR) and c) proposed joint wavelet and clustering sparse representation (WCSR)

As it is shown in the image high frequencies are retained better than CSR and the algorithm preserves the important information and edges at the same time.

4.3. Single Image Noise Level Estimation

In the iterative solution to the double header l_1-optimization problem using surrogate function, there are two regularization parameters τ_1, τ_2 defined which are inversely proportional to signal-to-noise – ratio (SNR).

<center>54</center>

$$\tau_1 = c_1 \frac{\sigma_\omega^2}{\sigma_a} \ , \tau_2 = c_2 \frac{\sigma_\omega^2}{\sigma_\gamma} \tag{4.23}$$

where σ_ω^2 is the noise variance and $\gamma = a - \beta$ and c_1, c_2 are constants. The algorithm uses these regularization parameters to update the cluster centers so it is really important to have a precise estimation of the noise variance. We have deployed a new algorithm to improve the noise variance estimation based on the sole noisy image [48]. Utilizing a patch based noise level estimation, the weak patches from a single noisy image are selected. The selection is based on the gradients of the patches and their statistics and then PCA is used to do the noise estimation from the selected patches. It is a fast and accurate method to estimate the variance of the noise in the lower variances (less than $\sigma = 10$) and improves the performance of estimation and denoising algorithm.

4.4. Proposed Algorithm

Utilizing the repeated patterns especially at the edges of medical images, a new image model based on sparse representation and clustering is proposed to denoise the low-dose CT images. In order to denoise the low-dose CT images, first we take advantage of the remaining structural details and information in the image and then the blurred edges and some of the detailed characteristics will be recovered. Sparse representation and an adaptive dictionary can be utilized to remove the noises effectively. A preprocessing method is proposed to improve the performance of the sparse representation and to learn a more accurate adaptive dictionary. Higher sparsity is reached by grouping the similar patches and therefore a denoising algorithm with high performance is achieved.

In order to extract important information and preserve edges which both are significant details for diagnosis in CT images, wavelet denoising as a preprocessor and sparse representation based on dictionary learning and clustering is proposed. As the results show in Figure 4.5 the proposed algorithm

55

has maintained and improved the high frequency information and while preserving and improving the detailed information and edges as well. Improved detailed information of the image helps to learn an adaptive dictionary with a reduced number of atoms, so fewer numbers of clusters are needed. This reduces the computational complexity of the algorithm as well. The algorithm will have the following structure with two dependent loops. In the first step we have to learn the dictionary which will be done by wavelet analysis and clustering and we repeat that for k times. The next step will be using the designed dictionary to denoise the image which will be done for l times. The initialization contains the definition of number of clusters with significant effect on the denoising process. The noise level estimation will be done in the second step each time to help the accurate update of τ_1, τ_2. The pseudo-algorithm of the proposed method and the algorithm flowchart is shown in Table 4.1 and Figure 4.6.

Table 4.1. Pseudo-Algorithm of the proposed method

Objective: Approximation of the solution to the denoising problem

Given: The input signal $\boldsymbol{y}$ (training signal)

Initialize: $\boldsymbol{I}_{den} = \boldsymbol{y}$, Number of clusters;

Main Iteration: $k = k + 1$

- **Wavelet and Clustering Step:**
  - Wavelet feature extraction and denoising;
  - Feature extraction using High Pass Filter (HPF);
  - K-means clustering;
  - Dictionary learning through PCA;
- **Sparse representation and noise estimation Step:**
  - Set $l = 1$, Begin Iterative regularization;
  - Single image noise σ_ω^2 estimation;
  - Sparse representation and estimation of τ_1, τ_2 through equation (7);
  - Centroid estimate update β_k;
  - Denoised Image update, $\boldsymbol{I}_{den}$;
  - Set $l = l + 1$;

Output: $\boldsymbol{I}_{den}$

The iteration is fixed at 3 iterations for the outer loop and 3 iterations for the inner loop and number of clusters is imported by the user. It is necessary to do the noise level estimation at the cluster update stage, since the clusters are closely dependent on the amount of the estimated noise. The single image noise estimation is used for lower noise variances to have an accurate estimation of the noise at each iteration before updating the cluster centers.

```
                    ┌─────────────────┐
                    │  Initialization │
                    └─────────────────┘
                            │ set k=1
                    ┌─────────────────┐
                    │ Wavelet feature │
                    │ extraction &    │
                    │ denoising  /HPF │
                    └─────────────────┘
                            │
                    ┌─────────────────┐
                    │ K-means Clustering│
                    │ / Dictionary Learning│
                    │ (PCA)           │
                    └─────────────────┘
                            │ set l=1
                    ┌─────────────────┐
                    │ Iterative       │
                    │ Regularization/ │
                    │ Noise Estimation│
                    └─────────────────┘
                            │
                    ┌─────────────────┐
                    │ Sparse          │
                    │ Representation/  │
                    │ Centroids Update│
                    └─────────────────┘
                            │
                    ┌─────────────────┐
                    │ Update the      │
        if l<3      │ denoised image  │   if l=3
                    └─────────────────┘
                            │ if k=3
                      Denoised Image
```
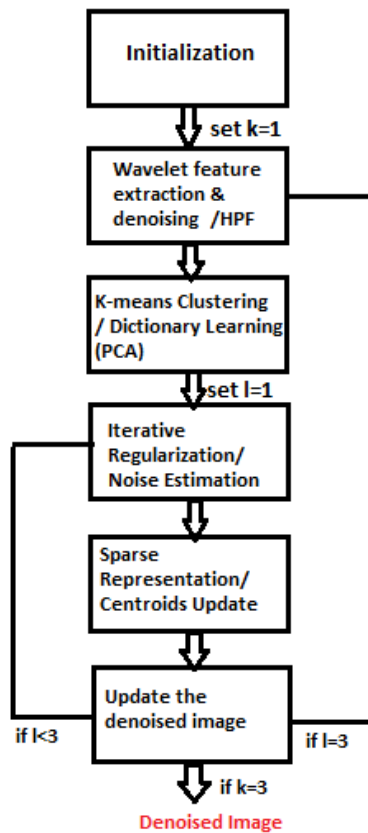
Figure 4.6.   Flowchart of the proposed method

4.5. Image Enhancement via Image fusion

Preforming the denoising algorithm on the low-dose medical images results in some blurring and loss of details, due to the similar texture at the edges. Hence, the proposed method is followed by an image

57

enhancement method based on fusion to highlight the detailed textures and edges. The selection of the enhancement method is based on the high performance of fusion of images from different modalities or instruments in different applications such as medical imaging, computer vision and etc.

Image fusion is expressed as the process to combine several images or some of their features together to form a single image. Image fusion can be done at different levels of information representation. Abidi and Gonzalez in [49] distinguished four different levels of information representation in fusion which are signal, pixel, feature and symbolic levels. The main intention of image fusion in the field of medical imaging is to provide an easier and enhanced interpretation for human observation. Therefore, the perception of the fused image is of paramount importance when evaluating different fusion schemes. There should be some requirements reserved on the fusion results. First, the fused image should contain all relevant information in the input images as closely as possible. Second, the process of fusion should not cause any artifacts or inconsistencies, which may result to distraction or misleading the human observer, or any subsequent image processing step. And finally the fused image should surpass irrelevant features and noise. In pixel level fusion, the input images are combined without having any pre-processing step. These algorithms can be very simple like image averaging or very complex like principal component analysis (PCA). Pyramid based image fusion can be categorized based on the domain of fusion either the spatial domain or the transform domain. After generating the fused image, there is a possibility to process further and extract the features of interest.

