Pedestrian Detection in Depth Images Using Framelet Regularization

Yan-Ran Li, Shiqi Yu*, and Shengyin Wu
College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, 518060, P. R. China.
*Corresponding author E-mail: shiqi.yu@szu.edu.cn.

***Abstract*—Pedestrian detection in depth images can reduce the effect of clothing in appearance and unpredictable illumination changes, and its applications include robotics and surveillance, etc. Due to the TOF depth images easily contaminated by noise, a parameter-free and framelet-based regularization approach is proposed to remove noise and preserve the object shape in depth images before feature detection and classifying. After noise removal, the histogram of depth difference (HDD) is utilized as a features descriptor and SVM with a linear kernel is adopted as a classifier. Experiments show that the framelet-based approach is adaptive and effective to denoise depth images and pedestrian detection in denoising depth images is feasible. The miss rate decreases from 9.1% of noisy images to 2.2% of denosing images at FPPW=10^{-4}.**

***Index Terms*—Pedestrian detection, depth images, tight framelet.**

I. Introduction

Pedestrian detection in images or video has attracted many researchers in computer vision community over the past few years, and its applications include robotics, entertainment, surveillance, care for the elderly and disabled, and content-based indexing [1]. Many techniques have been proposed and improved in terms of features, models, and general architectures [2].

The pedestrian detection system based on an overcomplete dictionary Haar wavelets was proposed to identify the important characteristics of the people class while ignoring noise present in the pixel-level representations [3]. Dalal and Triggs proposed the histograms of oriented gradients (HOG) as local feature descriptors, and adopted support vector machine (SVM) as a classifier of human detection [4]. Due to the weakness of the global or local feature descriptor in facing occlusions of pedestrian, part-based detection approach using edgelet features was proposed to handle this challenge [5]. Composing of local and global cues via a probabilistic top-down segmentation, the method was provided to detect pedestrian in crowded real-world scenes [6]. Wang *et al.* presented an approach capable of handling partial occlusion by combing HOG and local binary pattern (LBP) features descriptor.

Pedestrian detection in intensity images needs to face the challenges such as clothing in appearance, complex background, unpredictable illumination changes. However, it is still difficult to completely solve these problems. The depth camera provides a new way to tackle the above problems and captures the distance between objects and the camera. It is different from the traditional intensity camera and is not sensitive to light intensity. Krotosky *et al.* [7] used two color and two infrared cameras to obtain depth images and proposed a stero-based pedestrian detection approach. Multi-depth images captured by a multilayer laser scanner were fused to track pedestrian in urban environment [8]. Depth infirmations computed from a calibrated stereo camera and intensity images were combined together to detect pedestrian in [9]. Due to the efficiency of HOG as a local feature descriptor, histogram of depth difference (HDD) was inspired by HOG for pedestrian detection in depth images captured by time-of-flight (TOF) camera [10]. However, noise exists in the depth images by the TOF camera and causes the high miss rate for pedestrian detection. In order to overcome this problem, we propose a framelet-based method to remove the noise and reduce the miss rate.

The outline of the paper is as follows. A framelet-based method is proposed in Section II. Experiments are presented in Section III. Finally conclusions are given in Section IV.

II. Proposed Framelet-based Method

A. Observed Model for Depth Images

Pedestrian detection in depth images has been investigated [10]. However, the depth images shown in the first row of Fig. 1 are seriously contaminated by environment noises and measure errors, and the noise will disturb the features detected by local descriptors and reduce recognition rate for pedestrian detection. Thus, we need to remove the noise in depth images and propose a framelet-based denoising scheme for pedestrian detection. Let x and $n \in \mathbb{R}^{M\times N}$ be an original depth image and additive noise with image size of $M \times N$ pixels, respectively. The observed model of depth image can be presented as

$$y = x + n. \tag{1}$$

The goal of denoising in (1) is to retrieve original information from the observed data y, and this inverse process is an ill-posed problem. We utilize the framelet-based regularization model to make the problem into well-posed one and get satisfied results.

B. Framelet-based Regularization Model for Depth Images

Framelet system is the tight frame system with redundancy and the merit of multi-resolution analysis (MRA) [11], [12].

The piecewise linear tight framelet system in $L^2(\mathbb{R})$ is successfully applied to image processing [13], [14], [15] which has a low-pass filter τ_0 and two high-pass filters τ_1 and τ_2 as follows

$$\tau_0 = \frac{1}{4}[1,2,1],\ \tau_1 = \frac{\sqrt{2}}{4}[1,0,-1],\ \tau_2 = \frac{1}{4}[-1,2,-1]. \quad (2)$$

The filers τ_1 and τ_2 are the first-order and second-order differential operators, respectively. The one dimensional piecewise linear framelet operators $\{\tau_i\}_{i=0}^2$ can be extended to a two dimensional framelet system by tensor product. For $i,j \in \{0,1,2\}$, the coefficients of two dimensional framelet system are defined as

$$\tau_{i,j} := \tau_i^\top \tau_j, \quad (3)$$

where $\tau_i^\top$ is the transpose of τ_i, $\tau_{0,0}$ is a low-pass filter and others filters $\tau_{i,j}$ with $(i,j) \neq (0,0)$ are high-pass filters. Based on these filters, a geometry tight framelet system $\{h_\ell\}_0^{17}$ of $L^2(\mathbb{R}^2)$ with 18 filters was designed by Li *et al.* [14]. This geometry tight framelet system produces not only the first-order and second-order difference of an image in horizontal and vertical directions, but also in the $-\pi/4$ and $\pi/4$ directions. For example,

$$h_8 = \tfrac{1}{12}\begin{bmatrix} 0 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad h_9 = \tfrac{1}{12}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

are two second-order difference operators. The geometry tight framelet system $\{h_\ell\}_0^{17}$ is utilized in this paper to preserve edges of an image at local area and remove the noise in depth images.

Assume the matrixes $\mathcal{T}_L$ and $\mathcal{T}_\ell$ respectively represent the low-pass h_0 and framelet operators h_ℓ of the geometry tight framelet system $\{h_\ell\}_0^{17}$ in [14], and

$$\mathcal{T}_H = [\mathcal{T}_1^\top, \mathcal{T}_2^\top, \cdots, \mathcal{T}_{17}^\top]^\top.$$

According to the unitary extension principle of [12], $\mathcal{T}_L$ and $\mathcal{T}_H$ satisfy with

$$\mathcal{T}_L^\top \mathcal{T}_L + \mathcal{T}_H^\top \mathcal{T}_H = \mathcal{I}, \quad (4)$$

where $\mathcal{I}$ is an identity matrix. Γ defines a diagonal matrix as

$$\Gamma := \mathrm{diag}(\cdots, \gamma_l, \cdots), \quad \gamma_l \geq 0.$$

The framelet-based regularization optimal model is proposed as

$$\hat{x} = \arg\min_x \left\{ \frac{1}{2}\|x-y\|_2^2 + \|\Gamma \mathcal{T}_H x\|_1 \right\}, \quad (5)$$

where $\frac{1}{2}\|x-y\|_2^2$ is the fidelity term and $\|\Gamma\mathcal{T}_H x\|_1$ is the regularization term.

Assume the ℓ^{th} subband framelet coefficients

$$\tilde{x}_\ell = \mathcal{T}_\ell x, \ell = 1,2,\cdots,17,$$

and

$$\tilde{x}_H = [\tilde{x}_1^\top, \tilde{x}_2^\top, \cdots, \tilde{x}_{17}^\top]^\top = \mathcal{O}_H x.$$

Using the equation in (4), the optimal solution in (5) can be denoted by

$$\hat{x} = \mathcal{T}_L^\top \mathcal{T}_L y + \mathcal{T}_H^T \hat{\tilde{x}}_H, \quad (6)$$

where

$$\hat{\tilde{x}}_H = \arg\min_{\tilde{x}_H} \left\{ \frac{1}{2}\|\tilde{x}_H - \tilde{y}_H\|_2^2 + \|\Gamma\tilde{x}_H\|_1 \right\}, \quad (7)$$

with $\tilde{y}_H = \mathcal{T}_H y$. Let $\tilde{x}_{\ell,i,j}$ and $\tilde{y}_{\ell,i,j}$ be the coefficients of the ℓ^{th} subband at i^{th} row and j^{th} column of $\tilde{x}_H$ and $\tilde{y}_H$, respectively. Due to separability of every entry, the framelet-based optimal model in (7) can be decomposed into an optimal model of every $\tilde{x}_{\ell,i,j}$ as

$$\widehat{\tilde{x}}_{\ell,i,j} = \arg\min_{\tilde{x}_{\ell,i,j}} \left\{ \frac{1}{2}\left(\tilde{x}_{\ell,i,j} - \tilde{y}_{\ell,i,j}\right)^2 + \gamma_{\ell,i,j}|\tilde{x}_{\ell,i,j}| \right\}, \quad (8)$$

where $\gamma_{\ell,i,j}$ is the corresponding parameter in Γ. The optimal result for (8) is the well known soft thresholding operator [16] as

$$\widehat{\tilde{x}}_{\ell,i,j} = \mathrm{sign}(\tilde{y}_{\ell,i,j}) \max\{|\tilde{y}_{\ell,i,j}| - \gamma_{\ell,i,j}, 0\}. \quad (9)$$