4.5.1. Wavelet Transform Fusion

There are several 2D wavelet-based techniques for image fusion proposed in the literature such as [50], [51]. In all wavelet based image fusion algorithms the wavelet transforms W of the two registered input images $I_1(x, y)$ and $I_2(x, y)$ are computed and these transforms are combined using some kind of fusion rule ϕ (Figure 4.7). Then, fused image $I(x, y)$ is reconstructed using the inverse wavelet transform W^{-1}.

$$I(x,y) = W^{-1}\left(\phi\left(W(I_1(x,y)), W(I_2(x,y))\right)\right) \qquad\qquad (4.24)$$

Comparing wavelet transform fusion technique to pyramid based fusion technique; first, directional information is provided in the wavelet transform while the pyramid representation doesn't introduce any spatial orientation in the decomposition process; second, blocking effects usually happens in significantly different regions in the process of pyramid based image fusion. But wavelet based fusion results is clear of such artifacts; and finally the signal-to-noise ratios of the output images in wavelet fusion is better than images generated by pyramid image fusion when the same fusion rules are used.
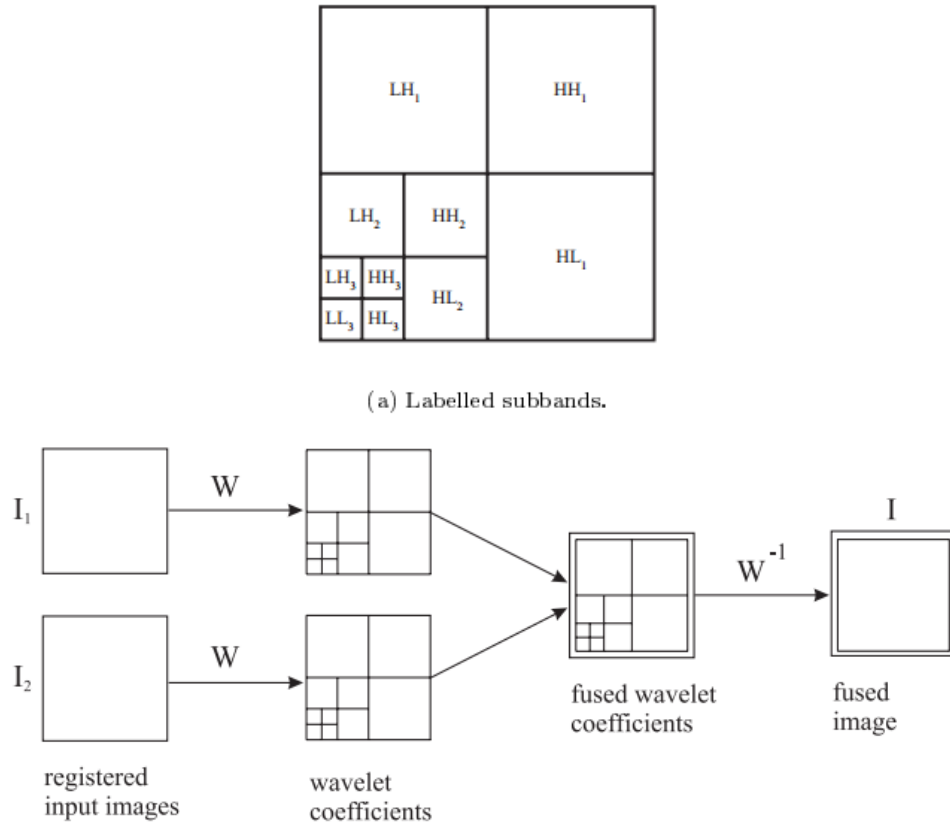

(a) Labelled subbands.

Figure 4.7. a) Labelled Subbands, b) Fusion of the wavelet transforms of two images.

Two dimension wavelet transform is obtained by filtering and down-sampling in the horizontal and vertical directions which result in four subbands. By indication of horizontal frequency first and vertical frequency second, the subbands are as followings: high-high (HH), high-low (HL), low-high (LH) and low-low (LL). The next step could be the application of same procedure on the low-low subband which results in a multi-resolution decomposition shown in Figure 4.7.a.

4.5.2. Pixel based Wavelet Transform Fusion

Multi-resolution fusion algorithm is inspired by the ability of human visual system to recognize local contrasts differences such as edges or lines. Using a specific fusion rule ϕ, all of the coefficients of the input image are combined in the pixel based wavelet transform fusion algorithm. Coefficients with the large absolute values are corresponding to the salient features of the image like corners and edges; therefore, designing a good fusion rule needs to extract this set of wavelet coefficients with large absolute value. An area based fusion algorithm extracts the maximum absolute value within a selected window and uses the measured value as the activity measure of the window central pixel. For recording the result of the selection rule, a binary decision map is created with the same size of wavelet transform [52]. The application of the weighted average coefficients of the image subband is another algorithm proposed to replace the binary decision map [53].

4.5.3. Proposed Image Enhancement based Wavelet Transform Fusion

Using the weighted average rule, the proposed fusion based enhancement method combines the second level (LL) wavelet coefficients of the noisy and the denoised image. The weighted average rule uses a normalized correlation between the two images subband combining with each other and reconstruction will be done on the resulting coefficients.

As explained earlier the denoising process affects the image with some blurring and loss of information at the edges. Therefore, by combining the second level (LL) wavelet coefficients of the noisy image which contains the important information about the texture not influenced by the noise, we can recover the lost or blurred information. The reason behind that is based on the fact that the high frequency subbands are mostly contacting the noise information, and then by selection of the LL subband we can eliminate the noise effect as much as possible. Then the proposed image enhancement algorithm shown in Figure 7.4.b. takes advantage of the highlighted detailed information and edges and an inverse wavelet transform conclude the algorithm with a high performance low-dose image denoising.

Chapter 5. Results and Discussion

5.1. Wavelet Preprocessing

In order to do the dictionary learning, the preprocessed image is filtered and high frequency features are extracted. K-means algorithm uses the extracted features to cluster the image and the dictionary is learned by obtaining the PCA of each clusters. As in low-dose CT images, some of the edges, lines and detailed information are blurred and have other artifacts due to noise, deploying a preprocessing method to help clustering the image more accurately will result in a better dictionary to represent the image. To evaluate the preprocessing method, Fourier spectrum of some of the famous image processing images is depicted compared with CSR method.

Figure 5.1 shows the preprocessed image and its Fourier transform of Lena, Barbara, Couple, and Man. The Symlet 4 wavelet with square root balance-sparsity norm thresholding technique and soft thresholding decision is implemented in this algorithm. As it is investigated by performing the Fourier transform, the high frequency information is well maintained in the preprocessed image which is the bone of the dictionary learning step. To overcome the smoothing problem square root balance-sparsity norm approach is used and it is shown in the results that there is no smoothing introduced in the preprocessing step. Another advantage that wavelet preprocessing introduces to the denoising algorithm is reducing the computational complexity by a decent factor since the noise is reduced in the output image. Therefore, by a reduced number of clusters, the more accurate dictionary can be learned.

5.2. Proposed Denoising Algorithm

As described previously, our method is proposed to improve the efficiency of the image denoising based on sparse representation and dictionary learning and design an image denoising technique for low-dose CT images. In order to obtain the higher efficiency and reduce the number of clusters, the preprocessing stage is introduced and the single image noise level estimation is developed to have an accurate update stage in the sparse representation step.

In order to have a reliable comparison between our proposed method, CSR and well-known K-SVD, we have developed these algorithms in the MATLAB software environment. The proposed image denoising method is tested on different images listed in Table 5.1. with different amount of noise variance. In all the experiments, the dictionary size for K-SVD algorithm is 64; CSR is tested with 30 and 64 dictionary atoms and WCSR method with just 30 atoms. The block size is 7 and $\lambda = 0.03$. The iterations for both loops in the proposed algorithm are selected to be 3. The comparison is based on the time of computations in minutes and quality in terms of peak signal to noise ratio (PSNR) in dB. For visual evaluation, Structural Similarity Index (SSIM) is used to perceptually compare the methods.
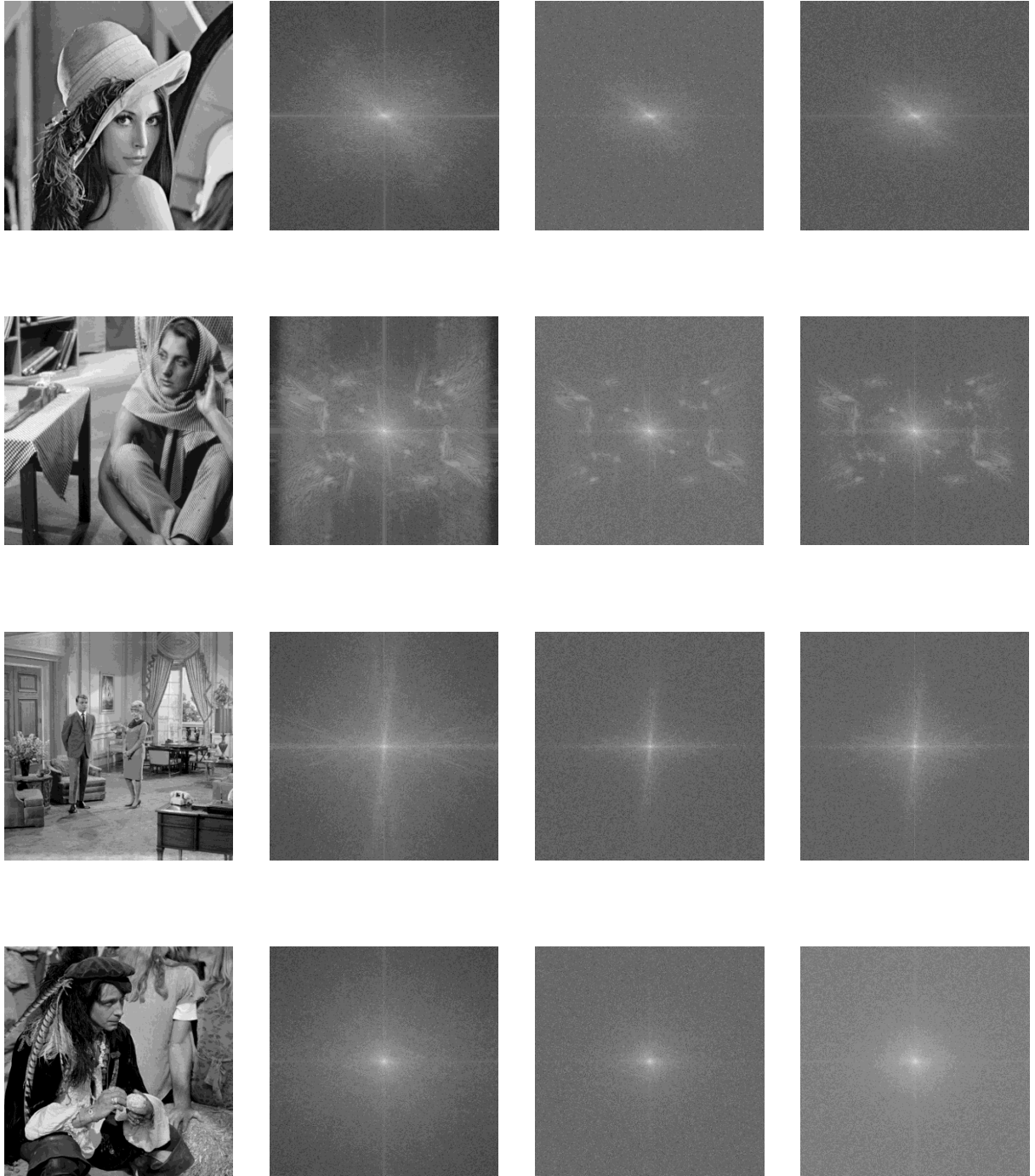
Figure 5.1. Displaying the result of Fourier transform of the preprocessing method. From left to right, first column: Original image, second column: Fourier transform of the original image, third column: Fourier transform of CSR and fourth column: Fourier transform of proposed method.

The PSNR results are demonstrated in Table 5.2 for images Lena, Barbara, Boat, Cameraman, Couple, Hill, Man and Peppers. It compares the K-SVD method [2] with CSR [1] and the proposed denoising method, WCSR. The SSIM results are indicated in Table 5.3 for images Lena, Barbara, Boat, Cameraman, Couple, Hill, Man and Peppers for CSR [1] along with the proposed denoising method, WCSR. Best quantitative results are represented in bold and it can be observed that the proposed denoising method achieves better performance in terms of PSNR and with fewer clusters, less than half of the previous number of clusters. In respect to SSIM, Table 5.3 demonstrates the improvement over CSR which implies better perceptual quality of the denoised images. The computation time for different images of size 512x512 with noise level of σ=20 for both proposed algorithm and CSR are summarized and compared in Table 5.1 which confirms that WCSR outperforms CSR.

Table 5.1. The computation time (MIN) comparison for the two algorithms for different noisy images σ=20

Algorithm	Lena	Man	Barbara	Couple
CSR	13.55 min	14.98 min	15.76 min	15.64 min
WCSR	**10.99 min**	**11.50 min**	**12.10 min**	**11.46 min**

All the experiments were done with an Intel core i7 processor of 1.88 GHz with 12 GB of memory, using MATLAB R2012a environment and under Windows operating system. The comparison is done using the CPU runtime; although CPU time is not exact measure, it gives a rough estimation of complexity.

Table 5.2. The PSNR (dB) results for different denoising methods. In each cell, the results of four denoising algorithms are reported. Top left: CSR(64 clusters) [1]; Top right: CSR(30 clusters) [1]; Bottom left: K-SVD [2]; Bottom right: WCSR(30 clusters)

σ/PSNR	5/34.15		10/28.13		15/24.61		20/22.11		25/20.17		30/18.59	
Lena512	38.74	38.75	35.90	35.88	34.20	34.19	32.96	32.94	31.98	31.96	31.16	31.14
	38.60	**38.82**	35.47	**35.96**	33.70	**34.28**	32.27	**33.06**	31.20	**32.07**	30.46	**31.23**
Barbara	38.43	38.42	**35.10**	35.05	**33.17**	33.13	**31.78**	31.73	30.66	30.66	29.72	29.72
	38.08	**38.47**	34.42	35.08	32.37	33.16	30.83	31.75	29.60	**30.66**	29.72	**29.72**
Boat	37.31	37.29	33.88	33.89	32.05	32.03	30.78	30.77	29.78	29.75	28.94	28.91
	37.22	**37.38**	33.64	**33.93**	31.73	**32.10**	30.96	**30.85**	29.28	**29.82**	28.43	**28.98**
C. Man	38.18	38.16	34.06	34.04	31.89	31.84	30.49	30.48	**29.48**	29.34	**28.64**	28.54
	37.85	**38.25**	33.70	**34.1**	31.44	**31.96**	29.96	**30.58**	28.93	29.47	28.07	28.57
Couple	37.41	37.40	33.95	33.93	32.00	31.94	30.60	30.59	29.52	29.51	28.62	28.63
	37.30	**37.46**	33.48	**33.95**	31.44	**32.03**	29.96	**30.67**	28.93	**29.62**	28.07	**28.74**
Hill	37.12	37.10	33.66	33.64	31.87	31.81	30.65	30.58	29.75	29.79	28.97	29.06
	37.03	**37.16**	33.38	**33.71**	31.46	**31.93**	30.15	**30.73**	29.19	**29.81**	28.37	**29.06**
Man	37.78	37.75	33.96	33.94	31.91	31.81	30.56	30.53	29.56	29.53	28.75	28.76
	37.50	**37.85**	33.55	**33.98**	31.44	**32.02**	30.05	**30.67**	29.02	**29.69**	28.23	**28.89**
Peppers	38.03	38.04	**34.64**	34.56	32.69	32.65	31.25	31.31	**30.14**	30.09	29.22	29.18
	37.78	**38.09**	34.25	34.59	32.22	**32.73**	30.77	**31.29**	29.69	30.13	28.82	**30.11**
Average	37.87	37.86	34.39	34.36	32.47	32.42	31.13	31.11	30.10	30.07	29.25	29.24
	37.67	**37.93**	33.98	**34.41**	31.97	**32.52**	30.61	**31.2**	29.48	**30.15**	28.77	**29.41**