The thresholding parameter $\gamma_{\ell,i,j}$ determines the sparsity of images and is estimated by the MAP estimator as

$$\gamma_{\ell,i,j} = \frac{\sqrt{2}\sigma_\ell^2}{\sigma_{y_{\ell,i,j}}}, \quad (10)$$

where σ_ℓ^2 is noise variance of the ℓ^{th} subband and $\sigma_{y_{\ell,i,j}}$ is the local signal variance. These two parameters are adaptively calculated in details as [15]. A framelet-based denoising algorithm is presented as follow.

Algorithm 1 (Framelet-based Denoising Algorithm):

1) Calculate $\tilde{y}_H = \mathcal{T}_H y$;
2) Calculate every $\gamma_{\ell,i,j}$ in (10);
3) Get $\widehat{\tilde{x}}_{\ell,i,j}$ by the thresholding operator in (9);
4) Reconstruct $\hat{x} = \mathcal{T}_L^\top \mathcal{T}_L y + \mathcal{T}_H^T \hat{\tilde{x}}_H$.

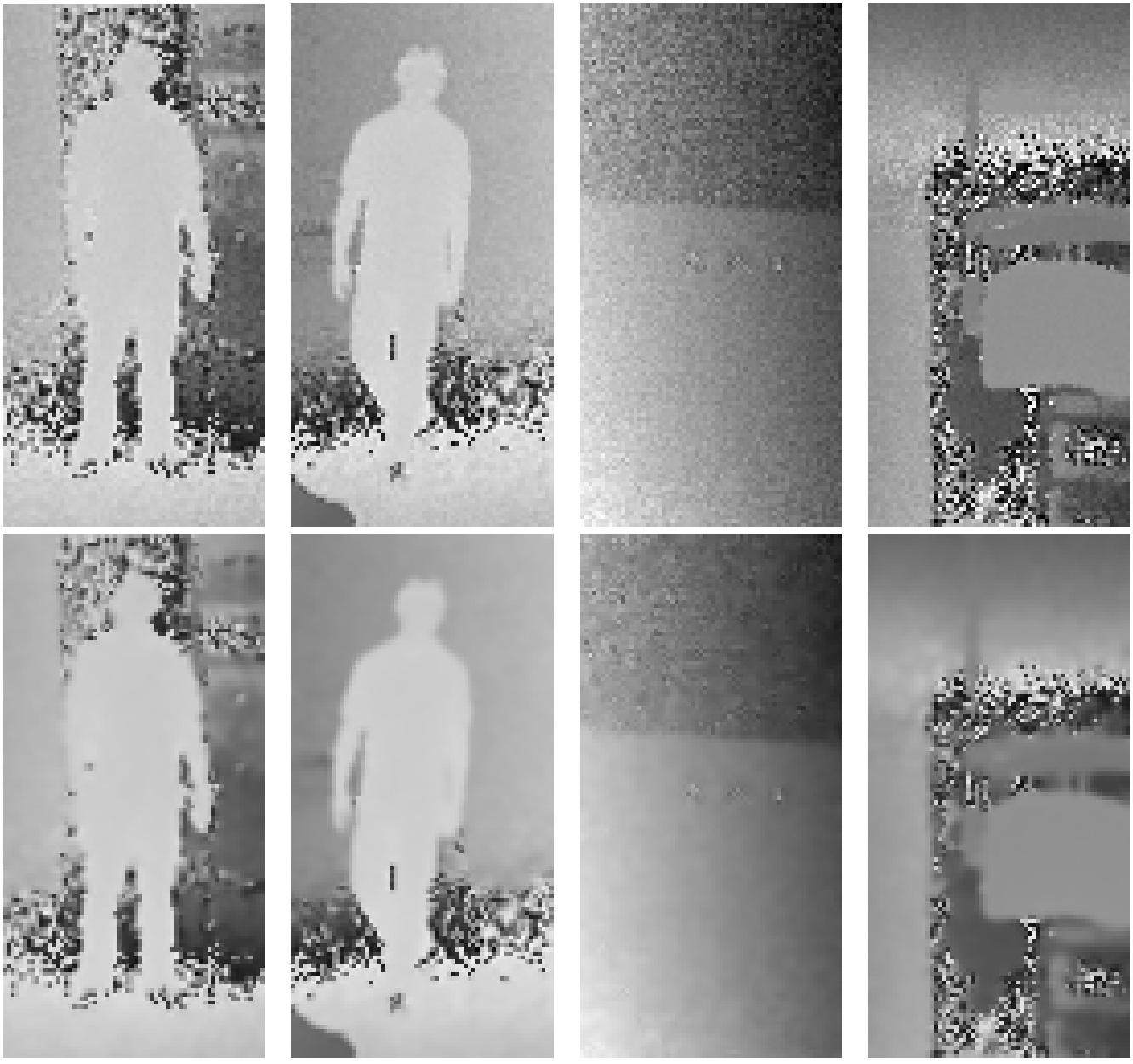
For pedestrian detection, the framelet-based denoising Algorithm 1 is applied to denoise depth images before features detection.

C. Descriptor and Classifier for Denoising Depth Images

1) HDD Descriptor: For detecting the features of denoising depth images, the histogram of depth difference (HDD) method in [10] is utilized as a local descriptor. Each depth image is dividend into overlapped blocks with four cells of size of 8×8 pixels, and any two adjacent blocks overlap half size of block. For a depth image with size of 64×128 pixels, there are $15 \times 7 = 105$ blocks for features detection.

Let $\hat{x}(i,j)$ be the value at i^{th} row and j^{th} column of the denoising image $\hat{x}$. At $(i,j)^{th}$ pixel of $\hat{x}$, we define

$$\begin{aligned} \Delta_h(i,j) &= \frac{\hat{x}(i+1,j) - \hat{x}(i-1,j)}{2} \\ \Delta_v(i,j) &= \frac{\hat{x}(i,j+1) - \hat{x}(i,j-1)}{2} \end{aligned} \quad (11)$$



(a) (b) (c) (d)

Fig. 1. The performance of framelet-based denoising Algorithm 1. The noisy depth images in the first row are captured by a TOF camera, and their corresponding denoising images are shown in the second row.

where $\Delta_h(i,j)$ and $\Delta_v(i,j)$ are the local variations in the horizontal and vertical directions, respectively. With these two components $\Delta_h(i,j)$ and $\Delta_v(i,j)$, the depth difference of HDD denotes by a magnitude $\mathcal{M}(i,j) = \sqrt{\Delta_h^2(i,j) + \Delta_v^2(i,j)}$ and an orientation $\theta(i,j) = \arctan(\Delta_v(i,j), \Delta_h(i,j))$. A histogram with orientation bins is used to describe statistical features of $\mathcal{M}(i,j)$ and $\theta(i,j)$ for each cell, and the bins are uniformly spaced over $[0°, 360°)$. The HDD features describe the distance variance in depth images and represent local geometrical structures, which will be fed into a classifier for pedestrian detection.

2) SVM Classifier: The support vector machine (SVM) is an outstanding classifier for many classification problems [2]. The SVM with a linear kernel in [4] is adopted in this paper as a classifier with two classes.

III. Experiments

The depth dataset is generated by a time-of-flight camera of Mesa Imaging AG, and the distances between objects and the camera rang from 0 to 5 meters. The depth dataset is composed of 4637 pedestrian (positive) and 56802 non-pedestrian (negative) images, which is divided into training and testing groups. The training group contains 3160 positive and 14199 negative

samples, and the testing group includes 1477 pedestrian and 56802 non-pedestrian images. This dataset is available at http://yushiqi.cn/research/depthdataset.

A. Denoising Depth Images

As shown in the first row of Fig. 1, the depth images are obviously noisy with the artifacts in the smooth area. We need to remove these artifacts and keep the edges in the depth images, so that the noise will not disturb feature detection by the HDD and discriminability by the SVM. Before feature detection, the framelet-based denoising Algorithm 1 is applied to depth images, and the four denoising results are presented in the second row of Fig. 1. Comparing the noisy images and denoising images in Fig. 1, the artifacts are effective to be removed and the edges of depth images are still preserved by our denoising Algorithm 1. This shows our Algorithm 1 can automatically threshold the geometry framelet coefficients by the sparsity formula in (10). Due to our parameter-free Algorithm 1, this makes our algorithm practical and attractive in real application of pedestrian detection.