Table 5.3. The SSIM results for different denoising methods. In each cell, the results of four denoising algorithms are reported. Top left: CSR(64 clusters) [1]; Top right: CSR(30 clusters) [1]; Bottom: WCSR

σ/PSNR	5/34.15		10/28.13		15/24.61		20/22.11		25/20.17	
Lena512	0.9448	0.9447	0.9145	0.9154	0.8915	0.8922	0.8725	0.8722	0.8564	0.8559
	0.9461		**0.9155**		**0.8950**		**0.8768**		**0.8597**	
Barbara	0.9645	0.9643	0.9413	0.9411	0.9212	0.9214	0.9027	0.9021	0.8841	0.8843
	0.9658		**0.9401**		**0.9233**		**0.9048**		**0.8864**	
Boat	0.9398	0.9399	0.8878	0.8877	0.8505	0.8491	0.8205	0.8204	0.7942	0.7941
	0.9429		**0.8912**		**0.8530**		**0.8237**		**0.7968**	
Cameraman	0.9603	0.9603	**0.9284**	0.9284	0.8995	0.8976	0.8754	0.8753	**0.8579**	0.8480
	0.9617		0.9286		**0.9025**		**0.8830**		0.8569	
Couple	0.9493	0.9493	0.9066	0.9065	0.8708	0.8693	0.8383	0.8385	**0.8289**	0.8097
	0.9510		**0.9087**		**0.8740**		**0.8420**		0.8144	
Hill	0.9418	0.9418	0.8851	0.8847	0.8405	0.8382	0.8037	0.7983	0.7732	0.7754
	0.9440		**0.8892**		**0.8443**		**0.8059**		**0.8581**	
Man	0.9532	0.9531	0.9058	0.9059	0.8646	0.8628	0.8306	0.8302	0.8013	0.8003
	0.9551		**0.9071**		**0.8702**		**0.8361**		**0.8067**	
Peppers	0.9549	0.9548	0.9259	0.9254	0.9030	0.9034	0.8828	0.8874	0.8647	0.8627
	0.9558		0.9226		**0.9054**		**0.8839**		**0.8652**	
Average	0.9510	0.9510	0.9119	0.9118	0.8802	0.8792	0.8533	0.8530	0.8325	0.8228
	0.9528		**0.9128**		**0.8834**		**0.8570**		**0.8430**	

In Figure 5.2 and 5.3, some examples of denoising of medical and natural images with additive white Gaussian noise, are displayed using CSR, K-SVD and the proposed methods. It can be seen that the proposed denoising scheme preserves the structures better and therefore has better perceptual image quality.

Figure 5.2. Displaying the result of proposed image denoising method. From left to right, First column: Original Image, Second column: Noisy Image ($\sigma = 20$), Third column: CSR denoising and Fourth column: WCSR denoising.
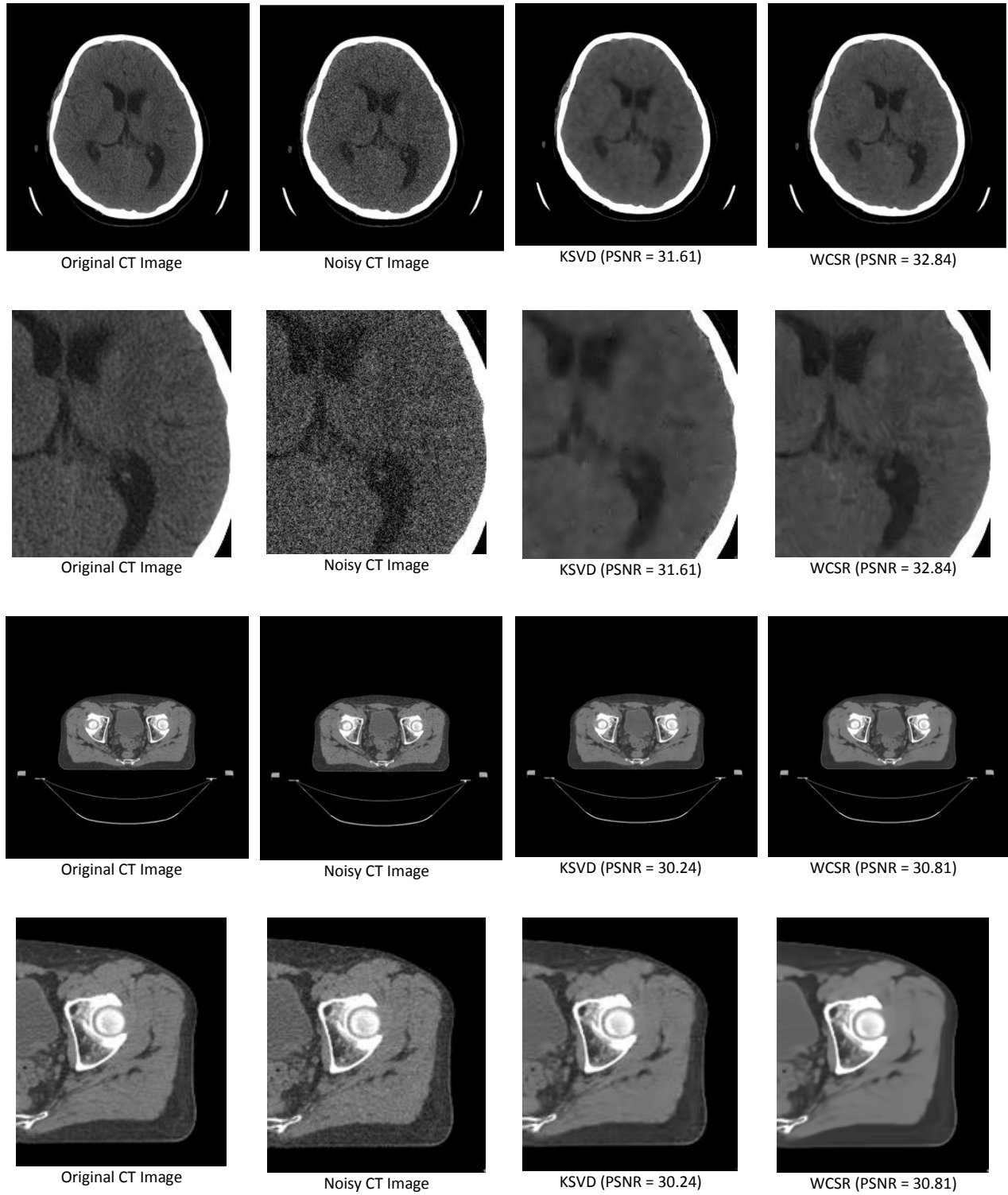
Figure 5.3. Displaying the result of proposed image denoising method on medical images. First column: Original Image, Second column: Noisy Image ($\sigma = 20$), Third column: KSVD denoising and Fourth column: WCSR denoising.

The images on the second and fourth columns represent a zoomed portion of the CT scans. As it can be observed, we have recovered most of the disappeared details in the noisy images in the denoised images and also our method outperforms K-SVD and CSR algorithms.

5.3. Proposed Low-Dose CT Image denoising

In this section we present the simulation results focusing on lower dose CT scans to demonstrate the performance of our algorithm. The performance of our algorithm is compared with K-SVD algorithm proposed [2]. The setting of experiments is the same as previous sections. For K-SVD algorithm, the experiments were performed with assumption of noise variance of 20 on low dose CT images and are based on 10 iterations.