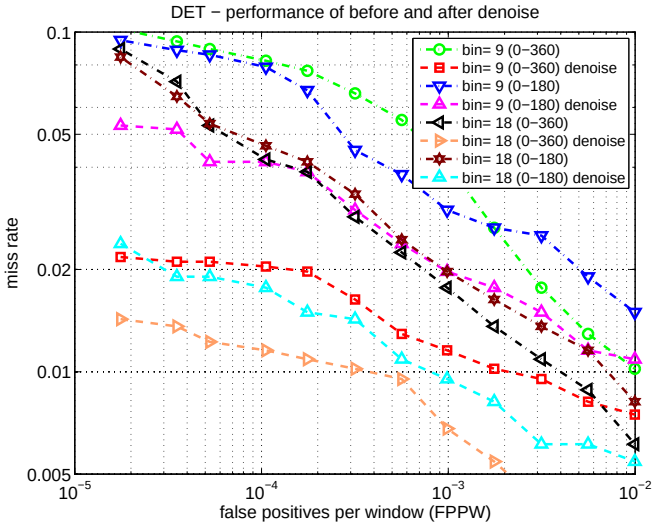


Fig. 2. The performance of pedestrian detection on noisy and denoising depth images.

B. Pedestrian Detection on Denoising Depth Images

The performances of pedestrian detection are compared on noisy and the denoising depth images by our framelet-based Algorithm 1. The experiments are carried out on two bin spaces $[0°, 360°)$ and $[0°, 180°)$ which are uniformly spaced into 9 or 18 bins. According to figure of miss rate with different cases shown in Fig. 2, the performance of pedestrian detection on denoising depth images is better than that on noisy depth images. For the case of bin space $[0°, 360°)$ with 9 bins, the miss rate of noisy images is 9.1%, but the miss rate of denoising images is greatly reduced to 2.2%. This shows that the framelet-based Algorithm 1 is effective to remove noise, attenuate noise effects for the feature descriptor HDD and classifier SVM, and decrease 6.9% miss rate for denoising depth images.

IV. Conclusions

The parameter-free denoising algorithm based on the geometry framelet system in [14] is employed to remove noise and preserve the object shape in depth images by shrinking the framelet coefficients. The HDD descriptor and SVM are respectively used to extract features and classify pedestrian problem in denoising depth images. The proposed framelet-based scheme is feasible and effective, and reduce the miss rate for pedestrian detection in noisy depth images.

Acknowledgements

This work is partly supported by Shenzhen Fundamental Research Program (Grant No. JC200903170431A, JC200903130300A, JC201005280432A) and China Postdoctoral Science Foundation (Grant No. 20090460532).

References

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[2] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[3] C. Papageorgiou and T. Poggio, "Trainable pedestrian detection," in *Proceedings of International Conference on Image Processing*, vol. 4, 1999, pp. 35–39.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[5] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proceedings of International Conference on Computer Vision*, vol. 1, 2005, pp. 90–97.

[6] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proceedings of International Conference on Computer Vision*, vol. 1, 2005, pp. 878–885.

[7] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, 2007.

[8] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine, "Pedestrian detection method using a multilayer laserscanner: application in urban environment," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 173–178.

[9] M. Rohrbach, M. Enzweiler, and D. M. Gavrila, "High-level fusion of depth and intensity for pedestrian classification," in *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, 2009, pp. 101–110.

[10] S. Wu, S. Yu, and W. Chen, "An attempt to pedestrian detection in depth images," in *Proceedings of the 3rd Chinese Conference on Intelligent Visual Surveillance*, 2011, pp. 97–100.

[11] R. Jia and Z. Shen, "Multiresolution and wavelets," in *Proceedings of the Edinburgh Mathematical Society*, vol. 37, 1994, pp. 271–300.

[12] A. Ron and Z. Shen, "Affine system in $L_2(R^d)$: the analysis of the analysis operator," *Journal of Functional Analysis*, vol. 148, pp. 408–447, 1997.

[13] J. F. Cai, R. H. Chan, and Z. Shen, "Simultaneous cartoon and texture inpainting," *Inverse Problems and Imaging*, vol. 4, no. 3, pp. 379–395, 2010.

[14] Y. R. Li, D. Q. Dai, and L. Shen, "Multiframe super-resolution reconstruction using sparse directional regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 7, pp. 945–956, 2010.

[15] Y. R. Li, L. Shen, D. Q. Dai, and B. W. Suter, "Framelet algorithms for de-blurring images corrupted by impulse plus Gaussian noise," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1822–1837, 2011.

[16] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.