For quality metric calculations and training purposes, we are provided with a set of registered matching high dose and low dose CT scans. With the information provided with the CT scans database, we can obtain which image is high dose and low dose using the dose reduction strategies explained in chapter two. In our experiments the phantom size and slice thickness is fixed; tube potential is set to 120 (kV), and position in the field of view is the same for all images. The tube-current-time product (mAs) is the only factor that has been changed to show the noise influence on the scans. The lower the mAs, the higher the noise will be; which result in lower dose CT images.

The proposed algorithm is tested on low-dose medical brain image as shown in Figure 5.4. This contains the results of the proposed denoising algorithm on a low-dose CT image and its zoomed portion. As it is shown we have recovered most of the edges and small structures in the image which is important for diagnosis. In the next experiment our proposed algorithm is implemented to denoise a 50% dose CT phantom (low dose) and we are using a 100% dose (high dose) CT phantom as the reference to calculate quality metrics. The images are registered and completely matched. As it can be seen in Figure 5.5, due

to similar structures throughout the phantom image, some edges are blurred in small extent which is due to the nature of any image denoising algorithm.

Furthermore, the proposed method is compared with conventional dictionary learning based K-SVD denoising method. The experiments were performed on low dose CT images with assumption of noise variance of 20. Even though quality-wise the results appear to be the same, the edges seem to be more blurred.
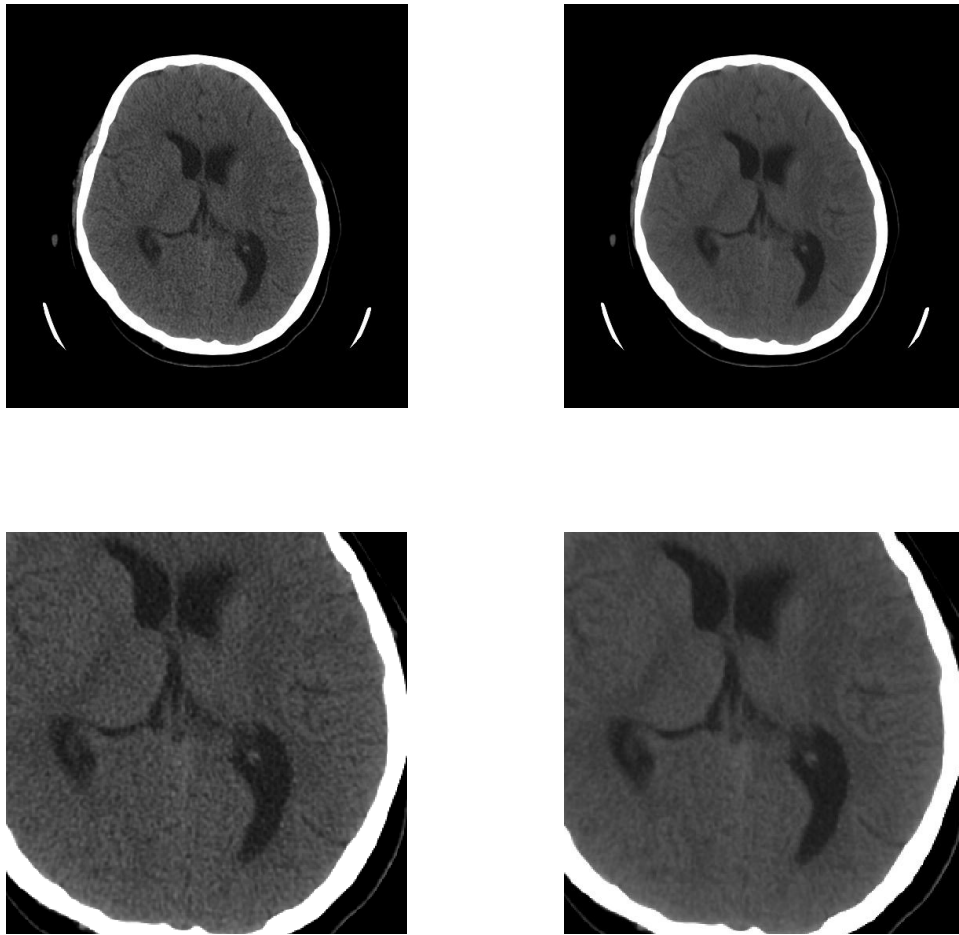

Figure 5.4. Top left: Low-Dose CT image Top right: Denoised image with WCSR, Bottom left: Zoomed part of Low-Dose image, Bottom right: Zoomed part of Denoised image with WCSR
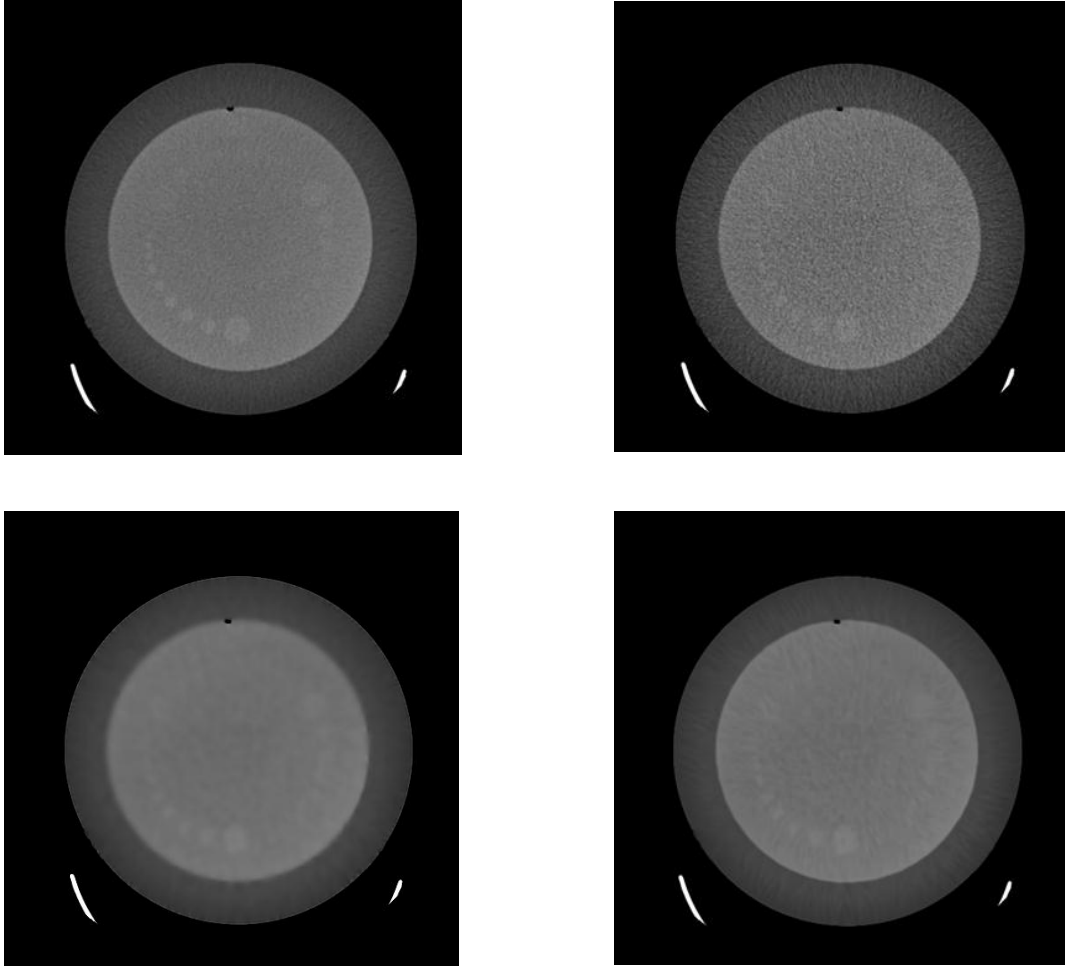
Figure 5.5. Top left: High-Dose Image, Top right: Low-Dose Image (PSNR = 27.54), Bottom left: Denoised image with K-SVD (PSNR = 28.68), Bottom right: Denoised image with WSCR (PSNR = 28.95)

5.4. Image Enhancement

In this section we have performed our proposed method of image enhancement on medical images in order to obtain an appropriate recovered image for medical diagnostic. As mentioned earlier, performing the denoising algorithm on the low-dose medical images may result in some blurring and loss of details. This blurring effect can cause loss of data that may lead to misdiagnosis. To enhance the image quality and overcome the mentioned issue, the denoised images are treated with the algorithm explained in section 4.5 as a post-processing step.

In our experiments, we can use both high dose and lose dose images in fusion step. In case that we have a set of high dose CT images we can use an average of them in the fusion step (method 1) and if we are not supported with these kinds of datasets we can use the noisy image itself and do the fusion step (method 2). In both approaches LL subband is used to extract the useful information not included in the denoised image. Fusion of the extracted structures with the denoised image concludes to a proper image containing all useful information for diagnosis.

For a better comparison we separate the denoised images of method 1 and 2 and compare them separately. The PSNR improvements are about 0.8 from in the first method and about 0.3 in the second methods. The reason behind that is although the image enhancement method used LL subband which is less affected by noise but still in the noisy image LL subband there are some artifacts remaining. Contrast map of the denoised image using the proposed method is similar to the contrast map of high dose image shown in Figure 5.7. The red window on each plot shows the chosen window for the purpose of demonstration of specific parts in the image. In general, the whole contrast distribution shows the improvement over the noisy image and how close we could get to the high dose image contrast distribution which is very close in this case. Figure 5.6 shows the results of denoised image using WCSR method enhanced by the proposed method using method 1 and 2. Table 5.4 summarizes the quality metric (PSNR) of both methods and the improvements comparing to the CSR.

Table 5.4. Denoised image enhancement results using the proposed method

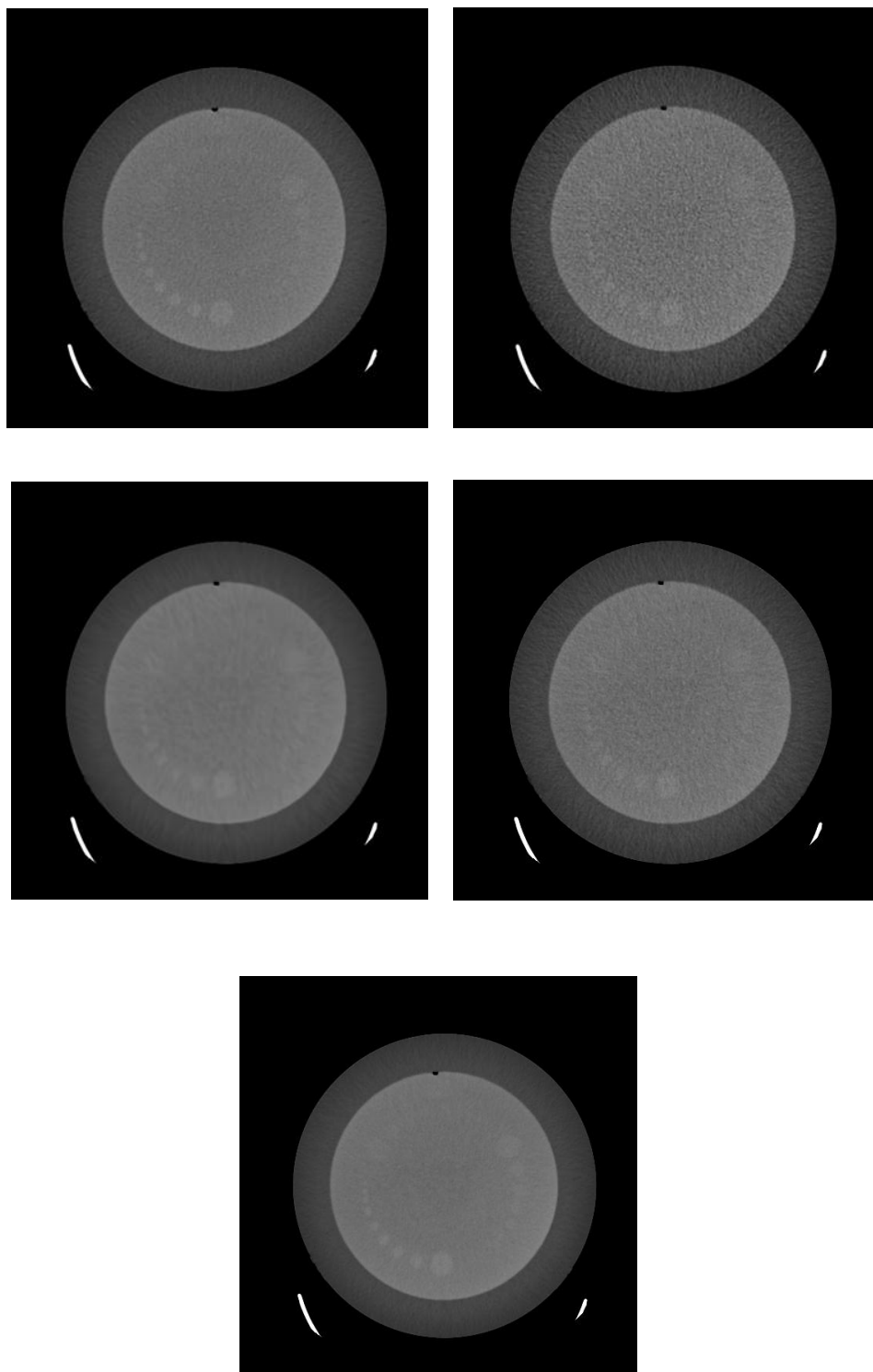| Method	PSNR (dB) before enhancement	PSNR (dB) after enhancement
Denoised using method 1	28.95	29.75
Denoised using method 2	28.95	29.25

Figure 5.6. First Row a)High-Dose Image, b) Low-Dose Image (PSNR = 27.54), Second Row: a) Denoised image with WSCR (PSNR = 28.95) and b) Denoised image with WCSR and enhancement method 1. Third row: Denoised image with WCSR and enhancement method 2.
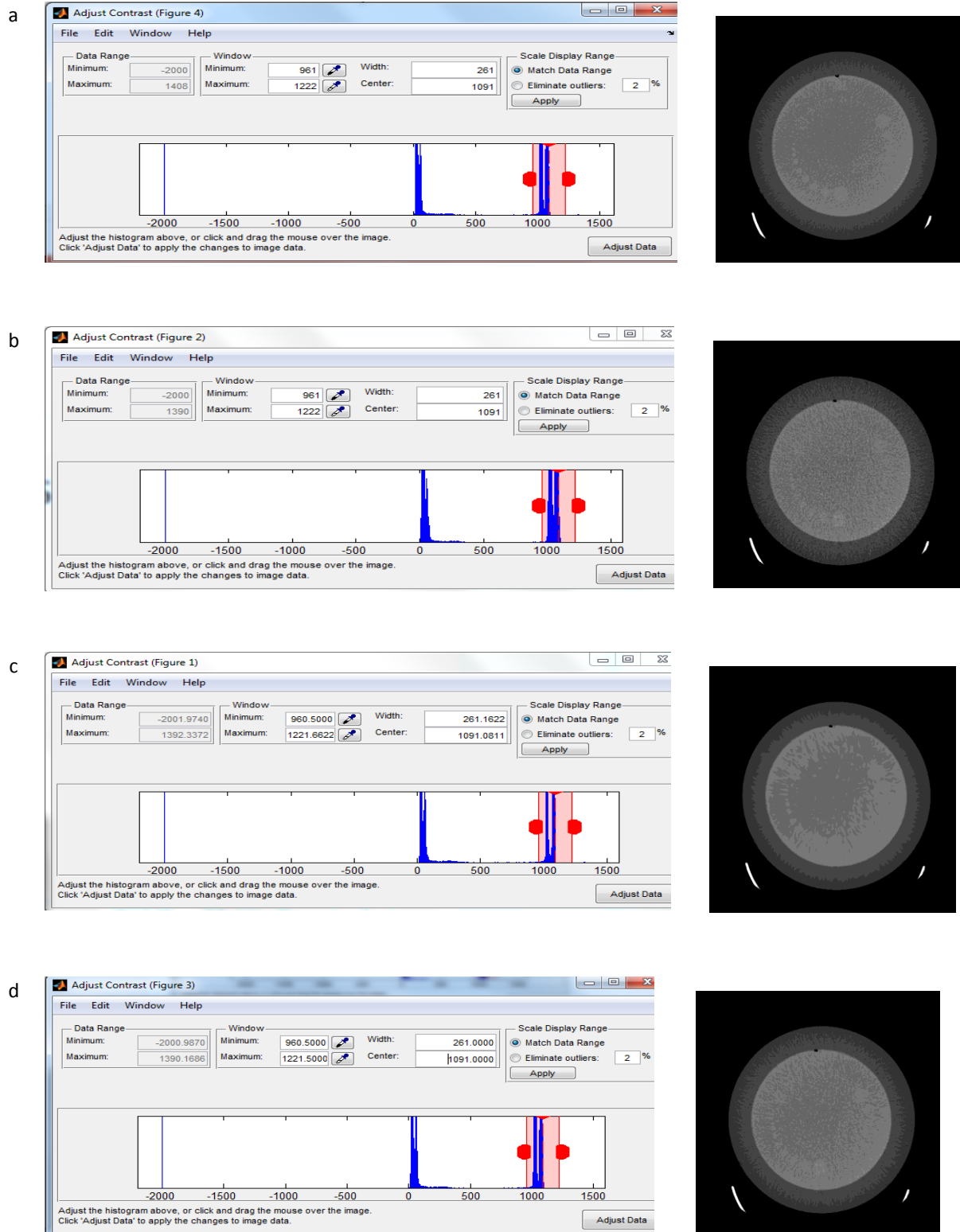
Figure 5.7. a) High dose Image, b) Low dose Image, c) Denoised image with WCSR and d) Denoised image with WCSR and enhancement

Chapter 6. Conclusion and Future Works

6.1. Wavelet Preprocessing Method

A preprocessing step based on the wavelet analysis was introduced in order to learn an accurate dictionary for the denoising algorithm which is based on sparse representation and dictionary learning. The proposed method is used to extract the suitable features for clustering step in the algorithm. Clustering step (K-means) utilizes the features extracted in preprocessing step which is high frequency information in this thesis and PCA helps to create the dictionary atoms from each clusters. In order to retain as much as high frequencies as possible from the noisy image, wavelet based method is used. Square root balance-sparsity norm thresholding technique is used in the wavelet analysis step to

preserve the detailed information and at the same time reduces the unwanted noise. Our proposed preprocessing method, by including most of the high frequency information, eliminates the need for having too many clusters to learn the dictionary and provides a more efficient and accurate dictionary learning. The results demonstrate that the utilization of the proposed method is a suitable technique to provide the necessary feature for dictionary learning as compared to other methods and it reduces the necessity of having too many clusters and therefore improves the computational cost by a good scale. As future improvements, it is possible to develop an adaptive strategy depending on the training data that maximizes the extraction of the useful features based on the image denoising application.

6.2. Proposed Image Denoising Algorithm Based on Joint Wavelet and Sparse Representation

In this thesis, we presented an image denoising algorithm based on the joint wavelet and sparse representation/dictionary learning. The first step as explained in the previous section is based on the utilizing wavelets to prepare the image for clustering and dictionary learning while retaining the important structures and preserving the edges. After the image is processed we take advantage of the repeated patterns especially at the edges in the image; and therefore an image model based on sparse representation and clustering is used for denoising. In this model, by grouping the similar patches higher sparsity can be obtained. Utilizing the remaining structural important information in the image in the presented model; a high performance denoising algorithm is achieved. The experimental results show that the performance of our algorithm is better than CSR algorithm in terms of both PSNR and SSIM with lower computational complexity.

The proposed method and clustering-based sparse representation are highly dependent on the initial number of the clusters in the algorithm. To overcome this problem in the future work, it is desired to design a new adaptive sparse representation via dictionary learning and clustering to automatically estimate the number of clusters. A gap statistics measurement can be used to find the optimal number of clusters in order to reduce the computational complexity and time consumption caused by the extra unnecessary clusters.

6.3. Proposed Low-Dose CT Image Denoising Algorithm Based on Joint Wavelet and Sparse Representation

In this thesis, we presented a main improvement in the proposed joint wavelet sparse representation denoising algorithm to make it suitable for denoising the low-dose CT images. Calculating the noise level in Low-dose CT images is targeted here. In order to accurately update the clusters at each iteration, a new noise level estimation is deployed to the algorithm to improve the performance especially for lower noise variances. It also confirms that our denoising algorithm works well for CT images, improving the quality of the image both with respect to PSNR and visually for diagnosis. The medical images are under ROC analysis review for evaluation and future improvements.

6.4. Image Enhancement via Fusion

In the proposed method, we solved the problem of fusion based image enhancement of the denoised images. The proposed fusion based enhancement method combines the second level (LL) wavelet

coefficients of the Low-dose/High-dose image and the denoised image utilizing the weighted average rule. By fusing the low dose/high dose CT images and the denoised image, we achieved an improvement in the quality and recovered some of the blurred edges and important details for diagnosis in the images. Taking advantage of any available high dose database and taking an average of them result in a better enhancement. This method is straightforward and easy to implement. In both approaches LL subband is used to extract the useful information not included in the denoised image. The LL subband is the best option to choose since noise is usually affecting high frequency information and therefore choosing a low frequency subband can eliminate the noise disturbance in the image.

Designing an adaptive fusion algorithm for different CT image categories can be a huge progress in the future. Each category of CT scans has their own specific characteristics that can be exploited in order to adapt the fusion method to the sensitive feature for diagnosis.

References

[1] Weisheng Dong, Xin Li, Lei Zhang, and Guangming Shi, "Sparsity based image denoising via dictionary learning and structure clustering," in *CVPR*, pp. 457-464, 2011.

[2] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing,* vol. 54, no. 12, pp. 4311-4322, 2006.

[3] M. Bethesda, "Ionizing radiation exposure of the population of the United States," National Council on Radiation Protection and Measurements, NCRP Report No.93, 1987.

[4] A. C. Kak and Malcolm Slaney, "Principles of Computerized Tomographic Imaging," *IEEE Press,* 1988.

[5] P.J. La Riviere, Junguo Bian, and P.A. Vargas, "Penalized-likelihood sinogram restoration for computed tomography," *Medical Imaging, IEEE Transactions on,* vol. 25, no. 8, pp. 1022-1036, Aug. 2006.

[6] Tianfang Li, Xiang Li, JingWang, JunhaiWen, Hongbing Lu, Jiang Hsieh, and Zhengrong Liang, "Nonlinear sinogram smoothing for low-dose x-ray ct. Nuclear Science," *IEEE Transactions on,* vol. 51, no. 5, pp. 2505-2513, Oct. 2004.

[7] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 12, no. 7, pp. 629-639, 1990.

[8] D.L. Donoho, "Denoising by soft-thresholding," *IEEE Trans. Inform. Theory,* vol. 41, no. 3, pp. 613-627, 1995.

[9] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. on Image Proc.,* vol. 15, no. 12, p. 3736–3745, December 2006.

[10] Antoni Buades, Bartomeu Coll, Jean-Michel Morel, "A non-local algorithm for image denoising," in *CVPR*, vol. 2, pp 60-65, 2005.

[11] P. Chatterjee and P. Milanfar, "Clustering-based denoising with locally learned dictionaries," *Image Processing, IEEE Transactions on,* vol. 18, no. 7, p. 1438–1451, 2009.

[12] J. Stelmark, "http://www.profstelmark.com/index.html," [Online], July 2014.

[13] Brenner DJ, Hall EJ. , "Computed tomography--an increasing source of radiation exposure.," *N Engl J Med.,* vol. 357, no. 22, p. 2277–2284, Nov 29, 2007.

[14] Cynthia H McCollough, Andrew N Primak, Natalie Braun, James Kofler, Lifeng Yu, Jodie Christner , "Strategies for reducing radiation dose in CT.," *Radiol. Clin. North Am.,* vol. 47, no. 1, pp. 27-40, 2009.

[15] Richard L. Morin, PhD; Thomas C. Gerber, MD; Cynthia H. McCollough, PhD, "Radiation Dose in Computed Tomography of the Heart," *Journal of American Heart Association,* no. 107, pp. 917-922, 2003.

[16] "The measurement, reporting and management of radiation dose in CT," *American Association of Physicists in Medicine,* no. 96, 2008.

[17] Haaga JR. , "Radiation dose management: weighing risk versus benefit," *AJR Am J Roentgenol,* vol.

177, no. 2, p. 289–291, Aug 2001 .

[18] Lifeng Yu, XIe Liu, Shuai Leng, James M Kofler, Juan C RAmirez-Giraldo, Mingliang Qu, Jodie Christner, Joel G Fletcher & CYnthia H McCollough, "Radiation dose reduction in computed tomography: techniques and future perspective," *Future Medicine,* vol. 1, no. 1, pp. 65-84, 2009.

[19] Joachim Weickert, "Theoretical Foundation of anisotropic diffusion in image processing," *In Theoretical Foundations of Computer Vision,* pp. 221-236, 1994.

[20] Chang, S.G., Bin Yu, Vetterli, M. , "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing,* vol. 9, pp. 1532 - 1546 , Sep 2000.

[21] L. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D,* vol. 60, pp. 259-268 , 1992.

[22] Guo W and Yin W, " EdgeCS: Edge Guided Compressive Sensing Reconstruction," Rice CAAM, 2010.

[23] R.Rubinstein, M.Zibulevsky, and M.Elad, "Double Sparsity: Learning Sparse Dictionaries for Sparse Signal Approximation," *IEEE Transaction on SIgnal Processing,* vol. 85, no. 3, pp. 1553-1564, 2010.

[24] Dominik Bartuschat, "A Parallel patch-based approach for the reduction of quantom noise in CT images," Lehrstuhl für Informatik,Master's Thesis, 2009.

[25] Catte F, Lions PL, Morel JM, Coll T, "Image Selective Smoothing and Edge Detection by Nonlinear Diffusion.," *SIAM J Numer Anal,* vol. 29, pp. 182-193, 1992.

[26] Alessandra Stagliano, Gabriele Chiusano and Curzio Basso, "Learning Adaptive and Sparse Representations of Medical Images," Universita`di Genova, Genova, 2010.

[27] T. Strohmer and R. Heath, "Grassmanian frames with applications to coding and communications," *in Applied and computational Harmonic Analysis,* vol. 14, p. 257–275, 2004.

[28] David Donoho and M. Elad, "Optimally sparse representations in general (non-orthogonal) dictioaries via l1 minimization," *Proc.Nat.Acad.Sci,* p. 2197–2202, 2002.

[29] S. G. Mallat and Z. Zhifeng, "Matching Pursuits With Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing,* vol. 41, no. 12, pp. 3397-3415, 1993.

[30] Y. C. Pati, R. Rezaiifar and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers,* vol. 1, pp. 40-44, 1993.

[31] I. T. Jolliffe, Principal component analysis, ebrary, Inc, 2002.

[32] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan and G. R. G. Lanckriet, "A Direct Formulation for Sparse PCA Using Semidefinite Programming," *SIAM Review,* vol. 49, no. 3, p. 434, 2007.

[33] V. Deepu, S. Madhvanath and A. G. Ramakrishnan, "Principal component analysis for online handwritten character recognition," *Proceedings of the 17th International Conference on Pattern Recognition,* vol. 2, pp. 327-330, 2004.

[34] P. J. B. Hancock, A. M. Burton and V. Bruce, "Face processing: Human perception and principal components analysis," *Memory & cognition,* vol. 24, no. 1, pp. 26-40, 1996.

[35] D. Mishra, R. Dash, A. K. Rath and M. Acharya, "Feature Selection in Gene Expression Data Using Principal Component Analysis and Rough Set Theory," *Biomedical and Life Sciences,* pp. 91-100, 2011.

[36] "Department of Cybernetics, Center for Machine Perception," [Online]. Available: http://cmp.felk.cvut.cz/˜hlavac, July 2014.

[37] H. Zou, T. Hastie and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,* vol. 15, no. 2, pp. 265-286, 2006.

[38] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society.Series B (Methodological),* vol. 58, no. 1, pp. 267-288, 1996.

[39] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 67, no. 2, pp. 301-320, 2005.

[40] I. T. Jolliffe, N. T. Trendafilov and M. Uddin, "A Modified Principal Component Technique Based on the LASSO," *Journal of Computational and Graphical Statistics,* vol. 12, no. 3, pp. 531-547, 2003.

[41] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. on Image Processing,* vol. 16, no. 8, p. 2080–2095, 2007.

[42] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE 12th International Conference on Computer Vision, pp 2272–2279*, 2009.

[43] E. J. Candes, J. K. Romberg, and T. Tao,, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory,* vol. 52, no. 2, p. 489–509, 2006.

[44] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics,* vol. 63, no. 1, p. 1–38, 2010.

[45] N. Galatsanos and A. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Transacactions on Image Processing,,* vol. 1, no. 3, p. 322 – 336, 1992.

[46] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourieranalysis and applications,* vol. 14, no. 5, p. 877–905, 2008.

[47] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Modeling and Simulation,* vol. 4, no. 2, p. 460–489, 2005.

[48] Xinhao Liu, Masayuki Tanaka and Masatoshi Okutomi,, "Noise Level Estimation Using Weak Textured Patches of a Single Noisy Image," in *ICIP*, pp 665 – 668, 2012.

[49] M.A.Abidi, R.C.Gonzalez, "Fusion of multi-dimensioanl data using regularization," *Data Fusion in Robotics and Machine Intelligence.,* pp. 415-455, 1992.

[50] Koren, I. ; Laine, A. ; Taylor, F., "Image fusion using steerable dyadic wavelet transform," in *International Conference on Image Processing*, Washington, DC, 1995.

[51] Laure J. Chipman; Timothy M. Orr; Lewis N. Graham, "Wavelets and image fusion," in *SPIE*, 1995.

[52] H. Li, B. S. Manjunath and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graphical Models and Image Processing,* vol. 57, no. 3, pp. 235-245, 1995.

[53] Burt, P. J. and Kolczynski, R. J., "Enhanced image capture through fusion," in *Fourth international conference of computer vision*, Los Alamitos, California, 1993.

List of Publications:

Samira Ghadrdan, Javad Alirezaie, Jean-Louis Dillenseger, *and* Paul Babyn, "Low-dose Computed Tomography Image Denoising based on Joint Wavelet and Sparse representation", accepted in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2014, Chicago, Illinois, USA*.